Filologia eta Geografi-Historia Fakultatea

Aditzen hautapen-murriztapenak: Kirol domeinura mugatutako ingeleseko hautapen-murriztapenak eta euren baliagarritasuna euskararako. Hastapeneko lana.

Ikaslea: ELISABETE POCIELLO IRIGOIEN
Zuzendaria: JOSE MARI ARRIOLA EGURROLA

Ikerkuntza-aldia gainditzeko lana, 2004ko abenduaren 2a.

Gaien aurkibidea

1 SARRERA ETA AURKEZPEN OROKORRA

1.1 Sarrera

Lan honekin semantika konputazionalaren ikerketa hedatu nahi dugu. Gaur egungo azterketa teoriko nahiz konputazionalean, osagai lexikalek garrantzi handia hartu dute. Osagai lexikalen artean aditza bereizten da, bera baita perpausa ulertzeko osagai inportanteena. Bestalde, aditzak dakarren informazioa oso baliagarri izan daiteke beste ataza konputazionaletarako: egitura sintaktikoen desanbiguazioa, hitzen adieren desanbiguazioa, anaforaren ebazpena, eta abar.

Ikuspegi honetatik abiatuz, eta aditzen jokaeran eta beraien ikerketan sakondu ahal izateko, aditzen objektu/subjektuen hautapen-murriztapenen azterketan murgildu gara, zehazkiago esanda, automatikoki ikasitako (eskuratutako) hautapen-murriztapenen[1] azterketan.

Horrela, Lengoaia Naturalaren Prozesamenduaren (aurrerantzean LNP[2]) arloan kokatutako lan honen helburua hautapen-murriztapenen azterketa automatikoa egitea da. Horretarako, ingeleserako automatikoki ikasi diren hautapen-murriztapenetan oinarritu gara lehenengo, gero hauek euskararentzat baliagarriak izan daitezkeen aztertu ahal izateko. Hau da, ingeleseko hautapen-murriztapenak eskuratzeko erabili diren teknika ezberdinak aurkeztu eta ebaluatu ditugu, hauen aplikazioa eleanitza izan daitekeela frogatu nahiaz.

Ikerlan hau mugatzearren, gure ustez kirol domeinuan nagusi diren aditz batzuetan oinarritu gara (*jokatu, entrenatu, irabazi, galdu* eta *berdindu*). Bestalde, EuroWordNet, adiera inbentario gisa erabili dugu, bertan ingeleseko eta euskarako aditz-adierak lotuak datozelako. Beraz, aditz hauen EuroWordNet-eko kirol adieretan funtsatu gara, eta ingeleseko itzulpenak hauetatik lortu ditugu[3]. Horrela bada, ikerlan honen parametro nagusiak domeinua eta adierak dira, kirol domeinuarekin bat datozen aditzen adieren (aditz-adieren[4]) hautapen-murriztapenak lortu ditugulako.

Ingelesa, abiapuntu gisa, ez dugu ausaz aukeratu. Azken urteotan LNPn corpusetan oinarritutako ikerkuntza suspertu da, makinak hizkuntzen egitura eta ezaugarri asko eta asko hauetatik eskura baititzake. Horretarako, garrantzitsua da corpus handiak izatea, zenbat eta corpus handiagoa izan, orduan eta informazio gehiago eta zehatzagoa lor daitekeelako. Egun, ingelesa da munduan gehien hitz egiten den hizkuntza, eta arrazoi horregatik, hizkuntza hau da euskarri informatikoan corpus handiena duena. Hala, LNPren ikuspegitik, ingelesa oso baliabide aberatsa da eta ondorioz, aurrerapen gehienak honen

[1] 2.1 atalean azalduko ditugu hautapen-murriztapen kontzeptua eta hautapen-murriztapen mota ezberdinak.

[2] Txosten honetan erabili ditugun laburdurak izen bereziak bezala deklinatuko ditugu. Maileguei dagokioenez, bukatzen diren hizkiaren (bokale eta kontsonante) arabera deklinatuko ditugu.

[3] EuroWordNet eta aditzen hautaketari buruzko argibideak zehazkiago emango ditugu 3. atalean.

[4] Aditzaren adiera jakin bati erreferentzia egiteko, aditz-adiera darabilgu. Aditzak adieraz ditzakeen adiera guztiei buruz ari garenean, aldiz, aditz-forma.

inguruan garatzen dira.

Hedapen gutxiagoko hizkuntzak (euskara, esate baterako), aldiz, informatikoki baliagarriak diren corpus txikiagoak dituzte, batzuen kasuan txikiegiak horietatik emaitza zuzenak lortzeko. Hori dela eta, eta hizkuntza hauen 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 ikasi izan dena bata bestearekin parekatu eta hizkuntza batekin bestea aberastea ahalbidetzen duena[5]. Hala, hizkuntza batentzat ikasitakoa beste hizkuntza batentzat baliagarria izan daiteke; eta normalean, abiapuntu gisa, konputazionalki baliabide gehiago dituen hizkuntza hartzen da. Hortaz, hipotesi honen arabera, jokabide linguistiko batzuk eleanitzak dira, eta ondorioz, hizkuntza batentzat automatikoki ikasitako 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 izanda, nahikoa litzateke makinak corpus aberatsenetatik ikastea, eta emaitzak zuzenean beste hizkuntzara itzultzea.

Ikerlan honen beste helburu nagusia hipotesi hau aztertzea da, eleaniztasunak izan ditzakeen aldaera eta parametroak kontuan hartuaz. Horretarako, ondorengo pausoak jarraitu ditugu:

(a) **Ingeleseko corpusetik eskuratutako hautapen-murriztapenen ebaluazioa:** ingeleseko hiru corpus aukeratu ditugu (EFE, SemCor eta British National Corpus), eta hauen gainean eskuratze-teknika desberdinak erabili ondoren, bakoitzetik lortutako emaitza ebaluatu dugu. Hots, ingeleseko hautapen-murriztapenen ikasketa automatikoa zuzena izan den ala ez aztertu dugu.

(b) **Eleaniztasuna:** Hiru corpus horien gainean erabilitako eskuratze-teknika ezberdin horietatik lortutako hautapen-murriztapenak euskarara itzuli eta euskararako egokiak diren egiaztatu dugu. Honekin frogatzen saiatu gara, hizkuntza baterako ikasitako hautapen-murriztapenak beste hizkuntza batean erabilgarriak izan daitezkeela.

(c) **Euskarako corpusetik eskuratutako hautapen-murriztapenen ebaluazioa:** euskarako corpus bat aukeratu dugu (Euskaldunon Egunkaria) eta honen gainean eskuratze-teknika desberdinak erabili ondoren, bakoitzetik lortutako emaitza ebaluatu dugu, hots, hautapen-murriztapenen ikasketa automatikoa zuzena izan den ala ez aztertu dugu.

