Euskal Herriko Unibertsitatea / Universidad del País Vasco

Euskal Filologia Saila

Euskararen ezagutza-base lexikala: Euskal WordNet

Elisabete Pociello Irigoyenek

Euskal Filologian Doktore titulua eskuratzeko aurkezturiko

Tesia

Donostia, 2.007ko urria.

Euskal Herriko Unibertsitatea / Universidad del País Vasco

Euskal Filologia Saila

Euskararen ezagutza-base lexikala: Euskal WordNet

Elisabete Pociello Irigoyenek Eneko Agirre Bengoaren eta Izaskun Aldezabal Rotetaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Euskal Filologian Doktore titulua eskuratzeko aurkeztua.

Donostia, 2.007ko urria.

Lan hau Eusko Jaurlaritzaren ikertzaileak prestatzeko beka batekin (BFI02.114) egin dut.

Gauza ederra duk hitzak suntsitzea. Jakina, aditzen eta izenondoen artean zagok zaborrik gehien, baina izenak ere ehunka zeudek baztertzeko modukoak. Eta ez sinonimoak bakarrik; antonimoak ere bota daitezkek zakarretara. Azken batean, zertarako behar diagu hitz bat beste baten alderantzizkoa baizik ez bada? Hitz batek bere baitan zaramak kontrako esanahia. Har ezak "on", esaterako. "On" baldin badaukak, zertarako demontre behar duk "txar"? "Ez-on" berak balio izango likek berdin; eta, hobeki gainera, zeren eta kontrako zehatza baituk, bestea ez bezala. Edo, bestela, "on"en aldaera indartsuagoa behar baldin baduk, zer zentzu zeukak "bikain", "gailen" eta gisako hitz lauso eta alferrekoen soka hori guztia edukitzeak? "Pluson" hitzak ematen dik esanahia, edo "bikoizpluson" hitzak, are esanahi indartsuagoa behar baduk. Jakina, dagoeneko erabiltzen dizkiagu forma horiek, baina hizketaberriaren azken bertsioan forma horiek besterik ez duk izango. Azkenean, ontasunaren eta txartasunaren eremu osoa sei hitzek bakarrik beteko ditek; hitz bakar batek egiazki.

[...]

Hala ere, hire bihotzean hizketazaharrari atxikita jarraitzea hobetsi duk, haren zehaztasun-gabezia eta esanahien abardura alferrekoak gorabehera.

(George Orwell, *1984*. Tafalla: Txalaparta, 2007)

"Profirió", "rezonó", "masculló", "remarcó"... Ikusten gaztelaniaren ugaritasuna? Gu, berriz, hor gabiltza beti "esan zuen" eta "esan zuen". Aldatu egin nahi, eta "bota zuen" darabilgu. Edo gehienera ere, "bota zion". Horrela nola idatz daiteke bizitasun pixka batez?

Eta abar? Neuk ere botatzen nituen antzekoak. Oker nengoen: zeure hizkuntzaren ispiluan begiratu behar dituzu zeure ahulezia eta bertute estilistikoak, ez beste hizkuntza baten ispiluan.

Anjel Lertxundi (Berria, 2007-04-28)

Aitari eta Amari

Eskerrik asko!

Tesi hau egin ahal izateko, jende askoren laguntza izan dut, eta hauei guztiei eskerrak eman nahi nizkieke:

- IXA taldeko kide guztiei, lan hau aurrera eramateko eskaini didazuen laguntza guztiagatik, eta batez ere, niretzat ezezaguna zen hizkuntzalaritza konputazionalaren munduan sartzeko aukera emateagatik.
- Zuzendariei, Enekori eta Izaskuni, gauzak izugarri errazteagatik, eta berez astuna dena arin bihurtzen laguntzeagatik.
- Ehundaka hitzen adierak editatu, etiketatu eta epaitu dituzuenoi (Larraitz, Karmele, Eli, Mikel, Jone eta Ainara), tesi hau gure eztabaida "semantiko-filosofiko-soziologikoen" emaitza ere badelako.
- Olatzi, nire erruz egiten ari zarena utzi eta datu-basean gora eta behera jardun behar izan duzulako; beti laguntzeko prest!
- A German, per respondre amb molta paciència a totes les meves preguntes, i així fer-me practicar el català.
- *Emakunde*n, nirekin batera, ordu piiiiiila pasa dituzuen bulegokideei (Aitziber, Olatz, Ruben, Klara, Maxux, Kike, Mikel, landare "bionikoa"...); urte guzti hauetan, lanaz gain beste mila bizipen partekatu ditugulako. Aiii, landare "bionikoak" hitz egingo balu...
- IXA-bulego nagusiko bulegokideei, tesiko azkeneko txanpan nire txorakeriak jasateagatik. Ah! eta bulegoan dardoak jartzeagatik!
- Inguruan izan ditudan informatikari gajoei, eta, batez ere, txosten honek itxura txukuna izateko latexekin lagundu didazuenei (Oier, Gorka, Aitor Soroa, Maite...), nirekin izan duzuen pazientzia handiiiiiagatik.

- *Gym* taldetxoari (Aitzpea, Bertol, Klara, Larraitz eta Ruben), estresaren aurkako formula erakusteagatik (kirol pixka bat + bazkari/afari ugari + "katxondeo" asko = estres gutxiago).

- Nereari eta Montseri; Nereari bere masajitoengatik eta *Emakunde*ko iskanbilak beheko solairutik "konpartitzeagatik"; eta Montseri per reir (i fer-me reir) tant (beeeeeh!).

- Tesiaren aldapa gogorra igo nahian zaudeten ixakide guztiei; eutsi goiari!! nik egin badut, zuek ere egingo duzue-eta!!

- Lagunei, tesia utzi eta garagardo bat zuekin hartzera joateko aitzaki ezin hobea izan zaretelako. Hurrengo potea nire kontu!

- "Eli, baina zuk unibertsitatean zer egiten duzu?" galdera ehundaka aldiz egin didazuenei. Hurrengoan, tesia oparituko dizuet, behingoz uler dezazuen, edo ez...

- Senide guztiei, beti hor egoteagatik.

- Etxekoei, nire lana ondo ulertu ez arren, zuek izan zaretelako, hasiera hasieratik, lan honen bultzatzaile nekaezinak.

- Ilobei, zuekin nagoenean ezinezkoa delako tesiarekin gogoratzea.

- Bertoli, txostentzar hau zuzentzen hartu duzun lanagatik; bide luze honetan, egunero-egunero, eman dizkidazun animoengatik; eta bereziki, lanak eta aisialdiak bateragarriak izan BEHAR dutela erakusteagatik.

Eskerrik asko denoi!

Laburtzapenak

Euskaraz:

DBL: Datu-Base Lexikala
EBL: Ezagutza-Base Lexikala
EDBL: Euskararen Datu-Base Lexikala
ELK: Egitura Lexikal-Kontzeptuala
EusWN: Euskal WordNet
HAE: Hitz Anitzeko Esapidea
HAUL: Hitz Anitzeko Unitate Lexikala
HEB: Hiztegi-Ezagutza Basea
HM: Hautapen-Murriztapena
LNP: Lengoaia Naturalaren Prozesamendua

Ingelesez:

BNC: British Nationa Corpus
c2c: class-to-class
EuroWN: EuroWordNet
ILI: Inter-Lingual-Index
LCS: Lexical Conceptual Structure
MCR: Multilingual Central Repository
MRD: Machine Readable Dictionary
s2semf: sense-to-semantic field
s2s: sense-to-sense
w2c: word-to-class
w2semf: word-to-semantic field
w2w: word-to-word
WN: WordNet

Glosategia

analisi semantiko
Analisi semantikoaren helburua esaldiaren esanahia lortzea da, hau da, bere edukiaren errepresentazio kontzeptuala sortzea. Horretan, esaldiaren esanahia egitura formal baten bidez adierazi beharko da.

autohiponimia
EBL batean hiperonimoa eta hiponimoa forma berekoak direnean, baina adiera desberdinekoak, hots, polisemikoak.

datu-base lexikal (DBL)
Lexikoaren gainean biltzen den ezagutza mota gehienbat gramatikala denean (kategoria, azpikategoria, morfotaktika...), *datu-base lexikal* (DBL) terminoa erabiltzen da.

desanbiguazio/desanbiguatu
Anbiguotasuna gertatzen denean, testuinguruari begiratzen zaio hitz batek aukeran dituen interpretazioen artean egokiena zein den jakiteko. Testuinguru jakin horri ez dagokion interpretazioa kentzea ala dagokiona besterik ez uztea da desanbiguatzea.

Domeinu-ontologia (*Domain Ontology*)
EuroWordNeten eta *The Multilingual Central Repository*n (MCRn), *synset*ak domeinuen arabera antolatzen dituen ontologia.

eremu semantiko (*semantic field*)
Eremu semantikoak WordNeten fitxategi batzuk dira, non WordNeteko klase semantiko bakoitza jasota dagoen.

eskuratu/eskurapen
Informazioa *eskuratu* dugula diogu, metodo automatikoetan oinarrituz, corpuse(ta)tik behar dugun informazioa lortzen dugunean. Esate baterako, tesi-lan honetan corpusetan oinarrituz aditz batzuen hautapen-murriztapenak lortu ditugu.

etiketatze
Zenbait markaketa linguistiko, hala nola hitzei kode bereziak atxikitzea haien zenbait ezaugarri adierazteko; eta ezaugarriei egokitzen zaizkien kodeei *etiketa* esaten zaie. Etiketatzea zenbait kontu markatzeko erabiltzen da. Eta horregatik maila desberdinetako etiketatzeak daude. Tesi-lan honetan etiketatze semantikoaz arituko gara, hau da, etiketa semantikoak erabilita hitzen adiera zehaztuko dugu, hots, desanbiguatuko dugu.

EuroWordNet (EuroWN)
Ezagutza-base eleanitza da (Vossen, 1998), Europako zortzi hizkuntzatara zabaltzen dena (ingelesa, nederlandera, italiera, gaztelania, alemana, frantsesa, txekiera eta estoniera), eta WordNet (Miller, 1985; Fellbaum 1998a) EBLan oinarritzen dena.

EuSemcor
IXA taldea semantikoki eskuz etiketatzen ari den euskarazko corpusa, Euskal WordNeteko *synset*etan oinarrituaz.

Euskal WordNet (EusWN)
IXA taldea garatzen ari den euskarako EBLa, WordNeten, EuroWordNeten eta *The Multilingual Central Repository*ren (MCR) ildotik sortutakoa.

ezagutza-base lexikal (EBL)
Hitz eta adierei buruzko informazioa duten lexikoia da. EBLen ezaugarri garrantzitsuena herentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen dira-eta.

Goi-ontologia (*Top Ontology*)
EuroWordNet eta *The Multilingual Central Repository*ko (MCRko) wordnet ezberdinetan, gehien erabilitako *synset*ak oinarrizko ezaugarri semantikoen arabera sailkatzea ahalbidetzen duen ontologia.

hautapen-murriztapen (HM)
HMak dira hitz baten adiera batek testuinguruan izan ditzakeen agerkidetzak. Zerrenda hau osatzen dute klase semantiko batean dauden hitzek, hau da, adiera zehatz batekin osagai gisa ager daitezkeen hitz guztiak. Horrela bada, aditz batek, bere adieraren arabera, argumentu bezala har ditzakeen izenen klase semantikoa mugatu dezake.

hiperonimia
Unitate lexikoen arteko edukitze-erlazioa, orokorragotik espezifikoagora doana. Honen kontrakoa *hiponimia* da. Adib., hegazti hitza txori hitzarekiko hiperonimiako erlazioan dago.

hiperonimo
Beste hitz batekiko *hiperonimiako erlazioan* dagoen hitzaz esaten da. Adib., hegazti hitza txori hitzaren hiperonimoa da.

hiponimia
Unitate lexikoen arteko edukitze-erlazioa, espezifikoagotik orokorragora doana. Honen kontrakoa *hiperonimia* da. Adib., txori hitza hegazti hitzarekiko hiponimiako erlazioan dago.

hiponimo
Beste hitz batekiko *hiponimiako erlazioan* dagoen hitzaz esaten da. Adib., txori hitza hegazti hitzaren hiponimoa da.

hitz anitzeko esapide (HAE)
Edozein hitz-konbinazio adierazteko; lexikalizatuak nahiz ez lexikalizatuak (Alegria *et al*, 2004).

hitz anitzeko unitate lexikal (HAUL)
Lexikalizaturiko hitz anitzekoak (Alegria *et al*, 2004).

hiztegi ezagutza-base (HEB)
HEBek hiztegietatik erauzitako informazioa jasotzen dute. Erauzitako informazioen artean, EBLetan bezala, hemen ere, adieren hierarkiak dira aipagarriak.

ikasi/ikasketa automatiko
Makinari emandako datu egokietan oinarrituz eta hauen gainean teknika estatistiko konplexuak aplikatuz, makinak *ikasi* egiten du; ikasketa honen ondorioz, gai da datu berriei buruz erabakiak hartzeko. Erabaki hauen zuzentasuna ikaste-prozesuaren egokitasunaren araberakoa izango da, noski; ikaste-prozesuaren egokitasuna, era berean, erabiltzen diren teknika estatistikoen eta ikasteko erabilitako datuen kopuruan eta egokitasunean datza.

informazio-erauzketa
Testuetatik edo hizketatik informazio adierazgarria automatikoki ateratzea.

interfaze
Gizakiaren eta makinaren arteko elkarrekintzan laguntzeko sistema.

***Inter-Lingual-Index* (ILI)**
Inter-Lingual-Index (ILI) honen bitartez, EuroWordNeten eta *The Multilingual Central Repository*n (MCRn) hizkuntza guztietako wordnetak lotuak daude.

ILI-record
*Inter-Lingual-Index*ean (ILIan) *ILI-record*ak daude, eta hauetako bakoitza WordNeteko *synset* bati dago lotua.

interpretazio semantiko
Testuingurua kontuan hartu gabe, esaldiaren esanahi abstraktua lortzen duen analisi-fasea. Forma logiko baten bitartez adierazten da esaldiaren esanahia.

Lengoaia Naturalaren Prozesamendua (LNP)
Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari Lengoaia Naturalaren Prozesamendua (LNP) esaten zaio, eta, batez ere, erabiliko diren teknika informatikoei erreparatzen dio: algoritmoak, konpilatzaileak, estrategiak, etab.

lexikalizazio
Morfema-segida bat unitate lexikal bilakatzen den prozesua, eta esanahi eta funtzioaren ikuspuntutik hitz bakar bat bezala funtzionatzen duena.

lexikoi
LNPren arloan informazio lexikalaren biltegiei edota hiztegiei erreferentzia egiteko erabiltzen den terminoa.

***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. Ezagutzabase honek EuroWordNeten eredua jarraitzen du. Bestalde, MCRk bost hizkuntzetako wordnetekin egiten du lan: euskara, katalana, ingelesa (Princetoneko WordNetaren 1.5, 1.6, 1.7 eta 1.7.1 bertsioekin), italiera eta gaztelania.

MRD (*Machine Readable Dictionary*)
Euskarri magnetikoan gordetzen den hiztegia. Hiztegi elektronikoa.

Oinarrizko Kontzeptu (*Base Concept*)
EuroWordNeten eta *The Multilingual Central Repository*n (MCRn) harreman semantikoen kopuru handiena duten *synset*ak dira. Gainera, hierarkian goi aldeko *synset*ak dira, eta EBL hauek osatzen duten wordnet guztietan agertuko dira.

ontologia
Mundu errealaren kontzeptualizazioak dira, hitzekin izendatzen ditugun kontzeptuak modu hierarkikoan antolatuta, mundu errealari buruzko inferentziak egiteko gaitasuna dutenak.

SemCor
WordNeteko *synset*ekin eskuz etiketatuko ingeleseko corpusa.

sinonimia
Termino *sinonimoen* arteko erlazioa.

sinonimo
Esaldi berean, beronen esanahia aldatu gabe, elkartruka daitezkeen terminoez (hau da, esanahi bera dutenez) esaten da.

synset (synonym set)
Synset bakoitza kontzeptu lexikal bati dagokio, eta hau osatuko duten hitz-multzoek kategoria berdinekoak eta testuinguru bereetan truka daitezkeenak dira.

urre-patroi (*goldstandard*)
Automatikoki eskuratutako emaitzak ebaluatu ahal izateko, eskuz sortzen diren emaitza prototipikoak.

variant
*Synset*a osatzen duten ale lexikalei *variant* deitzen zaie, eta, *synset* berean dauden *variant*ak sinonimoak dira.

WordNet
Kontzeptuen artean hainbat motatako harreman semantikoak ezarriz (hiperonimia, hiponimia, sinonimoa...) egiten diren ingeleseko sare semantiko ezagunenetakoa da (Miller, 1985; Fellbaum, 1998a).

wordnet
WordNeten (Miller 1985; Fellbaum, 1998a) oinarrituta garatu den edozein hizkuntzetako EBLari buruz hitz egiteko erabiltzen da. Hala, *WordNet* terminoarekin, ingeleseko wordnetari egingo zaio erreferentzia, eta *wordnet* terminoak aurretik zer hizkuntzetakoa den adierazia izan beharko du.

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

I. KAPITULUA

Tesi-lanaren aurkezpen orokorra

I.1 Gaiaren kokapena eta motibazioa

Lan hau Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldearen barruan kokatu behar da. IXA taldeak hogei urte inguru daramatza Lengoaia Naturalaren Prozesamenduan (aurrerantzean LNP) lanean. Arlo zabal horren barruan, euskararen gaineko ikerketa aplikatua da gure xede nagusia, eta helburu horrekin, orain arte, morfologia (Agirre *et al.*, 1992; Aduriz *et al.*, 1994, besteak beste) eta sintaxia (Aduriz *et al.*, 1998a; Aranzabe *et al.*, 2003; Aldezabal *et al.*, 2001b, besteren artean) landu ditugu batez ere. Arlo hauetan lan handia egiteke dagoen arren, hurrengo aurrerapauso garrantzitsua semantika jorratzea da.

Semantika beharrezkoa da hainbat ataza konputazionaletan aurrera egin ahal izateko, batez ere, hizkuntzaren ulermena beharrezkoa den atazetan (egitura sintaktikoen desanbiguazioan, hitzen adieren desanbiguazioan, anaforaren ebazpenean eta itzulpen automatikoan, adibidez). Arrazoi horregatik, IXA taldean dagoeneko hasiak gara ezagutza lexiko-semantikoaren ikasketan murgiltzen. Lan horietako batzuk jadanik doktoretza-tesiak sortu dituzte, eta beste lan batzuk, berriz, egin bidean dauden doktoretza-tesiak dira:

- Euskarako aditzen azpikategorizazioaren azterketa, hiztegi elebakar batean (Arriola, 2000; Arriola *et al.*, 1999) edo corpusetan oinarrituta (Aldezabal *et al.*, 2001b; Agirre *et al.*, 2004).

- Euskarako aditzen alternantzien eta klase semantikoen azterketa (Aldezabal, 2004).
- Hitzen adieren desanbiguazioa (Martínez, 2005).
- Erlazio lexiko-semantikoen gauzatze sintaktikoa (Lersundi, 2005).
- Ezagutza lexiko-semantikoa informazio-erauzketan (Ansa *et al.*, 2005).

Lan hauei guztiei etekin handiagoa aterako litzaieke erabilitako baliabide eta deskribapen linguistiko guztiak lexikoi berean egongo balira. Lexikoiak informazio lexikala jasotzen duten biltegi egituratuak dira. LNPren helburu nagusia, zentzu zabalean, hizkuntza automatikoki eskuratzea edo ulertzea da. Hori lortu ahal izateko, hizkuntza horren hiztegiaren ezagutza sakona jasota duen biltegi baten beharra dago, hots, lexikoi bat. Horrela, LNPn lexikoiak hizkuntzaren gordailu nagusi bihurtu dira, eta hauen eraikuntza arlo honetako funtsezko ataza dugu, gaur egun. Izan ere, LNPrako sistemek neurri errealeko testuekin lan egin behar badute, milaka sarrera dituzten baliabide lexikal aberatsak behar dituzte ezinbestean. Lexikoiei esker makinek itzulpen automatikoa, informazio-erauzketa eta hitzen adieren desanbiguazioa bezalako atazak burutu ditzakete.

IXA taldean, dagoeneko badugu informazio lexikala jasotzen duen gordailua: *Euskararen Datu-Base Lexikala* (EDBL) deritzoguna (Agirre *et al.*, 1994a; Aduriz *et al.*, 1998b; Aldezabal *et al.*, 2001a). EDBLn ale lexikal bakoitza bere kategoria eta azpikategoria lexikal edo morfosintaktikoaren arabera sailkatuta dago (kategoria morfosintaktikoak direnak, kategoriaz gain, dagokien informazioaz hornituta daude: kasua, aspektua, numeroa, mugatasuna, funtzioa...). Esan dezakegu, beraz, EDBLn jasotzen den ezagutza-mota gramatikala dela. Horrelako informazioa jasotzen duten lexikoiak izendatzeko *datu-base lexikal* (DBL) terminoa erabiltzen da.

Esan dugun bezala, IXA taldean dagoeneko morfologia eta sintaxia landu dira, eta horren fruitu dira, batetik, MORFEUS analizatzaile morfologikoa (Alegria *et al.*, 1996) —eta hau oinarrian duen XUXEN zuzentzaile ortografikoa (Agirre *et al.*, 1992)—, eta bestetik, garapenean dagoen euskarako analizatzaile sintaktikoa (Aranzabe *et al.*, 2004). Bi analizatzaile hauek EDBLn dute oinarria. Hau da, hitzak morfologikoki segmentatzeko eta analizatzeko behar den informazio gramatikala EDBLn dago jasota.

Hala ere, itzulpen automatikoa edota adieren desanbiguazioa egiteko informazio gramatikala ez da nahikoa, informazio semantikoa ere beharrezkoa

baita. Honen adierazgarri hurrengo adibidea dugu, zein itzulpen automatikoaren eremuan kokatu dugun.

(1) Eskusoinua jotzen dut.

Demagun (1) esaldia dugula, eta honen itzulpen automatikoa lortu nahi dugula. Horretarako, hasteko, nahitaezkoa izango da lexikoian esaldiko hitz bakoitzaren adierak zerrendatuta egotea, eta, are gehiago, adiera bakoitza dagokion erdarako ordainarekin zehaztuta etortzea. I.1 taulan aurreko adibideko hitzen adierak lexikoi hipotetiko batean aurkezten ditugu, bakoitzaren gaztelaniako itzulpenekin[1].

Hitza	***Adiera***	***Definizioa***	***Itzulpena***
eskusoinu	A1	musika-tresna, tekla edo botoiduna	acordeón
jo	A1	gauza batez beste bat halako indarrez ukitu	golpear/pegar
jo	A2	ukaldiak eman	golpear/pegar
jo	A3	musika-tresna bati soinua atera	tocar
jo	A4	tokiren baterantz joan	ir/dirigirse
jo	A5	kopuruei buruz, zenbatekoa, adierazten dena	estimar/calcular

I.1 Taula: (1) adibideko hitzen adierak eta itzulpenak.

Hala, hitzen itzulpena lortzeko tresnak, lehendabizi, esaldiko hitzen adierak kontsultatu beharko ditu oinarri gisa erabiliko duen lexikoian, eta, ondoren, hitzak esaldian zein adieratan erabiltzen diren aukeratu, hots, hitzen adieren artean desanbiguatu. Kasu honetan, lexikoian ditugu jo hitzaren hainbat adieren artean, 'musika-tresna bati soinua atera' (A3) adiera aukeratu beharko du makinak, horretarako beste guztiak gaitzetsiz. Joren adiera zuzena lortzeko beharrezkoa izango da esaldiko testuinguruari erreparatzea, eta jo eta eskusoinu hitzak semantikoki erlazionatzea: jok musika-instrumentuekin zerikusia du (A3), eta eskusoinua musika-instrumentu bat da (A1). Beste modu batean esanda, esaldi horretako jo hitzaren adiera desanbiguatzeko, eta, ondorioz, itzulpen zuzena emateko, jo eta eskusoinu hitzen eta hauen adieren arteko loturak zehaztuta egon behar dute lexikoian. Horrelako erlazioak dituzten lexikoiak, ordea, ez dira datu-base lexikalak, *hiztegi ezagutza-baseak* (HEB), *ezagutza-base lexikalak* (EBL) eta *ontologiak* baizik.

[1] Adibiderako *Euskal Hiztegi*ko (Sarasola, 1996) adierak erabili ditugu, eta hitzen adiera-kopurua eta definizioak laburtu egin ditugu.

Tesi-lan honetan EBLen alde egin dugu, hau da, euskararen informazio lexiko-semantikoa jasotzen duen lexikoia EBL gisa diseinatu dugu; II. kapituluan ikusiko dugun bezala, hauek sarrera lexikaletako informazioa egituratu egiten dute, erredundantzia konponduz, datuen kontrola eta kontsistentzia gauzatuz eta informazio-atzipena erraztuz. Hortaz, ezagutzaren errepresentaziorako eta biltegirako oso egokiak dira, eta gaur egun hauexek dira LNPn lexiko-semantikaren arloan nagusitzen direnak. EBLetan hitzei eta adierei buruzko informazioa dago, eta hauen ezaugarri garrantzitsuena herentzia izaten da, hitzak eta adierak klase/azpiklase hierarkien inguruan antolatzen baitira (Copestake, 1990).

Honenbestez, euskararen ikerketa semantiko aplikatua egiteko, eta datubase lexikal batek eskaintzen dituen analisi linguistikoetatik haratago joateko, euskararen informazio semantikoa egituratu eta antolatzen duen EBL baten beharra dago. Behar horri erantzuna emateko jaio zen tesi-lan hau, balizko EBL horren hezurdura garatzeko eta definitzeko, hain zuzen ere.

I.2 Helburuak

Hemen aurkezten dugun lanaren helburu nagusia, beraz, euskararen azterketa semantikoa ahalbidetzeko beharrezkoa den euskararako EBL bat sortzea da. Helburu hau gauzatzeko, eginkizun zehatzagoak ere bete behar izan ditugu:

- **IXA taldearen beharretara egokitzen den lexikoiaren ezaugarriak definitu:**

Lehenengo urratsa, IXA taldearen beharretara egokitzen den EBLaren ezaugarriak zerrendatzea izan da. Horretarako, kontuan hartu behar izan ditugu:

(a) EBLa non eta nola erabili nahi dugun.

Gure kasuan, konputazionalki inplementa daitekeen EBLa izatea nahi dugu.

(b) Zer informazio mota txertatu behar zaion EBLko sarrera bakoitzari.

Inplementatu beharreko EBLa izaki, geroz eta lexiko aberatsagoa izan, geroz eta emaitza hobeak izaten dira ataza konputazionaletan. Hala, hizkuntza bere osotasunean adierazten duen EBLa izan behar genuke, ahalik eta informazio gehiena jasotzen duena, bai semantikoa eta baita sintaktiko-semantikoa ere.

(c) EBLaren informazioa adierazteko aukeratzen den ereduak zein baldintza bete behar dituen.

Ez dago EBLaren eraikuntzarako eredu bakarra; eta, izatez, eredu bakarra jarraitzen duen EBLra mugatzea arriskutsua izan daiteke. Izan ere, askotan, EBLetan jasotako informazioa ez da berrerabilgarria eta, ondorioz, aplikazio berrien sorkuntza baldintza daiteke. Aukeratutako eredu honek ez ditu gainontzeko lan konputazionalak eragotzi behar, gure EBLa lan horien informazioarekin ere aberastu ahal izateko. Hala, gure EBLa informazio berrerabilgarria jasotzen duena izatea nahi dugu, eta bertan egindako deskribapen linguistikoekin ez baldintzatzea etorkizuneko aplikazioak.

Honekin batera, eleanitza den EBLa interesatzen zaigu, euskarako sarrera lexikalez gain, beste hizkuntzetako ordainak eskuragarri dituena. Itzulpen automatikorako, adibidez, ezinbesteko baldintza da hau.

- **Erdal hizkuntzetarako dauden ereduak aztertu, eta IXA talderako baliagarria izango den eredu bat aukeratu:**

Gure ereduaren izaera finkatuta, azterketa bibliografikoa egin dugu, aipatutako ezaugarrietara gehien egokitzen den formalismoaren bila. EBLen eraikuntzarako ereduak ugariak dira, eta ikerlan honen ezinbesteko muga dela-eta, azterketaren esparrua murriztu behar izan dugu.

- **Gure EBLa aukeratutako ereduari jarraituta garatzeko metodologia definitu:**

Euskarako EBLak jarraituko duen eredua aukeratu ondoren, eta EBLaren eraikuntzari ekin aurretik, garapenean eragina izango zuten hainbat erabaki hartu behar izan ditugu; hala nola, zein kategoria landuko genuen lehendabizi, edota zein ikuspegi erabiliko genuen sarrera lexikalak lantzeko garaian. Estaldura —sarrera lexikalen kopurua ahalik eta handiena izatea— eta kalitatea —sarrera lexikalen informazioa zuzena izatea— uztartzen saiatu gara, eta ezaugarri hauek izango dira, hain zuzen ere, EBLaren garapen-metodologia definituko dutenak.

- **Euskarako EBLaren garapenean sortutako zailtasunentzat irizpideak ebatzi:**

EBLa garatzeko metodologia zehaztu arren, EBL baten garapenean aurrera egin ahala, tratamendu berezia behar duten fenomeno linguistikoak agertzen dira. Hori gertatzean fenomenoa aztertu eta fenomeno linguistiko horrek EBLan izango duen tratamendua zehazten duen irizpide bat definitu behar dugu, fenomeno bera EBLko sarrera desberdinetan beti modu berean adierazia izan dadin.

- **Aukeratutako eredua informazio gehiagoz hornitu:**

Hizkuntza bere osotasunean adierazten duen EBLa izatea nahi dugunez, ahalik eta informazio gehien behar dugu, horrela, emaitza hobeak lor daitezkeelako. Hori dela eta, oinarri gisa aukeratutako eredutik jasotako informazioaz gain, informazio gehiagorekin aberasten saiatu gara gure EBLa; ingeleseko eta euskarako kirol-arloko aditz batzuen subjektu eta objektu hautapen-murriztapenekin, hain zuzen ere.

I.3 Tesi-txostenaren eskema

II. kapituluan, lexikoiez jardungo gara luze, hizkuntzalaritzan eta bereziki hizkuntzalaritza konputazionalean izan duen lekuaz eta berau lantzeko garaian izan diren gorabeherez. Lehenengo, lexikoiek izandako ibilbidea laburbilduko dugu. Gero, lexikoien ezaugarriak azaldu, lexikoien garapenean egungo joerak ikusi eta lexikoi mota desberdinak aztertuko ditugu. Honekin batera, egin diren hainbat lexikoien berri emango dugu.

III. kapituluan, batetik, egin nahiko genukeen EBLaren ezaugarriak zerrendatuko ditugu, eta bestetik, EBLen hainbat eredu edo formalismo aztertuko ditugu, gerora, egokiena iruditzen zaiguna euskarako EBLaren garapenean erabiltzeko. EBLen eraikuntzarako ereduak ugari daudenez, azterketaren esparrua murriztu behar izan dugu. Hortaz, lehenik eta behin, azterketarako aukeratutako formalismoen arrazoiak azalduko ditugu, eta formalismo bakoitzetik ezaugarri nagusienak ere aipatuko ditugu. Formalismo hauek aztertu ondoren, IXA taldearen beharretara hobekien egokitzen den EBL formalismoa zein den arrazoituko dugu, ***WordNet*** eta honen ildotik abiatuta garatu diren ***EuroWordNet*** eta ***The Multilingual Central Repository*** **(MCR)**, hain zuzen ere.

IV. kapituluan, WordNet, EuroWordNet eta MCR ereduen azterketa sakonagoa egingo dugu.

V. kapituluan, euskarako *wordnet*aren[2] (***Euskal WordNet***) garapenerako hartutako erabaki metodologikoak deskribatuko ditugu, eta erabaki hauen arabera, Euskal WordNetek izandako garapena ere deskribatuko dugu. Alde batetik, izenekin egindako azterketa azalduko dugu (garapen-aldiak zehazki deskribatuz), eta bestetik, oraindik hasiberria dugun aditzen azterketa eta garapenerako landu ditugun aukera metodologiko ezberdinak aurkeztuko ditugu.

VI. kapituluan, EBL eleanitz bat sortzeak ekartzen dituen fenomeno linguistiko batzuk aztertu ditugu. Lexikalizazioarekin zerikusia duten bereizgarriez eta hierarkiari dagozkion bereizgarriez arituko gara. Kapitulu honetan, fenomeno hauen adibideak emango ditugu eta hizkuntzen arteko ezberdintasun horiek nola konpondu ditugun azalduko dugu. Era berean, ereduari egindako hobekuntzak ere aurkeztuko ditugu.

VII. kapituluan, aukeratutako eredua informazio gehiagorekin hedatzeko egin dugun lehenengo saiakera azalduko dugu. Ingeleseko eta euskarako kirol-arloko aditz batzuen objektuen eta subjektuen hautapen-murriztapenen azterketa deskribatuko dugu. Azterketa honetan, erabilitako corpusei, eskuratze-tekniken azterketari eta ebaluazio linguistikoari erreparatuko diegu batez ere. Esan beharra dago azterlan hau eleaniztasunaren hipotesiaren ikuspegitik egina dagoela. Hots, ingeleserako automatikoki eskuratutako hautapen-murriztapenak euskaraz ere erabilgarriak izan daitezkeela frogatu nahi dugu. Horretarako, ingeleserako automatikoki eskuratu diren hautapen-murriztapenetan oinarritu gara lehenengo, gero hauek euskararentzat baliagarriak izan daitezkeen aztertu ahal izateko.

VIII. kapituluan, bukatzeko, zabaldu ditugun ikerlerroak, atera ditugun ondorio nagusiak eta aurrera begirakoak aipatuko ditugu.

[2] *WordNet* (letra larriz) erabiltzen dugu Miller-en taldeak (1985) egindako ingeleseko EBLa adierazteko; *wordnet* (letra xehez), aldiz, WordNeten oinarrituta garatu den edozein hizkuntzetako EBLari buruz hitz egiteko erabiltzen dugu. Hala, *WordNet* terminoarekin, ingeleseko wordnetari egingo zaio erreferentzia, eta *wordnet* terminoak aurretik zer hizkuntzetakoa den adierazia izan beharko du.

Gainontzean, hiru eranskinek osatzen dute tesi-lan hau:

- **A eranskina: Euskal WordNeteko editorearen eskuliburua.** Eskuliburu honetan Euskal WordNeteko editoreak *synset*ak lantzeko behar dituen argibide guztiak zehazten dira: alde batetik, interfazearen erabilerari buruzko azalpenak, eta bestetik, eleaniztasunak eragindako desberdintasun linguistikoetan erabili beharreko irizpideak.

- **B eranskina: Euskal WordNeteko aditzen hierarkiaz hierarkiako orrazketa.** Eranskin honetan {express_2, give_tongue_1, utter_1} klase semantiko osorako egindako hierarkiaz hierarkiako orrazketa aurkezten dugu. Honekin batera, orrazketa honen ondoren, lortutako ondorio nagusiak dakartzagu, baita ingelesa eta euskarako hierarkien arteko alderaketa bat ere.

- **C eranskina: Hautapen-murriztapenen azterketa eta ebaluazioa.** Hainbat eskuratze-teknika erabiliz, ingeleseko eta euskarako corpus ezberdinetatik eskuratutako hautapen-murriztapenak aurkezten ditugu, hauen zuzentasunari buruzko ebaluazioarekin batera. Bestalde, ebaluazioa egin ahal izateko, lehenengo hautapen-murriztapenen iturria aztertu dugu. Azterketa honen emaitzak eta honetarako erabilitako baliabideak ere zehazten dira.

I.4 Tesiarekin lotutako argitalpenak

Sarrera-kapitulu honi bukaera emateko, jarraian, argitalpenen zerrenda aurkezten dugu, eta I.2 taulan, argitalpen bakoitza zein kapitulurekin lotuta dagoen zehazten dugu[3].

- Agirre E., García E., Lersundi M., Martínez D., eta Pociello E. The Basque task: did systems perform in the upperbound? *Proceedings of the SENSEVAL-2 Workshop*, Tolosa (Frantzia), 2001.

[3]Hauek guztiak hurrengo web orrian daude atzigarri: `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016` (2007-07-02an atzitua).

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

- Agirre E., Aldezabal I., eta Pociello E. A pilot study of English selectional preferences and their cross-lingual compatibility with Basque. *Proceedings on International Conference on Text Speech and Dialogue (TSD)*, Ceske Budejovice (Txekiar Errepublika), 2003a.

- Agirre E., Aldabe I., Lersundi M., Pociello E., eta Uria L. The Basque lexical-sample task. *Proceedings on the 3rd ACL Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Bartzelona, 2004a.

- Pociello E. *Aditzen hautapen-murriztapenak: kirol domeinura mugatutako ingeleseko hautapen-murriztapenak eta euren baliagarritasuna euskararako. Hastapeneko lana.* Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2004a.

- Pociello E. *Sintaxi-semantika elkargunea zenbait teoriatan: euskararen ezagutza-basea lexiko-semantikorantz.* Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2004b.

- Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian M., eta Pociello E. *EuSemcor: euskarako corpusa semantikoki etiketatzeko eskuliburua: editatze- etiketatze- eta epaitze-lanak.* Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea, 2005a.

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

- Agirre E., Aldezabal I., eta Pociello E. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: Euskal WordNet. *GOGOA: Euskal Herriko Unibertsitateko Hizkuntza, Ezagutza, Komunikazio eta Ekintzari buruzko Aldizkaria*, 237–266, 2005c.

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

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

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

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

- Agirre E., Aldezabal I., Etxeberria J., eta Pociello E. A preliminary study for building the Basque PropBank. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa (Italia), 2006d.

- Agirre E., Aldezabal I., eta Pociello E. Lexicalization and multiword expressions in the Basque WordNet. In Fernández B. eta Laka I., editors, *Andolin gogoan: Essays in honour of the Professor Eguzkitza*, 51–68. Euskal Herriko Unibertsitatea, 2006d.

Kapitulua	***Argitalpenak***
III	Pociello (2004b) Agirre *et al.* (2005c)
IV	Agirre *et al.* (2005c) Agirre *et al.* (2005a)
V	Agirre *et al.* (2002) Agirre *et al.* (2005c) Agirre *et al.* (2006a) Agirre *et al.* (2006b)
VI	Agirre *et al.* (2006e) Agirre *et al.* (2006c)
VII	Agirre *et al.* (2003a) Pociello (2004a)

I.2 Taula: Kapitulu bakoitzarekin lotutako argitalpenak.

II. KAPITULUA

Lexikoiak

Kapitulu honetan lexikoiei buruz arituko gara eta hauek LNPn duten garrantzia ikusiko dugu. Lehenengo atalean lexikoiek izandako ibilbidea laburbilduko dugu. Gero, lexikoien ezaugarriak azalduko ditugu, lexikoien garapenean egungo joerak ikusiz eta lexikoi motak laburbilduz.

II.1 Lexikoiez historia apur bat

Jadanik aipatu dugu —I.1 atalean— lexikoien garapena funtsezko ataza dugula LNPn. Ordenagailuek itzulpen automatikoa, testuen-laburpena eta abar egin ahal izateko, milaka sarreratik gora dituzten biltegi lexikaletan oinarritu behar dute. Hots, neurri handi batean, esan dezakegu, gaur egun, LNPko edozein sistemaren muina biltegi lexikalean datzala.

Hala ere, lexikoien garrantzia ez da beti hain handia izan, ez hizkuntzalaritza teorikoan, ez LNPn. Neurri handi batean, azken hogeita hamar urteotan zabaldutako fenomenoa izan da. Hizkuntzalaritza teorikoaren zein konputazionalaren egungo joera hizkuntza-ezagutza gramatikalaren arlotik lexikoarenera lerratu da.

Hizkuntzalaritza teorikoak lexikoari buruz zuen ikuspegia 1980ko hamarkadatik aurrera aldatu egin da, sarrera lexikalaren edukiari garrantzi gehiago emanez. Hizkuntzalaritzako teoria askok eta askok (segur aski Chomskyrengandik hasita, 1970ean) joera lexikalistago bat hartu zuten. Hizkuntzalaritza konputazionalean ere lexikoien beharra aitortu zen, hauek guztiz beha-

rrezkoak baitziren aplikazio errealak garatzeko.

Hizkuntzalaritzaren ibilbidean sintaxia izan da aztergai nagusia, eta lexikoari ez zaio agian behar beste arretarik jarri, lexikoa hizkuntza bakoitzak berezkoa duen ataltzat hartu izan baita, erregela edo mekanismo linguistikoez orokortu ezin dena. Gramatika Sortzailearen hasierako eredua (Chomsky, 1965) eta ondorengo Gobernu eta Uztarduraren Teoria dira aipatutakoaren adibideak (Chomsky, 1987).

Teoria hauei egiten zaien gaitzespen azpimarragarrienetako bat da beraien erregelekin hizkuntzaren egitura orokorrenak bakarrik azal daitezkeela. Gabezia hau eta beste batzuk gainditzeko asmoz, gramatika aztertzeko ikuspuntua aldatu egingo da denborarekin, eta sintaxiaren arloan aritzen diren teoriak norabide lexikalistagoa hartuz joango dira (Hale eta Keyser, 1987; Jackendoff, 1990; Levin, 1993; Pustejovsky, 1995, ...).

> "Asistimos desde hace un cierto tiempo a un razonable aminoramiento — que no es a la aniquilación— de la idea de que no hay explicación gramatical posible si ésta no se puede formular en reglas estrictas y ordenadas. Con ello llega también el renacer de campos que, por razones próximas a ese punto de vista, habían sido relegados durante un largo período. La concepción de las teorías como sistemas modulares y de principios es buena prueba de lo primero; el retorno al estudio de las palabras en cuanto elementos portadores de un significado, de los conjuntos en que se organizan, de sus relaciones y su papel en la gramática, es signo de lo segundo." (Demonte, 1991, 24. or.)

Egile hauen ustez, lexikoa *salbuespenen zakua* izatetik, egitura konplexua duen atala izatera pasatu da, bertan sintaxi eta semantikaren arteko hartuemanak islatzen direlarik.

Hizkuntzalaritza konputazionalak teorikoaren antzeko ibilbidea izan du. Hastapenetan, 1950 eta 1960ko hamarkadetan, sistema konputazional gehienek *jostailuzko* lexikoak lantzen zituzten, oso aplikazio-domeinu zehatzei lotuak eta sarrera-kopuru murritzekoak. Askotan zerrenda soilak baino ez ziren izaten. B. Boguraev-ek eta T. Briscoe-k esaterako, hau diote:

> "Knowledge of words underlies these tasks, yet until very recently dictionaries (or lexicons, as linguists usually call them) for natural language processing systems have by and large been the poor sisters of computational linguistic research." (Boguraev eta Briscoe, 1989, 34. or.)

Oro har, ikertzaileak sintaxia eta erregela gramatikaletan jartzen zituzten beren indarrak.

1970 eta 1980ko hamarkadetan, LNPrekiko interesa areagotzeaz gain, hurbilpen-aldaketa gertatu zen: informazio- erauzketarako sistemek edo itzulpen automatikoko sistemek, baliabide lexikal sendoak behar zituzten, testu errealekin lan egitekoak baziren. Hurbilpen-aldaketa horren adierazgarri 1986ko Grosseto-ko mintegia (*Automating the Lexicon*) dugu, non mintegiaren bukaeran *Manifesto* dokumentua osatu zen, lexikoi sendoen beharra azpimarratuz. Gauzak horrela, 1980ko hamarkadaren bigarren erdian eta 1990eko hamarkadaren hasieran alderdi lexikoan arreta handiagoa jarri zen —*Generalized Phrase Structure Grammar* (Gazdar *et al.*, 1985), adibidez—, eta lexiko konplexu ugari proposatzen hasi ziren; esate baterako, Europan, lexikoien inguruan, hogei proiektu baino gehiago sortu ziren.

Hala ere, lexikoaren inguruan hainbat ikerketa eta proiektu garatu baziren ere, proiektu horietako ikerlariek lexikoa aztertzeko eta adierazteko, modu asko asmatu eta erabili zituzten. Nork berea —eta bere modura— egiten zuelarik, ordea, azkenean batek egindakoaz beste batek baliatu nahi zuenean, aurretik egindako lan guztia ez zen nahi litzatekeen bezain lagungarria suertatzen, eta, maiz, erabili ezina izaten zen ere bai.

B. Boguraevek eta T. Briscoek (1989) adibide baten bidez azaltzen dute aurrean aipatutako egoera. Hiru sistema desberdinek —BBN-CFG sistema (Ingria, 1988), IRUS sistema (Bates *et al.*, 1986) eta ALVEY sistema (Carroll eta Grover, 1989), hurrenez hurren— ingeleseko acknowledge hitzerako duten adierazpena azaltzen digutenean (ikus II.1 irudia).

II.1 irudiko hiru sarrerek acknowledge hitzari buruzko antzeko informazioa gordetzen dute: kategoria sintaktikoa, hitzaren azpikategorizazioa eta abar. Hala ere, informazio hori hain modu desberdinean dago adierazita, ia ezinezkoa bihurtzen dela hiru formalismo hauen arteko informazioa bateratzea.

Horrela, bada, garatutako lexikoi hauek behar bereziei aurre egiteko soilik diseinatzen ziren, proiektuen arteko elkarlana kontuan hartu gabe. Egoera honi aurre egiteko, informazio lexikalaren *berrerabilgarritasunaren* beharra azpimarratu zen. Calzolari-ren lanean (1994), egileak berrerabilgarritasunaren alde egiten du, nabarmen. Bere ustean, komunitate linguistikoak dagoeneko existitzen diren lexikoien informazioa berrerabiltzen eta estaldura zabala duten baliabide lexikalak eraikitzen ahalegindu beharko luke. 1990eko hamarkadaren lehen erdian, Europako Erkidegoko batzorde batek hiru baldintza aipatzen ditu lexikoiei etekin handiago atera ahal izateko:

```
[ACKNOWLEDGE
    Category:   V
    Base:       acknowledge
    Features:   (TRANSITIVE (REALNP) (PASSIVIZES))
                (CLAUSE (REALNP) (THATCOMP)
                (INDICATIVE: TENSE) (WH-))
                (NP-VP :AGR :AGRX (REALNP) :AGRX
                (PASSIVIZES) (INF) (WH-))]
```

```
[ACKNOWLEDGE
    FEATURES   (TRANS
                PASSIVE
                THATCOMP
                THATREQUIRED
                NPTOCOMP)
V S-D]
```

```
(acknowledge
    ((v +) (n -) (subcat npl)) acknowledge nil)
(acknowledge
    ((v +) (n -) (subcat sfin)) acknowledge nil)
    ;acknowledge that they were defeated
(acknowledge
    ((v +) (n -) (subcat se3)) acknowledge nil)
    ;acknowledge having been defeated
(acknowledge
    ((v +) (n -) (subcat or)) acknowledge nil)
    ;acknowledge him to do the best
```

II.1 Irudia: acknowledge hitzaren hiru adierazpen desberdin, BBN-CFG sistema (Ingria, 1988), IRUS sistema (Bates *et al.*, 1986) eta ALVEY sistema (Carroll eta Grover, 1989), hurrenez hurren.

- Baliabide lexikalen eraikuntza zabal onarturiko estandarretan egin beharra.

- Europako Erkidegoko hizkuntza guztietarako baliagarri izango diren oinarrizko lexikoien eraikuntza, adosturiko diseinu bat erabilita eraikiko dena.

- Sorturiko baliabide lexikalak komunitateak eskuragarri izan ditzan, distribuziorako politika baten beharra.

Egun, Europan, arlo honetako proiektu garrantzitsuenetakoek — *Expert Advisory Group on Language Engineering Standards* (EAGLES)[1], *Preparatory Action for Linguistic Resources Organization for Language Engineering* (PAROLE), *Trans-European Language Resources Infrastructure* (TELRI)[2] eta *European Language Resources Association* (ELRA)[3]— hiru alderdi horiek lantzea dute helburu nagusi.

Hortaz, lexikoi batek berrerabilgarria izan behar du; hau da, bere informazio lexikalaz baliatzeko aukera eman behar du, lexikoi berri bat garatu nahi denean edota dagoen lexikoiren bat aberastu nahi denean. Honekin batera, lexikoi bat berrerabilgarria izango da baldin eta *estandarra* bada. Hau da, baldin eta honen errepresentaziorako eskemak orokorrak eta aplikazioetarik independenteak badira. Modu horretan, bere baitan biltzen duen informazioaren adierazpidea formalismo berezi bati lotuegia egotea eragotz daiteke. Behar honi erantzuteko asmoz, hurrengo ekimenak aipa genitzake: *Text Encoding Initiative* (TEI), *The ACL Data Collection Initiative* eta *Consortium for Lexical Research*, besteak beste. Hala ere, tamalez, egun ezin da esan informazio lexikala kodetzeko formalismo estandar bat dugunik.

Nahiz eta albo batera utzi den hasiera bateko gehiegizko optimismoa, gaur egun, joera lexikalistak badirau, bai hizkuntzalaritza teorikoan bai konputazionalean. Lexiko konputazionalaren alorrean lexiko-sistemen azterketa, errepresentazioa eta erabilera, gero eta garrantzi handiagoa hartzen ari da. Azken hamarkadan lexikoigintzan aurrera egin da: erredundantziaren arazoa konponduz, datuen kontrola eta kontsistentzia gauzatuz, eta informazioatzipena erraztuz. Argi dago, beraz, hizkuntzen industriaren interesa lexikora lerratu dela, eta ez da harritzekoa, hortaz, lexikoi horien eraikuntza izatea

[1] `http://www.ilc.pi.cnr.it/EAGLES/home.html` (2007-07-02an atzitua).
[2] `http://www.ids-mannheim.de/telri/html` (2007-07-02an atzitua).
[3] `http://www.icp.grnet.fr/ELRA/home.html` (2007-07-02an atzitua).

LNPko gairik landuenetako bat. Hala, II.2. atalean lexikoien ezaugarriez arituko gara.

II.2 Lexikoiei buruz

Lexikoiei buruz hitz egin ahal izateko, *lexikoi* eta *hiztegien* artean desberdindu beharra dago. Bai lexikoiek eta bai hiztegiek hitz baten adierari buruzko deskribapena eta informazio lexikoa jasotzen dute, baina bakoitzaren erabileraren arabera, jasotzen den informazio mota eta informazio horren antolaketa aldatu egiten da.

Esate baterako, hiztegien erabiltzaileak gizakiok garenez, bertako informazioa gizakiok uler eta erabil dezagun dago antolatuta. Hala, hiztegisarrera bakoitzeko, orokorrean, hitz horren adierari buruzko azalpen bat eta adibide batzuk ematen zaizkigu.

Lexikoek, aldiz, informazio lexikala jasotzen duten biltegiak izateaz gain, aplikazio batekin lotura izan behar dute. Beste hitz batzuekin esanda, lexikoien erabiltzaileak ordenagailuak dira. Horregatik, lexikoi konputazionaletako sarrerek informazio linguistiko (morfologiko, sintaktiko eta semantiko) esplizituarekin hornituta egon behar dute, betiere LNPko sistema batean integratzeko moduan antolaturik. Hortaz, hiztegietan dugun informazioa lexikoietan aurkitzen duguna baino mugatuagoa da, hiztegietan hitz baten adiera ulertzeko behar den informazioa bakarrik eskaintzen baita.

Hurrengo definizioak ondo adierazten du lexikoiaren eta hiztegiaren arteko desberdintasuna:

> "[A lexicon is] a set of formalized entries to be used in conjunction with computer programs and by dictionary the physical printed text giving lexical information, including meaning descriptions." (Wilks *et al.*, 1996, 6. or.)

Wilks *et al.*-ek (1996), hiztegia testu inprimatu gisa definitzen badu ere, gaur egun jakina da testu inprimatua izateaz gain, euskarri elektronikoan ere egiten direla hiztegiak.

Lexikoiak zer diren zehaztu ondoren, aipa ditzagun lexikoak garatzeko erabiltzen diren hainbat iturri eta metodo.

II.2.1 Lexikoiak sortzeko hurbilpenak, metodoak eta iturriak

Lexikoia eskuratzeko bi hurbilpen nagusi erabili izan dira: *arauemailea* eta *deskriptiboa*. Hurbilpen arauemailean, marko zehatza definitzen da, eta informazioa marko horretan txertatzen da lehenengo. Hurbilpen deskriptiboan, aldiz, aurrez ez dago definiturik inongo marko zehatzik, eta ezaugarri multzoa osatuko duten elementuak aztertutako datuetan agertutakoak dira.

Lexikoiak sortzeko berebiziko garrantzia dauka, baita ere, lexikoiaren garapenean erabilitako metodoak, hau da, gizakiaren eskuratzea zenbaterainokoa izango den zehazteak. Hiru metodo erabil daitezke: *eskuzko metodoa*, *metodo automatikoa* edo *metodo erdiautomatikoa*. Metodoa erabakitzerakoan, kontuan izan behar dira, alde batetik, zeintzuk diren erabiliko diren iturriak, eta bestetik, helburu den aplikaziorako zein informazio zehaztu edo markatu behar den. Hala, eskuzko metodoek hurbilpen arauemailea darabilte. Hurbilpen deskriptiboan, aldiz, metodo automatikoa eta erdiautomatikoa erabil daitezke.

Hurbilpen arauemaileetan, esan bezala, eskuzko metodoa da nagusi, eta metodo honetan iturri nagusia introspekzioa da, hots, hizkuntzalariak munduari buruz duen jakinduria eta ezagutza. Lexikoia osatzeko garaian, gizakiek munduari buruz eta hizkuntzari buruz duten jakinduria erabiliz gero, sortutako datuen zuzentasuna bermatuko da. Hurbilpen honekin garatutako proiektuen arazo nagusienak dira, batetik, jende eta denbora ugari behar izatea, eta bestetik, jende ezberdin asko garai ezberdinetan proiektu batean lan egiterakoan, koherentzia arazoak sor daitezkeela.

Hemen aipatzen ditugu era honetan sortutako zenbait lexikoi: *Word Dictionary*, 10.000 sarrera dituen *Linguistic String Project*erako (LSP) sortutako lexikoia (Fox *et al.*, 1988); *WordNet*[4] (Miller, 1985; Fellbaum, 1998a) gaur egun 3.0 bertsioa da indarrean, eta 155.327 hitz daude bertan errepresentatuta eta baita euren arteko erlazio semantikoak dituzten 117.617 *synset* edo sinonimo-multzo ere; *Comlex* (Grishman *et al.*, 1994) ingeleseko 38.000 inguru hitzentzako informazio sintaktikoa dakarren lexikoi konputazionala; CYC ontologia (Lenat, 1995) 100.000 termino baino gehiago ditu. LDOCEren azken bertsioak, LDOCE3-NLP, 80.000 adiera ditu, eta hizkuntzalaritza konputazionaleko ikerkuntzarako laguntza gisa sortu dute *Longman*eko lexikografoek.

[4] `http://www.wordnet.princeton.edu` (2007-07-02an atzitua).

Era honetan sortutako lexikoietan, hasieran kontzeptuen ontologia sortzen da, eta ondoren kontzeptu horiei hitzak lotzen zaizkie. Lan egiteko sistema honi jarraitzen zaizkion proiektuen artean honakoak ditugu: CYC (Lenat eta Guha, 1990), WordNet (Miller, 1985; Fellbaum, 1998a), eta EDR (Yokoi, 1995), besteak beste.

Hurbilpen deskriptiboa arauemailearen kontrakoa da: lehenik hitzak sortzen dira, eta ondoren kontzeptuak lotzen zaizkie. Honezaz gain, hurbilpen deskriptiboetan, oinarri gisa aurretik sortuta dauden testu-baliabideak erabiltzen dira, eta horietan dagoen informazioa erauzten saiatzen dira metodo automatikoak edo erdiautomatikoak erabiliz.

II.1. atalean aipatu dugun bezala, LNPn, 1980 hamarkadarainoko sistemetan, ahaleginen handiena sintaxi-egituretara mugatzen zen. Garai horretan formalismo sintaktiko batzuk egitura sintaktikoen pisua lexikoira pasatzen hasi ziren, lexikoiaren egitura konplexuagoa bihurtuz. Garai berean, konturatu ziren LNPrako sistemen hedakuntzarako arazo nagusiena lexiko urriegia izatea zela eta lexikoa edukiz betetzea uste baino lan neketsuagoa zela. Lexiko zabal eta konplexuen eraikuntza eskuz egitea gehiegizko lana izango zela-eta, aldez aurretik sortuta zeuden testu-baliabideetan zegoen informazioa ustiatzen ahalegindu ziren, hala nola, *egitura gabeko oinarri lexikalak* (corpusak) edo *egituratutako oinarri lexikalak* (makinak irakurtzeko moduko hiztegiak (MRD[5]), thesaurusak eta entziklopediak).

Corpusak **egitura gabeko baliabidetzat** hartu izan ohi dira, baina erabilerari buruzko informazio-hitzak errealitatean dituzten maiztasunak, zenbait egitura sintaktikoek dituzten maiztasunak eta halako informazioa lortzeko oso erabilgarriak dira. Hitzak berak eta hitzarekin batera agertzen den testuinguruak informazio baliagarria ematen du askotan.

Honekin batera, makinaz tratatu daitezkeen testuen kopurua etengabe hazten ari da. Beraz, honetaz guztiaz jabetuta, 80ko hamarkadatik aurrera, baliabide hau geroz eta gehiago erabili izan da, eta egun, lexikoien aberasketan erabiltzeaz gain, hizkuntzalaritzaren beste alor guztietan ere erabiltzen da. Aitzindari gisa, milioi bat hitz inguru dituen *Brown* corpusa (Francis eta Kucera, 1982) aipatu beharra dago.

Corpusetan, berez, hitzak bakarrik daude (corpus gordina deritzoguna). Corpusa gordina baldin bada, erabilgarria da; baina are erabilgarriagoa da corpus berari informazio linguistikoa gehitzen badiogu. Badira lematizatuta-

[5]MRD, *Machine Readable dictionary*ren laburdura da. LNPn asko erabiltzen den laburdura da.

ko corpusak, hitzen kategoriak markatuta dituzten corpusak, morfologikoki analizatuta dauden corpusak, egitura sintaktikoak markatuta dituzten corpusak, semantikoki markatutako corpusak[6], eta abar. Beraz, corpus batean, gero eta informazio gehiago izan, orduan eta erabilgarriagoa izango da LNPn.

Esate baterako, *Penn Treebank* proiektuari esker, orain arte gehien erabili izan diren ingeleseko corpusak —jadanik aipatu dugun Brown corpusa (Francis eta Kucera, 1982), eta bi milioi hitz inguru duen *Wall Street Journal* corpusa— maila hauetan markatu dira: hitzaren kategoria (Marcus *et al.*, 1993) eta azaleko egitura sintaktikoa (Marcus *et al.*, 1994). Aurrekoez gain, 250.000 hitzetako Brown corpusaren testu zati bat hartu dute eta Princetoneko kategoria-etiketatzaile automatikoarekin etiketatu dute lehenik, eta semantikoki ondoren (eskuz) WordNeteko adierekin (Miller *et al.*, 1994).

Euskaraz ditugun corpusen artean azpimarratzekoak dira: batetik, UZEIk Euskaltzaindiarentzat sortutako *XX. Mendeko Euskararen Corpus Estatistikoa*[7] —XX. mendeko testuen laginez osatutako 4.650.000 hitzeko corpus estatistikoki lematizatua—, eta bestetik, *Ereduzko Prosa Gaur*[8]. IXA taldean, bestalde, ikerketarako erabiltzen dira *Euskaldunon Egunkaria* eta *Berria* egunkarien hemerotekekin sortutako corpusak, eta egun, informazio linguistiko aberatsa duen euskarako corpusa garatzen ari gara (Aduriz *et al.*, 2006).

Hala ere, corpusak ez dira beti elebakarrak, eta corpus elebidunak sarri erabiltzen dira LNPn. Corpus elebidun batek bi hizkuntza —gutxienez— parekatzeko aukera ematen du. Corpus elebidun hauek lerrokatuta baldin badaude, gainera, hizkuntza bateko esaldia beste hizkuntzako esaldi baliokidearekin parekatzeko aukera ematen digute. Honek, noski, itzulpen automatikorako eta antzeko aplikazioetarako baliagarri egiten ditu modu honetako corpusak.

Corpus elebidunei dagokienez, ikertzaileen artean gehien erabili izan dena *Hansard* corpusa izan da. Corpus honetan *Canadian Parlamentary Proceeding*ak daude, eta ingelesa eta frantsesa dira bertan aurkitzen ditugun hizkuntzak. Corpusak 3,5 milioi esaldiri dagozkien 97 milioi hitz ditu. Corpus lerrokatua da, hau da, hizkuntza bateko esaldi bakoitzaren beste hizkuntzako esaldi baliokidea zein den markatuta dago (Brown *et al.*, 1991).

[6] Semantikoki markatutako/etiketatuko corpusean, hitzak dagokien adierarekin desanbiguatuta daude. Hala, corpus bat *(semantikoki) etiketatua* dagoela diogunean, *(semantikoki) desanbiguatutako* corpus bat dela adierazi nahi dugu.

[7] `http://www.euskaracorpusa.net` (2007-07-02an atzitua).

[8] `http://www.ehu.es/euskara-orria/euskara/ereduzkoa` (2007-07-02an atzitua).

Corpusa egitura gabeko testu gisa definitu ondoren, ikus dezagun **egituratutako baliabide lexikalen** artean zer testu mota aurki daitezkeen: *makinak irakurtzeko moduko hiztegi* (MRD) elebakarrak eta elebidunak, *thesaurusak* eta *entziklopediak*. Guztiak baliabide egituratuak diren arren, corpus egituratuekin antzik ez dute, hauetan dagoen informazioa eta egitura oso ezberdinak direlako. Hiztegi, entziklopedia eta thesaurusetan hitzen kategoria, azpikategorizazioa, definizioa, erabilera-adibideak, eta abar aurki daitezke. Gainera, hitzen esanahiak antolatuak daude adieren bidez. Hiztegi elebidunen informazioa ere ustiatzen da, hizkuntza batetik besterako ordainak lortzeko adibidez.

Nahiz eta autore batzuk corpusak aztertzearen aldekoak izan —besteak beste, Grishman eta Sterling (1992)—, MRDak hartu izan dira nagusiki iturri lexikal aberatsentzat. Halaxe diote, behintzat, Donal Walker-ek eta Antonio Zampolli-k *Computational Lexicography for Natutal Language Processing* liburuaren sarreran:

> "The various kinds of existing dictionaries, and in particular the dictionaries available in machine-readable form, are obviously the richest and most valuable sources, based as they are on a long lexicographical tradition which encompass a treassure store of data, information and knowledge."
>
> (Boguraev eta Briscoe, 1989, xiv or.)

Hiztegietatik informazioa erauzteko metodoa ez da berria LNPn, eta honi buruzko laburpen interesgarriak ditugu Castellón (1992), Artola (1993), Agirre (1999), Rigau (1998) eta Arriola (2000) lanetan. Halako lanak 80ko hamarkadan hasi ziren. Amsler-en hainbat lanetan (Amsler eta White, 1979; Amsler, 1980) dagoeneko aipatzen da halakorik. Ondoren, *The Merriam-Webster New Pocket Dictionary* —Chodorow *et al.* (1985); Binot eta Jensen (1987), eta abar— eta *Longman Dictionary of Contemporary English* (LDOCE) —Michiels eta Nel (1994); Boguraev eta Briscoe (1993), besteak beste— hiztegien gainean egindako lanak argitaratu ziren. Hala, LNPko ikertalde askok jardun dute MRDez baliatzen, joan den hamarkadan.

Hiztegietatik ez ezik, badira egituratutako beste baliabide lexikaletatik informazioa erauzten duten lanak ere: Yarowsky-k (1992) eta Resnik-ek (1995), beste batzuen artean, *Roget's International Thesaurus*a erabili dute. Grefenstette-k (1993) *Macquarie's thesaurus*a erabili du. Sánchez-ek (1991) *Diccionario Ideológico de la lengua Española* thesaurus espainiarra erabili du. Eta Utsuro *et al.*-ek (1993) *Bunrui Goi Hyou* thesaurus japoniarra erabiltzen dute.

Entziklopediei dagokienez, berriz, Yarowskyk (1992) lexikoien sorkuntzarako *Grolier's Encyclopaedia* erabili du; eta Gómez *et al.*-ek (1994) *The World Book Encyclopedia* erabili dute.

Baina badira bi motatako metodoak erabiltzen dituzten lanak, eskuzko erauzteko metodoa eta automatikoa tartekatzea lexikoietako hutsuneak eragozteko asmoz. Esate baterako, biltegi lexikalak eraikitzerakoan iturri bakartzat hiztegiak ez direla erabili behar diote Ide eta Veronis autoreek (1993). Autore hauek ondorioztatzen dutenez, biltegi lexikalak eraikitzeko hiztegiak oso garrantzitsuak dira, baina, zenbaiten ustearen aurka, ezingo dira erabat automatikoki sortu, eta pertsonaren lana ezinbestekoa izango da, hainbat arazo ekiditeko.

Eskuratze-metodoak konbinatzeaz gain, bi iturriak elkarrekin erabili izan dira. Arrazoi nagusia da hiztegiek ez dutela —corpusek bai ordea— hitzen maiztasun erlatiboa jasotzen eta ezta hainbat testuingurutan hitzek duten erabilera ere. Horregatik, askotan bi iturriek emango duten informazioa uztartzea komeni da. Baliabideak konbinatzen dituzten lanak modu honetan sailkatzen dira:

- Iturri lexikal egituratuak konbinatzen dituztenak (Knight eta Luk, 1994): MRDak, ontologiak, thesaurusak, eta abar.

- Iturri egituratuak eta ez-egituratuak baliatzen dituztenak (Klavans eta Tzoukermann, 1996).

Beraz, lexikoiak sortzeko garaian hurbilpen eta iturri ugari daude, eta ondorioz, erabilitako hurbilpen eta iturri hauen arabera hainbat lexikoi mota lor daitezke. Hurrengo atalean, lexikoi mota nagusienak gainbegiratuko ditugu.

II.2.2 Ezagutza-base lexikalak, hiztegi ezagutza-baseak eta ontologiak

II.1 atalean esan bezala, hizkuntzalaritza konputazionalaren gaur egungo joeraren arabera hizkuntza-ezagutza gramatikaren arlotik lexikoarenera lerratu da, eta ikusmolde-aldaketa horrek gramatikak erraztea ekarri du. Baina informazioa lexikoan pilatzeak sarrera lexikalak informazio erredundanteaz hornitzea ekar lezake. Informazioaren kopuruak eta konplexutasunak informazioa bera kontrolatzeko arazoak sor ditzake. Beraz, beharrezkoa izango

da, sarrera lexikalek zein motatako informazioa behar duten erabakitzeaz gain, informazio hori guztia nola egituratu erabakitzea, erredundantzia ekiditeko eta portaera bereko hitz moten arteko pareko ezaugarriak antzemateko. Arazo horiei erantzuteko *ezagutza-base lexikalak* (EBLak)[9] garatzen dira.

Hala, EBLak ezagutzari buruzko informazioa gordetzen duten gordailu egituratuak dira. Amsler eta Walker egileek aipatzen dute EBLaren kontzeptua estreinako aldiz 1981-1982 tartean. Izan ere, lengoaia naturalen prozesamendu sintaktiko eta semantikoa egin ahal izateko, lexikoiak hitz-zerrenda izatetik ezagutza-base lexikal izatera pasatu behar dira, hitzei eta adierei buruzko informazioa duten ezagutza-base konplexuetara, alegia. Hala, ezagutza-base hauetan, entitateak eta beraien arteko erlazioak agerikoak dira, semantika lexikala errepresentatuz.

EBLen ezaugarri garrantzitsuena herentzia izaten da, adierak klase-azpiklase hierarkietan antolatzen dira-eta (Copestake, 1990). Esate baterako, WordNet —IV.1 atalean aztertuko duguna— hierarkia semantikoaren bidez antolatua dago. Hortaz, hitz moten hierarkia eta herentziaren nozioa EBLen ezaugarri garrantzitsuenetakoa da, eta hauei esker, mota bereko elementuek ezaugarri berak konpartituko dituzte. Horrela, herentzia-mekanismoak eta erregela lexikalak baliatuz, informazio lexikalaren erredundantzia ekiditea eta kontsistentzia bermatzea lortzen da. Esate baterako, ale lexikalak errepresentatzeko *Qualia Structure* teoria garatzen du Pustejovskyk (Pustejovsky, 1991). Teoria horren bidez, hitzek dakarten polisemia sistematikoki adierazten da lexikoian, behar ez den anbiguotasun lexikala ekidinez. Horrez gain, autore horrek dio egitura lexikal bakanak EBL zabalago batean integra daitezkeela herentzia lexikalaren teoriari esker. Teoria horrek lexikoiaren antolamendu orokorrerako behar diren printzipioak ditu, eta gure hizkuntzaren lexikoia osotasun kontzeptual batean integratzen laguntzen digu.

Bestalde, lexikoietako informazioa adierazteko ezaugarrien bidezko adierazpidea usu erabiltzen da. Ezaugarriak erabiltzen dituzten lexikoiekin, garai batean gramatika-erregeletan islatutako informazio kopuru handia maila lexikora lerratzea lortzen da. Hauen alde egiten dutenek argudiatzen dute informazio lexikalaren egitura konplexua herentziaren bidez errepresentatzea oso zaila izan daitekeela eta egokiagoak direla datu lexikalak errepresentatzeko ezaugarri-egituretan oinarritutakoak. Ematen dituzten arrazoiak hurrengoak dira (Aldezabal *et al.*, 2005):

[9] Ingelesez *Lexical Knowledge Base* (LKB) terminoa erabiltzen da.

- Informazioa atzitzeko eta maneiatzeko bide anitz.
- Hiztegi jakin baten antolaketa gordetzen ahal da, kontsultarako *transparente* eginez.
- Oinarri teoriko sendoa.
- Lexikoi konputazionalekiko bateragarritasuna.

Formalizazio honetan oinarritutako formalismo ugari garatu da, hala nola, LFG (*Lexical Functional Grammar*), HPSG (*Head-Driven Phrase Structure Grammar*), CUG (*Categorial Unification Grammar*) edo FUG (*Functional Unification Grammar*). Hurrengo kapituluan aztertuko ditugu sakonkiago hauetako batzuk.

Aurreko atalean aipatu dugun bezala, EBLak eskuz eraiki daitezke, adibidez, WordNet (Miller, 1985; Fellbaum, 1998a) eta EDR (Yokoi, 1995), baina askotan hiztegietatik erauzten dira (Copestake, 1990; Bruce *et al.*, 1992). EBLak eraikitzeko hiztegietatik erauzi izan den informazioz baliatuz gero, *hiztegi ezagutza-baseez* (HEB) hitz egiten da. Hortaz, HEBek hiztegietatik erauzitako informazioa jasotzen dute (Artola, 1993). EBLetan bezala, erauzitako informazioaren artean, adieren hierarkiak dira aipagarriak, baina HEB baten garrantzia hiztegiko informazioan datza. Hala ere, EBL batean dugun informazioa ez da hiztegi batean dugun bera, hiztegietako informazioaz gain, bestelako informazioa ere egoten baita; hala nola, sarrera lexikalen arteko lotura semantikoak, eta sarrera lexikalari buruzko hainbat informazio semantikoa (eremu semantikoa, adibidez) edo sintaktiko-semantikoa (rol tematikoak, adibidez).

Ontologiak, munduari buruzko ezagutzaren biltegiak dira, hau da, mundu errealaren kontzeptualizazioak dira, mundu errealari buruzko inferentziak egiteko gaitasuna dutenak. Gizakiok ezagutza hori lexikoaren bidez adierazten dugunez, baliabide lexikalen artean ere sarri aipatzen dira. Ontologiak aplikazio askotarako eraiki izan dira —softwarearen berrerabilgarritasuna, medikuntzako sistema adituak, hizkuntzaren sorkuntza, ulermena, itzulpena, eta abar—, eta normalean eremu espezifikoetarako eraiki ohi dira.

Ontologien izaera ez dago guztiz zehaztuta eta eztabaida handia dago honen definizioaren inguruan. Gruber (1993), Onyshkevych eta Nirenburg (1994) eta Guarino (1997) bat datoz ontologiak oso heterogeneoak eta norberearen beharren arabera eginak direla esaterakoan. Hala ere, ontologia guztiek dute kontzeptu zerrenda bat, eta kontzeptu horien arteko hierarkia

klase/azpiklase erlazioak egituratzen du. Hori da ontologien ezaugarririk garrantzitsuenetakoa.

Ontologien izaeraren inguruko eztabaidak zerikusia dauka EBL eta ontologien arteko mugak oso garbi ez egotearekin. Autore batzuk EBL eta ontologien arteko ezberdintasuna azpimarratzen saiatu diren arren, gu Lersundiren (2005) lanean defendatzen den ikuspegiarekin bat gatoz. Lan honetan, diferentzia nagusia orientazioan dagoela nabarmentzen da:

> "Ontologietan munduari buruzko informazioa dugu, kontzeptuen arteko erlazioek ez dute zertan motibazio linguistikorik eduki. Bestalde, EBLek hizkuntzaren ulermenerako eta sormenerako beharrei erantzun nahi diete, baina, azken finean, jakina da LNPren muturrera iristeko hizkuntzan agertzen diren arazo guztiak gainditu beharko direla, sen ona barne. Beraz, EBLetan munduari buruzko informazioak egon behar du. Adibide garbi bat hiperonimia erlazioa da. Izan ere, ontologietan eta EBLetan gordetzen den informazio semantikoa gainjarri egiten da; biak egitura isolatu bezala diseinatuko balira, ezagutza bera bi aldiz errepresentatu beharko litzateke, adibidez, hiperonimiari dagokion ezagutza." (Lersundi, 2005, 26. or.)

II.3 Laburbilduz

Kapitulu honetan lexikoien ibilbidea azaldu dugu, LNPn hartu duen garrantzia azpimarratuz. Horren adierazgarri dira, kapituluan zehar ikusi ahal izan dugun bezala, azken urte hauetan honetan egin diren lanak.

Bestalde, lexikoien garapenean dauden joerak aurkeztu ditugu (hurbilpen arauemailea eta deskriptiboa). LNPn bigarrenaren alde egin da, aldez aurretik sortuta dauden testu-baliabideetan (corpusak, MRDak, thesaurusak eta entziklopediak) dagoen informazioa ustiatzeko aukera ematen duelako.

Azkenik, hiru lexikoi mota ikusi ditugu: ezagutza-base lexikalak (EBLak), hiztegi ezagutza-baseak (HEBak) eta ontologiak. Gaur egun EBLa da LNPn lexiko-semantikaren arloan nagusitzen dena. Honek sarrera lexikaletako informazioa egituratu egiten du, erredundantzia konponduz, datuen kontrola eta kontsistentzia gauzatuz eta informazio-atzipena erraztuz. Horretaz gain, informazioa lexikala EBLetan gordez gero, EBLak eskaintzen dituen aukerei esker informazioaren mantentzea eta eguneratzea, eta bertsio desberdinen sorkuntza, besteak beste, oso modu ziurrean egin daitezke. Hortaz, ezagutzaren errepresentaziorako eta biltegirako oso egokia da

Arrazoi hauek guztiengatik, eta tesi-txosten honen izenburuak adierazten duen bezala, lan honetan EBLak izango dira aztergai. Euskararen azterketa

semantikoa ahalbidetzeko, euskararen informazio lexiko-semantikoa jasotzen duen lexikoia ezagutza-base gisa diseinatu dugu. Hala ere, esan beharra dago, IXA taldean honekin batera, paraleloki, euskararako HEB garatzen ari garela (Lersundi, 2005).