(d) **Ingeleseko eta euskarako corpusetatik eskuraturiko hautapen-murriztapenen konparaketa:** Emaitza guztiak baliatuz, euskarari zer bide (ingelesetik itzultzea ala euskarako corpusetan oinarritzea) hobeto egokitzen zaion ondorioztatu dugu.

Ikerlan 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 digun bidea aurkituz.

[5] Proiektu honi buruzko informazio gehiago, honekin batera entregatutako beste ikerkuntza lanean aurki daiteke (Pociello, 2004).

Azkenik, esan behar dugu ikerlan honetan eskuratze-tekniketatik lortutako emaitzen gainean egin dugula lan, hau da, emaitzen ebaluazio linguistikoan aritu gara, hain zuzen ere. Horregatik, txosten honetan ez dugu sakonduko eskuratze-teknika hauek garatzeko erabili diren hainbat prozesamendu eta algoritmo informatikoetan. Horrenbestez, azterketa honen ondorioz, informatikariek aditzen informazio lexikoa ikasteko baliabideak hobetzeko aukera izango dute.

1.2 Helburuak eta ekarpenak

Sarreran ikerlan honen egitasmo eta helburuen berri eman dugun arren, egokia iruditu zaigu helburuak eta ekarpenak zehazki zerrendatzea. Ikerlan honekin bi helburu bete nahi ditugu:

(a) Hainbat eskuratze-teknika erabiliaz ingeleseko eta euskarako corpus ezberdinetatik ikasitako hautapen-murriztapenak aztertu eta konparatu.

(b) Hautapen-murriztapenak eleanitzak izan daitezkeen aztertu.

Ikerlan ugari egin dira hautapen-murriztapenen ikasketa automatikoari buruz, baina ez dira hain ugariak ikasketa automatiko horren ebaluazio linguistikora mugatu diren lanak, are gutxiago euskarari dagozkionak. Lan honen ekarpen garrantzitsu bat horretan datza, hain zuzen ere. Egun erabiltzen diren hainbat eskuratze-tekniken azterketa eta ebaluazio linguistikoaren ondoren, lan honen bidez, euskarako hautapen-murriztapenen ikasketa automatikoa garatzeko aukera eta proposamen berriak eskaintzen dira.

Ikerlan honek dakarren beste ekarpen nagusia, eleaniztasunaren hipotesia bideragarria dela da, hots, hizkuntza baterako ikasitako hautapen-murriztapenak beste hizkuntza batean erabilgarriak izan daitezkeela. Euskararen LNPrako ekarpen garrantzitsua dugu hau, euskarak corpus eta baliabide kopuru txikiagoak dituelako, eta hipotesi honetaz baliatuz gero, baliabide gehiago duten hizkuntzetatik xurgatzeko aukera eskaintzen zaigulako.

Berriro ere aipatu beharra dago, honako hau hastapeneko lana dugula, hau da, aditz gutxi batzuekin egindako saiakera bat dela. Hemen aurkeztuko ditugun emaitzetatik eta ondorioetatik abiatuta, azterketa honen esparrua zabaltzeko asmoa dugu.

1.3 Proiektuaren testuingurua

Ikerketa hau IXA taldean semantika konputazionalaren alorrean egiten ari den lanaren barruan kokatzen da. IXA taldeak hamar urte baino gehiago daramatza euskararen tratamendu automatikoan lanean. Taldeak zenbait aplikazio eta tresna sortu ditu, horien artean aipagarrienak euskararen datu-base lexikala (EDBL) eta XUXEN zuzentzaile ortografikoa izanik. Besteak beste, ezagutza lexiko-semantikoaren eskuratzean ere hainbat lan burutu dira bertan. Lan horietako batzuk jadanik tesiak eman dituzte, eta beste batzuk bidean dauden tesietan kokatzen dira:

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

- Euskarako aditzen alternantzien eta klase semantikoen azterketa (Aldezabal, 2004).

- Aditzen hautapen-murriztapenen eskuratzen automatikoa WordNet eta corpusetan oinarrituta (Agirre eta Martínez, 2001, 2002)[6].
- Aditzen adieraren desanbiguazioa (Agirre *et al.*, 2001).
- Erlazio lexiko-semantikoen gauzatze sintaktikoa (Agirre eta Lersundi, 2001)[7].

Bestalde, esperimentu hau "MEANING: Developing Multilingual Web-Scale Language Technologies" (IST-2001-34460) proiektuaren (Rigau *et al.*, 2003) barne ere kokatzen da.

Lanaren testuinguruaren berri eman ondoren, 1.4 atalean proiektuaren ekarpena eta landutako alderdi nagusiak aurkezteari ekingo diogu.

1.4 Eduki nagusiak

Txosten hau sei atal nagusietan banatzen da. Sarrera honen ondoren, 2. atalean, hautapen-murriztapenen inguruan jardungo gara; zer diren eta beraien eskuratzea nola eta zertarako egiten den. 3. atalean, esperimentu honetan erabili diren baliabideen berri emango dugu (corpusak, eskuratze-teknikak eta EuroWordNet). 4. eta 5. ataletan ingeleseko eta euskarako hautapen-murriztapenen azterketan sakonduko dugu. Eta azkenik, 6. atalean, lanaren ondorioak eta etorkizuneko lanak aipatuko ditugu.

[6]Egin bidean dagoen tesia.
[7]Egin bidean dagoen tesia.

2 HAUTAPEN-MURRIZTAPENAK

2.1 Zer dira?

> Limitaciones de la combinabilidad sintagmática de las unidades léxicas, especialmente nombre y verbo, basados en los rasgos semánticos y sintácticos inherentes de las unidades, que regulan su combinabilidad (p. ej. verbos como leer, escribir, pensar, saber, pueden unirse generalmente sólo a un sujeto dotado del rasgo [+humano]). (Lewandowski, 1992, 301. orr.)

Hortaz, hautapen-murriztapenak hitz batek, honek duen adieraren arabera, testuinguruan har ditzakeen osagai linguistikoak murrizten ditu. Beste hitz batzuetan esanda, hautapen-murriztapena da **hitz baten adiera batek** testuinguruan izan dezakeen agerkidetzaren zerrenda. 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 murriztu dezake. Adibidez, *idatzi* aditzak, subjektu gisa [+gizaki] tasuna eskatzen du; [+gizaki] izango da bere subjektu hautapen-murriztapena, alegia.

Aditza ez da murriztapenak egin ditzakeen bakarra, izenek, adjektiboek eta postposizioek ere egin ditzakete.

> Las propiedades señaladas definen no sólo a los verbos sino también a los nombres, adjetivos y preposiciones, en la medida en que estas otras categorías léxicas pueden imponer límites sobre la clase semántica de los elementos (complementos) que los acompañan. Limitaciones de la combinabilidad sintagmática de las unidades léxicas, especialmente nombre y verbo, basados en los rasgos semánticos y sintácticos inherentes de las unidades, que regulan su combinabilidad. (Fernández, 1995, 91. orr.)

Esate baterako, adjektibo batek ezin ditu nahi adina izen modifikatu, izenaren klase semantikoaren arabera murriztuko ditu bere osagaiak. Adibidez, *goxo* adjektiboak, bere adiera hedatuenean ('zapore onekoa', hain zuzen ere), osagai gisa *janaria* edo *edaria* izango du beti.

Resnik-ek (1993) hautapen-murriztapenak azaltzen ditu, bere ezaugarri garrantzitsuenak zerrendatuaz:

1. **Hautapen-murriztapenak ezaugarri sintaktiko eta semantikoetan oinarritzen dira**. Gorago esan bezala, hautapen-murriztapenak hitz baten osagai sintaktikoak zeintzuk diren markatzen du, ezaugarri semantikoak abiapuntutzat hartuz. Adibidez, *idatzi* aditzarekin adierazten dugun ekintza, gizakiek bakarrik egin dezaketela badakigu. Hortaz, *idatzi* aditzarekin [+gizaki] tasuna daraman osagai bat agertzen bada, osagai horrek perpausean betetzen duen funtzio sintaktikoa subjektu dela jakingo dugu.

2. **Hautapen-murriztapenak ez daude joera orokorrera bakarrik mugatuak**. Badira hautapen-murriztapen batzuk oso zehatzak direnak, hau da, erlazio semantiko oso zehatzak egon daitezkeen neurrian, hauentzako hautapen-murriztapenak ere egongo dira. Adibidez, medikuntza domeinuan *onbera* adjektiboa erabiltzen denean, *tumore* izena izango da osagai gisa egon daitekeen bakarra. Beraz, adjektibo-adiera horren hautapen-murriztapena izen bakarrari dagokio.

3. **Hautapen-murriztapenak ez dira sintagmen buruetatik bakarrik osatzen, baita sintagma osoetatik ere**. Hala, sintagman dauden hitz guztien ezaugarri semantikoak izan beharko dira kontuan. Hau garbi asko ikus daiteke ondorengo adibideetan:

   (1) [Bizilagun bizarduna] alaba batez erditu zen.

   (2) [Bizilaguna] alaba batez erditu zen.

   Lehenengo adibidea, bigarrengoarekin alderatuz, arrotza egiten zaigu. Biek subjektu gisa, sintagma-buru bera duten arren (*bizilagun*), (1)en kasuan sintagma hori beste hitz batez osatua dago (*bizardun*), eta horixe da *erdituk* murrizten duen [+emakumezko] komunztadura semantikoa bortxatzen duena, hain zuzen ere.

4. **Hautapen-murriztapenak hausteak ez du adierazten perpaus horrek esanahirik ez duenik**. Nahiz eta (1) adibidea semantikoki arraroa izan, denok ondoriozta dezakegu bizilagunak (ez dakigu ziur gizonezkoa ala emakumezkoa den) alaba bat izan duela.

2.1.1 Hautapen-murriztapenei buruzko eztabaida

Hautapen-murriztapenak bateraezinak diren bi ikuspegietatik izan dira aztertuak. Batzuk hautapen-murriztapenak ikuspegi hertsi batetik aztertzen dituzte, besteek, ordea, ikuspegi zabalago batetik

1. **Ikuspegi hertsia - Katz eta Fodor (1964):** Katz eta Fodor-ek (1964) hautapen-murriztapenak semantikari bakarrik dagozkiola uste dute, eta inola ere ez, munduaren ezaguerari. Esanahi guztiak primitibo finko batzuekin antolatzen direla aldarrikatzen dute, eta primitibo hauek ez badira betetzen, ezin dela inolako interpretapenarik eskuratu. Beraz, Katz eta Fodor-en ustetan, hautapenak hertsiak dira, eta hauek balio jakin bat hartu behar dute perpausak interpretapen egoki bat izan dezan.

2. **Ikuspegi zabala - Wilks (1975):** Wilks-en (1975) proposamenak, aldiz, hautapen-murriztapenak ikuspegi zabalago bat behar dutela aldarrikatzen du. Esanahia antolatzen duten primitiboak ez dira horren finkoak eta nahiz eta hauek ez bete, perpausa ez da alde batera utziko. Teoria honek dioenez, semantikaren ereduak zehatz-mehatz betetzen ez dituzten perpausak (metaforak, adibidez) baztertu beharrean, hauek interpretapen semantikoa hartuko dute. Hortaz, hautapen-murriztapenak semantikaren mugak gainditzen dituzte, munduaren ezaguerari erreparatuaz.

Horrela bada, ikuspegi hertsiak (1) bezalako perpaus bat ez luke onartuko, semantikoki okerra delako. Ikuspegi zabalak, berriz, ez luke baztertuko, nahiz eta semantikoki arrotza izan, gizakiok horrelako perpausak ulertzeko gai garelako.

Honen arabera, Wilks-ek terminologia aldaketa bat proposatzen du, ikuspegien arteko desberdintasuna azpimarratu nahian. Ikuspegi hertsiko hautapenak **hautapen-murriztapenak** izaten jarraituko dute baina ikuspegi zabalekoak, aldiz, **hautapen-hobespenak**[1].

LNPn, eta ondorioz gure ikerlanean, nahiago izan da Wilks-en bidetik jarraitzea, hots, hautapen-murriztapenak hobespenen gisa ulertzea. Hauek corpusetako datuekin lan egiteko malgutasun gehiago eskaintzen dute, corpusetako datuak ez baitatoz beti bat gure intuizioekin.

> This notion of "preferring" is important. (...) We cannot enter such preferences as stipulations, as many linguistic systems do, such as Fodor and Katz's "selectional restrictions". For we can be said to drink gall and wormwood, and cars are said to drink gasoline. It is proper to prefer the normal (quite different from probabilistically expecting it, we shall argue), but it would be absurd, in an intelligent understanding system, not to accept the abnormal if it is described. Not only everyday metaphor but the description of the simplest fiction require it. (Wilks, 1975, 196. orr.).

Nahiz eta guk Wilks-en ildotik jarraitu, "hautapen-murriztapen" terminoari eutsiko diogu, hau baita ikuspegi batean zein bestean oro har erabili izan dena. Gainera, euskaraz emandako itzulpena ("hautapen-hobespen") arrotza da garatzen hasi berria den arlo honetan.