III. KAPITULUA

Ezagutza-base lexikalen azterketa kritikoa

Behin gure lexikoiak ezagutza-base lexikal (EBL) bat izan behar duela erabaki ondoren (irakurri berri duzuen atalean), eman beharreko lehenengo urratsa, erabaki beharreko EBL mota zehaztea da. Horixe egingo dugu kapitulu honetan: euskararako aukeratu dugun EBLa arrazoitu, eta egin nahiko genukeen EBLaren ezaugarriak zerrendatu.

II.2. atalean azaldu dugun bezala, informazio linguistikoa eredu edo formalismo jakinetan oinarrituta errepresentatzen da EBLetako sarreretan. Honenbestez, euskarako EBLa egiten hasi baino lehen, eredu edo formalismo horiek aztertu ditugu, ondoren guk eredu bat proposatzeko. Ikusiko dugun bezala, EBLen eraikuntzarako eredua ugari daude, eta ikerlan honen ezinbesteko muga dela-eta, azterketaren esparrua murriztu egin behar izan dugu. Hortaz, lehenik eta behin, aukeraketa horren zergatia azalduko dugu, eta, ondoren, formalismo bakoitzetik ezaugarri nagusienak aipatuko ditugu[1].

Formalismo hauek aztertu ondoren, IXA taldearen beharretara gehien egokitzen den EBL formalismoak *WordNet* eta honen ildotik garatu diren *EuroWordNet* eta *The Multilingual Central Repository* (MCR) direla arrazoituko dugu (III.3).

[1] Tesi-txosten honetan ez dugu formalismo bakoitzaren azalpen sakonik egingo. Eredu horien azterketa sakona eta azterketarako erabilitako metodologia eta irizpideak ezagutzeko, jo bedi Pocielloren lanera (2004b).

III.1 Gure EBLa definitzen

Euskararako nahi dugun EBLaren ezaugarriak definitzeko hainbat erabaki hartu behar izan ditugu: zein formalismoren arabera jasoko duen informazioa, zein informazio mota txertatuko dugun sarrera bakoitzean, non erabili nahi dugun, eta abar. Ataza honetan zenbait zailtasunekin topatu gara.

Batetik, EBLak egiteko eredu edo formalismo asko dago. II.1 atalean deskribatu dugun bezala, 1980ko eta 1990eko hamarkadetan lexikoien inguruan garatutako lanen gorakada gertatu zen, aurreikusitako estrategiarik edo formalismo garbirik izan gabe. Hortaz, lexikoa aztertzeko hamaika era desberdin erabili ziren. Horren adierazgarri dira bai hizkuntzalaritza teorikoa eta baita hizkuntzalaritza konputazionala ere. Esate baterako, 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 proposamen ugari ditugu (Bresnan eta Kaplan, 1982; Fillmore eta Baker, 2001; Miller, 1985; Kipper *et al.*, 2000, beste batzuen artean). Horietako asko fenomeno linguistiko zehatz bati mugatuak daude.

Formalismo-aniztasunari lotuta, aipatu beharra dago EBLetan maiz ez dagoela adostasunik ez hauek jaso behar duten informazioan, ez informazio hori errepresentatzeko moduan ere (Ingria, 1988). EBL baten diseinua definitzean, fenomeno linguistikoak zehaztu behar dira aldez aurretik, baina hauek ez daude argi. Esaterako, iritzi ezberdinak daude ale lexikalen izaera semantikoa definitzerakoan: ale lexikalak berezko semantika du ala testuinguru eraginaren ondorioz jasotzen du semantika hori? Hori horrela izanda, zein ezaugarri dira ale lexikalean berezkoak eta zeintzuk dira testuinguruarn eraginaren ondorioz sortutakoak?

Honen adierazgarri, adibidez, aditzen diatesi-alternantziak dira[2]. Demagun hautsi aditzaren sarrera lexikala lantzen ari garela, eta Leihoa hautsi da eta Maiderrek leihoa hautsi du bezalako esaldiak ditugula. Aditz honen argumentuak era ezberdinean azaleratu dira, eta arrazoi horregatik, bi esaldi hauetako esanahia ezberdina da: lehenengoan ‘norbaitek hausten dut leihoa’ eta bigarrenean ‘leihoa hautsi egiten da’. Honenbestez, hautsi aditza EBL

[2] *Alternantzia* kontzeptua definitzea ere ez da zailtasunik gabeko auzia. Levinek (1993), esaterako, horrelaxe azaltzen ditu: “Diathesis Alternations: alternations in the expressions of arguments, sometimes accompanied by changes of meaning.” (Levin, 1993, 2. or.)

batean adierazteko garaian, erabaki beharrekoa litzateke aditz honek berezko bi adiera dituen; ala berezko adiera bakarra duen, eta beste bi adierak testuinguru sintaktikoaren eraginez sortu diren. Hau horrela izanda, erabaki beharreko hurrengo gauza litzateke zein ezaugarri diren ale lexikalean berezkoak, eta zeintzuk testuinguruaren eraginaren ondorioz sortutakoak.

Ikus daitekeen bezala, semantika eta sintaxiaren arteko bereizketa ez da hain argia, eta gaur egun gauza onartua da bi atal hauen artean harremanik izan badela. Dena den, harreman hori nola gauzatzen den oso arazo eztabaidatua da. Bi maila hauen arteko lotura hori bideratzeko *sintaxi-semantika elkarguneaz* hitz egiten da.

> "In short, we come to see semantics not as *derived* from syntax, but as an independent generative system correlated with syntax through an interface."
> (Jackendoff, 2000, 124. or.)

Semantika eta sintaxiaren arteko harreman hau dela-eta, EBL batzuk ale lexikalen izaera semantikoa definitzeko, ezaugarri semantikoaz gain, ezaugarri sintaktiko-semantikoez ere baliatzen dira; hala nola, rol tematikoez, azpikategorizazioaz, eta hautapen-murriztapenez, besteak beste. Ezaugarri hauek, gainera, lexikoiko sarreretako informazioa orokortzen lagunatzen dute:

> "[...] consideramos que la interfaz sintáctico-semántica abarca conjuntos de piezas léxicas y que es factible organizar el léxico verbal en función de este criterio. En concreto, el objetivo final es conseguir determinar toda aquella información que pueda ser generizabla a un grupo de piezas léxicas verbales [...] con la intención de minimizar al máximo el contenido de una entrada léxica." (Vázquez *et al.*, 2000, 41. or.)

Zailtasun hauez guztiez jabetuta, eta nolabait hauek eragoztearren, euskararako EBLaren diseinua irizpide batzuetara mugatu dugu eta ereduak ondorengo baldintzak bete beharko dituela erabaki dugu:

- **Ahal dela, teoria edo ikerlan bakar bati lotua ez dagoen eredua izatea, hau da, beste eredu edo formalismo batzuetatik edan dezakeen EBLa izatea:**

Aipatu dugun legez, EBLaren eraikuntzarako ez dago eredu bakarra, ez hizkuntzalaritza teorikoan ezta konputazionalean ere; eta izatez, eredu bakarra jarraitzen duen EBLra mugatzea arriskutsua izan daiteke askotan, EBLan jasotako informazioa ez delako berrerabilgarria. Ondorioz, aplikazio berrien

sorkuntza baldintza daiteke. Beraz, ahalik eta *irekiena* eta *deskriptiboena* den eredua interesatzen zaigu. EBLa *deskriptiboa* bada, bertan jasoko den informazioa ez da arau-emailea izango eta EBL *irekia* izaten ahalbidetzen du. *Irekia* diogunean hauxe adierazi nahi dugu: aukeratutako eredu horretatik gertu beste lan konputazionalak egotea, gure EBLa horien informazioarekin ere aberastu ahal izateko. Hala, gure EBLa informazio berrerabilgarria jasotzen duena izatean nahi dugu, eta bertan egindako deskribapen linguistikoak etorkizuneko aplikazioak ez baldintzatzea.

- **Hizkuntza bere osotasunean adierazten duen EBLa izan behar du; ale lexikal bakoitzari dagokion adiera, klase semantikoa eta informazio sintaktiko-semantikoa zehaztuta dituen EBLa**:

Hizkuntzalaritza konputazionalaren ikuspegitik, geroz eta lexiko aberatsagoa izan, orduan eta emaitza hobeak lortzen dira ataza konputazionaletan. Guretzat, Pustejovsky-ren (1993) ildoa jarraituz, lexikoa aberatsa da baldin eta:

(a) Sarrera lexikalaren edukia oso landuta badago; hau da, sarrera horri dagokion informazio guztia egokiro adierazita badago.

(b) Lexikoaren antolaketa oso landuta badago, hots, lexikoa osatzen duten sarrerak beraien artean harreman egokiekin lotuta badaude.

Lehenengoak, sarrera lexikal zehatz bati dagokion informazio guztia eskuratzea ahalbidetzen du. Bigarrenak, berriz, hizkuntza bera ulertzeko behar diren inferentziak eskaintzen dizkigu, ale lexikalen arteko harremanari esker.

Hortaz, gure EBLak ahalik eta informazio gehiena jasotzea nahi dugu.

- **Konputazionalki inplementa daitekeen EBLa izatea, hots, LNPn erabilgarria. Honetaz gain, LNPko aplikazio bat baino gehiagorako baliagarria izatea, hau da, helburu askotarako baliagarria izatea.**

- **Eleanitza den EBLa izatea:** Euskarako sarrera lexikalez gain, beste hizkuntzetako ordainak eskuragarri dituena.Erabilera konputazionalari begira, oso egokia da ezagutza-baseak eleanitzak izatea, batik bat informazio-erauzketa elebakar eta elebidunerako, eta baita itzulpen automatikorako ere.

Laburbilduz, beraz, IXA taldeak nahi duen EBLak:

- euskal hizkuntzako ale lexikalen ahalik eta informazio gehien jaso behar du
- beste ereduetako informazioarekin bateragarria izan behar du
- IXA taldeko aplikazio ezberdinetan erabilgarria izan behar du: itzulpen automatikoa, sintaxi zuzentzailea, galdera-erantzun sistema, hitzen adieren desanbiguazioa, edo hizkuntzen arteko informazioaren bilatzailea
- EBL eleanitza izan behar du

III.2 Azterketarako aukeratutako formalismoak

EBL baten diseinurako proposamen ugari daude, eta hizkuntzalaritza konputazionalaren kasuan, proposamen hauek arloetan (sintaxian, semantikan, morfologian...) sakabanatzen dira. Egoera honen aurrean, eta ikerlan honen ezinbesteko muga dela-eta, azterketaren esparrua murriztu behar izan dugu.

Bereziki aztertu nahi ditugu semantika eta sintaxia aztertzen dituzten lanak, bi hizkuntza maila hauen arteko elkarreragina onartuta. Hala, sintaxia, semantika eta sintaxi-semantika elkargunea hiztegi baten bitartez azaltzen saiatu diren lan batzuk aztertuko ditugu. LNPren arloan jorratuak izan direnak interesatzen zaizkigu bereziki, baina askotan hauek lan teorikoetan oinarrituak daudenez, garrantzitsua iruditu zaigu lan teoriko hauen ezagutza ere izatea. Hortaz, hizkuntzalaritza teorikoko eta konputazionaleko formalismoak sartzen saiatu gara. Hala ere, formalismo batzuk ezin dira argi eta garbi ikuspegi baten pean kokatu. Hala, lau azpimultzo egin ditugu: *Hizkuntzalaritza teorikoan oinarritutako lanak* (III.2.1 atala), *Hizkuntzalaritza teoriko eta konputazionalaren erdibidean dauden lanak* (III.2.2 atala), *Hizkuntzalaritza konputazionalean oinarritutako lanak* (III.2.3 atala) eta *Corpusetan oinarritutako lanak* (III.2.5 atala)[3]. Azter ditzagun azpimultzo bakoitzeko ikerlanak.

[3]Hemen azpimultzo hauei buruzko puntu nabarmenenak azalduko ditugu, azalpen osorako, jo bedi Pocielloren lanera (2004b).

III.2.1 Hizkuntzalaritza teorikoan oinarritutako lanak

II.1 atalean aipatu dugun bezala, Gramatika Sortzailean eta Gobernu eta Uztarduraren Teorian, hizkuntzaren gaitasun sortzailea sintaxiari esker gertatzen da hein handi batean. Semantika eta fonologia, izan ere, sintaxiaren menpe dauden interpretazio mailak baino ez dira. Ikuspegi hau *sintaktozentrismo* bezala ezagutu izan da.

Beste ikuspegi berri batzuk ere badira lexikoan ere erregulartasunik badela argudiatzen dutenak. Erregulartasun hauek, hain zuzen ere, semantika eta sintaxiaren artean elkarreragina dagoen ideiatik etorriko dira. Hortaz, sintaktozentrismo ideiaren aurkako ikuspegiak dira. Horixe da Jackendoff (1990), Levin (1993) eta Pustejovsky (1995) autoreen kasua, hementxe aztertuko ditugunak.

Autore hauen ustez, ale lexikalek ezaugarri mota desberdin ugari dute beren baitan, eta ezaugarri horien guztien arteko harremanek ale lexikalaren gauzapen sintaktiko egokia baldintzatzen dute. Ikuspegi honekin, lexikoaren azterketa bilakatzen da aztergai nagusi, eta prozedura sintaktikoak horien arabera definitzen dira.

Autore hauen lanek oihartzun handia izan dute (gaur egun ere hala dute) hizkuntzalaritza konputazionalean, eta hauetatik abiatuta LNPrako lan ugari egin dira. Esate baterako, Dorr (1997, 1993) eta Fernández *et al.* (2002) Jackendoffen (1990) ereduan oinarritu dira; Buitelaar (1998) Pustejovskyrenean (1995), eta Saint-Dizier (1996) eta Poznanski eta Sanfilippo (1993) Levinenean (1993). Lan hauei buruz arituko gara autore bakoitzari eskaini diogun atalean.

Ikus ditzagun, bada, oso labur, autore hauen lexikoaren adierazpen proposamenak.

III.2.1.1 Jackendoff (1990)

Autore honek adierazpen-eredu abstraktu bat proposatzen du: *Egitura Lexikal-Kontzeptuala* (ELK)[4].

Egitura hau, batetik, hainbat egitura primitibo semantikoz osatzen da (*TO*, *FROM*, *TOWARD*, *AWAY-FROM*, *CAUSE*, *GO*, *VIA*...) eta bestetik, hainbat kategoria kontzeptualez (*Thing*, *Event*, *State*, *Action*, *Place*, *Path*, *Property*, *Amount*...). Egitura primitibo semantikoak kategoria kontzeptual horiekin lotzen dira. Adibidez, *TO*, *FROM*, *TOWARD*, *AWAY-FROM*

[4] *Lexical Conceptual Structure* (LCS).

eta *VIA* primitiboek *Path* kategoria kontzeptuala adieraz dezakete; eta *GO*, *STAY*, eta *CAUSE* primitiboek, berriz, *Event* kategoria kontzeptuala.

Kategoria sintaktikoak kategoria kontzeptualei lotzen zaizkie. Alegia, izen-sintagma batek *Thing* (the dog), *Event* (the war) edota *Property* (redness) kategoria kontzeptualei erreferentzia egin diezaieke, eta ildo beretik, preposizio-sintagma batek, *Place* (in the house), *Path* (to the kitchen) edota *Property* (in luck) kategoria kontzeptualei[5]. Primitibo semantikoak, beraz, aditzaren argumentuei lotzen zaizkie.

$$\left[\begin{array}{l} \text{run} \\ \text{V} \\ \underline{\quad}<\text{PP}_j> \\ \left[_{Event}\quad \text{GO}\ (\left[_{Thing}\quad\right]_i\quad\left[_{Path}\quad\right]_j)\right] \end{array}\right]$$

III.1 Irudia: run aditzaren ELKa.

III.1 irudian ikus daiteke run aditza Jackendoffen sarrera lexikal gisa[6]. Sarrera lexikal honek *GO* primitiboa du, eta Jackendoffek primitibo honekin definitzen ditu mugimenduzko egitura kontzeptualak[7]. Run mugimenduzko aditza izaki, bi argumentu eskatzen ditu: batetik, mugitzen den *gaia* (*Thing*) eta bestetik, mugitzen den horrek egin behar duen *ibilbidea* (*Path*). Lehenengoa i azpindize batez markatuko da (subjektua)[8] eta bigarrena, berriz, j azpindize batez, PSaren (PP) osagarria dela adieraziz. Azken hau, aukerazkoa izan arren, lexikoan agertzen da.

Esan dezakegu, beraz, lexikoan egitura kontzeptualaren eta sintaktikoaren arteko korrespondentzia gauzatzen dela, eta korrespondentzia hori ale lexikalen sarreretan ageri da.

[5] Adibideak Jackendoffen lanetik (1990) hartu dira.

[6] Txostenean aztertuko ditugun adibideak aztergai ditugun lanetatik hartutakoak dira. Hauetan autoreek erabiltzen duten terminologia agertzen denez, testuan hauek erabiliko ditugu. Bestalde, kontuan izanda autore hauen lanak ingelesez daudela, hizkuntzalaritzako termino arruntak (kategorien izenak-eta bezalakoak) adibidean ere ingelesez agertuko dira. Hala, nahiz eta azalpenean euskarako baliokideak erabili, adibideen azalpena ulerkorragoa egin ahal izateko euskarakoaren jarraian, hauen ingeleseko ordaina ere aipatuko dugu.

[7] *GO* primitiboa beti egongo da *Event* kategoria kontzeptualean: [EVENT] = [Event GO([Thing],[Path])].

[8] Jackendoffek i eta j azpindizeekin subjektu eta objektuen guneak adierazten ditu, hurrenez hurren (Jackendoff, 1990, 45. or.).

Jackendoff (1990) sintaxi-semantika elkargunearen adierazpenaz arduratu zenez, ELKak sortu zituenean arreta berezia jarri zion azpikategorizazioari, batez ere, aditzei eta preposizioei; beste kategoriak (izenak, adjektiboak eta adberbioak) alde batera utzi zituen. Adiera bigarren mailan dago lan honetan, hots, hitzen anbiguotasun semantikoa ez zuen esplizituki kontuan hartu.

Adierarekin bezala, klase semantikoak ere ez ditu esplizituki lantzen, nahiz eta batzuen berri ematen duen; adibidez, *ukipen-aditzak* (*contact verbs*) aipatzen ditu, baina ez du klase hau osatzen duten aditzen zerrenda ematen.

Horiek horrela, Jackendoffen lexikoaren ezaugarriak (zenbat sarrera dituen, ikusgarri dagoen ala ez, ...) ez ditugu ezagutzen; bai, ordea, honetatik abiatuta egin diren lexikoena. Esaterako, Dorrek (1993, 1997) Jackendoffen lanean oinarritutako aditzen eta preposizioen EBL bat sortu zuen, eta berarekin tutore-sistemak eta itzulpengintza automatikoa landu zituen. Aditzak sailkatzeko Levinen aditz-klaseak (Levin, 1993) erabili zituen eta klase hauek WordNeteko (Miller, 1985; Fellbaum, 1998a) aditzen adieretara lotuak daude. Bere txostenetan adierazten denez, erabilitako lexikoak 4.432 aditz zituen eta 492 aditz-klase. Preposizioei dagokienez, EBL horretan ingeleseko eta espainierako preposizioen interpretazioak (ELKak) ematen dituzte[9].

IXA taldean ere ikerlan batzuk egin dira Dorren lanetik abiatuta. Agirre eta Lersundi-ren lanean (2003) interpretazio berdina duten postposizio inbentario eleanitza sortu dute. Zerrenda honetako postposizioak interpretazioaren arabera multzokatuak daude, hau da, hartzen dituzten rol tematikoen arabera. Gaztelania eta ingeleseko preposizioen inbentarioa eta interpretazioak Dorren lanetik hartu dira, eta euskarakoak aldiz, Aldezabal-en ikerlanetik (2004). Dorren ELKetako interpretazioak Aldezabalenekin parekatu ondoren, ingeleseko, gaztelaniako eta euskarako postposizioen inbentario bakarra lortu dute. Hau oso erabilgarria izan daiteke bai itzulpen automatikorako, bai hizkuntza ezberdinetako postposizioen informazio sintaktiko-semantikoa aztertzeko.

Ildo beretik, *Volem* (Fernández *et al.*, 2002) proiektuak (ikus III.2.3.3 atala) garatutako EBLa dago. EBL hau gaztelaniako, frantseseko eta katalaneko aditz eta preposizioetara mugatzen da, aditz eta preposizio bakoitzaren izaera sintaktikoaren deskribapena (azpikategorizazioa, hautapen-murriztapenak eta alternantziak) eta informazio semantikoa (ELKa, rol tematikoak

[9]Informazio hau guztia, hurrengo web orrian dago ikusgarri: `http://www.umiacs.umd.edu/~bonnie/LCS/Database/Documentation.html` (2007-07-02an atzitua).

eta aditzen WordNeteko klase semantiko nagusiena) ematen duelarik.

Jackendoffen lanetik abiatutako bi EBL hauek Jackendoffen lanari alderdi semantikoa eta beste ikuspuntu teorikoak gehitu arren, aditz eta preposizioetara murrizten dira, eta, ondorioz, hauek ere ez dute hizkuntza bere osotasunean adierazten. III.1 atalean esan dugun bezala, euskararako nahi dugun EBLak, ordea, baldintza hau betetzea nahiko genuke.

III.2.1.2 Levin (1993)

Levinek bere lanean (Levin, 1993) ingeleseko aditzen sintaxia eta semantika sakonki aztertzen ditu. Liburuan bertan landutako aditzen zerrenda ematen du, bakoitzari buruzko informazio sintaktiko-semantikoarekin: klase semantikoa eta diatesi-alternantziak.

Beste teorietatik pixka bat aldendu egingo da, Levinek ez baitu zehazten sarrera lexikalaren itxurak nolakoa izan behar duen. Horren ordez, Levinek sarrera lexikal hori osatzeko bideak eskaintzen ditu.

Baina lan hau ez da harremanik gabeko aditzen klase semantiko eta diatesi-alternantzien zerrenda bat bakarrik; lan honi esker, Levinek aitzindari den hipotesi bat sortu eta erabili baitu: klase semantiko berean dauden aditzek, portaera sintaktiko bera dute (diatesi-alternantzia berak), osagai semantiko berdinak dituztelako. Esaterako, (1) adibideko sing eta chant aditzek, *performance verbs* klase semantikoan daudenez, izaera sintaktiko bera izango dute.

(1) You **sing/chant**. [IS + A]
You **sing/chant** a tune. [IS + A + IS]
You **sing/chant** me a tune. [IS + A + Izord + IS]
You **sing/chant** a tune to me. [IS + A + IS + PS]
You **sing/chant** a tune for me. [IS + A + IS + PS]

Teoria honen arabera, beraz, forma bera baina adiera desberdinak (klase semantiko desberdinak) dituen aditz batek, izaera sintaktiko desberdinak izango ditu. Adibidez, ingeleseko sing aditza, *performance verbs* klase semantikoari dagokionean, (1)eko edozein egitura sintaktikorekin ager daiteke. Aldiz, sing aditza, *verbs of sound emission* klase semantikoan dagoenean, beste adiera bat duenez, izan ditzakeen egitura sintaktikoak hurrengoak izango dira:

(2) A bird **sang** in the trees. [IS + A + PS]
The trees **sang** with birds. [IS + A + PS]
In the trees there **sang** the birds. [PS + Adlg + A + IS]
...

Horrela bada, Levinen teoriaren ardatza alternantziak eta klase semantikoak dira. Aditz batek bere portaera sintaktikoen arabera definituko ditu klase semantikoak, eta ondorioz, klase semantiko horri dagozkion osagai semantikoak.

Inplementazioari begira, Levinen lana erabilia izan da lexiko konputazionalak eraikitzeko, hala nola, *Acquilex* (Poznanski eta Sanfilippo, 1993). Poznanskik eta Sanfilippok ingeleseko diatesi-alternantziak definitu zituzten, ondoren Acquilex ezagutza-basean (Boguraev eta Briscoe, 1989) txertatzeko. Azterketa horren abiapuntua Levinen lana izan zen.

Bestalde, Levinen lanean oinarrituta itzulpengintza automatikoa ere egin izan da, esate baterako, *UNITRAN* (Dorr, 1993)[10]. Dorrek Levinen diatesi-alternantzietatik eta klase semantikoetatik abiatuz, patroi sintaktikoak sortzen ditu, eta horietako patroi bakoitzari Jackendoffen (1990) ELK bat egokitzen dio gutxienez.

Hauetaz gain, aditzen sailkapen automatikoa lortzeko ere erabili da Levinen lana. Saint-Dizierrek (1996), adibidez, Levinen sailkapen semantikoa frantsesera itzuli eta klase bakoitzerako diatesi-alternantziak definitzen ditu.

IXA taldean ere Levinen lana erabili da euskal aditzen azpikategorizazioa jorratzeko (Aldezabal, 2004), nahiz eta lan honetan Levinen teoriak hutsuneak dituela agerian geratu. Gogora dezagun, Levinen teoriak dioela diatesi-alternantzia berdinak dituzten aditzekin klase semantikoak egin daitezkeela. Baina Aldezabalek teoria honen aurka doazen adibideak topatu ditu; hau da, Levinen aditzen klase semantikoak ez dira beti osatzen konpartitzen dituzten alternantzien arabera. Adibidez, Levinek *put verbs* eta *remove verbs* klase semantikoak bereizten ditu. Beraz, Levinen teoriaren arabera, klase semantiko bateko eta besteko aditzek diatesi-alternantzia desberdinak izan behar dituzte. Levinek, aldiz, bi klase semantiko hauek deskribatzen ditu diatesi-alternantzia berdinekin. Aldezabalek Levinen diatesi-alternantzian oinarrituriko teoriaren trinkotasunik eza sakonkiago azaltzen du.

Bestalde, Levinen lanari beste ezaugarri batzuk gehitu bazaizkio ere, aditzen deskribapena soilik egiten duen eredua da, eta, ondorioz, ez du hizkuntza

[10] Argibide gehiagorako ikus Pocielloren lana (2004b).

bere osotasunean adierazten. Hala ere, ingeleseko aditzen deskribapen itzela da.

III.2.1.3 Pustejovsky (1995)

Pustejovskyk (1995) *Lexiko Sortzailea* (*Generative Lexicon*) proposatzen du, eta bere teoria hurrengo hiru hatsarretan oinarrituta dago:

- Egitura sintaktikoa kontuan hartu gabe, ezinezkoa da semantika lexikalean aurrera egitea. Adiera ezin da bere egituratik banatu.
- Ale lexikalaren adierazpenak rol tematikoen deskribapena baino zerbait gehiago izan behar du.
- Semantika lexikalak kategoria guztien adierazpen semantikoak landu behar ditu, eta ez aditzena bakarrik.

Pustejovskyk deskonposaketan oinarritutako teoria darabil, non sarrera lexikalaren deskonposaketa hiru adierazpen-mailatan islatzen den[11]:

- **Qualia-egituran** (*qualia structure*) ale lexikalaren semantika zehazten da.
- **Gertaera-egituran** (*event structure*) ale lexikalaren aspektua zehazten da.
- **Argumentu-egituran** (*argument structure*) ale lexikalaren azpikategorizazioa zehazten da.

Lehenago adierazi dugun bezala, Pustejovskyrentzat, egitura sintaktikoa kontuan hartu gabe ezinezkoa da ale lexikalaren adierazpena egitea. Hortaz, nahiz eta autore honen ustez ale lexikalaren adieraren muina qualia-egituran egon, beste egituretako informazioak mugatu egiten du.

Sarrera lexikalek III.2 irudiko itxura dute. Bertan, ingeleseko open aditzaren sarrera lexikala dugu. Ingeleseko aditz honek bi argumentu eskatzen ditu (1 eta 2 zenbakiekin markatuak), eta hauek zehaztuak datoz egitura bakoitzean. Qualia-egiturako (*QUALIA*) *AGENTIVE* ezaugarriak adierazten digu lehenengo argumentuak bigarrena *irekitzen* duela (*open act*), eta

[11] Alderdi hauetako bakoitza ezaugarri gehiagoz osatua dago Pocielloren lanean (2004b).

$$\begin{bmatrix} \text{open} \\ \text{EVENTSTR} - \begin{bmatrix} \text{E}_1 \text{ - } e_1\text{: process} \\ \text{E}_2 \text{ - } e_2\text{: state} \\ \text{RESTR - } <_x \end{bmatrix} \\ \text{ARGSTR} - \begin{bmatrix} \text{ARG1 - (1)} \\ \text{ARG2 - (2)} \begin{bmatrix} \text{physobj} \\ \text{FORMAL: entity} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} - \begin{bmatrix} \text{dc-lcp} \\ \text{FORMAL - P}\Big[e_2, \text{ o } \Big[\text{TELIC(2)}\Big]\Big] \\ \text{AGENTIVE - open\_act}(e_1, \text{ (1),(1))} \end{bmatrix} \end{bmatrix}$$

III.2 Irudia: open aditzaren sarrera lexikala Pustejovskyren teorian.

ekintza hori bukatua (*telic*) dela. Bestetik, argumentu-egituran (*ARGSTR*) bigarren argumentua gauza fisikoa (*physobj*) eta entitate bat (*entity*) dela zehazten zaigu. Eta azkenik, gertaera-egiturari (*EVENTSTR*) esker dakigu bi argumentuek jasaten duten ekintza prozesu bat dela (*process*) eta honen emaitza egoera (*state*) bat izango dela (atea irekia egotea, alegia).

Beraz, argumentu-egitura eta sintaxia erlazionatuak daude; batetik, azpikategorizatutako argumentuak aukerazkoak diren ala ez adierazten delako, eta bestetik, argumentuak beste maila guztiekin lotuak daudelako (zenbakien bidez). Honenbestez, teoria honetan sintaxi-semantika elkarguneak oso deskribapen aberatsa du.

Pustejovskyk hizkuntzalaritza konputazionalak eta teorikoak elkarren beharra dutela aldarrikatzen duen arren, bere lanean oinarritutako ikerlan konputazional gutxi ezagutzen dugu, eta bere teoriatik abiatuta lexiko erreal gutxi dago. Hortaz, euskararako EBLarentzat nahiko genukeen ezaugarrietako bat ez du. Eredu honetan oinarrituta ezagutzen den inplementazio garrantzitsuenetako bat Buitelaarrena da (Buitelaar, 1998); bere lanaren ondorioz izenen EBLa erdiautomatikoki sortu zen (CORELEX), 126 klase semantiko eta 40.000 izen inguru dituena. Hala ere, CORELEXek ez du Pustejovskyren teoria bere osotasunean adierazten, EBL hau garatzeko Buitelaarrek

Pustejovskyren teoriaren klase semantikoak bakarrik erabili baitzituen[12].

Oro har, hizkuntzaralitza teorikoan oinarritutako hiru ikerlan hauek ordura arte ez zegoen formalismo berri baten adierazle dira. Beraz, ez daude beste formalismoetatik gertu; bakarrak dira, eta hauen ondorengo lanek, inplementazioari begira, formalismo hauek beste formalismo ezberdinekin uztartu dituzte.

III.2.2 Hizkuntzalaritza teoriko eta konputazionalaren erdibidean dauden lanak

Aplikazio konputazionaletan baliatzeko helburuaz sortu diren formalismoen artean, garrantzitsuenak eta erabilienak *Lexical Functional Grammar* (LFG) (Bresnan eta Kaplan, 1982), *Generalized Phrase Structure Grammar* (GPSG) (Gazdar *et al.*, 1985) eta *Head-Driven Phrase Structure Grammar* (HPSG) (Pollard eta Sag, 1994) dira. Teoria hauek hizkuntzalaritza teoriko eta konputazionalaren artean kokatu ditugu, zeren oinarri teorikoak badarabiltzate ere, erabilpen konputazionala buruan zuten.

EBL eredu hauek interesgarriak iruditu zaizkigu, sarrera lexikalean informazio sintaktiko-semantiko ugari dakartelako, eta, gainera, ikuspegi konputazionalaren hastapenak direlako.

Hiru teoria hauek formalismo lexikalak dira eta Gobernu eta Uztardura Teoriaren atalkako egituran[13] oinarritzen dira. Dena den, teoria hauek Gobernu eta Uztardura Teoriarekiko diferentzia nabarmen bat dute: ez dute mugimendu edo transformaziorik; azaleko egitura adierazteko maila bakarra proposatzen da[14].

Hala, formalismo hauek asmo eraikitzaileaz eginak dira, eta testuingururik gabeko gramatiketan oinarritzen dira, egitura sintaktikoak osatzeko baterakuntza-erregelak erabiltzen dituztelarik. Baterakuntza-erregelak aplikatu ahal izateko, sarrera lexikalak ezaugarri-egitura modura planteatzen

[12]CORELEXi buruz argibide gehiago Pocielloren lanean (2004b).

[13]Gobernu eta Uztarduraren Teoria ez da erregela-multzo batez osatutako sistema, baizik eta hatsarre batzuen arabera parametrizatu daitekeen atalkako egitura; hots, gramatika atalka antolatua dago eta hauek hatsarre unibertsalez osatuak daude (Demonte, 1995, 10. or.).

[14]Esan behar da, Programa Minimalista (Chomsky, 1992) ere horretara doala. Eredu berri honek ekonomiaren baldintza hartuko du printzipio nagusitzat; hau da, gramatikako mekanismoak ahalik eta sinpleen, errazen (*minimalisten*) egitearena. Honen adierazle garbia, errepresentazio sintaktikorako maila bakarra eta bi interfaze-maila (Forma Logikoa eta Forma Fonetikoa) planteatzearena da (Sakoneko eta Azaleko mailak alboratuz).

dituzte[15]. Eta ikusiko dugunez, teorien arteko desberdintasun nagusia hautatzen dituzten ezaugarriak antolatzeko moduan datza.

HPSG GPSGren garapena denez, GPSG zaharkitua geratu da. Arrazoi horregatik, tesi-txostenean ez dugu honen berri emango.

III.2.2.1 Lexical Functional Grammar

Izenak adierazten duen bezala, teoria funtzioetan (subjektu, objektu eta antzekoetan) oinarritzen da. Lexikalismoan egin ohi den moduan, LFG esaldian ager daitezkeen egitura sintaktiko guztiak lexikoan zehazten saiatzen da. Ale lexikalak, besteak beste, ondoko informazioa izango du: funtzio gramatikala, kategoria sintaktikoak, eduki semantikoa, azpikategorizazioa, rol tematikoak eta hautapen-murriztapenak.

$$\mathrm{V} \rightarrow \begin{bmatrix} \textit{yawned} \\ (\uparrow \mathrm{PRED}) = \text{`YAWN<SUBJ >'} \\ (\uparrow \mathrm{TENSE}) = \mathrm{PAST} \end{bmatrix}$$

III.3 Irudia: yawned ale lexikalaren adierazpena LFGn.

III.3 irudian, yawned aditzaren egitura funtzionalaren adierazpena dugu eta honetan bi ezaugarri daude: adierari dagokiona (*PRED*), eta denborari dagokiona (*TENSE*). Hauen ondoan, bakoitzaren balioa dator zehazturik: *'YAWN<SUBJ>'* yawn aditzetik datorrela adierazteko eta aditzaren azpikategorizazioa zehazteko; eta *PAST* balioak, yawned iraganean dagoen adizkia dela adierazteko[16]. Bestalde, ↑ ikurraren bitartez, egitura sintagmatikoari buruzko informazioa jasotzen da, ↑ ikurrak ale lexikala menderatzen duen adabegia adierazten baitu. Yawned ale lexikala menderatzen duen lehen adabegia aditza da (*V*).

Orain arte, LFGren alderdi sintaktikoaz mintzatu gara, egitura sintaktikoei erreparatzen dien alderdiez, alegia. Baina teoria honek argumentu-

[15] Testuingururik gabeko gramatikak (*Context Free Grammar*) eta baterakuntza-erregelak erabiltzen dituzten gramatikei buruzko argibide gehiagorako jo bedi Gojenolaren (2000) lanera.

[16] Atal honetako adierazpenak Dalrymple (2001) lanetik hartutakoak dira. Bestalde, irudietako laburdurak eta terminologia LFG teorian erabiltzen diren bezala mantendu ditugu.

egituraren informazioa ere lantzen du. Are gehiago, sintaxiarekin duen harremana zehazten du rol tematikoak funtzio gramatikalekin lotuaz. Bresnanek eta Kaplanek (1982) sintaxi-semantika elkargunearen aurkezpena ondorengo irudian dugu ikusgarri:

$$\text{give} \begin{bmatrix} \text{SUBJ} & \text{OBJ} & \text{OBL}_{goal} \\ \text{—} , & \text{—} , & \text{—} \\ \text{AGENT} & \text{THEME} & \text{GOAL} \end{bmatrix}$$

III.4 Irudia: Sintaxi-semantika elkargunea LFGn (Bresnan eta Kaplan, 1982).

III.4. irudian ikus daitekeen bezala, give aditzak hiru argumentu ditu, eta bakoitzaren rol tematikoak adierazita datoz. Bestalde, rol tematiko hauei funtzio gramatikalak esleitzen zaizkie: *egileari* subjektua, *gaiari* objektua eta *helburuari* zehar objektua. Hortaz, Bresnanek eta Kaplanek funtzio gramatikalak eta rol tematikoen arteko hartu-emana egitura funtzionaleko *PRED* ezaugarrian eransten dute. Beraz, hiztegi-sarreren muina *PRED* ezaugarria da, bertan definitzen baita sarreraren adiera. Hala ere, eremu hau xehetasun gehiagorekin dator aditzaren kasuan, eta, bertan dagoen informazio rol tematikoetara bakarrik mugatzen da semantika.

LFGk inplementazio batzuk izan ditu. Hemen horietako batzuk aipatuko ditugu. Alde batetik, LFG formalismoko egitura funtzionalak erabilita corpus etiketatuak daude, esate baterako Cahill *et al.*-ek (2002) egitura funtzionaleko informazioarekin ingeleseko 100.000 ale lexikal eta 50.000 esaldiko corpusa etiketatu dute erdiautomatikoki. King *et al.*-ek (2003) ere ingeleseko corpus etiketatu bat egin dute, LFG analizatzaile sintaktiko (LNPn *parser* edo *gramatika* bezala ere ezagutzen direnak) bat erabilita eta ale lexikalen dependentziak ere islatzen dituena: *PARC 700 Dependency Bank* (*PARC 700 DEPBANK*)[17].

Horrelako analizatzaile sintaktikoak erabilita itzulpen automatikorako saiakerak ere egin dira, Way (2003) adibidez.

Hala ere, ezin da esan formalismo honen semantika aberatsa denik, zeren eta nahiz eta informazio sintaktiko aberatsa izan, semantika rol tematikoetara mugatzen da.

[17] *PARC 700 Dependency Bank* `http://www2.parc.com/istl/groups/nltt/fsbank/default.html` web orrian dago eskuragarri (2007-07-02 atzitua).

III.2.2.2 Head-Driven Phrase Structure Grammar

Head-Driven Phrase Structure Grammar (HPSG aurrerantzean) formalismoak, Lexical Functional Grammar (LFG) eta Generalized Phrase Structure Grammar (GPSG) teorien eragin handia jaso du. Hortaz, hauetatik abiatutako teoria da. Hala ere, ezin da HPSG aurreko bi formalismoekin parekatu, hau aurrekoen garapena baita; alde batetik, hiztegi aberatsagoa du, eta bestetik, aldarrikapen unibertsalagoak lortzen ditu.

HPGSren adierazpenaren muina *zeinuan* (*sign*) datza. Zeinua informazio fonologikoa, sintaktikoa eta semantikoa jasotzen duen unitatea da. Zeinu hauek matematikako antzeko matrizeekin adierazten dira (*attribute-value matrix* deiturikoekin) non ezaugarri bakoitzak bere balioa duen. Bestalde, zeinuak ale lexikalak edo sintagmak izan daitezke.

$$
\textit{gives}\left[\begin{array}{ll}
\text{CAT} & \left[\begin{array}{ll}\text{HEAD} & \textit{verb}[\textit{fin}] \\ \text{SUBCAT} & <\text{NP}[\textit{nom}]_{(1)[\textit{3rd, sing}]}\ ,\ \text{NP}[\textit{acc}]_{(2)}\ ,\ \text{NP}[\textit{acc}]_{(3)}\ >\end{array}\right] \\
\text{CONTENT} & \left[\begin{array}{ll}\text{RELN} & \textit{give} \\ \text{GIVER} & (1) \\ \text{GIVEN} & (2) \\ \text{GIFT} & (3)\end{array}\right]
\end{array}\right]
$$

III.5 Irudia: gives aditzaren adierazpena HPSGn.

Adibide gisa, irudian[18] gives aditzaren sarrera lexikala dakargu III.5. *CATEGORY* ezaugarriak, hitzaren kategoria adierazteaz gain, honek eskatzen dituen argumentuak ere zehazten ditu. Gives aditz burutua da (*verb*[*fin*] (*finite*) balioekin adierazita) eta hiru argumentu hartzen ditu: 3. pertsonan dagoen izen-sintagma nominatibo bat (irudian *NP[nom1[3rd,sing]]*) eta bi izen-sintagma akusatibo (irudian *NP[acc]2* eta *NP[acc]3*).

CONTENT ezaugarrian ale lexikalaren irakurketa semantikoa zehazten da. Hemen jasoko da ale lexikalak adierazten duen *egoera* esaldi osoaren

[18] Adierazpen guztiak Pollard eta Sag (1994) lanetik hartuak daude. Bestalde, sarrera lexikal hauek matrize osoen laburpen bat dira. Matrize osoen azalpena ikusteko jo bedi Pollard eta Sagautoreen (1994) eta Pocielloren lanera (2004b).

osotasunetik ikusita[19]. III.5 irudian *CONTENT* ezaugarriaren bitartez adierazten zaigu, batetik, ingeleseko gives aditza *give* erlazioarekin harremanetan dagoela, honen rolak *GIVER*, *GIVEN* eta *GIFT* direlarik. Eta bestetik, *GIVER*, *GIVEN* eta *GIFT* rolak 3. pertsonan dagoen izen-sintagma nominatiboari (*NP[nom1[3rd,sing]]*) eta bi izen-sintagma akusatiboei (*NP[acc]2* eta *NP[acc]3*) dagozkiela, hurrenez hurren. Hortaz, azpikategorizazioan dagoen osagarri bakoitza rol batekin lotuta dago, eta lotura hau azpindize berdinekin dator adierazita[20].

HPSG inplementazio handia duen formalismoa da, eta hurrengoak dira erabilera ezagunenak[21].

Bestetik, HPSG formalismoak corpus etiketatuak ditu, ingeleserako (Oepen *et al.*, 2002, edo *LinGO Redwoods* deiturikoa) eta baita beste hizkuntza batzuetarako ere, hala nola, bulgarierarako (Osenova eta Simov, 2003).

Eta bestetik, HPSGk analisi sintaktikoak automatikoki egiten dituen analizatzaile sintaktikoak ere baditu (Minnen, 1999; Nishida *et al.*, 1999; Popowich eta Vogel, 1990; Copestake eta Flickinger, 2000). Esate baterako, Copestakek eta Flickingerrek (2000) ingeleserako analizatzaile sintaktiko bat egin dute, eta honen aplikazioetako bat itzulpen automatikoa izan da. Proiektu horretan bileren egitaraua eta bidaia-erreserbak ziren itzuli beharreko gaiak edo domeinuak.

Hala ere, eta LFGri buruz esan dugun bezala, HPSGn, nahiz eta adierazpen semantikoa eraiki, ale lexikalaren tasun semantikoak rol tematikoetara bakarrik mugatzen dira.

Honezaz gain, hizkuntzalaritza teorikoaren eta konputazionalaren erdibidean dauden lan hauen inguruan, hizkuntzalaritza teorikoko lanei buruz esandako gauza bera errepikatuko dugu: lan hauek ordura arte ez zegoen formalismo berri baten adierazle dira. Beraz, ez daude gainontzeko formalismoetatik gertu, eta bertan egindako deskribapen linguistikoak etorkizuneko aplikazioak baldintzatzen ditu.

[19]HPSGko semantika *Situation Semantics* teorian oinarritua dago (Barwise eta Perry, 1983), eta HPSGko *CONTENT* ezaugarria *Situation Semantics* teoriaren ikuspuntuaren ildotik sortutako ezaugarria da. Teoria honen ideia nagusiena Pocielloren lanean (2004b) dator azalduta.

[20]Rol tematikoak *Situation Sematics* teoriako *egoera* horren ikuspegi desberdinak lirateke.

[21]HPSGren erabileraren berri `http://hpsg.stanford.edu` web orrian ematen da (2007-07-02an atzitua).

III.2.3 Hizkuntzalaritza konputazionalean oinarritutako lanak

FrameNet (Fillmore eta Baker, 2001), *WordNet* (Miller, 1985; Fellbaum, 1998a), *EuroWordNet* (Vossen, 1998), *The Multilingual Central Repository* (MCR) (Rigau *et al.*, 2003), *Volem* (Fernández *et al.*, 2002) eta *PropBank* (Palmer eta Kingsbury, 2003), iturri desberdinetan oinarrituta sortutako EBLak dira. Hau da, EBL baterako hiztegi-eredu bat landu beharrean, besteen ereduetatik abiatuta beraiena sortu dute. Gaur egun, LNPn ikertalde gehienek (nahiz eta beraien ikuspegi teorikoa askotan guztiz bat ez etorri) EBL hauek ezagutu eta erabiltzen dituzte.

Hizkuntzalaritza konputazionalean oinarritutako ikerlan gehiago badaude (Gómez, 1998; Vázquez *et al.*, 2000, eta abar), baina hautatu ditugun ereduetatik nahiko gertu daudenez, ez ditugu azalduko.

III.2.3.1 FrameNet

FrameNet proiektuan (Fillmore eta Baker, 2001) ingeleserako baliabide lexikografikoa eraikitzen ari dira, *Frame Semantics* (Fillmore, 1985) teorian oinarritua eta corpus errealeko datuekin lagunduta. *Frame Semantics*ak aldarrikatzen dituen printzipio nagusienak hauek dira:

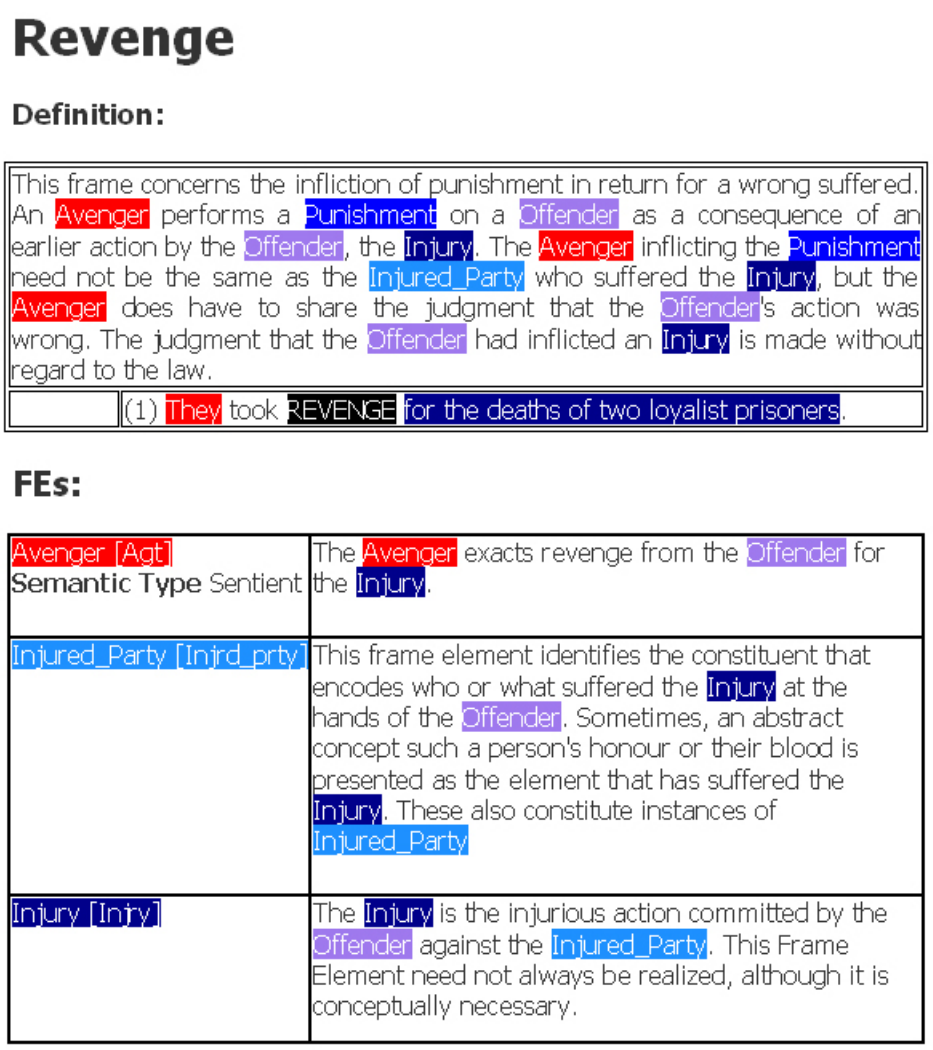
- Ale lexikalen semantika eta funtzio gramatikala ***frame***etatik (egitura kontzeptual aberatsetatik) dator.
- Kontzeptualki erlazionatuak dauden ale lexikalek, ***frame*** bereko alderdi desberdinak erakus ditzakete.

Bi printzipio hauetan oinarrituaz, FrameNeten ale lexikal bakoitza beraiek sortutako *frame*etan sailkatzen dute, batetik, ale honen semantika eta sintaxia definitzeko, eta bestetik, *frame*ko beste osagaiekin duen harremana zehazteko. Teoria honetan sakontzearren har dezagun (3) adibidea oinarri gisa:

(3) Hook tries to avenge himself on Peter Pan by becoming a better father.

Esaldi hau, avenge aditzaren eraginez, *Mendeku*aren esparruari dagokiola esango genuke; hots, *Revenge frame*ari (ikus III.6 irudia).

*Avenger, Injured party, Punishment, Injury... Revenge frame*aren alderdiak edo partehartzaileak dira —*frame elements* (FE hemendik aurrera)



Revenge

Definition:

This frame concerns the infliction of punishment in return for a wrong suffered. An Avenger performs a Punishment on a Offender as a consequence of an earlier action by the Offender, the Injury. The Avenger inflicting the Punishment need not be the same as the Injured_Party who suffered the Injury, but the Avenger does have to share the judgment that the Offender's action was wrong. The judgment that the Offender had inflicted an Injury is made without regard to the law.

(1) They took REVENGE for the deaths of two loyalist prisoners.

FEs:

Avenger [Agt] **Semantic Type** Sentient	The Avenger exacts revenge from the Offender for the Injury.
Injured_Party [Injrd_prty]	This frame element identifies the constituent that encodes who or what suffered the Injury at the hands of the Offender. Sometimes, an abstract concept such a person's honour or their blood is presented as the element that has suffered the Injury. These also constitute instances of Injured_Party
Injury [Injry]	The Injury is the injurious action committed by the Offender against the Injured_Party. This Frame Element need not always be realized, although it is conceptually necessary.

Lexical Units

avenge.v, avenger.n, get_back.v, get_even.v, payback.n, retaliate.v, retaliation.n, retribution.n, retributive.a, retributory.a, revenge.n, revenge.v, revengeful.a, revenger.n, sanction.n, vengeance.n, vengeful.a, vindictive.a

III.6 Irudia: *Revenge frame*a.

deiturikoak—, eta hauek ale lexikal desberdinez egongo dira adierazita. 4. adibidean ikus daitekeen bezala, *Avenger* FEa Hook ale lexikalak adierazten

du, *Offender* FEa Peter Pan ale lexikalak, eta abar.

(4) [Hook *Avenger*] tries to avenge [himself *Injured party*] [on Peter Pan *Offender*] [by becoming a better father *Punishment*].

Bestalde, *frame* bakoitzak bere FEak zehaztuta izango ditu. III.6 irudian *Revenge frame*aren alderdi bakoitza definituta dator. Esate baterako, *Avenger* FEaren definizioa hurrengoa da: *The Avenger exacts revenge from the Offender for the Injury.* Honebestez, *frame*ak dira:

> "[...] schematic representations of situations involving various participants, props, and other conceptual roles, each of which is a frame element (FE)." (Johnson eta Fillmore, 2000, 56. or.)

Frame bakoitzarekin batera, *frame* hori onartzen duten ale lexikalen zerrenda ematen da. *Revenge frame*aren kasuan, hauexek: avenge, avenger, get back, get even, retaliate, retaliation, retribution, retributive, retributory, revenge, revenger, sanction, vengeance, revengeful, vengeful eta vindictive. Hala, *frame*etan oinarritzeak orokortzeko aukera ematen du, hau da, *frame* bera osatzen duten ale lexikalek klase semantiko bat osatzen dute, eta hori dela eta, *frame*a definitzen duten ezaugarri kontzeptualak klase semantiko osoari egokitzen zaizkio, baita ezaugarri sintaktiko-semantikoak ere. Klase semantikoa, beraz, beti dator zehaztua berau onartzen duten ale lexikalen zerrendarekin.

Hau esanda, FrameNet proiektuan egiten dutena hurrengoa da: ale lexikal bakoitza bere adieraren arabera sailkatu honi dagokion *frame*an. Hala, *frame*en funtsa adieran dago: ale lexikal beraren adieretako bakoitza *frame* ezberdin batean egongo da.

> "It is not that every word has its own frame, but every sense of every word has its own frame." (`http://www.icsi.berkeley.edu/framenet/book.html`)

Frame bakoitzari dagokion informazio guztia zehazteko (*frame*aren alderdiak, *frame*ko ale lexikalen zerrenda, *frame*aren informazio sintaktiko-semantikoa...), *etiketatze semantikoa* baliatzen dute. Esaldi bakoitzaren etiketatzea *target*en (esaldiko ale lexikal baten) ikuspuntutik eginda dago. Hau da, esaldiko ale lexikal baten *frame*a oinarri hartuta[22], esaldiko beste elementuak *frame* horren alderdiei lotuko zatzaizkie. Esaterako, (4) esaldiaren

[22] Ale lexikal hauek aditzak, objektuak edo adjektiboak izango dira, hots, gobernatzaileak izan daitezkeen ale lexikalak.

etiketatzean, avenge aditza izan da etiketatzeko abiapuntua (*target*a). Beraz, esaldiko beste ale lexikalak avengeri dagokion *frame*aren alderdiekin etiketatu dira.

Alderdi semantikoarekin batera, osagaien funtzio eta kategoria sintagmatikoak ere etiketatzen dira, eta *target*aren ikuspuntutik egingo denez, esaldiko ale guztiek berarekin duten lotura sintaktikoa adieraziko dute.

Ondorioz, esaldien etiketatze semantikoaren emaitza izango da esaldiko ale lexikal bakoitza etiketatua egotea FE batekin, funtzio sintaktiko batekin eta kategoria sintaktiko batekin. Hala, bada, esaldiko ale guztiek *target*arekiko duten lotura sintaktiko-semantikoa adieraziko dute.

Honezaz gain, corpus erreal bat etiketatzetik lortzen dituzten datuak erabiltzen dituzte, *frame* bakoitzaren egitura sintaktikoak proposatzeko. Esaterako, corpuseko agerpenetan oinarrituaz *Revenge frame*an dagoen avenge aditzaren azpikategorizazioa III.1 irudikoa litzateke. Hau da, avenge aditzarekin batera, corpusean agertu diren osagaien zerrenda dugu, hauen FEa, kategoria eta funtzioa, maiztasunarekin batera, zehazten direlarik.

Informazio sintaktiko-semantikoaren adierazpenaz gain, FrameNeten *frame*en arteko harreman semantikoak ere adierazten dira, hau da, *frame* guztiekin hierarkia bat osatzen dute, eta hierarkia horretan *frame* konplexuagoek zehatzagoak direnak barnean hartzen dituzte. Esate baterako, avenge aditza *Revenge frame*ari dagokio, eta *frame* hau *Reward and Punishments frame*aren *subframe* bat da. Eta azken hau, aldi berean, *Intentionally affect frame*aren azpian kokatzen da hierarkian.

Hortaz, formalismo hau, nahiz eta teoria bati lotua egon, corpus errealeko datuetan oinarritzen da; beraz, inplementa daitekeen EBLa da. EBLa sortu eta lantzearekin batera, corpus etiketatu bat eratzen ari dira eta horrek hainbat erabilerari bidea zabaltzen die (baita konputazionalei ere). Horren adierazgarri da, FrameNet batzuk ari direla garatzen hainbat hizkuntzatan: alemana (Boas, 2002), gaztelaniakoa (Subirats-Rüggeberg eta Petruck, 2003) eta japoniarra (Ohara *et al.*, 2003), hain zuzen ere.

Hala ere, esan beharra dago, FrameNeten corpusaren erabilera mugatua egiten dutela: aldez aurretik aukeratutako corpusaren lagin bat erabiltzen dute, sortutako *frame*ak zuzenak diren ala ez egiaztatzeko, eta hauei adibideak lotzeko:

Number Annotated	**Patterns**				
2 total	***Avenger***	***Injured Party***	***Injury***	***Offender***	***Punishment***
1	NP Ext	NP Obj	PP[for] Comp	– –	PPing [by] Comp
1	NP Ext	NP Obj	PP[of] Comp	– –	PPing [by] Comp
11 total	***Avenger***	***Injured Party***	***Injury***	***Offender***	***Punishment***
2	– –	NP Ext	– –	– –	
1	– –	NP Ext	PP[on] Comp	– –	
6	NP Ext	NP Obj	– –	– –	
1	NP Ext	NP Obj	– –	PPing[by] Comp	
1	NP Ext	NP Obj	PP[on] Comp	PPing [by] Comp	
19 total	***Avenger***	***Injured***	***Offender***	***Punishment***	
3	– –	NP Ext	– –	– –	
1	– –	NP Ext	– –	PP[by] Comp	
10	NP Ext	NP Obj	– –	– –	
2	NP Ext	NP Obj	– –	PP[with] Comp	
2	NP Ext	NP Obj	– –	PPing[by] Comp	
1	Poss Ext	– –	PP[against] Comp	– –	

III.1 Taula: avenge aditzaren egitura sintaktikoak corpuseko agerpenetan oinarrituta.

> "Because FrameNet is primarily lexicographic, we are not attempting to annotate whole texts or even a random sample of sentences which include each lemma. Rather, we want to annotate a set of sentences which exemplify the range of combinatorial possibilities of a lexical unit, including all the types of syntactic constituents which can embody the frame elements."
> (Ruppenhofer *et al.*, 2002, 371. or.)

Beraz, beraien helburua ez da corpus oso bat *frame*ekin etiketatzea. Aldiz, LNPren ikuspegitik interesgarriagoa litzateke corpusa bere osotasunean erabiliko balute, honek aplikazio berrietarako aukera handigoak emango lituzkeelako.

Aztertzen ari garen EBL hau oso interesgarria da batez ere ikuspegi konputazionaletik, LNPren arlo ezberdinen azterketarako oso baliagarria delako[23]. Baina epe luzerako EBLa da; hau da, eremu batzuetara (komunikazioa, legedia, hezkuntza...) mugatutako lexikoa da, denborarekin hizkuntza bere osotasunean adierazteko helburua duena. Gure euskararako EBLa, ordea, ezin da eremu zehatz horietara mugatu. Aitzitik, hizkuntza bere osotasunean adierazteko gai izan behar du.

Kopuruez mintzatuz gero, FrameNetek gutxi gorabehera, 450 *frame*, 6.000 ale lexikal eta 130.000 esaldi etiketatu ditu eta handitzen jarraitzen du. FrameNet EBL publikoa da[24].

III.2.3.2 WordNet eta WordNetetik abiatutakoak

WordNet (Miller, 1985; Fellbaum, 1998a) teoria psikolinguistikoetan oinarritua dagoen ingeleseko ezagutza-base lexikala da.

WordNetek ingeleseko izen, aditz, adjektibo eta adberbioei buruzko informazioa dauka, eta informazio hau *sinonimo-multzo* (***synonym set*** edo ***synset*** deiturikoa) ideiaren arabera antolatuta dago. *Synset* bakoitza kontzeptu lexikal bati dagokio, eta hau osatuko duten hitz-multzoek kategoria berdinekoak eta testuinguru bereetan truka daitezkeenak dira.

Esaterako, {car, auto, automobile} hitz-multzoak[25] *synset* bat osatzen dute, kontzeptu bera adierazten dutelako. *Synset*aren adiera, normalean, glosa baten bidez adierazten da: a motor vehicle with four wheels.

[23] FrameNeten erabilera konputazionalari buruzko argibide gehiagorako, jo bedi Pociello ren lanera (2004b).

[24] `http://www.icsi.berkeley.edu/framenet` (2007-07-02an atzitua).

[25] Adierazpen guztiak WordNet 3.0 bertsiotik hartu ditugu — `http://www.wordnet.princeton.edu` (2007-07-02an atzitua)—, eta gehienetan, leku arazoak direla-eta, adibidearen informazioa laburtu egin dugu.

(5) {car, auto, automobile} (a motor vehicle with four wheels)

Ildo honetatik, WordNeteko 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. Honezaz gain, sinonimia ez den beste erlazio semantikoei esker, *synset*en arteko harremanak daude. Erlazio semantiko garrantzitsuena **hiperonimia-hiponimia** erlazioa da.

Hiperonimia-hiponimia erlazioak *synset* orokorrenak *synset* zehatzagoekin lotzen ditu[26]. (6) eta (7) adibideetan (5)en hiperonimoak eta hiponimoak ikus ditzakegu, hurrenez hurren:

(6) {car, automobile} (a motor vehicle with four wheels)
=> {self-propelled vehicle} (a wheeled vehicle that carries...)
=> {wheeled vehicle} (a vehicle that moves on wheels...)
=> {vehicle} (a conveyance that transports people or...)
=> {conveyance, transport} (something that serves...)
=> {instrumentation} (an artifact that is...)
=> {artifact} (a man-made object taken as a...)
=> {...}

(7) {car, automobile} (a motor vehicle with four wheels)
=> {ambulance} (a vehicle that takes people to and from hospitals)
=> {cab, taxi, hack, taxicab} (a car driven by a person whose...)
=> {limousine, limo} (large luxurious car)
=> {jeep, landrover} (a car suitable for traveling over rough...)
=> {sedan} (a closed car that has front and rear seats...)
=> {...}

(6) adibidean car izenaren hiperonimoak ditugu. *Synset* hau self-propelled vehicle bezala definitzen da; self-propelled vehicle, wheeled vehicle mota bat bezala; wheeled vehicle, aldi berean, vehicle mota bat bezala, eta abar.

Hiponimoak hiperonimoen zehaztapenak dira. Hortaz, (7) adibidean, car izenaren zehaztapen gisa auto motak agertzen dira (ambulance, taxi...). 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 erlazio semantiko nagusi baten arabera antolatzen da. Izen eta aditzen kasuan erlazio semantiko

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

nagusia hiperonimia-hiponimia da[27]. Adjektibo eta adberbioek, berriz, sinonimia-antonimia dute ardatz beraien antolakuntzan.

WordNeteko sailkapena, beraz, *synset*etan eta beraien erlazio semantikoetan datza. Erlazio semantiko hauen bidez, *synset*ak hierarkikoki multzokatzen dira, edo, beste era batera esanda, klase semantikoak osatzen dira. Autoen klase semantikoa, adibidez, {car, auto, automobile} *synset*aren azpian egongo da jasota.

WordNeten ildotik jarraituta, beste EBL batzuk garatu dira: *EuroWordNet* (Vossen, 1998) eta *The Multilingual Central Repository* (MCR) (Rigau *et al.*, 2003). Oinarri bera erabili arren, bakoitzak aurreko EBLa aberastu du.

EuroWordNet

EuroWordNet (Vossen, 1998) ezagutza-base eleanitza da, Europako zortzi hizkuntzatara zabaltzen dena (ingelesa, nederlandera, italiera, gaztelania, alemana, frantsesa, txekiera eta estoniera), eta WordNeten eredua jarraitzen duena.

Proiektu honetan parte hartu duen hizkuntza bakoitzak wordnet *independente* bat du, eta EuroWordNeten helburua wordnet desberdin hauek guztiak ezagutza-base eleanitz bakarrean elkartzea da. Beste hitz batzuetan esanda, *synset* bera ingelesez, nederlanderaz, italieraz, gaztelaniaz, alemanez, frantsesez, txekieraz eta estonieraz ikusteko aukera ematen du.

The Multilingual Central Repository

The Multilingual Central Repository (MCR) interfaze eleanitza da, non Europa Batzordeko *MEANING: Developing Multilingual Web-Scale Language Technologies* (IST-2001-34460) proiektuan euskararako, katalanerako, ingeleserako, italierarako eta gaztelaniarako (Rigau *et al.*, 2003) aztertu den informazio guztia integratzen den. Ezagutza-base honek EuroWordNeten eredua jarraitzen du. Horregatik, honetan ere, hizkuntza bateko *synset* batekin beste hizkuntzetakoa ere ikusgarri dago.

MCR EuroWordNeten bertsio aurreratuagoa da, hau da, MCR eta EuroWordNet oinarrian gauza bera dira, baina MCR EuroWordNet *aberatsago* bat da. Honenbestez, MCR WordNet eta EuroWordNeten informazioaz

[27] Aditzen kasuan, eta gero IV.1.2 atalean ikusiko dugun bezala, *hiperonimia-troponimia* erlazioaz hitz egiten da.

baliatzen da, eta honetaz gain, informazio berria dakar: hautapen-murriztapenak, *The Suggested Upper Merged Ontology* (SUMO) delakotik hainbat informazio, eta abar.

Hurrengo kapituluan, WordNet, EuroWordNet eta MCRren azalpen sakonagoa emango dugu.

Oro har, hiru EBL hauek hizkuntza bere osotasunean adierazi nahi duten EBL publikoak dira[28]. Esate baterako, WordNetek 117.617 *synset* ditu (81.426 izen, 13.650 aditz, 18.877 adjektibo eta 3.664 adberbio). Eta baldintza honi esker, eta EuroWordNet eta MCRk eskaintzen duten eleaniztasuna kontuan hartuta, hiru EBL hauek oso erabiliak izan dira LNPren arlo oso ezberdinetan: galdera-erantzun sistemetan, informazio-erauzketan, itzulpen automatikoan, eta abar (argibide gehiago IV. kapituluan).

Dena den, WordNeti egin zaion gaitzespen garrantzitsuenetako bat informazio sintaktiko-semantiko urria duela izan da.

> "Many users of WordNet have lamented the lack of syntactic information that would match the detail of the semantic treatment in WordNet. Indeed, WordNet contains very little syntax, because it was conceived as a semantic database only.[...] Applications in knowledge engineering and inferencing especially would benefit from information linking verbs and nouns."
>
> (Fellbaum, 1998a, 11. or.)

Behar hau ikusita WordNeten informazio sintaktiko-semantikoarekin aberasteko saiakerak egon dira, adibidez, aditzen alternantziak gehitu dira (Kohl *et al.*, 1998). MCRko interfazeak berak (hurrengo kapituluan ikusiko dugun bezala), informazio sintaktiko-semantikoaren beharraz jabetuta, informazio hau txertatzeko baliabideak eskaintzen ditu.

---

[28] WordNet: `http://www.wordnet.princeton.edu` (2007-07-02an atzitua).
EuroWordNet: `http://ixa2.si.ehu.es/mcr/wei.html` (2007-07-02an atzitua).
MCR: `http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl` (2007-07-02an atzitua).

III.2.3.3 Volem

Volem proiektuaren (Fernández *et al.*, 2002) helburua zera da: Pirinio inguruko hizkuntza batzuetako (gaztelania, katalana eta frantsesa) aditz eta preposizioen ezaugarriekin EBL bat eraikitzea hurrengo informazioarekin:

- Gaztelania, katalana eta frantseseko aditz eta preposizio bakoitzaren izaera sintaktikoaren deskribapena: azpikategorizazioa, hautapen-murriztapenak eta aditzen alternantziak.
- Gaztelania, katalana eta frantseseko aditz eta preposizio bakoitzaren informazio semantikoa: *Egitura Lexikal-Kontzeptuala* (ELKa), rol tematikoak eta aditzen kasuan, WordNeteko klase semantiko nagusiena.

Fernández *et al.*-en lanetik (2002) hartutako adibidean ikus daitekeen bezala[29], informazio hau guztia adierazteko eta antolatzeko Jackendoffen (1990) ELKak erabiltzen dituzte.

(8) **Common part to the three languages:**

thematic grid: *[inic(ag, tc), th]*
(e.g. agent or causal iniciator, theme)

LCS: Literally: I (subject) caused an object J to undergo a change of state in its ontological universe, BECOMING (achievement) STATE.

[event CAUSE([thing I],
[event BECOME+ char,+ident([thing J],
[state STATE])]

Spanish lexical database:
Spanish verb: cerrar (to close)
Sense number: 75

Alternations + examples:
caus-2np:
El viento cerró las ventanas de golpe
(the wind closed the windows)
...

[29] (8) adibideak ez du sarrera lexikal guztia jasotzen. Sarrera lexikal osoa, Pocielloren lanean (2004b) edota hurrengo webgunean dago: `http://www.irit.fr/recherches/ILPL/Site-Equipe/demonstrations.html` (2007-07-02an atzitua).

Catalan lexical database: Catalan verb: tancar (to close)
Sense number: 75

Alternations + examples:
caus-2np:
El vent va tancar les finestres de cop
(the wind closed the windows)
...

French lexical database:
French verb: fermer (to close)
Sense number: 75

Alternations + examples:
caus-2np:
Le vent ferme les feneêtres d'un coup
(the wind closed the windows)
...

Lehenik, adiera bereko ale lexikoek (kasu honetan, cerrar, tancar eta fermer) hizkuntza guztietan duten antzekotasun semantikoa (*Common part to the three languages*) rol tematiko eta ELK baten bidez definitzen dute, eta, gero, hizkuntza bakoitzean sarrera lexikal horrek (zehaztutako adiera horrekin) izan ditzakeen alternantziak zerrendatzen dituzte. Ezagutza-base eleanitza denez, azalpenak eta argibideak ingelesez ematen dituzte.

Hala, Volem proiektuan Levinen (1993) hipotesia jarraitzen dute: adiera bakoitzeko ezaugarri sintaktikoak aldatzen dira. Aditzen adierak WordNeteko klase semantiko nagusienen arabera antolatzen dituzte (*verbs of possession, verbs of movement, verbs of consumtion...*). Oso klase orokorrak direnez, eta hori denez adierari buruz zehazten den informazio semantiko bakarra, aditzen semantika ez da beti argi ikusten. Volemen, beraz, Jackendoff (1990), Levin (1993) eta WordNeteko informazioa txertatzen da. Hala ere, aditz eta preposizioetara mugatzen da, eta, hori dela eta, ez du hizkuntza bere osotasunean adierazten.

Gerora, proiektu honen bigarren zatiari ekin zaio (*Volem2*), zeinetan Volemeko aditz eta preposizioei euskara eta okzitanieraren informazioa gehitzen zaion.

Ezagutza-base hau LNPrako interesgarria izan daitekeen arren, egun oraindik eraikitze bidean dagoenez, honekin ez dira aplikazio ugari ezagutzen. Dena den, hasiberria den proiektu bat aipatu dezakegu: *SenSem* (*Sentence Semantics*) (Alonso *et al.*, 2005) proiektua MCyT (BFF2003-06456). Proiektu honetan corpus etiketatu bat eraikitzen ari dira erdiautomatikoki eta bere

helburu nagusiena, Volemeko informazioa orraztu eta aberastea da. Horretarako, corpusean predikatuekin agertzen diren alternantziak Volemen zerrendatuak dituzten predikatuekin erkatzen dituzte, ezagutza-basean dituztenak zuzenak diren ala ez egiaztatzeko eta ez dituztenak gehitzeko.

III.2.4 PropBank

PropBank proiektuan (Palmer eta Kingsbury, 2003) *Penn Wall Street Journal Treebank II* corpusa —300.000 tokeneko corpusa— etiketatu dituzte predikatu-argumentu erlazioekin. Horrelako, aditzen adierak eta adiera horien dependentziak (argumentuak) markatzen dituzte.

PropBank ereduan bi maila bereizten dituzte: batetik, argumentu eta adjuntuen maila, eta bestetik, rol semantikoen maila. Argumentu gisa etiketatzen diren ale lexikalak *Arg0*tik *Arg5*era zenbakitzen dira. Etiketa hauek ez daude funtzio gramatikal bati lotuak. Aditz desberdin edota aditz beraren adiera desberdin bakoitzean etiketa hauek informazio desberdina adieraz dezakete. Adibidez, Johnek leihoa hautsi zuen eta Leihoa hautsi zen esaldietan, leihoa hitzak argumentu-etiketa bera izango du, bi esaldiak aditz-adiera beraren alternantziak direlako.

Dena den, oro har, zenbaki baxuenak dituzten argumentuen artean erregulartasun bat ageri da. Esaterako, aditz iragankorren subjektuek *Arg0* marka izaten dute eta objektu zuzenek *Arg1*.

Rol semantikoen mailan, PropBankek bi rol mota erabiltzen ditu: aditz bakoitzari dagozkion rol zehatzak —ingeleseko buy aditzaren rolak *buyer* eta *thing bought* bezalakoak izango dira—, eta rol orokorrak —*agent* eta *theme* bezalakoak. Azken hauek *VerbNet* (Kipper *et al.*, 2000) lexikoiari lotuta daude. III.2 taulan PropBankeko argumentu markekin agertzen diren rol eta funtzio sintaktikoak ikus daitezke.

VerbNet aditzen lexikoi zabala da, non aditzak Levinen (1993) sailkapenaren arabera antolatuta dauden. Aditzak hierarkikoki antolatzen dira eta aditz bakoitzean informazio sintaktikoa eta semantikoa egoteaz gain, aditz horrek WordNeten duen adiera ere adierazten da. Hortaz, esan daiteke, VerbNet eta WordNet osagarriak direla.

Corpus horrekin batera, lexikoia garatzen ari dira, non etiketatutako aditz bakoitzaren adiera eta argumentuak zerrendatzen diren. Sarrera bakoitza aditz-adiera bat da, ***roleset*** deritzaiona, eta bertan aditzaren alternantziak, —***frame*** deiturikoak— honek hartzen dituen argumentuekin zehazten dira. III.7 taulan tell.01 *roleset*a dugu; aditz-adiera honek lau alternantzia ditu

Arguments	***VerbNet roles***	***Syntactic function***
Arg0	agent, experiencer	subject
Arg1	patient, theme, attribute, extension	direct object, attribute, predicative, passive subject
Arg2	attribute, benficiary, instrument, extension, final state	attribute, predicative, indirect object, adverbial complement
Arg3	benficiary, instrument, attribute, cause	predicative, circumstantial complement
Arg4	destination	adverbial complement
Adjuncts	***VerbNet roles***	***Syntactic function***
ArgM	location, extension, destination, cause, time, manner, direction	adverbial complement

III.2 Taula: PropBankeko argumentu markekin agertzen diren funtzio sintaktikoak eta VerbNeteko rolak.

(*ditransitive, odd ditransitive, prepositional arg2* eta *fronted*). Nahiz eta informazio osoa lehenengo *frame*ari informazio osoa bakarrik jarri, sarrera bakoitzeko *frame* guztiek izango dute argumentuen informazioa.