2.2 Hautapen-murriztapenen eskuratzea

2.2.1 Eskuratze-metodo desberdinak

LNPn, hautapen-murriztapenak ikasteko garaian, hiru metodo dira aipagarrienak: lehenengoa, **introspekzioa**; bigarrena, **hiztegietan oinarrituriko eskuratze automatikoa**[2]; eta azkenik, **on-line corpusetan oinarrituriko eskuratze automatikoa**.

Introspekzioa

Orain dela hamarkada bat, hautapen-murriztapenak eskuz egiten ziren, hizkuntzalariaren iritzi eta intuizio linguistikoen arabera (Lenat eta Guha, 1990). Pertsonen intuizioetan oinarritzeak baditu bere arriskuak: egindako lana hizkuntzalariaren subjektibotasunaren menpe egongo da, baita honen akats, ahazte, eta kontraesanen menpe ere. Bestalde, eskuratze-mota hau garestiegia da, eta datu-kopuru bera eta gehiago lortzeko badaude beste metodo merkeago batzuk.

Arrazoi hauengatik, gaur egun, LNPn metodo hau alde batera geratu da. Haatik, introspekzioa eskuratze-metodo gisa guztiz *fidagarria* izan ez arren, automatikoki ikasitako hautapen-murriztapenak ebaluatzeko erabiltzen da. Guk geuk esperimentu honetan introspekzioaz baliatu gara eskuratutako emaitzak ebaluatzeko[3].

[1] Ingelesez, *selectional restrictions* eta *selectional preferences*, hurrenez hurren.
[2] Ingelesez *automatic acquisition from machine-readable versions of dictionaries (MRD).*
[3] Honi buruz, 4.2 eta 5.1 ataletan mintzatuko gara.

Eskuratze automatikoa hiztegietatik

Lexikografikoak hiztegigintzan hiztegi-sarrera bat definitzerakoan sarrera horrek hartzen dituen hautapen-murriztapenen azterketa eta adierazpena egiten du. Hiztegi hauek informatikoki baliagarriak direnean, makinak hiztegi hauetatik bertatik erauz ditzake lexikografoak hiztegi-sarrera bakoitzari egokitu dion hautapen-murriztapena (Montemagni, 1994).

Hala ere, metodo honen bidez lortutako hautapen-murriztapenak ez dira guztiz fidagarriak, lexikografoek sortutako hiztegiak baitira, eta gorago esan dugun bezala, pertsonen intuizioetan oinarritzeak bere alde txarrak dauzka: objektibotasun falta eta hutsegite sistematikoak egotea adibidez.

Bestalde, hiztegietatik informazio interesgarria lor daitekeen arren, hiztegietako sarrera guztiek ez dute hautapen-murriztapenak eraucteko adina informazio ematen, informazio hori ez delako esplizituki agertzen hiztegi-sarrera guztietan.

Eskuratze automatikoa corpusetik

Metodo honen bitartez makinak automatikoki eskura ditzake hitz bati dagozkion hautapen-murriztapenak, hitz horrek corpusean dituen agerpen guztien testuinguruan oinarrituz.

Aipatu ditugun metodoen artean, corpusean oinarritutako ikasketa da LNPn adostasun handiena lortu duen metodoa:

1. Corpusen tamaina handiari esker, aztertu beharreko hitzaren adibide nahikoak eskuratu ahal izango ditugu.

2. Corpusa domeinuka dagoenean, domeinu zehatz bati dagokion informazio linguistikoa eskuratzeko aukera izango dugu.

3. Hiztegiek ez bezala, metodo honek corpus batetik eskuratutako datuen maiztasuna ere eskaintzen digu.

Guk egindako esperimentuak corpusak ere hartu ditu ardatz gisa.

2.2.2 Formalizazioa

Atal honetan, corpusean oinarritutako eskuratze-metodoan erabiltzen diren eskuratze-teknika nabarmenenei buruz jardungo gara: **hitzean oinarritzen direnak** eta, **klase semantikoan oinarritzen direnak**[4].

Hitzean oinarritzen diren eskuratze-teknikak

Ikerlari batzuk (Church *et al.*, 1991; Hindle, 1990; Hindle eta Rooth, 1991; Pereira *et al.*, 1993, esate baterako) predikatu eta argumentu baten arteko harreman semantikoak atzitzeko, hitzean bertan funtsaturiko saiakuntzak egin dituzte.

Hurbilpen hau semantika berdintsua duten hitzek testuinguru berdintsuetan agertzeko duten joeraz baliatzen da.

[4]Ingelesez, *word-based* eta *class-based*, hurrenez hurren.

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

1. irudia: *Drink* aditzaren objektuak hitzen hurbiltasunean oinarritutako teknika erabiliaz (Hindle, 1990).

> [...] 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. orr.)

Harreman linguistiko askok semantikoki parekoak diren hitzak eskatzen dituzte. Hala, 2.1 atalean aipatu den legez, adjektibo batek ezin ditu nahi adina izen modifikatu, izenaren klase semantikoaren arabera murriztuko baititu bere osagaiak. Adibidez, *goxo* adjektiboak, bere adiera hedatuenean, osagai gisa *janaria* izango du beti. Horrela bada, teknika hauek hizkuntzak eskaintzen dizkigun distribuzioaz baliatuko dira hautapen-murriztapenak eskuratu ahal izateko.

Hindle-ek (1990), adibidez, izenen arteko antzekotasuna neurtzeko teknika hau landu zuen corpuseko aditz, subjektu eta objektuen distribuzioari begiratuaz. Aditz baten subjektu/aditza eta objektu/aditza bikote-agerkidetzak estatistikaren arabera neurtu zituen, *co-occurrence score* delakoarekin (*mutual information*-en parekoa)[5]. Honela, izenen arteko antzekotasuna neurtzeaz gain, aditz baten argumentu gisa agertzen diren izenen zerrenda lortzen du agerkidetza altuenetik baxuenera.

1. irudiak, *drink* aditzarekin maizen gertatzen diren objektu/aditz bikoteetako batzuk erakusten ditu, *co-occurrence score*-aren arabera zerrendatuta. Eta hain zuzen ere, objektu/aditz bikote hauexek dira "zer edan daiteke?" galdera erantzuten dutenak.

Hala eta guztiz ere, Hindle-ek lortutako hautapen-murriztapenak oraindik mugatuak dira, ez baita orokortzen. Azken batean, aztertzen ari garen hitzaren ezaugarri lexikoak, hitz-zerrenda batek adieraziko ditu, 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. orr.)

[5]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. orr.).