PropBank proiektuko emaitzak publikoak dira[30], eta LNPn asko erabiltzen ari den EBLa da, batez ere rolen etiketatze automatikoaren oinarri gisa (Pradhan *et al.*, 2003; Carreras eta Màrquez, 2004). Erabilera hau dela eta, egun, beste hizkuntza batzuentzat ere garatzen ari da eredu hau: txinerarako (Palmer eta Xue, 2003), gaztelania eta katalanerako (Civit *et al.*, 2005a), errusierarako (Civit *et al.*, 2005b), eta euskararako (Agirre *et al.*, 2006d).

Hala ete guztiz ere, eredu emankorra izan arren, aditzen deskribapena soilik egiten duen eredua da, eta, ondorioz, ez du euskararako EBLrako zehaztu dugun baldintzetako bat betetzen, hots, ez du hizkuntza bere osotasunean adierazten.

[30] `http://www.cis.upnn.edu/ace` (2007-07-02an atzitua).

Roleset tell.01 “pass along information”:
Roles:
Arg0: *Speaker*
Arg1: *Utterance*
Arg2: *Hearer*
Frames:
distransitive (-)
The score tell you what the characters are thinking and feeling
Arg0: The score
REL: tell
Arg2: you
Arg1: what the are thinking and feeling
odd ditransitive (-)
prepositional arg2 (-)
fronted (-)

III.7 Irudia: tell.01 sarrera lexikala PropBanken.

III.2.5 Corpusetan oinarritutako lanak

Kapitulu honetan zehar, EBLak eraikitzeko hainbat proposamen azaldu ditugu, hizkuntzalaritza teorikoa eta konputazionalaren ikuspegiak kontuan hartuz. EBLak garatzean, normalean, corpusak ere erabiltzen direla ikusi dugu. Atal honetan, aipatutako corpusak bere osotasunean komentatuko ditugu.

Dagoeneko aipatu dugu II.2.1 atalean, LNPn corpusek hartu duten garrantziaz. Alde batetik, erabilerari buruzko informazioa, hitzak dituzten maiztasun errealak, egitura sintaktiko zenbaitek dituzten maiztasunak, eta halako informazioa lortzeko oso erabilgarriak dira. Bestetik, informazio linguistikoa baldin badute —esate baterako, corpusak lematizatuta badaude, kategoriak markatuta badituzte, semantikoki markatuta badaude, eta abar— hauetatik informazio linguistikoa erauzi eta aberasteko erabil daitezke. Eta, azkenik, corpusen bidez, hipotesien zuzentasuna froga daiteke; hau da, eredu baten zuzentasuna egiaztatzeko era bakarra, eredu hori corpus errealean frogatzea da.

Horren adierazgarri ditugu aurreko ataletan aipatutako ia eredu guztiekin garatzen ari diren corpusak. Adibidez, LFG formalismoko egitura funtzionalak erabilita corpus etiketatuak daude, esate baterako Cahill *et al.* (2002). HPSG formalismoak corpus etiketatuak ere baditu, ingeleserako (Oepen *et al.*, 2002, edo *LinGO Redwoods deiturikoa*) eta baita beste hizkuntza batzuetarako ere, hala nola bulgarierarako (Osenova eta Simov, 2003).

EBL eta corpusen arteko harremanaren adibide garbia FrameNet proiektuan ikus daiteke. III.2.3.1 atalean azaldu dugun bezala, FrameNet proiektuan (Fillmore eta Baker, 2001) ingeleserako baliabide lexikografikoak eraikitzen ari dira. *Frame Semantics* (Fillmore, 1985) teorian oinarrituta eta corpus errealeko datuekin lagunduta. FrameNeten ale lexikal bakoitza beraiek sortutako *frame*etan sailkatzen dute (*Revenge frame*a, *Commercial Transaction frame*a, *Criminal Process frame*a, *Perception frame*a, eta abar,) batetik ale honen semantika eta sintaxia definitzeko, eta bestetik, *frame*ko gainontzeko osagaiekin duen harremana zehazteko. *Frame*a, *frame*aren parte-hartzaileak (*frame elements* deiturikoak), eta *frame*a osatzen duten ale lexikalak sortu ondoren, corpus errealera jotzen dute *frame*aren zuzentasuna egiaztatzeko, hau da, etiketatze semantikoa baliatzen dute, introspekzioz sortutako *frame* horiek egokiak diren ala ez ziurtatzeko. Corpuseko datuak eta *frame*a bat etorriko ez balira, *frame*aren ezaugarriak corpusaren informazio berri horretara egokituko lirateke. FrameNeteko corpusak gutxi gorabehera, 130.000

esaldi etiketatu ditu eta handitzen jarraitzen du.

WordNetek ere badu etiketatuko corpus bat: *SemCor* (Miller *et al.*, 1994; Fellbaum *et al.*, 2001). Hala ere, FrameNeten ez bezala, WordNet eta SemCor ez dira aldi berean garatu. Lehenengo WordNet sortu zen eta gero, 250.000 hitzetako *Brown* corpusaren testu zati bat hartu, eta Princetoneko kategoria-etiketatzaile automatikoarekin etiketatu ondoren, eskuz etiketatu zen WordNeteko adierekin (Miller *et al.*, 1994).

Volem proiektuaren jarraipen gisa *SenSem* (*Sentence Semantics*) proiektua garatzen ari dira. Proiektu honetan gaztelaniako corpus etiketatu bat eraikitzen ari dira erdiautomatikoki eta bere helburu nagusiena, Volem EBLko gaztelaniako informazioa orraztea eta aberastea da. Horretarako, corpusean predikatuekin agertzen diren alternantziak Volemen zerrendatuak dituzten predikatuekin erkatzen dituzte, ezagutza-basean dituztenak zuzenak diren ala ez egiaztatzeko eta ez dituztenak gehitzeko. Volemetik abiatutako gaztelaniako EBL berritu honi SenSem deitu diote. SenSem EBLan 788 aditzen 1.092 adiera daude, eta beraien izaera sintaktiko-semantikoa adierazita dago. Bestalde, aditzen adierak WordNeteko *synset*ekin lotzen ari dira[31].

Aipatutako *PropBank* proiektua (Palmer eta Kingsbury, 2003) ere horixe bera da: *Penn Wall Street Journal Treebank II* corpusa etiketatzea predikatu-argumentu egiturekin. Horretarako, aditzen adierak eta adiera horien dependentziak (argumentuak) markatzen dituzte. Corpus horrekin batera, lexikoia garatzen dute, non etiketatutako aditz bakoitzaren adiera eta argumentuak zerrendatzen diren. Inplementazioari begira, PropBank corpusari VerbNeteko informazioa gehitu zaio (Kipper *et al.*, 2002)[32].

III.3 Gure aukera eta arrazoiak

III.1 atalean zehaztu ditugu euskararako garatu nahi dugun EBLak izan beharko lituzkeen baldintzak. Ikusi dugun bezala, zaila da baldintza hauek guztiak jasotzen dituen EBLa topatzea. Hala ere, baldintza horietan oinarrituta, hain zuzen ere, arrazoituko dugu IXA taldearen beharretara gehiago egokitzen den EBL formalismoak WordNet, eta honen ildotik abiatuta garatu diren EuroWordNet eta MCR direla.

[31] SenSem kontsultagarri dago hurrengo web orrian: `http://gril.uab.es/demo` (2007-07-02an atzitua).

[32] PropBank hurrengo web orrian dago ikusgarri (2007-07-02an atzitua): `http://www.rochester.edu/gildea/PropBank/Sort/C.html`.

- **Eredu irekia eta deskriptiboa:**

WordNet ez dago teoria bakar bati lotua; hots, teoria ezberdinek erabil dezaketen EBLa da. Bestalde, EuroWordNet eta MCR WordNeten garapenak dira, WordNet beste oinarri eta ikuspuntu teoriko eta konputazionaletatik informazio gehiagorekin aberastu dutenak.

Aurreko atalean aipatutako formalismo eta lan teoriko askok ere gerora WordNet eta EuroWordNet adierekin edo/eta klase semantikoekin aberastu dituzte[33]; esate baterako, Dorrek (1997) Jackendoffen lanarekin. Dorrek Jackendoffen ELKetan oinarritutako EBLa eraiki du. ELK hauek WordNeteko adieretara lotuak daude. Lan horretan bertan, Dorrek Levinen aditz-klaseetako aditzak ere WordNeteko aditzekin lotzen ditu. Ildo honetatik jarraitu duen formalismoa Volem izan da: gaztelaniako, frantseseko eta katalaneko aditzen informazio sintaktiko-semantikoari (azpikategorizazioa, hautapen-murriztapenak eta alternantziak), ELKa, rol tematikoak eta WordNeteko klase semantiko nagusienak eransten dizkiote. Bestalde, Pustejovskyren lexikoaren ezaugarri batzuk WordNetekoekin lotzeko saiakera ere egin da (Buitelaar, 1998). Formalismo ezberdin hauen arteko uztardura oso baliagarria eta aberatsa da. Izan ere, WordNeten ildotik euskararako egingo den EBLa hauetaz guztiez balia daiteke (neurri handi batean behintzat), eta horrela euskararako EBLa aberastu. Beraz, garbi dago WordNet eta EuroWordNet LNPren arloan baliabide oso erabiliak izan direla, eta egun oraindik hainbat esperimentu eta ikerlanetarako iturburu direla.

- **Hizkuntzaren ikuspuntu orokorra:**

WordNet (EuroWordNet eta MCR) lexiko zabal eta garatua da. Era berean, adieran oinarritutako ontologia da, hizkuntzaren lexikoa ezagutza-base batean jaso nahi duena, ale lexikalak, ale lexikalen adierak, klase semantikoak, kategoriak, eta hauen guztien arteko erlazio semantikoak kontuan izanda (III.2.3.2 atalean azaldu dugun bezala). Noski, hizkuntzaren lexikoak ez du mugarik. Horregatik, etengabe garatzen dauden ezagutza-baseak dira lexikoi hauek. Hala ere, hizkuntzaren ikuspuntu orokorra eman dezaketen ezagutza-baseak ditugu. Esate baterako, WordNetek 117.617 *synset* ditu (81.426 izen, 13.650 aditz, 18.877 adjektibo eta 3.664 adberbio)[34]. MCRk WordNet ezagutza-basearen tamaina berdina du, baina erlazio semantiko gehiagorekin (1.600.000 erlazio inguru).

[33]MCR orain dela gutxiko EBLa izanda, oraindik ez da horrela erabili.

[34]WordNeten azkeneko bertsioaz ari gara, 3.0 bertsioaz, alegia.

- **Inplementazioa:**

WordNet, EuroWordNet eta MCR inplementatutako EBLak dira, hots, praktikoak direla asko frogatua dago. Gainera, ezagutza-base publikoak dira, kontsultagarriak, alegia, eta hainbat erabilera izan ditzakete (hiztegi eta thesaurus gisa adibidez).

EuroWordNeten eta MCRren aukerak areago doaz, EBL hauek **eleanitzak** direlako, ingeleseko WordNeti beste hainbat hizkuntza gehitu baitzaizkio (nederlandera, italiera, gaztelania, alemana, frantsesa, txekiera, estoniera...), eta horien artean —tesi honetan arrazoitutakoari jarraiki— euskara txertatzen hasi garelako (Agirre *et al.*, 2002).

Hiru EBL hauek oso erabiliak izan dira LNPren arlo oso ezberdinetan: galdera-erantzun sistemetan, informazio-erauzketan, itzulpen automatikoan... (argibide gehiago IV.1 ataletan). Honen adierazgarri da WordNeten oinarrituta egin diren publikazioen kopurua. WordNeteko web orriak[35] batzuk jasotzen ditu, eta 422 inguru dira gaur egun.

Azpimarratu beharra dago WordNetek paper garrantzitsua jokatu duela adiera-desanbiguazioan. Adiera-desanbiguazioko sistemak estaldura handiko baliabide lexikaletan (lexikoietan, corpusetan, ontologietan, etab.) oinarritu behar dira, baliabide hauei esker sistema bera garatu eta ebalua daitekeelako. Geroz eta estaldura handiagoko baliabideak izan, orduan eta emaitza hobeak lortuko dira. WordNet estaldura handiko EBLa izateaz gain (gorago aipatu ditugu EBL honen kopuruak) bere *synset*ak baliatuta, eskuz etiketatuta 250.000 hitzeko corpusa dago: *SemCor* (Miller *et al.*, 1994). WordNetek SemCorren duen estaldura %96 da. SemCorrek testuinguru egokia eskaintzen du adiera-desanbiguazioko sistemak bertatik *ikasteko*[36] eta gero ebaluatzeko. Hala, semantikoki etiketatutako corpusen arrakasta eta erabilgarritasuna ikusita, beste hizkuntzetako wordnetak ere beraien corpus etiketatuak garatzen ari dira. Honen adibide da *MultiSemcor* (Bentivolgi eta Pianta, 2005) proiektua, non ingeleseko SemCor italierara itzultzen ari diren eta ingeleseko corpuseko hitzen etiketa semantikoak zuzenean italierako hitzei esleitzen dizkieten. Honen emaitza semantikoki etiketatutako italierako corpusa izango da.

[35] `http://www.wordnet.princeton.edu` (2007-07-02an atzitua).

[36] Makinari emandako datu egokietan oinarrituz eta hauen gainean teknika estatistiko konplexuak aplikatuz, makinak *ikasi* egiten du; ikasketa honen ondorioz, gai da datu berriei buruz erabakiak hartzeko. Erabaki hauen zuzentasuna ikaste-prozesuaren egokitasunaren araberakoa izango da, noski; ikaste-prozesuaren egokitasuna, era berean, erabiltzen diren teknika estatistikoen eta ikasteko erabilitako datuen kopuruan eta egokitasunean datza.

Honekin batera, tesi-lan honetan aurrerago (V. kapituluan) aurkeztuko dugun *EuSemcor* proiektua ere aipa dezakegu: semantikoki etiketatzen ari den euskarako corpusa.

Beste arrazoi batzuk ere baditugu eredu hauen alde egiteko:

- **EuroWordNetek datuak eguneratzeko eskaintzen dituen erraztasunak:** ILIaren bidez lortzen den hizkuntzen arteko lotura horri esker (argibide gehiago IV.2 ataldean), EuroWordNeti lotuta dauden beste hizkuntzetako wordnetetako batean aldaketaren bat egiten bada *synset*en batean, aldaketa hori euskarako wordnetean ere gertatzen da.

- **WordNet ereduak EBLaren eta corpusaren garapena aldi berean egiteko aukera ematea:** Hots, ez da EBLa amaitua izatera itxaron behar honen informazioarekin corpus bat etiketatzeko.

Atal honetan azaldu ditugun abantailak direla eta, wordnet berrien kopurua handitzen ari da (katalana, portugesa, grekoa, suediarra, errumaniarra, bulgariarra, norvegiarra, lituaniarra, errusiarra...). Hala, geroz eta gehiago dira eredu hau jarraituta EBLak garatzen dituztenak.

Informazio sintaktiko-semantikoa, batez ere aditzetan, mugatua duela, horixe da WordNeti egin zaion gaitzespen nagusiena. Adibidez, ez dituzte azpikategorizazioa, hautapen-murriztapenak eta rol tematikoak zehazten. Hau oztopo bat da euskararako EBL bat hauetan oinarrituta egiteko, lanaren hasieratik esan dugun bezala (III.1 atalean), euskararako EBLan, ale lexikalen adierez gain, hauen informazio sintaktiko-semantikoa adierazita etortzea nahiko genukeelako.

EuroWordNet WordNeten bertsio aurreratua izaki, tankera honetako informazio gehiagorekin hornitu da (kategoria ezberdineko *synset*en loturekin adibidez)[37]. Are gehiago MCR, EuroWordNeten gapapena baita. Azken honetan, adibidez, hautapen-murriztapenak txertatzeko asmoa dago. IV.3 atalean ikusiko dugun bezala, MCRko interfazeak hautapen-murriztapenak kontsultatzeko aukera ematen du, baina oraindik ez da informazio hau atzitu eta EBLan txertatu. Txosten honen VII. kapitulua lan honi dagokio, hain zuzen ere. Gerora, hautapen-murriztapenez gain, MCRren sintaxi-semantikari buruzko informazio gehiago txertatu nahi da, hala nola, funtzio gramatikalak. Beraz, esan daiteke, MCRk WordNet eta EuroWordNeten hezurdura duela, baina informazio sintaktiko-semantikoa jasotzeko aukerarekin.

[37] IV.2 atalean hitz egingo dugu erlazio semantiko hauei buruz.

Honenbestez, euskararako EBLa MCRren ereduan eraikiz gero, honek WordNet eta EuroWordNeten hezurdura izango luke, hots, adieraka antolatutako EBL semantiko eleanitz baten abantailak izango genituzke, eta, gainera, bi ezagutza-base hauetan dagoen informazioarekin batera, MCRn gehituko den informazio sintaktiko-semantikoa eskuragarri izango genuke.

Aipatu diren arrazoi horiek guztiak direla medio, euskararako EBLa MCRren eredua jarraituz egingo dugu eta, MCRk beste iturrietako informazioa jasotzeko oinarri sendoa duenez, ikerlan honetan landu ditugun beste formalismoetatik baliagarri zaigunari probetxua atera ahal izango diogu, MCRn behar bezala txertatuz gero. Alde batetik, EBLan ale lexikalak sailkatzeko erabiltzen dituzten ezaugarri batzuk, MCRn ez daudenak aprobetxa genitzake. Bestetik, MCRn sarrera lexikalak jasotzen ez duen informazioa jaso genezake[38].

Jarraian, tesi-lan honetan landutako ikerlan eta formalismoetatik MCRn sartzeko baliagarri izan daitekeen informazioa dagoen ala ez ere adieraziko dugu.

Hizkuntzalaritza teorikotik hiru lan aztertu ditugu: Jackendoff (1990), Levin (1993) eta Pustejovsky (1995).

Jackendoffen kasuan (III.2.1.1 atalean), Dorrek (1997) eta Fernández *et al.*-ek (2002) Jackendoffen eredua konputazionalki inplementatu (eta aberastu) dute, aditzen klaseak WordNeteko adieretara lotuz. Lotura hau euskarako aditzen sailkapenerako erabilgarri izan daiteke, noski, lehendabizi bertan dagoen informazioa euskararen izaera sintaktiko-semantikora egokitzen dela egiaztatu eta gero. IXA taldean Volem proiektuaren jarraipenean parte hartu duenez, horrelako esperimentuak egiteko aukera izan dugu. Aldezabalen (2004) lanean aztertutako ehun aditzak Volemeko eredura egokitu ditugu, eta aditz hauen adiera bereko frantseseko, gaztelaniako eta katalaneko ordainen errepresentazioarekin erkatu ditugu. Kasu gehienetan, hizkuntza guztietan, aditz-adiera berak egitura sintaktiko-semantiko bera du. Hala ere, ikerketa hauek tesi-lan honetatik kanpo geratu dira.

Jackendoffen ereduarekin esan dugun bezala, Dorrek Levinen klase semantikoak WordNetera lotuak ditu. Horrela bada, MCRren ildotik eginda-

[38]Kontuan izan behar da lan hauek ingeleserako pentsatuak daudela. Horregatik, EBL hauen informazioa euskararako EBLari gehitu baino lehen, informazio hori hizkuntzatik independentea den (unibertsala den), edo behintzat euskararako baliagarria den, frogatu beharko genuke. VII. kapituluan horrelako saiakera baten berri ematen dugu. Ingeleserako corpusetatik automatikoki lortutako hautapen-murriztapenak euskaratu, eta euskararako baliagarriak diren aztertu dugu (Agirre *et al.*, 2003a; Pociello, 2004a).

ko euskarako EBLrako, Levinen lanetik zuzenean informazioa atera ordez, Dorren lanetik abiatzea errazagoa litzaiguke. Horretarako, bete beharreko lehenengo pausua, Levinen aditz-klaseak eta MCRkoak zer puntutaraino pareka daitezkeen aztertzea litzateke.

Horrekin batera, Aldezabalen (2004) tesi-lanean Levinen lana erabili da euskal aditzaren azpikategorizazioa jorratzeko. Hortaz, eredu honen euskararako egokitzapena balia dezakegu MCR aberasteko.

Betalde, Agirre eta Lersundiren lanean (2003) Dorren ELKetako interpretazioak Aldezabalenekin parekatu ondoren, ingeleseko, gaztelaniako eta euskarako postposizioen adiera-inbentario bakarra lortu dute, eta postposizio bakoitza MCRra lotu dute. MCRn ez dago preposizioen/postposizioen *synset*ik, beraz, lotura hau era honetara egin dute: postposizioa jaso duen eratorriaren (zilargile) eta oinarriaren (zilar) arteko erlazio semantikoa ('IZEak ADIt(z)en dituena') adierazten dute MCRn. Lan honetako informazioa dagoeneko MCRn txertatuta dago.

Hizkuntzalaritza teorikoari dagokion atalean, aztertutako azken lana Pustejovskyrena (1995) izan da. Ezagutzen den inplementazioetako bat Buitelaarrena da (1998). Buitelaarrek Pustejovskyren alderdi semantiko batzuk (*alderdi konstitutiboa* adibidez) WordNeten dauden antzeko harreman semantikoekin erkatzen ditu. Berriro ere, euskarako EBLari begira, WordNeterako lotura hau ondo etor dakiguke Buitelaarren lanetik lortutako emaitzak gure EBLan eransteko.

Hizkuntzalaritza teoriko eta konputazionalaren erdibidean dauden lanek (LFG, GPSG eta HSPG) ingelesari buruzko informazio sintaktiko-semantikoaren deskribapen aberatsa dute. LFG, GPSG eta HPSG euskararako erabiltzeko saiakera bat egin da (Gojenola, 1998), eta hortik baliagarri izan dakigukeen informazioa lor genezake.

Azkenik, hizkuntzalaritza konputazionaleko lanak izan ditugu aztergai: FrameNet (Fillmore eta Baker, 2001), WordNet eta honen ildotik etorritakoak (Miller, 1985; Fellbaum, 1998a; Vossen, 1997; Atserias *et al.*, 2004), Volem eta PropBank proiektua (Palmer eta Kingsbury, 2003) (Fernández *et al.*, 2002). WordNet, EuroWordNet, MCR eta Volemi buruzko ondorioak gorago aipatu ditugunez, zuzenean FrameNet eta PropBanki buruz jardungo gara.

Esan bezala (III.2.4 atalean), PropBankeko sarrera lexikalak VerbNeten hauei dagokien sarrerarekin lotuta daude. Aldi berean, VerbNeteko sarrera hori WordNeteko *synset* batekin (edo gehiagorekin) loturik dago. Hortaz, lotura honi probetxu atera geniezaioke gure EBLko aditzak VerbNet eta

PropBankeko informazio sintaktiko-semantikoarekin aberasteko.

FrameNeten kasuan ere antzeko zerbait egin daiteke. LNPn rolen informazio sintaktiko-semantikoa erauzteko eta markatzeko oso ezagunak dira, bai PropBank, bai VerbNet eta baita FrameNet ere. Arrazoi honengatik, hiru baliabideetako informazio bateratua erabiltzeko saiakerak egon dira. Giuglea eta Moschittiautoreek (2004), adibidez, PropBank eta FrameNeten arteko lotura egiteko VerbNet erabili dute. Horretarako, VerbNeteko klase semantikoen eta FrameNeteko *frame*en mapaketa egin dute. Adibidez, VerbNeteko *Judgement* klase semantikoa FrameNeteko *Rewards and punishments, Judgement communication, Sentencing, Notification of charges, Arrangement, Court examination, Pardon, Try defendant, Forgiveness, Jury deliberation* eta *Judgement direct address frame*ekin parekatu dituzte. Hala, klase semantiko bakoitzeko hiru EBLen informazioa dute eskuragarri. Mapaketa hau corpusean rolak automatikoki ezagutzeko egin da.

Horrela, bada, FrameNet VerbNetekin lotuz gero, VerbNeteko aditzak WordNeteko *synset*ekin parekatuak daudenez, EBL hauetako guztietako informazioa izango genukeen eskuragarri.

III.4 Ondorioak

Kapitulu honetan arrazoitzen saiatu gara euskararako EBLa egiteko WordNeten eredua (zehazkiago, MCRrena) jarraitzea dela biderik egokiena. Erabaki hori hartu dugu euskarako EBLrako nahiko genituzkeen ezaugarriak ondo definitu ondoren —konputazionalki inplementa daitekeena izatea, hizkuntza bere osotasunean adierazten duena izatea, eleanitza izatea, eta informazio berrerabilgarria jasotzen duena izatea—, ezaugarri hauen arabera mugatu dugu gure proposamena:

- WordNet eta honen ildotik garatu diren EuroWordNet eta MCR ez daude teoria bakar bati lotuta, bestelako eredu eta teoria ezberdinekin erabil daitezke. Horren proba da formalismo eta lan teoriko asko, gerora, WordNeten adiera edo/eta klase semantikoekin aberastu dituztela.

- WordNet, EuroWordNet eta MCR lexiko zabala eta garatua dute; sarrera bakoitzean ale lexikalaren adiera, klase semantikoa, kategoria eta beste sarrerekin izan ditzaken erlazio semantikoak jasotzen dituzte. Esate baterako, WordNeten 3.0 bertsioan 117.617 *synset* daude (81.426 izen, 13.650 aditz, 18.877 adjektibo eta 3.664 adberbio).

- WordNet, EuroWordNet eta MCR inplementatutako EBLak dira. Honen adierazgarri dira WordNeten oinarrituta egin diren publikazioen kopurua (gaur egun, WordNeteko web orriak[39] 422 inguru jasotzen ditu).

- WordNet EBL elebakarra izan arren, honen ildotik sortutako EuroWordNet eta MCR eleanitzak dira.

Aukeraketa hau, halere, ikerkuntzaren ikuspegian, helburuen edota ematen zaizkion erabileren mende dago.

Bestalde, behin MCRren aldeko aukera eginda, eredu hau beste lan eta formalismoetako informazioarekin osa dezakegula ikusi dugu. Hala ere, formalismo desberdinak direnez eta batzuetan beraien artean kontraesanean daudenez, hauen artean hautu bat ere egin beharko genuke. Hau da, MCRren ildotik egingo den euskararako EBLa hauetako zeinekin osatzea komeniko litzatekeen erabaki beharko genuke.

Dena den, lan hori ez dugu tesi-lan honetan jorratuko; etorkizunerako lan gisa proposatuko dugu.

[39] `http://www.wordnet.princeton.edu` (2007-07-02an atzitua).

IV. KAPITULUA

WordNet, EuroWordNet eta MCR

Kapitulu honetan *WordNet* (IV.1), *EuroWordNet* (IV.2) eta *The Multilingual Central Repository* ereduaren (MCR) (IV.3) azterketa sakonago bat egingo dugu.

IV.1 WordNet eta WordNetetik abiatutakoak

IV.1.1 Sarrera

WordNet (Miller, 1985; Fellbaum, 1998a) teoria psikolinguistikoetan oinarritua dagoen ingeleseko ezagutza-basea da. Princeton-eko Unibertsitatean eskuz garatzen ari da —*Cognitive Science Laboratory* delakoan— George A. Millerren ardurapean.

Ingeleseko izenak, aditzak, adjektiboak eta adberbioak ***synonym set*** edo ***synset***etan (sinonimo multzotan) antolatuak daude, hauetako bakoitza kontzeptu lexikal bati (adiera bati) dagokiolarik. Esaterako, ingeleseko tree izenak WordNeten bi *synset*[1] ditu[2]:

[1] Aurrerantzean *synset* terminoa erabiliko dugu, adiera edo kontzeptu lexikalaren pareko.

[2] Kapitulu honetako WordNeteko adierazpen guztiak WordNet 3.0 bertsiotik hartu ditugu —`http://www.wordnet.princeton.edu` (2007-07-02an atzitua)—, eta leku-arazoengatik adibide batzuk moztu egin ditugu.

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

Lehenengoa 'landare' (plant) *synset*ari dagokio, eta bigarrena, berriz, 'diagrama' (diagram) *synset*ari. *Synset*ak desberdindu ditzakegu hauen ondoan gehienetan datorren glosei esker. (1) adibidean tree izenaren 'landare' adieraren glosa a tall perennial woody plant having a main trunk and branches da. Lehenengo *synset* hau ale lexikal bakar batez osatua dago (tree); hots, tree izenak, *synset* horretan, ez du sinonimorik. Bigarrenak, ordea, tree ale lexikalaz gain, beste ale bat ere badu *synset*ean (tree diagram). Bi ale lexikal horiek (tree eta tree diagram) sinonimoak dira. *Synset*a osatzen duten ale lexikalei ***variant*** deitzen zaie, beraz, *synset* berean dauden *variant*ak sinonimoak dira.

Hain zuzen ere, sinonimia da WordNeteko erlazio semantiko garrantzitsuenetarikoa. Izan ere, ezagutza-basearen oinarria ale lexikala izanik, adiera batek ale lexikal bat baino gehiago dituenean, ale lexikalak multzokatu egiten ditu sinonimia erlazioak.

WordNeteko sinonimiaz hitz egiterakoan, kontuan izan behar da ez dela gauza bera sinonimia eta hitzak bata bestearekin elkar trukatzea. Hau da, WordNeteko *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, 1998a, 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, WordNetek 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:

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

(2) Sense 1
{tree} (a tall perennial woody plant having a main trunk and...)
=> {woody plant, ligneous plant} (a plant having hard lignified...)
=> {vascular plant, tracheophyte} (green plant having a...)
=> {plant, flora, plant life} (a living organism...)
=> {life form, organism, being, living thing}
=> {entity, something} (anything having existence)

Sense 2
{tree , tree diagram} (a figure that branches from a single root)
=> {plane figure, two-dimensional figure} (a 2-dimensional shape)
=> {figure} (a combination of points and lines and planes...)
=> {shape, form} (the spatial arrangement of something...)
=> {attribute} (abstraction belonging to a...)
=> {abstraction} (a general concept formed by ...)

(3) Sense 1
{tree} (a tall perennial woody plant having a main trunk and...)
=> {yellowwood, yellowwood tree} (any of various trees having...)
=> {lancewood, lancewood tree} (source of most of the lancewood...)
=> {Guinea pepper, negro pepper} (tropical west African tree...)
=> {anise tree} (any of several evergreen shrubs ...)
=> {winter's bark tree, Drimys winteri} (South American tree...)
=> {zebrawood tree} (any of various trees ... having mottled or ...)
=> {granadilla tree, Brya ebenus} (West Indian tree yielding ...)
=> {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)
=> {cladogram} (a tree diagram used to illustrate phylogenetic...)

(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 da.

Treeren 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[4]. 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, right} (in the right manner)
=> {improperly} (in an improper way)

WordNeteko sailkapena, beraz, *synset*etan eta beraiek harremanetan jartzen dituzten erlazio semantikoetan datza. Erlazio semantiko hauen bidez, *synset*ak hierarkikoki multzokatzen dira, edo, beste era batera esanda, klase semantikoak osatzen dira. Horrela, WordNetek izenak hierarkiatan banatzen ditu, eta hierarkia hauetako bakoitza klase semantiko bati dagokio. Klase semantiko hauetako bakoitzean, klase horretako izenen antolaketaren hastapena dago, *unique beginner* deritzona. Hau izango da klase semantiko horren hierarkian mailarik altuena eta orokorrena, eta bere ezaugarri guztiak bere hiponimoek heredatuko dituzte. (5)eko taulan WordNeteko izenak sailkatzen dituzten 25 *unique beginner*rak datoz zerrendatuta. Aldi berean, *unique beginner* horiek WordNeteko izenek osatzen dituzten klase semantikoak adierazten dituztela esan dezakegu, *unique beginner* bakoitzaren azpian klase horri dagozkion izen guztiak jasotzen baitira. Esate baterako, food *unique beginner*raren azpian janariarekin zerikusia duten izenak egongo dira hierarkikoki antolatuta. Ondorioz, multzo horrek janariari dagokion klase semantikoa osatzen du.

(5)

{act, action, activity}	{animal, fauna}	{artifact}
{tribute, property}	{body, corpus}	{cognition, knowledge}
{communication}	{event, happening}	{feeling, emotion}
{food}	{group, collection}	{location, place}
{motive}	{natural object}	{natural phenomenon}
{person, human being}	{plant, flora}	{possession}
{process}	{quantity, amount}	{relation}
{shape}	{state, condition}	{substance}
{time}		

[4]Aditzen kasuan, eta gero IV.1.2 atalean ikusiko dugun bezala, *hiperonimia-troponimia* erlazioaz hitz egiten da.

Honezaz gain, izenak klase semantikoetan banatuak egoteak badu beste arrazoi praktiko bat: klase semantiko bakoitza fitxategi batean jasota dago (*semantic field* deiturikoa)[5]. WordNet garatzeko lexikografoek hogeita bost fitxategi hauek beraien artean banatu eta fitxategiz fitxategi ingeleseko WordNet osatzen joan ziren[6]. Hala, lexikografo bakoitzak eremu semantiko bereko kontzeptuak lantzen zituen.

Ondoren (IV.1.2 atalean), ikuspegi sintaktiko-semantikoan sakontzearren, aditzaren azterketan murgilduko gara.

IV.1.2 Aditza eta informazio sintaktiko-semantikoa

Askotan aipatu dugun bezala, sintaxi-semantika elkargunearen muina aditza da, esaldiaren antolakuntza hartzen baitu bere baitan. Arrazoi honengatik, WordNeten jasota dagoen informazio sintaktiko-semantikoa aditzari lotua dago.

WordNeten aditzen *synset*ak, irizpide semantikoan oinarrituz, 14 klase semantikotan banatuak daude (*motion; perception; contact; change; communication; competition; cognition; consumption; creation; emotion; perception; possession; bodily care and functions; verbs referring to social behaviour and interaction*). Bestetik, 14 klase semantiko horietan lekurik ez duten aditzen multzoa dago (*verbs denoting states* delakoan), eta aditz hauek (be, belong, resemble...) egoera adierazten dute[7].

Izenekin ikusi dugun bezala, klase semantiko hauetako bakoitzean aditz horien antolaketaren hastapena dago, *unique beginner* deritzona. Esaterako, *communication* klase semantikoak *unique beginner* bezala communicate *synset*a du eta honetatik hasten da klase semantiko honetako aditzen sailkapena.

[5] Euskaraz *eremu semantiko* deritzogu.

[6] Hogeita bost *unique beginner*ren artean hainbat multzo egin dira. Esate baterako, horietatik zortzi tangible things bezala sailkatu dituzte, bost abstraction bezala; eta hiru psychological features bezala. Hala, *unique beginner*ren kopurua hogeita bostetik hamaikara murriztu dute.

[7] Izenekin bezala, klase semantiko bakoitza fitxategi batean jasota dago.

Klase semantiko hauek aditzen sailkapenerako aproposak izan arren, euren arteko muga ez da guztiz hertsia. Hori dela eta, aditz batzuk klase semantiko bat baino gehiagotan egon daitezke; adibidez, ingeleseko The bullet whistled past him[8] esaldian, whistle aditza *communication* klaseari dagokion *synset* bat du (make whistling sounds glosaduna), eta *motion* klase semantikoari dagokion beste *synset* bat du (move with, or as with, a whistling sound glosaduna).

Gorago azaldu dugun bezala (IV.1.1), WordNet *synset*en arabera dago antolatua, eta, beraz, *synset*a osatzen duten sinonimoak beraien artean truka daitezke testuinguru konkretu batzuetan. Aditzen kasuan trukatze hau bideratzea zaila gertatzen da. Batzuetan aditzek —end/terminate eta rise/ascend bezalako anglosaxoi/grekolatindar hitz pareek adibidez— adiera bera izan arren, erregistro ezberdina eskatzen dute. Adibidez, anglosaxoi/grekolatindar hitz pareen kasuan grekolatindarrek besteak baino erabilera jasoagoa dute.

Beste batzuetan, ordea, aditzen arteko adiera-aldaketa hautapen-murriztapen ezberdinekin azaleratzen da. Esaterako, ingeleseko rise eta fall aditzek entitate abstraktuak (temperature, prices...) har ditzakete argumentu gisa; aurrekoen adieraren oso antzekoa duten ascend eta descend aditzek, berriz, ezin dute argumentu mota honekin agertu (Fellbaum, 1998a). Horrelako kasuetan, WordNeten irizpide nagusia aditzak *synset* desberdinetan banatzea da, hau da, rise eta ascend bi *synset*etan kokatzea.

Hortaz, hautapen-murriztapenak kontuan hartzen dituzte hierarkia osatzeko garaian, baina ontologian oraindik ez dago adierazita zeintzuk diren aditz bakoitzak hartzen dituen hautapen-murriztapen konkretuak. Hau da, WordNeteko interfaze informatikoak ez du eskaintzen rise eta entitate abstraktuak (WordNeten abstraction ale lexikala daraman *synset*aren bitartez adierazten dena) hautapen-murriztapen gisa lotzeko biderik.

Hautapen-murriztapenekin bezala, ale lexikal baten *synset*ak ezberdintzerakoan azpikategorizazioa kontuan hartzen dute, informazio hau aditzaren adiera bakoitzeko proposatuz, baina rol tematikorik aipatu gabe:

[8] Adibidea Fellbaumen lanetik (1998a) hartua da.

(6) 4 senses of "descend"
Sense 1
{descend, fall, go down} (move downward but not necessarily all the way)
EX: The airplane is sure to descend
Sense 2
{derive, come, descend} (come from; be connected by a blood relationship)
Something is — -ing PP
Somebody — -s PP
Sense 3
{condescend, descend} (do something that one considers to be below...)
Somebody — -s to INFINITIVE
Sense 4
{stoop, descend} (to sink in status or dignity, or worsen in condition)
Somebody — -s PP

Horrela, bada, WordNet, aurretik ikusi ditugun lanen eredutik banandu egiten da, *semantika deskonposatzailea* jarraitzen dutenetatik alegia. Jackendoff-ek bere lanean (1990), adibidez, primitiboak baliatuta egiten du aditzen azterketa (*TO*, *FROM*, *TOWARD*, *AWAY-FROM*, *CAUSE*, *GO*, *VIA*...). WordNeten ale lexikalak ez daude unitate txikiagoetan deskonposatuak. WordNetek *loturazko semantikaren* (*relational semantics*) ildotik jorratzen ditu aditzak; hortaz, *synset*ak hitzekin osatzen dira eta ez tasun edo primitiboekin. Hala eta guztiz ere, *synset*en arteko harreman semantikoek deskonposaketaren alderdi batzuk ere eskain ditzakete. Nahiz eta WordNetek primitiboak edo antzeko tasun txikiagoak ez erabili, hauetako batzuk agerian geratzen dira harreman semantikoen bidez. Adibidez, *semantika deskonposatzailean* oihartzun gehien duen tasunetako bat *kausa* da (*CAUSE* primitiboa deitzen duena Jackendoffek). WordNeten informazio hau *cause* erlazio semantikoarekin ikus dezakegu, eta bere bitartez learn aditza teach aditzaren ondorioa dela jakin dezakegu:

(7) 1 of 6 senses of "learn"
Sense 5
{teach, learn, instruct} (impart skills or knowledge to)
=> {learn} (acquire or gain knowledge or skills)

Bestetik, mugimendua adierazten duen tasunak (Jackendoffek (1990) *GO* deitzen duenak) hierarkiaren hastapen diren *unique beginner*rek adieraz ditzakete. Run aditza adibide gisa hartuz gero, bere hiperonimo garaiena —*motion* klase semantikoaren *unique beginner*ra dena—, {go, move, travel,

locomote} *synset*az osatzen da[9], eta honek erakusten digu run mugimenduzko aditza dela.

(8) Sense 1
{run} (move fast by using one's feet, with one foot off the ground at any...)
=> {travel rapidly, speed, hurry, zip} (move very fast)
=> {travel, go, move, locomote} (change location)

Amaitzeko, aditzen *modua*ren berri hierarkian bertan dugu. Arestian hitz egin dugu hiperonimia-hiponimia erlazio semantikoaz. Aditzek erlazio honen antzekoa duten arren, Fellbaumek (1998b) hiponimiaren ordez **troponimia** erabiltzea erabaki zuen. Honen arrazoia da aditzek dutela *IS-A* erlazioa betetzen. Honen ordez, *to x is to y in some particular manner* definitzen da aditzen hierarkiak osatzeko. Hortaz, aditz hiperonimo baten (walk) troponimoak aditz hiperonimoak adierazten duena egiteko moduak izango dira (trot, march...). Hala, WordNetek hitzaren kategoriaren arabera baliabide semantiko desberdinak erabiltzen ditu ezagutza sintaktiko-semantikoa berri emateko. Ezagutza-baseko sarrera lexikal bakoitza ez dator zehaztuta tasun zerrenda batekin; zehaztuta etorri beharrean, bere zehaztapena hierarkiatik jasotzen dituen tasunetatik dator.

IV.1.3 Bestelako erlazio semantikoak

Sinonimia eta hiperonimia-hiponimia/troponimia erlazio semantikoez gain, WordNetek beste asko landu ditu. Hemen batzuen aipamen laburra egingo dugu[10].

Izenak lotuak egon daitezke ondorengo erlazio semantikoen bidez:

- ***Part-whole relations:***

Zatia eta osotasuna harremanetan jartzen dituen erlazioak dira. Batetik, **meronimia** dago, *X is a meronym of Y if Ys are parts of X* definizioari jarraitzen diona; hatzak (9. adibidean, finger) eskuen (adibidean, hand) zati bat dira, eta eskua, aldi berean, besoarena (adibidean, arm):

[9] *Motion* klase semantikoak bi *unique beginner* ditu, bata {go, move, travel, locomote} (change location), eta bestea, {move, displace} (cause to move); lehenengoan 'norbait/zerbait mugitzen da', bigarrenean 'norbaitek/zerbaitek norbait/zerbait mugitzen du'.

[10] Argibide gehiago Fellbaumen (1998a) eta Millerren (1985) lanetan.

(9) 1 of 2 senses of "finger"
Sense 1
{finger} (any of the terminal members of the hand)
PART MERONYM: {hand, manus} (the extremity of the superior limb)
PART MERONYM: {arm} (the part of the superior limb between...)

Bestetik, **holonimia** kontrako erlazioa da, *x has a y (as a part)* definizioarekin bat datorrena. Adibidez, eskuek (10. adibidean hand) hatzak dituzte (10. adibidean, finger):

(10) 2 of 14 senses of "hand"
Sense 1
{hand} (the extremity of the superior limb)
PART HOLONYM: {finger} (any of the terminal members of the hand)

- ***Antonimia*:**

Izen batzuek antonimoak dituzte eta erlazio semantiko honek lotzen ditu:

(11) 1 sense of "victory"
Sense 1
{victory, triumph} (a successful ending of a struggle or contest)
ANTONYM: {defeat, licking} (an unsuccessful ending)

- ***Inplikazioa*:**

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

(12) 1 sense "snore"
Sense 1
{snore} (breath noisely during one's sleep)
ENTAILMENT: {sleep} (be asleep)

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

IV.1.4 Erabilera

WordNetek 117.617 *synset* ditu (81.426 izen, 13.650 aditz, 18.877 adjektibo eta 3.664 adberbio)[11].

WordNeten erabilerak era askotakoak izan dira. Alde batetik, hiztegi eta thesaurus gisa erabili izan da. Hiztegi tradizionaletan bezala, WordNetek *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.

Esan beharra dago, WordNet ezaugarri psikolinguistikoetan oinarrituta egon arren, psikolinguistek ez dutela kontu handian hartu eta hizkuntzalari konputazionalei interesgarriagoa iruditu zaiela. Hala, LNPri begira, WordNetek erabilera ugari izan ditu. WordNeteko web orrian agertzen den bibliografian[12] hau erakusten duten 2.000 artikulu inguru daude. Guk arlo bakoitzetik garrantzitsuenak baino ez ditugu aipatuko:

- **Hitzen adieren desanbiguazioan:** WordNet adieran oinarritutako ontologia denez, WordNeteko informazioak, hau da, adierak hierarkikoki antolatuta egoteak desanbiguazioaren atazan lagundu egiten du. Hots, hitzaren testuinguruan dauden beste hitzei erreparatuta, eta desanbiguatu nahi den hitzaren WordNeteko erlazio semantikoak ezagututa, hitzaren adiera zuzen posibleen aukera aukera txikitu egiten da. Adibidez, This letter has no address esaldian, letter hitzak, gutxienez, bi adiera izan ditzake: bata, 'gutun' adiera, eta bestea 'hizki' adiera. Hiztegi arruntetan, hitz hauen adieraren definizioa izango genuke. Aldiz, WordNetek bi adiera hauen glosak emateaz gain, hiztegietan ez dagoen, eta desanbiguaziorako oso erabilgarria den, informazio gehigarria ematen digu: erlazio semantikoak. Esate baterako, 'gutun' adiera duen *synset*a address *synset*arekin lotua dago meronimia erlazioaren bitartez. Kasu honetan, desanbiguazio algoritmoak WordNeteko erlazioak eta testuinguruan duen informazioa erabilita, letter hitzari 'gutun' adiera egokituko dio. Arlo honetan esperimentu ugari egin dira (Miller *et al.*, 1994; Banerjee eta Pedersen, 2002; Agirre eta Martínez, 2000; Matwin *et al.*, 1995).

[11] WordNeten azkeneko bertsioaz ari gara, 3.0 bertsioaz, alegia: `http://www.wordnet.princeton.edu` (2007-07-02an atzitua).

[12] Ikus `http://engr.smu.edu/rada/wnb/web` orrian (2007-07-02an atzitua).

- **Itzulpen automatikoan:** Itzulpen automatikorako sistemek hiztegi edo EBL bat behar dute, batetik, hitzen adieren desanbiguaziorako, eta bestetik, desanbiguatutako adierari dagokion erdarako ordaina egokitzeko. Letter adibidearekin ikusi dugun bezala, WordNetek hitzen adieren desanbiguazioan lagun dezake, baina ingeleseko EBLa izaki, ezin ditu erdarako ordainak esleitu; hau da, ezin du letter izena gutun edo carta bezala itzuli. Horretarako, beste hizkuntzetako hiztegi eta EBLekin bateratu behar da, eta horixe izan da zenbait lanetan egin dena: Dorr (1993, 1997) Rigau *et al.* (1995), Knight (1993), Moon eta Kim (1995) eta abar. Esate baterako, Knightek (1993) WordNetez gain, *The Harper Collins Spanish-English/English-Spanish Dictionary* (Collins, 1971) eta gaztelaniako ULTRA lexikoia erabili ditu. Hala ere, itzulpen automatikoko erabilera areagotu egin da, WordNeten ondorengo ereduekin (EuroWordNet eta MCR), hauek EBL eleanitzak baitira.

- **Informazio-erauzketan:** WordNet lagungarria izan daiteke erabiltzaileari beharrezkoa zaion edukia bere barne daukan dokumentua aurkitzeko. Bilaketan erabilitako hitzek indexatutako dokumentuetan daudenen berdinak izan behar dute[13], emaitza egokia lortzeko. Baina, askotan gertatzen da erabiltzaileak galderan erabilitako hitza ez egotea indexatua. Kasu horretan, WordNeten erlazio semantikoek lagun dezakete, informazio-erauzketa sistemaren emaitzak hobetuz: sistemak erabiltzaileak idatzitakoa (demagun, dog dela) *hedatu* egiten du; hau da, hitz horren sinonimoak (canis familiaris), hiponimoak (puppy, hunting dog, dalmatian, Pekinese...) eta hiperonimoak (canine, domestic animal...) bilatzen ditu. Hala, dog hitzari buruzko galdeketa eginez gero, sistemak hitz honi lotutako dokumentuak zerrendatzen ditu. Zenbait saiakera egin dira. Esaterako, Magnini eta Strapparava (2001), Mandala *et al.* (1998), Milhacea eta Moldovan (2001), besteak beste.

- **Galdera-erantzun sistemetan:** WordNeteko *synset*en arteko harremanek galdera bati dagozkion erantzunak ezagutzen laguntzen dute (Pasca eta Harabagiu, 2001; Harabagiu eta Moldovan, 1996; Mann, 2002; Ansa *et al.*, 2005, eta abar). Galdera-erantzun sistemek erabiltzaileak idatzitako galderaren (adibidez, Nor da Kubako gobernuburua?) erantzuna lortzen dute. Horretarako, informazio-erauzketan

[13] Informazio-erauzketa egin ahal izateko, aldez aurretik, dokumentuak egituratu behar dira, gero sistemari bilaketak errazteko.

bezala, galderan erabilitako hitzak indexatutako dokumentuetan agertu behar dute, hauetatik erantzun zehatza lortu ahal izateko. Hala, galde-erantzunean informazio-erauzketa beharrezkoa da, galderaren erantzunak indexatutako dokumentuetan bilatzen baitira. Beraz, hemen ere WordNeten erlazio semantikoak erabilita galdera *hedatu* egiten da: esate baterako, gobernu-bururen hiponimoak lehendakari eta presidente dira, eta hiperonimoak ordezkari, pertsona eta abar. Hauei esker, galderaren erantzuna bilatzeko erabili behar diren dokumentuen esparrua handitu egiten du. Hau da, Kubako gobernu-burua duten dokumentuak begiratzeaz gain, sistemak Kubako presidentea edota Kubako lehendakaria duten dokumentuetan ere begiratuko du erantzunaren bila.

Azkenik, nabarmendu nahi dugu, **WordNetekin etiketatutako corpusa** —*SemCor* (Miller *et al.*, 1994; Fellbaum *et al.*, 2001)— oso lagungarria gerta daitekeela ataza hauentzat guztientzat. Sistemek corpusetik ikasi egiten dute. Arestian aipatutako adibidearekin jarraituz, letter hitza 'gutun' adierarekin etiketatutako agerpenetan zein testuingurutan agertu den ikasiko du. Hau da, letter hitza gutun adierarekin agertu den bakoitzean, bere testuinguruko hitzak (eta hitz horien adierak) zein diren *memorizatu* egingo du nolabait makinak. Honela, letter hitzaren hurrengo agerpenetan, memorizatutako informazio honetan oinarrituko da makina erabaki bat hartzeko. Hau guztia teknika estatistiko konplexuak erabiliz egiten da.

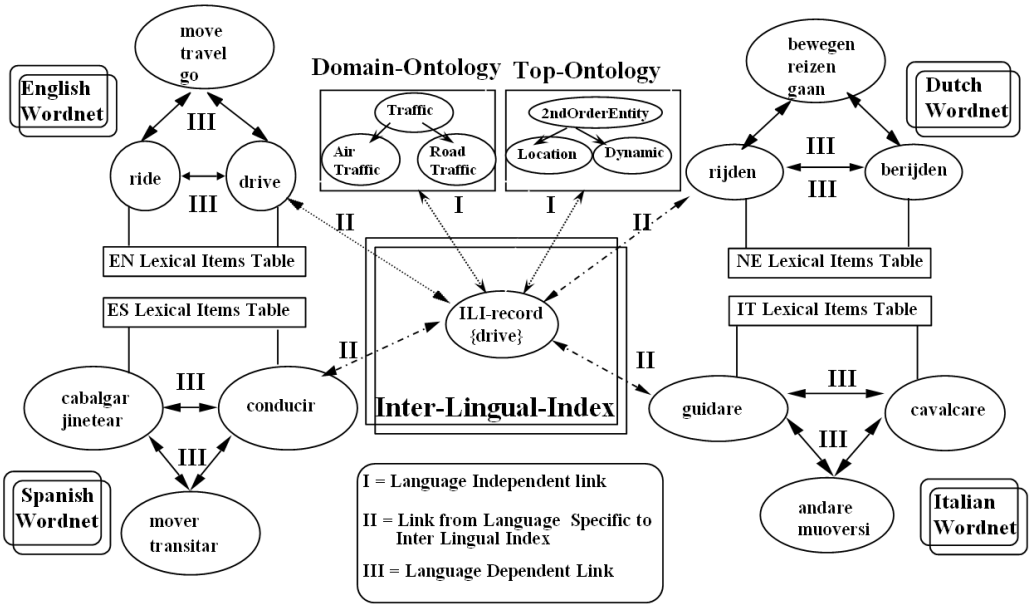
Aipatu dugun bezala, WordNet abiapuntutzat hartuta, beste ezagutza-base batzuk egin dira: EuroWordNet (Vossen, 1998) eta The Multilingual Central Repository (MCR)(Atserias *et al.*, 2004). Oinarri bera erabili arren, aberasketa batzurekin hornitu dute WordNet. Hurrengo ataletan (IV.2 eta IV.3) hauen berri emango dugu, oso laburki bada ere. Batez ere, sintaxi-semantika elkarguneari erreparatuko diogu.

IV.2 EuroWordNet

EuroWordNet proiektua (Vossen, 1998) 1996an hasi eta 1999 urteraino luzatu zen proiektu europarra da. Ezagutza-base eleanitza da, Europako zortzi hizkuntzatara zabaltzen dena (ingelesa, nederlandera, italiera, gaztelania, alemana, frantsesa, txekiera eta estoniera).

EuroWordNetek Princetoneko WordNetaren eredua jarraitzen du (ikus IV.1 atala); hots, Princetonen ingeleserako egindako WordNetaren hierarkiaren ideia bera darabil honek ere, eta modu berean adierazten dira, bai *synset*ak, bai erlazio semantikoak.

Nahiz eta EuroWordNeten hizkuntza bakoitzak wordnet "independente" bat izan, EuroWordNeten helburua wordnet desberdin hauek guztiak ezagutza-base eleanitz bakarrean elkartzea da. Horretarako, hizkuntza guztien wordnet guztiek elkargune bat dute, *Inter-Lingual-Index*a (hemendik aurrera ILI) deritzona, Princetoneko WordNet 1.5 bertsioari lotua dagoena. ILI honen bitartez, hizkuntza guztietako wordnetak lotuak daude. Beste hitz batzuetan esanda, *synset* bera ingelesez, nederlanderaz, italieraz, gaztelaniaz, alemanez, frantsesez, txekieraz eta estonieraz agertzen da.



IV.1 Irudia: EuroWordNeteko arkitektura.

IV.1 irudiak eskematikoki wordnet desberdinen eta ILIaren arteko harremanak azalerazten ditu. Erdian ILIa dago, non *ILI-records* deiturikoak jasotzen diren. *ILI-record* bakoitza wordnetetako *synset* bati lotua dago[14]. Esate baterako, irudiko *ILI-record*a gaztelaniako conducir *synset*ari lotua da-

[14]ILIko adierak Princetoneko WordNet 1.5 bertsiotik ateratakoak dira.

go, eta baita adiera hori bera duten beste hizkuntzetako driveri, rijdeneri eta guidareri ere. Hala, ILIan *ILI-record*ak daude, eta hauek hierarkian antolatu gabeko adieren zerrenda osatzen dute. ILIan adierak antolatu gabe egotean, ILIaren mantentze-lanak erraztu egiten ditu (bertsioen eguneraketak eta bestelako aldaketak eragozten dira honela). Hala ere, *ILI-record*en egitura erauzi daiteke wordnet independenteetatik; hots, irudiko *ILI-record*aren harreman semantikoak wordnet bakoitzean zeintzuk diren jakin dezakegu, ILI horrek wordnet independente guztiekin lotura duelako, eta wordnet independenteetako *synset*ak hierarkikoki antolatuta daudelako.

WordNeten egitura, erlazio semantikoetan eta *synset*etan oinarritu arren, WordNetek ez zituen ezaugarri batzuk EuroWordNeten gaineratu dira. Aldaketarik aipagarrienak hurrengoak dira[15]:

- **Erlazio semantikoen aberasketa:**

WordNeteko erlazio semantiko batzuk findu egin dituzte eta beste erlazio semantiko batzuk aberastu. Batez ere, morfologikoki aldatzen diren kategoria ezberdinen arteko erlazioak ugaritu dituzte (nice eta niceness bezalakoak, alegia).

Bestalde, EuroWordNetek ez du WordNeten interfaze informatikoa; EuroWordNetena interfaze berria da, hizkuntza bakoitzeko wordnetak erlazio berriak gehitzeko aukera duelarik.

- **Hierarkiaren aberasketa:**

WordNetek 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[16].

Bigarrenak, wordnet ezberdinetan gehien erabilitako *synset*ak oinarrizko ezaugarri semantikoen arabera sailkatzea ahalbidetzen du[17], eta nolabait

[15] Argibide gehiago Vossenen lanean (1998).

[16] Domeinuen sailkapena ez da EuroWordNeteko interfazean ikusten, beste fitxategi batzuetan daude.

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

esateko, EuroWordNeteko domeinuen antza badute ere, hauen garapenean motibazio linguistiko sakonagoak hartu dira kontuan. Hau da, tasun linguistikoak ([+/- bizidun], [+/- egile] adibidez) kontuan hartzen dituen ontologia da eta wordnetak tasun hauen arabera eraikitzen dira. Hortaz, ale lexikal bat [+biziduna] bada Goi-ontologiaren [+biziduna] adabegiaren azpian kokatuko da eta [-biziduna] bada, aldiz, [-biziduna] ezaugarriaren azpian. Hala, WordNeten hierarkia mantentzen dute, baina, hierarkia hau ontologia linguistikoago batekin aberasten dute.

Oinarrizko ezaugarri semantikoak definitzerakoan, EuroWordNeten sortzaileak hizkuntzalaritzan egon diren zenbait sailkapen semantikoen eredutan oinarritu dira: Vendler (1967), Verkuyl (1972), Dowty (1979), Pustejovsky (1991), Levin (1993), Lyons (1977) eta Pustejovsky (1995) autoreen ereduetan, 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 daitezkeen eta denboran/lekuan antzeman daitekeen entitateak dira (animalia, objektu, substantzia eta antzeko 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.

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

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 (orotarikoak), aditzak eta adjektiboak
- **Hirugarren mailako entitateak (*3rd Order Entity*)**: izen abstraktuak

IV.1 irudian Goi-ontologiaren hierarkia guztia zerrendatua ikus dezakegu.

Goi-ontologiak EBLaren informazioa aberasteaz gain, beste zeregin bat ere badu: wordnet independenteen bateragarritasuna ziurtatzea. Esan dugun bezala, gehien erabiltzen diren *synset*ak oinarrizko ezaugarri semantikoen arabera sailkatzen ditu goi-ontologiak. Gehien erabiltzen diren *synset* hauei *oinarrizko kontzeptuak* (*Base Concepts*) deitzen zaie eta beraien ezaugarriak hurrengoak dira:

- Harreman semantikoen kopuru handiena duten *synset*ak dira.
- Hierarkian goi aldeko *synset*ak dira.
- Wordnet guztietan agertuko dira.

EuroWordNeteko datu-baseak hizkuntza bakoitzeko wordnet independente bat egiteko aukera ematen du, baina gutxieneko bateragarritasun bat ziurtatzeko oinarrizko kontzeptuen zerrenda adostu egin zen, eta wordnet bakoitzaren garapena *synset* horiekin hasi zen. Hortaz, wordnet guztiek izango dituzte oinarrizko kontzeptu berdinak, eta hierarkikoki era berean antolatuak egongo dira.

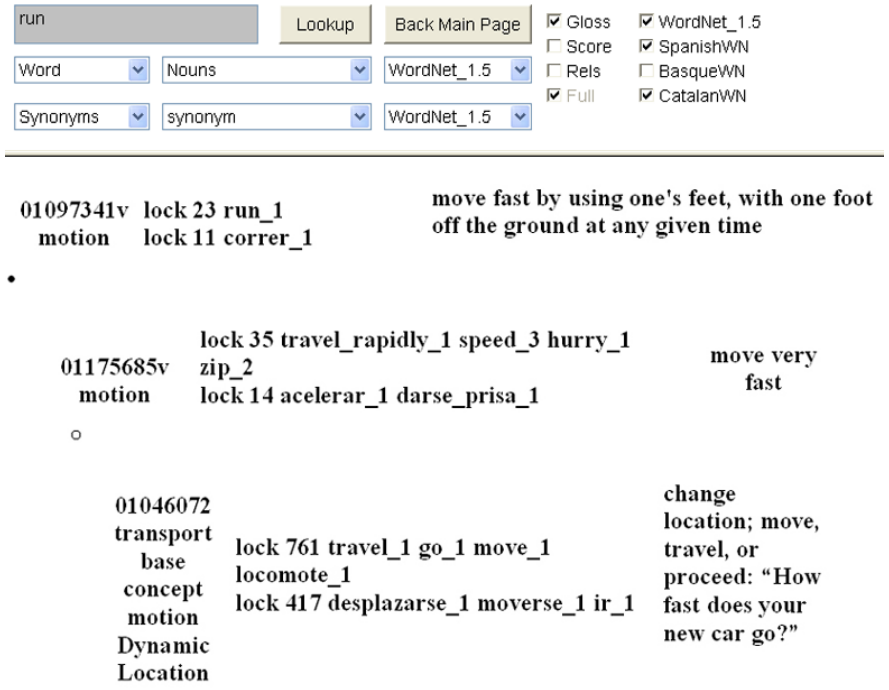
IV.1 irudian ikus daitekeen bezala, Goi-ontologia eta Domeinu-ontologia wordnetetatik independente dauden moduluak dira. Hauen ezaugarriak *ILI-record*ek jasoko dituzte, eta *ILI-record* horien bitartez wordnetetako *synset*ek. Esate baterako, *Location* eta *Dynamic* goi-ezaugarriak drive *ILI-record*ari daude zuzenean lotuta, eta, ondorioz, ezaugarri hauek zeharka jasotzen dituzte *ILI-record* horrekin harremanetan dauden wordnet desberdinetako kontzeptuek (guidare, conducir, drive, rijden).

EuroWordNet WordNeten oinarritutako ezagutza-basea denez, informazio sintaktiko-semantikoa, WordNeten parekoa da (ikus IV.1.2 atala). Hala eta guztiz ere, EuroWordNeteko Goi- eta Domeinu-ontologiari esker, informazio sintaktiko-semantikoa aberatsagoa du. Hau da, WordNeten *synset* batek bere tasun sintaktiko-semantikoak hierarkiatik jasotzen ditu; EuroWordNetek *synset* hauek guztiak ditu, eta gainera Goi- eta Domeinu-ontologiatik datozkionak.

Top	
1st Order Entity	***2nd Order Entity***
Origin Natural Living Plant Human Creature Animal Artifact **Form** Substance Solid Liquid Gas Object **Composition** Part Group **Function** Vehicle Representation Money Representation Language Representation Image Representation Sofware Place Occupation Instrument Garment Furniture Covering Container Comestible Building	**Situation Type** Dynamic Bounded Event Unbounded Event Static Property Relation **Situation Component** Cause Agentive Phenomenal Stimulating Communication Condition Existence Experience Location Manner Mental Modal Physical Possession Purpose Quantity Social Time Usage
3rd Order Entity	

IV.1 Taula: EuroWordNeteko Goi-ontologia.

IV.1.2. atalean run aditza hartu dugu adibide gisa, WordNeten dagokion *unique beginner*rak ({travel, go, move, locomote} *synset*ak) mugimendu tasuna ematen diola ikusteko. EuroWordNeten run *synset* honek berak, tasun hori izango du (IV.2 irudian *motion*), baina horretaz gain, interfazean bertan (ikus IV.2 irudia) Goi-ontologiako *dynamic* eta *location* tasunak ere ikusten ditugu[18].



IV.2 Irudia: Run aditzaren *synset* bat eta bere hiperonimoak EuroWordNeteko interfazean.

IV.2 irudian EuroWordNeteko *synset*ek interfazean duten itxura ikus dezakegu, eta bertan gorriz dauden *Dynamic* eta *Location* dira Goi-ontologiako markak. Nahiz eta Goi-ontologiako tasunak run aditzaren *synset*ean bertan ez egon, bere hiperonimoetatik jasotzen ditu. EuroWordNeten 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 bar-

[18] Aditz honek Goi-ontologiako bi adabegietan du hastapena.

nean erabiltzeko, batik bat, informazio-erauzketa elebakar eta elebidunerako (Cuypers *et al.*, 1997; Gilarranz *et al.*, 1996; Vossen, 1997). Arrazoi horregatik, gaur egun, hainbat wordnet berri sortzen ari dira (katalana, portugesa, grekoa, suediarra, errumaniarra, bulgariarra, norvegiarra, lituaniarra, errusiarra...), EuroWordNeten ezagutza-basean oinarrituta. IXA taldean ere, tesi honetan arrazoitutakoari jarraiki, euskararako wordneta garatzen hasi gara (Agirre *et al.*, 2002). EuroWordNet kontsultarako interfazea publikoa da[19].

IV.3 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 EuroWordNeten eredua jarraitzen du.

MCRk bost hizkuntzetako wordnetekin egiten du lan: euskara, katalana, ingelesa (Princetoneko WordNetaren 1.5, 1.6, 1.7 eta 1.7.1 bertsioekin), italiera eta gaztelania. MCR bost hizkuntza horien izen, aditz, adjektibo eta adberbioen adieren inbentarioa da, eta EuroWordNeten ereduari jarraiki, hizkuntza guztiak lotuta daude. Horregatik, hizkuntza bateko *synset* batekin beste hizkuntzetakoa ere ikusgarri dago.

MCR EuroWordNeten bertsio aurreratuagoa da. Hortaz, EuroWordNeten gisa, MRCn ILIak (kasu honetan WordNet 1.6n oinarritutakoa), Goi-ontologia eta Domeinu-ontologiak erabiltzen ditu. MCR WordNet eta EuroWordNeten informazioaz baliatzen da, eta honetaz gain, informazio berria dakar:

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

EuroWordNeteko domeinuak ugaritu eta orraztu dituzte[20], hierarkian egon zitezkeen irregulartasunak gainditzeko. Bestalde, entitate edo izen bereziei

[19] `http://ixa2.si.ehu.es/mcr/wei.html` (2007-07-02an atzitua) web orrian dago eskuragarri.

[20] EuroWordNeteko hainbat domeinu gehiago zehaztu dituzte, "azpidomeinuak" sortuaz. Esate baterako, jokatu aditzak kirol adiera duenean, EuroWordNeteko *free time* domeinua, domeinuaren barruko *sport* azpidomeinuarekin zehaztu dute.

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

- ***The Suggested Upper Merged Ontology*:**

The Suggested Upper Merged Ontology (SUMO) (Niles eta Pease, 2001) *Terminology Corporation*en sortutako goi-ontologia da, *IEEE Standard Upper Ontology Working* taldean abiapuntu gisa erabiltzen dutena. SUMO, ontologia ezberdinen bilkuraren emaitza da — Sowa-ren (2000) goi-ontologia, Allen-en (1984) denbora-axiomak, Guarino-ren *mereotopologia formala* (Guarino, 1997; Borgo *et al.*, 1996), WordNet 1.6...—, eta termino orokorren definizioak jasotzen dira.

MCRn, oraingoz, SUMOko hiperonimia erlazioak eta etiketak bakarrik daude.

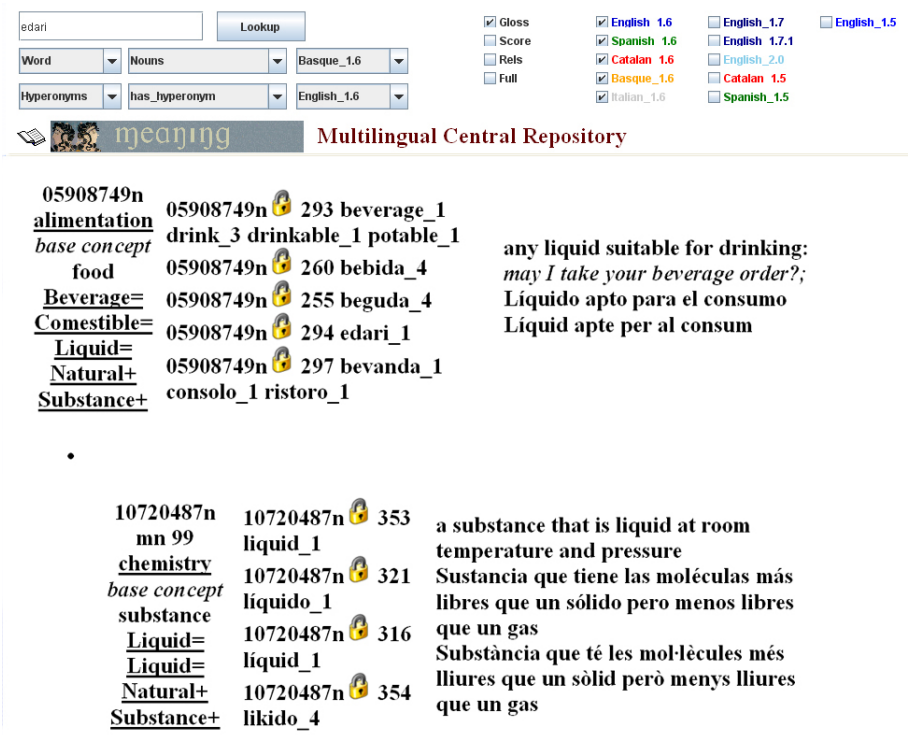
- **Hautapen-murriztapenak:**

MCR ezagutza-baseak aditzen hautapen-murriztapenak kontsultatzeko aukera ematen du *Role* erlazio semantikoa erabilita. Zazpi *Role* mota daude: *agentea* (*Role agent*), *norabidea* (*Role direction*), *baliabidea* (*Role instrument*), *kokalekua* (*Role location*), *gaia* (*Role patient*), *abiapuntua* (*Role source location*) eta *helmuga* (*Role target direction*).

Hala ere, nahiz eta interfazeak hautapen-murriztapenak jasotzeko aukera izan, *Role* harreman semantiko hauek hutsik daude; hots, oraindik ez da informazio hau eskuratu eta interfazean txertatu. Dena den, *synset*en arteko hautapen-murriztapenak eskuratzeko, dagoeneko saiakera batzuk egin dira: Carroll *et al.* (2003) eta tesi-txosten honen VII. kapituluan dakarkiguna. Bi lan hauetan hautapen-murriztapenen azterketa automatikoa egin da; hau da, teknika konputazional desberdinak erabiliaz zenbait corpusetatik (*British National Corpus* eta SemCorretik, hain zuzen ere) aditzen hautapen-murriztapen batzuk eskuratu eta ebaluatu dira. Eskuratze automatikorako baliabide eta teknika konputazional ezberdinak baliatzen dira, konbinazio ezberdinen emaitzak alderatzeko. Hala, emaitzarik onenak ematen dituen teknika-baliabideen konbinazioa definitu ondoren, hautapen-murriztapenen eskuratzea masiboa egingo da, gerora, MCRn txeratzeko.

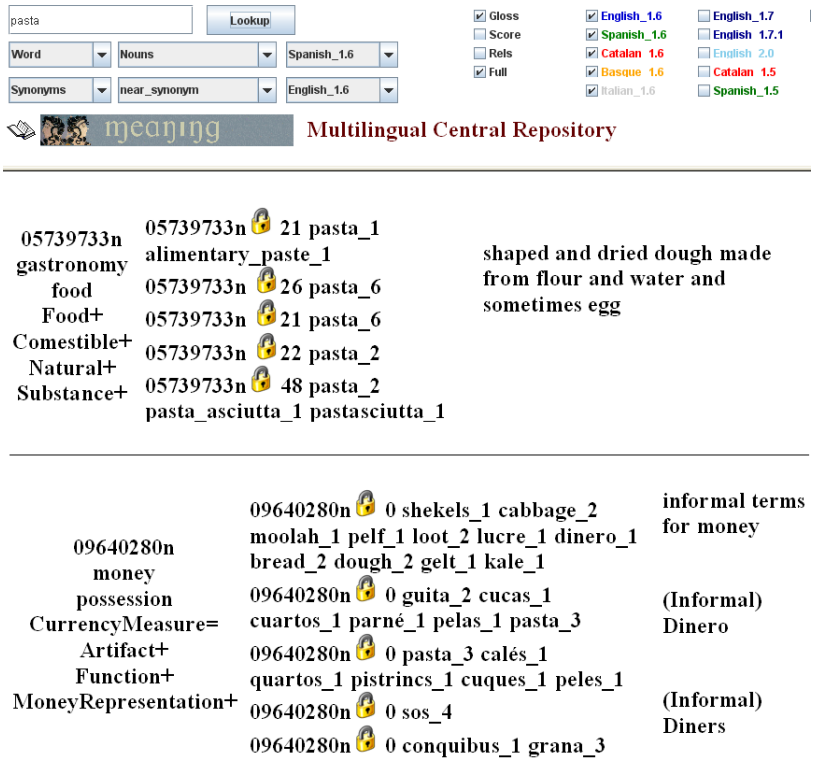
Hala, corpusetako datuetan oinarrituz, *Role* erlazio semantikoen bitartez aditz batekin ager daitezkeen ale lexikoak eta har ditzaketen rol tematikoak bereizteko gai dira. Ondorioz, MCRn aditzaren rol tematikoen berri ematen duen erlazio semantikoa genuke.

MCRn, ale lexikalak kategoriaka antolatuta daudenez (WordNet eta EuroWordNeten bezala) *Role* erlazioak inplizituki azpikategorizazioaren berri ere eman dezake. Esate baterako, eta IV.3 irudian adierazten den bezala, *Role patient* erlazioaren bidez jakin genezake edari izena edan aditzaren hautapen-murriztapena dela[21], izena dela bere kategoria eta *gaia* bere rol tematikoa. IV.3 irudiak *Role patient* erlazioa MCRn nola adieraziko litzatekeen erakusten du.



IV.3 Irudia: edari izenari dagokion *Role patient* erlazioa MCR interfazean.

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



IV.4 Irudia: Gaztelaniako pasta izenaren bi *synset* MCR interfazean.

MCRren kontsultarako interfazea publikoa da[22]. IV.4 irudian, MCRko *synset*ek duten itxura ikus dezakegu. EuroWordNeteko interfazearen oso antzekoa izan arren, interfaze hau informazio gehiagorekin aberastu da (Goi-ontologia, Domeinu-ontologia, SUMO, etab.). Kasu honetan, gaztelaniako pasta izenaren bi *synset* ditugu: bata 'jaki' adierari dagokiona (shaped and dried dough made from flour and water and sometimes egg glosaduna), eta bestea 'diru' adierari dagokiona (informal terms for money). Kontzeptu hauek guztiak ingelesez, katalanez, euskaraz eta italianoz ere ikus daitezke. Goi-ontologia, Domeinu-ontologia eta SUMOk *synset* hauen adiera ezberdintasuna hobeto ulertzen laguntzen dute. EuroWordNeten bezala, interfazearen ezkerretara eta gorriz Goi-ontologiako ezaugarriak adierazten dira: *Comestible, Natural* eta *Substance* 'jaki'ari dagokion *synset*arentzat; eta

[22] `http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl` (2007-07-02an atzitua).

Artifact, Function eta *Money Representation* 'diru'ari dagokion *synset*arentzat. Aipatu dugun bezala, EuroWordNeten tasun hauek ez dituzte *synset*ez *synset* adierazten, hierarkiari esker herentziaz jaso baitaitezke. MCRn, ordea, tasun hauek *synset* guztietan ikus daitezke. Horrela, informazio hori jakiteko ez dago hiperonimoetara jo beharrik. Lila kolorea baliatuz, SUMOren tasunak azaltzen dira: *Food*, 'jaki'ari dagokion *synset*arentzat; eta *Currency Measure*, 'diru'ari dagokionarentzat. Beltzez, Domeinu-ontologiari dagozkion tasunak adierazten dira: *gastronomy*, 'jaki'ari dagokion *synset*arentzat, eta *money* 'diru'ari dagokionarentzat. Adibide honetan ez dago hautapen-murriztapenei buruzko informaziorik, baina hauen berri IV.3 irudian eman dugu.