Synset-zenbakia	*Synset-eko hitzak*	*Definizioa*
05110805	tea	dried leaves of the tea shrub; used to make tea
078546754	tea, camellia sinensis	extensively cultivated in e.g. China and Japan and India; source of tea leaves

2. irudia: *Te* hitzari dagozkion bi *synset*-ak WordNet-en.

Haatik, hurbilpen honek beste bi arazo ekartzen ditu:

1. Hitzean oinarritutako teknikek lortzen dutena hitz-formak dira eta ez hitz-adierak, azken hauek direlarik semantikan hautapenak zehazten dituztenak. Adiera desanbiguazioa, adibidez, ezinezkoa litzateke hitz-formetan bakarrik oinarrituz gero.

2. Lortutako hautapen-murriztapenak corpusean gertatu diren agerpenetara bakarrik mugatuko dira, hau da, corpusetik at dauden antzeko adibideentzako ezingo dira orokortu.

4.3.1. atalean ikusiko dugun bezala, gure proiektuan honen antzeko zerbait erabili dugu, baina ez eskuratze-teknika bezala, baizik eta eskuzko lanerako baliabide bezala.

Klase semantikoan oinarritzen diren eskuratze-teknikak

Teknika hauek klase semantikoak baliatzen dituzte bi hitzen arteko hautapen-murriztapena 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 semantikoa hartzen dute oinarri gisa (Resnik, 1993).

Lan honetan guk horixe egingo dugu: WordNet-et (Miller, 1985; Fellbaum, 1998), edo bere parekoa den EuskalWordNet (Agirre *et al.*, 2002) (euskarako WordNet-a) erabiliko dugu eskuratze-teknika mota hau aplikatzeko[6]. Horretarako, WordNet-en inguruko azalpen txiki bat beharrezkoa da (nahiz eta 3.1.2 atalean sakonago ikusiko dugun). WordNet-en, ingeleseko hitzen adierak kontzeptuetan antolatzen dira, hau da, adiera bera duten hitzek kontzeptu bat osatzen dute, eta kontzeptu hauei *synonym set* (aurrerantzean *synset*) deitzen zaie. Adibidez, ingeleseko *tea* hitzak, WordNet-i jarraituz, bi adiera ditu: bata 'edaria' eta bestea 'landarea'. Beraz, *tea* hitzak bi *synset* ditu WordNet-en (ikus 2. irudia). 'Landarea' adierazi nahi denean, badago *tea* hitzaren sinonimo bat: *camellia sinensis*. Hortaz, bai *te* eta bai *camellia sinensis synset* berean egongo dira, hots, kontzeptu bera adierazteko bi hitz posible egongo dira. *Synset* guztiak elkarrengandik bereizteko, *synset-zenbaki* batekin kodetzen dira.

WordNet hierarkikoki antolatua dago, eta *synset* bakoitzak klase semantiko bat definitzen du: *synset* hori eta bere azpian dauden *synset* guztiak (hiponimo gisa) osatuko dute klase semantikoa. 1. irudiko *wine*, *champagne*, *beer*, *Pepsi*, *tea* eta abar multzokatzen dituen *synset*-ak edarien klase semantikoa osatzen du; esan genezake, *synset* hau eta bere azpian dauden *synset* guztiak [+edangarri] tasuna duen klase semantikoa direla. Honakoan, gainera, klase hori multzokatzen duen *synset*-a hitz baten bidez adierazten

[6]Ikus 3.1.2 atala.

da: *beverage*[7]. Beraz, WordNet-en [+edangarri] kontzeptua *beverage 05074818 synset*-ak adierazten du, eta hona hemen, honen hiponimoetako batzuk:

(3) => **05074818 beverage**
=> 05045640 milk
=> 05076795 alcohol
=> 05081539 wine
=> 05078694 beer
=> 05083066 champagne
=> ...
=> 05102501 soft drink
=> 05103971 Pepsi
=> ...
=> 05108061 juice
=> 05110805 tea
=> ...

Ikus daitekeen bezala, *alcohol 05076795 synset*-ak *wine 05081539*, *champagne 05083066* eta *beer 05078694* multzokatzen ditu, edari alkoholdunen klasea sortuz; *Pepsi 05103971*, aldiz, *soft drink 05102501 synset*-aren azpian dago, freskagarriak diren edarien klasearen azpian[8]. Baina ez *alcohol 05076795 synset*-ak eta ezta *soft drink 05102501 synset*-ak ez dituzte (3)ko edari mota guztiak multzokatzen, eta denak multzokatzen dituena behar dugu: *beverage 05074818*.

WordNet-i buruzko azalpen labur honen ondoren, klase semantikoan oinarritzen diren eskuratze-tekniken aplikazioa zertan datzan hobeto ulertu ahal izango dugu: behin hitz batek (adibidez, *drink* aditzak) corpusean dituen osagai posibleak lortu ondoren (ikus 1. irudia), osagai horiei dagozkien *synset*-ak bilatzen dira WordNet-en, gerora, *synset* horiek guztiak multzokatzen dituen hiperonimo *synset*-a (klase semantikoa) eskuratzeko, eta hortik, aditz horren hautapen-murriztapena lortzeko. Beste hitz batzuetan esanda, *beverage 05074818*ren azpian dauden *synset* guztiak (eta hauei dagozkien hitz guztiekin, noski) ezaugarri semantiko komunak izango dituzte ([+edangarri]) eta ondorioz, agerkidetza sintaktiko bera izango dutela suposatzen da; adibideari jarraituz, guztiak *drink* aditzarekin ager daitezkeela. Honenbestez, [+edangarri] tasuna edo klase semantikoa (*beverage 05074818*) izango da *drink* aditzaren hautapen-murriztapena.

Resnik-ek (1993) teknika hau erabiltzen du, WordNet-en (Miller, 1985; Fellbaum, 1998) hierarkia kontzeptualean eta *association score*[9] neurri estatistikoan oinarrituaz.

Ondorioz, bere hautapen-murriztapenek 3. irudikoen antza dute. Hitzean oinarritzen diren teknikekin ez bezala, klase semantikoa ez da adierazten hitz-zerrenda baten bidez (ikus 1. irudia), baizik eta klase semantiko horren azpian dauden hitz guztiak multzokatzen dituen *synset*-aren bidez: 3. taulan, *beverage*-n bidez, adibidez.

[7]Hizkuntzan kontzeptu batzuk ez daude lexikalizatuak, eta WordNet-en hauek parafrasiaren bitartez moldatzen dituzte. Beraz, *synset* batzuk parafrasiaren bidez adieraziak daude, hots, ez dute hitz zehatzik. Esate baterako, ura duten gauza guztiak (ibaiak, lakuak, putzuak...) multzokatzen dituen *synset*-a *body of water* bezala adierazi dute.