Horrela bada, WordNet eta EuroWordNet ezagutza-baseen ildotik jarraituz, MCRk erakutsi du hasieran egitasmo semantiko eta psikolinguistiko soilekin burutu zen ezagutza-basea baliagarria izan daitekeela informazio sintaktiko-semantikoa jasotzeko ere. Proiektu honen hurrengo urratsetan MCR informazio sintaktiko-semantiko gehiagorekin (azpikategorizazioa, erlazio semantiko konplexuagoak diatesi-alternantziak, Dorren ELKak (1997), eta abar) osatzeko asmoa dago.

IV.4 Laburbilduz

Kapitulu honetan WordNet ereduaren azterketa sakonago bat aurkeztu dugu. WordNeten ardatza *synset*a eta hiperonimia-hiponimia harremana dela azaltzeaz gain, eredu honek dituen beste harreman semantiko eta ezaugarri batzuk ere aipatu ditugu. EuroWordNet eta MCR WordNeten hedapen eleanitzak izaki, eredu batetik bestera zer aberasketa egon diren deskribatu dugu.

V. KAPITULUA

Euskal WordNeten eraikuntzarako metodologia

IXA taldearen beharretara gehien egokitzen den EBL formalismoa *WordNet, EuroWordNet* eta *The Multilingual Central Repository*ren (MCR) ildotik sortutako euskal EBLa ***Euskal WordNet*** deitu dugu.

Kapitulu honetan, Euskal WordNeten garapenean hartutako erabaki metodologikoak deskribatuko ditugu, eta, erabaki hauen arabera, Euskal WordNeten garapenak izandako urratsak ere azalduko ditugu.

Lehenik eta behin, Euskal WordNet nola garatu behar zen erabaki behar genuen. Izan ere, nahiz eta WordNeten egitura eta oinarriak izan, hainbat ikuspegi eta metodologia erabil zitezkeen garapenerako:

- WordNeten hierarkia jarraitzea eta bertako *synset*ei zuzenean esleitzea euskarako ordainak.

- Guk geuk sortzea euskarako adieren inbentarioa eta hierarkia, eta gero *Inter-Lingual-Index*ari (ILIari) (ikus IV.2 atala) lotzea.

Bi aukera hauek aztertu ditugu, eta lehenengoaren alde egin dugu. Erabaki horren berri V.1 atalean emango dugu.

Bestetik, Euskal WordNet garatzeko diseinatu dugun metodologiak irizpide batzuk behar zituen. Alde batetik, eta aurrerago aipatu izan dugun bezala (ikus III.1), Euskal WordNet estaldura handikoa izan behar zuen, hots, lexiko zabalekoa eta ikuspegi orokorrekoa. Bestetik, kalitate onekoa. Bi irizpide

hauen arabera, Euskal WordNeten garapena aldi eta modu ezberdinetan burututako prozesua izan da: aberasketa automatikoa eta eskuzkoa konbinatuz; eta hainbat hiztegi elebakar eta elebidunenez baliatuz eta corpusetik jasotako informazioa baliatuz.

Metodologia hauek izenen aberasketarako erabili dira, Euskal WordNeten garapenaren lehenengo urratsak izenetan oinarritu baitziren. V.2 atalean sakonduko dugu fase hauetako bakoitzean. Izenen aberasketarekin amaitu ondoren[1], orain aditzen aberasketarekin hasteko garaia da. Hala ere, aditzek duten informazio aberatsa dela-eta (azpikategorizazioa, hautapen-murriztapenak...), hauen orrazketarako eta aberasketarako hainbat metodologia aztertu ditugu.

V.3 atalean, batetik, aditzen lanketak arreta berezia zergatik merezi duen azalduko dugu; eta bestetik, aditzak garatzeko zer metodologia probatu ditugun deskribatuko dugu, hauetatik zein aukeratu dugun ondorioztatuko dugularik.

Beraz, kapitulu honetan, Euskal WordNeten hastapenaren nondik norakoak azalduko ditugu. Azken urteotan izenen garapenean izandako faseak zehazki deskribatuko ditugu, eta oraindik hasi gabe dugun aditzen garapenerako landu ditugun metodologia ezberdinak aurkeztuko ditugu.

Azkenik, esan beharra dago, adjektiboen eta adberbioen lanketa tesi-lan honen etorkizunerako lan bezala utzi dela.

V.1 Diseinua eta metodologia

Euskarako EBLa egiteko oinarrituko garen eredua erabaki ondoren, eta EBL hori —aztertutako EBL gehienak bezala— ingeleserako sortuta dagoela ikusita, beste erabaki berri baten aurrean gaude: euskaraz dauden corpusetatik eta hiztegietatik abiatuta euskarako wordneta sortzea, ala euskararako EBLa egitea, erdaretarako egin diren wordnetez baliatuta.

Lehenengo aukeran, sortu beharreko adierak eta hierarkiak WordNeteko hierarkiekiko independenteak izango lirateke, eta horrek adieren inbentarioa eta hierarki bera gure irizpideen arabera garatzeko eta kontrolatzeko askatasun guztia emango liguke. Baina, bestalde, hurbilpen horrek

[1] Lan lexikografikoen antzera, EBLen aberasketa-lanak ez dira inoiz amaitzen. Hala ere, egindako orrazketa guztien ondoren, Euskal WordNetek euskarako izen gehienak jasotzen dituela esan dezakegu.

lan lexikografiko handia eskatuko luke, eta, horrez gain, hizkuntzen arteko adieren loturak adierazteko ILIra lotzeko bideak sortu beharko lirateke. Vossen-ek (1999) *merge approach* deitu du metodologia hau.

Bigarren aukeran, MCRko hizkuntza bateko wordneta abiapuntu gisa hartuz gero, nahiz eta guk ez kontrolatu adieren sorkuntza eta antolamendu hierarkikoa, lan lexikografikoa beste aukeran baino askoz ere txikiagoa da. Izan ere, askotan, lana euskarako hitzak ILIari lotzera mugatzen da; hots, euskarako ordainak zuzenean *synset* egokiei esleitzea litzateke egin beharreko lana. Honezaz gain, MCRko ILIari esker, euskarako ordainak ingeleseko kontzeptuei lotuta geratuko lirateke. Gainera, modu honetan hizkuntzen arteko adieren loturak egiteko bidea ere ematen zaigu. Vossenek (1999) *expand approach* bezala izendatu du metodologia hau.

Tesi-lan honetan, bigarren aukeraren alde egin dugu; hau da, Euskal WordNeten garapena MCRn oinarritu dugu, eta bertako ingeleseko kontzeptuak abiapuntutzat harturik, euskarako ordainak lotzen ditugu, eta ez dauden euskarako kontzeptuak (sagardoa, trikitixa eta abar) txertatzen ditugu[2]. Hala ere, IXA taldean lehenengo aukerarekin saiakerak egin dira (Agirre *et al.*, 2003c; Lersundi, 2005), etorkizunean bi hurbilpenen abantailak baliatzeko asmoa baitago. Honetaz gain, beste euskarako hiztegietatik erauzitako hierarkiak eta erlazio semantikoak ere txertatuko zirela erabaki zen, eta, egun, egin dira horren inguruko hainbat saiakera IXA taldean (Agirre *et al.*, 2003c), baina hori ez da tesi honen eremuan sartuko.

Hizkuntza askotako wordnetak egonik (katalanez, gaztelaniaz, frantsesez, ingelesez, italieraz, estonieraz, nederlanderaz, txekieraz, estonieraz...), Euskal WordNet sortzeko hauetako edozeinetan oinarritu gintezkeen. Ulermenari begira, lan lexikografiko urriagoa litzateke *synset*en adierak gaztelaniaz ulertzea ingelesez baino. Bestalde, gaztelania-euskarako hiztegi elebidun gehiago daude ingelesa-euskarakoak baino. Baina ezin da ahaztu, MCRk *ILI-record*ak WordNet 1.6tik hartu dituela, eta hizkuntzen arteko bateragarritasunari begira, WordNet 1.6eko hierarkian oinarritu zirela proiektuan parte hartutako hizkuntza guztiak. Arrazoi hauengatik, Euskal WordNet Princetoneko WordNet 1.6 bertsioaren gainean garatzea erabaki genuen, WordNeteko ingeleseko kontzeptuak abiapuntutzat hartuz, euskarako ordainak hauei lotuz, eta ez dauden euskarako kontzeptuak txertatuz.

Euskal WordNeten eraikuntzan metodologia aldatuz joan da. Metodo-

[2]MCRn ez dauden euskarako kontzeptuak (trikitixa, ikastola eta abar), momentuz, zerrendatzen ditugu etorkizunean lantzeko.

logian egondako aldaketa hauek estaldura eta kalitatea uztartzearren izan dira. Estalduraz hitz egiterakoan, kontzeptu, sarrera lexikal, kategoria, hitz-adiera eta sinonimoen kopuruaz ari gara. Kalitateaz hitz egiterakoan, *synset* eta *variant*en zuzentasunari, osotasunari eta egokitasunari buruz ari gara. Laburbilduz:

- **Zuzentasuna:** *synset*ean dauden *variant* eta hitz-adierak zuzenak izatea.
- **Osotasuna:** *synset*ari dagozkion *variant* eta hitz-adiera guztiak egotea.
- **Egokitasuna:** *synset*ean dauden *variant* eta hitz-adiera guztiek espezifikotasun maila bera izatea.

Badago faktore bat batzuetan eragina izan duena estaldurari edo kalitateari garrantzia emateko garaian: baliabide gutxiko eta abiadura handiko garapenaren beharra. Hau dela eta, hasieran estaldurari garrantzia eman genion eta kalitatea bermatzea bigarren urrats gisa definitu genuen.

Kategoriei begira, WordNeteko lau kategorietatik (izenak, aditzak, adjektiboak eta adberbioak) lehenengo izenak eta gero aditzak landuko genituela erabaki zen, hauek informazio lexiko oso garrantzitsua jasotzen dutelako, eta, ondorioz, LNPn gehien landu direnak direlako.

Hala, hartutako erabakiei jarraituz, Euskal WordNet eraikitzen joan gara. Jarraian bereizitako fase bakoitza sakonkiago aztertuko ditugu.

V.2 Izenen garapenerako urratsak

V.2.1 Estaldura helburu: garapen automatikoa eta oinarrizko kontzeptuak

Lehenengo urratsak oinarrizko Euskal WordNet eraikitzea izan zuen xede, eta, horregatik, estaldura izan genuen helburu nagusi. Hala, garapenaren lehenengo urratsean bi bide jorratu genituen:

- Oinarrizko kontzeptuei (*Base Concepts* izenekoei) euskarako ordainak eskuz lotu.

- Ingeleseko *synset*en euskal ordainak hiztegi elebidunak baliatuz —euskara-ingelesa Morris (1998); Aulestia eta White (1990)— automatikoki sortzea. Garapen automatikoa zer teknika informatikoekin egin zen eta zer nolako kalitatea lortu zen ikusteko, jo bedi Agirre *et al.*-era (2002).

V.2.2 Kalitatea helburu: eskuzko orrazketa eta corpus baten etiketatzea

Hurrengo urratsetan, kalitateari eman zitzaion garrantzi handiago. Kalitatea lantzeko ere metodologia ezberdinak erabili dira. Hasieran, automatikoki sortu ziren euskarako *synset* horien eskuzko orrazketa egin genuen hizkuntzalariok. Gero, beste orrazketa bat egin genuen *Elhuyar Hiztegi Txikia* (Elhuyar, 1998) hiztegiko adierak Euskal WordNeten zeudela ziurtatzeko eta *synset*ean zeuden ordainak egokiak zirela egiaztatzeko. Gaur egun, Euskal WordNeteko *synset*ekin eskuz etiketatzen (desanbiguatzen) ari garen euskarako corpus baten (*EuSemcor*) informazioa baliatzen ari gara EBLa orrazteko[3].

V.2.2.1 Kontzeptuz kontzeptuko eskuzko orrazketa

Orrazketa honetan hizkuntzalariok, alde batetik, *synset*aren euskarako ordaina egokia zen ala ez berrikusten genuen; bestetik, *synset*ean euskarako beste ordainik behar zen egiaztatzen genuen.

Prozesu hau guztia erraztearren hurrengo pausoak jarraitu ziren:

- **Hizkuntzalariontzat lan egiteko erabilerraza den interfazea sortu:**

EBLari lotutako interfaze bat sortu zen (Benítez *et al.*, 1998), batetik, hizkuntzalarioi adierazpide intuitiboa eskaintzeko eta bestetik, aldi berean hizkuntzalari batek baino gehiagok lan egin ahal izan zezan.

[3] A eranskinean Euskal WordNeteko *synset*ak editatzeko jarraitzen ditugun irizpideak datoz.

- **Orraztu beharreko *synset*ak tratatzeko ordena antolatu:**

*Synset*en orrazketa nolabait antolatu beharra zegoen. Nondik hasi behar genuen hizkuntzalariok *synset*ak orrazten? Aukera ugari zeuden: hierarkiak goitik behera jarraituta edota alderantziz (behetik gora), oinarrizko kontzeptuak lehenengo eta ondoren bestelakoak, ingeleseko edo euskarako ordainaren arabera, eta abar. Gure ustetan, orrazketaren abiadura azkartuko zen, baldin eta hizkuntzalariak antzeko *synset*ak jarraian berrikusten bazituen. Hau da, berrikusitako *synset* baten ondoren, berrikusi beharreko hurrengo *synset*a klase berekoa bazen, prozesua azkartuko litzatekeela iruditzen zitzaigun. Hala, *synset*en orrazketa hiperonimo kateak jarraituta antolatu zen: hierarkia bakoitzeko *synset* altuenetatik —orokorrenetatik— hasi (*unique beginner* deritzona) eta azkeneko hiponimoraino. Orrazketa mota hau ahalbidetzeko, interfazean aparteko botoi bat gehitu zen, eta hau sakatuz gero, hiperonimo katean behera jarraituta, orraztu gabe zegoen hurrengo *synset*a agertzen zen interfazean.

Orrazketarekin hasi ahala, interfazean beste botoi batzuk gehitu ziren, interfazea hizkuntzalarion beharretara egokitzeko. Esate baterako, hasiera batean, hizkuntzalariok zalantzazko *synset*ei buruzko oharrak eskuz idazten genituen. Gerora, interfazean botoi bat txertatu zen zalantzazko *synset*ak markatzeko. Era honetara, errazagoa zen zalantzazko *synset*ak berrikusteko garaian hauek aurkitzea. Botoi hauen guztien berri A eranskinetan ematen da.

	Izenak	***Synset***	***Variant***	***Variant synseteko***	***Lema***	***Variant lemako***
EusWN 0.1	**BC eskuz**	228	-	-	-	-
	auto.	27.641	291.011	10,5	46.164	6,3
	Kontz. eskuz	23.486	41.107	1,7	22.166	1,8
WN 1.6	**eskuz**	66.025	116.364	1,7	95.135	1,2

V.1 Taula: Euskal WordNeteko izenen kopuruak WordNet 1.6koekin alderatuta, oinarrizko kontzeptuak, sorkuntza automatikoa eta kontzeptuz kontzeptuko orrazketak egin ondoren.

Kontzeptuz kontzeptuko orrazketarekin amaitzean, aurreko urratsetako emaitzen ebaluazioa (V.2.1 atalean aipatutakoena) egin genuen. V.1 taulan, orain arte aipatutako garapen-urratsetan —garapen automatikoa (*auto.* taulan) eta kontzeptuz kontzeptuko eskuzko orrazketa (*Kontz. eskuz* taulan) ize-

netarako lortu diren kopuruak daude: *synset*ak, *variant*ak, lemak, *synset*eko dauden *variant*en batezbestekoa, eta lemako dauden *varian*en batezbestekoa. Hauekin batera, WordNet 1.6 bertsioaren kopuruak ere aurkezten dira (Euskal WordNet garatzen hasi ginenean bertsio honekin hasi baikinen).

Alderdi kuantitatiboari begira, kontzeptuz kontzeptuko orrazketaren ondoren Euskal WordNet 0.1 bertsioan dauden izenen *synset*en kopurua (ikus *EusWN 0.1 Kontz. eskuz* errenkada V.1 taulan) ez da WordNet 1.6 bertsioan daudenen kopuruaren erdira iristen (ikus *WN 1.6* errenkada). Kontzeptuen estaldura % 38koa izan zen, eta lemena, 22.166 lemekin, % 25ekoa.

Garapen automatikoan *synset* bakoitzeko dauden *variant*en eta lemako dauden adieren batezbestekoa oso handia da (ikus V.1 taulako *EusWN 0.1 auto.* errenkadan: 10,5 *variant synset*eko eta 6,3 adiera lemako). Hau sorkuntza automatikoan arrunta bada ere, honen beste arrazoi bat hauxe izan daiteke: garapen automatikorako erabilitako hiztegietako batek (Aulestia eta White, 1990) aldaera ortografiko eta dialektal ugari jasotzen ditu, asko eta asko azken urteotan Euskaltzaindiak onartutako arauekin bat ez datozenak. Kontzeptuz kontzeptuko orrazketaren ondoren, batezbesteko hauek 1,7 eta 1,8ra jaitsi dira (ikus *EusWN 0.1 Kontz. eskuz* errenkada), eta WordNetekoekin ia berdindu (ikus *WN 1.6* errenkadan: 1,7 eta 1,2).

Bestalde, aipatu beharra dago eskuzko orrazketaren ondoren *synset*, lema eta *variant* kopuruak jaitsi direla nabarmen, eta bereziki azken hauena. Honek adierazten du garapen automatikoan, estaldura handia lortu arren, forma desegoki asko sartzen dela kalitatearen kaltetan. Kapitulu honetan zehar aipatu izan dugun bezala, eskuzko orrazketarekin arazo hau konpondu nahi izan dugu.

Hala ere, hobetu beharreko zenbait puntu antzeman genituen:

- Nahiz eta Euskal WordNeten hitz bat egon, horrek ez zuen ziurtatzen hitz honen adiera guztiak EBLan zeudenik. Kontzeptuz kontzeptuko orrazketa amaitzean, Euskal WordNeten ez zeuden hitz-adieren kopurua % 20koa zen. Kalkulu hori egiteko *Euskal Hiztegia* (Sarasola, 1996) eta Euskal WordNeten arteko konparaketa bat egin zen (Agirre *et al.*, 2002).

- *Synset* barruko *variant*en espezifikotasun-maila ez zen guztiz egokia. Askotan, euskarako *variant*ak hierarkiaren maila desegokian zeuden. Adibidez, religious kontzeptuak (a member of a religious order glosaduna), euskaraz erlijioso, serora eta lekaide ordainak zituen. Bai serorak eta

bai lekaidek adiera hori izan dezakete (erlijio-talde baten kide baitira), baina erlijioso mota bat direnez —bata gizonezko erlijiosoa eta bestea emakumezkoa—, hierarkian ez dagokie leku hori. Aitzitik, erlijioso *synset*aren hiponimo bana beharko lukete.

- Euskal WordNeteko hitzen adieren espezifikotasun-maila erreferentzia gisa erabilitako hiztegiena baino finagoa da. Esate baterako, *Hiztegi Modernoa*k (Elhuyar, 2000) lantegi izenarentzat hiru adiera ematen ditu:

(a) Eskuzko lanen bat egiten den tokia, tailerra; *Zurgin-lantegia*.

(b) Fabrika; *Hegazkin-lantegi batean*.

(c) Lana, egitekoa, lanbidea; *Lantegi gogorra baso-mutilarena*.

Eta Euskal WordNeten hitz horrek sei *synset* ditu:

(a) Industria-lana egiteko eraikina; *Beraiek autoak produzitzeko lantegi bat eraiki zuten*.

(b) Eskulanak edo fabrikazioa egiten den eraikin txikia; *Osaba bere lantegian espartigintzan ari da*.

(c) Jarduera profesionala egiten den tokia; *Bere lantegira eraman behar duzu mezua*.

(d) Talde txiki batentzako ikastaro labur eta trinkoa; arazo bat konpontzera bideraturik; *Gorputz adierazpeneko lantegi*.

(e) Ahalegina eginiko lanaren parteetako bat. *Haur eta gazte literatura zituen beste zenbait lantegi*.

(f) Pertsona baten bizitzako aktibitate nagusia, zeinek dirua irabazteko aukera ematen duen; *Aurrez ezagutzen zuten lantegiari lotu ziren: ardiari, alegia*.

Orrazketa honen emaitzak eta ondorioak ikusita, ondoren azalduko dugun eskuzko orrazketari ekin genion.

V.2.2.2 Hitzez hitzeko eskuzko orrazketa

Hitzak *Elhuyar Hiztegi Txikia*n (Elhuyar, 1998) zituen adiera guztiak Euskal WordNeten zituela ziurtatzea zen urrats honen helburua, eta, era berean, *synset*ean zeuden ordainak egokiak zirela egiaztatzea. Azken finean, aurreko urratseko lan berbera egiten genuen, baina beste ikuspegi osagarri batetik begiratuz.

Garapen-urrats honetarako, lehenengo *Elhuyar Hiztegi Txiki*ko izenen sarrerak corpuseko (*Euskaldunon Egunkaria*[4] eta *XX. mendeko euskararen corpus estatistikoa*[5]) maiztasunaren arabera ordenatu ziren: maiztasun handienetik txikienera. Hala, euskaraz gehien erabiltzen ziren izenak EBLan landuta zeudela ziurtatzen genuen. Ondoren, zerrendako izen bakoitzarekin hurrengoa izan zen hizkuntzalarion lana:

- **Adieren estaldura ziurtatzea:** hitzaren adiera arruntenak Euskal WordNeten sartu.
- ***Variant*en estaldura ziurtatzea:** *Sinonimoen Hiztegia* baliatuz (UZEI, 1999), *synset*ean *variant*/sinonimo guztiak daudela ziurtatu.
- **Hitzaren adieren zuzentasuna bermatzea**: Euskal WordNeten dauden adiera guztiak zuzenak direla ziurtatzea.
- **Hitzaren adieren estaldura bermatzea:** hitzaren adiera guztiak Euskal WordNeten daudela ziurtatzea.
- ***Synset* barruko *variant*en espezifikotasun-maila egokia ziurtatzea**: euskarako *variant*ak hierarkiaren maila egokian egon daitezen, honen hiperonimo eta hiponimoei begiratzea. Hala, religious kontzeptuarekin aipatutako arazo mota hori eragozten da.
- **Hitzen adieren espezifikotasun-maila:** lantegi adibidearekin ikusi dugun bezala, askotan Euskal WordNeteko hitzen adieren espezifikotasun-maila erreferentzia gisa erabilitako hiztegiena baino finagoa da. Hiztegietan ez dauden adiera edo *synset* horiei euskarako ordaina sartuko zaie, baldin eta egiaztatzen badugu adiera horiek euskaraz ezagunak direla, eta LNPko atazetarako beharrezkoak direla. Adibidez, Euskal

[4] `http://www.egunero.info` (2007-07-02an atzitua).
[5] `http://www.euskaracorpusa.net` (2007-07-02an atzitua).

WordNeteko lantegiren (c) eta (d) adierak ('jarduera profesionala egiten den tokia' eta 'talde txiki batentzako ikastaro labur eta trinkoa; arazo bat konpontzera bideraturik') ez daude *Hiztegi Modernoa*n, ezta *Elhuyar Hiztegi Txikia*n ere. Hala ere, adiera hauen erabilera egiaztatzen dugu corpusetan —hala nola, *XX. mendeko euskararen corpus estatistiko*an eta *Ereduzko Prosa Gaur* corpusean[6]— eta beste hiztegietan —*Elhuyar Hiztegia: euskara-gaztelania*[7] (Elhuyar, 1996) hiztegian, eta *Euskal Hiztegia*n, adibidez. Kasu honetan, bi adiera hauek *Elhuyar Hiztegi* elebidunean agertzen direnez, zuzentzat jo ditugu eta Euskal WordNet txertatu ditugu.

Orrazketa honen erdibidean ginela, eta WordNet eta LNP komunitatean corpus desanbiguatuak hartzen ari ziren indarra ikusita (Fellbaum *et al.*, 2001; Palmer eta Kingsbury, 2003; Marcus *et al.*, 1993), hitzez hitzeko eskuzko orrazketa metodologia corpus baten etiketatze semantikoarekin osatzea erabaki genuen. Erabaki hau IXA taldean jorratzen ari den lan-ildo batekin bat etortzearren hartu zen. Izan ere, IXA taldean maila linguistiko desberdinetan etiketatuko den erreferentziazko corpusa garatzen ari gara (Aduriz *et al.*, 2006): *Euskararen Prozesamendurako Erreferentziazko Corpusa* (EPEC). Corpus hau 300.000 hitzekoa da; heren bat *XX. mendeko euskararen corpus estatistiko* hartua dago, eta beste guztia *Euskaldunon Egunkaria* corpusetik. EPEC corpusen morfosintaxia, sintaxia, Euskal WordNeteko adierak eta PropBankeko rolak (Agirre *et al.*, 2006d) etiketatuko dira eskuz.

Lan-ildo honetatik abiatuta, Euskal WordNeten ondorengo garapen-fase berrian hasi ginen: corpus baten etiketatze semantikoan.

V.2.2.3 Corpus baten etiketatze semantikoa

Orrazketa eta etiketatzea uztartuz, corpuseko informazioa erabil dezakegu Euskal WordNet garatzeko eta aberasteko. Aldi berean, eskuz etiketatutako euskarako corpus semantikoa sortzen ari gara: EuSemcor (Agirre *et al.*, 2006a). Alegia, EPEC corpusa maila semantikoan, Euskal WordNeteko *synset*ak erabilita, etiketatzen ari gara.

Beraz, lan honen helburua 300.000 hitzeko corpusa etiketatzea da, eta hauxe da gaur egun egiten ari garena. Izenak, adjektiboak eta aditzak etiketatu nahi dira. Aldi berean, eta corpusetik lortzen den informazioan oina-

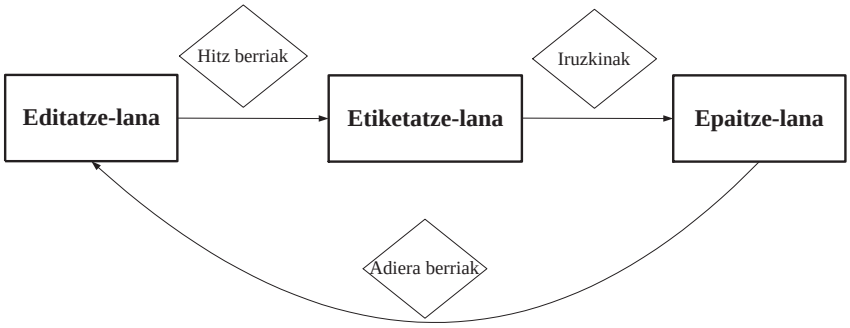
[6] `http://www.ehu.es/euskara-orria/euskara/ereduzkoa` (2007-07-02an atzitua).
[7] `http://www1.euskadi.net/hizt_el/indice_e.htm` (2007-07-02an atzitua).

rrituz, Euskal WordNeteko *synset*ak orraztuko dira; hau da, 300.000 hitzeko corpusaren etiketatze semantikoa amaitu ondoren, Euskal WordNetek corpusean agertu diren adiera horiek guztiak izan beharko ditu.

Lan-taldea bost hizkuntzalariz osatua dago: gainbegirale bat, **editore** bat, bi **etiketatzaile** eta **epaile** bat. Editorea Euskal WordNet *editatzen* duena da, hots, Euskal WordNeteko *synset*ak lantzen dituena. Etiketatzaileek etiketatu beharreko hitzaren agerpen berak corpus berean etiketatzen dute (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 landu beharreko hitzak aukeratzen ditu, eta hitz hauen Euskal WordNeteko *synset*ak lantzen eta berrikusten ditu hitzez hitzeko metodologian oinarrituz (ikus V.2.2.2 atala). 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, hitzei dagozkien agerpenak etiketatzen hasiko dira etiketatzaileak. Etiketatze-lana amaitzean, *synset*en glosak[8] ere ingelesetik euskarara itzultzen dituzte. Lan hauek guztiak bukatu ondoren, editorearekin eta epailearekin egindako bilera batean etiketatzean izan dituzten gorabeherak azaltzen dituzte. Gero, epaileak programa informatiko baten laguntzarekin, bi etiketatzaileen lana erkatzen du, eta ezberdin etiketatuta dauden agerpen horiek ebazten ditu. Gainera, corpusean adiera berriren bat agertzen bada, horren berri ematen dio editoreari. Azkenik, editoreak corpusean agerturiko adiera berri horien egokitasuna aztertzen du hauek Euskal WordNeten sartzea erabaki baino lehen. V.1. irudian adierazten den bezala, metodologia ziklikoa da.

[8]III. kapituluan azaldu dugun bezala, *synset*aren adiera, normalean, glosa edo azalpen txiki baten bidez adierazten da. Glosa hauen itzulpenetarako jarraitzen diren irizpideak Agirre *et al.* lanean (2005b) daude.



V.1 Irudia: EuSemcorreko etiketatze semantikoaren metodologia.

Editoreak, etiketatzaileak eta epaileak metodologia honen urrats bakoitzean bere zeregina zehaztuta dute eskuliburu batean (Agirre *et al.*, 2005b).

V.2 taulan EuSemcorren kopuruak ageri dira. Alde batetik, dagoeneko landu diren izenen kopuruak, eta bestetik, corpuseko izen guztien kopuruak. Azken honetan (*guztira* zutabean) agertzen diren kopuruak, corpuseko izenak Euskal WordNetekin parekatuta lortu dira. Esate baterako, corpusean 1.810 izen monosemiko daudela diogunean, corpuseko 1.810 izenek Euskal WordNeten *synset* bakarra dutelako da, eta corpusean, izen hauek guztien agerpen kopurua 16.606 da. Hala ere, 1.810 izen hauetatik, 192 bakarrik, berrikusi dira Euskal WordNeten eta etiketatu EuSemcorren. Beraz, lantzeko dauden 1.618 izenak Euskal WordNeten berrikusi ondoren, baliteke izen horietatik batzuk polisemikoak izatea (Euskal WordNeten garapenaren aurreko faseren batean izen horri ez zuen adiera bat egokitu zitzaiolako[9]). Hala, *guztira* zutabeko hitzei dagozkien kopuruak ez dira guztiz zehatzak, baina corpusa etiketatzeko geratzen zaigun lanaren balioespena egiteko oso erabilgarria zaigu.

Bestalde, V.2 taulan ikus daitekeen bezala, 442 izen polisemiko etiketatu ditugu dagoeneko, eta, agerpen-kopuru handikoak direla egiaztatu dugu. Hauek corpusean 39.208 aldiz agertu dira (izen polisemiko guztien agerpenen % 56a). Monosemikoei dagokienez, oraingoz, 192 izen sailkatu dira *synset* bakarrekoak bezala, eta izen hauen agerpenak automatikoki etiketatu dira (izen monosemiko guztien agerpenenen % 45a). Orain arte, corpusean 83 izen agertu dira Euskal WordNeten ez daudenak, eta gehitu egin behar izan ditugu. 83 izen hauek corpuseko 487 agerpeni dagozkie. *Guztira* errenkadan,

[9] Adieren lanketari buruzko argibide gehiago A eranskinean.

	Eginak		***Guztira***	
	Hitz	**Agerpen**	**Hitz**	**Agerpen**
Polisemikoak	442	39.208	3.330	68.871
Monosemikoak	192	7.281	1.810	16.606
EusWNen ez daude	83	487	11.070	39.936
Guztira	717	46.976	16.210	125.413

V.2 Taula: EuSemcor: izenei dagozkien kopuruak.

deigarria da 16.210 izenetatik 11.070 ez egotea Euskal WordNeten. Honen arrazoia corpusean agertzen diren izen berezietan datza, eta horiek, oraingoz, ez ditugu Euskal WordNeten gehituko.

Izenen garapen-urratsekin amaitzeko, V.1 taula dakargu berriro, orain arte aipatutako garapen-urratsak —garapen automatikoa (*auto.* taulan), kontzeptuz kontzeptuko eskuzko orrazketa (*Kontz. eskuz* taulan) eta hitzez hitzeko orrazketa eta corpus baten etiketatze semantikoa (*Hitzez. eskuz* taulan)— erabilita Euskal WordNeteko egungo kopuruak aurkezteko (ikus V.3 taula: *synset*ak, *variant*ak, lemak, *synset*eko dauden *variant*en batezbestekoa eta lemako dauden *variant*en batezbestekoa). Hauekin batera, WordNet 1.6 bertsioaren kopuruak ere aurkezten dira.

Euskal WordNet 0.1 bertsioaren estaldurarekin erkatuz gero (23.486 *synset* eta 41.107 adiera), egungo Euskal WordNet 0.2 handitu den arren (28.943 *synset* eta 40.848 *variant*), oraindik WordNet 1.6n *synset* eta *variant* kopurua ia Euskal WordNeten bikoitza baino gehiago da (66.025 *synset* eta 116.364 *variant*).

Bestalde, kapitulu honetan zehar aipatu izan dugun bezala, eskuzko orrazketarekin kalitatearen alde egin nahi izan dugu. Baina kalitatearen alde eginez gero, oso mantso egiten dugu aurrera: astean hamabi *synset* editatzen ditugu batezbeste.

Synset eta lema bakoitzeko dauden *variant*en batezbestekoa antzekoa da euskarako eta ingeleseko eskuzko orrazketetan. Automatikoan, aldiz, desegokia diren *variant* asko sartzen dira. Hala, eskuzko orrazketak kalitate handiagokoak direla garbi ikusten da taula honetan, *variant synset*eko eta *variant* lemako zutabeei erreparatuz gero.

	Izenak	*Synset*	*Variant*	*Variant synseteko*	*Lema*	*Variant lemako*
EusWN 0.1	**BC eskuz**	228	-	-	-	-
	auto.	27.641	291.011	10,5	46.164	6,3
	Kontz. eskuz	23.486	41.107	1,7	22.166	1,8
EusWN 0.2	**Hitzez. eskuz**	28.943	40.848	1,4	23.137	1,7
WN 1.6	**eskuz**	66.025	116.364	1,7	95.135	1,2

V.3 Taula: Euskal WordNeteko izenen kopuruak WordNet 1.6koekin alderatuta, oinarrizko kontzeptuak, sorkuntza automatikoa, kontzeptuz kontzeptuko orrazketa eta hitzez hitzeko orrazketa egin ondoren.

V.3 Aditzen garapenerako urratsak

EuSemcorren maiztasun handieneko izenak lantzen joan ahala, aditzen aberasketari ere ekin zaio, baina neurri txikiagoan.

Esan daiteke aditza dela hizkuntzako kategoria lexiko eta sintaktiko garrantzitsuena. Esaldi gehienek aditz bat badute gutxienez, eta aditza da esaldia semantikoki eta sintaktikoki antolatzen duena. Aditzean zehazten dira: esaldian egon daitezkeen egitura sintaktiko posibleak (azpikategorizazio hertsia); argumentuak rol tematikoekin lotzean, esaldian adierazten diren ekintza edo egoeren adierak; hautapen-murriztapenak (aditz horrekin ager daitezkeen izen-klaseen ezaugarriak).

Hiztunok geure baitako lexikoian informazio sintaktiko eta semantiko hau guztia jasota dugunez, hau guztia aditzaren sarrera lexikalean gorde beharreko informazioa dela pentsatu izan da. LNPren ikuspegitik begiratuta, aditzekin batera datorren informazio hori guztia EBL batean jasota izanez gero, hainbat atazatan oso baliagarria izango litzateke.

Baina, nahiz eta aditzak informazio ugari eraman, informazio hori oso konplexua da, eta arrazoi horregatik da horren zaila aditza aztertzea eta bere informazioa adieraztea.

> "This syntactic and semantic information is generally thought to be part of the verb's lexical entry, that is to say, part of the information about the verb that is stored in a speaker's mental lexicon. Because of the complexity of this information, verbs are probably the lexical category that is most difficult to study." (Miller *et al.*, 1993, 40. or.)

III.2.3.2 atalean esan dugun bezala, WordNeteko aditzek informazio sintaktiko-semantiko mugatua dute:

> "WordNet was designed to model lexical memory rather than represent lexical knowledge, so it excludes much of a speaker's knowledge about both semantic and syntactic properties of verbs. There is no evidence that the syntacitc behavior of verbs [...] serves to organize lexical memory."
>
> (Miller *et al.*, 1993, 55. or.)

Hori dela eta, izenak lantzeko eta aditzak lantzeko jarraitutako urratsak desberdinak izan dira.

Estaldurari dagokionez, izenen oinarrizko kontzeptuekin (*Base Concept* delakoekin) batera, ingeleseko aditzen oinarrizko kontzeptuei ere euskarako ordainak eskuz lotu zitzaizkien. Izan ere, V.2 atalean esan dugun bezala, Euskal WordNeten eraikuntzaren lehenengo urratsetan oinarrizko estaldurari eman zitzaion garrantzia.

Kalitateari begira jarri ginenean, hainbat gauza zeuden kontuan hartzeko modukoak. Tesi-txosten honen hasieratik esan dugun bezala (III.1 atalean), euskarako EBLan ale lexikalen adierez gain, hauen informazio sintaktiko-semantikoa adierazita etortzea nahiko genuke. MCRn horrelako informazioa esplizitu egiten saiatzen badira ere, aditzen antolaketa eta hierarkia WordNeterako egindakoa da. Honela, aditzen lanketa masiboarekin hasi baino lehen, hauxe da egin dugun azterketa:

- Aditzak WordNeten landuta nola dauden ikustea: adiera-bereizketak eta hierarkiaren nondik norakoak.

- Euskarako aditzak MCRn txertatzeko erarik egokiena eta azkarrena aztertzea.

V.3.1 Aditzak WordNeten

Aditzen lanketarako, izenetan kontuan hartu ez zen baldintza bat guztiz beharrezkoa da: informazio sintaktiko-semantikoa (azpikategorizazioa, rol tematikoak, hautapen-murriztapenak...). Aditzen semantika aztertzeko sintaxia kontuan hartu behar da zalantzarik gabe. Esate baterako, Levin (1993) eta Pustejovskyren (1995) lanak (ikus III. kapitulua) argi erakusten dute adiera ezin dela aditzaren egituratik banatu. Hau da, egitura sintaktikoa kontuan hartu gabe, hauen ustez ezinezkoa da ale lexikalaren adierazpena egitea. Hortaz, forma bera baina adiera desberdinak dituen aditz batek, izaera sintaktiko desberdina ere izango du.

WordNetek ere informazio sintaktiko-semantikoa erabiltzen du *synset*ak osatzeko: *synset*eko osagaiek hautapen-murriztapen eta azpikategorizazio bera izan behar dute. Hori ez bada betetzen, aditzak *synset* desberdinetan banatzen dira.

(1) Mary ate an apple.

(2) Mary ate.

Adibide honetan ikus daitekeen bezala, ingeleseko eat aditza iragankor edota iragangaitz gisa erabil daiteke. Nahiz eta bi adibideetan aditz-forma bera izan, izaera sintaktiko desberdina izanda, eat aditzak mota bakoitzeko *synset* bat izango du, eat_1 eta eat_2:

(3) {eat_1} (take solid food; "She was eating a banana")
{eat_2} (eat a meal; "We did not eat until 10 P.M.")

Informazio sintaktiko-semantikoak ez du *synset* mailan bakarrik eragiten. *Synset*a jasotzen duen hierarkian edo klase semantikoan ere badu eragina: ingeleseko eat aditza bi klase semantikotan banatua dago, bata iragankorra eta bestea iragangaitza. Hortaz, eat_1 klase semantikoa osatzen duten troponimoak iragankorrak izango dira (gobble, gulp, devour eta abar bezalakoak, euskaraz irentsi aditzaren parekoak direnak), eta eat_2renak iragangaitzak (dine, breakfast eta abar bezalakoak, euskaraz afaldu, gosaldu direnak hurrenez hurren).

Fellbaum eta Kegleren ustez, (1989) izaera sintaktiko ezberdin hau ez da iragankor-iragangaitz alternantziagatik bakarrik: semantikak ere badu eragina. Beste hitz batzuetan esanda, Fellbaum eta Keglek defendatzen dute bi aditz hauek leku desberdinetan daudela taxonomian: (2) adibidean, eat iragangaitzak 'otordu bat jan' adiera du. Hala, aditz honen aditz-troponimok asko (dine, breakfast, snack, picnic...) bere baitan daramate otordua:

(4) They breakfasted hurriedly.
I hate dining alone.
I have been snacking all day.
There were several families picnicking on the river bank.

Bestalde, (1) adibidean bezala eat iragankorra denean, bere adiera 'nolabait irentsi' litzateke. Horregatik, bere troponimo guztiek 'jateko erak' adierazten dituzte (gobble, gulp, devour... bezalakoak).

Vázquez *et al.*-ek (2000) fenomeno honi *infraespezifikazioa* deitzen diote:

> "La infraespecificación consiste en la no expresión sintagmática de un miembro de la valencia combinatoria del verbo, produciéndose una oposición semántica entre una contrucción más específica y otra más general, [...] donde los elementos infraespecificados son aquellos que contienen menos información, es decir, los más generales." (Vázquez *et al.*, 2000, 126. or.)

Fenomeno honetaz gain, *synset*-mailan eta hierarkia-mailan eragina duten beste fenomeno batzuk ere jasotzen dituzte WordNeten. Esate baterako, alternantzia kausatibo/inkoatiboa.

> "WordNet contains CAUSE pointers from causative, transitive verbs to the corresponding antiacusative (inchoative), intransitive sense of the same word." (Fellbaum, 1998a, 83. or.)

Hala, (5) adibideko aditzak nahiz eta forma berekoak izan, polisemikotzat joko dira, eta ondorioz, hierarkian *synset* ezberdinetan kokatuko dira, semantikoki eta sintaktikoki ezberdinak direlako. Gainera, break_2 *synset*aren troponimoek inkoatibo izaera izango dute (The plastic bottle crushed/ cracked) eta break_5 kontzeptuarenek, aldiz, kausatiboak (He smashed/ shattered a plate).

(5) {break_2} (become separated into pieces; "The figurine broke")
{break_5} (cause to separate into pieces; "He broke the plate")

Honela bada, Fellbaum eta Keglek — Levinek (1993) eta Pustejovskyek (1995) bezala— adiera hartzen dute oinarri gisa ezaugarri sintaktikoak definitzeko:

> "Thus, the semantics of the troponyms in each case provide a classification in terms of two distinct hierarchies matching the syntactic distinction between the two verb groups." (Fellbaum eta Kegl, 1989, 97. or.)

Hala, Euskal WordNeteko aditzen adierak zehazteko hiztegietako adierek bakarrik ez dute balio, izaera sintaktikoa ere guztiz beharrezkoa da *synset*en arteko desberdintasunak egiteko. Hautsi eta jan aditzen kasuan, esate baterako, gorago aipatu dugun eat eta break aditzen fenomeno bera gertatzen da: forma iragankorra eta forma iragangaitza bi *synset* desberdinetan daude. Ondorioz, hautsi_1 iragankorra denez (Platera puskatu zuen esaldian, adibidez), honen azpian dauden troponimoak iragankorrak izango dira (birrindu eta txikitu bezalakoak). Aldiz, hautsi_2 iragangaitza denez (Platera berotzean hautsi zen), honen troponimoak iragangaitzak dira (esate baterako, zaratatu).

V.3.2 MCRn aditzak txertatzeko azterketa

Argi dago, beraz, aditzak Euskal WordNeten lantzean adiera-banaketan eta hierarkian zerikusia duten ezaugarri sintaktiko-semantiko hauek guztiak kontuan hartu behar ditugula. Hori dela eta, izenekin egun erabiltzen ari garen orrazketa motaz (**hitzez hitzekoa**) gain, beste orrazketa mota bat ere probatu nahi izan dugu aditzekin: **hierarkiaz hierarkiakoa**. Hala, bost aditz (hitzez hitzeko eskuzko orrazketaren kasuan) eta hierarkia bat (hierarkiaz hierarkiako eskuzko orrazketaren kasuan) aukeratu eta landu ondoren, aditzen lanketa masiborako zein orrazketa mota den egokiena ondoriozta dezakegu.

Lehendabizi, ordea, bost aditzen hitzez hitzeko eskuzko orrazketa zertan izan den azalduko dugu.

V.3.2.1 Bost aditzen hitzez hitzeko eskuzko orrazketa

Izenekin egindako orrazketa mota bera da: aditz batek hiztegietan dituen adierak Euskal WordNeten daudela ziurtatzea eta *synset*ean dauden beste ordainak egokiak direla egiaztatzen saiatzea. Orrazketan erabilitako baliabide eta iturriak ez dira izenekin erabilitako berdinak izan, eta metodologia aldetik ere aldaketa batzuk egon dira. Hasteko, orrazketa mota hau aditz batzuekin bakarrik probatu da. Hau da, orrazketa mota hau aditzen lanketarako baliagarria den aztertzeko, bost aditz bakarrik landu ditugu (esan, banandu, banatu, abestu eta ekarri), gero ondorio batzuk atera ahal izateko.

Azterketarako hautatutako aditzen artean, ezaugarri eta jokaera guztietako aditzak sartzen saiatu gara: maiztasun handikoak eta txikikoak, eta joera sintaktiko desberdinekoak (iragankorrak eta iragangaitzak, adibidez).

Aditzak aukeratzeko beste irizpide garrantzitsua *Volem2* proiekturako aztertutako euskal aditzen artean egotea zen. Proiektu honetan Volemeko (III.2.3.3) aditz eta preposizioei euskara eta okzitanieraren informazioa gehitu zaie, beti ere Volemerako definitutako formalismoari jarraituz. Euskarako aditzei dagokionez, Aldezabalek (2004) aztertutako aditzen informazioa txertatu zen. Hala, Aldezabalek bere ikerlanerako aukeratutako ehun aditzetatik berrogei Volem EBLan zeudenez, horietatik abiatu gara hitzez hitzeko orrazketaren azterketarako.

Bestalde, aukeratutako aditzak Aldezabalen lanean eta Volemen aztertutakoak izanik, Euskal WordNeteko, Aldezabalen laneko eta Volemeko EBLak lotzea ekarri du erabaki honek, bakoitza bestearen informazioarekin aberastuz.

Adierak zehazteko erabilitako baliabideen artean, *Elhuyar Hiztegia* —elebiduna— (Elhuyar, 1996) eta *Elhuyar Hiztegi Modernoa* (Elhuyar, 2000) —elebakarra— erabili dira. Hauek dakarten aditzei buruzko informazio sintaktikoa murritza da gure lanerako. Hori dela eta, Aldezabalek (2004) egindako aditz horien sailkapenean oinarritzea erabaki dugu, non aditzaren adiera bakoitzeko azpikategorizazio zehatza definitzen den.

V.3.2.2 Aditz-hierarkia baten orrazketa

Hitzez hitz lantzean lortzen duguna da orrazten ari garen hitzaren adiera guztiak finkatzea eta zehaztea. Hala, hitz horren adiera guztiak orraztuak geratzen dira. Baina, bestalde, beste huts egite bat egin daiteke metodologia horrekin: hierarkiaren egokitasunari nahikoa ez erreparatzea; hierarkia desorekatua gera daiteke kasuren batean, metodologia horrekin ez baita funtsezkoa hierarkia lantzea, landu beharreko hitza baizik. Hortaz, ematen du menderatu beharreko eremua murritzagoa dela.

Horretaz gain, *synset* mailan arituta, *synset* horiek adierazten dutena ulertu ahal izateko, hizkuntzalarioi nahitaezkoa izan zaigu hauek beraien hierarkian kokatzea. Hau da, *synset*aren hiperonimoak eta hiponimoak aztertzea. Hala, *synset*aren klase semantikoari buruzko informazioa lor daiteke, eta, ondorioz, orraztu beharreko *variant*ak klase semantiko horretan egokiak diren ere jakin dezakegu. Hain zuzen ere, horixe egin behar izan dugu (4) eta (5) adibideetan aipatu ditugun eat eta break aditzen kasuan; bere hiperonimoetara eta troponimoetara jo bi *synset* hauen arteko desberdintasuna zertan datzan jakiteko.

Desoreka hauetaz jabetuta, orrazketa era berri batekin saia gintezkeela iruditu zitzaigun: **hierarkiaz hierarkiako orrazketa**. IV. kapitulua esan bezala, WordNeteko aditzak 15 klase semantiko nagusitan banatuak daude. Hauetako bakoitzean aditz horien antolaketaren hastapena dago, *unique beginner* deiturikoak, hain zuzen. Hierarkiaz hierarkiako orrazketarekin hierarkia osoak orraztu ditugu *unique beginner*retatik hasita, hierarkiako azken troponimora arte.

Orrazketa mota hau probatzeko {express_2, give_tongue_1, utter_1} *unique beginner*ra aukeratu genuen hierarkia honen troponimo kopurua, beste hierarkienarekin parekatuz gero, tartekoa zelako. *Unique beginner* askok berrehun troponimo baino gutxiago dituzte, eta beste batzuk, aldiz, bostehun baino gehiago. Guk aukeratutako hau, 198 troponimoekin, erdibidean kokatzen denez, egokia iruditu zaitzaigun orrazketa mota honen lehenengo

ondorioak ateratzeko.

Hurrengo atalean, azterketa honetatik lortutako ondorio nagusienak dakartzagu. Dena den, hierarkiaz hierarkiako orrazketa hau guztia B eranskinean dator, baita ingeleseko eta euskarako hierarkien arteko alderaketa bat ere.

V.3.2.3 Hitzez hitzeko orrazketa ala hierarkiaz hierarkiakoa?

Azterketa honen ondorioz, esan dezakegu hierarkiaz hierarkiako orrazketa, hitzez hitzeko orrazketa baino lan zabalagoa dela. Izan ere, hierarkiaz hierarkiako orrazketan, hitz horrek dituen hiperomino eta troponimo guztiak aztertu behar dira, eta bakoitzaren adiera hierarkia horretan egokia den ala ez egiaztatu. Gainera, hierarkia orekatua eta logikoa den ere aztertu behar da. Troponimo baten ordaina ezin da hiperonimo batena baino orokorragoa izan, adibidez. Orduan, hierarkia osoaren ikuspegia edukitzea oso mesedegarria da. Hala ere, gerta daiteke *synset* bakoitzean dagoen hitzaren zein adiera den ondo ez menderatzea, beharrezkoa baita horretarako hitz horrek dituen gainontzeko adierak ezagutzea. Hortaz, hierarkiaz hierarkiako metodologia egokiagoa dirudi eremu zabalagoa orraztea lortzen delako, baina ez dira, ahal den neurrian, hitz bakoitzak dituen adiera desberdinak alde batera utzi behar.

Hala, ez dirudi erraza erabakitzea zein orrazketa mota den aditzen lanketarako mesedegarriena. Bien artean erabaki ordez, hitzez hitzeko orrazketa eta hierarkiaz hierarkiakoa aldi berean egitea dirudi egokiena. Baina horrek eskuzko lan ugari eskatzeaz gain, aditzen EBLaren garapena mantsotuko luke.

Aztertzeko dugun beste aukera bat da WordNeteko aditzak PropBankeko aditzekin (Civit *et al.*, 2005a) batera garatzea. Arestian aipatu bezala (V.2.2.2 atalean), EPEC corpusa morfosintaktikoki, sintaktikoki, Euskal WordNeteko adierekin eta PropBankeko rolekin etiketatzen ari gara IXA taldean. PropBanken aditz-adiera bakoitza sarrera bat da, eta *VerbNet* (Kipper *et al.*, 2000) EBLko sarrara bati lotuta dago[10]. VerbNeteko sarrera hori, aldi berean, WordNeteko *synset* batekin lotuta dago. Hala, euskarako PropBankeko aditzak garatzean (gerora hauen rolekin EPEC corpusa etiketatzeko), VerbNeteko informazioa erabilita, aditz hauen WordNeteko baliokideak izango genituzke zuzenean.

[10]PropBanki eta VerbNeti buruz III. kapituluan aritu gara.

Lehenago aipatu izan dugun Euskal WordNetekin batera euskarako corpusa semantikoki ere etiketatzen ari gara: EuSemcor (Agirre *et al.*, 2006a). Euskal WordNeten landutako hitza corpusean etiketatzeaz gain, corpusetik ere Euskal WordNeten ez dagoen adiera berriren bat lor daiteke, eta, ondorioz, Euskal WordNet corpus errealeko adiera berriekin aberastu. EuSemcor proiektuan, izenen etiketatzea amaitzean aditzekin hasiko gara. Hortaz, corpuseko aditzen agerpenak Euskal WordNeteko *synset*ekin etiketatu ahal izateko, aldez aurretik, aditzen *synset*ak orraztu egin beharko dira Euskal WordNeten. Hori dela eta, arrazoi praktikoengatik, aditzen hitzez hitzeko orrazketarekin hasiko ginateke, nahiz eta hurrengo faseren batean hierarkiaz hierarkiako orrazketa erabiltzea ez dugun baztertzen.

Azterketarako bi orrazketa hauek kontuan izanda, V.4 taulan Euskal WordNetek dituen aditzen kopuruak ekartzen ditugu.

	Aditzak	***Synset***	***Variant***	***Variant synseteko***	***Lema***	***Variant lemako***
EusWN 0.1	**BC eskuz**	792	-	-	-	-
EusWN 0.2	**eskuz**	3.751	9.510	2,5	3.496	2,7
WN 1.6	**eskuz**	12.127	22.073	1,8	10.326	2,1

V.4 Taula: Euskal WordNeteko aditzen kopuruak WordNet 1.6koekin alderatuta, oinarrizko kontzeptuak, hitzez hitzeko orrazketa eta hierarkiaz hierarkiako orrazketak egin ondoren.

Kopuruetan ikus daitekeen bezala, oraindik oso urruti gaude ingeleseko WordNetetik (WordNet 1.6 bertsioak 12.127 *synset*, 22.073 *variant* eta 10.326 lema dituen bitartean, Euskal WordNetek 3.751 *synset*, 9.510 *variant* eta 3.496 lema ditu, bakarrik).

V.4 Ondorioak

Kapitulu honetan, Euskal WordNeten garapenerako zein metodologia erabili eta nola diseinatu dugun azaldu dugu. Estaldura eta kalitatea izan dira metodologiaren diseinuaren ardatzak, eta hauen arabera banatu ditugu Euskal WordNeteko izen eta aditzen garapena, fase ezberdinetan. Izenen garapenean, esate baterako, lau fase nagusi aipatu ditugu: garapen automatikoa eta oinarrizko kontzeptuen eskuzko aberasketa, kontzeptuz kontzeptuko orrazketa, hitzez hitzeko orrazketa, eta azkenik, hitzez hitzeko orrazketa EuSemcor

corpusaren etiketatze semantikoarekin bateratuta. Hasierako urratsetan estaldura hartu bagenuen abiapuntu gisa, gerora kalitatearen alde jo dugu, eta arrazoi hori dela eta Euskal WordNeten aberasketa mantsotu egiten dela ikusi dugu.

Aditzen kasuan ez gara mintzatu hauen garapenaz —ez baikara oraindik aditzen lanketa masiboarekin hasi—, baizik eta nahiko genukeen garapenaren azterketaz. Aditzen lanketarekin hasi aurretik, aditzen konplexutasuna delaeta —hauek daramaten informazio sintaktiko-semantikoagatik—, hauen garapenerako metodologia proposatu dugu. Horretarako, saiakera batzuk egin ditugu bi orrazketa motekin: izenekin erabilitako hitzez hitzeko orrazketarekin eta hierarkiaz hierarkiako orrazketarekin. Hitzez hitzeko orrazketak ez du hierarkiaren ikuspegia, eta, aldiz, hierarkiaz hierarkiako orrazketak ez ditu hitzaren adierak kontuan hartzen. Dirudienez, bata bestearen osagarria da. Hala, epe laburrean EuSemcor proiektuan aditzen etiketatzea hasiko garenez, aditzen hitzez hitzeko orrazketarekin hasiko ginateke, nahiz eta hurrengo faseren batean hierarkiaz hierarkiako orrazketa erabiltzea ez dugun baztertzen.

VI. KAPITULUA

WordNetetik Euskal WordNetera: bereizgarriak eta hobekuntzak

Euskal WordNeten egon diren orrazketetan, eta kontuan izanda euskarako wordneta ingelesekoaren gainean garatzen ari garela, ingelesaren eta euskararen arteko hainbat bereizgarri linguistiko azaleratu dira. Kapitulu honetan hauen berri emateaz gain, hizkuntzen arteko ezberdintasun horiek nola kodetu ditugun ere azalduko dugu, kasu batzuetan ereduaren hobekuntzak aurkeztuaz.

Hasteko, lexikalizazioari dagozkion bereizgarriak azalduko ditugu (VI.1 atalean). Ingeleseko kontzeptuak antolatzen dituen EBLa izaki, hainbat kontzeptu ez dira lexikalizatuak euskaraz, gure kulturan ez ditugulako erabiltzen. Alderantziz ere gertatzen da; euskal kulturari dagozkion kontzeptu batzuk ez dira ingeleseko hierarkian agertzen. Honetaz gain, maiz gertatzen da ingeleseko kontzeptu bat euskaraz flexio-atzizkidun hitz batekin edota hitz anitzeko esapide batekin adieraztea, eta askotan ez dago garbi horiek euskaraz lexikalizatuak dauden ala ez. Hala, hauen lexikalizazioaren inguruan zalantzak sortzen dira, eta hauei aurre egiteko irizpideak behar dira.

Beste bereizgarri nagusia hierarkiari dagokio (VI.2 atalean). Gure euskarako wordneta ingeleseko hierarkiaren gainean garatzen ari garenez, bi hierarkien arteko aldeak agertzen dira. Esate baterako, ingeleseko hierarkiak oso zehaztapen aberatsa du: *synset* orokorrenetik zehatzeneraino, *synset* kopuru ugari aurkitzen dira (askotan hamar eta hamasei). Horien euskal ordainen bila jotzen dugunean, ordea, askotan ez dugu hitz desberdinik topatzen, eta

horregatik, askotan, ingeleseko hierarkiako *synset* ugari hiperonimoaren ordain bera erabilita, edota hiperonimoarekin batera beste izen, adberbio edota adjektibo bat gehituta itzultzen dira.

Bi bereizgarriz ari bagara ere, esan beharra dago hierarkia-bereizgarrietan ere lexikalizazioaz ari garela, baina hierarkiaren egituraren ikuspegitik. VI.2 atalean, fenomeno honen adibideak emango ditugu eta honen inguruan erabakitako hainbat irizpide azalduko ditugu.

Bi fenomeno hauei heltzeko definitutako irizpideek *The Multilingual Central Repository*k (MCRk) duen errepresentazioaren hedapena eskatzen dute. Hori dela eta, MCRn hobekuntza batzuk proposatu ditugu ingeleseko eta euskarako wordnetak bateratu ahal izateko. Hala, bereizgarri linguistikoen azalpenarekin batera, bereizgarri hauek eragin dituzten errepresentazio-hobekuntzak ere aipatuko ditugu VI.1, VI.2 ataletan zehar eta VI.3 ataletan.

VI.1 Lexikalizazioa

Lexikalizazioa zer den hobeto ulertzeko Lewandowski-ren hitzetara (1992) jo dugu:

> "El término *lexicalización* se refiere a la transformación de un elemento, o una unión de elementos, en un elemento léxico o conceptual único, p. ej. *camino de hierro/ferrocarril*." (Lewandowski, 1992, 208. or.)

Hortaz, lexikalizazioaren transformazioaren ondorioa elementu bat (guk hitz bat esango dugu[1]: ferrocarril) izan daiteke, edota aleen multzo bat (hitz bat baino gehiago), hots, hitz anitzeko esapide bat (camino de hierro).

Autore batzuek diotenez (Calzolari *et al.*, 2002), lexikalizazioa *continuum* gisa ulertu behar da: batetik, produktiboak eta konposizionalak diren egiturak daude, bestetik, finko eta izoztuta dauden egiturak. Honen arrazoia da lexikalizazioa faktore desberdinen emaitza dela. Batzuetan faktore hauek guztiak gerta badaitezke ere, beste batzuetan ez dute inolako eraginik.

Faktore hauen kopurua adostuta ez dagoen arren, faktore garrantzitsuenak *continuum* horretan ondoko ordenan gertatzen dela esaten da: *kolokazioa* > *fijazioa* > *espezializazio semantikoa* > *idiomatizazioa*. Faktore guztiak zeharo betetzen direnean —hots, lexikalizatu beharreko adierazpideak

[1] *Hitza* ulertuta zuriguneen artean dagoen karaktere multzo gisa (Fontenelle *et al.*, 1994).

faktore guztien eragina jaso badu—, orduan, adierazpide izoztu bat (edo *frozen expression* delakoa) izango genuke (adarra jo eta larru bizirik, adibidez). Aldiz, faktore guztiak ez direnean gertatzen —hots, lexikalizatu beharreko sekuentziak faktore guztien eragina jasotzen ez duenean—, adierazpide hori *continuum*aren edozein puntutan gera daiteke (adibidez, janaria egin eta sakelako telefonoa). Hala, adierazpide hauek *continuum*aren puntu batean ala bestean geldituz gero, ezaugarri desberdinak izango dituzte, adierazpide mota desberdinak sortuz.

Lexikalizaturiko hitz anitzekoen kasuan, hurrengo ezaugarriak dituztela esaten da (Calzolari *et al.*, 2002):

- sintaktikoki eta semantikoki guztiz gardenak ez izatea
- konposizionaltasun mugatua izatea
- gutxi gorabeherako esapide finkoak izatea
- arau sintaktikoak guztiz ez betetzea
- lexikalizazio-maila handia izatea
- konbentzionalitate-maila handia izatea

Datu errealekin lan egitean, ordea, lexikografoek ezaugarri hauekiko duten iritzia ez da bateratua. Batzuetan oso lan zaila da hitz bat edo hitz segida bat *continuum* horretako zein puntutan dagoen erabakitzea, hots, lexikalizatuta dagoen ala ez zehaztea. Eta zailtasun hau agerian geratzen da bi hizkuntza konparatzerakoan, edota, gure kasuan bezala, hizkuntza baterako egindako EBLtik abiatuz (*WordNet*), beste hizkuntza bateko lexikoia garatu behar denean (*Euskal WordNet*).

VI.1.1 WordNet, lexikalizazioa eta hizkuntzen arteko aldeak

Askotan aipatu izan dugu WordNet (Fellbaum, 1998a) teoria psikolinguistikoetan oinarritutako lexikoia dela:

> "WordNet is a semantic dictionary that was designed as a network, partly because representing words and concepts as an interrelated system seems to be consistent with evidence for the way speakers organize their mental lexicons." (Fellbaum, 1998a, 7. or.)

Horrela bada, WordNeten kontzeptuak eta hitzak erlazionatzen dira, hau da, kontzeptuen hierarkia bat da eta kontzeptu horien esanahia adierazteko hitzak erabiltzen dituzte. Jatorrizko WordNetek, lexikoi gehienek bezala, kontzeptu eta sarrera lexikalizatuak bakarrik jasotzen ditu, direla hitz bakarrekoak, direla hitz anitzekoak[2]:

(1) {girlfriend, girl} (a woman with whom a man is involved...)
{house} (a dwelling that serves as living quarters)
{scissors} (a cutting implement having two crossed blades)
{sleep} (be asleep)
{simnel} (eaten at mid-Lent or Easter or Christmas)
{forties, mid-foties} (the time of life between 40 and 50)
{cook} (prepare a hot meal)
{pet} (a domesticated animal kept for companionship or...)
{lyrics, words, language}(the text of a popular song)
{furnishing} (the instrumentalities that make a home livable)
{parent} (a father or mother)
{cold} (feeling a sensation of coldness)
{commodity, goods} (articles of commerce)
{waif} (a homeless child especially one forsaken)
{Alps} (a large mountain in south-central Europe)
{military man, serviceman}(someone who serves the forces)

Salbuespen bakarrak hierarkia antolatzen laguntzen duten kontzeptu orokorrak dira, esate baterako, body of water edota visual property. Asmaturiko kontzeptu hauek ez daude lexikalizatuak, baina oso baliagarriak dira klase semantiko bat multzokatzeko eta izendatzeko. Hauei buruzko azalpen gehiago VI.2.1 atalean emango dugu.

[2]Kapitulu honetan aurkezten diren adibideetan, espazio-arazoak direla-eta, *synset*etako *variant* kopurua txikitu edota glosak murriztu ditugu, eman beharreko azalpenak nahikoak iruditu zaizkigunak soilik utziz.

V.1 atalean azaldu dugun bezala, Euskal WordNet WordNetaren gainean garatzen ari gara, Vossen-en (1998) *expand approach*a jarraituz; hots, ingeleseko *variant*ak –(1) adibidekoen moduak— euskarakoekin ordezkatzen ditugu.

Lan hori egiterakoan, editoreak lexikalizazio-arazoak maiz topatzen ditu, bi hizkuntzen artean **kontzeptu-mailako desorekak** eta **adierazpide-mailako desorekak** baitaude.

Desoreka kontzeptualen artean *kontzeptu kulturalak* deritzotenak daude: kultura bati loturik agertzen diren kontzeptuak, beste hizkuntzetan existitzen ez direnak. Adibidez, simnel 'Ingalaterran Eguberrietan jaten den gozokia' da, edota off-sales 'edari alkoholikoak sal ditzaketen Ingalaterrako dendak, non hauek edatea debekatua dagoen'. Hauek Ingalaterrako kontzeptu kulturalak lirateke. Euskaraz ere gertatzen da hori jakina: trikitixa, ikastola, txakolina eta abar Euskal Herriko kontzeptu kulturalak dira. Horrelako kontzeptu kulturalak ditugunean, hizkuntza batean ez da egongo hori adierazteko hitzik. Kasu hauek *hutsune kultural* (*cultural gaps*) bezala izendatzen ditu Vossenek (1999).

> "A cultural gap is a concept not known in [another] culture."
>
> (Vossen, 1999, 39. or.)

Hutsune kulturalak ezin dira hitz bat edo hitz anitzeko esapide baten bidez adierazi; hauek azalpen edo definizio gisa adierazten dira edo bere horretan itzultzen dira (abiapuntuko hizkuntzaren hitz bera erabilita). Horregatik, editoreak hutsune kulturalen lexikalizazioa ez du zalantzatan jarriko, horrelakoak lexikalizatu gabeko kontzeptuak baitira. Hala ere, gero ikusiko dugun bezala, kasu hauek Euskal WordNeten nola landu behar diren erabaki behar izan dugu (ikus VI.1.4).

Adierazpide-mailako desoreka gertatzen da, berriz, bi hizkuntzatan kontzeptua ezagutzen denean, baina bata eta bestean adierazpide desberdinak erabiltzen direnean. Esate baterako, batzuetan ingeleseko *synset*ak euskaraz hitz anitzeko esapideen bidez itzultzen dira:

(2) pet → konpainia-animalia
sleep → lo egin
cook → janaria egin

Alderantziz ere gerta daiteke, hots, euskarako *synset* bat ingelesez hitz anitzeko batekin adieraz daiteke:

(3) polizia → police officer, policeman
abeltzaintza → livestock farming
soinujole → accordion player

Vossenek (1999) horrelakoei *hutsune pragmatikoak* (*pragmatic gaps*) deitzen die:

> "Pragmatic gaps are caused by lexicalization differences between languages, in the sense that in this case the concept is known but not expressed by a single lexicalized form in English:
>
> Dutch: *doodschoppen* (to kick to death)
>
> Spanish: *alevin* (young fish)
>
> Italian *rincasare* (to go back home)"
>
> (Vossen, 1999, 39. or.)

Vossenek, ikusten dugun bezala, hutsune pragmatikotzat jotzen du kontzeptua bi hizkuntzetan egon eta adierazpide-mailan desoreka egotea.

Dena den, ez da erraza hutsune pragmatiko hauen lexikalizazioa ebaztea, batez ere hiztegietan oinarriatuz gero: lo egin hiztegi-sarrera da, aldiz, janaria egin ez; etxe-abere hiztegi-sarrera da, konpainia-animalia, ordea, ez. Hizkuntza sortzailea den heinean, hitz-konbinazio berriak sortzen doaz, eta ulertzen ditugun arren, zaila da esaten lexikalizatuak dauden ala ez. Honek, noski zailtasunak dakartza hitz hori Euskal WordNeten sartu ala ez erabakitzeko.

Zailtasun hau areagotu egiten da aldi berean semantikoki etiketatutako (desanbiguatutako) corpusa sortzen ari bagara (gogoratu V. kapituluan aipaturiko *EuSemcor*). Bertan hitz anitzeko esapide lexikalizatu baten osagai diren corpuseko agerpen guztiak markatu egiten dira. Adibidez, mutil izenaren agerpenak etiketatzen egonez gero, eta corpusean honi lagun izenak jarraitzen badio, mutil, agerpen horretan, hitz anitzeko baten osagarri gisa markatzen da[3]. Hala ere, etiketatzaileak maiz ez daude ados hitz anitzeko esapide lexikalizatua zer den erabakitzeko orduan.

Horregatik, gure ustez bada beste desoreka mota bat: kontzeptu bat existitzea hizkuntza batean (bere adierazpen lexikalarekin; gehienetan hitz bakarrekoa), eta beste hizkuntzan zalantzan egotea kontzeptu hori bereziki bereizten dugun (hots, lexikalizatua dagoen), edo, besterik gabe, sintaxi askeko beste edozein adierazpide gisa ulertzen dugun. Aurreko adibideez gain (konpainia-animalia, janaria egin), horrelakoak izaten dira flexio-atzizkia edo numeroaren marka daramaten ordainak:

[3]Etiketatze semantikoari buruzko argibide gehiago Agirre *et al.*-en lanean (2005b).

(4) words → hitzak
furnishing → altzariak
goods → salgaiak
cold → hotzez

WordNeten hitz hauen adiera flexio-atzizkian edo pluraltasunean oinarritzen da. Hau da, flexio-atzizkia dutenean edota pluralean erabiltzen direnean adiera bat dute, eta gainontzean beste bat edo beste batzuk. Esate baterako, editoreak, WordNetetik abiatuta, singularreko *synset*ak euskaratzean (furniture → altzari, adibidez), ez du lexikalizazio-zalantzarik izaten euskarako ordain hori (altzari) hiztegi-sarrera denean hiztegi elebakar edo elebidunetan. Baina bestela gertatzen da hiztegietan sarrerarik ez dagoenean eta gainera adiera bereizketa argia ez denean. Azken hau (adiera-bereizketa eza) gertatzea arruntagoa da etiketatze semantikoan jardutean WordNeten hitzak editatzen jardutean baino. Editorearen ikuspegitik, errepresentazio-arazoa da gehiago gertatzen dena. Adibidez, furnishing kontzeptua adierazteko, altzari formaren adieretako bat balitza bezala landuko dugu, pluralean erabiltzen dela nolabait markatuz? edo altzariak hitz desberdina erabiliko dugu, horrekin ulertaraziz hitz hori (adiera horrekin) beti pluralean erabiltzen dela? Horrelakoetan editoreak kontzeptuaren lexikalizazioari buruzko zalantzak ditu. Ondorioz, ez daki *synset* horiek nola landu.

Etiketatze semantikoarekin arazo hau areagotu egiten da, testuetako adibideen aurrean ez delako argi ikusten bi formen arteko bereizketaren beharra. Demagun, altzariak (furnishing adierazteko) lexikalizatutzat jotzen dugula. Orduan, altzari eta altzariak adiera desberdineko bi *synset* direla adierazten egongo ginateke eta hori corpusean ere halaxe izan beharko litzateke. Baina etiketatzaileek (5) adibideko agerpenen aurrean zalantzak dituzte. Hau da, ez dakite horrelako agerpenei altzari kontzeptua, altzariak kontzeptua, edo biak dagozkien. Gauza bera hitz eta hitzak, salgai eta salgaiak eta hotz eta hotzez kontzeptuekin.

(5) Etxeko **altzariak** saldu behar izan ditut.
Ez dira nik idatzitako **hitzak**.
Salgaiez beteriko dendak.
Hotzez hil dela salatu dute.

Ingelesetik euskarara itzuli beharrean, alderantziz egingo bagenu arazo bera izango genuke; esate baterako, euskarako guraso hitzak hiztegietan bi

adiera ditu: bata, 'aita edo ama' (gurasoetako bat, alegia) adierazten duena, eta bestea 'aita eta ama' (bi gurasoak, alegia). Bigarrenean, WordNeteko words eta goods kontzeptuekin gertatzen den bera gertatzen zaigu: 'aita eta ama' adiera adierazteko beti plurala erabili behar da, eta honek bereizten ditu bi adierak, hain zuzen ere. Ingelesez, 'aita edo ama' adierazteko parent hitza darabilte. 'Aita eta ama' adiera, aldiz, ez dute hiztegietan jasota eta hiztunak hori adierazteko modua parents da, beste edozein izenekin bezala plurala erabiltzen dute. Guraso 'aita eta ama' adierazten duen *synset*ean, zer beharko luke parent ala parents? Gauza bera euskarako gazteria hitzarekin; ingelesez, kontzeptu hori adierazteko youngs edo young people bezalako bat beharko litzateke, baina *synset*ean young edo youngs jartzea erabaki beharko litzakete.

(6) Parents are asked not to come.
Youngs are the victims of the war on drugs.

Hiztegietan oinarrituz, pluralaren kasuan, hiztegi-sarrera bezala izen bereziak daude (Alpeak, Estatu Batuak eta antzekoak). Izen bereziak ez diren beste pluraletan, hiztegiak askotan ez datoz bat. *Hiztegi Batua*k[4], esate baterako, seme-alabak, senar-emazteak eta damak ('dama-joko'a adierazteko) hiztegi-sarrera gisa proposatzen ditu.

(7) *Hiztegi Batua*
seme-alabak: seme-alabak
senar-emazteak: senar-emazteak
damak: (joko-izena)

Guraizeak, aiton-amonak eta prakak formak, aldiz, ez dira hiztegi-sarrera, hots, dagokien hiztegi-sarrera singularrean dago (guraize, aiton-amona eta praka); baina flexioaren erabilerari buruzko nolabaiteko azalpena dator.

(8) *Hiztegi Batua*
guraize: pl.
aiton-amona: pl.
praka: pl., praka-pare bat

Azkenik, mobiliario eta mercancía bezalakoak adierazten dituzten euskal ordain pluralak (altzariak eta salgaiak), hiztegi-sarrera singularrean dute (salgai eta altzari) inolako beste azalpenik gabe. Beraz, dirudienez, *Hiztegi Batua*k hitz hauen erabilera plurala ez du bereziki markatzen.

[4] `http://www.euskaltzaindia.net/hiztegibatua` (2007-07-02an atzitua).

(9) *Hiztegi Batua*
salgai: 1. pred.: salgai dagoen liburua
2. iz: Europa guztiko salgaiak itsasoz zabaltzen zituen
altzari: altzari

(9)ko adibide hauek berak beste hiztegietan era ezberdinean datoz adierazita. Hala ere, esan beharra dago gehienetan hiztegi-sarrera gisa lema soilik erabiltzen dutela. (8) adibidekoak bezalako azalpenak ere oso era aldakorrean ematen dira hiztegi batetik bestera. Horren adierazgarri (10) eta (11) ditugu, non *Hiztegi Modernoa*k (Elhuyar, 2000) eta *Elhuyar Hiztegi* elebidunak (Elhuyar, 1998)[5] (hurrenez hurren) (9)ko adibide berdinak nola adierazten dituzten ikus dezakegun[6]:

(10) *Hiztegi Modernoa*
seme-alaba: Gizonezkoa edo emakumezkoa bere gurasoekiko
senar-emazte: Elkarrekin ezkondurik dauden gizon eta emakumea
dama: *ez dago horrelako sarrerarik joko-izena adierazteko*[7]
guraize: Erdialdean giltzatzen diren eta alde batean ahoa eta punta...
aiton-amona: *ez dago horrelako sarrerarik*
praka: galtzak
salgai: 1. Saltzeko dagoen gauza. 2. Saltzeko
altzari: [...] hainbat zereginetarako erabiltzen den objektu higigarria

(11) *Elhuyar Hiztegia*
seme-alaba: ez sing.; Hijos [hijos e hijas]
senar-emazte: ez sing.; Marido y mujer, esposos, cónyugues
dama: *ez dago horrelako sarrerarik joko-izena adierazteko*[8]
guraize: pl.; tijera(s)
aiton-amona: *ez dago horrelako sarrerarik*
praka: pl. pantalones
salgai: batez ere pl.; mercancía, género
altzari: mueble; (pl.) mobiliario, enseres

Flexio-atzizkidun hitzetan ere gertatzen dira halako zalantzak: hotzik hiztegi-sarrera da, baina hotzez ez; edota buruz hiztegi-sarrera da, baina eskuz ez.