[8]Adibide honetako edarien hierarkia ez dago bere osotasunean. Hierarkia osoa WordNet-en duzue ikusgarri.

[9]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. orr.)

Association score	*verb*	*object classes*
3.58	drink	**beverage** [beverage, drink, drinkable, potable]

3. irudia: *Drink* aditzaren objektuak, kategoria semantikoan oinarritutako teknika erabiliaz (Resnik, 1992).

Klase semantikoan oinarritutako teknikek dituzten abantailak, aurkeztutako beste hurbilpenarekin erkatuz gero, hurrengoak dira:

1. Nahiz eta corpus txikia izan, esanguratsuak izan daitezkeen datu estatistikoak lor daitezke.

2. Corpusean lortutako hautapen-murriztapenek, bertan azaltzen ez diren adibideentzako ere balio dute.

3. Klase semantikoek, ikasitako hautapen-murriztapenen interpretapena errazten dute.

4. Klase semantikoak hierarkikoki antolatuta egoteak, hautapen-murriztapen orokorrak lortzen laguntzen du.

Dena den, eskuratze-teknika mota honek desabantailak ere baditu:

1. Klase semantikoen bidez tasun semantikoak adieraztea ez da beti zuzenekoa, batzuetan ez datoz bat. Adibidez, [+edangarri] tasunak modu egokian adierazten du WordNet-eko *beverage 05074818*ri 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 neketsua da; irekitzen diren gauzen kasuan, WordNet-en *container 01990006 synset*-a klase semantikoa izan daiteke behar bada aproposena.

2. Klase semantikoaren barnean tasun semantiko hori ez duten *synset*-ak ager daitezke. Esaterako, *hegazti* klase semantikoak [+hegan] tasuna eskatzen du. EuskalWordNeten, berriz, *hegazti* hitzari dagokion *synset*-aren azpian hegan egin ezin dutenak ere badaude (*pinguinoa* eta *oiloa*, adibidez). Edota, irekitzen diren gauzen klasea adierazteko *container 01990006* erabiliz gero, honen azpian irekitzen ez diren gauzak ager daitezke; eta alderantziz, irekitzen diren hainbat gauza *container 01990006*k ordez, beste *synset* batek jaso ditzake. Arazo hau adimen artifizialean ezaguna den arren, ez du berehalako ebazpenik. Konponbide posible bat, klase semantikoaren tasun bera daramaten *synset* guztiak zerrendatzea izan daiteke.

2.3 Zertarako dira?

Hautapen-murriztapenak hizkuntzaren arlo ezberdinetan lagungarri izan dira. Atal honetan LNPn eta lexikografian izan duten erabilera aipatuko dugu.

2.3.1 LNP

Hautapen-murriztapenen ezagutzari esker LNPren ataza askotan aurrera jo daiteke.

Egitura sintaktikoen desanbiguazioa

Egitura-anbiguotasuna konpontzeko baliagarria den informazio lexikoa eskaintzen du. Adibidez, postposizio-sintagmekin gertatzen den anbiguotasuna ekiditeko erabili izan da:

(4) Zapatos de piel de señora.

(5) Emakumeen larruzko zapatak.

Kasu hauetan, *de señora* edo bere euskarako itzulpena *emakumeen* postposizio-sintagmek zer izen modifikatzen duten zehaztu beharko genuke: *zapatos/zapatak* ala *piel/larru*, honen arabera esaldiak esanahi desberdina izango baitu:

(4') Zapatos [de piel] [de señora] vs Zapatos [de piel [de señora]]

(5') [Emakumeen] [larruzko] zapatak vs [[Emakumeen]larruzko] zapatak

Zapatak zerez eginda daude, larruarekin ala emakumeen larruarekin? Irakurketa arruntena lehenengoa da, hots, *de señora/emakumeen* postposizio-sintagmak *zapatos/zapatak* izena modifikatzen duena. *Zapato/Zapata* hitzak dituen hautapen-murriztapenen artean bai *de piel/larruzko* eta bai *de mujer/emakumeen* egongo dira, baina *de piel de señora/emakumeen larruzko* bezalakorik ez da inoiz agertuko. Beraz, hautapen-murriztapenek honen inguruan badute zeresanik.

Haatik, (4) eta (5) bezalako adibideekin ikus dezakegu hizkuntza bateko hautapen-murriztapenak beste hizkuntzetakoen antzekoak edo berdinak izan daitezkeela, eleanitzak alegia. Horixe izango da lan honen hipotesi nagusia eta hori dela eta, suposatuko dugu hizkuntza baterako lortu dugun hautapen-murriztapena, beste hizkuntza baterako erabilgarria izan daitekeela.

Hitzen adieren desanbiguazioa

Hautapen-murriztapenek hitzen adieren desanbiguazioan ere eragina badute (Ribas, 1995).

(6) Play the piano.

(7) Play football.

Adibideetan ikus daitekeen bezala, *play* aditzak adiera bat baino gehiago izan dezake, eta hautapen-murriztapena lagungarria izan daiteke aditzaren adiera egokia erabakitzeko. *Play* kirola adierazteko erabiltzen denean, objektu gisa kirolarekin zerikusia duen ekintza bat du (*football, baseball, golf...*). Objektuak musika-instrumentuak direnean (*piano, guitar, violin...*), aldiz, *play* aditzak 'musika-instrumentua jo' adiera du. Beraz, *play* aditzak [+kirol-ekintza] hautapen-murriztapena duenean, euskarako 'jokatu' adieraziko luke (kirol domeinukoa); [+musika-instrumentua] hautapenarekin, aldiz, *play* aditza euskarako 'jo' (musika-instrumentua) aditzaren parekoa litzateke.

Anafora

Anaforaren fenomenoari ere irtenbidea eman diezaiokegu, hautapen-murriztapenak erabiliaz:[10]

(8) The packet contained chocolate but nobody was allowed to **open** it.

(9) The packet contained chocolate but nobody was allowed to **eat** it.

Bai (8)n eta bai (9)n, ez dugu inolako dudarik *it*-ek zeri egiten dion erreferentzia; (8)n *packet*-i eta (9)n *chocolate*-ri. Ezagutza aditzetik datorkigu, eta baita honek normalean eskatzen duen argumentu motatik, hautapen-murriztapenetik, alegia. Hau da, normalki *paketeak* irekitzen diren gauzen artean egongo dira, eta *txokolatea* jaten den janarien artean.

Halere, egitura eta hitzen desanbiguazioarekin ez bezala, anafora automatikoki lantzea oso zaila da, corpus etiketatu baten beharra baitago, besteak beste.