[5] `http://www1.euskadi.net/hitz_e/indice_e.html` (2007-07-02an atzitua).
[6] Hiztegietako definizioak eta azalpenak laburtu egin dira.
[7] 'Joko-izena' adierazteko dama-joko sarrera dago.
[8] Ikus 6. oin-oharra.

WordNetek eta hiztegiek lexikalizaturiko kontzeptuak jasotzen badituzte ere, eta Euskal WordNeteko hasierako helburua horixe bazen ere, argi dago kasuistika honen aurrean, kontzeptuen lexikalizazioa ebaztea zaila dela, are gehiago, corpusarekin lan egitean. Horregatik, eta lexikalizazioaren zailtasunaz jabetuta, lana ahalik eta modu erosoenean egiteko irizpideak lantzea erabaki genuen.

VI.1.2 Zalantzazko lexikalizazioa duten adierazpideen beharra

Zerk erabakitzen du kontzeptu bat lexikalizatua dagoen ala ez; hiztegietako hiztegi-sarrera izateak ala ez izateak? Normalean, ordain batzuk lexikoian sartzeko edo ez erabakitzeko erabiltzen diren irizpideak beste faktore eta baldintzen arabera zehazten dira; gehienetan, lexikoiari eman nahi zaion erabilerak erabakitzen du zer ordain mota behar diren lexikoian. Gure kasuan, Euskal WordNetek euskararen interpretazio semantikoa eskaintzen duen EBLa izatea nahi dugu, LNPko hainbat atazetan lagungarria izan dadin. Hori dela eta, lexikalizaturiko ordainez gain, zalantzazko lexikalizazioa duten ordainak ere Euskal WordNeten gehitzea beharrezkoa iruditu zaigu. Arrazoietan sakonduko dugu segidan.

Arrazoi nagusiena da gure lanaren helburuen artean ez dagoela lexikalizaziori buruzko hausnarketa sakona egitea, baizik eta Euskal WordNet ahalik eta ordain kopuru handienarekin aberastea. Gainera, ordain bakoitzaren lexikalizazioa erabakitzen gehiegi luzatuz gero Euskal WordNeten garapena izugarri motelduko genuke.

Bestalde, ingeleseko *variant*ak euskarakoekin ordezkatzeko hiztegiak bakarrik kontuan hartuko bagenitu, (hots, hiztegi-sarrera direnak ordain gisa eman eta hiztegi-sarrera ez direnak ez) aipatutako *synset* horiek guztiak (furnishing → altzariak; pet → konpainia-animalia eta abar) euskaraz hutsik geratuko lirateke. Aldiz, ordain horiek Euskal WordNeten egonez gero, oso erabilgarriak izan daitezke, adibidez, itzulpengintza automatikorako.

Bestalde, interpretazio semantikoa eta adieraren desanbiguazioa egiteko ere oso baliagarriak dira: zenbat eta ordain gehiago egon Euskal WordNeten, orduan eta errazagoa izango zaio programa bati adierak desanbiguatzea.

Hitz anitzeko esapideen kasuan, zalantzazko lexikalizazioa dutenak EBLan txertatzeko ikuspegi hau dagoeneko erabilia izan da Bentivogli eta Piantaren lanean (2002). Autore hauek *maiz errepikatzen diren konbinazio askeak* deitzen dituztenak italierako wordnetean txertatzen dituzte.

(12) a. WordNet {toilet roll}
Italierako WordNet {rotolo di carta igienica}
b. WordNet {bike}
Italierako WordNet {andare in bicicletta}

Hortaz, Bentivogli eta Piantak (2002) *maiz errepikatzen diren konbinazio askeak* sartzen dituzte bakarrik italierako wordnetean. Hitz anitzeko bat maiz errepikatzen den konbinazio askea den ala ez jakiteko, aldez aurretik neurtu behar dira hitz anitzeko esapide horrek corpus orekatu batean dituen agerpenak eta hitz anitzekoen osagaien arteko asoziazio-maila.

Euskal WordNeten sartuko ditugun zalantzazko hitz anitzekoak, aldiz, ez dira bakarrik maiztasun handikoak izango. VI.1.4 atalean azalduko dugun bezala, hauek Euskal WordNeten sartzeko, beste ezaugarri batzuk ere hartuko ditugu kontuan.

Euskal WordNeteko *variant* lexikalizatu, zalantzazko lexikalizatu, eta ez-lexikalizatuak koherenteki lantzeko, hauei buruzko terminologia zehaztu behar izan dugu, eta baita hainbat irizpidetan oinarritutako metodologia bat definitu ere.

VI.1.3 Terminologiaren azterketa eta gure aukera

VI.1 atalean esan dugun bezala, adierazpideek, *continuum*aren puntu batean ala bestean geldituz gero, ezaugarri desberdinak dituzte, eta horrek literaturan hainbat sailkapen egitea ekarri du. Horietako batzuen berri emango dugu hemen.

Segidan aurkeztuko dugun sailkapena hitz anitzekoei dagokie. Hitz bakarren eta hitz anitzekoen lexikalizazioaz aritu bagara ere, lexikalizazio-arazoak gehienetan hitz anitzekoekin aztertzen dira, hauetan konplexuagoa baita lexikalizazio-mugak zehaztea.

Sag *et al.*-en (2002) ustez, bi hitz anitzeko mota daude: *hitz anitzeko esapide lexikalizatuak* (*lexicalized phrases*) eta *hitz anitzeko esapide institutzionalizatuak* (*institutionalized phrases*). Hitz anitzeko esapide lexikalizatuak horrela deskribatzen dituzte:

> "Lexicalized phrases have at least partially idiosyncratic syntax or semantics, or containing "words" which do not occur in isolation."
>
> (Sag *et al.*, 2002, 3. or.)

Ezaugarri hauek betetzen dituzten hitz anitzeko lexikalizatuen artean hurrengoak daude: **lokuzioak** (*proper idioms*) (13. adibidean), **bana daitezkeen lokuzioak** (*decomposable idioms*) (14. adibidean), **hitz elkartuak eta hitz anitzeko esapide terminologikoak** (*compound nominals and terminological multiword expressions*) (15. adibidean), **izen bereziak** *proper names* (16. adibidean), **aditz-partikula egiturak** (*verb-particle constructions*) (17. adibidean) eta **aditz arin egiturak** (*light verb constructions*) (18. adibidean)[9].

(13) a. to kick the bucket
'hil'; lit: 'ontziari ostikada eman'
b. to pull somebody's leg
'adarra jo'; lit: 'norbaiten hankatik tira egin'
c. adarra jo
'to pull somebody's leg'; lit: 'to play the horn'
d. larru bizirik
'stark naked'; lit: 'raw-skinned'

(14) a. to spill the beans
'agerian utzi'; lit: 'sekretuak ezaguturazi'
b. to sweep something under the carpet
'ezkutatu'; lit: 'alfonbra azpira erraztatu'
c. burua jan
'to brainwash'; lit: 'to eat the head'
d. muturra sartu
'to stick somebody's nose'; lit: 'to put the muzzle in'

(15) a. car park
'aparkaleku'; lit: 'auto parke'
b. central processing unit
'prozesatzeko unitate zentral'; lit: 'prozesatzeko unitate zentral'
c. buruhauste
'problem'; lit: 'broken head'
d. sudur-zapi
'handkerchief'; lit: 'nose-cloth'

[9]Ingelesko adibideak Sag *et al.*-etik (2002) hartutakoak dira, baina hauekin batera euskarako batzuk ere proposatzen ditugu.

(16) a. Los Angeles
b. Chicago Bulls
c. Euskal Herri
'Basque Country'
d. Europako Banku Zentrala
'European Central Bank'

(17) a. do without
'moldatu'; lit: 'gabe moldatu'
b. go after
'-en atzetik joan'; lit: '-en atzetik joan'
c. -tzat hartu
'to take someone for'; lit: 'to take as'
d. -i eutsi
'defend'; lit: 'to hold to something'

(18) a. make a mistake
'akats bat egin'; lit: 'akats bat egin'
b. fall asleep
'lo hartu'; lit: 'lo hartu'
c. hitz eman / berba eman
'to promise'; lit: 'to give the word'
d. min hartu
'to hurt'; lit: 'to take hurt'

Lokuzioak egitura izoztuak dira. Beraz, beraien adiera ezin da konposizionalki osatu hitz anitzekoaren osagai bakoitzetik. Gainera, hitz anitzeko osagai bakoitza ezin da beste sinonimo batengatik ordezkatu. Esate baterako, (13c) adibideko adarra jo lokuzioa ezin da ulertu konposizionalki, kasu horretan adarra hitzak ez baitu zerikusirik hiztegietan duen adierekin (animaliarena, zuhaitzarena...). Honen adierazgarri dugu, hitz anitzeko adarra osagaia ezin dela hiztegietan duen adiera horietako baten sinonimoarengatik ordezkatu: *adarkia jo.

Bana daitezkeen lokuzioak, ordea, maiz elkarrekin agertzen edo erabiltzen diren hitz multzoak dira, eta beraien adiera konposizionaltzat jotzen dute. Esate baterako, berari ez dagokion arazo batean muturra sartu du esaterakoan, hitz anitzekoaren adiera konposizionalki uler daiteke, nahiz eta muturra sartu ekintza fisikoaren adiera metaforikoa izan (koldarrak amaitzearren muturra katiluan sartu zuen). Hala ere, mota honetako hitz anitzekoen osagaiek badute halako ezaugarri semantiko bat euren sinonimoengatik ordezkaezinak egiten

dituena. Hala nola, berari ez dagokion arazo batean muturra sartu du esan badezakegu ere, arraroa litzateke berari ez dagokion arazo batean musua sartu du erabiltzea. Antzeko fenomenoa ikus daiteke aipatutako beste hitz anitzeko motetan ere. Adibidez, hitz eman eta berba eman sinonimoak dira, biek promes egin adierazten dute. Aldiz, ele izena hitz eta berbaren sinonimoa izan arren, ezin da ele eman erabili hitz eman edo berba emanen sinonimo gisa, ele eman hitz anitzekoak beste adiera bat baitu: 'hizpidea eman'.

Sag *et al.*-ek (2002) *hitz anitzeko esapide instituzionalak* sintaxiaren erregelak jarraitzen dituzten hitz konbinazioak baino ez direla argudiatzen dute. Hala ere, osagaien adierak konposizionalki elkartzen badira ere, ezin dira beti sinonimo batengatik ordezkatu (ikus 19. adibidea). Dirudienez, konbentzionalizatutako egiturak dira, eta, horregatik, gauza bera adierazteko erabil litezkeen beste hitz anitzeko batzuk baino maiztasun handiagoa dute. Esate baterako, euskaraz nortasun-agiri erabiltzen da 'norbaiten identitatea ziurtatu ahal izateko balio duen txartela/agiria' adierazteko. Honen ordez, identitate-agiri berdin-berdin erabil zitekeen. Are gehiago, hala beharko luke, 'pertsona bat nor den adierazten duen datu multzoa' adierazteko hobetsitako ordaina identitate baita, eta ez nortasun. Hala eta guztiz ere, nortasun-agiri izan da gure artean zabaldu dena, nahiz eta nortasun hitzaren adiera hori hobetsia ez egon. Antzekoa gertatzen da telefono mugikor hitz anitzekoarekin: telefono higikor, telefono higigarri edo sakelako telefono erabiliz gero, edonork ulertuko baligu ere, konbentzionalizatutako forma telefono mugikor izan da.

(19) a. traffic light
'semaforo'; lit: 'trafiko argi'
b. telephone box
'telefono-kabina'; lit: 'telefono-kabina'
c. telefono mugikor
'cellphone'; lit: 'mobile phone'
d. nortasun-agiri
'identity card'; lit: 'identity document'

Horrela, bada, Sag *et al.*-en (2002) ustetan, *hitz anitzeko esapide instituzionalizatuak* semantikoki eta sintaktikoki konposizionalak dira, baina estatistikoki instituzionalak.

Bentivogli eta Piantak (2002) *hitz anitzeko esapide lexikalizatuak* (*lexicalized multiword expression*) eta *maiz errepikatzen diren konbinazio askeak* (*recurrent free combination*) bereizten dituzte.

Sag *et al.*-en (2002) *hitz anitzeko esapide lexikalizatuak* eta Bentivogli eta Piantarenak (2002) bat datoz. Hala ere, Bentivogli eta Piantak (2002) hauen azpian bi azpimultzo bakarrik egiten dituzte: **lokuzioak** (*idioms*) eta **kolokazio mugatuak** (*restricted collocations*). Azken hauek Sag *et al.*-en (2002) *hitz anitzeko esapide lexikalizatu* izenaren azpian multzokatutako guztiak onartzen dituzte. Bentivogli eta Piantaren ustetan (2002), *lokuzioek* eta *kolokazio mugatuek* analisi linguistikoaren mailaren batean unitate gisa jokatzen dute eta hitz anitzeko esapide lexikalizatuak dira. Hala ere, beraien artean badago nolabaiteko desberdintasuna. *Lokuzioak* egitura izoztuak dira, eta beraien adiera ez da konposizionala (ikus 13. adibideko kasuak). *Kolokazio mugatuak*, aldiz, maiz elkarrekin agertzen edo erabiltzen diren hitz multzoak dira, eta beraien adiera konposizionala da (14. adibideko kasuekin azaldu dugun bezala).

Bestalde, *maiz errepikatzen diren konbinazio askeek* sintaxiaren erregelak jarraitzeaz gain, adiera konposizionala dute eta osagai bat sinonimo batez ordezkatzea onartzen dute. Adibidez, ingeleseko toilet roll hitza euskaraz komuneko paper-erroilu itzultzen da *Euskaltermen*[10] arabera (ikus (20b) adibidea), eta italieraz rotolo di carta igienica. Dena den, erroilu izenaren sinonimo bat erabil dezakegu gauza bera adierazteko: biribilki. Eta aldi berean italieraz, rotolo osagaiaren sinonimo bat ere erabil dezakegu: bobina. Hori dela eta, Bentivogli eta Piantak (2002) horrelako formak ez-lexikalizatu bezala deskribatzen dituzte, eta, ondorioz, hauek ez dira hiztegi-sarrerak izango.

(20) a. bizikletan ibili/joan
andare in bicicletta
'to bike'; lit: 'to go on a bicycle'
b. komuneko paper-erroliu, komuneko paper-biribilki
rotolo di carta igienica, bobina di carta igienica
'toilet roll'; lit: 'toilet paper roll'

Azkenik, Alegria *et al.*-ek (2004) *hitz anitzeko esapidea* terminoa erabiltzen dute **edozein** hitz-konbinazio adierazteko; lexikalizatuak nahiz ez lexikalizatuak. Bestetik, *hitz anitzeko unitate lexikal* darabilte lexikalizaturiko hitz anitzekoei buruz bakarrik hitz egiteko, hau da, semantikoki ez-konposizionalak eta sintaktikoki idiosinkratikoak diren hitz anitzeko horiek izendatzeko; hala nola, (13)tik (18)ra aipatutako adibide guztiak. Ikuspegi hau, hain zuzen ere, IXA taldean garatzen ari den tesi-lan batean hartu da

[10] `http://www1.euskadi.net/euskalterm` (2007-07-02an atzitua).

(Urizar, *Kolokazioak euskaraz*), non hitz anitzekoen azterketa sakona egiten den, gero LNPko hainbat atazetan automatikoki ezagutu ahal izateko.

Gurean, hitz anitzeko esapideez hitz egiterakoan, Alegria *et al.*-en (2004) terminologia erabiltzearen alde egin dugu, orokorra izanik erabilerrazagoa zaigulako, eta berean, IXA taldekoarenarekin bat egiten genuelako. Hala, aurrerantzean, **hitz anitzeko esapideak** (**HAEak**) eta **hitz anitzeko unitate lexikalak** (**HAULak**) bereiztuko ditugu. Beste hitz batzuetan esanda, HAE adierazpidea lexikalizatu nahiz ez-lexikalizatuentzako termino orokor gisa erabiliko dugu, eta, aldiz, zehazki lexikalizatutakoei erreferentzia egiterakoan, HAUL. Hortaz, (21)eko guztiak HAEak dira, baina horietako batzuk bakarrik dira HAULak.

Dena den, eta aipatutako tesi-lan horren emaitzak iritsi bitartean, beste hainbat terminologiaren beharra izan dugu.

Esan dugun bezala, simnel eta off-sales bezalakoak hutsune kulturalak dira, eta hutsune kulturalak ezin dira hitz bat edo HAE batez adierazi (behintzat jatorrizkoa ez den hizkuntzan). Aitzitik, azalpen antzeko bat behar dute. Beraz, HAEen artean, beste maila bateko bereizketa behar dugu: abiapuntu den hizkuntzako hitzaren ordaina *kategoria sintaktiko berarekin* itzulitakoak, eta, lexikalizatzeko modurik ez daukatenez, *azalpen batekin* itzuli behar direnak.

Kategoria sintaktiko berdinarekin itzultzen direnen artean, berriz, bi motakoak egongo dira:

- Lexikalizatuak, HAULak deritzogunak.
- Zalantzazko lexikalizazioa dutenak.

Azken hauei *adierazpide sintagmatiko* (*phrasal concepts*) deitu diegu:

> "*Phrasal concepts* constitute the representation of phrase structures that are composed by several concepts with semantic content."
>
> (Agirre *et al.*, 1994b, 1.394. or.)

Hona hemen adierazpide sintagmatikoen adibide batzuk:

(21) a. WordNet: {corkscrew}
Euskal WordNet: {kortxo-kentzeko}
b. WordNet {bike}
Euskal WordNet: {bizikletan ibili}

Beraz, dagoeneko badakigu zein kasuistika izango dugun. Baina nola jakingo dugu, kasuan kasu, *variant* bat HAUL gisa, adierazpide sintagmatiko gisa, hutsune kultural gisa, hitz bakar lexikalizatu gisa ala ez-lexikalizatutako hitz gisa landu behar den? Horretarako, hurrengo ataleko irizpideak definitu behar izan ditugu.

VI.1.4 Euskal ordainak Euskal WordNeten sartzeko eta markatzeko irizpideak

VI.1.1 atalean lexikalizazioaren inguruko arazoak aurkeztu ditugu, baita hauen hiztegietako adierazpideei buruzkoak ere. Atal honetan, forma hauek Euskal WordNeten sartzeko eta errepresentatzeko finkatu ditugun irizpideak azalduko ditugu.

Euskal WordNeteko editoreak hiztegi-sarrera den beste ordain batekin itzultzen badu *synset*a, ez du inolako zalantzarik ez bere lexikalizazioaz, ez EBLan adierazteko moduaz. Aldiz, hiztegi-sarrera ez denean, orduan sortzen dira lexikalizazioari buruzko zalantzak. Beraz, lehenengo irizpide argia horixe dugu:

- **Lehenengo iripizdea:** Euskarako adierazpidea *Elhuyar Hiztegia*n, *Hiztegi Modernoa*n, *Euskal Hiztegia*n, *Euskaltermen* edota *Hiztegi Batua*n[11] **hiztegi-sarrera bada**, orduan, editoreak adierazpide hori lexikalizatutzat hartuko du eta *synset*ean sartuko du. Adibidez, ingeleseko sleep aditza euskaraz lo egin esaten da. Forma hau gutxienez aipatutako hiztegi batean hiztegi-sarrera bada, editoreak *synset*ean sartuko du *variant* gisa eta **lexikalizatu** gisa markatuko du (*LEX* markarekin):

(22) **Synset-zenbakia:** 00009805
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** Lo-egoeran egon
=> **Sinonimoak:**
=> lo egin

Lehenengo irizpideak hiztegi-sarrera diren HAEei egiten die erreferentzia. Beste guztientzat ere irizpide batzuk behar ditugu nolabait kodetzeko eta bereizteko.

[11] Aipatu beharra dago, hiztegi hauek hautatu izanaren arrazoia. Alde batetik, IXA taldeak hiztegigileekin duen harreman estuarengatik, euren hiztegiak euskarri elektronikoan erabiltzeko aukera ematen digutelako. Bestetik, hiztegi espezializatu (*Euskalterm*) eta orokor gisa erabilera handia duten hiztegiak direlako.

- **Bigarren irizpidea:** Euskarako adierazpidea HAE bat bada, eta *Elhuyar Hiztegi*an, *Hiztegi Modernoa*n, *Euskal Hiztegi*an, *Euskalterm*en edota *Hiztegi Batua*n hiztegi-sarrera ez bada:

  (a) kontzeptu hori euskaraz **kategoria sintaktiko berarekin** itzul badaiteke, orduan, editoreak adierazpide hori *variant* gisa sartuko du, eta **lexikalizatu** (*LEX*) eta **adierazpide sintagmatiko** gisa (*IXALEX*) markatuko du. 23. adibidean, ingeleseko to cook *synset*ari lotutako euskarako *variant*ak ditugu (janaria prestatu eta janaria egin). Euskaraz, to cook adierazteko hiztegi-sarrera ez den, baina ingeleseko kontzeptuaren kategoria sintaktiko bera duen HAE bat darabilgu.

  (b) kontzeptu hori adierazteko **kategoria sintaktiko desberdineko** HAE konplexu bat —definizio edo azalpen gisakoa— erabili behar badugu, orduan, editoreak HAE hori ez du *variant* gisa txertatuko baizik glosa gisa. Hauek *hutsune lexikal —lexical gaps* (Vossen, 1999)— izendatu ditugu, eta **ez-lexikalizatu** gisa markatu ditugu (*NOLEX*) (ikus 24. adibidea).

(23) **Synset-zenbakia:** 01143604
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** elikagaiak jateko prestatu
=> **Sinonimoak:**
=> janaria prestatu (IXALEX)
=> janaria egin (IXALEX)

(24) **Synset-zenbakia:** 05678078
=> ***Synset*aren lexikalizazio-egoera:** NOLEX
=> **Glosa:** Ingalaterran Eguberrietan jaten den gozokia
=> **Sinonimoak:**
=> -

- **Hirugarren irizpidea:** Kontzeptu bat adierazteko **plurala** edo **flexio-atzizkia** duen forma erabili behar bada, orduan, editoreak *variant*a pluralaren edota flexioaren atzizkirik gabe sartuko du, eta alboan interfazeak eskaintzen duen *PLU* marka (ikus 25. adibidea) edo *FLEX* marka (ikus 26. adibidea) aukeratuko du, kontzeptu horrek pluraleko tasuna edo flexio-atzizkia, hurrenez hurren, hartzen duela adierazteko.

(25) **Synset-zenbakia:** 02729592
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** Hainbat zereginetarako erabiltzen diren objektu higigarriak.
=> **Sinonimoak:**
=> altzari (PLU)

(26) **Synset-zenbakia:** 01199751
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** Bero-gabeziak gorputzean eragiten duen sentsazioa.
=> **Sinonimoak:**
=> hotz (FLEX)

Hala, ez gara forma pluralaren lexikalizazioari buruzko eztabaidetan sartzen. Ingeleseko kontzeptu bat euskaraz adierazteko plurala behar dugula bakarrik adierazten dugu, eta horretarako darabilgu *PLU* etiketa.

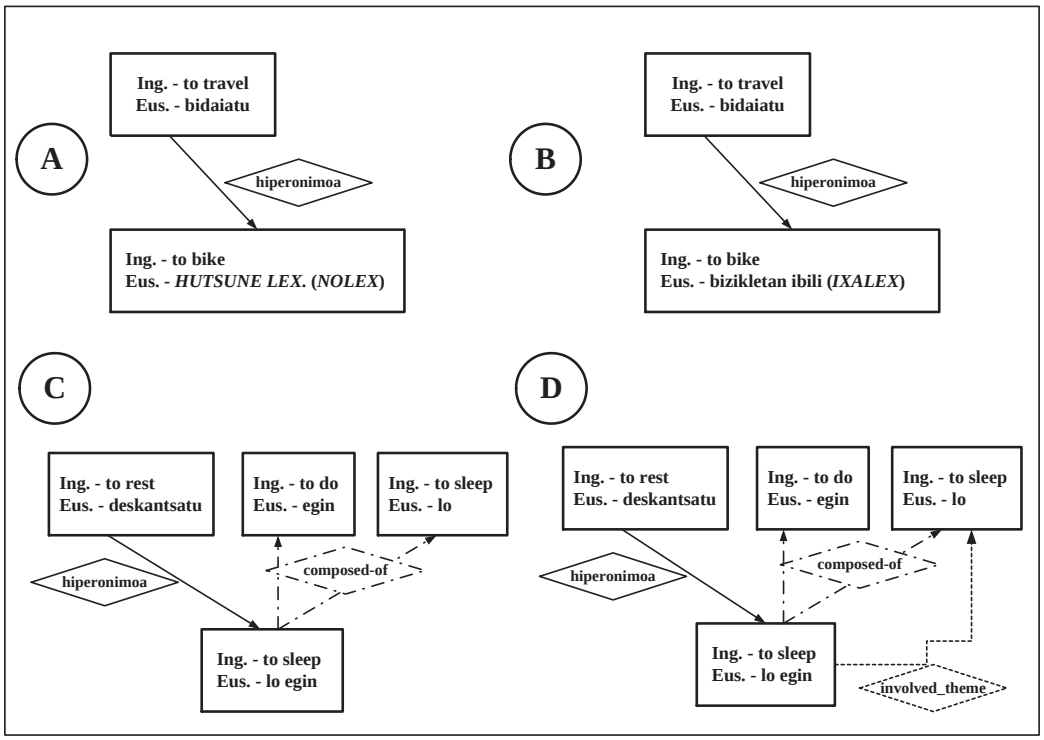
Nahiz eta oraingoz izen eta aditzekin lan egin dugun, dagoeneko aurreikusten dugu, hirugarren iripizpide honek etorkizunean landuko ditugun beste kategorien (adjektibo eta adberbioen) adierazpenetarako ere balioko digula, hotzik/hotzez bezalakoak adierazteko, adibidez.

VI.1.4.1 Barne-errepresentazio semantikoa Euskal WordNeten

HAEak Euskal WordNeten lantzeko irizpideak hauen lexikalizaziora bakarrik mugatzen dira. Irizpide hauek ez dute HAEei buruzko bestelako informaziorik ematen, hala nola, HAEa osatzen duten osagaien arteko harreman semantikoei buruzkoa. Sag *et al.*-en ustez, (2002) HAEen analisi sintaktikoa eta interpretazio semantikoa lotu ahal izateko, HAEen barne-errepresentazio semantikoa beharrezkoa da; batez ere, konposizionalki uler daitezkeen HAE horiena, edota, Sag *et al.*-en (2002) terminologiari jarraituz, *bana daitezkeen esapideena* (*decomposable idioms*) (14. adibidean), *hitz elkartuak eta hitz anitzeko esapide terminologikoena* (*compound nominals and terminological multiword expressions*) (15. adibidean), *aditz arin egiturena* (*light verb constructions*) (18. adibidean) eta *hitz anitzeko esapide instituzionalizatuena* (*institutionalized phrases*) (19. adibidean).

Bentivogli eta Piantak (2002), italierako wordneteko HAEetan oinarrituta, barne-errepresentazio eredu bat proposatzen dute. Autore hauek ***composed-of*** lotura erabiltzen dute HAEa den *synset*a eta honen osagaien artean (ikus VI.1 irudiko c) atala). Beste hitz batzuetan esanda, *synset*a HAE bat bada, HAE hau bere osagaiei dagokion *synset*ekin lotuta egongo da *composed-of* harremanaren bitartez. 3. irudiko c) atalean, adibide gisa,

lo egin HAEa dugu. *Synset* hau, beste edozein *synset* bezala, bere hiperonimo (deskantsatu) eta troponimoei (siesta egin, kuluxka bat egin, hibernatu...) lotuta egongo da. Baina, honetaz gain, *synset*eko HAEa osatzen duen osagai bakoitzari (lo eta egin) dagokion *synset*arekin *composed-of* lotura bat izango du, HAEa den *synset*a beste bi *synset*ez osatua dagoela adieraziz.



VI.1 Irudia: HAEen barne-errepresentazio ezberdinak.

Euskal WordNeten *composed-of* harreman semantikoa erabiliko dugu, konposizionalki osatzen diren HAEen osagaiak errepresentatzeko aproposak iruditzen zaizkigulako. Hala ere, harreman honetaz gain, HAEa osatzen duten osagaien barne-errepresentazioa gehiago zehaz daiteke. Esate baterako, *composed-of* harreman semantiko honek ez du HAEen osagaien arteko harreman sintaktiko-semantikoa adierazten. Har dezagun umeak lo egin zuen esaldia adibide gisa, non aditz arineko egitura bat dugun: lo egin. Semantikoki, esaldi honetan *composed-of* harremanak ez du adierazten lo egin ekintzaren azpian lo egotearen egoera dagoenik. Sintaktikoki ere ez du adierazten HAUL honen osagai nominala (lo) hitz anitzeko aditz-esapidearen (lo egin) objektu

sintaktikoa denik. Hala, HAEko lo osagaia lo egin aditzaren objektua bada, honek rol tematiko bat hartuko du. Rol hau bi osagaien arteko harreman semantikoen bidez adierazita etorriko balitz, umeak lo egin zuen esaldiaren interpretazio sintaktiko-semantiko osoa genuke.

Nahiz eta WordNeten erlazio gutxi egon, *EuroWordNet*en orain erabilgarriak izan daitezkeen erlazioak definitu ziren (ikus IV.2 atala). Horien artean, kategoria desberdinetako *synset*ak lotzen dituzten harreman semantikoak zeuden: ***involved relation*** deiturikoak, hain zuzen ere.

> "The INVOLVED relation is used to encode data on arguments or adjuncts lexicalized within the meaning of a 2nd order entity."
>
> (Alonge *et al.*, 1998, 29. or.)

Harreman hauek lehenengo, bigarren eta hirugarren mailako entitateen arteko harremanak bideratzen dituzte. IV.2 atalean azaldu bezala, lehenengo mailako entitateak izen konkretuak dira; bigarren mailakoak ekintzak, prozesuak eta egoerak adierazten dituzten izen, aditz eta adjektiboak; eta azkenik, hirugarren mailakoak izen abstraktuak dira. *Involved* harremana aditz edo ekintza bat adierazten duen izen batetik abiatzen da, izen konkretu edo abstraktu batekin lotzeko. Adibidez, ingeleseko to hammer aditza hammer izenari lotuko zaio *involved instrument* harremanaren bidez.

Zortzi *involved* harreman mota daude: *agent, patient, instrument, result, location, direction, source direction* eta *target direction*.

Gure ustez, *involved relation* harremana barne-egiturak errepresentatzeko oso egokia da. VI.1 irudiko d) atalean, lo egin HAEaren errepresentazioa dugu non *composed-of* harremanaz gain, *involved relation* harremana ere erabiltzen dugun: lo HAEaren *gaia* (*involved patient*) da, eta honi esker jakin dezakegu *lo egiteko*, *lo egotea* beharrezkoa dela.

Harreman semantiko hauei esker, Euskal WordNeten ezagutza aberas daitekes: HAEaren osagaietako bakoitzari adiera emateaz gain, HAEak berak daraman informazio sintaktiko-semantikoari buruzko argibideak ere adierazten dira. Informazio hau guztia oso baliagarria zaigu LNPko hainbat atazatan, hala nola, itzulpen automatikoan eta adieraren desanbiguazioan.

Orain arte, Euskal WordNeteko HAEak diren izen eta aditzak dagozkien lexikalizazio-estatusarekin markatu ditugu; hots, lexikalizatu edo HAUL gisa, adierazpide sintagmatiko gisa eta hutsune lexikal gisa. Sailkapen hau VI.1 irudiko b) atalean dator adierazita. Kasu honetan, adierazpide sintagmatiko baten errepresentazioa dugu (*IXALEX*); ingeleseko to bike aditza euskaraz bizikletan ibili HAEaren bitartez adierazten dugu. HAE hau ez denez

hiztegi-sarrera eta ingelesekoren kategoria sintaktiko berarekin itzul daitekeenez, Euskal WordNeten adierazpide sintagmatiko gisa sartu dugu.

Gerora begira, ordea, HAEen barne-errepresentazioa adierazteari ekin nahi diogu VI.1 irudiko d) eredua jarraituta. Horretarako, dagoeneko eratorpenarekin erabili diren (Agirre eta Lersundi, 2001) metodo erdiautomatikoak erabiltzea pentsatzen dugu. Horrela, barne-egiturako *synset*ak eta beraien arteko harreman semantikoak automatikoki desanbiguatu ahal izango ditugu. Harreman berri hauei esker, MCRa informazio gehiagorekin aberastu ahal izango dugu. Gainera, kategoria desberdineko osagaiak dituzten HAEez gain, kategoria berdineko osagaiak dituzten HAEen osagaien arteko harremanak ere adierazi ahal izango ditugu.

VI.1 taulan Euskal WordNeteko datuez gain, Euskal WordNeten HAE mota bakoitzak dituen kopuruak ikus daitezke. Orain arte, izenek eta aditzek HAE kopuru antzekoa dute (2.935 eta 2.439, hurrenez hurren). Hala ere, gogoratu beharra dago aditzen garapena hasi baino ez dugula egin: Euskal WordNeteko izenen *synset*ak 28.705 dira, eta aditzena, berriz, 3.751. Hala, aditzekin HAE gehiago behar ditugula dirudi. Gauza bera esan dezakegu hutsune lexikal eta adierazpide sintagmatikoei buruz. Honen arrazoia ingeleseko hierarkiaren espezifikazio-maila izan daiteke, baina fenomeno honen berri VI.2.2 atalean emango dugu.

	Guztira	***Izenak***	***Aditzak***
Variant	50.670	41.160	9.510
Lema	26.565	23.069	3.496
Synset	32.456	28.705	3.751
Hutsune lexikal	2.499	2.198	301
Izen berezi	722	722	0
HAE	5.374	2.935	2.439
Adierazpide sintagmatiko	352	79	273

VI.1 Taula: Euskal WordNeteko datuak, eta HAE moten kopuruak.

VI.2 Bereizgarri hierarkikoak

V. kapituluan aipatu dugun bezala, EuroWordNeten garapena den MCR eredua aukeratuta, Euskal WordNeten garapena *expand approach* eta *merge approach* metodologietan oinarrituta egin zitekeen. Lehenengoan, euskarako ordainak, WordNeteko hierarkiari jarraituz, bertako *synset*ei zuzenean esleitzen zaizkie. Bigarrenean, aldiz, guk geuk sortu behar dugu euskarako adieren inbentarioa eta hierarkia, eta *Inter-Lingual-Index*ari (ILIari) lotu ondoren. Gure kasuan *expand approach* erabiltzearen alde egin genuen.

Bide bat ala bestea aukeratzeak kasuistika ezberdina ekar dezake. *Merge approach*ean oinarritutako wordneteko kontzeptuak ILIarekin lotzean, kontzeptualizazio-mailako arazoak ekar ditzake, hizkuntza horretarako egindako kontzeptuen sailkapena beste wordnetetako sailkapenarekin bat ez etortzea gerta daiteke, hau da, kontzeptuen diseinua era ezberdinean egin delako. Esate baterako, WordNeten dog izena ugaztun gisa adierazten da, hots, mammal *synset*aren hiponimo gisa sailkatua dago. Italierako wordnetak ere sailkapen hau egiten du cane izenarekin. Baina nederlanderako wordnetean hond izena, ugaztun gisa sailkatzeaz gain, konpainiako animalia gisa ere sailkatzen dute. Bai EuroWordNetek eta bai MCRk ezberdintasun hierarkiko hauek konpontzeko aukera eskaintzen dute. Hala, EBL eleanitzak izan arren, hizkuntza ezberdinen informazio elebakarrari ere garrantzia ematen diote, eleaniztasuna eta elebakartasuna uztartuz.

*Expand approach*ean oinarrituz gero, gertatzen diren hierarkia-bereizgarriak beste batzuk dira. Kasu honetan, WordNetaren sailkapen hierarkikoa jarraitzen denez, ingelesetik datorren hierarkia onartu egiten da, euskarako ordainak bertan txertatuz. Hala ere, euskarako ordainak ezin dira *synset* batean sartu *synset* horretako ingeleseko *variant* baten itzulpena izateagatik bakarrik; hasteko, adiera bera izan behar dute, eta gainera koherentzia bat mantendu behar da hierarkian. Horren adierazgarri (27) adibidea dugu.

(27) {associate} / {adiskide, lagun, kide} (who joins with others in an activity)
=> {ally, friend} / {aliatu, adiskide, lagun} (an associate who provides...)

Kasu honetan, {adiskide, kide, lagun} *synset*aren hiponimo gisa {aliatu, lagun, adiskide} ordainak ditugu. Lehenengo begiratuan, {aliatu, lagun, adiskide} *variant*ek *synset* horretan zuzenak dirudite, ingeleseko ally eta friendden baliokideak baitira. Baina hiru *variant*ak ez dira maila berekoak, lagun eta adiskide, aliatu baino orokorragoak dira. Hiperonimoari erreparatuz gero

({lagun, adiskide, kide}) {aliatu, lagun, adiskide} kontzeptuaren hiperonimoa dela ikusten dugu. Hala, gure susmoa egiaztatzen da: lagun, adiskide eta aliatu ez dira maila berekoak eta euskarako *synset*ak ez da hierarkiaren ordenarekin koherentea. Hiperonimia-hiponimian oinarritutako hierarkia izaki, honi ere erreparatu behar zaio euskarako *variant*ak itzultzeko momentuan, euskarako *synset*en sailkapena koherentea dela ziurtatuz. Hala, (27)ko hiperonimo-hiponimoaren adierazpen egokia (28) adibidean dakarkigu:

(28) {associate} / {adiskide, lagun, kide} (who joins with others in an activity)
=> {ally, friend} / {aliatu} (an associta who provides assistance)

Ikuspegi honetatik abiatuta, hierarkia euskaratzeak eragin ditzakeen bi kasu nagusienak aztertuko ditugu: hierarkia antolatzeko lexikalizaturik ez dagoen ordain bat asmatu behar denean (*kontzeptu antolatzaileak* deituko duguna), eta ingeleseko hiperonimo-hiponimo *variant*ak euskarako ordain berarekin lexikalizatzen direnean (*autohiponimia* bezala (Cruse, 2000) ezagutzen dena). Hala, bereizgarri hierarkikoak izan arren, lexikalizazioarekin oso lotuta daude: aurreko atalean (VI.1) *synset*-mailako lexikalizazioaz aritu gara, eta oraingoan WordNeteko antolakuntza hierarkikoak eragindako lexikalizazio-bereizgarriez.

VI.2.1 Kontzeptu antolatzaileak

Esan dugun bezala, *kontzeptu antolatzaile* deitzen diegu hierarkia antolatzeko asmatu diren kontzeptu orokorrei. Hierarkiaren goi-aldean egon ohi dira, eta beharrezkoak dira klase semantikoen sailkapenerako.

> "Unlike dictionaries in book format, WordNet contains short phrases, such as *bad person*, that are not paraphrasable by a single word. These phrases reflect lexical gaps and are a product of WordNet's relational structure, [...] that happens not to be lexicalized in English."
>
> (Fellbaum, 1998a, 6. or.)

Esate baterako, ikusmenaren bidez bereizten ditugun ezaugarri motak (kolorea, iluntasuna, ehundura...) multzokatzen dituen ingeleseko *synset*a visual property dugu. Kontzeptu hau ez dago lexikalizatuta; artifiziala da. Ikusmenezko ezaugarri motak adierazten duten *synset* guztiak batera jasotzen dituen klase-semantikoari izena emateko balio du (guztira 150 hiponimo).

(29) {color property} (an attribute of vision)
=> {texture} (the characteristic appearance of a...)
=> {lightness} (the visual effect of illumination on objects as...)
=> {dulness} (a lack of visual brightness)
=> {color} (a visual attribute of things that results from the...)
=> {achromatism} (the visual property of being without color)
=> {color property} (an attribute of color)
=> {...}

WordNetean salbuespen gisa zerrendatzen dira, EBL honetan hauek baitira ez-lexikalizatutako *synset* bakarrak, eta HAE bat behar dute hauen adiera adierazteko. Lexikalizazioari buruz aritzean, ikusi dugu Euskal WordNeteko hutsune pragmatikoak adierazpide sintagmatiko gisa (*IXALEX* gisa) ebatzi ditugula. Kasu honetan, nahiz eta ez-lexikalizatutako kontzeptuak izan, beste marka bat erabiliko dugu, hierarkiari dagokiola bereizteko: kontzeptu antolatzailean asmaturiko euskarako *variant* bat sartuko dugu eta *OROKORRA* marka jarriko diogu.

(30) **Synset-zenbakia:** 03871460
=> ***Synset*aren lexikalizazio-egoera:** lexikalizatugabea
=> **Glosa:** ikusmenak duen ezaugarria
=> **Sinonimoak:**
=> ikusmenezko ezaugarri (OROKORRA)

Horrela, kotzeptu sintagmatikoetatik bereizten ditugu. (30) adibidean ikusmenezko ezaugarri *variant*a dugu, eta *OROKORRA* markak adierazten du *synset* hori kontzeptu antolatzaile bat dela. Kontzeptu antolatzaileak lexikalizaturik ez dauden kontzeptuak direnez, *NOLEX* marka ere jarriko zaio. (31) adibidean kontzeptu antolatzaileen adibide gehiago dakartzagu:

(31) a. {psychological feature} → {ezaugarri psikologiko}
b. {representational process} → {irudikapen-prozesu}
c. {natural phenomenon} → {gertakari natural}

VI.2.2 Hierarkiak eta espezifikotasun lexikala

Ale lexikal polisemiko baten adierak elkarren hiperonimo/hiponimo izan daitezke, edota, beste hitz batzuetan esanda, hiperonimo-hiponimo harremana ale lexikal berarekin adieraz daiteke. Euskal WordNeten, esate baterako, hurrengo adibibidea dugu:

(32) {pertsona_1, gizabanako_1, lagun_15} (gizon-emakumeen multzoko bakoitza)
=> {adiskide_7, lagun_10} (ondo ezagutzen den pertsona)

Lagun_15 hiperonimoa da, adiera zabalagoa duena: 'pertsona' adiera duena; eta lagun_10 hiponimoa 'adiskide' adierarekin bakarrik erabiltzen da. Hala, ale lexikal berak bi adiera desberdin ditu, eta, gainera, bata bestearen hiperonimo-hiponimoak dira. Crusek (2000) polisemia mota honi *autohiponimia* deritzo:

> "Autohyponymy occurs when a word has a default general sense, and a contextually restricted sense which is more specific in that it denotes a subvariety of the general sense." (Cruse, 2000, 110. or.)

Aditzetan ere autohiponimia gerta daiteke: hiperonimoa eta hiponimoa diren bi *synset* forma berekoak izan daitezke, baina adiera desberdinekoak, hots, polisemikoak. Gainera, adiera ezberdintasuna azpikategorizazioan ere azalera daiteke:

(33) {abestu_4, kantatu_5} ("Jonek ondo abesten du")
=> {abestu_5, kantatu_7] ("Bertsoak abestu ditu")

Hiperonimoak (abestu_4) adiera orokorragoa du: 'ahotsez musika-soinuak egin'. 'Ahotsez musika-soinuak' abestu aditzaren barruan dagoen abesti izen orokorrak adierazten dituela dirudi (abestu aditzaren barruan dagoela, alegia), eta, ondorioz, oso arrunta da objekturik gabe geratzea sintaxian (Jonek ondo abesten du). Aldiz, bere hiponimoa 'abesti motak' edo 'abesti espezifikoak' onartuko dituen *abestu* izango da, 'musika-konposizioa' adieraziko duten objektuak (bertsoak, umetako abestiak, Eguberritako kantak...) hartzen dituena, alegia (Jonek bertsoak abestu ditu).

Hortaz, nahiz eta forma bereko hitzak izan, semantikoki desberdinak dira, eta hori hierarkiaren puntu desberdinean jarriz adierazten da.

Hala ere, Euskal WordNet ingeleseko hierarkian oinarrituta eraikitzen denez, autohiponimia *faltsua* sor dezakegu; alegia, gehiegizko autohiponimia. Egondako orrazketetan *synset*ak itzultzen joan ahala, ingeleseko bi adiera (edo gehiago) bazeuden eta euskaraz horietarako hitz bera erabiltzen bazen, autohiponimia baliatzen genuen beti (hiponimoak hiperonimoaren ordain bera), euskaraz adiera horiek benetan bereizten ziren kontuan hartu gabe. Aldiz, euskarako adierei erreparatuta, askotan, ez zegoen desberdintasun semantikorik. Hitzez hitzeko eskuzko orrazketarekin hastean (ikus V.2.2.2

atala), *synset*ak lantzeko garaian hierarkiari gehiago erreparatzen hasi ginen, eta orduan konturatu ginen euskarako hierarkian *synset* autohiponimoen kopurua ingelesekoan baino askoz ere handiago zela (euskaraz 4.500 autohiponimo genituen eta ingelesez 26 bakarrik). Desoreka honen arrazoiak aztertzerakoan, ingeleseko wordnetak duen espezifikotasun-maila xeheagatik zela konturatu ginen. (34) adibidean {merrymaking} *variant*aren hiponimoak ditugu:

(34) {celebration, festivity} (any festival or other celebration)
=> {merrymaking} (boisterous celebration)
=> {revel, revelry} (noisy partying)
=> {bout, spree} (a drunken revel)
=> {bender, bust} (an occasion for heavy drinking)
=> {carouse} (a merry drinking party)
=> {orgy} (a wild gathering involving drinking and promiscuity)
=> {whoopee} (noisy and boisterous revelry)

(35) adibidean Euskal WordNeteko editoreak emandako ordainak ditugu:

(35) {festa, jai} (zerbait ospatzeko antolatzen den ekitaldia edo jaia)
=> {parranda} (jai zaratatsua)
=> {parranda} (jai zaratatsua)
=> {parranda} (asko edanez egiten den jaia)
=> {parranda} (asko edanez egiten den jaia)
=> {parranda} (asko edanez egiten den jaia)
=> {orgia} (gehiegikeriak egiten diren jaia)
=> {parranda} (jai zaratatsu)
=> {...}

Hierarkia hauek erkatuz gero, ikusten dugu ingelesez, *synset* orokorrenetik zehatzenerainoko bidean, *synset* guztiak hiperonimoa ez den beste hitz batez lexikalizaturik daudela (merrymaking, bout, bender eta abar)[12].

Ingelesa ama-hizkuntza izan gabe, etengabe hiztegi elebidunetara — euskara-ingelesa (Morris, 1998) eta gaztelania-ingelesa (Oxford, 2003; Collins, 1998)— jo behar dugu *synset*en lanketarako. Kasu honetan celebration kontzeptuak edozein ospakizun adierazten du, horregatik egokitu zaizkio festa eta jai ordainak. Jai-moten artean 'jai zaratatsuak' ditugu, ingelesez merrymaking deritzona. *Morris Hiztegia*ren arabera, kontzeptu hau euskaraz parranda itzultzen da; gaztelania-ingelesa hiztegien arabera juerga edo

[12] Adibideko klase semantiko osoak 22 hiponimo ditu, baina adibidean merrymaking hiponimoaren hiponimo zuzenak bakarrik jarri ditugu. Gainera, espazio-arazoak direla-eta, *synset*etako *variant* kopurua ere txikitu dugu.

jolgorio gisa. Merrymakingek hiponimo bat dauka eta hiperonimoa bezalaxe (parranda) itzultzen da *Morris Hiztegia*ren arabera, eta juerga edo jolgorio gaztelania-ingelesa hiztegien arabera. Gauza bera gertatzen da revelen hiponimo gehienekin.

Hala, espezifikazio-maila xehea dela-eta, askotan, ingeleseko hierarkiako *synset* ugari hiperonimoaren ordain bera erabilita itzultzen dira. (34) eta (35) adibideetan argi eta garbi ikus daiteke fenomeno hau. Beraien hiperonimoa bezala itzultzen diren hiponimoak (revel, bout, bender, carouse, whoope eta abarri dagozkien itzulpenak) autohiponimotzat har genitzake: euskaraz hirurak hitz berarekin (parranda) adierazten ditugulako. Baina, euskaraz parranda ordainak kontzeptu hauetan guztietan adiera bera du.

Horrelako kasuetan, benetako autohiponimia autohiponimoa faltsutik bereizteko, hiponimo baxuenak (hiperonimoarekin itzultzen diren neurrian) *variant* gabe utziko ditugula erabaki dugu, hots, hutsune lexikal gisa utziko ditugu. Aipatu izan dugu, hutsune lexikal gisa uzten ditugula euskaraz ez ditugun kontzeptu kultural horiek (forties, simnel eta abar). Azaldu berri dugun kasu hau, antzekoa da baina kontzeptua adierazteko hiperonimoa dugu (eta ez azalpen bat): ingelesez hiperonimoaren espezifikazio bat da, baina euskaraz hiperonimoa eta bere hiponimoa maila berean ulertu eta itzultzen ditugu. Autohiponimo faltsuak hutsune kulturaletatik bereizteko, ingeleseko hitz hiponimoaren *synset*ean *ESPEZIFIKOA HIPERONIMOAZ* marka ezartzen dugu, eta era berean, lexikalizatugabea bezala (*NOLEX*). (36) adibidea ingeleseko revel *synset*aren euskarako baliokidea dugu:

(36) **Synset-zenbakia:** 00328944
=> ***Synset*aren lexikalizazio-egoera:** *NOLEX*
=> **Glosa:** jai zaratatsua
=> **Sinonimoak:**
=> - (ESPEZIFIKOA HIPERONIMOAZ)

Ingelesearen eta euskararen arteko espezifikotasun-mailen arteko aldea ikustearren, beste adibide bat aurkezten dugu:

(37) {vesell}/ {ontzi} (an object used as a container (especially for liquids)
=> {barrel} / {upel} (a cylindric container that holds liquids)
=> {butt} / ESPEZIFIKOA HIPERONIMOAZ
=> {hogshead} / {bukoi} (a large cask especially one...)
=> {keg} / {barrika} (small cask or barrel)
=> {firkin} / ESPEZIFIKOA HIPERONIMOAZ (a small barrel)
=> {tun} / ESPEZIFIKOA HIPERONIMOAZ (a large cask...)

(37) adibidean, upel moten sailkapen bat dugu. Berriro ere, ingelesez *synset* bakoitzeko lexikalizaturiko ordain bat dago, eta euskaraz, berriz, hiperonimoak (upel) balio digu kontzeptu horietako asko adierazteko. Hots, termino orokorrarekin nahikoa dugu termino espezifikoagoak adierazteko.

Beti ere, kontuan izan beharrekoa da, *synset* batek *ESPEZIFIKOA HIPERONIMOAZ* marka duen ala ez erabakitzeko, hiztegiak hartzen ditugula oinarri gisa. Euskara estandarizazio-bidean dagoen hizkuntza izanik, baliteke hiztegietatik kanpo kontzeptu hauentzat ordainen bat egotea, hainbat euskalki eta domeinuetako hitzak gure hiztegietara ez baitira heldu.

Bestalde, oroitu beharra dago Euskal WordNet aberasteko prozesua ingeleseko *synset*eta oinarrituz egin dela. Aztertu behar litzateke alderantzizko prozesua egingo bagenu zer neurritan gertatuko liratekeen antzeko kasuak ingeleserako. Dena den, gai honek azterketa sakonagoa mereziko lukeela iruditzen zaigun, eta beste tesi-lan bat izan daitekeela uste dugu.

Irizpide hau erabili ondoren, autohiponimo faltsuen kopurua 4.500etik 3.378ra murriztu da. Ingeleseko WordNet 1.6 bertsioan 41 autohiponimo daude, eta gaztelaniako wordnet 1.6 bertsioan 971. Lanean jarraitu ahala, kopuru hauek etengabe aldatuz doaz (ikus VI.2 taula).

	0.1 bertsioa	***0.2 bertsioa***
Euskal WordNet	4.500	3.378
WordNet	-	41
Spanish WordNet	-	971

VI.2 Taula: Autohiponimoen kopuruak.

Bestalde, WordNeten espezifikazio-mailak beste ondorio bat izan dezake euskarako hierarkietan: batzuetan, euskarako hiperonimoaren ordainarekin batera beste izen, adberbio, edota adjektibo bat ere hartzen dute *synset*ek kontzeptu hori adierazteko. (38) adibidean, vintage kontzeptua euskaratzeko hiperonimoari (ardo) izenlagun bat (erreserbako) gehitu behar izan zaio.

(38) {wine, vino} / {ardo} (fermented juice (of grapes especilly))
=> {vintage} / {erreserbako ardo} (a season's yield of wine from a vineyard)

Fenomeno hau, aditzetan oso nabaria da. Hauetan, hiperonimoa eta hiponimoa ordain bera izan ordez, gehiagotan gertatzen da hiponimoak hiperonimoaren ordainaz gain beste osagai baten beharra izatea, ingeleseko unitateak

barneraturik duen osagaia euskaraz aditzetik aparte adierazten delako. Aditzen hiperonimia-hiponimia erlazio hau zehatzago adierazteko, hiperonimia-troponimia terminoa erabiltzen da (ikus IV. kapitulua). Hau da, A1 aditza (hiponimoa) A2 aditza (hiperonimoa) era berezi batean egitea da. Esate baterako, 'herrenka ibiltzea' ibiltzeko era berezi bat da. (39) adibidean ikus dezakegu, ingeleseko troponimoentzat lexikalizatutako hitz bakarreko ordain bat dutela eta euskaraz HAE baten beharra dugula, askotan ez-lexikalizatua dirudiena (eta hiztegietan agertzen ez dena).

(39) {walk}/ {ibili} (advance by steps)
=> {lollop} / {baldar ibili} (walk clumsily and with a bounce)
=> {bumble} / {estropezu eginez ibili} (walk unsteadly)
=> {perambulate} / {noraezean ibili} (stroll)
=> {creep} / {behatz puntetan ibili} (to go stealthily)
=> {wade} / {uretan ibili} (walk through relatively shallow water)
=> {sleepwalk} / {lotan ibili} (walk in one's sleep)
=> {slink} / {isilean ibili} (walk stealthily)
=> {hitch} / {herrenka ibili} (walk impeded by some physical injury)
=> {skulk} / {inguruan ibili} (move stealthily)
=> {...}

HAE mota hauen errepresentazioa VI.1.4 atalean aipatu dugu, eta bertan esandakoari jarraituz, HAE hauek adierazpide sintagmatiko bezala lantzen ditugu. Hots, herrenka ibili Euskal WordNeten sartu egingo dugu adierazpide sintagmatiko gisa, nahiz eta hiztegi-sarrera bat ez izan.

Honenbestez, eta orain artekoa laburbilduz, argi dago Euskal WordNet garatzeko ingeleserako egindako hierarkia kontzeptuala jarraitzeak eraginak dituela: bi hizkuntzetako kontzeptuen sailkapena ez dator beti bat, ezta kontzeptu horiek lexikalizatzeko modua ere.

VI.2.3 Bestelako espezifikotasun lexikalak

Batzuetan WordNeteko espezifikazio-mailaren xehetasuna, hiperonimo-hiponimo ez diren *synset*en artean ere agertzen da, hots, hierarkiko harremanik ez duten *synset*en artean.

V. kapituluan aipatu dugu dagoeneko, WordNet granularitate xeheko EBLa dela. Hau da, WordNeten hiztegietan baino adiera gehiago agertzen dira, edo beste hitz batzuetan esanda, hiztegietako adierak adiera espezifikoagoetan banatzen dira. Adibide gisa, herri hitzaren adiera bat dakarkigu, 'jende multzoari' dagokiona. Adiera honek *Hiztegi Modernoa*n hurrengo definizioak ditu:

- Hainbat ohitura eta erakunde komun dituzten gizon-emakumeen multzoa, gehienetan taldean eta lurralde jakin batean bizi dena. *Munduko herri eta etniak. Herri kurdua*.

- Herri bateko kideen gehiengoa (maiz goi-klaseei, eliteari edo agintariei kontrajarririk erabilia).

- Unitate politiko bateko biztanleen osotasuna, botere politikoa datzaneko multzotzat hartua. *Herriak aukeratutako parlamentariak*.

Eta Euskal WordNeten herri hitzaren adiera horrek sei *synset* ditu. (40) adibidean sei *synset*ak aurkezten ditugu, beraien ingeleseko, gaztelaniako eta euskarako ordainekin:

(40)

Ing: {common people, folk}
Gazt: {plebe, vulgo, pueblo}
Eus: {herri, populu}
Glosa: biztanleen gehiengoa osatzen duen gizaki multzoa

Ing: {country, land, nation, nationality}
Gazt: {pueblo, nación}
Eus: {herri, nazio}
Glosa: jatorri bera duten nazio edo herrialde bateko biztanleak

Ing: {res publica, country, land, nation}
Gazt: {estado, país}
Eus: {herri, estatu, nazio, erresuma}
Glosa: enitate politiko bakarraren baitan dagoen gizaki multzoa

Ing: {public, world, populace}
Gazt: {pueblo, mundo}
Eus: {herri, mundu}
Glosa: pertsona multzoa osotasun gisa harturik

Ing: {people, multitude, mass}
Gazt: {masa, gente}
Eus: {herri, jende, masa, populu}
Glosa: herri xeheak osatzen duen multzo handia

Ing: {town, townsfolk, townspeople}
Gazt: {pueblo}
Eus: {herri}
Glosa: hiria baino txikiagoa den udalerri bateko biztanleria

Ing: {villate, settlement}
Gazt: {pueblo}
Eus: {herri}
Glosa: hiria baino txikiagoa den udalerri bateko biztanleria

Espezifikazio-maila dela-eta, batzuetan zaila egiten da *synset*en arteko desberdintasuna ikustea, batez ere, corpuseko agerpen errealak hauekin etiketatu behar direnean:

(41) Pinochetek eskualde honetako herriei egin dien kaltea konpontzen hasi da.
Herria nekatuta dago bete gabeko promesekin.
Herriak elkarrizketa eskatzen digu alderdiei.
Europako sindikatuek herrietan oinarritutako Europa soziala aldarrikatu dute.
Presoen auziari herri gisa eman behar zaio aterabidea.

Agerpen hauei (40)ko *synset* bakarra egokitzea lan zaila da, adiera askoren arteko muga lausoa delako. Gainera, testuinguruak ez badu laguntzen, *synset* bat baino gehiagorekin etiketatu daitezke, eta, ondorioz, anbiguoak izaten jarrai dezakete.

WordNeten granularitate finak ez du laguntzen LNPren hainbat atazetan, eta, batez ere, adieraren desanbiguazioan.

> "The granularity of word senses in current general purpose sense inventories is often too fine-grained, with narrow sense distinctions that are irrelevant for many NLP applications. This has particularly been a problem with WordNet which is widely used for word sense disambiguation (WSD)."
> (McCarthy, 2006, 17. or.)

Arrazoi horregatik, WordNeteko adierak elkartzeko hainbat saiakera egon dira: Milhacea eta Moldovan (2001), Tomuro (2001), Agirre eta Lopez de la Calle (2003). Guk ere bide hau jarraitzea erabaki dugu: antzeko adiera duten *synset*ak multzokatu ditugu eta corpuseko agerpenak *synset* horiekin guztiekin etiketatzen ditugu[13].

VI.3 Errepresentazioaren hedapena

Kapitulu honetan zehar, hainbat lexikalizazio-arazo aurkeztu ditugu eta hauei aurre egiteko irizpide batzuk proposatu ditugu. Irizpide hauek eraginda *synset*en errepresentaziorako EBLan marka edo ezaugarri berriak sortu ditugu. Hots, EBLa informazio gehiagorekin aberastu dugu. VI.3.1 atalean, marka hauek guztiak laburbilduta dakartzagu.

Bestalde, VI.1.4.1 atalean ikusi dugun bezala, HAEen barne-errepresentazio aberatsago baten proposamena ere egin dugu, non HAEaren barne-osagaiak harreman semantikoen bidez erlazionatzen diren. Hau VI.3 atalean laburki gogoraraziko dugu.

VI.3.1 Lexikalizazioaren errepresentazioari dagozkion markak

EuroWordNeten ereduari jarraituta, *synset* bat lexikalizatua dagoen ala ez markatu egiten dugu. Adibidez, (42) lexikalizaturiko kontzeptu bat da eta (43) ez.

(42) **Synset-zenbakia:** 06079949
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** pertsona multzoa osotasun gisa harturik
=> **Sinonimoak:**
=> mundu
=> herri

(43) **Synset-zenbakia:** 03871460
=> ***Synset*aren lexikalizazio-egoera:** NOLEX
=> **Glosa:** ikusmenak duen ezaugarria
=> **Sinonimoak:**
=> ikusmenezko ezaugarri (OROKORRA)

[13]Etiketatze semantikoari buruzko argibide gehiagorako jo bedi Agirre *et al.*-en lanera (2005b).

EuroWordNetek sortutako marka hauei, guk beste batzuk gehitu dizkiogu:

- ***PLU* marka:** kontzeptu bat adierazteko pluralezko ordaina erabiltzen denean, *variant* horri *PLU* marka erantsiko zaio.

(44) **Synset-zenbakia:** 03773162
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** Ebakitzeko tresna, erdialdean giltzatzen diren eta alde...
=> **Sinonimoak:**
=> guraize (PLU)

- ***FLEX* marka:** kontzeptu bat adierazteko flexio-atzizkia erabiltzen denean, *variant* horri *FLEX* marka erantsiko zaio.

(45) **Synset-zenbakia:** 01199751
=> ***Synset*aren lexikalizazio-egoera:** lexikalizatua
=> **Glosa:** Bero-gabeziak gorputzean eragiten duen sentsazioa.
=> **Sinonimoak:**
=> hotz (FLEX)

- ***IXALEX* marka:** Adierazpide sintagmatiko deitu ditugun HAEak markatzeko sortutako marka da. Honekin hiztegietako hiztegi-sarrerak ez diren HAEak baina Euskal WordNeten sarrera gisa sartu ditugunak markatzen ditugu. Horrela, hiztegi-sarrera diren HAEak hiztegi-sarrera ez direnetatik ezberdintzen ditugu.

(46) **Synset-zenbakia:** 01143604
=> ***Synset*aren lexikalizazio-egoera:** LEX
=> **Glosa:** elikagaiak jateko prestatu
=> **Sinonimoak:**
=> janaria prestatu (IXALEX)

- ***OROKORRA* marka:** kontzeptu antolatzaileei ezartzen zaien marka, hutsune kulturaletatik ezberdintzeko (ikus (43) adibidea).

- ***ESPEZIFIKOA HIPERONIMOAZ* marka:** Autohiponimo faltsuak hutsune kulturaletatik bereizteko sortutako marka da. Ingeleseko hitz hiponimoaren *synset*ean *ESPEZIFIKOA HIPERONIMOAZ* marka ezartzen dugu, hiperonimoa bezala lexikalizatzen dela adierazteko. Marka honekin batera, derrigorrezkoa da *synset*a ez-lexikalizatu bezala markatzea.

(47) **Synset-zenbakia:** 00328944
=> ***Synset*aren lexikalizazio-egoera:** NOLEX
=> **Glosa:** jai zaratatsua
=> **Sinonimoak:**
=> - (ESPEZIFIKOA HIPERONIMOAZ)

VI.3.2 HAEen barne-errepresentazio aberatsagoa

Bentivogli eta Piantak (2002), italierako wordneteko HAEetan oinarrituta, HAEen barne-errepresentazio eredu bat proposatzen dute: *composed-of* deiturikoa. Lotura hau erabiltzen dugu HAEa den *synset*a eta honen osagaiak lotzeko (ikus VI.1 irudiko c) atala).

Kategoria desberdinez osatutako HAEen osagaien arteko *synset*ak lotzeko EuroWordNeten *involved relation* erabiltzea proposatzen dugu: VI.1 irudiko d) atalean, lo egin HAEren errepresentazioa dugu non *composed-of* harremanaz gain, *involved relation* harremana ere erabiltzen dugun: lo (izena) HAEaren *gaia* (*involved patient*) da, eta honi esker jakin dezakegu *lo egiteko lo egotea* beharrezkoa dela.

PLU, *IXALEX*,*OROKORRA* eta *ESPEZIFIKOA HIPERONIMOAZ* markak ez bezala, HAEen barne-errepresentazioa adierazteko modu hau proposamena baino ez da. Hau da, oraindik ez dugu proposamen hau erabili, baina VI.1.4.1 esan bezala, etorkizunean Agirre eta Lersundiren (2001) metodo erdiautomatikoak erabiltzea pentsatzen dugu, barne-egiturako *synset*ak eta beraien arteko harreman semantikoak automatikoki desanbiguatu ahal izateko.

VI.4 Ondorioak

Kapitulu honetan, wordnet eleanitzekin lan egiteak hizkuntzen arteko ezberdintasunak gainditu beharra dakarrela erakutsi dugu. Gure kasuan, ingeleseko wordnetaren gainean lan egiteak ekartzen dituen ondorio batzuk aurkeztu ditugu. Alde batetik, lexikalizazioarekin zerikusia duten bereizgarriak ikusi ditugu, eta hitz-mailan eta hitz anitzeko esapideen mailan lexikalizatu eta ez-lexikalizatuen kasuistika zabala aztertu dugu. Azterketa horretan, argi geratu da lexikalizazioaren mugak lausoak direla, eta askotan lan zaila dela hitz bat edo hitz anitzeko bat lexikalizatua dagoen ala ez ebaztea. Lexikalizazioaren eztabaidak eragoztearren, eta LNPko atazen erabilgarritasunari

begira, VI.1.4 atalean zehaztu dugu Euskal WordNeten zer adierazpen mota txertatu behar genuen: lexikalizaturiko adierazpideez gain, *adierazpide sintagmatiko* deitu ditugunak Euskal WordNeten ere txertatzearen alde egin dugu, honetarako, hainbat irizpide eta marka proposatuz. Etorkizunean, landuko ditugun beste kategorien (adjektibo eta adberbioen) errepresentaziorako ere (hotzik/hotzez bezalakoak) balioko digu irizpide honek.

Honetaz gain, HAEen kasuan errepresentazio hau aberastu dugu HAEen osagaien barne-errepresentazio bat proposatuz: alde batetik, Bentivogli eta Piantaren (2002) *composed-of* harremana, eta bestetik, EuroWordNeteko *involved relation* harremana erabilita.

Bestalde, ingeleseko hierarkiak duen espezifikotasun maila handia dela eta, *synset*ak euskaratzean sortzen diren arazoei (hala nola, *autohiponimia faltsua* deitu duguna) aurre egiteko irizpideak eta markak ere definitu ditugu.

Honenbestez, abiapuntu gisa hartu dugun EBLa irizpide, marka eta errepresentazio berriekin aberastu dugula esan dezakegu.

VII. KAPITULUA

Euskal WordNet eta hautapen-murriztapenak

Kapitulu honetan, MCR eredua informazio gehiagorekin hedatzeko egin dugun lehenengo saiakera azalduko dugu. Ingeleseko eta euskarako kirol-arloko aditz batzuen objektuen eta subjektuen hautapen-murriztapenen azterketa deskribatuko dugu. Azterketa honetan, erabilitako corpusei, eskuratze-tekniken azterketari eta ebaluazio linguistikoari erreparatuko diegu batez ere. Esan beharra dago azterlan hau eleaniztasunaren hipotesiaren ikuspegitik egina dagoela. Hots, ingeleserako eskuratutako hautapen-murriztapenak euskaraz ere erabilgarriak izan daitezkeela frogatu nahi dugu. Horretarako, ingeleserako automatikoki eskuratu diren hautapen-murriztapenetan oinarritu gara lehenengo, gero hauek euskararentzat baliagarriak izan daitezkeen aztertu ahal izateko.

VII.1 Sarrera

III.1 atalean zehaztu dugun bezala, argi genuen gure EBLak hizkuntza bere osotasunean hartu behar zuela. Horretarako, ale lexikal bakoitza dagokion adierarekin, klase semantikoarekin eta informazio sintaktiko-semantikoarekin (rol tematikoak, azpikategorizazioa, hautapen-murriztapenak, funtzio gramatikalak, kategoriak, besteak beste) hornitzea da gure asmoa. Baldintza hauek kontuan hartuta, *WordNet, EuroWordNet* eta *The Multilingual Central Repository* (MCR) aukeratu ditugu eredu gisa (ikus III.3), eta honetan oinarrituta *Euskal WordNet* garatzeari ekin genion (lehendabizi izenak eta ondoren

aditzak). Izenen EBLen artean, WordNeten eredua ezaguna da eskaintzen duen informazio aberatsarengatik. Aditzen adierazpena, aldiz, behin baino gehiagotan esan dugun bezala, mugatua da, WordNeten azpikategorizazioa, hautapen-murriztapenak eta rol tematikoak bezalako informazio sintaktiko-semantikoa ez baita zehazten.

Gabezia honetaz ohartuta, WordNeten oinarritutako hurrengo ereduek (batez ere, MCRk) informazio sintaktiko-semantikoa txertatzeko aukera gehiago eskaintzen dituzte. IV.3 atalean esan dugun bezala, MCR ezagutza-baseak aditzen hautapen-murriztapenak kontsultatzeko aukera ematen du *Role* erlazio semantikoa erabilita. Hala ere, nahiz eta interfazeak hautapen-murriztapenak jasotzeko aukera izan, *Role* harreman semantiko hauek hutsik daude; hots, oraindik ez da informazio hau eskuratu eta EBLan txertatu.

Ikuspegi honetatik abiatuz, aditzen objektu/subjektuen hautapen-murriztapenen azterketan murgildu gara, Euskal WordNet informazio sintaktiko-semantikoarekin aberasteko asmoarekin. Hautapen-murriztapenak lortzeko abiapuntu gisa, beste batzuk egindako lana balia genezakeen —esate baterako, tesi-lan honetan aipatu ditugun hainbat lan eta formalismo (ikus III.3)—, edota euskarako corpusetan eta bestelako baliabide informatikoetan oinarrituz, guk geuk eskura genitzakeen.

Lehenengo aukeraren kasuan, kontuan izan beharrekoa da lan gehienak ingeleserako pentsatuak daudela, eta hauetan dagoen informazioa euskararako EBLan gehitu baino lehen, informazio hori hizkuntzatik independentea den (unibertsala den) edo behintzat euskararako baliagarria den frogatu beharko genukeela. Aukera honetan eskuzko lana ikaragarria litzateke. *LONGMAN Dictionary of Contemporanean English* (LDOCE)[1] lexikoian gehitutako hautapen-murriztapenak dira honen adibide. Baina esan beharra dago maila orokorreko hautapen-murriztapenak direla.

Bigarren aukera egingarriagoa da, eta hauxe izan da azken urteotan LNPn suspertu dena, hizkuntzen egitura eta ezaugarri asko eta asko corpusetatik eskura baititzake makinak. Baina, horretarako, garrantzitsua da corpus handiak izatea; zenbat eta corpus handiagoa izan, orduan eta informazio gehiago eta zehatzagoa lor daitekeelako. Hedapen urriko hizkuntzek (euskarak, esate baterako), aldiz, informatikoki balia daitezkeen corpus txikia dituzte; batzuetan txikiegiak horietatik emaitza zuzenak lortzeko. Hori dela eta,

[1] `http://pewebdic2.cw.idm.fr` (2007-07-02an atzitua).

beste hizkuntzetan dauden lanetako informazioa berrerabiltzeko eta hedapen urriko hizkuntzen baliabide falta konpontzearren, berriki, *MEANING: Developing Multilingual Web-Scale Language Technologies* (IST-2001-34460) proiektuarekin (Rigau *et al.*, 2003), ezagutza lexiko-semantikoaren eskuratzeari buruzko ikuspuntu berri bat sortu da: ezagutza lexiko **eleanitzaren** aberasketan oinarritzen dena. Hots, hizkuntza ezberdinetarako eskuratutakoa bata bestearekin parekatu eta hizkuntza batekin bestea aberastea ahalbidetzen duena[2]. Izan ere, hizkuntza batentzat eskuratutakoa beste hizkuntza batentzat baliagarria izan daiteke; eta, normalean, abiapuntu gisa, konputazionalki baliabide gehiago dituen hizkuntza bat hartzen da. Gaur egun, ukaezina da ingelesak arlo guztietan duen indarraz, eta arrazoi horregatik, hizkuntza honek euskarri informatikoan ere corpus handiena (edo handienetakoa) du. Hala, LNPren ikuspegitik, ingelesak oso baliabide aberatsak ditu, eta, ondorioz, aurrerapen gehienak ere hizkuntza honetarako garatzen dira. Hortaz, aipatutako eleaniztasunaren hipotesi berri honen arabera, jokabide linguistiko batzuk eleanitzak dira, eta, ondorioz, hizkuntza batentzat automatikoki eskuratutako datuak beste batzuentzat ere erabilgarriak izan daitezke. Adibidez, ingeleseko play aditzak ('instrumentu bat jo' adieran) objektu gisa musika-instrumentua adierazten duten izenak hartzen baditu (I play the piano), aditz horren euskarako ordainak ere (jo) izen mota horiek hartuko ditu objektu gisa (Nik pianoa jotzen dut). Hori horrela balitz —aztertu egin beharko da zenbateraino betetzen den fenomeno hau—, nahikoa litzateke makinak corpus aberatsenetatik informazioa eskuratzea (kasu honetan, play aditzaren adiera batek objektu gisa musika-instrumentuak hartzen dituela automatikoki eskuratzea). Honela, itzulpen-automatikoa egiterakoan adibidez, play aditza musika-instrumentuekin doanean, euskaraz jo bezala itzultzea lortuko genuke, bere hautapen-murriztapenean oinarrituz, hain zuzen ere.

MEANINGeko ikuspuntuari jarraituz, aditzen objektu/subjektuen hautapen-murriztapenen azterketarekin batera, eleaniztasunaren hipotesia aztertzeari ekin diogu, hizkuntzen artean egon daitezkeen aldaera eta parametroak kontuan hartuaz. Horrela, kapitulu honetan hautapen-murriztapenen azterketa automatikoaz arituko gara. Horretarako, ingeleserako automatikoki eskuratu diren hautapen-murriztapenetan oinarritu gara lehenengo, gero hauek euskararentzat baliagarriak izan daitezkeen aztertu ahal izateko. Hau da, ingeleseko hautapen-murriztapenak eskuratzeko erabili diren tekni-

[2]Proiektu honi buruzko informazio gehiago, Pocielloren lanean (2004b).

ka ezberdinak aurkeztu eta ebaluatu ditugu, hauen aplikazioa eleanitza izan daitekeela frogatu nahian, gerora, Euskal WordNeten txertatu ahal izateko. Azterketa honen ondoren, ingeleserako erabilitako eskuratze-teknika bat euskarako corpus batean erabili dugu, ingeleseko emaitzekin erkatzeko.

Azterketa hau mugatzearren, gure ustez kirol-domeinuan gehien agertzen diren aditz batzuetan oinarritu gara (jokatu, entrenatu, irabazi, galdu eta berdindu). Bestalde, MCR adiera-inbentario gisa erabili dugu, bertan ingeleseko eta euskarako aditz-adierak lotuak datozelako. Beraz, aditz hauen MCR-ko kirol-adieratik abiatuz ingeleseko itzulpenak lortu ditugu. Horrela bada, azterketa honen parametro nagusiak domeinua eta adierak dira, kirol-domeinuarekin bat datozen aditzen adieren hautapen-murriztapenak aztertu eta eskuratu ditugulako.

Hala, laburbilduz, kapitulu honetan azalduko dugun azterketaren helburuak hurrengoak dira:

- Hainbat eskuratze-teknika erabiliz ingeleseko eta euskarako corpus ezberdinetatik eskuratutako hautapen-murriztapenak aztertzea eta konparatzea.

- Hautapen-murriztapenak eleanitzak izan daitezkeen aztertzea.

Azterketa hau hastapenekoa da; emaitzak ez dira behin betikoak. Lan honetatik abiatuta, euskararako jorratzen hasiberriak garen hautapen-murriztapenen arlo hau garatu nahi dugu, emaitzarik egokienak eskaintzen dizkigun bidea aurkituz.

Azkenik, esan behar dugu azterlan honetan eskuratze-tekniketatik lortutako emaitzekin egin dugula lan, hau da, emaitzen ebaluazio linguistikoan aritu gara. Horregatik, txosten honetan ez dugu sakonduko eskuratze-teknika hauek garatzeko erabili diren hainbat prozesu eta algoritmo informatikoetan[3]. Alderantziz, azterketa honen ondorioz, informatikariek aditzen informazio lexikoa aztertzeko baliabideak hobetzeko aukera izango dute.

Tesi-txosten honen sarreran (VII.1 atalean) hautapen-murriztapenen ezaugarri eta erabilerari buruzko informazioa eman dugunez, kapitulu honetan eskuratze-automatikoaz jardungo gara. Dena den, hautapen-murriztapenen izaera eta erabilerari buruzko azterketa sakonagoa Pocielloren (2004a) lanean dago ikusgai. Kapitulu hau sei atal nagusitan banatzen da. Sarrera honen

[3]Horien berri izateko jo bedi hurrengo lanera: Agirre eta Martínez (2002).

ondoren, VII.2 atalean, hautapen-murriztapenen eskuratzearen inguruan jardungo gara. VII.3 atalean, azterlan honetan erabili diren baliabideen berri emango dugu (corpusak eta eskuratze-teknikak). VII.4 eta VII.5 ataletan ingeleseko eta euskarako hautapen-murriztapenen azterketan sakonduko dugu. Eta, azkenik, VII.6 atalean, lanaren ondorioak eta etorkizuneko lanak aipatuko ditugu.

Kapitulu honetan zehar, jokatu/play aditzak erabiliko ditugu adibide gisa saiakera honen xehetasun guztiak emateko, baina C eranskinean aditz guztien hautapen-murriztapenak eta beraien ebaluazioa zehaztuta datoz.

VII.2 Hautapen-murriztapenak eta hauen eskuratzea

Hitz batek, honek duen adieraren arabera, testuinguruan har ditzakeen osagai linguistikoak murrizten ditu hautapen-murriztapenak (aurrerantzean, HM). Beste hitz batzuetan esanda, HMak dira **hitz baten adiera batek** testuinguruan izan ditzakeen agerkidetzak. Zerrenda hau osatzen dute klase semantiko batean dauden hitzek, hau da, adiera zehatz batekin osagai gisa ager daitezkeen hitz guztiak.

Horrela bada, aditz batek, bere adieraren arabera, argumentu bezala har ditzakeen izenen klase semantikoa mugatu dezake. Adibidez, idatzi aditzak, subjektu gisa [+gizaki] tasuna eskatzen du; [+gizaki] izango da bere subjektu HMa, alegia[4].

VII.2.1 Eskuratze-metodoak

LNPn, HMak eskuratzeko garaian, hiru metodo dira aipagarrienak: lehenengoa, *introspekzioa*; bigarrena, *hiztegietan oinarrituriko eskuratze automatikoa*[5]; eta, azkenik, *corpusetan oinarrituriko eskuratze automatikoa*.

VII.2.1.1 Introspekzioa

HMak eskuratzeko introspekzioa erabiliz gero, HMak eskuz sortzen dira, hizkuntzalariaren iritzi eta intuizio linguistikoen arabera. Eskuratze-metodo hau izan da erabiliena orain dela hamarkada bat arte (Lenat eta Guha,

[4]HMei buruzko argibide gehiagorako jo bedi Pocielloren lanera (2004a).

[5]Ingelesez *automatic acquisition from machine-readable versions of dictionaries* (MRD).

1990). Pertsonen intuizioetan oinarritzeak baditu bere arriskuak: egindako lana hizkuntzalariaren subjektibotasunaren mende egongo da, baita honen akats, ahazte, eta kontraesanen mende ere. Bestalde, eskuratze-mota honek eskuzko lan handia eskatzen du, eta datu-kopuru bera edo handiagoa lortzeko badaude beste metodo azkarrago batzuk.

Arrazoi hauengatik, gaur egun, LNPn metodo hau alde batera geratu da. Haatik, introspekzioa eskuratze-metodo gisa guztiz *fidagarria* izan ez arren, automatikoki eskuratutako HMak ebaluatzeko erabiltzen da. Gu geu, saiakera honetan, introspekzioaz baliatu gara eskuratutako emaitzak ebaluatzeko[6].

VII.2.1.2 Eskuratze automatikoa hiztegietatik

Lexikografikoak hiztegian hiztegi-sarrera bat definitzerakoan, sarrera horrek hartzen dituen HMen azterketa eta adierazpena egiten du. Hiztegi hauek informatikoki baliagarriak direnean, makinak hiztegi hauetatik bertatik erauz ditzake lexikografoak hiztegi-sarrera bakoitzari egokitu dion HMa (Montemagni, 1994).

Hala ere, metodo honen bidez lortutako HMak ez dira guztiz fidagarriak, pertsonen intuizioetan oinarritutako hiztegiak baitira hauek ere, eta gorago esan dugun bezala, honek bere alde txarrak dauzka: objektibotasun falta eta eskuzko lan handia, adibidez.

Bestalde, hiztegietatik informazio interesgarria lor daitekeen arren, hiztegietako sarrera guztiek ez dute HMak erauzteko adina informazio ematen, informazio hori ez delako esplizituki agertzen hiztegi-sarrera guztietan.

VII.2.1.3 Eskuratze automatikoa corpusetik

Metodo honen bitartez makinak automatikoki eskura ditzake hitz bati dagozkion HMak, hitz horrek corpusean dituen agerpen guztien testuinguruan oinarrituz.

Metodo hau da eskuratze automatikorako adostasun handiena lortu duena, ondoko arrazoiengatik:

- Corpusen tamaina handiari esker, aztertu beharreko hitzaren adibide nahikoak eskuratu ahal izango ditugu.
- Corpusa domeinuka dagoenean, domeinu zehatz bati dagokion informazio linguistikoa eskuratzeko aukera izango dugu.

[6]Honi buruz, VII.4.1 eta VII.5.1 ataletan mintzatuko gara.

- Hiztegiek ez bezala, eskuratutako datuen maiztasuna ere eskaintzen digu.

Guk egindako saiakerak ere corpusak hartu ditu ardatz gisa.

VII.2.2 Formalizazioa

Atal honetan, corpusean oinarritutako eskuratze-metodoan erabiltzen diren eskuratze-teknika nabarmenenei buruz jardungo gara: *hitzean oinarritzen direnak* eta, *klase semantikoan oinarritzen direnak*[7].

VII.2.2.1 Hitzean oinarritzen diren eskuratze-teknikak

Ikerlari batzuk (Hindle, 1990; Church *et al.*, 1991; Hindle eta Rooth, 1991; Pereira *et al.*, 1993, esate baterako) predikatu eta argumentu baten arteko harreman semantikoak atzitzeko, hitzean bertan oinarrituriko saiakuntzak egin dituzte. Hurbilpen hau semantika berdintsua duten hitzek testuinguru berdintsuetan agertzeko duten joeraz baliatzen da.

> "[...] the lexical relationships between given words are modeled by analogy with other words that present a similar distribution in the training corpus." (Ribas, 1995, 7. or.)

Harreman linguistiko askok semantikoki parekoak diren hitzak eskatzen dituzte. Hala, adjektibo batek ezin ditu nahi adina izen modifikatu, izenaren klase semantikoaren arabera murriztuko baititu bere osagaiak. Adibidez, goxo adjektiboak, bere adiera hedatuenean ('zapore onekoa', hain zuzen ere), bere ondoan, osagai gisa janaria edo edaria izango du beti. Horrela bada, teknika hauek hizkuntzak eskaintzen dizkigun distribuzioaz baliatuko dira HMak eskuratu ahal izateko.

Hindlek (1990), adibidez, izenen arteko antzekotasuna neurtzeko teknika hau landu zuen, corpuseko aditz, subjektu eta objektuen distribuzioari begiratuz. Aditz baten subjektu/aditza eta objektu/aditza bikote-agerkidetzak estatistikaren arabera neurtu zituen, *co-occurrence score* delakoarekin (*mutual information*en parekoa)[8]. Honela, izenen arteko antzekotasuna neurtzeaz gain, aditz baten argumentu gisa agertzen diren izenen zerrenda lortzen du agerkidetza altuenetik baxuenera.

[7] Ingelesez, *word-based* eta *class-based*, hurrenez hurren.

[8] "Mutual information, $I(x; y)$, compares the probability of observing word x and word y *together* (the joint probability) with the probabilities of observing x and y *independently*." (Church *et al.*, 1991, 118. or.).

Co-occurrence score	*verb*	*object*
11.75	drink	tea
11.75	drink	Pepsi
11.75	drink	champagne
10.53	drink	liquid
10.20	drink	beer
9.34	drink	wine
7.65	drink	water

VII.1 Taula: Drink aditzaren objektuak hitzen hurbiltasunean oinarritutako teknika erabiliaz (Hindle, 1990).

VII.1. taulak, drink aditzarekin maizen gertatzen diren objektu/aditz bikoteetako batzuk erakusten ditu, *co-occurrence score*aren arabera zerrendatuta. Hain zuzen ere, objektu/aditz bikote hauexek dira "zer edan daiteke?" galdera erantzuten dutenak.

Hala eta guztiz ere, Hindlek lortutako HMak oraindik mugatuak dira. Azken batean, aztertzen ari garen hitzaren ezaugarri lexikoak, hitz-zerrenda batek adieraziko ditu; hau da, ez ditu inolako etiketa edo tasun semantikoren bidez biltzen. Horrela bada, hitzaren agerkidetzan oinarritzeari jarri zaion eragozpenetako bat, honen zorroztasun falta izan da. Ribas-ek halaxe azaltzen du:

> "[...] it is by no means obvious that the distribution of words will directly provide a useful semantic classification, at least in the absence of considerable human intervention, and especially for low-frequency words." (Ribas, 1995, 17. or.)

Haatik, hurbilpen honek beste bi arazo ekartzen ditu:

- Hitzean oinarritutako teknikek lortzen dutena hitz-formak dira eta ez hitz-adierak, azken hauek direlarik semantikan hautapenak zehazten dituztenak. Hitzen adieren desanbiguazioa, adibidez, ezinezkoa litzateke hitz-formetan bakarrik oinarrituz gero.

- Lortutako HMak corpusean gertatu diren agerpenetara bakarrik mugatuko dira: hau da, corpusetik at dauden antzeko adibideentzako ezingo dira orokortu.

VII.4.2.1. atalean ikusiko dugun bezala, gure azterketan honen antzeko teknika bat erabili dugu, baina ez eskuratze-teknika bezala, baizik eta eskuzko lanerako baliabide bezala.

VII.2.2.2 Klase semantikoan oinarritzen diren eskuratze-teknikak

Teknika hauek klase semantikoak baliatzen dituzte bi hitzen arteko HMa adierazteko. Klase semantiko bat ezaugarri komunak dituzten hitzek osatzen dute, eta, normalean, hierarkikoki antolatuta daude. Zenbait autorek, Grishman eta Sterling-ek (1992) esaterako, eskuz egin dituzte klase semantikoak; beste zenbaitek, berriz, zailtasunak ikusita, egina dagoen ezagutza semantiko bat hartzen dute oinarri gisa: Resnik-ek (1993), esate baterako, WordNet darabil. Azterlan honetan guk ere horixe egingo dugu: MCR edo Euskal WordNet erabiliko dugu eskuratze-teknika mota hau aplikatzeko. Hala, eskuratze-teknika honen azalpenerako, WordNet ereduak eskaintzen dituen klase semantikoetan oinarrituko gara.

Behin hitz batek (adibidez, drink aditzak) corpusean dituen osagai posibleak lortu ondoren (ikus VII.1. taula), osagai horiei dagozkien *synset*ak bilatzen dira WordNeten, gerora, *synset* horiek guztiak multzokatzen dituen hiperonimo *synset*a (klase semantikoa) eskuratzeko. Eta hiperonimo horixe izango da aditz horren hautapen-murriztapena. VII.1. taulako *synset*en klase semantikoa (1) adibidean dator:

(1) => {liquid}
=> {beverage}
=> {milk}
=> {alcohol}
=> {wine}
=> {beer}
=> champagne}
=> {...}
=> {soft drink}
=> {Pepsi}
=> {...}
=> {juice}
=> {tea}
=> {...}

Ikus daitekeen bezala, alcohol *synset*ak wine, champagne eta beer multzokatzen ditu, edari alkoholdunen klasea sortuz; Pepsi, aldiz, soft drink *synset*aren azpian dago, freskagarriak diren edarien klasearen azpian[9]. Baina ez alcohol *synset*ak, eta ezta soft drink *synset*ak ez dituzte VII.1. taulako

[9] Adibide honetako edarien hierarkia ez dago bere osotasunean. Hierarkia osoa WordNeten dago ikusgarri: `http://www.wordnet.princeton.edu` (2007-07-02an atzitua).

Association score	*verb*	*object classes*
3.58	drink	**beverage** [beverage, drink, drinkable, potable]

VII.2 Taula: Drink aditzaren objektu hautapen-murriztapena, WordNet eta klase semantikoan oinarritutako teknika erabiliz (Resnik, 1992).

edari mota guztiak multzokatzen, eta denak multzokatzen dituena behar dugu: beverage, alegia. Beste hitz batzuetan esanda, beverageren azpian dauden *synset* guztiak (hauei dagozkien hitz guztiekin, noski) ezaugarri semantiko komunak izango dituzte ([+edangarri]), eta, ondorioz, agerkidetza sintaktiko bera izango dutela suposatzen da; adibidearekin jarraituz, guztiak drink aditzarekin ager daitezke. Honenbestez, [+edangarri] tasuna edo klase semantikoa (beverage) izango da drink aditzaren HMa.

Resnikek (1993) teknika hau erabiltzen du, WordNeten hierarkia kontzeptualean eta *association score*[10] neurri estatistikoan oinarrituaz. Ondorioz, bere hautapen-murriztapenek VII.2. taulakoen antza dute. Hitzean oinarritzen diren teknikekin ez bezala, klase semantikoa ez da adierazten hitz-zerrenda baten bidez (ikus VII.1. taula), baizik eta klase semantiko horren azpian dauden hitz guztiak multzokatzen dituen *synset*aren bidez: VII.2 taulako beveragen bidez, adibidez.

Klase semantikoan oinarritutako teknikek dituzten abantailak, aurkeztutako beste hurbilpenarekin erkatuz gero, hurrengoak dira:

- Nahiz eta corpus txikia izan, esanguratsuak izan daitezkeen datu estatistikoak lor daitezke.
- Corpusean lortutako HMek, bertan azaltzen ez diren adibideentzako ere balio dute.
- Klase semantikoek eskuratutako HMen interpretazioa errazten dute.
- Klase semantikoak hierarkikoki antolatuta egoteak HM orokorrak lortzen laguntzen du.

[10] "The association score takes the mutual information between the verb and a class, and scales it according to the likelihood that a member of that class will actually appear as the object of the verb." (Resnik, 1992, 328. or.)

Dena den, eskuratze-teknika mota honek desabantailak ere baditu:

1. Klase semantikoen bidez tasun semantikoak adieraztea ez da beti zuzena, batzuetan ez baitatoz bat. Adibidez, [+edangarri] tasunak modu egokian adierazten du WordNeteko beverageri dagokion klasea. Baina ez da beti posible tasun semantikoari dagokion klase semantikoa topatzea. Esate baterako, ireki aditzak irekitzen diren gauzak behar ditu argumentu gisa (kaxak, paketeak, poteak eta abar). Eta irekitzen diren gauzak zer klase semantikoren barnean daude? Horrelakoentzat, tasun zehatz bat ezartzea nahiko zaila da; irekitzen diren gauzen kasuan, WordNeten container (something that holds things) *synset*a jo daiteke, behar bada, klase semantiko aproposena bezala.

2. Batzuetan, klase semantikoaren barnean tasun semantiko hori ez duten *synset*ak ager daitezke. Esaterako, hegazti klase semantikoak gehienetan [+hegan] tasuna eskatzen du, baina klase honetan hegan egin ezin dutenak ere badaude: pinguinoa eta oiloa, adibidez, hegan egin ez arren, hegaztiak dira. Horrelako salbuespenen errepresentazioa arazo bat da, eta arazo hau adimen artifizialean ezaguna den arren, ez du berehalako ebazpenik. Konponbide posible bat klase semantikoaren tasun bera daramaten kontzeptu guztiak multzokatzea izan daiteke.

VII.3 Baliabideak

Sarreran aipatu dugun bezala, azterlan honen helburu nagusiena honako hau da: corpus eta eskuratze-teknika desberdinak erabiliz, ingeleseko kirol-aditz batzuentzat automatikoki eskuratutako HMak aztertzea, gero hauek euskararentzat baliagarriak izan daitezkeen ikusi ahal izateko. Horrela, ikerlan honetan ondorengo ataza hauek egin ditugu:

- **Ingeleseko aditz batzuen HMenak lortzeko erabili diren eskuratze-teknika automatikoen emaitzak hartuta, hauen azterketa eta ebaluazioa egin teknika bakoitzaren alderdi on eta txarrak aipatuz.**

Beste era batera esanda, HMen eskuratze-teknika desberdinen ebaluazio bat egin dugu, eta, honetarako, bi parametro hartu ditugu kontuan: **domeinua eta adiera**.

Domeinuak azken urte hauetan garrantzi handia hartu du. Hasieran HMak aditzen adierentzat definitu baziren ere (Wilks, 1973), lehenengo ahalegin automatikoetan aditz formetara mugatu ziren (Resnik, 1993). Geroago, aditzen adierak kontuan hartzen dituzten eskuratze-teknikak proposatu dira (Agirre eta Martínez, 2002; McCarthy, 2001). Gaur egun, HMen eskuratzea domeinu zehatz bati buruz aritzen diren corpusetara mugatzen hasi dira, aditzaren adiera eta bere HMena corpusaren domeinutik lortu daitekeela ikusi dugu (Agirre *et al.*, 2003b; McCarthy, 2001).

Gure azterketan ere bide hau jarraitu dugu, eta bi corpus mota erabili ditugu: kirol-domeinuarekin harremanetan daudenak eta domeinu zehatzik ez dutenak; hauetatik lortutako HMak parekatzea interesgarria iruditu zaigulako.

Adierari dagokionez, eskuratze-teknika batzuk aditzaren HMak eskuratzen dituzte aditz-adiera kontuan izanda, eta beste batzuk, aldiz, aditzforman oinarritzen dira. Eskuratze-teknika hauen arteko aldean ere sakonduko dugu.

- **Ingeleseko aditzentzat eskuratze-teknika bakoitzetik lorturiko HMak euskarako ordainen HMak izan daitezkeen aztertzea, bi hizkuntzetarako egokiak diren ala ez, hots, HMak eleanitzak izan daitezkeen ala ez egiaztatzeko**.

Beraz, ingeleserako lortu diren datuak euskaraz berrerabili ditugu, eta berrerabilera hau egokia den ala ez aztertu dugu. Honetarako, MCRz baliatu gara, bertan ingeleseko ordain bakoitza euskarakoarekin lotua baitator.

- **Ingeleserako erabilitako eskuratze-teknika batzuk euskarako corpus batean erabili (a) eta (b)ko emaitzekin erkatzeko.**

Ingeleseko corpusetik lortutako HMak eta euskarako corpusetik lortutakoak konparatzea, alegia. Hemen ere, kirol-domeinuari dagozkion corpusak eta corpus orekatuak erabili ditugu, beraien artean zer desberdintasun agertzen diren aztertzeko.

Kapitulu honetan jokatu aditza erabiliko dugu saiakeraren metodologia eta garapena azaltzeko[11], baina aipatutako aditz guztiekin egin dugu azterlan bera[12].

[11] VII.4 eta VII.5 ataletan saiakera hau urratsez urrats aipatzen badugu ere, Pociello (2004a) lanean urrats bakoitzari buruzko xehetasun gehiago datoz.

[12] Aditz guztiekin jasotako emaitzak C eranskinean datoz.

Hurrengo ataletan saiakera hau egiteko beharrezkoak izan diren corpusez (VII.3.1 atala) eta eskuratze-teknikez (VII.3.2 atala) jardungo gara.

VII.3.1 Azterketarako erabili diren corpusak

HMak ondorengo corpusetatik lortu ditugu:

VII.3.1.1 Ingeleseko corpusak

- ***SemCor:*** Ingeleseko corpus hau (Fellbaum *et al.*, 2001) semantikoki eskuz etiketatutako corpusik handiena da. Semantikoki etiketatuko corpusa dela adierazten dugunean, hitzen adierak dagokien adierarekin desanbiguatuta daudela esan nahi dugu. Hala, corpus bat *(semantikoki) etiketatua* dagoela diogunean, *(semantikoki) desanbiguatutako* corpus bat dela adierazi nahi dugu. *Brown Corpus*aren zati batez eta Stephen Craig-en *The Red Badge of Courage* eleberriaz osatuta dago eta 350.000 hitz inguru ditu. Corpuseko hitz bakoitza WordNeteko *synset* batekin desanbiguatuta dago, eta arrazoi honengatik LNPn oso erabilia izan da.

- ***The British National Corpus* (BNC):** BNC 100 milioi hitzetako corpus orekatua da, hots, jatorri ezberdinetako corpusekin osatutakoa, baina eskuz etiketatu gabea.

- ***EFE*:** EFE agentziaren corpusa da, 70 milioi hitz baino gehiago dituena. Kazetaritzari dagokion corpusa da eta kazetaritzaren gaien edo domeinuen arabera antolatua dago. Horregatik, domeinu zehatz bateko agerpenenak kontsultatzeko oso lagungarria da, baina ez dago eskuz etiketatuta.

VII.3.1.2 Euskarako corpusa

- *Euskaldunon Egunkaria*: Egunkari honetako berriekin osatutako corpusa da, 7 milioi hitz inguru dituena. EFEren antzera, corpus domeinuka antolatuta dago. Hala, euskarako hitz baten testuingurua corpus osoan zehar ala domeinu zehatz batean kontsulta daiteke. Orain ari gara, *EuSemcor* proiektuaren baitan (Agirre *et al.*, 2006a), corpus hau eskuz desanbiguatzen Euskal WordNeteko *synset*etan oinarrituta. Proiektu

hori amaitu gabe dagoenez, saiakera honetan eskuz etiketatu gabeko bertsioa erabili dugu.

VII.3.2 Azterketarako erabili diren eskuratze-teknikak

Azterlan honetan klase semantikoan oinarritzen diren eskuratze-teknikak erabili dira (ikus VII.2.2.2. atala) eta MCR baliatu dugu klase semantiko horiek adierazteko. Horrela bada, eskuratze-teknika hauek aditzen objektu/subjektuen HMak adierazteko MCRko klase semantikoak darabiltzate. Hala ere, teknika honen barruan aldaerak egon daitezke. Gu lau eskuratze-teknika ezberdinez jardungo gara, bi multzo nagusitan banatu ditugunak hauen azalpena ulergarriagoa egin ahal izateko:

- *Synset* batekin adierazitako HMak.
- Domeinu-eremu semantiko bikote batekin adierazitako HMak.

VII.3.2.1 *Synset* batekin adierazitako HMak

Mota honetako eskuratze-teknikek aditz baten HMak ***synset*** batez adierazten dituzte, *synset* hau klase bezala kontsideratzen dutelarik; hau da, *synset*a bera eta honen hiponimo guztiak izango dira aditz horren objektu/subjektuen HMak.

Aditzari dagokionez, ikuspuntu ezberdinetik landu daiteke, eta hori izango da multzo honetako eskuratze-teknikak ezberdinduko dituena.

Aditzaren HMak eskuratzean, HM hauek aditzaren adiera guztientzako izan daitezke, **aditz-formarentzat**, alegia. Demagun irabazi aditz-forma dugula. Aditz honek adiera ezberdinak ditu ('lehiaketa irabazi', 'dirua irabazi' eta abar). Kontuan izanda eskuratze-teknikak irabazi aditzaren HMak eskuratzean aditz horrek izan ditzakeen adiera guztietan oinarritzen dela, aditz horren edozein adierari dagokion HMak eskura ditzake: objektuaren kasuan, [+lehiaketa] edo [+jabetza], esate baterako.

HMak aditzaren adiera bakarrarentzat ere lor daitezke, **aditz-adierarentzat**, alegia. Adibidez, irabazi aditzaren objektu HMak eskuratzerakoan, eskuratze-teknikak aditz-forma honen adiera bakarra har dezake kontuan[13]

[13]Corpusa etiketatua badago, eskuratze-teknikak zuzenean hartzen du corpusetik adiera hori. Bestela, hitzen adieren desanbiguazioan erabiltzen diren teknikak erabili behar dira. Argibide gehiagorako jo bedi Agirre eta Martínezen lanera (2002).

(adibidez, 'lehiaketa irabazi' kirol-adiera). Hala, eskuratze-teknika honek adiera horri bakarrik dagozkion objektuen HMak eskuratuko ditu: [+lehiaketa], [+kirola], eta abar.

Aditz-forman oinarritzen den eskuratze-teknikari ***word-to-class*** (aurrerantzean, w2c) deritzo, eta aditz-adieran oinarritzen denari ***class-to-class*** (aurrerantzean, c2c)[14]. Izenak adierazten duen bezala, w2c teknikak hitzetik abiatuta (aditz-formatik) klaseak diren HMak lortzen ditu; c2c-ek, aldiz, aditz-klase batetik abiatuta klaseak diren HMak lortzen ditu.

HMak adierazteko *synset*a darabilten eskuratze-teknika hauen ezberdintasun nagusiena azaldu ondoren, HM hauek eskuratzeko jarraitzen diren urratsak eta irizpideak aipatuko ditugu. Nahiz eta w2c-en eta c2c-en eskuratze prozesua oso antzekoa izan, nahiago izan ditugu banandurik azaldu.

Berriro ere, azpimarratu beharra dago lan honetan ez garela eskuratze-teknika hauen azterketa sakonean murgilduko. Ikerlana hauetatik abiatuta egin dugu eta hauei buruzko azalpen labur bat bakarrik emango dugu[15].

Class-to-class (c2c)

HM mota hau zertan datzan ulertu ahal izateko, lehendabizi nola lortzen den ulertzea garrantzitsua da.

Aditz baten c2c HMak eskuratzeko, lehenengo corpusaren gainean *Minipar* analizatzaile sintaktikoa (Lin, 1993) erabili behar da, aditz horren corpuseko agerpen bakoitza [IZENA, (izena eta aditzaren arteko) ERLAZIO SINTAKTIKOA, ADITZA] hirukote modukoetan adierazteko. Adibidez, Miniparrek corpusean irabazi aditzaren (2)ko agerpena, (3) adibideko hirukotean bilakatuko luke:

(2) Futbol-taldeak irabazi zuen.

(3) $\begin{bmatrix} \text{Futbol-talde } (\textit{Izena}) \\ \text{Subjektua } (\textit{Erlazio sintaktikoa}) \\ \text{Irabazi } (\textit{Aditza}) \end{bmatrix}$

[14]Eskuratze-tekniken laburdurak ingelesez mantendu ditugu, hizkuntzalaritza konputazionalean horrela ezagutzen direlako. Esaterakoan, ordea, hauek euskaraz ***hitza-klase*** eta ***klase-klase*** bezala aipa daitezke.

[15]Argibide gehiagorako jo bedi hurrengo lanetara: Agirre eta Martínez (2001, 2002); Pociello (2004a).

Ondoren, hirukote bakoitzean dauden izenak MCRn kontsultatzen dira. Horrela, aditza bera, eta aditz horrekin agertu den izen bakoitzaren adiera (bere *synset*-zenbakiarekin) desanbiguatuko da automatikoki (Agirre eta Martínez, 2002). SemCor corpusaren gainean ari bagara, hirukote hau corpusetik zuzenean datorkigu, corpusa bera WordNeteko *synset*-zenbakiekin eskuz etiketatuta baitago. Hortaz, orain hirukotea [IZENA eta bere SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKOA, ADITZA eta bere SYNSET-ZENBAKIA] motakoa izango da.

(4) $\begin{bmatrix} \text{Futbol-talde/05167683 } (\textit{Izena/Synset-zenbakia}) \\ \text{Subjektua } (\textit{Erlazio sintaktikoa}) \\ \text{Irabazi/00620486 } (\textit{Aditza/Synset-zenbakia}) \end{bmatrix}$

Azkenik, hirukote bakoitzaren probabilitatea kalkulatzen da, corpusean duten maiztasunaren arabera[16]. Hirukoteak daraman kopuru hau 1 zenbakitik geroz eta gertuago egon, orduan eta ziurrago egon gaitezke hirukoteak aditzarekiko adierazten duen harremana egokia dela.

Beraz, [IZENA/SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKOA, ADITZA/ SYNSET-ZENBAKIA] motako hirukoteak dauzkagu, ondoan HMaren egokitasuna markatzen duen probabilitatearekin. (5) adibidean (2), (3) eta (4)ko hirukote bera dakargu, baina probabilitatea gehituta (kasu honetan, 0.085) eta prozesuaren ondorioz ikus ahal izango dugun itxurarekin[17]:

(5) **c2c.subj** (*Eskuratze-teknika eta erlazio sintaktikoa*)
irabazi 00620486 (*Aditza eta bere synset-zenbakia*)
05167683 0.085 futbol-talde "Futbolean jokatzen duen taldea"
(*Synset-zenbakia, probabilitatea, synseteko sinonimoak eta definizioa*)

Esan bezala, eskuratze-teknika honetan HMak **izen klaseen** bidez datoz adierazita. Eskuratze-eredu honetako algoritmoak corpusetik jasotzen dituen objektu/subjektuen izenak MCRn kontsultatzen ditu, gerora izen horiek guztiak multzokatzen dituen klase semantikoa aukeratzeko; normalean hauen hiperonimo bat. Horrela, corpuseko izen hori orokor dezakeen beste izen bat lortzen da, aditz batekin joan daitekeen izen multzo bat mugatzen duena, hain zuzen ere. (2) adibidearekin jarraituz, ezin da ukatu futbol-talde izena irabazi aditzaren subjektua izan daitekeela, baina era berean esan dezakegu:

[16] Argibide gehiago hurrengo lanetan: Agirre eta Martínez (2001, 2002).

[17] Azalpena ulergarriagoa izan dadin, atal honetako HMen adibide, glosa eta *synset* asmatuak euskaraz jarri ditugu. Hala ere, hurrengo ataletan ingelesez aurkeztuko ditugu, azterlan honetan eskuratze-tekniken emaitza guztiak ingelesez daudelako.

(6) Saskibaloi-taldeak irabazi zuen.

(7) Errealak irabazi zuen.

Esandakoaren arabera, (5) ez da eskuratze-prozesuaren azken emaitza, futbol-talde izenaren ordez, hau orokortzen duen hiperonimo bat agertuko zaigulako:

(8) **c2c.subj**
irabazi 00620486
04771851 0.101 0.145 gizatalde "Mota bereko izaki bizidunen multzoa"

HM honetatik abiatuta badakigu, irabazi 0062486 aditzaren subjektu mota batek gizakia izan behar duela ([+gizakia]), eta gainera gizaki horiek talde bat osatu behar dutela ([+talde]). Horrela bada, eskuratze-eredu honekin HMak izen klaseak izango dira.

Bestalde, esan dugun bezala, eskuratze-teknika honek aditzaren adiera ere kontuan hartzen du. c2c eskuratze-teknikak lortzen dituen HMak aditzaren adiera jakin baterako dira. Beraz, MCR kontsultatzean irabazi aditzari 00620486 *synset*-zenbakia egokitu bazaio ('lehiaketa baten irabazlea izan'), automatikoki eskuratutako HMak irabazi aditzaren adiera horrentzat bakarrik izango dira, eta inolaz ere aditzaren beste adierentzat. Arrazoi horregatik, (5) eta (8) adibideetan aditzaren ondoren honen *synset*-zenbakia dator zehaztuta: 00620486 *synset*ari dagokion adieraren ('lehiaketa baten irabazlea izan') HMak direla adierazteko.

(9) adibidean irabazi aditzaren objektu HMen adibide bat dugu, 00620486 *synset*ari dagokion adierarekin, hots, kirol-adierarekin ('lehiaketa baten irabazlea izan').

(9) **c2c.subj**
irabazi 00620486
04771851 0.101 lehiaketa "Sari bat irabazteko elkarren lehian egiten den jarduna"
00597858 0.066 talde-ekintza "Talde batek aurrera daraman ekintza"

Gainera, eskuratze-teknika honek aditza klase bezala ere ulertzen du, hau da, lortutako HMak baliagarriak dira aditz horrentzat, bere *synset*ean dituen sinonimo guztientzat, eta bere troponimoentzat. (8)ren kasuan, HM horiek irabazi 0060486 *synset*ari eta honen azpian dauden beste *synset* guztiei dagozkio. Horrela, bada, eskuratze-teknika honen HMak aditz-klase oso bati dagozkie. SemCor semantikoki etiketatutako corpus bat izaki, eskuratze-

teknika honek, corpusean irabazi 0060486 *synset*aren troponimo bat agertuko balitz, bere hiperonimoarekin erlazionatzeko gai izango litzateke, eta klase guztiari HM berdinak egokituko lizkioke[18].

Azkenik, aipatu beharra dago, eskuratze-teknika honekin (eta besteekin) ez dela aditz bakoitzarentzat HM bakarra lortzen, aditz bakoitzak probabilitate kopuru altuenetik baxuenera ordenaturiko HMen zerrenda bat izango baitu. Horrela, aditz baten objektu/subjektu argumentu gisa agertzen diren izenen zerrenda izango dugu probabilitate altuenetik baxuenera.

Zerrenda hau oso luzea izan daiteke, eta hamar HM baino gehiagok osatzen dutenean lehenengo hamarretara bakarrik mugatzen gara lan honetan. Irizpide hau azterlan honetako eskuratze-teknika guztiekin erabili dugu.

Word-to-class (w2c)

Eskuratze-teknika honen prozesua aurrekoaren oso antzekoa da. Ezberdintasun bakarra da w2c ereduan aditzaren adiera guztiak kontuan hartzen direla. Hala, lehenik, Minipar analizatzaile sintaktikoaren bitartez [IZENA, (izena eta aditzaren arteko) ERLAZIO SINTAKTIKOA, ADITZA] hirukote modukoak ateratzen dira; bigarren pausoa MCRn kontsulta egitea da, baina oraingo honetan, hirukoteko izenak bakarrik begiratzen dira MCRn, aditza bere adiera guztiekin kontuan hartzen baita. Hala, izen horiek adierarekin edo *synset*-zenbakiarekin desanbiguatuta izango ditugu. Beraz, orain hirukotea [IZENA/SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKOA, ADITZA] motakoa izango da.

(10) $\begin{bmatrix} \text{Futbol-talde } (\textit{Izena}) \\ \text{Subjektua } (\textit{Erlazio sintaktikoa}) \\ \text{Irabazi } (\textit{Aditza}) \end{bmatrix}$

c2c eskuratze-teknikarekin ikusi dugun bezala, SemCor WordNeteko *synset*ekin etiketatutako corpusa izaki, honen gainean aritzean, hirukoteko izenen *synset*ak corpusetik zuzenean datozkigu.

Azkenik, c2c eskuratze-teknikan bezala, hirukote bakoitzaren probabilitatea kalkulatu egiten da, corpusean duten maiztasunaren arabera[19].

Horrela bada, HM hauek duten itxura c2c teknikarekin lortutakoaren oso antzekoa da:

[18] Honen adibideak VII.4 atalean ikusiko ditugu.

[19] Argibide gehiago hurrengo lanetan: Agirre eta Martínez (2002, 2001).

(11) **w2c.subj** (*Eskuratze-teknika eta erlazio sintaktikoa*)
irabazi (*Aditza*)
05167683 0.070 futbol-talde "Futbolean jokatzen duen taldea"
(*Synset-zenbakia, probabilitatea, synseteko sinonimoak eta definizioa*)

w2c eskuratze-teknikan, c2c-en gertatzen den bezala, izenen HMak **izen klaseen** bidez datoz adierazita, hots, corpusean irabazi aditzak subjektu edo objektu gisa hartzen dituen izenak, algoritmoak automatikoki dagokien hiperonimoarekin multzokatzen ditu.

(12), (13) eta (14) adibideetan irabazi aditz-formarekin objektu gisa agertu diren izen klaseen zerrenda dugu (15) adibidean, probabilitate altuenetik baxuenera ordenaturik. Bertan ikus daiteke oso garbi w2c eskuratze-teknika honek eskaintzen dituen HMak aditzaren adiera guztiei erreparatzen dietela. Honela bada, lehiaketa izen-klasea kirol-adierari dagokio, eta jabegoa, aldiz, finantza adierari.

(12) **partidua** irabazi (hiperonimoa: *lehiaketa*)

(13) **futbolean** irabazi (hiperonimoa: *talde-ekintza*)

(14) **dirua** irabazi (hiperonimoa: *jabego*)

(15) **w2c.obj**
irabazi
04771851 0.101 lehiaketa "Sari bat irabazteko elkarren lehian egiten den jarduna"
00597858 0.066 talde-ekintza "Talde batek aurrera daraman ekintza"
00017394 0.037 jabego "Norbaitek berea duen zerbaitekiko duen eskubidea"

VII.3.2.2 Domeinu eta eremu semantiko batekin adierazitako HMak

Mota honetako eskuratze-teknikek aditz baten HMak domeinu-eremu semantiko bikote batez adierazten dituzte, bikote hau klase bezala kontsideratzen dutelarik, hau da, domeinu hori eta eremu semantiko hori dituzten izen guztiak izango dira aditz horren objektu/subjektuen HMak.

IV. kapituluan azaldu dugun bezala, *synset*arekin domeinua eta eremu semantikoari buruzko informazioa dator. Alde batetik, MCRko klase semantiko bakoitza fitxategi batean jasota dago, *eremu semantiko* deritzogun fitxategia, hain zuzen (ingelesez, *semantic field*): *gertaera, jabetza, taldea,*

pertsona, ekonomia, lekua eta abar bezalakoak. Bestalde, domeinu-ontologia dugu, eta honekin *synset*ak domeinuen arabera antolatzen dira: *kirola*, *jatetxea*, edo *trafikoa*, esate baterako[20].

Synset batekin adierazitako HMetan barruan w2c eta c2c eskuratze-teknikekin gertatzen zen bezala, hemen ere eskuratze-teknikak ezberdintzen dira HMak aditz-formatik edo aditz-adieratik abiatuta eskuratzearen arabera.

Aditzaren HMak eskuratzean, HM hauek aditzaren adiera guztientzako izan badaitezke, (**aditz-formarentzat**, alegia) ***word-to-semantic-field*** (aurrerantzean, w2semf[21]) eskuratze-teknikaz hitz egingo dugu, hots, hitzetik abiatuta domeinu-eremu semantiko bikoteak lortzen dituenaz.

HMak aditzaren adieraren arabera ere lor badaitezke (**aditz-adierarentzat**, alegia), orduan, ***sense-to-semantic-field*** (aurrerantzean, s2semf) eskuratze-teknikaz baliatu garela esango dugu, hau da, aditz-adieratik[22] abiatuta domeinu-eremu semantiko bikoteak lortzen dituenaz.

Har ditzagun, berriro ere, irabazi aditza eta (12), (13) eta (14) adibideak. Aditz honen w2semf objektu HMak aditzaren adiera guztientzat lirateke.

(16) **w2semf.obj** (*Eskuratze-teknika eta erlazio sintaktikoa*)
irabazi (*Aditza*)
obj ekonomia-jabetza 33
obj kirola-gertaera 28
(Erlazio sintaktikoa, domeinu-eremu semantiko bikotea eta probabilitatea)

(16)ko HMak (15)ekoen berdinak dira; adierazpidea da aldatzen dena. (15)ekoak *synset* bidez adierazten ditu klase semantikoak, eta (16)koak, berriz, domeinu-eremu semantiko bikotearen bitartez. Adibidean ikus daitekeen bezala, gauza bera adierazteko, (15)ekoak hiru *synset* behar izan ditu eta (16)koak bi domeinu-eremu semantiko.

Aditz horren kirol-adieran oinarrituz gero (irabazi 00620486), s2semf eskuratze-teknikak aditz-adiera horren kirol domeinuarekin harremanetan

[20] Azalpena ulergarriagoa izan dadin, adibideko eremu semantikoak eta domeinuak euskaraz jarri ditugu. Hala ere, hurrengo ataletan ingelesez aurkeztuko ditugu, azterlan honetan eskuratze-tekniken emaitza guztiak ingelesez daudelako.

[21] Eskuratze-tekniken terminologia ingelesez mantendu dugu, hizkuntzalaritza konputazionalean horrela ezagutzen direlako. Hala ere, hauek euskaraz ***hitza-domeinu-eremu semantiko bikotea*** eta ***adiera-domeinu-eremu semantiko bikotea*** esan daitezke.

[22] c2c eta s2semf ezberdintzen dira, aditzaren izaeran. Lehenengoak aditzaren *synset*eko sinonimoak eta troponimoak kontuan hartzen ditu; eta bigarrenak, aditzaren *synset*eko sinonimoak bakarrik.

dauden objektuen HMak bakarrik eskuratuko lituzke[23]:

(17) **s2semf.obj** (*Eskuratze-teknika eta erlazio sintaktikoa*)
irabazi 00620486 (*Aditza eta bere synset-zenbakia*)
obj joko-ekintza 33
obj kirola-gertaera 28
(Erlazio sintaktikoa, domeinu-eremu semantiko bikotea eta probabilitatea)

(17)ko HMak (9)koen berdinak dira baina lehenengoak domeinu-eremu semantiko bikoteekin adieraziak, eta bigarrenak *synset*ekin.

Atal honen hasieran esan bezala, bikote hauek klase semantikoak dira: *kirola* domeinua eta *gertaera* eremu semantikoa duten izen guztiak izan daitezke irabazi aditzaren objektuak.

Domeinu-eremu semantiko bikoteen bidez adierazitako izen klase hauek corpusetatik erauzteko, w2c eta c2c eskuratze-tekniketan erabilitako aurreprozesu bera erabiliko da w2semf-ekin eta s2semf-ekin ere. Lehenengo, corpusaren gainean Minipar analizatzaile sintaktikoa (Lin, 1993) erabili behar da, aditz horren corpuseko agerpen bakoitza [IZENA, (izena eta aditzaren arteko) ERLAZIO SINTAKTIKOA, ADITZA] hirukote modukoetan adierazteko. (2) adibidearen hirukotea (3)koa litzakete.

Ondoren, hirukote bakoitzean dauden izenen MCRko eremu semantikoak eta domeinuak kontsultatzen dira. Hortaz, orain hirukotea [IZENA eta bere DOMEINUA/EREMU SEMANTIKOA, ERLAZIO SINTAKTIKOA, ADITZA] motakoa izango da. Adibidez, (12)ko irabazi aditzaren agerpena, (18) adibideko hirukotean bilakatuko litzateke:

(18) $\begin{bmatrix} \text{Futbol-talde/football/group } (\textit{Izena/Domeinua/Eremu semantikoa}) \\ \text{Subjektua } (\textit{Erlazio sintaktikoa}) \\ \text{Irabazi } (\textit{Aditza}) \end{bmatrix}$

HMa aditzaren adiera bakarrarentzat lortzen denean, hirukote hau aditzaren *synset*arekin zehaztuta dator.

Azkenik, hirukote bakoitzaren pisua kalkulatzen da corpusean duten maiztasunaren arabera[24]. Hirukoteak daraman pisua geroz eta handiagoa izan, orduan eta fidagarritasun handiagoa. Azkeneko emaitza (16) eta (17)koen itxurakoa da.

[23] Corpusa semantikoki etiketatua badago, eskuratze-teknikak zuzenean hartzen du corpusetik adiera hori. Bestela, hitzen adieren desanbiguazioan erabiltzen diren teknikak erabili behar dira. Argibide gehiagorako jo bedi Agirre eta Martínezen lanera (2002).

[24] Argibide gehiago hurrengo lanetan: Agirre eta Martínez (2002, 2001).

VII.3.2.3 Baliabideak laburbilduz

Ikusi dugun bezala, saiakera honetan ingeleseko lau corpus ditugu (ingeleserako SemCor, BNC eta EFE eta euskararako *Euskaldunon Egunkaria*), eta lau eskuratze-teknika erabili ditugu.

Eskuratze-teknika hauek guztiak ez dira corpus guztietan aplikatu. c2c eta w2c tekniken emaitzak Martínezen (2005) lanari lotutakoak dira; bi teknika hauekin landu zuen berak HMen eskuratze automatikoa. Martínezek teknika hauek SemCorren eta BNCn bakarrik erabili zituen.

s2semf eta w2semf teknikako emaitzak MEANING proiektuari dagozkionak dira. Proiektu honetan ingeleserako SemCor eta EFE corpusak baliatu ziren, eta euskararako *Euskaldunon Egunkaria*. Eskuratze-teknikari dagokienean, s2semf eta w2semf erabiltzea erabaki zen, hasiera batean (etorkizunean gainontzeko teknikak probatzeko asmoarekin). s2semf-ek desanbiguatutako corpusa behar duenez, SemCorren bakarrik erabili ahal izan zen. Aldiz, EFEn eta *Euskaldunon Egunkaria*n w2semf erabiltzea erabaki zen.

Hala, saiakera honen emaitzak behin-behinekoak dira, eskuratze-teknika guztiak ezin izan direlako corpus guztien gainean erabili. Hortaz, honako hau hastapeneko lana dugu, eta hemen aurkeztutako emaitzetatik eta ondorioetatik abiatuta, azterketa honen esparrua zabaltzeko asmoa dugu.

VII.4 Ingeleseko HMak

Aipatu izan dugun bezala (ikus VII.3. atala), azterketa honetarako, kirol-domeinuko bost aditz aukeratu ditugu: jokatu, galdu, irabazi, entrenatu eta berdindu. Hala ere, aditz hauek kirol-adieraz gain beste adiera batzuk izan ditzakete ('zuzen jokatu, dirua irabazi/galdu...'). Hauetako bakoitzak dituen adierez jabetzeko, MCRra jo dugu, eta adiera horietatik guztietatik kirolarekin zerikusia zutenetan bakarrik oinarritu gara.

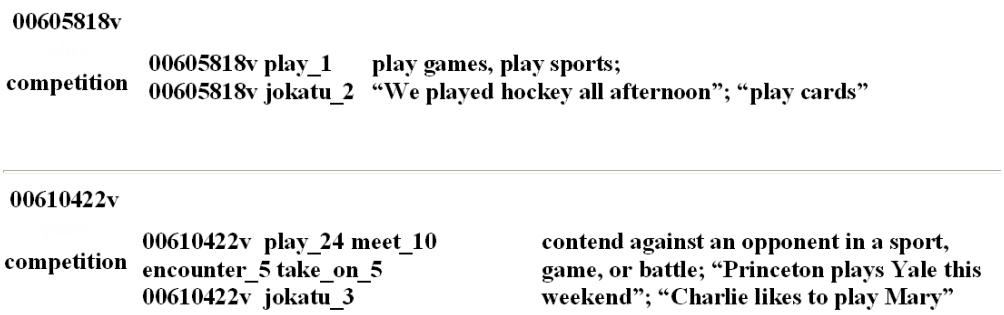
Nola jakin *synset* bat kirol-adierari dagokiola? Batetik, *synset*arekin batera datorren glosari eta eremu semantikoari esker, eta bestetik, *synset* horri dagokion domeinua begiratuta. Kirol-adiera duten *synset* guztiek *sport* domeinua izan behar dute.

Har dezagun jokatu aditza. MCRn kirolarekin harremanetan dauden bi *synset* ditu; batek, 'zerbaitetan jokatu' adierazten du (jokatu 00605818), eta, besteak, '-ren aurka jokatu' (jokatu 00610422). VII.1. irudian jokatu aditzaren bi kirol *synset*ak ditugu, eta adiera-ezberdintasun hau glosan adierazia

Synset-eko hitza(k)	*Kategoria*	*Synset-zenbakia*	*Domeinua*	*Domeinua*
jokatu, jokoan jardun	Aditza	00605818	play	**sport**
jokatu	Aditza	00610422	play	**sport**

VII.3 Taula: jokatu aditzaren kirol *synset*ak eta beraien domeinuak MCRn.

dator. Bestalde, irudian ikus daitekeen bezala, bi *synset*en eremu semantikoak (*competition*) agertzen dira.



VII.1 Irudia: jokatu aditzaren bi kirol *synset*ak.

Domeinuari erreparatuz (ikus VII.3 taula), bi *synset* hauek *sport* domeinuaren marka daramate[25]. Dena den, *synset*ek domeinu bat baino gehiago izan ditzakete, bi *synset* hauen kasuan ikus daitekeen bezala[26]. Ikusiko dugun bezala, honek HMetan ondorioak izango ditu.

Euskarako aditz-formen kirol-adierak mugatu ondoren, aukeratutako bost aditzen kirol-adierei honako zortzi *synset* hauek dagozkie (bai euskararako eta bai ingeleserako), eta hauetan oinarritu gara HMen azterketarako:

[25]Domeinuen sailkapena ez da MCR interfazean ikusten, beste fitxategi batzuetan daude. Hala ere, aurki jarriko dira interfazean.

[26]*Play* eta *Sport* domeinuek antzekoak diruditen arren, gauza ezberdinak adierazten dituzte. *Sport* domeinuak ekintza fisikoarekin edota joko konpetitiboekin zerikusia duenari egiten dio erreferentzia; *play* domeinuak, ordea, apustua edota jokoarekin zerikusia duen edozeri. Euskarako itzulpenak *jokoa* eta *kirola* izan daitezke.

- 00605818 {play_1/jokatu_2}; "play games, play sports"[27]
- 00610422 {encounter5, meet10, play24, take on5/jokatu3}; "contend against an opponent in a sport or game"
- 00468052 {coach_2, train_7/entrenatu_1}; "teach and supervise, as in sports or acting"
- 00059698 {train_8/entrenatu_3}; "exercise in order to prepare for an event or competition"
- 00630097 {equalize_1, get even_1/berdindu_16}; "compensate; make the score equal"
- 00630097 {draw_25, tie_2/berdindu_15}; "finish a game with an equal number of points, goals..."
- 00620486 {win_1/irabazi_3}; "be the winner in a contest or competition"
- 00620218 {lose_2/galdu_9}; "fail to win"

Atal honetan, corpus eta teknika desberdinak erabiliz, ingeleserako eskuratutako HMak aztertu eta ebaluatuko ditugu, hau da, MCRko *synset* horietan dauden ingeleseko *variant*ak hartuko ditugu ingeleseko HMen azterketa eta ebaluazioa egiteko[28]. Baina, atal honetan, *synset* hauetatik play 00605818 *synset*a baliatuko dugu adibide gisa ingeleseko aditz hauekin guztiekin erabilitako metodologia ulergarriago egitearren. Honenbestez, atal honetan jokatu 00605818 *synset*aren azterketaz arituko gara; beraz, hemendik aurrera, bere ingeleseko *variant*a (play 00605818) hartuko dugu oinarri gisa. Hala ere, aditz guztien azterketa eta emaitzak C eranskinean daude.

[27] MCRn *synset*ek zenbaki bat daramate (00605818), baita *synset* barruko ordainek ere (play_1). Lehenengoa *synset* osoari dagokio, osatzen duten ordainak barne. Bigarrenak hitzaren adiera zehazten du, hau da, hitz polisemikoen adierak zenbakituak datoz. Bigarren hauei *variant* deitzen zaie (ikus IV.1.1 atala). Hala ere, biekin gauza bera adieraz daiteke: play_1ek MCRko play hitzaren lehenengo adiera adierazten du; eta play 00605818k, play hitzak 00605818 *synset*eko adiera duela, hots, play_1.

[28] Jokatu 00605818 *synset*ak ingelesez *variant* bakarra duenez (play), *variant* horren HMak aztertuko ditugu bakarrik. Baina, Jokatu 00610422ren kasuan, adibidez, bere ingeleseko *variant*ak lau dira ({encounter, meet, play, take on}), hau da, kontzeptu hori adierazteko ingelesez sinonimo horiek erabil daitezke. Azterlan honetan *synset* berean dauden ingeleseko *variant* guztien HMak aztertu ditugu.

VII.4.1 Ingeleseko HMetarako irizpideak

Eskuratze-teknika desberdinen HMak ebaluatzeko, *synset* bakoitzeko eskuratze-teknika bakoitzaren emaitza prototipikoak eskuz sortu ditugu (*urre-patroi* deitu duguna[29]), eta kasu honetan, play 00605818 *synset*arentzat.

Urre-patroiak eskuratze-teknika bakoitzaren ereduan sortuko dira. Hau da, guk sortutako urre-patroiek teknika hauen emaitzek hartzen duten itxura hartuko dute: alde batetik, HMak adierazteko *synset*ean oinarritzen direnenak (w2c eta c2c), eta bestetik, domeinu-eremu semantikoetan oinarritzen direnenak (w2semf eta s2semf). Hala, urre-patroiak ere bi azpimultzo hauetan banatu ditugu; patroi batzuk *synset* bidez adieraziko ditugu w2c eta c2c tekniketatik lortutako HMak ebaluatzeko, eta beste patroiak domeinu-eremu semantiko bikoteen bidez definituko ditugu, w2semf tekniketatik lortutako HMak ebaluatu ahal izateko.

Hortaz, argi dago urre-patroi hauek proposatu ahal izateko MCR erabili behar izan dugula, VII.2 atalean ikusi dugun bezala, bertan oinarritzen baitira eskuratze-teknikak HMak adierazteko (*synset*, eremu eta domeinu semantikoen bidez).

Honezaz gain, erabilitako corpusetan ere oinarritu gara saiakeran. Corpus hauetatik hartutako esaldietatik, aztertu beharreko aditz-adiera bakoitzaren jokaera linguistikoa orokortzen saiatu gara, gerora, orokortasun horiek (HMak, alegia) MCRko *synset* eta domeinu-eremu semantiko batzuen bidez adierazteko. Corpuseko izen bat HM batean orokortzeko, gehienetan izen horrek MCRn duen hiperonimoetara jo dugu. Azken finean, makinak eskuratze-tekniken bidez egin beharko lukeena egiten saiatu gara eskuz. Esan dezakegu, beraz, MCRko *synset* eta domeinu-eremu semantikoetan oinarrituta, introspekzioaz baliatu garela urre-patroiak sortzeko.

(19)n ditugu play 00605818 aditz-adieraren urre-patroiak eta (20)n patroien adibideak[30]:

[29] Izen hau ingeleseko *goldstandard*etik itzuli dugu.

[30] Eskuratze-teknikek ematen dituzten emaitzak ingelesez daude, MCRko informazioa ingelesez dagoelako. Hau da, MCRko euskarri informatikoa ingelesez dago; ingelesez ez dagoen bakarra beste hizkuntzetako *variant*ak eta glosak dira. Euskarako glosak oraindik ez daude guztiz itzulita, horregatik, ingelesekoetan oinarritzen gara.

(19) **play 00605818 Objektuak**
w2c, c2c:
00240760 {sport, athletics} "an active diversion requiring physical exertion and..."
00254052 {game} "a contest with rules to determine a winner"
04771851 {contest, competition} "an occasion on which a winner is selected from..."
09065837 {amount of time, period, period of time} "time period a length of time"

s2semf, w2semf:
sport-event
time-period_time
sport-act
play-act

play 00605818 Subjektuak
w2c, c2c:
00004865 {person, individual, someone, somebody, human soul} "a human being"
00017008 {group, grouping} "any number of entities (members)considered..."

s2semf, w2semf:
person-person
factotum-group[31]

(20) Objektuak:

John played **football**.
John played **a match**.
John played **five minutes**.
John played **a game**.

Subjektuak:

John played football.
The football-team played a match.

Kontuan izan beharrekoa da MCR hierarkia bat dela eta batzuetan ez dela horren erraza HMa adierazten duen *synset egokia* aukeratzea, gerta litekeelako *synset* hori orokorregia izatea (hierarkian goregi egotea) edo zehatzegia izatea (hierarkian beheregi egotea). Esate baterako, play aditzarentzat {contest, competition}[32] HMa proposatu ordez, MCRko bere hiponimoa

[31] Adiera batek domeinurik ez duenean *factotum* markarekin adierazten da.

[32] *Synset* berean ordain bat baino gehiago agertzen direnean, azalpenetan *synset*a adierazteko bi *variant*ak giltzen artean adieraziko ditugu.

(match "a formal contest in which two or more persons or teams compete") proposatuz gero, aditz horren objektuen aukeraketa gehiagi mugatuko genuke, eta {contest, competition} bezalakoak ezingo genituzke zuzentzat jo. Alderantziz ere berdin: {contest, competition} HMaren ordez, bere hiperonimoa social event ("an event characteristic of persons forming groups") proposatu izan bagenu, aukera gehiegi izango genituzke eta zuzenak ez diren HMak ere agertuko lirateke (adibidez, play 00605818 aditzak social event horren hiponimoa den ballet HMa onartuko luke).

Arazo hau bera areagotu egiten da domeinu-eremu semantiko bikoteen bidez adierazitako HMak ebaluatzean. Domeinu-eremu semantiko bikote hauek *synset*ak baino orokorragoak dira. Adibidez, Errealak partidua jokatu zuen esaldian, subjektuaren HMa sport-group bikote gisa adieraz daiteke. Baina kirol-aditzak ez dira kirolarekin harremanetan dauden izenetara bakarrik mugatzen (Donostiarrek partidua jokatu zuten). Horregatik domeinu-eremu semantiko bikote orokorragoak onar daitezke (factotum-group, adibidez).

HMak adierazteko arazo hau dela eta, hauek ebaluatzeko maila desberdineko markak erabili ditugu:

- **Zuzena:** Urre-patroiarekin bat datorrenean.
- **Onargarria:** Urre-patroiaren hiperonimoa edo hiponimoa denean. Domeinu-eremu semantiko bikoteen bidez adierazitako HM kasuan, onargarri bezala kontsideratu ditugu urre-patroia baino orokorrago edota zehatzago direnak.
- **Okerra:** Urre-patroiarekin bat ez datorrenean eta MCRko hierarkian ere loturarik ez dutenean.

Marka hauek ez digute inolako arazorik eman *synset*ekin adierazitako HMak ebaluatzerakoan. Haatik, domeinu-eremu semantiko bikoteekin adierazitakoak ebaluatzeko, batzuetan onargarriak ala okerrak diren erabakitzeko zailtasunak izan ditugu. Esate baterako, play 00605818 *synset*ak [+gizaki] motako subjektuak har ditzake; *synset*ekin adierazita, 00004865 {person, individual, human} "a human being"[33] HMa litzateke, eta domeinu-eremu semantiko bikoteekin adierazita, person-person. Eskuratze-tekniken emaitzetan hauexek agertuz gero, play 00605818ren urre-patroietan definituak

[33]Batzuetan, toki-arazoak direla-eta, *synset*ak laburtu egin ditugu, *variant* kopurua edota glosa txikituz.

daudenez, ez legoke inolako arazorik, eta zuzentzat joko genituzke. Hala ere, emaitzetan hauen aldaerak ager daitezke, hau da, urre-patroiaren hiperonimo/hiponimoak diren *synset*ak (06441015 young man "an adolescent male", adibidez) edo urre-patroiko domeinu-eremu semantiko bikotea baino orokorrago/zehatzago[34] diren bestelako bikoteak (transport-person, administration-person, basketball-person...). Demagun, eskuratze-teknika baten emaitza 06441015 young man "an adolescent male" dela, orduan, onargarri gisa ebaluatutako dugu hau urre-patroiko 00004865 {person, individual, human} "a human being" *synset*aren hiponimo bat delako. Aldiz, eskuratze-teknikaren emaitza transport-person, administration-person, basketball-person... denean, zenbaitetan zalantza dugu. Lehenengo begiratuan, basketball-person domeinu-eremu semantikoa play 00605818ren kirol adierarekin zerikusia duenez[35], onargarritzat joko genuke, eta transport-person eta administration-person, berriz, okertzat —play 00605818ren adierarekin bateragarriak ez direlako (?Administrators played football), eta *transport* eta *administration* ez direlako *sport* domeinuaren hiponimoak edo hiperonimoak MCRn. Hala ere, datuak eta corpusak aztertuz, konturatu gara hauek Brazilians, cyclist eta gisa horretako agerpenetatik datozela, eta play 00605818rekin onargarriak direla (Brazilians played football). Baina, Brazilians bezalako kasu hauek gutxienekoak dira, eta hauek sortutako *administration* HMa onargarritzat joz gero *administration* domeinuaren azpian dauden beste hitz guztiak ere (chairman, chancellor...) jokatu aditzaren (kirol-adieraren) subjektu/objektu prototipiko gisa ager daitezkeela baieztatzen ariko ginateke. Hori, bistan da, ez litzateke oso egokia.

Ikus daitekeen bezala, domeinu-eremu semantiko bikoteekin *synset*ekin baino arazo gehiago sortu zaizkigu, eta horren ondorioa izan da ebaluaziorako irizpide zehatzagoen beharra:

- Domeinu-eremu semantiko bikote bat onargarritzat hartuko dugu, urre-patroia baino orokorrago edota zehatzago bada, **eta domeinuko beste izen gehienak aditz horren argumentu izan badaitezke**. Irizpide honen arabera, zuzentzat hartuko ditugu, urre-patroia baino orokorrago edota zehatzago diren HMak, baldin eta domeinuko beste izen gehienak aditz horren argumentu izan badaitezke. Aurreko adibidearen kasuan, *administration* domeinuaren azpian MCRko chairman, adminis-

[34] Domeinu hierarkia izanik, domeinuak hiperonimia/hiponimiaren arabera antolatuak daude.

[35] MCRko domeinu hierarkian *basketball* domeinua *sport* domeinuaren hiponimoa da.

trator, chancellor eta abar bezalakoak daude sailkatuak; hauek ezin dute play 00605818ren HMak izan (ez testuinguru arruntetan behintzat). Beraz, domeinu-eremu semantiko bat onargarria den erabakitzeko, lehendabizi domeinu horrek hartzen dituen izenak aditz horren argumentu gisa ager daitezkeen aztertu beharko dugu.

- Izen-bereziak (x baten bidez adieraziak datozenak), pronominalak (pro baten bidez adieraziak datozenak), eta factotum-Tops bikoteak erreferente orokorregia dute, eta ezinezkoa da jakitea beraien jatorria corpusean. Arrazoi horregatik nahiz eta onargarri bezala ebaluatu, ez dira estatistiketan kontuan hartuko. Esate baterako, factotum-Tops bikote honek ia edozer gauza adieraz dezake, *factotum*ekin domeinurik ez duten hitzak adierazten direlako, eta *Tops* eremuak MCRko hierarkian oso goian dauden *synset*ak jasotzen dituelako. Beraz, oso orokorra diren kontzeptuak dira.

- Zuzen/onargarri bezala ebaluatutako HM batekin, bi urre-patroi eskuratu daitezke, baldin eta eremu semantikoa bera duten. Esate baterako, factotum-act HMarekin play-act eta sport-act urre-patroiak eskuratzen dira, adibidez.

VII.4.2 HMen azterketa eta ebaluazioa

Corpus desberdinetatik eskuratutako HMen azterketa egin aurretik, orain arte jarraitutako pausoak laburbilduko ditugu. Gogora dezagun azalpenerako jokatu 00605818 *synset*ean oinarritu garela adibide gisa:

- Euskarako jokatu aditz-formatik abiatu gara eta honek dituen kirol-adierak (*synset*ak) bilatu ditugu MCRn (jokatu 00605818 eta jokatu 00610422).

- *Synset* hauek kirol-adiera dutela egiaztatzeko beraien domeinua *sport* dela egiaztatu dugu.

- *Synset* bat hartu dugu –gure kasuan jokatu 00605818 eta bere ingeleseko ordaina hartu dugu (play 00605818)– aditz-adiera honen HMak ingeleseko corpusetatik lortzeko.

- Eskuratze-tekniken emaitzak ebaluatu ahal izateko, ingeleseko corpusetan oinarrituta aditz-adiera horrek hartzen dituen HMen urre-patroiak eskuz sortu ditugu landutako eskuratze-teknika mota guztientzako.

Emandako urrats hauekin, eskuratze-teknika mota bakoitzaren emaitza ebaluatzeko gai gara. Eskuratze-teknika hauek programa informatikoak dira, eta jarraian, eskuratze-teknika hauek automatikoki lortutako emaitzen (HMen) ebaluazio linguistikoa egingo dugu. Hurrengo ataletan lan honen azalpenari ekingo diogu, eta, horretarako, azalpena corpusen arabera antolatu dugu. Horrela, VII.4.2.1. atalean SemCor corpusetik eskuratutako HMen azterketa egingo dugu, VII.4.2.2. atalean BNCtik eskuratutakoena, eta, azkenik, VII.4.2.3. atalean EFEtik eskuratutakoena.

VII.4.2.1 SemCorretik eskuratutako HMen azterketa eta ebaluazioa

Corpus honetan c2c, w2c eta s2semf eskuratze-teknikak erabili dira. Hauekin irizpide metodologiko berdintsuak baliatu ditugun arren, beraien artean bada berezitasunik.

c2c SemCorretik

c2c eskuratze-teknikak lortzen dituen objektuen edo subjektuen HMak aditzaren adiera jakin baterako dira: play 00605818. Eskuratze-teknika honetan HMak aditz-adiera horrentzat baliagarri diren neurrian, *synset*ean dituen sinonimoentzat eta bere troponimoentzat ere baliagarri dira.

Eskuratze-teknika honen emaitza ebaluatzeko, hurrengo urratsak jarraitu ditugu:

- **HM bakoitzaren jatorria ezagutu:** HMak lortzeko corpusaren agerpen zehatzetan oinarritzen garenez —zehazkiago esanda, corpusean aditzarekin batera agertu diren izenetan (objektu eta subjektu direnetan)—, gure lehenengo lana corpuseko jatorria zein den jakitea da. Hala, eskuratze-teknikaren lana oinarritik ebaluatu dezakegu, gerta baitaiteke corpuseko objektu/subjektu izen horri okerreko HMa egokitzea (geroago ikusiko dugun bezala). Horretarako, corpusean aditz horrekin subjektu edo objektu gisa agertu diren izenen zerrenda oso baliagarria litzaiguke. Arrazoi horregatik bi tresnatxo sortu dira lan hau guztia erraztearren: w2w eta s2s deiturikoak (w2c eta c2c teknikentzat, hurrenez hurren). Corpusetik agerpen horiek guztiak eskuz ateratzen jardun ordez, w2w eta s2s baliabideen bidez automatikoki ematen zaizkigu fitxategi batean (fitxategi hauek jasotzen duten informazioa

C eranskinean dago ikusgarri)[36].

- **Izena corpuseko testuinguruan kokatu:** Aditzaren agerpen zehatzak ezagutu ondoren, corpusean hauen testuingurua bilatzen dugu, hauek guztiak aztertzen ari garen kirol aditzarekin bateragarriak diren ala ez eskuz egiaztatzeko.

- **HMen ebaluazioa:** Eskuratze-tekniken HMen eta hauen corpuseko jatorria aurrean izanda, ebaluazioa egiten has gaitezke.

Pauso hauek jarraituta, play 00605818 *synset*aren objektu eta subjektu HMak ditugu (21)en; s2s zerrendako[37] izenetatik abiatutako HMak letra lodiz adierazi ditugu, dagokien corpuseko agerpenak (izenak) ere zehaztuz:

(21) **c2c.obj**
play 00605818
002289900.215 {**activity**} **"any specific activity or pursuit"**
PLAY: *football, basketball, golf, game3...*
00004865 0.117 {person, individual, human} "a human being"
00017008 0.102 {**group, grouping**} **"any number of entities considered as..."**
PLAY: *The Owls*
00009469 0.071 {**object, physical object**} **"a physical entity"**
PLAY: *ball, card, rightfield*
04771851 0.035 {**contest, competition**}**"an occasion on which a winner is..."**
PLAY: *game*
03875944 0.029 {interest, involvement} "a sense of concern with curiosity about..."
08162378 0.014 {cost} "the total spent for goods [...] including money and time..."
01691640 0.011 {horse} "solid-hoofed herbivorous quadruped domesticated..."

c2c.subj
play 00605818
00017008 0.517 {**group, grouping**} **"any number of entities considered as..."**
PLAY: *The Mustangs, Texans, line...*
00004865 0.507 {**person, individual, human**} **"a human being"**
PLAY: *mate, Bill Kunkel, Nelson, youngman...*
00009469 0.079 {object, physical object} "a physical (tangible and visible) entity"

[36]Hitzean oinarritzen den eskuratze-teknikaren antza handia dute (ikus VII.2.2.1. atala), baina hauek corpuseko agerpenak zuzenean hartzen ditu, inolako probabilitaterik eskaini gabe. Ez dira eskuratze-teknikak, hizkuntzalariaren lana errazten duten baliabideak baizik. Hauei buruzko argibide gehiago Agirre eta Martínez (2001, 2002) lanetan.

[37]Fitxategi hauek jasotzen duten informazioa C eranskinean dago.

08413915 0.032 {**digit**} **"one of the elements that form a system of..."**
PLAY: *nine*
03953834 0.032 idea, thought "the content of cognition"

Letra lodiz markatu gabe HM ugari geratu dira. Gogoratu beharra dago c2c eskuratze-teknika aditz *synset* horren HMak eskuratzeaz gain, bere troponimoenak ere eskuratzen dituela. SemCor, semantikoki etiketatutako corpus bat izaki, eskuratze-teknika honek corpusean play 00605818 *synset*aren troponimo bat agertuko balitz, bere hiperonimoarekin (play 00605818) erlazionatzeko gai izango litzateke, eta klase guztiari HM berdinak egokituko lizkioke. Hortaz, pentsa daiteke jatorria zehaztu gabe geratu diren horiek; play 00605818ren troponimoetatik datozela. Hipotesi hau egiaztatzeko, s2s datuen aldaera diren s2s-hype fitxategiko datuak erabiliko ditugu. Honek corpusean agertu diren play 00605818 *synset*aren troponimoak zehaztuko dizkigu, hauekin agertu diren izenekin batera. Hala, play 00605818rekin orain arte jarraitu dugun metodologia bera erabiliko dugu troponimo hauekin ere.

Lehenengo, troponimoak eta beraien domeinuak ezagutu behar ditugu (ikus VII.4. taula). Ondoren, s2s-hype erabilita troponimoen agerpenak corpusean zehaztu eta hauen testuinguruak aztertu behar ditugu, kirol-adiera dutela egiaztatzeko eta gero ebaluatzeko. (22)n letra lodiz markatu ditugu corpuseko izenetatik eratorritako HMak eta beraien azpian zerrendatuak datoz corpuseko agerpenak (bai play 00605818renak eta bai honen troponimoenak).

(22) **c2c.obj**
play 00605818
00228990 0.215 {**activity**} **"any specific activity or pursuit"**
PLAY: *football, basketball, golf, game3...*
STAKE: *career*
00004865 0.117 {**person, individual, human**} **"a human being"**
START: *mate*
00017008 0.102 {**group, grouping**} **"any number of entities considered as..."**
PLAY: *The Owls*
FIELD: *team*
00009469 0.071 {**object, physical object**} **"a physical entity"**
PLAY: *ball, card, rightfield*
04771851 0.035 {**contest, competition**} **"an occasion on which a winner..."**
PLAY: *game2*
03875944 0.029 {interest, involvement} "a sense of concern with curiosity about..."

Synset-eko hitza(k)	*Synset-zenbakia*	*Domeinua*	*Domeinua*
start	00607112	play	**sport**
field	00611046	play	**sport**
bet on	00646526	baseball	**sport**
stake	00646526	play	**sport**
parlay	00646865	play	**sport**

VII.4 Taula: play 00605818 *synset*aren troponimoak eta bere domeinuak Euskal WordNeten.

08162378 0.014 {**cost**} **"the total spent for goods [. . .] including money and. . . "**
PARLAY: *earnings*
01691640 0.011 {**horse**} **"solid-hoofed herbivorous quadruped domesticated. . . "**
BET ON: *pony*
c2c.subj
play 00605818
00017008 0.517 {**group, grouping**} **"any number of entities considered as. . . "**
PLAY: *The Mustangs, Texans, line. . .*
FIELD: *The Oriols*
textbf00004865 0.507 {person, individual, human} "a human being"
PLAY:*mate, Bill Kunkel, Nelson, youngman. . .*
START: *Haddix*
BET ON: *Berry*
00009469 0.079 {object, physical object} "a physical (tangible and visible) entity"
08413915 0.032 {**digit**} **"one of the elements that form a system of numbers"**
PLAY: *nine*
03953834 0.032 {idea, thought} "the content of cognition. . . "

Horrela, bada, troponimoak kontuan izanda, ia HM guztien jatorria lor dezakegu. Hau da, uler dezakegu makinak zein pauso jarraitu dituen HM horiek eskuratzeko. Dena den, oraindik geratu dira HM batzuk jatorria zehaztu gabe, letra lodiz ez dauden horiek, hain zuzen ere. Horiek nondik eskuratu diren ikertzeke dugu oraindik.

Orain arte, eskuratze automatikoan ematen diren pausoak azaldu ditugu. Hemendik aurrera eskuratze-teknika honen ebaluazio linguistikoaz jardungo gara. Zenbateraino fida gaitezke metodo honek egin duen eskuratzeaz?