2.3.2 Lexikografia

Hautapen-murriztapenek hitzen adiera eta agerkidetzari buruzko informazioa ematen dutenez, lexikografiarekin harremanetan dauden atazetan ere lagungarri izan dira. Hala nola, informazio hau guztia hiztegi-sarreretan gehitu daiteke, horrela, erabiltzaileak hitz-adiera hori nola erabili behar duen jakin baitezake.

Hurbilpen hau lexikografian *defining in context* bezala ezagutzen da, eta ikasleentzako hiztegietan arrakasta handiz erabili izan da (Hornby, 1991).

Azken finean, hautapen-murriztapenak hitzaren beste ezaugarri linguistikoak esplikatzeko aukera ematen du, eta, ondorioz, semantikarekin lotura duen edozein arlori interesa dakioke.

[10] Adibidea McCarthy (2001) lanetik hartua da.

3 ESPERIMENTUAREN INGURUAN

Sarreran (1.1) aipatu dugun bezala, ikerlan honen helburu nagusiena hurrengoa da: corpus eta eskuratze-teknika desberdinak erabiliaz, ingeleseko kirol aditz batzuentzat automatikoki ikasitako hautapen-murriztapenak aztertzea, gero hauek euskararentzat baliagarriak izan daitezkeen ikusi ahal izateko. Beste era batera esanda, ingeleseko aditzentzat automatikoki eskuratutako hautapen-murriztapenak, euskarako itzulpenekin berrerabili daitezkeen egiaztatu nahi dugu. Horrela, ikerlan honetan ondorengo eginkizun nagusiak beteko ditugu:

(a) **Ingeleseko aditz batzuen hautapen-murriztapenenak lortzeko erabili diren eskuratze-teknika automatikoen emaitzak hartuta, hauen azterketa eta ebaluazioa egin teknika bakoitzaren alderdi on eta txarrak aipatuaz.** Beste hitz batzuekin esanda, hautapen-murriztapenen eskuratze-teknika desberdinen ebaluazio bat egin dugu. Honetarako, bi parametro hartu ditugu kontuan: **adiera eta domeinua**. Hasieran hautapen-murriztapenak aditzen adierentzat pentsatuak definitu baziren ere (Wilks, 1973), lehenengo ahalegin automatikoetan aditz formetara mugatu behar izan zuten (Resnik, 1993). Geroago, aditzen adierak kontuan hartzen dituzten eskurapen-teknikak proposatu dira (Agirre eta Martínez, 2002; McCarthy, 2001). Eta gaur egun, hautapen-murriztapenen ikasketa domeinu zehatz bati buruz aritzen diren corpusetara mugatzen hasi dira, aditzaren adiera eta bere hautapen-murriztapenena corpusaren domeinutik lortu daitekeelarik (Agirre *et al.*, 2003; McCarthy, 2001). Esperimentu honetan ere, bi corpus mota erabili ditugu: kirol domeinuarekin harremanetan daudenak eta domeinu zehatzik ez dutenak. Hauetatik lortutako hautapen-murriztapenak corpus orekatuetakoekin parekatzea interesgarria iruditu zaigu. Adierari dagokionez, eskuratze-teknika batzuk aditzaren hautapen-murriztapenak ikasten dituzte aditz-adiera kontuan izanda, eta beste batzuk aldiz, aditz-forman oinarritzen dira. Eskuratze-teknika hauen arteko aldean ere sakonduko dugu.

(b) **Ingeleseko aditzentzat eskuratze-teknika bakoitzetik lorturiko hautapen-murriztapenak euskarako ordainen hautapen-murriztapenak izan daitezkeen frogatzea**, bi hizkuntzetarako egokiak diren ala ez egiaztatzeko, hots, hautapen-murriztapenak eleanitzak izan daitezkeen ala ez aztertzeko. Beraz, ingeleserako lortu diren datuak euskaraz berrerabiliko ditugu eta euskaraz erabilgarriak diren ala ez ikusi. Honetarako, EuroWordNet-az baliatu gara, bertan ingeleseko ordain bakoitza euskarakoarekin lotua baitator. Noski, ikusiko dugun bezala, hizkuntzen arteko konparaketa honek zenbait fenomeno linguistiko agerian utziko ditu.

(c) **Ingeleserako erabilitako eskuratze-teknika batzuk euskarako corpus batean erabili (a) eta (b)ko emaitzekin erkatzeko.** Ingeleseko corpusetik lortutako hautapen-murriztapenak eta euskarako corpusetik lortutako hautapen-murriztapenekin konparatzea, alegia. Hemen ere, kirol domeinuari dagozkion corpusak eta corpus orekatuak erabili ditugu, beraien artean zer desberdintasun agertzen diren aztertzeko.

Zer aditz aukeratu ditugu esperimentu hau aurrera eramateko? Domeinu bereko aditzak aukeratu ditugu, gure ustez **kirol domeinuko** aditz esanguratsuenak, hain zuzen ere: *jokatu, galdu, irabazi, entrenatu* eta *berdindu* eta hauen ordainak ingelesez[1].

Txosten honetan *jokatu* aditza erabiliko dugu esperimentuaren metodologia eta garapena pausoz pauso azaltzeko, baina aipatutako aditz guztiekin egin dugu azterlan bera.

Hurrengo ataletan esperimentu hau egiteko beharrezkoak izan diren baliabide (3.1 atala) eta eskuratze-teknikez (3.2 atala) jardungo gara.

3.1 Esperimenturako erabili diren baliabideak

3.1.1 Corpusak

Hautapen-murriztapenak ondorengo corpusetatik lortu ditugu:

Ingeleseko corpusak

- SemCor: Ingeleseko corpus hau (Fellbaum *et al.*, 2001) eskuz etiketatutako corpusik handiena da. Brown Corpus-aren zati batez eta Stephen Craig-en *The Red Badge of Courage* eleberriaz osatuta dago eta 350.000 hitzen inguru ditu. Corpuseko hitz bakoitza WordNet-eko *synset* bati dagokio, eta arrazoi honengatik LNPn oso erabilia izan da.

- The British National Corpus (BNC): BNC 100 milioi hitzetako corpus orekatua da, hots, iturri ezberdinetako corpusekin osatutakoa.

- 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.

Euskarako corpusa

- Euskaldunon Egunkaria: Egunkari honetako berriekin osatutako corpusa da, 7 milioi hitz inguru dituena. EFEren antzera, corpus domeinuka antolatuta dago, hau da, kazetaritzaren domeinu bakoitzak bere corpusa du. Hala, euskarako hitz baten testuingurua corpus osoan zehar ala domeinu zehatz batean kontsulta daiteke.

[1] 3.2 atalean ikusiko dugun bezala, ingeleseko ordainak lortzeko EuroWordNet erabili dugu.