Ebaluazio honekin hasi baino lehen, ekar dezagun gogora hasieratik eskuratze-teknika mota hauentzako proposatutako urre-patroiak, hauekin parekatu behar baititugu c2c HM hauek:

(23) **play 00605818 Objektuak**
w2c, c2c:
00240760 {sport, athletics} "an active diversion requiring physical exertion and..."
04771851 {contest, competition} "an occasion on which a winner is selected from..."
00254052 {game} "a contest with rules to determine a winner"
09065837 {amount of time, period, period of time} "time period a length of time"

play 00605818 Subjektuak
w2c, c2c:
00004865 {person, individual, someone, somebody, human soul} "a human being"
00017008 {group, grouping} "any number of entities (members) considered as a unit"

(24)n letra lodiz markatu ditugu zuzentzat jo ditugun HMak; beste guztiak okertzat jo ditugu:

(24) **c2c.obj**
play 00605818
00228990 0.215 activity "any specific activity" ONARGARRIA
00004865 0.117 person, individual, human "a human being"
00017008 0.102 group, grouping "any number of entities considered..."
00009469 0.071 object, physical object "a physical entity"
04771851 0.035 contest, competition "an occasion on..." ZUZENA
03875944 0.029 interest, involvement "a sense of concern with curiosity..."
08162378 0.014 cost "the total spent for goods [...] including money..."
01691640 0.011 horse "solid-hoofed herbivorous quadruped..."

c2c.subj
play 00605818
00017008 0.517 {**group, grouping**} **"any number of entities..." ZUZENA**
00004865 0.507 {**person, individual, human**} **"a human being" ZUZENA**
00009469 0.079 {object, physical object} "a physical entity"
08413915 0.032 {digit} "one of the elements that form a system of numbers"
03953834 0.032 {idea, thought} "the content of cognition"

Onargarri marka daraman bakarra activity objektu HMa da, eta hauxe da probabilitate-neurri handieneko HMa (0.215), berez, eskuratze-teknikak egokitzat proposatzen duena. *Synset* hau football, basketball eta abarren hiperonimoa da, baina tartean badaude HM gisa egokiagoak direnak, urre-patroian proposaturiko {sport, athletics}, adibidez. Hizkuntzalaritzari begira, activity klase semantikoa ezin da beti izan play 00605818ren objektua: ezin da edozein ekintzetan jokatu, baina bai, ordea, ekintza batzuetan (kirola adierazten duten ekintzetan, hain zuzen ere).

Objektuen artean zuzena den bakarra {contest, competition} objektu HMa da, eta hau probabilitate-neurriaren zerrendan ez da lehenengoetakoa (bosgarrena da).

Beste HM guztien jatorria ez da aditz-adiera honentzat egokia. Esate baterako, person HMa ez dagokio play 00605818ri baizik eta play 00610422ri. Azken *synset* honek objektu gisa [+pertsona] tasuna daramatenak hartzen ditu bere MCRko glosan adierazten den bezala (contest against an opponent). Zergatik azaltzen dira play 00610422ren HMak play 00605818koekin nahastuta? SemCorren etiketatze-erroreak daudelako, eta horren adibide play 00605818 eta play 00610422ren arteko nahasketa delako. Hau da, play kirol-adierarekin agertzen denean, SemCorren hau play 00605818 bezala etiketatu dute. Hortaz, SemCorreko play 00605818 *synset*eko HMetan play 00610422renak ere azaldu dira. VII.4.3 atalean azalduko ditugu errore hauen arrazoia sakonkiago.

Okerrak diren object eta digit HMen azalpena VII.4.3 atalean dago.

Azkenik, esan beharra dago troponimoetatik etorritako HM gehienak okerrak direla. Zuzenak direnak troponimo gabe lortu dira; play 00605818ren kasuan bet on, parlay eta stake bezalako troponimoak ditu, hots, apustua domeinuarekin zerikusia dutenak. Honenbestez, *play* domeinua dute, *sport*ekin batera. *Play* domeinuak indar gehiago duela dirudi eta honek HMetan eragina izan du. Hauen HMak play 00605818renekin zeharo ezberdinak dira. Esate baterako, aditz hauen objektu arruntenetako bat 'dirua' izango da (cost HMetan). Horse HMa, adibidez, bet on a pony testuingurutik dator. Beraz, ez dirudi aditz batek eta bere troponimoek HM berak dituztenik (behintzat MCR hierarkian oinarritzen bagara).

w2c SemCorretik

VII.3.2.1. atalean adierazi dugun bezala, eredu honekin aditz-formaren (hitzak izan ditzakeen adiera guztiak kontuan hartuta) objektu edo subjektu HMak lortzen dira. Beraz, gure adibidearekin jarraituz, HM hauekin play aditzaren adiera guztiak izan beharko ditugu kontuan. Hala ere, behin eta berriro esan dugun bezala, ikerlan hau kirol-domeinuko aditzetara mugatu dugu. Horregatik, nahiz eta w2c eskuratze-teknikan adiera guztiak kontuan hartu, adiera guzti horien artean guk kirol-adiera dutenak soilik hartuko ditugu kontuan. Horrela, eskuratze-teknika hau HMak kirol-adierarentzat bakarrik eskuratzen dituztenekin (c2c-ekin, adibidez) erkatu ahal izango dugu.

HM hauen ebaluazioa egin baino lehen, bakoitzaren jatorria ezagutzen saiatu gara, eta, berriro, s2s-ko datuak erabili ditugu[38].

Hala eta guztiz ere, w2c eskuratze-teknika honekin zaila da lotzea HM bakoitza bere jatorriarekin, ez baitakigu HM hori zein adierari dagokion. Esaterako, (26) adibidean begiratzen badugu, play 00605818ren subjektua izateko probabilitate handiena duen HM, {person, individual, human} *synset*ak adierazten duena da, [+pertsona] alegia. Hortaz, badakigu play 00605818k orokorrean subjektu gisa [+pertsona] adierazten duen izen bat hartuko duela. Baina, guk badakigu, play aditz-formaren adiera gehienek hartzen dutela subjektu mota hau: I play the piano, I play football, I play cards, I play Hamlet, eta abar.

SemCorreko s2s izen-zerrendari esker, HM bakoitzaren jatorria zehazteko gai izan gaitezke. s2s zerrendan dauden izen guztien hiperonimoak begiratuta zer HMetan bilakatu diren asma genezake. Baina lan honek gure saiakerari ez lioke abantaila handirik ekarriko, eta, gainera, erabilera konputazional mugatua lortuko genuke. Itzulpen automatikoan edo adiera desanbiguazioan, adibidez, w2c ez litzateke horren erabilgarria, aditz-forma baten aurrean ezingo genukeelako honen HMetatik bere adiera mugatu. Horregatik adiera batean oinarritzearen garrantzia.

HM hauetan adiera guztiak nahasturik daudenez, ezinezkoa zaigu aditz-adiera baten HMak ebaluatzea, aditz horren adiera posible guztiak kontuan hartuta daudelako. Horregatik, w2c motako HMak aztertzerakoan, play 00605818rekin zerikusia duten HMak ezberdintzen saiatu gara, gerora play 00605818rekin egindako beste eskuratze-tekniken emaitzekin bat datozen ikusteko. Hala, (26) adibidean play aditz-formaren w2c objektu/subjektu HMak ditugu. Letra lodiz markatu ditugu gure ustez play aditzaren kirol-adieraren objektu/subjektuak izan daitezkeenak, (25)eko urre-patroiekin bat datozenak, alegia. Urre-patroia bera edo antzekoa denean (hiperonimo edo hiponimo bat, adibidez), zuzen edo onargarri bezala kontsideratu dugu; baina bat ez datozenak ez ditugu okertzat hartu, hauek, berez, beste aditz-adiera baten HMak izan daitezkeen heinean, zuzenak izan daitezkeelako. Bestalde, HMen azpian SemCorreko play 00605818rekin batera corpusean agertu diren objektu/subjektu izenak zerrendatuak datoz.

[38] Ikus s2sko datuak C eranskinean.

(25) **play 00605818 Objektuak**
w2c, c2c:
00240760 {sport, athletics} "an active diversion requiring physical exertion and..."
00254052 {game} "a contest with rules to determine a winner"
04771851 {contest, competition} "an occasion on which a winner is selected from..."
09065837 {amount of time, period, period of time} "time period a length of time"

play 00605818 Subjektuak
w2c, c2c:
00004865 {person, individual, human} "a human being"
00017008 {group, grouping} "any number of entities (members) considered as..."

(26) **w2c.obj**
play
002289900.148 {**activity**} **"any specific activity or..." ONARGARRIA**
PLAY 00605818: *football, basketball,golf, game3...*
00004865 0.105 {person, individual, human} "a human being"
00009469 0.040 {object, physical object} "a physical (tangible and visible) entity"
00017008 0.031 {group, grouping} "any number of entities (members) considered..."
00018599 0.029 {communication} "something that is communicated between people..."
00021098 0.028 {action} "something done (usually as opposed to something said)"
00018966 0.008 {measure, quantity} "how much there is of something that you can..."
00015437 0.007 {state} "the way something is with respect to its main attributes"
00017586 0.007 {attribute} "an abstraction belonging to or characteristic of an entity"
04771851 0.006 {**contest, competition**} **"an occasion on..." ZUZENA**
PLAY: *game*

w2c.subj
play
00004865 0.308 {**person, individual, human**} **"a human being" ZUZENA**
PLAY: *mate, Bill Kunkel, Nelson, youngman...*
00017008 0.125 {**group, grouping**} **"any number of entities..." ZUZENA**
PLAY: *The Mustangs, Texans, line...*
00009469 0.059 {object, physical object} "a physical (tangible and visible) entity"
00012670 0.043 {abstraction} "a general concept formed by extracting common..."
06467898 0.029 {physical phenomenon} "a natural phenomenon involving the physics..."
08522741 0.016 {situation, state of affairs} "the general state of things"
08125923 0.011 {community} "common ownership"
00012878 0.008 {cognition knowledge} "the psychological result of perception..."

Ikus daitekeen bezala, urre-patroiko HM gehienak azaldu egiten dira. Subjektuen kasuan ez da harritzekoa, beste adieren subjektuek ere HM horiek onar baititzakete. Arrazoi horregatik daude probabilitate altueneko postuetan. Objektuen artean, kirolari bakarrik dagokion HMa {contest, competition} da, eskuratze-tekniken proposamenean azkena, probabilitate baxuenarekin agertu dena, alegia. Bestalde, objektuetan probabilitate handiena activityk du. Play 00605818k ekintza bat har dezake objektu gisa (activityk jasotzen dituen football, basketball, eta abar), baina aditz honen beste adieretan ere HM hau ager daiteke (play cards, adibidez).

s2semf SemCorretik

Eskuratze-teknika honek aditzaren adiera bakoitzarentzat HMak domeinu-eremu semantiko bikoteekin adierazten ditu. Honek orain arte erabilitako metodologia baldintzatzen du, ezin jakin baitezakegu zeintzuk diren HM zehatzak. Honen arrazoi nagusiena izen berak domeinu eta eremu semantiko bat baino gehiago har ditzakeela da. Esaterako, football izenaren domeinuak bi dira: *play* eta *sport*; eta bere eremu semantikoa *act* da. Hortaz, play-act eta sport-act bikoteak agertuz gero, HM desberdin hauek izen beretik abiatutakoak izan daitezke. Hala, gehienetan ezinezkoa zaigu ziurtasunez jakitea HM hauen corpuseko jatorri zehatza zein den.

Bestalde, bikote hauek adierazten dutena ulertzea ez da begibistakoa. Domeinuaren eta eremu semantikoen informazioa *synset*ena baino orokorragoa da eta gehienetan MCRra jo behar dugu hauen azpian zer dagoen ulertu ahal izateko.

Beraz, ezin dugu eskuratze-teknika honen ebaluazio sakon bat egin, baina s2s datuak aurrean izanda[39], subjektiboki bada ere, horietatik zuzenak zein diren aipa dezakegu.

Ebaluazioarekin hasi baino lehen, komeni da gogora ekartzea zeintzuk diren eskuratze-teknika mota honentzat proposatutako urre-patroiak:

(27) **play 00605818 Objektuak**
s2semf, w2semf:
sport-event
time period-time
sport-act
play-act

[39]Fitxategi honek jasotzen duen informazioa C eranskinean dago.

play 00605818 Subjektuak
s2semf, w2semf:
person-person
factotum-group

(28)n letra lodiz markatu ditugu zuzenak/onargarriak iruditu zaizkigun HMak:

(28) **s2semf.obj**
play 00605818
obj play-act 3.5 ZUZENA
obj sport-act 1.5 ZUZENA
obj baseball-artifact 1
obj factotum-Tops 1
obj card-artifact 1
obj play-artifact 0.5
obj golf-act 0.5 ONARGARRIA
obj anthropology-Tops 0.5
obj basketball-act 0.5 ONARGARRIA
obj sport-artifact 0.5

s2semf.subj
play 00605818
subj number-quantity 1
subj sport-person 1 ONARGARRIA
subj factotum-group 1 ZUZENA
subj factotum-Tops 1 ONARGARRIA
subj person-person 1 ZUZENA
subj biology-Tops 0.5
subj anthropology-Tops 1

Objektuen HMetako play-act, sport-act urre-patroietan daudenez ez dugu inolako zalantzarik zuzen bezala ebaluatzeko. Hauen zehaztapen gisa har daitezke golf-act eta basketball-act, domeinuen hierarkian golf eta basketball, *sport* domeinuen jasota baitaude. Arrazoi horregatik onargarri bezala hartu ditugu, urre-patroia baino zehatzagoak direlako. Urre-patroiko beste bi objektuen HMak ez dira s2semf HM hauetan agertu. Zuzen bezala ebaluatu ditugunak zerrendako lehenengo bi postuetan daude, onargarri gisa ebaluatutakoek, berriz, probabilitate gutxiago dute.

Azkenik, *artifact* eremu semantikoa daramatenen artean, nondik etorri diren susmatzen dugu; card-artifacten kasuan, play 00605818 aditzaren glosari erreparatuz gero, play cards bezalakoak onartzen dituela badakigu. Hortaz,

synset berean 'kartetan jokatu' eta 'futbolean jokatu' elkarrekin daudela dirudi. Card izenaren eremu semantikoa MCRn *artifact* da, eta arrazoi horregatik agertu da HM hori.

Beste HM bat play ball (play-artifact) dugu. Oraingo honetan ball izena football, basketball... bezala ulertu beharko genukeen, hots, ekintza bat bezala. Hala, *act* eremu semantikoa izan beharko luke eta ez *artifact*. MCRn kontsultatuz gero, ball *synset* ugaritan dago baina horietako batek ere ez du ekintza-adiera hori[40]. Beraz, eskuratze-teknikak horren ordez beste bat hartu du ausaz, *artifact* eremu semantiko duena, hain zuzen ere.

Subjektuei dagokionez, s2semf eskuratze-teknikak urre-patroian proposaturiko bi HMak lortu ditu. Horietaz gain, onargarri bezala ebaluatu ditugun sport-person eta factotum-Tops ere baditu. Lehenengoa, person-person horren zehaztapena da, eta honen jatorria mate izenaren agerpena izan daiteke, honen domeinua *sport* delako. Hala ere, errepikatu beharra dago HM hauen jatorria zehaztea ez dela lan batere erraza. Bigarrena, oso HM orokorra da[41] eta honen jatorria edozer izan daiteke.

Probabilitate altueneko subjektua, number-quantity HMa, ez da zuzena, baina honek c2c eskuratze-teknikako digit HMenarekin zerikusia duela uste dugu (azalpen zehatzagoa VII.4.3 atalean).

VII.4.2.2 BNCtik eskuratutako HMen azterketa eta ebaluazioa

Corpus honetan c2c eta w2c eskuratze-teknikak erabili dira. Erabilitako irizpide metodologikoa orain artekoaren ezberdina izan da. BNC corpusa ez dago adierekin etiketatua, hots, desanbiguatuta, ezta domeinuka antolatuta ere. Honek guztiak HMak nondik datozen zehaztea ezinezkoa egiten du. SemCorrekin eskuratze-teknikak aztertzerakoan, s2s (eta s2s-hype) fitxategiak genituen non aditzaren adierak (*synset*-zenbakia) zehaztuak zeuden eta baita izenenak ere. BNC semantikoki etiketatu gabeko corpusa da eta nahiz eta w2w fitxategi bat izan, bertan play aditz-formarekin objektu/subjektu gisa agertu diren hitzen zerrenda luze bat besterik ez zaigu ematen[42]. Mila hitzetik gora osatutako zerrendak dira, eta izugarrizko eskuzko lana litzateke bakoitzaren testuinguruak aztertu eta ki-

[40]Kontuan izan beharrekoa da, *WordNet* eta MCR etengabe eguneratzen dauden ezagutza-baseak direla, eta batzuetan horrelako hutsuneak aurki daitezkeela.

[41]Bikote honek ia edozer adieraz dezake, *factotum*ekin domeinurik ez duten hitzak adierazten direlako, eta *Tops* eremuak MCRko hierarkian oso goian dauden *synset*ak jasotzen dituelako. Beraz, oso orokorra den kontzeptu baten aurrean gaude.

[42]Ikus C eranskina.

rolaren domeinuari dagozkionak aukeratzea, gero horren arabera beraien MCRko *synset* eta hiperonimo posibleak zehazteko. Arrazoi horregatik, eta datu enpirikoetan oinarritu gabe, BNC gainean aplikatutako eskuratze-teknika hauen HMak zuzenean gure urre-patroiekin erkatu ditugu.

w2c BNCtik

Teknika honekin playren adiera guztien objektuen edo subjektuen HMak lortzen dira. Eskuratze-teknika honen HMak gure urre-patroiekin erkatu ditugu (ikus (29) adibidea), kirol-adierarekin bat datozenak nabarmentzeko –letra lodiz (30) adibidean. Urre-patroia bera edo antzekoa (hiperonimo edo hiponimo bat adibidez) denean zuzen edo onargarri bezala kontsideratu dugu hurrenez hurren; baina bat ez datozenak ez ditugu okertzat hartu. Izan ere, hauek, berez, beste aditz-adiera baten HMak izan daitezkeen heinean, zuzenak izan daitezke.

(29) **play 00605818 Objektuak**
w2c, c2c:
00240760 {sport, athletics} "an active diversion requiring physical exertion..."
04771851 {contest, competition} "an occasion on which a winner is selected from..."
00254052 {game} "a contest with rules to determine a winner"
09065837 {amount of time, period, period of time} "time period a length of time"

play 00605818 Subjektuak
w2c, c2c:
00004865 {person, individual, human} "a human being"
00017008 {group, grouping} "any number of entities (members) considered as a unit"

(30) **w2c.obj**
play
00228990 0.082 activity "any specific activity or..." ONARGARRIA
00009469 0.077 object, physical object "a physical (tangible and visible) entity"
00004865 0.070 person, individual, human "a human being"
00012670 0.028 abstraction "a general concept formed by ..."
00021098 0.020 action "something done (usually opposed to something said"
00597858 0.012 group action "action taken by a group of people"
00012878 0.012 cognition, knowledge "the psychological result of perception..."
04771851 0.009 contest, competition "an occasion on..." ZUZENA
05650477 0.009 part, piece "a portion of a natural object"
04690182 0.008 happening, occurrence, natural event "an event that happens"

w2c.subj
play
08813320 0.16 helium "a very light colorless element that..."
00004865 0.12 person, individual, human "a human being" ZUZENA
04455766 0.06 he "the 5th letter of the Hebrew alphabet"
00011607 0.04 artifact, artefact "a man-made object"
05149489 0.03 organization, organisation "a group of...' ONARGARRIA
04313427 0.02 message, content, subject "what a communication that is about..."
00016649 0.01 act, human action, "something that people do or cause to happen"
00018966 0.01 measure, quantity, "how much there is of something that..."
00014314 0.01 location "a point or extent in space"
00012878 0.01 cognition, knowledge "the psychological result of perception..."

Ikus daitekeen bezala, urre-patroiko HM gehienak azaltzen dira. Objektuen artean, kirolari dagokion HM bakarra {contest, competition} da. Onargarri marka daraman HMa (activity) urre-patroiko {sport, athletics}en hiperonimoa da. Nahiz eta play 00605818k ekintza bat har dezakeen objektu gisa (activityk jasotzen dituen football, basketball eta abar), beste adieretan ere HM hau ager daiteke (He played Hamlet esaldian, adibidez), eta horregatik du probabilitate-neurri altuena.

Subjektuen kasuan, {organisation, organization} onargarritzat jo dugu, {group, grouping} *synset*aren hiponimo bat delako, talde mota zehatzagoa, alegia. Zuzentzat hartu dugun bakarra (eta probabilitate-neurri altuenetakoa duena) person HMa da. Hau baino probabilitate-neurri handiagoa he izenordainak du, baina honi egotzi zaizkion *synset*ak ez dira izenordainak. Aurreprozesu lanetan ez zirenez izenordainak markatu, analizatzaile sintaktikoak ez ditu detektatzen, eta, gainera, MCRn izenordainik ez dagoenez, makinak he izenordainaren idazkera antzekoa duten beste bi *synset*ekin parekatu ditu —helium ('elementu kimikoa') eta he ('hebrear alfabetoko bosgarren letra'). Arrazoi horregatik dira probabilitate handiena dituzten HMak. Honi buruz, VII.4.3 atalean mintzatuko gara.

Bestalde, location bezalako subjektu HMak agertzen direnean, eta w2w fitxategietan begiratuta, leku izen berezietatik etor daitezkeen (Argentina, Madril...) susmoa dugu. Horrelakoekin corpusean kirol taldeak adierazi nahi dira eta MCRn leku-izen berezi bezala daude. Hori dela eta, location bezalako HMak ditugu play aditzarekin.

Beraz, kirol-adierari dagokion HM bakarra {contest, competition} dela dirudi.

c2c BNCtik

Eskuratze-teknika honek lortzen dituen objektu edo subjektuen HMak play 00605818 adierarako dira (ikus VII.4.2.1. atala).

(31)n dugun urre-patroiekin erkatuta, (32)n letra lodiz markatu ditugu zuzenak iruditu zaizkigun HMak; beste guztiak okerrak dira:

(31) **play 00605818 Objektuak**
w2c, c2c:
00240760 {sport, athletics} "an active diversion requiring physical exertion..."
04771851 {contest, competition} "an occasion on which a winner is selected from..."
00254052 {game} "a contest with rules to determine a winner"
09065837 {amount of time, period, period of time} "time period a length of time"

play 00605818 Subjektuak
w2c, c2c:
00004865 {person, individual, someone, somebody, human soul} "a human being"
00017008 {group, grouping} "any number of entities (members) considered as a unit"

(32) **c2c.obj**
play 00605818
09065837 0.006 {**period, amount of time**} **"an indefinite length..." ZUZENA**
08813320 0.004 {helium} "a very light colorless element that..."
08520394 0.004 {condition, status} "a condition or state at a particular time"
08534455 0.001 {status, position} "the relative position of persons in a society"
08745609 0.001 {opportunity, chance} "a possibility due to a favorable..."
08522741 0.001 {situation, state of affairs} "the general state of things"
08781633 0.001 {material, stuff} "the tangible substance that goes into..."
08523811 0.0007 {relationship} "a state involving mutual dealings..."
09164158 0.0006 {**playing period, play**} **"time during..." ONARGARRIA**

c2c.subj
play 00605818
08813320 0.14 {helium} "a very light colorless element that..."
09065837 0.005 {period, amount of time} "an indefinite length of time"
08520394 0.003 {condition, status} "a condition or state at a particular time"
09069911 0.002 {now} "the momentary present"
08807415 0.001 {metal} "any of several chemical elements that..."
08534455 0.001 {status, position} "the relative position of persons in a society"
08525534 0.001 {friendship, friendly, relationship} "the state of being friends"
08781633 0.001 {material, stuff} "the tangible substance that goes into..."
08522741 0.001 {situation, state of affairs} "the general state of things"

Objektuaren HMetan denborazkoak bakarrik harrapatu ditu, bata zuzena (zerrendatik probabilitate-neurri handiena duena, gainera) eta bestea onargarria (aurrekoaren hiponimo bat). Eta subjektuaren HMetan ez du bat bera

ere harrapatu. Berriro ere, aipatu behar dugu, subjektuaren HMetako helium *synset*a ingeleseko he izenordainari dagokiola, eta hauxe dela subjektu HMen artean probabilitate-neurri altuena duena.

Horrela bada, eskuratze-teknika honen emaitzak ez dira batere onak izan. Corpusarengatik izan daiteke (etiketatua ez egotea, kirol domeinukoa bakarrik ez izatea...), baina, hala ere, harritzekoa da subjektuetan HM zuzen bat bera ere ez lortzea, subjektuen HMen eskuratzean aukerak askoz gutxiago izanik (aditzen objektuak mota askotakoak izan daitezke; aditzen subjektuak, aldiz, askotan [+pertsona] dira). Objektuekin ere harritzekoa da kirol domeinuan arruntak diren {contest, competition} edo {sport, athletics} objektu HMen ordez denborazkoak bakarrik eskuratu izana. Arrazoi posible bat izan daiteke, kirol-domeinuari buruz aritzean, kirol-ekintzari buruzko informazioa inplizitua egotea, irakurleak informazio hori ez duelako behar testua ulertzeko. Hala, nahiz eta testuan bertan ez zehaztu (Liverpool will play next match on Wednesday), irakurleak badaki "zertan" jokatzen duten albisteko protagonistek (kasu honetan, irakurleak badaki Liverpool futboltalde bat dela, eta ondorioz, futbolean jokatuko dutela).

Bestalde, troponimoen eraginak zerikusirik baduela pentsa dezakegu. Baina SemCor ez bezala, BNC etiketatu gabeko corpusa denez, oso zaila egiten zaigu hipotesi hori zehatz-mehatz egiaztatzea.

VII.4.2.3 EFEtik eskuratutako HMen azterketa eta ebaluazioa

EFE domeinuka antolatutako corpusa da, eta guk kirol-domeinuari dagokion atala erabili dugu saiakera honetarako. Corpus honetan w2semf eskuratze-teknika aplikatu dugu. Aipatu dugun bezala, teknika honek eskuratzen dituen HMak aditz-forma osoarentzat dira, aditzaren adiera guztientzat, alegia. Gogoratu probabilitate kopuru altuenetik baxuenera ordenaturiko domeinueremu semantikoen bikoteak direla.

BNCren antzera, corpus hau ez dago semantikoki etiketatuta, eta horrek HMen jatorria zehaztea zaildu egiten du. Corpus honetarako ere w2w tresnatxoa sortu da. Honi esker fitxategi batean EFE corpuseko kirol domeinuan play aditz-formarekin agertu diren hitzen zerrenda dugu, hauen maiztasunaren arabera ordenaturik[43].

Hirurehun hitzetik gorako zerrendak dira, eta izugarrizko eskuzko lana litzateke bakoitzaren testuinguruak aztertu eta kirolaren domeinuari dagoz-

[43]Fitxategi honek jasotzen duten informazioa C eranskinean dago.

kionak aukeratzea, gero horren arabera beraien MCRko *synset*, eremu semantiko eta domeinu posibleak zehazteko.

Honekin batera, corpus honekin erabili dugun w2semf eskuratze-teknikak ematen dituen HMek ez dute laguntzen HMen jatorria bilatzen. Izan ere, ez dira ulerterrazak, hau da, domeinuak eta eremu semantikoen informazioa *synset*ena baino orokorragoa da, eta gehienetan MCRra jo behar dugu hauen azpian zer *synset* jasotzen diren jakiteko. Gainera, hitz berak domeinu eta eremu semantiko bat baino gehiago har ditzake (VII.3.2.2 atalean ikusi dugun bezala). Honezaz gain, EFE corpusean erabilitako eskuratze-teknikak aditzforma osoa hartzen du kontuan.

Arrazoi hauengatik guztiengatik, eta datu enpirikoetan oinarritu gabe, zuzenean EFE gainean aplikatutako eskuratze-teknika hauen HMak gure urre-patroiekin erkatu ditugu.

w2semf EFEtik

Eskuratze-teknika honentzat proposatu ditugun urre-patroiak daude (32)n, eta (33)n play aditzaren w2semf objektu/subjektu HMak ditugu (letra lodiz gure ustez play 00605818 aditzari dagozkionak):

(32) **play Objektuak**
w2semf:
sport-event
time period-time
sport-act
play-act

play Subjektuak
w2semf:
person-person
factotum-group

(33) **w2semf.play.kirola.obj**
obj x 100
obj play-act 50.013 ZUZENA
obj factotum-act 30.390 ONARGARRIA
obj time period-time 29.009 ZUZENA
obj zoology-animal 25.2
obj factotum-artifact 25.026
obj sport-event 23.514 ZUZENA
obj sport-act 23.038 ZUZENA
obj number-quantity 22.957
obj geography-location 16.918

w2semf.play.kirola.subj
subj x 372 ONARGARRIA
subj administration-group 168.64
subj chemistry-substance 52.66
subj sport-group 44.01 ONARGARRIA
subj zoology-group 40.5
subj linguistics-communication 38.72
subj physics-substance 34.66
subj geography-location 33.35
subj administration-location 32.31
subj number-quantity 26.64

Urre-patroiaren antzekoa (domeinu edo eremu semantiko orokorrago edo zehatzago bat edo urre-patroi bera duenean, adibidez) denean zuzen edo onargarri bezala kontsideratu dugu (esaterako, sport-group). HM batzuk zalantzan jar daitezke. Sport-groupen kasuan ez dago dudarik kirol-adierarekin zerikusia duela; administration-groupen kasuan, nahiz eta lehenengo begiratuan okerra zela iruditu, w2w zerrendak eta corpusak aztertuz, konturatu ginen Colombians, Brazilians eta abar bezalako agerpenetatik zetorrela. Izen hauen domeinua MCRn *administration* da. Horregatik dugu administration-group bezalako HM bat. Hala ere, okertzat jo dugu, VII.4.1 atalean finkatutako irizpidearengatik: domeinu-eremu semantiko bikote bat onargarritzat hartuko dugu, urre-patroia baino orokorrago edota zehatzago bada, **eta domeinuko beste izen gehienak aditz horren argumentu izan badaitezke**. Argi dago administration-group HMak ez duela azken baldintza hau betetzen. Administration-group HMa onargarritzat joz gero *administration* domeinuaren azpian dauden beste hitz guztiak ere (chairman, chancellor...) jokatu aditzaren (kirol-adieraren) subjektu/objektu prototipiko gisa ager daitezkeela baieztatzen ariko ginateke. Hori, bistan da, ez litzateke oso egokia.

Bestalde, gogorazi beharra dago eskuratze-teknika honek izen bereziak x batez adierazten dituela.

Aditzaren adiera guztiak kontuan hartzen dituen eskuratze-teknika izateko, kirolari dagozkion HM ugari daude. Urre-patroiko objektu HM guztiak daude eta oso probabilitate-neurri altuekin, gainera. Dirudienez, eta aditzforman oinarritutako beste eskuratze-tekniken emaitzekin erkatuz gero, kirol domeinuan oinarritutako corpus baten gainean aritzeak badu eraginik. Izan ere, neurri txikiagoan agertuko dira kirol-domeinukoak ez diren adierak.

Orain arteko eskuratze-teknikekin aipatu ditugun erroreak ikus daitezke w2semf honetan ere (gero VII.4.3 atalean azalduko ditugunak). Esate baterako, ingeleseko he eta heliumen arteko nahasketa. Subjektu HMetan

chemistry-substance eta physics-substance bezala ageri da. Beste adibide bat, leku-izen bereziak (Argentina, Madril...) –geography-location bezala eskuratzen direnak– eta kirol taldeen izen berezien arteko nahasketa da (Argentina played well).

Hala eta guztiz ere, eskuratze-teknika honekin aurrekoekin detektatu ez dugun errore mota bat aurkitu dugu (anbiguotasuna), hurrengo atalean azalduko duguna.

VII.4.3 Erroreen azterketa

Eskuratzean erroreak badaudela ikusi dugu, eta hauek, batez ere, etiketatu gabe dauden corpusetatik datoz. Errore hauek kontuan izan beharrekoak dira eskuratze-teknikak findu ahal izateko. Horregatik, horien guztien berri emango dugu atal honetan.

Atal honetan ez gara troponimiaz eta aditzaren adiera guztietan oinarritzen diren eskuratze-teknikez (c2c, w2c eta w2semf) jardungo, azterketan zehar hauek sortzen dituzten arazoak aipatu ditugulako.

VII.4.3.1 Etiketatze-erroreak

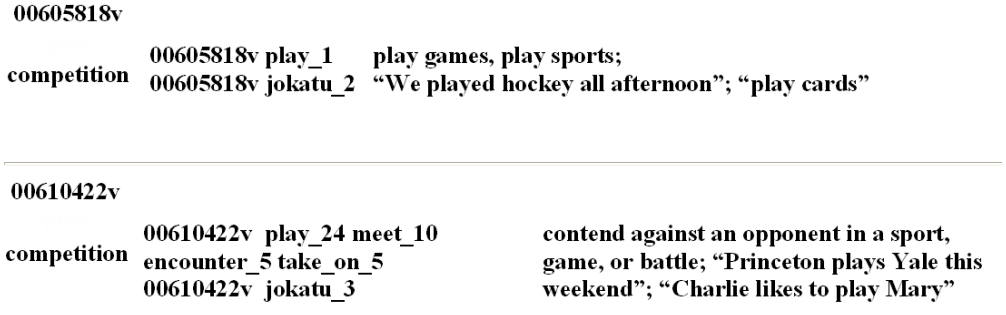
Errore mota hau SemCor corpusean bakarrik gertatu da, hau baita erabili dugun corpus etiketatu bakarra. Eskuz etiketatutako corpusa izan arren, etiketatze-erroreak gertatzen direla nabarmendu beharra dago. Esate baterako, arraroa badirudi ere, SemCorren play 00605818 eta play 00610422 (ikus VII.2 irudiko glosak) ez dituzte bereizi, hau da, play aditzaren agerpen guztiak play 00605818 *synset*arekin etiketatuak daude. Hortaz, (34) bezalako esaldiak, nahiz eta berez play 00610422ren adibide bat izan, play 00605818 gisa hartzen dira.

(34) SMU will play **the Owls** at Rice Stadium in Houston.

Nahasketa horrek objektuaren HMetan ondorioak izan ditu. Hala nola, play 00605818ren objektuen arten person eta group ageri zaizkigu, [+gizaki] tasuna daramatenak, hain zuzen ere. Objektu mota hauek play 00610422ren HMak izan beharko lukete.

Etiketatze-erroreak ez dira aditzekin bakarrik gertatzen, izenenekin ere gertatzen dira.

(35) Our interior **line** and out linebackers played exceptionally well.



VII.2 Irudia: jokatu aditzaren bi kirol *synset*ak.

(36) For a serious **young man** who plays golf with a serious intensity.

(35)en kasuan line linebacker izenaren (futbol jokalaria) laburdura bat da, eta a formation of people (pertsonen errenkada, multzoa) adierarekin etiketatua dago.

(36)ko young man "a man who is the lover of a girl or young woman" bezala etiketatu dute, hots, euskarako 'mutil-lagun' adierarekin, "an adolescent male" adierarekin etiketatu ordez.

Hala ere, bi adibide hauek, subjektuaren HMetan ez dute eragin handirik izan. Beraien hiperonimoak group eta person direnez, makinak HM horietan bilakatu ditu; urre-patroian zuzentzat definitu ditugunak.

VII.4.3.2 Falta diren adierak

HMak MCRn oinarrituta adierazi ditugu (corpuseko izenen *synset*en hiperonimoak edota domeinu eta eremu semantikoak erabilita). Gerta liteke MCRn adiera-inbentarioan baten bat ez egotea. Esate baterako, football, basketball... bezala uler daiteke ball ingelesez, ekintza bat bezala, alegia:

(37) I play football/basketball/ball...

MCRn kontsultatuz gero, *synset* ugaritan dago ball, baina horietako batek ere ez du ekintza-adiera hori. SemCor etiketatzerakoan, antzekoena izan zitekeen beste *synset* batekin etiketatu behar izan zuten.

(38) 02103632 ball "round object that is hit or thrown or kicked in games"

Makinak corpusean ball izena 02103632 bezala (ikus (38) adibidea) topatzen badu play 00605818 horren objektu gisa, honen HMa eskuratzeko zuzenean hiperonimora joko du, eta {sport, recreation}en (edo sport-act domeinu-eremu semantikoaren) ordez, object *synset*a (play-artifact domeinu-eremu semantikoa) lortzen du objektu HM gisa.

EFE eta BNCn, semantikoki etiketaturik ez dauden corpusetan, antzeko prozesua gertatzen da. Makinak corpusean ball izena topatzen duenean play 00605818ren objektu gisa, eta honen HMa eskuratu behar duenean, MCRtik ball 'ekintza' adierazten duen horren ordezko bat hartzen du, 'objektu' adiera duena hain zuzen ere. Hala, honen hiperonimotik abiatuta object *synset*a (edo *artifact* eremu semantikoa) lortzen du objektu HM gisa, berez dagokion {sport, recreation} *synset*aren (edo sport-act domeinu-eremu semantikoaren) ordez.

Antzeko beste adibide bat, leku-izen bereziak dira (Argentina, Madril eta abar). MCRn leku-izen berezi bezala bakarrik daude, baina corpusean hauekin kirol-taldeak adierazi nahi dira. Hori dela eta, location edo geography-location bezalako HM okerrak ditugu play 00605818 aditzarekin.

VII.4.3.3 Anbiguotasuna

Gure ustez, hau izan daiteke HMen eskuratzean gehienetan gerta daitekeen fenomenoa; semantikoki etiketatu gabeko corpusen gainean aritzean, noski. Baina, errore hau antzematen zailenetakoa da.

Corpuseko izenek adiera bat baino gehiago izan dezakete, eta semantikoki etiketatu gabe daudenean, eskuratze-teknikak adiera horietako bat aukeratu behar du MCRtik. Gerta daiteke ez dagokion adiera aukeratzea, eta, ondorioz, zuzena ez den HMa sortzea. Esate baterako, ingeleseko game izenak bost adiera ditu MCRn:

a. 00254052 {game_1} "a contest with rules to determine a winner"

b. 00254326 {game_2} "a single play of a game; the game lasted 2 hours"

c. 00256308 {game_3} "an amusement or pastime"

d. 01485683 {game_4} "animal hunted for food or sport"

e. 00341531 {game_5} "informal terms for your occupation"

Kirol-adierak lehenengo biak izan daitezke (a eta b). VII.4.2.3 atalean aztertutako HMen artean zoology-group eta zoology-animal bezalakoak genituen, eta okerrak bezala ebaluatu ditugu. Horien atzean anbiguotasunaren arazoa dago, makinak game izena game_4 bezala etiketatu du ('animalia' bezala, alegia), eta ondorioz, *synset* horren HM gisa lortu dira HM okerrak (ikus 21. eta 26. adibideak).

VII.4.3.4 Analizatzaile sintaktikoak eragindako erroreak

VII.3.2.1. atalean ikusi dugun bezala, aditz baten HMak eskuratzeko, lehenengo corpusaren gainean Minipar analizatzailea edo analizatzaile sintaktikoa (Lin, 1993) erabili dugu. Analizatzaile sintaktikoak errore batzuk izan ditzake, eta ondorioz, honek HMetan eragina izan du. Honen adibide argi bat da play 00605818ren (39)ko subjektuaren HMa; (40) adibidean honi dagokion SemCorreko jatorria dugu:

(39) 08413915 0.032 {digit} "one of the elements that collectively forms..."

(40) **Nine of the league's teams** play in baseball parks and therefore...

Subjektuaren burua ez da nine, baizik eta teams, baina analizatzaile sintaktikoak nine zenbakia hartu du burutzat, eta horregatik dugu honen hiperonimoa subjektuaren HM gisa.

VII.4.3.5 Izen berezien ezagutza eta anaforaren ebazpena

Bi errore hauek eragotziko lirateke hauen ezagutzarako prozesu informatikoren bat erabili izanez gero. Esate baterako, entitateen ebazpenarekin corpuseko izen bereziak pertsona-izen, erakunde-izen edo talde-izen bezala sailkatuko lirateke, hauetatik MCRko lotura egin daitekeelarik.

Anaforak berarekin informazio linguistiko asko darama, baina hau ezin da eskuratu baldin eta corpus batean semantikoki etiketaturik ez dagoen. Aipatu dugu subjektuaren HM batzuetan agertutako helium ('elementu kimikoa') eta he ('hebrear alfabetoaren bosgarren letra'), ingeleseko he izenordainarekin nahasten direla. MCRn ez daudenez izenordainak, makinak izenordain horren antzekoak diren beste bi *synset*ak aukeratzen ditu. Hortik, HM okerrak izatea. Anafora automatikoki landu izanez gero, anaforaren aurrekariaren informazioa jaso ahal izango genuke, eta honela, horrelako erroreak desagertuko lirateke.

VII.4.4 Ebaluazioaren azterketa

Play 00605818n oinarrituta, pausoz pausoz azaldu dugu ingeleseko aditzekin egindako ikerlana. Hainbat eskuratze-teknika aipatu ditugu, eta hauetako askok corpus ezberdinetan (SemCor, BNC eta EFE) objektu eta subjektuentzat zer nolako HMak eman dituzten ere aztertu dugu. Ebaluazio honen laburpenaren berri VII.5 taulan ematen dugu, hau da, corpus bakoitzean erabili den eskuratze-teknika bakoitzetik play 00605818ren zenbat objektu/subjektuen HM diren zuzenak (urre-patroiarekin bat datozenak), zenbat diren onargarriak (urre-patroiaren hiperonimo edo hiponimoak direnak) eta urre-patroietatik zenbat ez diren eskuratu (eskuratu gabe bezala izendatu ditugunak)[44]. Datu hauek kopuru zehatzak erabiliz adierazi ditugu; esaterako, eskuratze-teknika bakoitzaren objektu/subjektuen HMetatik (gehienez hamar) zenbat diren zuzenak edo onargarriak zenbakitu ditugu; eta baita eskuratze-teknika bakoitzarentzat proposatutako urre-patroietatik zenbat geratu diren eskuratu gabe ere. Taula bat egin dugu saiakera honetan erabilitako kirol-aditz bakoitzarentzat, hots, MCRtik aukeratutako zortzi *synset*entzat (00605818 {play_1/jokatu_2}; 00610422 {encounter5, meet10, play24,take on5/jokatu3}; 00468052 {coach_2, train_7/entrenatu_1}; 00059698 {train_8/entrenatu_3}; 00630097 {equalize_1, get even_1/berdindu_16}; 00630097 {draw_25, tie_2/berdindu_15}; 00620486 {win_1/irabazi_3}; 00620218 {lose_2/galdu_9})[45]. VII.5 taularen antzeko eredua jarraituta, ingeleseko aditz guztiak kontuan hartuta lortu diren emaitzak ditugu VII.6 taulan, oraingoan ehunekotan adierazita.

VII.6 taulan eskuratu gabeen zerrendan datu azpimarragarriena % 0 zenbakira hurbiltzen dena da, honek eskuratze-teknikak urre-patroiko HM guztiak lortu dituela esan nahi duelako. Emaitzek adierazten dutena ulerterra-

[44]Domeinu-eremu semantiko bikoteen ebaluazioan erabilitako irizpide nagusia VII.4.1 atalean aipatu dugu. Honekin batera, eskuratu gabeak diren ala ez neurtzeko, beste irizpide batzuk finkatu ditugu: batetik, zuzen/onargarri bezala ebaluatutako HM batekin, bi urre-patroi eskuratu daitezke. Adibidez, play 00605818ren objektuen urre-patroiak (domeinu-eremu semantiko bikoteentzako) play-act, sport-act, sport-event eta time period-time badira, eta eskuratze-teknikaren emaitza sport-act bada, aurreko lau urre-patroietatik bi (sport-act eta play-act) eskuratu direla esaten dugu, *act* eremu semantikoa daramaten biak, hain zuzen ere. Gauza bera, factotum-act HMarekin. Eta bestetik, alderantziz ere gerta daiteke, onargarritzat jo dugun HMa eskuratu gabea bezala ebaluatzea; esate baterako, izen bereziak (x baten bidez adieraziak datozenak), pronominalak (pro baten bidez adieraziak datozenak), eta factotum-Tops bikotea.

[45]Taula hauek guztiak C eranskinean daude ikusgarri.

zagoa egitearren, zuzenak/onargarriak kopuruen batura ere adierazi dugu eta taulan *Batura z/o* bezala izendatu dugu. Zuzen eta onargarrien zerrendan, aldiz, datu nabarmenenak % 100era gerturatzen direnak dira, eskuratze-teknikak eskuratutako HM guztiak zuzenak/onargarriak direla adierazten duelako. Taula hauek aurrean izanda, hurrengo atalean, hauetatik ondoriozta ditzakegun emaitzak komentatuko ditugu.

		Objektua			***Subjektua***		
Jatorria	*Teknika*	*Zuzena*	*Onargarria*	*Eskuratu gabe*	*Zuzena*	*Onargarria*	*Eskuratu gabe*
SemCor	w2c	10etik 1	10etik 1	4tik 1	5etik 2	0	0
SemCor	c2c	8tik 1	8tik 1	4tik 1	5etik 2	0	0
SemCor	s2semf	10etik 2	10etik 3	4tik 2	7tik 2	7tik 2	0
BNC	w2c	10etik 1	10etik 1	4tik 1	10etik 1	10etik 1	0
BNC	c2c	10etik 1	10etik 1	4tik 3	0	0	2tik 2
EFE (kirola)	w2semf	10etik 4	10etik 1	0	0	10etik 1	2tik 1

VII.5 Taula: Corpus ezberdinetatik play 00605818rentzat eskuratutako HMen emaitzak.

		Objektuak				***Subjektuak***			
Jatorria	*Tek.*	*Zuz.*	*Onarga.*	*Batura z/o*	*Eskuratu gabe*	*Zuz.*	*Onarga.*	*Batura z/o*	*Eskuratu gabe*
SemCor	w2c	% 16,3	% 18,5	% 34,8	% 29,5	% 26,6	% 9	% 35,6	% 18,1
SemCor	c2c	% 6,9	% 26,4	% 33,3	% 44	% 38	% 7,1	% 45,1	% 3,5
SemCor	s2semf	% 14,2	% 42,8	% 57	% 64,2	% 7	% 37,6	% 44,6	% 60
BNC	w2c	% 9	% 13,6	% 22,6	% 15,9	% 11,1	% 6,3	% 17,4	% 13,6
BNC	c2c	% 1,4	% 0	% 1,4	% 96,4	% 0	% 0	% 0	% 100
EFE (kir.)	w2semf	% 14,1	% 10	% 24,1	% 45,4	% 2,7	% 21,8	% 24,5	% 41

VII.6 Taula: Kirol-aditz guztientzat, corpus eta eskuratze-teknika ezberdinak erabiliz, lortutako emaitzak.

VII.4.4.1 SemCorretik eskuratutako HMak

Corpus honetatik hiru HM mota jaso ditugu:

- **w2c:** Eskuratze-teknika honek aditz-forma osoa kontuan hartzen duenez, zehazten zaila da zein HM diren kirolaren domeinuari dagozkionak. Urre-patroiarekin bat etorri direnak kontsideratu ditugu domeinu horretakoak. Horregatik, urre-patroietatik gutxi geratzen dira eskuratu gabe, baina zuzen eta onargarrien kopurua ez da oso handia.

- **c2c:** Teknika honen emaitzak w2c-en antzekoak badira ere (esate baterako, c2c-en *Batura z/o* objektuen kasuan,% 33,3a da eta w2c-en % 34,8a), eta kontuan izanda eskuz etiketatutako (desanbiguatutako) corpusa dela, ez dira espero bezain emaitza onak, lortutako HM gehienak okerrak baitira. Dena den, w2c-ek baino zuzen eta onargarri gehiago lortzen ditu eta eskuratu gabeen kopurua antzekoa da, objektuen eta subjektuen kopuruen batura kontuan hartzen badugu. HM okerrak lortzearen arrazoia corpuseko etiketatze-erroreetan, analizatzaile sintaktikoaren analisi okerrean, eta corpusean agertu diren baina MCRn ez dauden adieretan egon daiteke.

  Bestalde, errore asko troponimoetatik datoz. Zuzentzat jo ditugunak troponimoak kontuan izan gabe lortu dira. Troponimia kontuan hartuta domeinu eta ezaugarri desberdinak hartzen dituzten aditzak nahasten direla ikusi dugu. Esate baterako, aztergai izan dugun play 00605818ren kasuan, honek bet on, parlay eta stake bezalako troponimoak ditu, hots, apustua domeinuarekin zerikusia dutenak. Hauen HMak play 00605818-rekin zeharo ezberdinak dira. Esate baterako, aditz hauen objektu arruntenetako bat 'dirua' izango da (cost HMetan). Beraz, ez dirudi aditz batek eta bere troponimoek HM berdinak dituztenik (behintzat, MCR hierarkian oinarritzen bagara).

  Bestalde, aipagarria da eskuratze-teknika honek subjektuekin eman dituen emaitza onak, eskuratu gabe % 3,5a bakarrik utzi baitu. Honen arrazoia corpus etiketatua izatea da. Hau da, entitateak landuta eta semantikoki etiketatuta daude, eta eskuratze-teknikak ez ditu desanbiguatu behar.

  Objektuetan ez dira emaitza hain onak lortzen eskuratu gabeei dagokienez, objektu HMen kopurua subjektuen HMena baino handiagoa delako. Honen erakusle garbia da bakoitzaren urre-patroien kopurua

(playren kasuan, subjektuek, oro har, bi HM dituzte, eta objektuek, aldiz, lau).

- **s2semf:** HM hauek domeinu-eremu semantiko bikoteekin definitua datozenez, eta hitzak domeinu edo eremu semantiko bat baino gehiago izan ditzakeenez, batzuetan zaila da zehazten corpuseko zein agerpenetan dagoen HM hauen jatorria, eta, ondorioz, ezinezkoa zaigu zuzenak diren ala ez jakitea. Hori dela eta, eskuratze-teknika honen ebaluazio subjektiboago bat egin dugu. VII.5 taulako emaitzei erreparatuz, aurreko biak baino HM hobexeak lortzen dituela esan genezake. VII.6 taulan, aditz guztiak kontuan hartuta, ezberdintasuna ez da horrenbestekoa: zuzen eta onargarrien batura altua (% 57 eta % 44,6) da, baina baita eskuratu gabeena ere (% 64,2 eta % 60).

VII.4.4.2 BNCtik eskuratutako HMak

Semantikoki etiketatu gabeko corpus honen gainean w2c eta c2c eskuratze-teknikak erabili ditugu.

- **w2c:** Teknika honen HMak, aditzaren adiera guztietan oinarritzen direnez, zein adierari dagozkion asmatzen oso zaila da, baita hauen jatorria aurkitzea ere. Honenbestez, BNCren gainean aplikatuta HM batzuk lortu ditu (objektuen *Batura z/o* % 22,6a eta subjektuena % 17,4a), baina hauek SemCorren gainean lortutakoak baino kalitate baxuagoa dutela nabarmendu behar da. Izan ere, aipatu dugunez, w2c teknikek adiera guztiak hartzen dituzte kontuan. Bestalde, eskuratu gabeen kopuru txikiena honek du.

- **c2c:** Teknika honek espero baino emaitza okerragoak eman ditu, play 00605818ren HM bakarra asmatu baitu, eta beste aditz guztiekin ere hala-moduzko emaitzak izan ditu (ikus VII.6 taula). Corpusaren osaerak izan dezake eraginik honetan. Izan ere, gogora dezagun corpus hau ez dagoela etiketatua eta kirol domeinuarena bakarrik ez dela, besteak beste. Bestalde, troponimoen eraginak zerikusirik duela pentsa dezakegu, baina SemCor ez bezala, BNC etiketatu gabeko corpusa denez, oso zaila egiten zaigu hipotesi hori zehatz-mehatz egiaztatzea. Teknika hau, berez, corpus ez-etiketatuekin edo domeinu batera mugatua ez dauden corpusekin ez dela oso erabilgarria esan daiteke.

VII.4.4.3 EFEtik eskuratutako HMak

Kirol-domeinuko eta semantikoki etiketatu gabeko corpus honetan w2semf eskuratze-teknika erabili da.

- **w2semf:** Nahiz eta HM hauek aditzaren adiera guztientzat izan, teknika honekin emaitza onak lortu dira. SemCorreko w2c eta c2c-ekin alderatuz, corpus honetan w2semf-en zuzen/onargarrien batura txikiagoa bada ere (% 24,1 eta % 24,5, objektu eta subjektuei dagozkienak, hurrenez hurren), kontuan izanda eskuz etiketatu gabeko corpusa dela, azpimarratu beharreko emaitzak dira. Corpusaren domeinuak (kirola) beste adierak baztertzen lagundu duela dirudi. Dena dela, esan beharra dago, eskuratu gabeen kopurua ere handi xamarra dela.

VII.4.5 HMen erkaketa

VII.5 eta VII.6 tauletatik abiatuta, batetik eskuratze-teknikak erkatuko ditugu, eta bestetik corpusak.

VII.4.5.1 Eskuratze-teknikaren arabera

- **w2c eta c2c:** Emaitzei erreparatuz, c2c-ek HM zuzen/onargarri gehiago eskuratu ditu SemCorren (objektuen *Batura z/o* % 33,3a da, eta subjektuena % 45,1a); BNCn, aldiz, w2c-ek gehiago lortu ditu (objektuen *Batura z/o* % 22,6a da, eta subjektuena % 17,4a), c2c-ek baino (SemCorren objektuen *Batura z/o* % 33,3a da eta subjektuena % 45,1; BNCn objektuen *Batura z/o* % 1,4a eta subjektuena % 0 da). Hala ere, w2c teknikak ez du informazio gehiegirik ematen, HM hauek aditz-formarentzat baitira, eta erabilera konputazionalerako (hala nola, adieren desanbiguaziorako edota itzulpen automatikorako) aditz-adierari buruzko informazioa lagungarria baitzaigu.

  c2c-ek, ordea, w2c-ek baino emaitza hobeak eman ditu SemCorreko subjektuen eskuratzean, eskuratu gabe % 3,5a bakarrik utzi baitu. Honen arrazoia corpus etiketatua izatea da. Hau da, entitateak landuta eta etiketatuta daude, eta eskuratze-teknikak ez ditu desanbiguatu behar. w2c teknikak ez du abantaila hau guztia aprobetxatzen. Izan ere, hitzaren adiera guztiak hartzen ditu kontuan.

Ondorioz, esan daiteke, c2c dela teknikarik egokiena corpus etiketatua erabiltzen den kasuetan. Dena dela, gerta daiteke desanbiguaturiko corpusik ez izatea. Kasu horretarako, egokiago da w2c teknika.

- **w2semf/s2semf eta c2c/w2c:** s2semf eta w2semf-en HMak zailak dira beste biekin erkatzeko, batean klasean eta bestean domeinu-eremu semantikoak erabiltzen direlako. SemCorreko corpusean s2semf-ek beste bi eskuratze-tekniken emaitzak baino hobeak eskaintzen dizkigu (objektuen *Batura z/o* % 57a da, eta subjektuena % 44,6a). Baina, esan dugun bezala, eskuratu gabekoen ehunekoa oso altua da (% 64,2 eta % 60) eta beste eskuratze-teknikena baino okerragoa. Bestalde, EFEko corpusaren gainean, kontuan izanda etiketatu gabeko corpusa dela, w2semf HMak nahiko onak dira. Baliteke, corpusari esker izatea, EFE corpusa kirol-domeinuari baitagokio. Hala ere, w2c-ekin gertatzen den antzera, HM hauek ez dute informazio gehiegirik eskaintzen, aditz-formarentzat baitira.

VII.4.5.2 Corpusaren arabera

- **BNC eta SemCor corpusen erkaketa:** SemCorren gainean erabilitako w2c eta c2c eskuratze-teknikek, BNCn baino emaitza hobeak lortu dituzte. Hala ere, desberdintasun handiagoa espero genuen, SemCor semantikoki etiketatutako corpusa dela kontuan hartuz. Honen arrazoia corpusen tamaina izan daiteke; hau da, SemCor corpus txikia da BNCkin parekatuta, eta hori dela eta:

  (a) SemCorren aditz bakoitzeko agerpen gutxiago daude, eta ondorioz, eskuratze-teknikek ezin dituzte HM batzuk eskuratu; hau da, urre-patroi batzuk eskuratu gabe geratzen dira.

  (b) BNCn eskuratze-teknikak agerpen gehiagotan oinarritu daitezke. Horrela, urre-patroi gehiago eskuratzen dira. Dena den, BNC etiketatu gabeko corpusa izaki, HM hauen kalitatea ez da SemCorrekoa bezain ona.

  Ondorioz, desanbiguatutako corpus handiagoa beharko litzatekeela esan dezakegu, emaitza hobeak lortu ahal izateko.

- **EFE:** Corpus honetatik emaitza onak lortu dira. Baliteke, corpusari esker izatea, EFE corpuseko kirol-domeinuari bakarrik baitagokio. Domeinu jakin batekin lan eginda, aditzaren adiera eta bere HMena corpusaren domeinutik lortu daitekeela deritzogu. Dena den, hau gehiago aztertu beharrekoa litzateke, kasuistika handia baitago. Aditz batzuek domeinu batekiko harreman gehiago dute beste batzuek baino. Horren adierazgarri, saiakera honetako ingeleseko meet eta equalize aditzekin lortutako emaitzak dira[46]. Nahiz eta EFEko kirol corpusera mugatu, badirudi aditz hauen beste adierek —kirol-arlokoak ez direnak— indar edo erabilera handiagoa dutela. Beraz, ikusteko dago domeinua aditz jakin batzuekin bakarrik den baliagarria ala aditz guztietara orokortu daitekeen.

VII.4.5.3 Ingeleseko HMen emaitzen laburpen orokorra

SemCor eta BNCren gainean erabilitako teknikak (c2c eta w2c, hurrenez hurren) dira HM gutxien eskuratu gabe utzi dituztenak: objektuen HMetan BNCko w2c (% 15,9) eta SemCorreko w2c (% 29,5) teknikek lortutakoak dira emaitzarik onenak, eta subjektuen HMetan SemCorreko c2c (% 3,5) eta BNCko w2c (% 13,6) teknikenak. Datu hauek hasierako susmoekin bat egiten dute:

- SemCor corpus desanbiguatua izanda, besteak baino emaitza hobeak izan behar zituela (hala ere, espero baino emaitza kaxkarragoak lortu dira).

- BNC corpus handiena izaki, eskuratu gabe oso HM gutxi geratu behar zirela.

Corpus desberdinen erabilerari dagokionez, argi geratu da, beraz, geroz eta corpus etiketatu handiagoa izan, orduan eta emaitza hobeak lortuko direla.

Esan beharra dago, domeinu-eremu semantiko bikoteekin adierazitako HMen emaitzak oso aldakorrak direla ebaluatzeko irizpideen arabera. Hauetatik jasotako emaitzak kuantitatiboki nahiko onak izan arren, neurketa hauek modu objektibo batean egiteko erraztasun falta, eta *synset*ekin parekatzeko duten zailtasuna kontuan izanda, saiakera honetatik abiatuta au-

[46] C eranskinean aditz guztien emaitzak daude.

rrerantzean egingo diren beste lanetan, domeinu-eremu semantiko bikoteekin adierazitako HMak alde batera utziko direla erabaki dugu.

VII.5 Euskarako HMak

Ingelesekoez gain, euskarako HMak eskuratzeko saiakera bat ere egin dugu. Bi bide erabili ditugu honetarako:

Batetik, ingeleseko zortzi *synset* horientzat eskuratutako HMak *synset* horietako euskarako ordainentzat berrerabiliko ditugu, euskararentzat erabilgarriak diren ala ez ikusteko. Berrerabilpenerako ez dira eskuratze-teknika guztietako HMak hartu. Azterketa hau hastapenekoa izaki, honen emaitzak ikusteko lagin bat erabiltzearekin nahikoa dela iruditu zaigu. Ingelesetik euskarara zuzenean itzuli behar genituen HMak aukeratzerakoan bi irizpide hauetan oinarritu gara:

- **SemCorretik eskuratutako HMak izatea, eta, gainera, aditz-adiera bakarrari egokitzea.** Horrela, MCR baliatuta, zuzenean itzul ditzakegu euskarara bai ingeleseko corpuseko hitzak (*synset*ekin etiketatutakoak), eta bai HMak (*synset*ekin adieraziak). Izan ere, MCRko *synset*a abiapuntu izanda, zuzenean beraien euskarako ordainera pasa gaitezke eta horrek itzulpen lana errazten. SemCor da erabili dugun corpus etiketatu bakarra, eta honen gainean aditza-adiera hautapenak eskuratzeko, c2c eta s2semf eskuratze-teknikak aplikatu dira.

- **Domeinu konkretu bateko corpus bateko HMak erabiltzea (gure kasuan, EFE)**. Honetatik lortutako HMak beste corpus orekatuetakoekin parekatzea interesgarria iruditzen zaigulako. EFE gainean w2semf eskuratze-teknika erabili dugu.

Hala, guztira, ingeleseko c2c, s2semf eta w2semf HMak berrerabili ditugu euskararako.

Bestetik, w2semf eskuratze-teknika euskarako corpus batean erabili dugu. Eskuratze-teknika hau aukeratu dugu, inplementatzeko sinpleena zelako. Horrela, teknika honen ingeleseko eta euskarako emaitzak baliatuz, euskarari zein bide (ingelesetik itzultzea ala euskarako corpusetan oinarritzea) egokitzen zaion hobeto ondoriozta dezakegu.

Erabili dugun corpusa *Euskaldunon Egunkaria* da. Domeinuka antolatutako corpusa denez (kirolak, ekonomia, kultura, eta abar), kirol-domeinutik

eskuratzeko aukera ematen digu. Hortaz, euskarako HMak kirol-domeinuan oinarritutako corpusetik lortu ditugu. Hala ere, kirol domeinuarekin erabilitako eskuratze-teknika bera erabili dugu corpus osoaren gainean, hau da, domeinurik zehaztu gabe. Emaitzek domeinuaren eragina zenbaterainokoa izan daitekeen aztertzea ahalbidetuko digute.

Euskarako HM hauen guztien azalpenerako, ingelesekoekin bezala, 00605818 play1/jokatu2; "play games, play sports" *synset*eko euskarako ordainean (jokatu 00605818n) oinarrituko gara.

VII.5.1 Euskarako HMetarako irizpideak

Ingeleseko urre-patroiak (VII.4.3 atala) sortzeko metodologia bera jarraitu dugu:

- Kirol-aditz bakoitzeko urre-patroi batzuk zehaztu dira, kasu honetan jokatu 00605818rentzat. Bestalde, urre-patroiak eskuratze-teknika bakoitzaren ereduan sortuko dira. Hala, euskarako azterketan, alde batetik, HMak adierazteko *synset*ean oinarritzen den teknika dugu (c2c), eta bestetik, domeinu-eremu semantikoetan oinarritzen direnak (w2semf eta s2semf).

- Urre-patroiak proposatu ahal izateko corpusetan oinarritu gara, aditz-adiera bakoitzaren jokaera linguistikoa orokortzeko. Corpuseko izen bat HM batean orokortzeko, gehienetan izen horrek MCRn duen hiperonimoetara jo dugu, eta, hala, HMak MCRko *synset* eta domeinu-eremu semantiko batzuen bidez adierazi ditugu.

Corpusean ikusitakoaren arabera, jokatu 00605818 aditzak lehiaketa, txapelketa eta abar bezalako objektuak hartzen ditu, orain arte HMetan {contest, competition} bezala agertutakoak[47]:

(41) Objektua:
Sidneyko **Joko Olinpikoak** jokatuko baitira irailaren.
Aste Santuan jokatuko da **Euskal Herriko txapelketa**.
Klub Arteko **Munduko Txapelketa** jokatuko da Brasilen.
Euskadiko Kopako finalerdia jokatuko du Zarautzen.

[47] 04771851 *synset*ean {contest, competition} izenak daude, eta *synset* bereko euskarako ordainak {lehiaketa, txapelketa} dira. Orain arte HMak ingelesez eman ditugu, eskuratze-tekniken emaitzak hizkuntza horretan ematen direlako. Euskaraz ere, eskuratze-tekniken emaitzak ingelesez daudenez, bere horretan mantenduko ditugu.

Joko Olinpikoak eta finalerdia izenak {contest, competition} *synset*aren hiponimoak dira. Beraz, hiperonimoaz baliatu gara jokatu 00605818ren objektuak orokortu ahal izateko.

Subjektuen kasuan, taldeak eta pertsonak izan dira nagusi:

(42) Subjektua (taldea):
Realak datorren asteazkenean jokatu behar duten partidua...
textbfKataluniako Eskubaloi Selekzioa jokatu gabe geratu zen...
Adiskidantzazko partidu gehiago jokatuko ditu **Bidasoak**.
Bestalde, hilak 14ean, hiruko torneoa jokatuko du **Bidasoak** Bermeon.

(43) Subjektua (pertsona):
Gutxienez bi partidu egongo da **Rider** jokatu gabe.
Agirresarobe - Iriatek jokatuko dute.
Iruk jokatuko du hasieratik.
Dmitri Khokhlov errusiarrak hasieratik jokatutako partidu nagusia.

Ingeleseko play 00605818k ez bezala, euskarako jokatu 00605818 aditzak ez ditu futbol, golf eta abar bezalako objektuak hartzen, ez behintzat absolutibo kasuan. Berez, jokatu 00605818k argumentu bezala onartzen ditu, baina beste kasu batekin: inesiboarekin.

(44) Objektua (inesiboa):
FutboleaN jokatzen badakitela erakutsi zuten Lotinaren jokalariek.
Banekien han dena ezberdina zela, **futboleaN** ere han jokatuta bainengoen.
Rafa Alkortak [...] **golfeaN** jokatuko duela dio irribartsu.

Euskarako subjektuen eta objektuen argumentuak, ergatiboarekin eta absolutiboarekin agertzeaz gain, beste kasu-marka batzuekin ere ager daitezkeela ikusita (jokaturen kasuan objektua inesiboa izan daiteke), euskarako HMen eskuratzea funtzio gramatikaletan oinarritu ordez —ingeleserako egin dugun bezala—, **kasu-marketan oinarrituta** egitea erabaki dugu. Hala, ergatiboen, absoltuiboen, inesiboen eta bestelako kasu-marken HMei buruz jardungo gara.

(45)en ditugu jokatu 00605818 aditzaren c2c-rako urre-patroiak eta (46)n w2semf eta s2semf teknikentzako lortutakoak:

(45) **jokatu 00605818 Absolutiboa**
c2c:
04771851 contest, competition "an occasion on which a winner is selected..."
00254052 game "a contest with rules to determine a winner"
09065837 amount of time, period, period of time "time period a length of time"

jokatu 00605818 Ergatiboa
c2c:
00004865 individual, someone, somebody, mortal, human soul "a human being"
00017008 group, grouping "any number of entities (members) considered as a unit"

jokatu 00605818 Inesiboa
c2c:
00240760 sport, athletics "an active diversion requiring physical exertion and ..."

(46) **jokatu 00605818 Absolutiboa**
s2semf, w2semf:
sport-event
time period-time

jokatu 00605818 Ergatiboa
s2semf, w2semf:
factotum-group
person-person

jokatu 00605818 Inesiboa
s2semf, w2semf:
sport-act
play-act

Beraz, ikus dezakegun bezala, ez ditugu objektu/subjektuen HMak eskuratu, deklinabide kasuan oinarritzen diren HMak baizik. Ingeleseko eta euskarako "funtzio-kasu" desoreka hau aditz bat baino gehiagorekin gertatu zaigu. Esate baterako, play 00610422rekin (ikus VII.1 irudia): ingeleseko Princeton plays Yale, euskaraz, Princetonek Yaleren aurka jokatzen du itzuliko genuke. Ingeleseko objektua (Yale), euskaraz -en kontra postposizioarekin adierazten dugu. Horregatik, play 00610422ren HMak eskuratzerakoan, ingelesarekin egin bezala objektuen eta subjektuen HMak lortu ordez, euskararako ergatiboaren eta -en kontra postposizioaren HMetan oinarritu gara.

Desoreka honek ingeleseko HMak euskarara itzultzeko zailtasunak sortu ditu. Hau da, ingeleseko argumentuak ezin dira zuzenean euskara itzuli, ingelesez funtzio gramatikalei dagozkien HMak baitauzkagu eta euskaraz deklinabide kasu-markei dagozkienak. Hortaz, ingeleseko argumentuak ezin

dira zuzenean euskarara itzuli. Horregatik, HMen erkaketa egin ahal izateko, bi hizkuntzetako aditzen argumentuak parekatu behar izan ditugu lehendabizi, aditz bakoitzaren izaera sintaktiko-semantikoa definituz. Oro har, esan dezakegu ingeleseko subjektuen HMak euskarako ergatibo eta absolutibo HMak izango direla — aditz iragankor eta iragangaitzei dagozkienak, hurrenez hurren—, eta ingeleseko objektu HMak euskarako absolutiboak emango digula. Hala ere, aditz bakoitzaren izaera sintaktiko-semantikoa kontuan izanda objektuen artean bestelako kasu-markak ere egon daitezke: esate baterako, inesiboa.

Bestalde, ingeleseko HMekin bezala, urre-patroi hauen arabera HMak ebaluatzeko maila desberdinak definitu ditugu:

- **Zuzena:** Urre-patroiarekin bat datorrenean.
- **Onargarria:** Urre-patroiaren hiperonimoa edo hiponimoa denean. Domeinu-eremu semantiko bikoteen bidez adierazitako HM kasuan, onargarri bezala kontsideratu ditugu urre-patroia baino orokorrago edota zehatzago direnak.
- **Okerra:** Urre-patroiarekin bat ez datorrenean eta MCRko hierarkian ere loturarik ez dutenean.

Euskararako eskuratutako HMak domeinu-eremu semantiko bikoteetan oinarrituak dira, eta hauen ebaluazioa irizpide batzuen arabera egin dugu; ingelesekoekin erabilitako berdinak direnez ez ditugu errepikatuko (ikus VII.4.4 atala).

VII.5.2 *Euskaldunon Egunkari*tik eskuratutako HMen azterketa eta ebaluazioa

Atal honetan eskuratze-teknika batek (w2semf) euskarako corpus batetik (*Euskaldunon Egunkaria*) eskuratutako HMak aztertu eta ebaluatuko ditugu.

VII.5.2.1 w2semf *Euskaldunon Egunkari*tik

Eskuratze-teknika hau VII.3.2.2 atalean azaldu dugu. Aditz-forma osoaren HMak erauzten dituen eskuratze-teknika da eta HMak domeinu-eremu semantiko bikote batez adieraziak datoz, bikote hau klase bezala kontsideratzen delarik. Bi proba desberdin egin ditugu. Batetik, teknika hau corpus osoan (domeinuak kontuan hartu gabe) aplikatu dugu. Bestetik, kirol-

domeinuari bakarrik dagokion zatian erabili da. Horrela, domeinuaren eragina zenbaterainokoa den ikusteko aukera izan dugu.

Nahiz eta ingeleserako eskuratze-teknika bera erabili, euskararako aldatu egin behar izan dugu pixka bat: objektu eta subjektu funtzio sintaktikoen HMen ordez, inesibo, absolutibo eta ergatibo deklinabide kasuen HMak eskuratu ditugu.

Abiapuntuko metodologia orain arte erabilitakoaren parekoa izan arren (HMaren jatorria eta corpuseko testuinguruak bilatu, HMa bera ebaluatzen hasi baino lehen), arestian gertatu zaigun bezala (VII.3.2.2), eskuratze-teknika honekin zaila da jatorria zein den zehaztea. Batetik, HMak aditz-formarentzat direlako eta hauen jatorria aztertzeko agerpenak banan-banan berrikusi beharko genituzkeelako. Bestetik, HMak adierazteko domeinu-eremu semantiko bikoteak erabiltzen dituen eskuratze-teknika izaki, eredu honen informaziotik jatorrira iristeko, nahitaez MCRra jo behar dugu domeinu eta eremu semantiko bakoitzaren azpian zein *synset* dagoen jakiteko.

Hala ere, w2w moduko zerrendak ditugu, non jokatu aditz-formarekin agertu diren hitzen zerrenda (maiztasunaren eta kasu-marken arabera ordenaturik) eskaintzen zaigun; fitxategi batean corpus osoko agerpenak daude eta bestean kirol-domeinukoak bakarrik[48].

Oso zerrenda luzeak dira, eta lan handia litzateke bakoitzaren testuinguruak aztertu eta kirolaren domeinuari dagozkionak aukeratzea, gero horren arabera beraien MCRko *synset*, eremu semantiko eta domeinu posibleak zehazteko.

Arrazoi hauengatik guztiengatik, eta datu enpirikoetan oinarritu gabe, zuzenean *Euskaldunon Egunkaria*ren gainean aplikatutako eskuratze-teknika hauen HMak gure urre-patroiekin (ikus (46)) erkatu ditugu.

(47)n jokatu aditzaren w2semf absolutibo (abs), inesibo (ine) eta ergatibo (erg) deklinabide kasuen kirol-domeinuko corpuseko HMak ditugu (letra lodiz gure ustez jokatu 00605818 aditzari dagozkienak)[49].

Bestalde, esan beharra dago eskuratze-teknika honek izen bereziak x batez adierazten ditu, anafora pronominalak pro batez eta elipsiak 0 batez.

[48] Ikus C eranskina.

[49] Ingelesekoekin gertatzen zen bezala, HMen zerrenda oso luzea izan daiteke, eta aditz baten HMak hamar baino gehiago direnean, lehenengo hamarrak (probabilitate-neurri handienekoak) bakarrik aztertu ditugu.

(47) **w2semf.jokatu.kirola**
abs x 33
abs sport-event 18.933 ZUZENA
abs anthropology-group 6.6
abs number-quantity 6.515
abs politics-group 6.504
abs sociology-group 5.671
abs history-group 5.6
abs factotum-act 2.853
abs sport-act 2.646
abs 0 2

ine x 28
ine time period-time 7.062
ine tourism-time 4
ine buliding industry-artifact 3.009
ine factotum-act 2.3 ONARGARRIA
ine number-quantity 2.272
ine factotum-location 2.138
ine 0 2
ine play-act 1.983 ZUZENA
ine sport-act 1.900 ZUZENA

erg pro 128 ONARGARRIA
erg x 25 ONARGARRIA
erg number-quantity 7
erg 0 3
erg transport-person 1.5
erg geography-person 1
erg administration-person 1
erg basketball-person 1 ONARGARRIA
erg time period-time 0.6
erg cycling-person 0.25 ONARGARRIA

(48)n corpus osoa erabilita lortutako HMak ditugu:

(48) **w2semf.jokatu.corpus_osoa**
abs x 40
abs sport-event 31.933 ZUZENA
abs sport-act 13.646
abs number-quantity 8.515
abs anthropology-group 8.131
abs politics-group 7.004
abs sociology-group 6.671
abs history-group 5.6

abs time period-time 4.632 ZUZENA
abs factotum-act 3.907

ine x 32
ine time period-time 7.437
ine factotum-act 4.020 ONARGARRIA
ine tourism-time 4
ine 0 4
ine building industry-artifact 3.609
ine factotum-location 2.361
ine number-quantity 2.272
ine factotum-state 2.081
ine factotum-group 2.068

erg pro 204 ONARGARRIA
erg x 33 ONARGARRIA
erg number-quantity 7
erg 0 3
erg linguistics-communication 2
erg politics-person 1.601
erg person-person 1.53 ZUZENA
erg transport-person 1.5
erg administration-person 1.365
erg basketball-person 1 ONARGARRIA

Ingelesekoekin bezala, urre-patroiaren berdina edo antzekoa (domeinu edo eremu semantiko orokorrago edo zehatzago bat adibidez) denean zuzen edo onargarri bezala kontsideratu dugu; baina bat ez datozenak ez ditugu okertzat hartu, hauek berez, beste aditz-adiera baten HMak izan daitezkeen heinean, zuzenak izan daitezkeelako.

Aditzaren adiera guztiak kontuan hartzen dituen eskuratze-teknika izateko, kirolari dagozkion HM ugari daude bi corpusetan. Urre-patroiko objektuen HM guztiak daude eta nahiko probabilitate-neurri altuekin, gainera.

Corpus osoko eta kirol-domeinuko HMak erkatuz gero, ez dago horrenbesteko alderik bata eta bestearen artean; desberdintasun nabarmenena inesibo deklinabide kasuko HMek erakusten dute. Kirol-domeinutik eskuratutako inesiboaren HMetan urre-patroian proposaturiko HM guztiak daude: sport-act, play-act. Corpus osotik eskuratutakoetan hauek baino orokorragoa den factotum-act bakarrik dago. Bestalde, kirol-domeinuko corpuseko inesiboen HMetan, deigarria da sport-act, play-act HMak probabilitate-neurri txikienarekin agertzea; probabilitate-neurri handienarekin izen bereziak edo x (Anoetan jokatu dute adibidez) eta time period-time (Bigarren zatian jokatu

du; Igandean jokatuko dute eta abar) daude, jokatu 00605818ren adjuntuak direnak. Kirol-domeinuko albisteak izanda (ez ahaztu *Euskaldunon Egunkaria* egunkari bat dela), berez, baliteke informazio asko inplizitu egotea, irakurleak testua ulertzeko ez dituelako behar. Hau da, nahiz eta albistean bertan ez zehaztu, irakurleak badaki "zertan" jokatzen duten albisteko protagonistek, egunkariko atal berezi batean, izenburu eta guzti, zehaztuta datorrelako (futbola, adibidez), edota pertsonak ezagutzen dituelako (Errealak Madrilen jokatuko du eta ez Errealak Madrilen futbolean jokatuko du).

Ergatibo HMetako (corpus osoko eta kirol domeinukoak) probabilitate-neurri handienak izen bereziek (x) eta anafora pronominalek (pro) dute. Esan beharra dago, transport/administration/geography-person HMekin zalantzak izan ditugula. Nahiz eta lehenengo begiratuan okerrak iruditu, w2w zerrendak eta corpusak aztertuz, konturatu ginen hauek ondorengo agerpenetatik zetozela:

(49) **Italiarrek** bi jokalari gutxiagorekin jokatu dute.
5 kilometroko erlojupekoa jokatu dute **txirrindulariek**.

Italiar izenaren domeinuak MCRn administration eta geography dira; eta txirrindulari izenarena, transport. Horregatik ditugu geography-person, administration-person eta transport-person bezalako HMak. Hala ere, arestian aipatutako irizpideari jarraituz, *transport*, *geography* eta *administration* domeinuetako izen gehienak jokatu aditzaren argumentu ezin dutenez izan, okertzat jo ditugu. Horrela, domeinu hauetako hitzak (salbuespenak salbuespen) ez direla jokatu aditzarekin agertzen adierazten dugu.

Haatik, politics-person okertzat jo dugu ergatiboko w2w zerrenda aztertuta errore bat dela ikusi dugulako; w2w zerrendako ergatiboen artean, *politics* domeinua har dezakeen bakarra defentsa baita:

(50) **Defentsak** ondo jokatu zuen.

Testuingurua zuzena da eta esaldiko defentsa izenaren domeinua *sport* da. Hortaz, honen HMa sport-person izan beharko litzateke. Nondik lortu da politics-person HMa? Izen horrek MCRn hamar *synset* inguru ditu, eta horietako bat *politics* domeinuari dagokio. Beraz, anbiguotasun errore bat egon da.

Hala, badirudi ingeleseko eskuratze-teknikekin aipatu ditugun erroreak euskarako w2semf teknikarekin ere gertatzen direla (ikus VII.4.3 atala).

VII.5.3 Ingelesetik itzulitako HMen azterketa eta ebaluazioa

Ingeleserako erabilitako eskuratze-teknika batzuekin eskuratutako HMak euskarara itzuli ditugu, HMak eleanitzak izan daitezkeen frogatzeko asmoz. Horretarako, eta VII.5 atalean azaldu ditugun irizpideak jarraituta, SemCorreko c2c eta s2semf eskuratze-tekniken emaitzak euskaratu ditugu, EFEko s2semf-ekoekin batera.

VII.5.3.1 SemCorreko c2c euskarara itzulita

VII.4.2.1 atalean azaldutako c2c objektu/subjektuen HMak (51) adibidean ipini ditugu (zuzentzat eta onargarritzat jo ditugunak bakarrik, beraien ebaluazio eta guzti), euskarako jokatu 00605818 aditzarentzat ere baliagarriak diren egiaztatzeko. Buruan izan, c2c eskuratze-teknikak lortzen dituen objektuen edo subjektuen HMak aditzaren adiera jakin baterako direla. Beraz, gure kasuan, HM hauekin play 00605818 aditza bakarrik izan beharko dugu kontuan. HM hauek euskaratzerakoan, beraz, jokatu 00605818 aditz-adierarentzat bakarrik izango dira.

(51) **c2c.obj**
play 00605818
00228990 0.215 {**activity**} **"any specific activity or pursuit" ONARGARRIA**
04771851 0.035 {**contest, competition**} **"an occasion on which..." ZUZENA**

c2c.subj
play 00605818
00017008 0.517 {**group, grouping**} **"any number of entities..." ZUZENA**
00004865 0.507 {**person, individual, human**} **"a human being" ZUZENA**

Atal honen sarreran esan dugun bezala, ingeleseko argumentuak ezin dira zuzenean euskarara itzuli. Horregatik, HMen erkaketa egin ahal izateko, bi hizkuntzetako argumentuak parekatu behar izan ditugu: ingeleseko subjektu HMak euskarako ergatibo HMak izango dira, eta ingeleseko objektu HMak euskarako absolutibo eta inesibo HMak izango dira[50]. (52)n, deklinabide kasuak kontuan hartuta egindako urre-patroiak dakartzagu:

[50] Jakina, parekatze hau aditzaren izaera sintaktiko-semantikoaren araberakoa da.

(52) Objektua:
jokatu 00605818 Absolutiboa
c2c:
04771851 {contest, competition} "an occasion on which a winner is selected from..."
00254052 {game} "a contest with rules to determine a winner"
09065837 {amount of time, period, period of time} "time period a length of time"

jokatu 00605818 Inesiboa
c2c:
00240760 {sport, athletics} "an active diversion requiring physical exertion and..."

Subjektua:
jokatu 00605818 Ergatiboa
c2c:
00004865 {person, individual, someone, somebody, human soul} "a human being"
00017008 {group, grouping} "any number of entities (members) considered as a unit"

Euskarako jokatu 00605818rentzat proposaturiko urre-patroiak (ikus (52)), ingeleseko HMekin guztiz bateragarriak dira (ikus (53)):

(53) **c2c.obj**
jokatu 00605818
00228990 0.215 {**activity**} **"any specific activity or pursuit" ONARGARRIA**
04771851 0.035 {**contest, competition**} **"an occasion on which..." ZUZENA**

c2c.subj
jokatu 00605818
00017008 0.517 {**group, grouping**} **"any number of entities..." ZUZENA**
00004865 0.507 {**person, individual, human**} **"a human being" ZUZENA**

VII.5.3.2 SemCorreko s2semf euskarara itzulita

VII.4.2.1 atalean azaldutako s2semf objektu/subjektu HMak (54)n ipini ditugu (bakarrik zuzentzat eta onargarritzat jo ditugunak, beraien ebaluazio eta guzti), euskarako jokatu 00605818 aditzarentzat ere baliagarriak diren egiaztatzeko.

Eskuratze-teknika honek aditzaren adiera bakoitzarentzat HMak domeinu-eremu semantiko bikoteekin adierazten ditu.

(54) **s2semf.obj**
play 00605818
obj play-act 3.5 ZUZENA
obj sport-act 1.5 ZUZENA
obj golf-act 0.5 ONARGARRIA
obj basketball-act 0.5 ONARGARRIA

s2semf.subj
play 00605818
subj sport-person 1 ONARGARRIA
subj factotum-group 1 ZUZENA
subj factotum-Tops 1 ONARGARRIA
subj person-person 1 ZUZENA

Euskarako jokatu 00605818rentzat proposaturiko urre-patroiak (ikus (55)), ingeleseko HMekin guztiz bateragarriak dira (ikus (56)):

(55) Objektua:
jokatu 00605818 Absolutiboa
sport-event
time period-time

jokatu 00605818 Inesiboa
sport-act
play-act

Subjektua:
jokatu 00605818 Ergatiboa
factotum-group
person-person

(56) **s2semf.obj**
jokatu 00605818
obj play-act 3.5 ZUZENA
obj sport-act 1.5 ZUZENA
obj golf-act 0.5 ONARGARRIA
obj basketball-act 0.5 ONARGARRIA

s2semf.subj
jokatu 00605818
subj sport-person 1 ONARGARRIA
subj factotum-group 1 ZUZENA
subj factotum-Tops 1 ONARGARRIA
subj person-person 1 ZUZENA

VII.5.3.3 EFEko w2semf euskarara itzulita

VII.4.2.1 atalean azaldutako w2semf objektu/subjektu HMak (ebaluazio eta guzti) (57)n ipini ditugu (bakarrik zuzentzat eta onargarritzat jo ditugunak), euskarako jokatu 00605818 aditzarentzat ere baliagarriak diren egiaztatzeko.

EFE domeinuka antolatutako corpusa da, eta guk kirol-domeinuari dagokiona erabili dugu saiakera honetarako. Corpus honetan w2semf eskuratze-teknika aplikatu dugu, euskarako HMak eskuratzeko erabili duguna. Teknika honek eskuratzen dituen HMak aditz-formarentzat dira, aditzaren adiera guztientzat, alegia. Gainera, probabilitate kopuru altuenetik baxuenera ordenaturiko domeinu-eremu semantiko bikoteak dira.

(57) **w2semf.play.kirola.obj**
obj play-act 50.013 ZUZENA
obj factotum-act 30.390 ONARGARRIA
obj time period-time 29.009 ZUZENA
obj sport-event 23.514 ZUZENA
obj sport-act 23.038 ZUZENA

w2semf.play.kirola.subj
subj x 372 ONARGARRIA
subj sport-group 44.01 ONARGARRIA

Euskarako jokatu 00605818rentzat proposaturiko urre-patroiak (ikus (58)), ingeleseko HMekin guztiz bateragarriak (ikus (59)) dira:

(58) Objektua:
jokatu 00605818 Absolutiboa
sport-event
time period-time

jokatu 00605818 Inesiboa
sport-act
play-act

Subjektua:
jokatu 00605818 Ergatiboa
factotum-group
person-person

(59) **w2semf.jokatu.kirola.obj**
obj play-act 50.013 ZUZENA
obj factotum-act 30.390 ONARGARRIA
obj time period-time 29.009 ZUZENA
obj sport-event 23.514 ZUZENA
obj sport-act 23.038 ZUZENA

w2semf.jokatu.kirola.subj
subj x 372 ZUZENA
subj sport-group 44.01 ONARGARRIA

VII.5.4 Ebaluazioaren azterketa

VII.7 taulak laburbiltzen du euskararako jokatu 00605818rentzat eskuratutako edo itzulitako HMen emaitzen kalitatea. Corpus bakoitzean erabili den eskuratze-teknika bakoitzetik, zenbat objektu/subjektuen edo absolutibo/ergatibo/inesiboen HM diren zuzenak (urre-patroiarekin bat datozenak), zenbat diren onargarriak (urre-patroiaren hiperonimo edo hiponimo bat direnak) eta urre-patroietatik zenbat ez diren eskuratu (eskuratu gabeak deitu duguna) erakusten du taulak. Datu hauek kopuru zehatzak erabiliz adierazi ditugu; esaterako, eskuratze-teknika bakoitzaren objektu/subjektuen HMetatik (gehienez hamar) zenbat diren zuzenak eta onargarriak zenbakitu ditugu; eta baita eskuratze-teknika bakoitzarentzat proposatutako urre-patroietatik zenbat geratu diren eskuratu gabe ere. Horrelako taula bana egin dugu saiakera honetan erabilitako kirol-aditz bakoitzarentzat, hots, MCRtik aukeratutako zortzi *synset*entzat[51].

VII.8 taulan euskararako zortzi aditzentzat eskuratutako edo itzulitako HMen emaitzak laburbildu ditugu, oraingoan ehunekotan adierazi ditugularik[52]. Taula honetan zuzenen eta onargarrien kopuruak batu ditugu (*Batura z/o* zutabean).

Eskuratu gabeen zerrendan datu azpimarragarriena % 0 zenbakira hurbiltzen dena da, honek eskuratze-teknikak urre-patroiko HM guztiak lortu dituela esan nahi duelako. Zuzen eta onargarrien zerrendan, aldiz, datu nabarmenenak % 100era gerturatzen direnak dira, noski. % 100 lortzeak eskuratze-teknikak eskuratutako HM guztiak zuzenak/onargarriak direla adieraziko

[51] Taula hauek guztiak C eranskinetan daude ikusgai.

[52] Taula honetan absolutiboaren eta ergatiboaren datuak bakarrik adierazi ditugu, aditz guztiekin agertu zaizkigunak, hain zuzen ere.

Corpusa	*HMak*	*Kasua*	*Zuzena*	*Onargarria*	*Eskuratu gabea*
Egunkaria osoa	w2semf	abs	10etik 2	0	0
		ine	0	10etik 1	0
		erg	10etik 1	10etik 3	2tik 1
Egunkaria kirola	w2semf	abs	10etik 1	0	2tik 1
		ine	10etik 2	10etik 1	0
		erg	0	10etik 4	2tik 1
SemCor	c2c	obj	8tik 1	8tik 1	4tik 1
		subj	5etik 2	0	0
SemCor	s2semf	obj	10etik 2	10etik 3	4tik 2
		subj	7tik 2	7tik 2	0
EFE kirola	w2semf	obj	10etik 4	10etik 1	0
		subj	0	10etik 4	2tik 1

VII.7 Taula: Euskararako eskuratutako eta ingelesetik itzulitako jokatu 00605818ren HMen emaitzak.

luke.

Taula hauek aurrean izanda, hurrengo atalean, hauetatik ondoriozta ditzakegun emaitzak komentatuko ditugu.

VII.5.4.1 *Euskaldunon Egunkari*tik eskuratutako HMak

*Euskaldunon Egunkari*tik, w2semf teknikarekin, eskuratutako objektuen (euskarako kasuan, absolutiboen) HMak ingelesekoenak baino hobexeak dira, urre-patroi gehienak eskuratu direlako (% 3,5 dira eskuratu gabeak). Dena den, datu hau aztertu beharrekoa da, susmoa baitugu euskarako objektua beste kasu-markekin adierazita datorrenean, emaitzak ez direla horren onak (adibidez, entrenatu aditzaren kasuan inesibo HMen emaitzak oso txarrak dira[53]). Baliteke honen arrazoia hauek inplizituki adieraziak datozela izatea. Hau da, irakurleak testua ulertzeko beraien beharrik ez duenez, baliteke testuan argumentu hauek ez azaltzea. Hala balitz, eskuratu gabeko urre-patroien kopurua handiagoa litzateke[54].

Hala ere, *Euskaldunon Egunkari*tik eskuratutako HM asko onargarriak diren arren, subjektuen kasuan, gehienak (% 75) eskuratu gabe geratu di-

[53]Ikus C eranskina.

[54]Honi buruz VII.5.2.1 atalean mintzatu gara.

Corpusa	*HMak*	*Kasua*	*Zuzena*	*Onargar.*	*Batura z/o*	*Eskuratu gabea*
Egunkaria osoa	w2semf	abs	% 25,7	% 25,7	% 51,4	% 3,5
		erg	% 3,7	% 62,5	% 66,2	% 81,2
Egunkaria kirola	w2semf	abs	% 25,7	% 31,4	% 57,1	% 3,5
		erg	% 2,8	% 62,5	% 65,3	% 75
SemCor	c2c	obj	% 6,9	% 26,4	% 33,3	% 44
		subj	% 38	% 7,1	% 45,1	% 3,5
SemCor	s2semf	obj	% 14,2	% 42,8	% 57	% 64,2
		subj	% 7	% 37,6	% 44,6	% 60
EFE kirola	w2semf	obj	% 14,1	% 10	% 24,1	% 45,4
		subj	% 2,7	% 21,8	% 24,5	% 41

VII.8 Taula: Euskararako eskuratutako eta ingelesetik itzulitako HMen emaitzen portzentaiak, MCRtik aukeratutako zortzi *synset*entzat.

ra. Zergatia ez dugu sakonki aztertu baina susmoa dugu hurrengo arrazoiek zerikusia dutela: euskarako corpusaren tamaina txikiegia dela eta euskarako analizatzaile sintaktikoa ez deka ingelesekoa bezain ona. Bestalde, aurreprozesuan entitateak ez lantzeak ere izan du eraginik. Ergatiboen HMetako gehienak izen bereziak (x) edo pronominalak (pro) dira. Hauek onargarritzat jo ditugun arren, ezin dira urre-patroiekin parekatu, eta, ondorioz, ezin ditugu eskuratu gisa kontsideratu. Arrazoi horregatik, euskarako HMetan, ergatiboaren kasuan, eskuratu gabeen kopurua asko handitu da.

Bestalde, ingeleseko HMekin gertatu ez den bezala, euskararen kasuan, corpusa domeinu zehatz batean egoteak ez du aditzaren adiera desanbiguatzen. Corpus osoko eta kirol-domeinuko euskarako HMen emaitzak oso antzekoak dira. Are gehiago, kasu askotan, kirol corpusean eta corpus osoan, HMak berdin-berdinak dira. Hots, aztergai dugun aditz horren agerpenak kirol-domeinuko corpusean bakarrik daudenez, corpus osoko datuak kirol atalaren berdinak dira. Hala ere, euskarako corpus handiago batean saiatuz gero, corpusaren domeinuaren eragina nabarituko litzatekeela pentsatzen dugu.

VII.5.4.2 SemCorretik eskuratutako HMak

Corpus honetan bi eskuratze-teknika erabili ditugu: c2c eta s2semf. Bi eskuratze-teknikek eskuratutako HMak euskararentzat baliagarriak dira (HM zuzenak eta onargarrietaz ari gara, noski).

Ikus daitekeen bezala, teknika hauen emaitzak berdin-berdinak dira ingeleserako eta euskararako. Hortaz, eleaniztasunaren hipotesia egiaztatu egiten da; hau da, saiakera honetarako aukeratutako ingeleseko aditzen HMak berberak dira euskararako aditz homologoentzat. Hala eta guztiz ere, itzulpena egiterakoan, kontuan izan beharrekoa da bi hizkuntzetan argumentuak ez direla deklinabide kasu berarekin gauzatzen. Aipagarriak dira ingeleseko c2c eskuratze-teknikak lortutako subjektuentzako emaitza onak. Honen arrazoia corpusean entitateak markatuak egotea izan daiteke. Hala, entitate horiek person, group, location eta abar bezalako *synset*ekin adierazten dira.

Ingeleseko emaitzak azaltzerakoan esan dugun bezala, kontuan izanda SemCor semantikoki etiketatutako corpusa dela, emaitza hobeak espero genituen. Corpusaren tamaina (erabilitako corpus txikiena dugu hau) eta etiketatze-erroreak izan daitezke zergatiak. kasu honetan.

VII.5.4.3 EFEtik eskuratutako HMak

Corpus honetan eskuratze-teknika bakarra erabili dugu: w2semf. Bai ingelesez eta bai euskaraz, emaitza nahiko onak lortu ditugu. SemCorreko c2c-ekin alderatuz, EFEren w2semf-en zuzen/onargarrien batura txikiagoa da. Baina kontuan izanda semantikoki etiketatu gabeko corpusa dela, azpimarratu beharreko emaitzak dira. Corpusaren domeinuak (kirola) beste adierak baztertzen lagundu duela dirudi. Aipatu bezala, euskarako kirol-aditzen agerpen gehienak kirol-domeinuari dagokion corpus-atalean bakarrik azaldu dira.

VII.5.5 Euskarako HMen emaitzen laburpena

Oro har, emaitzei erreparatuz, *Euskaldunon Egunkaria* corpusaren gainean aplikatutako w2semf teknikak eskaintzen dizkigu emaitzarik onenak, batez ere, objektuei dagozkienak. SemCorreko c2c eskuratze-teknikaren subjektuen HMak azpimarragarriak dira, % 3,5a soilik uzten baitu eskuratu gabe. Hala, badirudi teknika hauen arteko ebakidura eginez gero, lortuko genituzkeela emaitzarik onenak.

Amaitzeko, esan dezakegu ingeleserako HMak euskarara itzul daitezkeela. Izan ere, ikusi dugu kirol-domeinuko aditzekin, *synset* berean dauden aditzek argumentu mota berdinak hartzen dituztela, hots, aditzen argumentuen tasunak eleanitzak direla. Hala ere, hizkuntza bakoitzak tasun hauek era ezberdinetan azaleratzen ditu. Gogoratu, jokatu aditzak, adibidez, objektua inesiboarekin adierazten duela. Argumentuen tasunak parekatzeko

garaian, beraz, ezberdintasun hauek kontuan izan beharko dira.

VII.6 Ondorioak

Kapitulu honetan azaldu dugun azterlanak bi helburu nagusi zituen:

- Hainbat eskuratze-teknika erabiliz ingeleseko eta euskarako corpus ezberdinetatik eskuratutako HMak aztertzea eta konparatzea.
- Ingeleserako eskuratutako HMak euskararako baliagarriak diren aztertzea.

Azterketa ugari egin dira HMen eskuratze automatikoari buruz, baina ez hainbeste eskuratze automatiko horren ebaluazio linguistikoari buruz; are gutxiago euskarari dagozkionak. Lan honen ekarpen garrantzitsu bat horretan datza, hain zuzen ere. Egun erabiltzen diren hainbat eskuratze-tekniken azterketa eta ebaluazio linguistikoa egin ondoren, lan honen bidez, euskarako HMen eskuratze automatikoa garatzeko aukera eta proposamen berriak eskaintzen dira.

Azterlan honek dakarren beste ekarpen nagusia eleaniztasunaren hipotesiaren bideragarritasunari buruzkoa da; hots, ingeleserako eskuratutako HMak euskararako erabilgarriak izan daitezkeela frogatu dugu. Honenbestez, hizkuntza batentzat eskuratutako HMak beste edozein hizkuntzatarako baliagarriak direla esatera ausartzen gara, nahiz eta baieztapen hau guztiz frogatzeko azterketa osoago bat egitea komeni den. Izan ere, aztertu ditugun aditzak kirol-domeinukoak dira eta beste domeinuetan begiratu beharko litzateke hipotesi hau baieztatzeko. Gainera, hizkuntza desberdinekin portaera hori errepikatzen den egiaztatu beharko litzateke. Hala ere, badirudi ingelesak eta euskarak konpartitzen duten portaera hau, errazago beteko dela elkarren antza handiagoa (edo gutxienez jatorri bera) duten bi hizkuntzen artean; adibidez, frantsesa eta ingelesa edota gaztelania eta frantsesa.

Euskararen LNPrako ekarpen garrantzitsua dugu hau, euskarak corpus eta baliabide kopuru txikiagoak dituelako, eta hipotesi honetaz baliatuz gero, baliabide gehiago dituzten hizkuntzenak erabiltzeko aukera eskaintzen zaigulako.

Saiakera honen emaitzak behin-behinekoak dira, aditz-adiera batzuk bakarrik aztertu baititugu, eta eskuratze-teknika guztiak ezin izan direlako corpus guztien gainean erabili. Hortaz, honako hau hastapeneko lana dugu,

eta hemen aurkeztutako emaitzetatik eta ondorioetatik abiatuta, azterketa honen esparrua zabaltzeko asmoa dugu.

Ingeleseko HMetatik, bestalde, honako hauek ondorioztatu ditugu:

- **Corpus bakoitzak bere idiosinkrasia du eta hori emaitzetan islatzen da.** SemCor eta BNCn eskuratze-teknika berak erabili dira, eta SemCorretik eskuratutakoak BNCkoak baino hobeak dira, SemCor semantikoki etiketatutako corpusa delako. Hala ere, emaitza hobeak espero ziren. Corpus txikiagoa izatea, etiketatze-erroreak izatea eta corpuseko adiera batzuk MCRn ez egotea izan daitezke arrazoiak. Azkenik, EFE corpusa domeinu zehatz batekin erabiltzeak emaitza nahiko onak eman ditu.

- **c2c eskuratze-teknikak ez dira w2c-renak baino askoz hobeak**. Lehenengoaren kasuan, c2c, aditza klase bezala kontsideratzeak (troponimoaz baliatuz) ez dirudi emaitza hobeak lortzen laguntzen duenik. Eskuratze-teknika hau oinarri egokia iruditzen zitzaigun HMen eskuratze eleanitza egiteko, hau da, hizkuntza bateko HMak zuzenean beste batera itzultzeko. Emaitza ikusita, bide honetatik jarraitu aurretik, honek ikerkuntza gehiago behar duela argi dago. Bigarrenaren kasuan, aldiz, w2c, HMen kalitatea nahiko ona izan arren, hauek aditzaren adiera guztientzat dira, eta erabilera konputazional mugatua dute. Eskuratze-teknika hau domeinu konkretu bateko corpusean erabilita emango lituzkeen emaitzak ikustea interesgarria izan daiteke.

- **Domeinu-eremu semantiko bikoteekin adierazitako HMak interpretatzeko zailagoak dira, *synset*ekin adierazitakoak baino**. Hala ere, baliabide gutxien eskatzen duten eskuratze-teknikak dira, eta hauek EFE corpusaren gainean (kirol-domeinuaren gainean), emaitza nahiko onak lortu dituzte.

- **Domeinu batean oinarritutako eskuratze-teknikek HM hobeak eskuratu dituzte, eta domeinuaren arabera aditz horren adiera mugatu daiteke**. Hala ere, beste aditzekin frogatu beharko litzateke; dirudienez, aditz batzuk domeinu batekin beste batzuek baino lotura gehiago izan baitezakete.

- **Izenen anbiguotasuna arazo bat da**. Ikusi ditugu game eta defentsa bezalako izenekin gertatu diren nahasketak. Beraien MCRko *synset* edo domeinu-eremu semantiko egokia hartu ordez, makinak beste *synset* edo domeinu-eremu semantiko bat aukeratu du, eta ondorioz, HM okerra lortu du.

- **Erroreen azterketatik ondoriozta dezakegu, prozesaketa linguistiko hobe batekin, HM hobeak lortuko genituzkeela**. Hau da, analizatzaile sintaktikoan aurkitutako erroreak konponduz gero, eta anafora eta izen berezien tratamendua landuz gero, okerrak ziren HM asko eragotziko genituzkeela uste dugu.

Ingeleseko eta euskarako HMen konparaketari dagokionez:

- **Euskarako HMen kalitatea ingelesekoena baino zertxobait handiagoa da.** Baliteke argumentuak kasu-marketan banatu izanak eraginik izatea. Susmoa dugu euskarako objektua beste kasu-markekin adierazita datorrenean, emaitzak ez direla horren onak.

- **Ingeleseko aditzen HMak euskarara zuzenean itzul daitezke**. Hala ere, gerta daiteke ingeleseko objektua euskarako kasu ezberdinekin agertzea (inesiboan adibidez). Beraz, moldaketaren bat beharrezkoa litzateke.

Oro har, domeinuetaz baliatuz gero, aditz-adieraren HM hobeak lortuko ditugu. Bestalde, emaitzek erakusten dute HMak hizkuntza batetik bestera itzul daitezkeela. Horrela, baliabide gehiago dituen hizkuntzaz baliatu gaitezke euskararen eskuratze automatikorako. Dena den, hizkuntzen argumentuen ezaugarri linguistikoak batzuetan ez datoz bat eta moldatu egin behar dira.

Etorkizuneko lanari begira, eta honako hau hastapeneko lan bat izaki, badaude sakonago lantzeko hainbat puntu. Hasteko, kirolaren domeinuaz gain beste domeinu batzuetako aditzak ere aztertu nahiko genituzke (finantzaren domeinukoak, adibidez). Bestalde, domeinu bakarreko corpusean erabili ez diren eskuratze-teknikak (w2c eta c2c) mota horretako corpusekin probatu nahiko genituzke. Hori egin baino lehen, ordea, eskuratze-teknika hauen algoritmoak hobetzen saiatuko gara. Izan ere, SemCorren oinarrituta izandako emaitzak ikusita, eskuratze-teknika hauek berriro erabili baino lehen, antzemandako erroreak gainditzea komeni da (analizatzaile sintaktikoaren akatsak

konpondu, anafora eta izen berezien tratamendua egin, aditz klaseetan troponimia kontuan ez hartu, eta abar).

Hurrengo saiakeretan, domeinu-eremu semantiko bikoteekin adierazitako HMak alde batera utziko dira. Hauek lortutako emaitzak oso aldakorrak dira ebaluatzeko irizpideen arabera. Gainera, ebaluatzean izandako arazoetaz jabetu gara, baita *synset*ekin parekatzeko duten zailtasunez ere. Horiengatik guztiengatik, beste eskuratze-tekniketan oinarritzea erabaki dugu.

Bestalde, ingeleserako eta euskararako eskuratutako HMen ebakidura eginez gero, errore ugari desagertuko liratekeela uste dugu, eta hipotesi hau egiaztatu nahiko genuke.

Euskararako HMei dagokienez, w2semf eskuratze-teknikatik lortutakoetaz gain, mota gehiago probatu nahi ditugu. Hasiera batean, w2c eta c2c teknikekin hastea pentsatu dugu. Horrela, euskarako datu gehiago izango dugu ingelesekoekin erkatzeko. Honekin batera, euskarako eskuratze-teknikak hobetzeko, semantikoki etiketatzen ari garen corpusa (EuSemcor) erabiltzea pentsatu dugu. Azkeneko helburua eskuratze-teknika egokiarekin jo ondoren, eskuratzen diren HMak Euskal WordNeten txertatzea da.

VIII. KAPITULUA

Ondorioak eta etorkizuneko lanak

Ikerlan honen emaitza gisa euskararen semantikaren azterketa aplikaturako oinarrizkoa den EBL eleanitza diseinatu eta garatu dugu: ***Euskal WordNet***.

EBL hau, IXA taldeak garatutako gainerako tresnak bezalaxe, euskararen azterketa aplikaturako egitasmo orokor baten barruan kokatzen da, eta bide horretan aurrera egiteko oinarrizko baliabidetzat jo daiteke, batez ere, hizkuntzaren ulermena beharrezkoa duten atazetan; hala nola, hitzen adieren desanbiguazioan, itzulpen automatikoan, egitura sintaktikoen desanbiguazioan, informazioaren erauzketan eta galdera-erantzun automatikoan.

Erabilera konputazionalez gain, Euskal WordNeten kontsultarako interfazea publikoa denez[1], hiztegi eta thesaurus gisa ere erabil daiteke; batetik, hiztegi elebakar gisa, hitzen adierak kontsultatzeko, hiztegi tradizionalen antzera, Euskal WordNetek *synset* bakoitzeko definizio edo glosa bat baitu (gehienetan adibide eta guzti[2]); eta bestetik, hiztegi elebidun gisa, *synset* bakoitzak dagokion ingeleseko, gaztelaniako, katalaneko eta italierako ordainak baititu. Honetaz gain, *synset* bakoitzean hizkuntza bakoitzeko ale lexikal bat baino gehiago egon daitezkeenez, thesaurus bezala balia daiteke, adiera berdina adierazteko sinonimo desberdinak ditugulako. Hala, erabilera orokorreko baliabidea garatu dugula esan daiteke.

[1] `http://ixa2.si.ehu.es/mcr/wei.html` (2007-07-02an atzitua).

[2] Glosak EuSemcor proiektuaren barruan lantzen ari gara; *synset*a editatu, eta honen agerpenak etiketatu ondoren, *synset*aren glosa gehitzen dugu.

VIII.1 Ondorio nagusiak

Tesi-lan honetan, Euskal WordNet sortzeko eta garatzeko jarraitu dugun ibilbidearen berri eman dugu, eta bertatik zenbait ondorio atera ditugu, hurrengo ataletan laburbildu ditugunak.

VIII.1.1 EBLen azterketa kritikoa

EBLen ereduei dagokionez, ez dago eredurik, oraindik, hizkuntzaren ulermenerako beharrezkoa den informazio guztia duenik. Arrazoi horregatik, guretzat garrantzitsua izan da orotariko informazioa bil dezakeen EBL bat egitea. Horretarako, urrats hauek eman ditugu:

- Batetik, IXA taldearen beharretara egokitzen den lexikoiaren ezaugarriak zerrendatu ditugu: non eta nola erabili nahi dugun, horretarako zer informazio-mota txertatuko dugun sarrera bakoitzean, eta zein eredu edo formalismoren arabera jasoko duen informazio hori.

- Bestetik, erdal hizkuntzetako LNPren arloan oihartzuna izan duten hainbat EBLen ereduak aztertu ditugu, aipatutako ezaugarrietara gehien egokitzen den formalismoaren bila. Horretarako, eredu hauen arteko azterketa konparatiboa egin dugu.

- Azkenik, IXA talderako baliagarria izango den eredu bat aukeratu dugu —***WordNet***, eta honen ildotik sortutako ***EuroWordNet*** eta ***The Multilingual Central Repository*** **(MCR)**—, eta hartutako erabaki honen arrazoiak azaldu ditugu:

  (a) Eredu hauek ez daude teoria bakar bati lotuta, bestelako eredu eta teoria ezberdinekin erabil daitezke. Horren proba da formalismo eta lan teoriko asko, gerora, WordNeten adiera edo/eta klase semantikoekin aberastu dituztela.

  (b) Eredu hauek lexiko zabala eta garatua dute; sarrera bakoitzean ale lexikalaren adiera, klase semantikoa, kategoria eta beste sarrerekin izan ditzaken erlazio semantikoak jasotzen dituzte.

  (c) Inplementatutako EBLak dira. Honen adierazgarri dira WordNeten oinarrituta egin diren publikazioen kopurua (gaur egun, WordNeteko web orriak[3] 422 inguru jasotzen ditu).

[3] `http://www.cogsci.princeton.edu/cgi-bin/webwn` (2007-07-02an atzitua).

(d) WordNet EBL elebakarra izan arren, honen ildotik sortutako EuroWordNet eta MCR eleanitzak dira.

VIII.1.2 Euskal WordNeten eraikuntzarako diseinua eta metodologia

WordNet, eta honen ildotik sortutako EuroWordNet eta MCR ereduei lotutako euskal EBLari ***Euskal WordNet*** deitu diogu. Euskal WordNetek hauen egitura eta oinarriak izan arren, honen garapena metodologia eta ikuspegi ezberdinak baliatuta egin zitekeen. Hauek guztiak aztertu ditugu, eta hauexek dira, orain arte, Euskal WordNeten garapenean hartu ditugun erabaki metodologikoak:

- Alde batetik, Euskal WordNet sortzeko diseinua definitu dugu: euskarako adieren inbentarioa eta hierarkia guk geuk sortu ordez, WordNeteko hierarkiari jarraitu eta bertako *synset*ei zuzenean esleitu dizkiegu euskarako ordainak.

- Bestetik, *synset*ei euskarako ordainak esleitzeko garaian, estaldura —sarrera lexikalen kopurua ahalik eta handiena izatea— eta kalitatea —sarrera lexikalen informazioa zuzena izatea— uztartzeko garrantzia nabarmendu dugu. Ezaugarri hauek izan dira, hain zuzen ere, EBLaren garapen-metodologia definitu dutenak, eta Euskal WordNeten garapenaldi eta orrazketa ezberdinak eragin dituztenak.

Beste ereduetan egindakotik ondorioztatu dugu, EBLa sortzearekin batera, corpus bat etiketatzea beharrezkoa dela EBL hori aberasten joateko. Izan ere, corpusean adibide, adiera eta erabilera errealak agertzen dira. Hala, EBLaren garapenari lotuta, Euskal WordNeteko *synset*ak erabiliz eskuz etiketatzen dugun euskarako corpus semantikoa aurkeztu dugu: ***EuSemcor***. Euskarako corpus semantiko bat izate hutsak berez daukan garrantziaz gain, corpus honek Euskal WordNet etengabe orrazteko, garatzeko eta aberasteko balio digu.

VIII.1.3 Euskal WordNet eta kontzeptuen errepresentazioa

Wordnet eleanitzekin lan egiteak hizkuntzen arteko ezberdintasunak gainditu beharra dakarrela erakutsi dugu. Gure kasuan, ingeleseko wordnetaren gainean lan egiteak tratamendu berezia behar duten eta *synset*en adierazpenean eragina duten bi fenomeno linguistiko azaldu ditugu:

- **Lexikalizazioa**: Ikusi dugun legez, hizkuntzen arteko lexikalizazioa ez dator beti bat; hau da, hizkuntza bateko kontzeptuak ez dira beti era berdinean lexikalizatzen beste hizkuntzetan. Honi aurre egin ahal izateko, lexikalizazioaren eta fenomeno honen kasuistikaren adibideak aurkeztu eta aztertu ditugu. Azterketa horretan, argi geratu da lexikalizazioaren mugak lausoak direla, eta askotan lan zaila dela hitz bat edo hitz anitzeko bat lexikalizatua dagoen ala ez ebaztea. Lexikalizazioaren eztabaidak eragoztearren, eta LNPko atazen erabilgarritasunari begira, Euskal WordNeten zer adierazpen mota txertatu behar genituen zehaztu dugu. Laburbilduz, irizpide hauen bitartez, honako hau arrazoitu dugu: lexikalizaturiko ordainez gain, zalantzazko lexikalizazioa duten ordainak ere Euskal WordNeten gehitzea beharrezkoa dela.

- **Hierarkia kontzeptualen antolaketa**: Hizkuntza ezberdinetako EBLak bateratzean, bi hierarkien artean aldeak daudela ere azpimarratu dugu. Honen adierazgarri, Euskal WordNeten gertatzen den *auto-hiponimia faltsua* dugu: hizkuntza batean bi hitz desberdinekin adierazten den desberdintasun hierarkikoa, bete hizkuntzetan ez da hain argia hitz bera erabiltzen delako. Hierarkiaren eraginez ematen diren fenomenoak eta kasuistika aztertu ditugu, eta hauek guztiek Euskal WordNeten izango duten tratamendurako irizpideak ere definitu ditugu, fenomeno bera beti era berean adierazia izan dadin EBLan.

- Irizpide hauek eraginda, *synset*en errepresentaziorako Euskal WordNeten marka edo ezaugarri berriak sortu ditugu. Horrekin batera, hitz anitzeko esapideen (HAEen) barne-errepresentazio aberatsago baten proposamena ere egin dugu, non HAEaren barne-osagaiak harreman semantikoen bidez erlazionatzen diren. Honenbestez, abiapuntu gisa hartu dugun EBLaren errepresentazioa hedatu eta aberastu dugula esan dezakegu.

VIII.1.4 Euskal WordNet eta hautapen-murriztapenak

Euskal WordNet hautapen-murriztapenen informazioarekin hedatu ahal izateko egin dugun ikerketa azaldu dugu. Ingeleseko eta euskarako kirol-arloko aditz batzuen objektuen eta subjektuen hautapen-murriztapenen azterketa deskribatu dugu, eta honako emaitza hauek lortu ditugu:

- Hainbat eskuratze-teknika erabiliz, ingeleseko eta euskarako corpus ezberdinetatik eskuratutako hautapen-murriztapenen ebaluazioa eta azterketa konparatiboa.

- Hautapen-murriztapenak eleanitzak direnaren zantzuak topatu ditugu; zehazkiago esanda, ingeleserako eskuratutako hautapen-murriztapenak euskaraz ere erabilgarriak izan daitezkeela egiaztatu dugu.

VIII.2 Ekarpenak

Tesi-lan honen ekarpen nagusienak euskararen semantika lantzeko EBL eleanitza (Euskal WordNet) eta honen corpus osagarria (EuSemcor) dira. VIII.1 taulan, gaur egun, Euskal WordNeten dauden izenen eta aditzen kopuruak aurkezten ditugu; VIII.2 taulan EuSemcorrekoak.

	Guztira	***Izenak***	***Aditzak***
Adierak	50.670	41.160	9.510
Lemak	26.565	23.069	3.496
Synsetak	32.456	28.705	3.751
Hutsune lexikalak	2.499	2.198	301
Izen bereziak	722	722	0

VIII.1 Taula: Euskal WordNet: kopuruak

Bestalde, EBLen garapenaren arloan lagungarri izan daitezkeen hurrengo ekarpenak ere aurkezten ditugu:

- EBLen azterketa kritikorako bibliografia-bilketa eta azterketa konparatiboa egin dugu, non egun LNPren arloan oihartzuna duten EBL-ereduen ezaugarri nagusiak aurkeztu ditugun.

	Eginak		*Egingabeak*		*Guztira*	
	Hitz	**Agerpen**	**Hitz**	**Agerpen**	**Hitz**	**Agerpen**
Polisemikoak	442	39.208	2.888	29.663	3.330	68.871
Monosemikoak	192	7.281	1.618	9.325	1.810	16.606
EusWNen ez daude	83	487	10.987	39.449	11.070	39.936
Guztira	717	46.976	15.493	78.437	16.210	125.413

VIII.2 Taula: EuSemcor: kopuruak

- Euskal WordNeten diseinuaren deskribapenarekin batera, estaldura eta kalitatea uztatzea helburu duen eraikuntza-metodologia proposatu dugu.

- EBL eleanitz bat sortzean azaltzen diren fenomeno linguistikoen deskribapena egin dugu, eta, gainera, hauek EBLan lantzeko eta adierazteko irizpideak zehaztu ditugu.

- EBL baten eta semantikoki etiketatutako corpus baten garapenak bateratzeko bideak erakutsi ditugu.

- MCRren ereduaren aberasketa: HAEen osagaiak semantikoki erlazionatzen dituen errepresantazio-eredu bat proposatu dugu.

- Hautapen-murriztapenen eskuratze automatikoaren ebaluazio linguistikoa egin dugu. Honi esker, hautapen-murriztapenen eskuratze automatikoa garatzeko aukera eta proposamen berriak eskaini ahal izan ditugu, gerora, lortuko dugun informazio hori EBLan txertatzeko asmoarekin.

VIII.3 Etorkizuneko lanak

Euskal WordNet egunez egun handitzen eta eguneratzen ari da, eta horrekin batera, Euskal WordNeteko *synset*ekin eskuz etiketatzen ari garen euskarako corpusa (EuSemcor). Egun, maiztasun handieneko izenen lanketa amaitzen ari gara, eta, dagoeneko, aditzen aberasketari ere ekin zaio. Etorkizunean, gure asmoa aditzak, adjektiboak eta adberbioak (ordena horretan) lantzea da.

Bestalde, Euskal WordNeten aberasteko hurrengo ikerlerroak proposatzen ditugu:

- **Euskal WordNet kontzeptu berriekin aberastea**:

WordNeten ez dauden eta zerrendatuta ditugun, euskarako kontzeptuak (trikitixa, ikastola... bezalakoak) EBLan sartu nahi ditugu. Egitasmo hau betetzeko, bestelako wordnetetan ataza hau nola egiten duten ezagutu eta gure metodologia definitu beharko dugu, sortzen diren zailtasun berriei aurre eginez.

- **HAEen barne-errepresentazioa zehaztea**:

HAEen barne-errepresentazioaren proposamena EBLan gauzatu nahi dugu. Horretarako, Agirre eta Lersundiren (2001) metodo erdiautomatikoak erabiltzea pentsatzen dugu, barne-egiturako *synset*ak eta beraien arteko harreman semantikoak automatikoki desanbiguatu ahal izateko. Eratorpenaren azterketarako sortutako metodo erdiautomatiko horrek, hiztegietako definizioetan oinarrituta, eratorritako hitza eta bere erroaren arteko harreman semantikoa zehazten laguntzen du. Hala, metodo hau HAEen osagaien arteko harremanak zehazteko erabili aurretik, metodoaren berrikuspena egin beharko genuke, hau da, HAEen azterketarako egokitu beharko genuke.

- **Euskal WordNet informazio gehiagorekin aberastea:**

Aipatu izan dugun bezala, nahiz eta gure EBLaren garapena WordNeten egitura eta oinarriak izan, ikuspegi eta metodologia ezberdinak erabilita egin zitekeen:

(a) WordNeten hierarkian jarraituta eta bertako *synset*ei zuzenean esleituta euskarako ordainak.

(b) Guk geuk sortuta euskarako adieren inbentarioa eta hierarkia.

Tesi-lan honetan Euskal WordNeten garapena lehenengoan oinarritu dugu, eta ingeleseko kontzeptuak abiapuntutzat harturik, euskarako ordainak lotu ditugu. Hala ere, (b) hurbilpena ez dugu baztertu. Izan ere, azken helburu gisa, bi hurbilpenen abantailak baliatzea erabaki dugu; beste euskarako hiztegietatik erauzitako hierarkiak eta erlazio semantikoak ere Euskal WordNeten txertatu nahi ditugu. Dagoeneko *Euskal Hiztegi*tik (Sarasola,

1996) hierarkiak eta erlazio semantikoak erauzi dira (Agirre *et al.*, 2003c), eta emaitza horietako batzuk Euskal WordNeten txertatzen hasiak bagara ere, etorkizunean lan hori masiboki egin nahiko genuke.

Honetaz gain, ez dugu baztertzen Euskal WordNeten euskarako edo erdarako beste lan eta formalismoetako informazioa gehitzea; esate baterako, dagoeneko IXA taldean ezagutza lexiko-semantikoaren arloan lortutako emaitzak (Arriola, 2000; Aldezabal, 2004; Martínez, 2005; Lersundi, 2005; Ansa *et al.*, 2005), edota WordNeten eredutik gertu dauden beste lan konputazionaletako informazioa —azterketa bibliografikoan aipatutakoena, adibidez— oso baliagarria izan dakiguke.

Aditzen kasuan, esate baterako, ia eredu guztiak bat datoz multzokatze semantiko zabalagoak egitearekin, adiera oso zehatzak izanda corpus bat etiketatzea oso zail izaten baita. Ildo honetatik, III.2.4 atalean azaldu dugun *PropBank* aipa dezakegu. EBL honetako sarrera lexikalak *VerbNet*en (Kipper *et al.*, 2000) dagozkien sarrerekin lotuta daude. Aldi berean, VerbNeteko sarrera bakoitza WordNeteko *synset* batekin (edo gehiagorekin) loturik dago. Hortaz, lotura honi probetxua atera geniezaioke gure EBLko aditzak VerbNeteko eta PropBankeko informazio sintaktiko-semantikoarekin aberasteko. Arrazoi honengatik eta LNPn rolen etiketatze automatikoak hartu duen indarrarengatik, IXA taldea ere aditzentzat eredu hau garatzen hasi da euskararako (Agirre *et al.*, 2006d), eta etorkizunean Euskal WordNetekin lotzeko asmoa dago.

- **Hautapen-murriztapenen aztertzea:**

Euskal WordNeten aberasketan zabaldutako beste ikerlerroa hautapen-murriztapenena da. IXA taldean arlo honen inguruan lortutako emaitzak (Martínez, 2005) Euskal WordNeten txertatu aurretik ebaluatu ditugu. Azterketa hau hastapenetan dago eta etorkizunean gehiago sakondu nahi dugu. Alde batetik, kirolaren domeinuaz gain, beste domeinu batzuetako aditzak ere aztertu nahiko genituzke (finantzaren domeinukoak, adibidez). Bestalde, eskuratze-tekniken algoritmoak hobetzen saiatuko gara, eta eskuratze-teknika mota gehiagorekin ere probatu nahi dugu.

Euskararen hautapen-murriztapenei dagokienez, euskarako eskuratze-teknikak hobetzeko semantikoki etiketatzen ari garen corpusa (EuSemcor) erabiltzea pentsatua dugu. Azken helburua, eskuratze-teknika egokiarekin jo ondoren, eskuratzen diren hautapen-murriztapenak Euskal WordNeten txertatzea da.

Bibliografia

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., eta Urkia M. A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages*, Granada, 1998a.

Aduriz I., Aldezabal I., Ansa O., Artola X., eta Díaz de Illarraza A. EDBL: a multi-purposed lexical support for the treatment of Basque. *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, 1998b.

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

Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Illarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque taggen at morphological and syntactic levels fot the automatic processing. In Wilson A., Rayson P., eta Archer D., editors, *Corpus Linguistics Around the World*, Book series: Language and Computers, 1–15, Rodopi (Holanda), 2006.

Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian M., eta Pociello E. Euskal WordNet: euskararako ezagutza-base lexiko-semantikoa. *Euskalingua*, (7), 2005a. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian M., eta Pociello E. Improving the Basque WordNet by corpus annotation. *Proceedings of Third International WordNet Conference*, Jeju (Korea), 2006a. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian M., eta Pociello E. A methodology for the joint development of the Basque WordNet and Semcor. *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*, Genoa (Italia), 2006b. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Agirre E., Aldezabal I., eta Pociello E. A pilot study of English selectional preferences and their cross-lingual compatibility with Basque. *Proceedings on International Conference on Text Speech and Dialogue (TSD)*, Ceske Budejovice (Txekiar Errepublika), 2003a. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Agirre E., Aldezabal I., eta Pociello E. Lexicalization and multiword expressions in the Basque WordNet. *Proceedings of Third International WordNet Conference*, Jeju (Korea), 2006c. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

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

Agirre E., Ansa O., Arregi X., Arriola J., Díaz de Ilarraza A., Pociello E., eta Uria L. Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. *Proceedings of First International WordNet Conference*, Mysore (India), 2002. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Agirre E., Arregi X., Arriola J., eta Artola X. EDBL: euskararen datu-base lexikala. Barne-txostena (LSI/TR 8-94), Euskal Herriko Unibertsitatea, 1994a.

Agirre E., Atserias J., McCarthy D., Real F., Rigau G., eta Rodríguez H. MEANING: developing multilingual web-scale language technologies. Working paper 5.2a. Barne-txostena, 2003b.

Agirre E., Atutxa A., Gojenola K., eta Sarasola K. Exploring portability of syntactic information from English to Basque. *Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC)*, Lisboa (Portugal), 2004.

Agirre E. eta Lersundi M. Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. *Proceedings of the Annual SEPLN Meeting*, 2001.

Agirre E. eta Lersundi M. Semantic interpretations of postpositions and prepositions: a multilingual inventory for Basque, English and Spanish. *Workshop on The linguistic dimensions of prepositions and their use in computational linguistics formalisms and applications*, Tolouse, France, 2003.

Agirre E. eta Martínez D. Learning class-to-class selectional preferences. *Proceedings of the Workshop "Computational Natural Language Learning"*, Tolosa (Frantzia), 2001.

Agirre E. eta Martínez D. Integrating selectional preferences in WordNet. *Proceedings of First International WordNet Conference*, Mysore (India), 2002.

Agirre E. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, Donostia, 1999.

Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian M., eta Pociello E. Eusemcor: euskarako corpusa semantikoki etiketatzeko eskuliburua: editatze- etiketatze- eta epaitze-lanak. Barne-txostena, Euskal Herriko Unibertsitatea, 2005b.

Agirre E., Aldezabal I., Etxeberria J., eta Pociello E. A preliminary study for building the Basque Propbank. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa (Italia), 2006d. URL http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016. (2007-07-02an atzitua).

Agirre E., Aldezabal I., eta Pociello E. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: Euskal WordNet. *GOGOA: Euskal Herriko Unibertsitateko Hizkuntza, Ezagutza, Komunikazio eta Ekintzari buruzko Aldizkaria*, V-2:237–266, 2005c. URL http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016. (2007-07-02an atzitua).

Agirre E., Aldezabal I., eta Pociello E. Lexicalization and multiword expressions in the Basque WordNet. In Fernández B. eta Laka I., editors, *Andolin gogoan: Essays in honour of the Professor Eguzkitza*, 51–68. Euskal Herriko Unibertsitatea, 2006e. URL http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016. (2007-07-02an atzitua).

Agirre E., Ansa O., Arregi X., Artola X., Zubillaga X., Díaz de Ilarraza A., eta Lersundi M. A conceptual schema for a Basque lexical-semantic framework. *Conference on Computational Lexicography and Text Research*, Budapest (Hungaria), 2003c.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Edvard F., eta Sarasola K. Lexical knowledge representation in an intelligent dictionary help system. *Proceedings of COLING'94*, 544–550, Kyoto (Japonia), 1994b.

Agirre E. eta Lopez de la Calle O. Clustering WordNet word senses. *Proceedings of the conference of Recent Advances in Natural Laguage Processing*, Borovets (Bulgaria), 2003.

Agirre E. eta Martínez D. Exploring automatic word sense disambiguation with decision lists and the Web. *Proceedings of the Semantic Annotation And Intelligent Annotation Workshop organized by COLING*, Luxenburgo, 2000. URL http://arXiv.org/abs/cs/0010024. (2007-07-02an atzitua).

Aldezabal A., Ansa O., Arrieta B., Artola X., Ezeiza N., Hernández G., eta Lersundi M. EDBL: a general lexical basis for the automatic processing of Basque. *Proceedings of the IRCS Workshop on Linguistic Databases*, Filadelfia (EEBB), 2001a.

Aldezabal I. *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa. Levin-en (1993) lana oinarri hartuta eta metodo informatikoak baliatuz.* Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2004.

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K., eta Goenaga P. Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus. *Actas del XVII Congreso de la SEPLN Universidad de Jaén*, Jaén, 2001b.

Aldezabal I., Arriola J.M., Díaz de Ilarraza A., eta Sarasola K. *Hizkuntzalaritza Konputazionala.* Udako Euskal Unibertsitatea, 2005.

Alegria I., Ansa O., Artola X., Ezeiza N., Gojenola K., eta Urizar R. Representation and treatment of multiword expressions in basque. *Proceedings of the ACL on Multiword Expressions*, 48–55, Bartzelona, 2004.

Alegria I., Artola I., Sarasola K., eta Urkia M. Automatic morphological analysis of Basque. *Proceedings of the Annual SEPLN Meeting*, Sevilla, 1996.

Allen J. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.

Alonge A., Calzolari N., Vossen P., Bloksman L., Irene Castellón T.M., eta Peters W. The linguistic design of the EuroWordNet database. *Computers and the Humanities*, 32 lib., 91–115. 1998.

Alonso L., Capilla J., Castellón I., Fernández A., eta Vázquez G. The Sensem project: syntactic-semantic annotation of sentences in Spanish. *Proceedings of the International Conference RANLP*, Borovets (Bulgaria), 2005.

Amsler R. *The Structure of the Merriam-Webster Pocket Dictionary.* Doktoretza-tesia, University of Texas, 1980.

Amsler R. eta White J. Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries. Barne-txostena, National Science Foundation, University of Texas, 1979.

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

Aranzabe M., Arriola J., Atutxa A., Balza I., eta Uria L. Guía para la anotación sintáctica manual de Eus3LB (corpus del euskera anotado a nivel sintáctico, semántico y pragmático). Barne-txostena, Eukal Herriko Unibertsitatea, 2003.

Aranzabe M., Arriola J.M., eta Díaz de Illaraza A. Towards a dependency parser of Basque. *Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar*, Geneva (Suitza), 2004.

Arriola J. *EUSKAL HIZTEGIAren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira. Aditz-adibideen analisia Murriztapen-Gramatika baliatuz, azpikategorizazioaren bidean*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2000.

Arriola J., Artola X., Maritxalar A., eta Soroa A. A methodology for the analysis of verb usage examples in a context of lexical knowledge acquisition from dictionary entries. *Proceedings of EACL'99, Linguistically Interpreted Corpora*, Bergen (Norvegia), 1999.

Artola X. *HIZTSUA: Hiztegi-sistema urgazle adimenduaren sorkuntza eta eraikuntza*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 1993.

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

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

Banerjee S. eta Pedersen T. An adapted lesk algorithm for word sense disambiguation using WordNet. *Proceedings of the Third International*

Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02), Mexiko, 2002.

Barwise J. eta Perry J. *Situations and Attitudes*. Bradford Books. MIT Press, 1983.

Bates M., Moser M., eta Stallard D. The IRUS transportable natural language database interface. In Kershberg L., editor, *Expert Database Systems*. Benjaming/Cummings, Menlo Park (Kalifornia), 1986.

Benítez L., Escudero G., Farreras J., eta Rigau G. WWI: a multilingual WordNet interface using the web. Barne-txostena, Departament de LSI, Universitat Politécnica de Catalunya, 1998.

Bentivogli L. eta Pianta E. Extending WordNet with syntagmatic information. *Proceedings of Second Global WordNet Conference*, 47–53, Brno (Txekiar Errepublika), 2002.

Bentivolgi L. eta Pianta E. Expliting parallel texts in the creation of multilingual semantically annotated resources: The Multisemcor Corpus. *Natural Language Engineering*, 11:247–261, 2005.

Binot J. eta Jensen K. A semantic expert using an on-line standard dictionary. *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI)*, 709–714, Milan (Italia), 1987.

Boas H.C. Bilingual FrameNet Dictionaries for Machine Translation. *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, 2002.

Boguraev B. eta Briscoe T. *Computational Lexicography for Natural Language Processing*. Longman - John Wiley and Sons, London - New York, 1989.

Boguraev B. eta Briscoe T. Large lexicons for Natural Language Processing. *Computational Linguistics*, 13(3-4):203–218, 1993.

Borgo S., Guarino N., eta Masolo C. A pointless theory of space based on strong connection and congruence. In Aiello L.C. eta Doyle J., editors, *Principles of Knowledge Representation and Reasoning*. Morgan Kauffman, 1996.

Bresnan J. eta Kaplan R.M. Introduction: grammars as mental representations of language. In Bresnan J., editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge (Massachussetts), 1982.

Brown P., Lai J., eta Mercer R. Aligning sentences in parallel corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, 169–176, Berkeley (Kalifornia), 1991.

Bruce R., Wilks Y., Guthrie L., Slator B., eta Dunning T. NounSense – a disambiguated noun taxonomy with a sense of humour. Barne-txostena, Computer Research Laboratory, New Mexico State University, Las Cruces, NM, 1992.

Buitelaar P. *Systematic Polysemy and Underspecification*. Doktoretza-tesia, Brandeis University, 1998.

Cahill A., McCarthy M., Genabith J., eta Way A. Parsing with PCFGs and automatic F-structure annotation. *Proceedings of the LFG02 Conference*, 2002.

Calzolari N. Issues for lexicon building. In Zampolli A., Calzolari N., eta Palmer M., editors, *Current Issues in Computational Linguistics: Essays in Honour of Don Walker*, 267–281. Giardini Editori e Stampatori - Kluwer Academic Publishers, Pisa - Dordrecht, 1994.

Calzolari N., Charles J.F., Grishman R., Ide N., Lenci A., MacLeod C., eta Zampolli A. Towards best practice for multiword expressions in computational lexicons. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 1934–1940, 2002.

Carreras X. eta Màrquez L. Introduction to the CoNLL-2004 shared task: semantic role labeling. *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning*, Boston, 2004.

Carroll J., Rigau G., Magnini B., Agirre E., Rodríguez H., eta Atserias J. MEANING: cycle 1: Acquisition. Barne-txostena, 2003.

Castellón I. *Lexicografía computacional: adquisición automática de conocimiento léxico*. Doktoretza-tesia, Universitat de Barcelona, 1992.

Chodorow M., Byrd R., eta Heidorn G. Extracting semantic hierarchies from a large on-line dictionary. *Proceedings of the 23rd Annual Meeting Association for Computational Linguistics (ACL)*, 299–304, Chicago (Illinois), 1985.

Chomsky N. *Aspects of the Theory of Syntax*. MIT Press, Cambridge (Massachussetts), 1965.

Chomsky N. *Lectures on Government and Binding. The Pisa Lectures*. Mounton de Gruyter, Berlin - New York, 1987.

Chomsky N. A minimalist program for linguistic theory. *MIT Occasional Papers in Lingusitics*, (1), 1992.

Church K., Gale W., Hanks P., eta Hindle D. Using statistics in lexical analysis. *Lexical Acquisition: Exploting On-Line Resources to Build a Lexicon*, 115–164. Lawrence Erlbaum Associates, Hillsdale (New Jersey), 1991.

Civit M., Aldezabal I., Pociello E., Taulé M., Aparicio J., Màrquez L., Navarro B., Castellví J., eta Martí M. 3LB-LEX: léxico verbal con frames sintácticos-semánticos. *Revista de la Asociación Española para el Procesamiento del Lenguaje Natural*, Granada, 2005a. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Civit M., Castellví J., Morante R., Oliver A., eta Aparicio J. 4LEX: A multilingual lexical resource. *Cross- Language Knowledge Induction Workshop*, Errumania, 2005b.

Collins. *The Harper Collins Spanish-English/English-Spanish Dictionary*. William Collins Sons and Co. Ltd., 1971.

Collins. *Collins Master*. Grijalbo, 1998.

Copestake A. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. *Proceedings of the First International Workshop Inheritance in NLP*, 19–29, Tilburg (Holanda), 1990.

Copestake A. eta Flickinger D. An open source grammar development environment and broad-coverage English grammar using HPSG. *International Conference on Language Resources and Evaluation (LREC)*, Atenas, 2000.

Cruse A. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press, 2000.

Cuypers I., Sánchez A., Schippers L., Adriaens G., Louw M., eta Forest P. Test specifications for EuroWordNet: internal data quiality and application in multilingual information retrieval. Barne-txostena, University of Amsterdam, 1997.

Dalrymple M. *Lexical Functional Grammar*, 34. lib. of *Syntax and Semantics*. Academic Press, Londres (Ingalaterra), 2001.

Demonte V. *Detrás de la palabra. Estudios de gramática del español*. Alianza Editorial, Madril, 1991.

Demonte V. *Teoría sintáctica: de las estructuras a la rección*. Colección Lingüística. Síntesis, 1995.

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

Dorr B. Large-scale acquisition of LCS-based lexicons for foreign language tutoring. *Proceedings of the ACL Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, 1997.

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

Elhuyar. *Elhuyar Hiztegia: euskara-gaztelania*. Elhuyar Kultur Elkartea, 1996.

Elhuyar. *Elhuyar Hiztegi Txikia*. Elhuyar Kultur Elkartea, 1998.

Elhuyar. *Hiztegi Modernoa*. Elhuyar Kultur Elkartea, 2000.

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

Fellbaum C. eta Kegl J. Taxonomic structures and cross-category linking in the lexicon. *Proceedings of the Sixth Eastern States Conference on Linguistics*, 93–104, Columbus, 1989.

Fellbaum C. A semantic network of English verbs. In Fellbaum C., editor, *WordNet: An Electronic Lexicak Data-base*. MIT Press, 1998b.

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

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

Fillmore C.J. Frames and the semantics of understanding. *Quaderni di Semantica*, 6.2 lib. 1985.

Fillmore C.J. eta Baker C.F. FrameNet: Frame semantics meets the corpus. *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh, 2001.

Fontenelle T., Adriaens G., eta de Brackeleer G. The lexical unit in the metal MT system. *MT*, 9:1–19, 1994.

Fox E., Nutter T., Ahlswede T., Evens M., eta Marcowitz J. Building a large thesaurus for information retrieval. *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLP)*, 101–108, Austin (Texas), 1988.

Francis W. eta Kucera H. *Frequency Analysis of English Usage*. Houghton Mifflin Company, Boston (Massachusetts), 1982.

Gazdar G., Klein E., Pullum G., eta Sag I. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge (Massachussetts), 1985.

Gilarranz J., Gonzalo J., eta Verdejo F. An approach to conceptual text retrieval using the EuroWordNet multilingual semantic database. *Proceedings of AAAI-96 Spring Symposium Cross-Language Text and Speech Retrieval*, 1996.

Giuglea A.M. eta Moschitti A. Knowledge discovergin using FrameNet, VerbNet and PropBank. *Proceedings of the Workshop on Ontology and Knowledge Discovering at ECML*, Pisa (Italia), 2004.

Gojenola K. Guneak zuzendutako egitura sintagmatikoen gramatika (HPSG) eta euskararako aplikazioa. Barne-txostena, Euskal Herriko Unibertsitatea, 1998.

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

Gómez F., Hull R., eta Segami C. Acquiring knowledge from encyclopedic texts. *Proceedings of the 4th Conference Applied Natural Language Processing (ANLP)*, 84–90, Sttugart (Alemania), 1994.

Gómez F. Linking WordNet verb classes to semantic interpretation. In Harabagiu S., editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, 58–64. Association for Computational Linguistics, Somerset (New Jersey), 1998.

Grefenstette G. Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. *Proceedings of SIGLEX Workshop on Acquisition of lexical knowledge from text*, Columbus, 1993.

Grishman R., Macleod C., eta Reyers A. Complex syntax: building a computational lexicon. *Proceedings of the 15th annual meeting of the Association for the Computational Linguistics (COLING)*, 268–272, Kyoto (Japonia), 1994.

Grishman R. eta Sterling J. Acquisition of selectional patterns. *Proceedings of COLLING-92*, Nantes (Frantzia), 1992.

Gruber T.R. Towards principles for the design of ontologies for knowledge sharing. *Proceedings of the International Workshop on Formal Ontology*, Padova (Italia), 1993.

Guarino N. Semantic matching: formal ontological distinctions for information organization, extraction and integration. *Information Extraction*, 139–170. Springer, Berlin (Alemania), 1997.

Hale K.L. eta Keyser S.J. A view from the middle. Barne-txostena, Center of Cognitive Science, Cambridge, Massachussetts, 1987.

Harabagiu S.M. eta Moldovan D.I. An intelligent system for question answering. *Proceedings of the 5th Conference on Intelligent Systems*, Reno, 1996.

Hindle D. Noun classification from predicate-argument structures. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 327–329, 1990.

Hindle D. eta Rooth M. Structural ambiguity and lexical relations. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 229–236, 1991.

Ide N. eta Veronis J. Extracting knowledge bases from machine-readable dictionaries: have we wasted our time? *Proceedings of the International Conference on Building and Sharing of Very Large-Scale Knowledge Bases*, 257–266, Japonia, 1993.

Ingria R. Lexical information for parsing systems: points of convergence and divergence. In Walker D., Zampolli A., eta Calzolari N., editors, *Automating the Lexicon: research and Practice in a Multilingual Environment*. Cambrigde University Press, Cambridge, 1988.

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

Jackendoff R.S. *Brain, Meaning, Grammar, Evolution*. Oxford University Press, 2000.

Johnson C.R. eta Fillmore C.J. The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, Seattle (Washington), 2000.

King T.H., Crouch R., Riezler S., Dalrymple M., eta Kaplan R.M. The PARC 700 Dependency Bank. *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest (Hungaria), 2003.

Kipper K., Dang H.T., eta Palmer M. Class-based construction of a verb lexicon. *Proceedings of the Seventh National Conference on Artificial Intelligence*, 691–696, 2000.

Kipper K., Palmer M., eta Rambow O. Extending PropBank with VerbNet semantic predicates. *Workshop on Applied Interlinguas*, Tiburon (Kalifornia), 2002.

Klavans J. eta Tzoukermann E. Dictionaries and corpora: combining corpus and machine-readable dictionary for building lexicons. *Journal of Machine Translation*, 10(3-4):185–218, 1996.

Knight K. Building a large ontology for machine translation. *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, 1993.

Knight K. eta Luk S. Building a large-scale knowledge base for machine translation. *Proceedings of the 12th American Association for artificial intelligence (AAAI)*, 773–778, Seatle (Washington), 1994.

Kohl K.T., Jones D.A., Berwick R.C., eta Nomura N. Representing verb alternations in WordNet. In Fellbaum C., editor, *WordNet: an Electronic Lexicak Data-base*. MIT Press, 1998.

Lenat D. Steps to sharing knowlegde. *Toward very large knowledge bases*, 1995.

Lenat D.B. eta Guha R.V. *Building Large Knowledge-Based Systems*. Addison Wesley, 1990.

Lersundi M. *Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2005.

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

Lewandowski T. *Diccionario de la Lingüística*. Cátedra, 1992.

Lin D. Principle based parsing without overgeneration. *31st Annual Meeting of the Association for Computational Linguistics*, Columbus (Ohio), 1993.

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

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

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

Mann G. Building proper noun ontologies for question answering. *Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks"*, 2002.

Marcus M., Kim G., Marcinkiewicz M., MacIntyre R., Bies A., Ferguson M., Katza K., eta Schasberger B. The Penn Treebank: annotating predicate argument structure. *Proceedings of ARPA Workshop on Human language technology*, San Frantzisko, 1994.

Marcus M., Santorini B., eta Marcinkiewicz M. Building a large annotated corpus of English: The Penn TreeBank. *Computational Linguistics*, (19): 313–330, 1993.

Martínez D. *Supervised Word Sense Disambiguation: facing Current Challenges*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2005.

Matwin S., Szpakowicz S., eta Li X. A WordNet-based algorithm for word sense disambiguation. 1995. URL `http://citeseer.ist.psu.edu/155268.html`. (2007-07-02an atzitua).

McCarthy D. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Doktoretza-tesia, University of Sussex, 2001.

McCarthy D. Relating wordnet senses for word sense disambiguation. *Proceedings of the EACL2006 Wordkshop Making Senses of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, 17–24, Trento (Italia), 2006.

Michiels A. eta Nel J. Approaches to thesaurus production. *Proceedings of the Ninth International Conference on Computational Linguistic*, 227–232, Amsterdam, 1994.

Milhacea R. eta Moldovan D.I. Word Semantics for Information Retrieval: moving one step closer to the semantic web. *International Conference on Tools in Artificial Intelligence*, 2001.

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

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

Miller G.A., Fellbaum C., eta Katherine J.M. Five papers on WordNet. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`. (2007-07-02an atzitua), 1993.

Minnen G. Selective magic HPSG parsing, 1999. URL `http://citeseer.ist.psu.edu/minnen99selective.html`. (2007-07-02an atzitua).

Montemagni S. Extracting typical subjects and objects of verbs from mono- and bi-lingual dictionaries. Barne-txostena, ESPRIT BRA-7315 Acquilex-II, 1994.

Moon Y.J. eta Kim Y.T. Concept-based verb translation in the Korean-English machine translation system. *Journal of the Korea Information Science Society*, 1995.

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

Niles I. eta Pease A. Towards a standard upper ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, 17–19, 2001.

Nishida K., Torisawa K., eta Tsujii J. Efficient HPSG parsing algorithm with array unification, 1999. URL `http://citeseer.ist.psu.edu/408471.html`. (2007-07-02an atzitua).

Oepen S., Flickinger D., Toutanova K., eta Manning C.D. A rich and dynamic Treebank for HPSG. *In Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol (Bulgaria), 2002.

Ohara K.H., Fujii S., Saito H., Ishizaki S., Ohori T., eta Suzuki R. The Japanese FrameNet project: a preliminary report. *Proceedings of Pacific Association for Computational Linguistics (PACLING03)*, 2003.

Onyshkevych B. eta Nirenburg S. The lexicon in the scheme of KBMT things. Barne-txostena, Computing Research Laboratory, New Mexico State Laboratory, 1994.

Osenova P. eta Simov K. The Bulgarian HPSG Treebank: specialization of the annotation scheme. *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT2003)*, Växjö, 2003.

Oxford. *The Oxford Spanish Dictionary*. Oxford University Press, 2003.

Palmer M. eta Xue N. Annotating the propositions in the Penn Chinese Treebank. *Proceedings of the Second Sighan Workshop*, Sapporo (Japonia), 2003.

Palmer M. eta Kingsbury P. From TreeBank to PropBank. 2003. URL `http://citeseer.ist.psu.edu/574953.html`. (2007-07-02an atzitua).

Pasca M. eta Harabagiu S.M. The informative role of WordNet in open-domain question answering. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, 2001.

Pereira F., Tisgby N., eta Lee L. Distributional clustering of English words. *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 183–19, 1993.

Pociello E. Aditzen hautapen-murriztapenak: kirol domeinura mugatutako ingeleseko hautapen-murriztapenak eta euren baliagarritasuna euskararako. Hastapeneko lana. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2004a. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Pociello E. Sintaxi-semantika elkargunea zenbait teoriatan: euskararen ezagutza-basea lexiko-semantikorantz. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2004b. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/kidearen_argitalpenak?kidea=1000809016`. (2007-07-02an atzitua).

Pollard C. eta Sag I. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, 1994.

Popowich F. eta Vogel C. Chart parsing Head-Driven Phrase Structure Grammar. Barne-txostena 90-1, 1990.

Poznanski V. eta Sanfilippo A. Detecting dependencies between semantic verb subclasses and subcategorization frame in text corpora. *Procceedings of the ACL-SIGLEX WSHP on Extracting Lexical Knowledge from Text*, 1993.

Pradhan S., Hacioglu K., Ward W., Martin J., eta Jurafsky D. Semantic role parsing: adding semantic structure to unstructured text. *Proceedings of the International Conference on Data Mining (ICDM-2003)*, Melbourne, 2003.

Pustejovsky J. The generative lexicon. *Computational Linguistics*, 17(4), 1991.

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

Pustejovsky J., editor. *Semantics and the Lexicon*. Kluwer Academic Publishers, 1993.

Resnik P. A class-based approach to lexical discovery. *Prooceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 1992.

Resnik P. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Doktoretza-tesia, University of Pennsylvania, 1993.

Resnik P. Disambiguating noun groupings with respect to WordNet senses. *Proceedings of the 3rd Workshop on Very Large Corpora*, MIT, 1995.

Ribas F. *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy*. Doktoretza-tesia, Universitat Politècnica de Catalunya, 1995.

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

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

Rigau G. *Automatic Acquisition of Lexical Knowledge from MRDs*. Doktoretza-tesia, Universitat Politècnica de Catalunya, 1998.

Ruppenhofer J., Baker C., eta Fillmore C. The FrameNet database and software tools. *Proceedings of the Tenth Euralex International Congress*, 1. lib., 371–375, Copenhage, 2002.

Sag I., Baldwin T., Bond F., Copestake A., eta Flickinger D. Multiword Expressions: A pain in the neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, 1–15, Mexiko, 2002.

Saint-Dizier P. Constructing verb semantic classes for French: methods and evaluation. *Proceedings of the COLING*, 1996.

Sánchez A. Informatización de diccionarios convencionales: un sistema de consulta para el "Diccionario Ideológico de la lengua española"de J. Casares. *Proceedings fo the 7th Annual Meeting de la Sociedad Española del Procesamiento del Lenguale Natural (SEPLN)*, Valentzia, 1991.

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

Sowa J. *Kwoledge Representation*. Brooks/Cole - Pacific Grove, 2000.

Subirats-Rüggeberg C. eta Petruck M.R.L. Surprise: Spanish FrameNet! *Workshop on Frame Semantics, International Congress of Linguists*, Praga (Txekiar Errepublika), 2003.

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

Tomuro N. Tree-cut and a lexicon based on systematic polysemy. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburg, 2001.

Utsuro T., Matsumoto Y., eta Nagao M. Verbal case frame acquisition from bilingual corpora. *Proceedings of International Joint Conference of Artificial Intelligence (IJCAI)*, Chambery (Frantzia), 1993.

UZEI. *Sinonimoen Hiztegia*. UZEI, 1999.

Vázquez G., Fernández A., eta Martí M.A. *Clasificación Verbal. Alternancias de diátesis*. Quaderns de Sintagma 3. Edicions de la Universitat de Lleida, 2000.

Vendler Z. *Linguistics in Philosophy*. Cornell University Press, Ithaca (New York), 1967.

Verkuyl H. *On the Compositional Nature of the Aspects*. Reidel, Dordrecht, 1972.

Vossen P., editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1998.

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

Vossen P. EuroWordNet general document. URL `http://www.illc.uva.nl/EuroWordNet/docs.html`. (2007-07-02an atzitua), 1999.

Way A. Translating with examples: the LFG-DOT models of translation. *Recent Advances in Example-Based Machine Translation*. Kluwer Academic Publishers, 2003.

Wilks Y. Preference semantics. In Keenan E., editor, *The Formal Semantics of Natural Language*. Cambridge University Press, 1973.

Wilks Y., Slator B., eta Guthrie L. *Electric words: dictionaries, computers and meanings*. The MIT Press, 1996.

Yarowsky D. Word sense disambiguation using statistical models of Rogets categories trained on large corpora. *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, 454–460, Nantes (Frantzia), 1992.

Yokoi T. The impact of the EDR electronic dictionary on very large knowledge bases. *Toward very large knowledge bases*, 1995.