3.1.2 WordNet, EuroWordNet eta EuskalWordNet

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

Ingeleseko izen, aditz, adjektibo eta adberbioak sinonimo multzotan (*synonym set* edo ***synset*** deiturikoetan) antolatuak daude, hauetako bakoitza kontzeptu lexikal bati dagokiolarik.

Esaterako, ingeleseko *tree* izenak WordNet-en bi *synset* (adiera)[2] ditu:

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

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

Hortaz, WordNet-eko erlazio semantiko garrantzitsu bat **sinonimia** da; ezagutza-basearen oinarria hitzaren adieran dago, eta adiera hori hitz batek baino gehiago adierazten dutenean, hitzak multzokatu egiten dira.

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

Hiperonimia-hiponimia erlazioak *synset* orokorrenak *synset* zehatzagoekin lotzen ditu[3], eta, 2.2.2. atalean aipatu dugun bezala, hiperonimia-hiponimia hierarkia honek izen klaseak esplizitu adierazten ditu, hots, hiponimoak onartzen dituen *synset*-ak klase bat osatzen du. (11) eta (12) adibideetan (10)ren hiperonimoak (kontzeptu orokorragoak) eta hiponimoak (kontzeptu zehatzagoak) ikus ditzakegu hurrenez hurren[4]:

(11) Sense 1
tree (a tall perennial woody plant having a main trunk and branches...)
=> woody plant, ligneous plant – (a plant having hard lignified tissues...)
=> vascular plant, tracheophyte – (green plant having a vascular system...)
=> plant, flora, plant life – (a living organism lacking the power of locomotion)
=> life form, organism, being, living thing – (any living entity)
=> entity, something – (anything having existence (living or nonliving))
Sense 2
tree , tree diagram (a figure that branches from a single root; "genealogical tree")
=> plane figure, two-dimensional figure (a 2-dimensional shape)
=> figure (a combination of points and lines and planes that form a visible palpable shape)
=> shape, form (the spatial arrangement of something as distinct from its substance)
=> attribute (an abstraction belonging to or characteristic of an entity)
=> abstraction (a general concept formed by extracting common features...)

[2]Aurrerantzean *synset* terminoa erabiliko dut.

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

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

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

WordNet-eko aditzen hierarkia ez da izenena bezalakoa (hiperonimia-hiponimia edo *x is a kind of y* motakoa); aditzen hierarkia hiperonimia-**troponimia** motakoa da (*to x is to y in some particular manner*). Hortaz, aditz hiperonimo baten (*walk*, esaterako) troponimoak aditz hiperonimoak adierazten duen hori egiteko moduak izango dira (*trot*, *march*...).

(13) Sense 1
walk (use one's feet to advance; advance by steps; "Walk, don't run")
=> tramp down, trample, tread down – (walk on and flatten; "ramp down the grass")
=> lollop – (walk clumsily and with a bounce)
=> tap – (walk with a tapping sound)
=> stumble, falter, bumble – (walk unsteadily; "The drunk man stumbled about")
=> toe – (walk so that the toes assume an indicated position or direction)
=> traipse, shlep – (walk or tramp about)
=> perambulate, walk about, walk around – (walk with no particular goal)
=> sneak, mouse, creep – (to go stealthily or furtively)
=> ...

WordNet-en, hierarkia eta *synset*-ez gain, eremu semantikoak jasotzen dituzten fitxategi batzuk daude[5]. Fitxategi hauetan testuinguru antzekoetan gertatzen diren kategoria sintaktiko bereko *synset* zerrendak daude, hots, eremu semantiko berdinari (*food, person, plant, body, communication*, eta abar) dagozkienak. Esate baterako, *play* aditzak kirola adierazten duenean (*play football* diogunean adibidez), *synset* hori *competition* eremu semantikoaren fitxategian dago; *play Hamlet* esan nahi dugunean, ordea, adiera horri dagokion *synset*-a *creation* domeinuaren fitxategian dago.

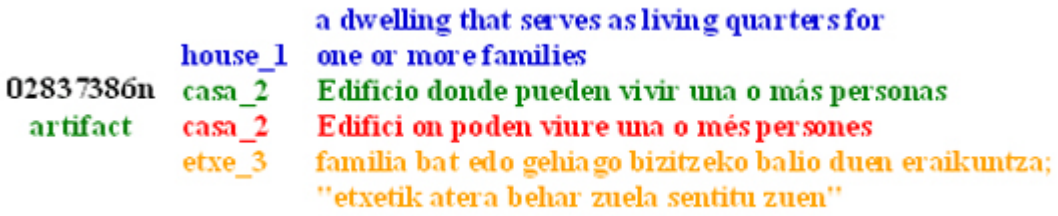
WordNet ugari eraiki dira hainbat hizkuntzatarako (ingelesa, daniera, italiera, gaztelania, alemana, frantsesa, txekiera eta estoniera), eta EuroWordNet (Vossen, 1998) ereduak hizkuntza guztietako kontzeptuak lotzea ahalbidetzen du. Proiektu horren barruan EuskalWordNet (Agirre *et al.*, 2002)) eraikitzen ari gara Donostiako Informatika Fakultateko IXA ikerkuntza taldearen barruan. Egun, EuskalWordNet-ek ondoko taulan zehazten den hitz eta kontzeptu kopurua du, baina etengabe ari dira hazten (ikus 4. irudia).

Nahiz eta EuroWordNet-en hizkuntza bakoitza WordNet "independente" bat izan, EuroWordNet-en helburua WordNet desberdin hauek guztiak ezagutza-base eleanitz bakarrean elkartzea da, eta horri esker, EuskalWordNet WordNet-a duten gainontzeko hizkuntzetara lotuta egon daiteke[6]. 5. irudian ikus daitekeen bezala, *synset* berean ingeleseko, gaztelaniako, katalaneko eta euskarako ordainak ditugu.

[5] WordNet-en *lexicographer's files* bezala izendatzen dituzte.
[6] Ikus http://siuc02.si.ehu.es/wei2004-06-21/wei.html EuskalWordNet-eko edukiak arakatzeko.

	HITZAK	SYNSET-AK	ADIERAK
IZENAK	22146	27649	48214
ADITZAK	3155	3240	9295

4. irudia: EuskalWordNet-ek egun dituen *synset* eta hitzen kopurua.



5. irudia: *Etxe* izenaren *synset* bat eta bere ordainak EuroWordNet-eko interfazean.

WordNet-en egitura, harreman eta *synset*-etan oinarritu arren, WordNet-ek ez zituen ezaugarri batzuk EuroWordNet-en gehitu dira. Esaterako, eremu semantikoak aparteko fitxategietan egon ordez, interfazean bertan ikusgarri daude. 5. irudiko *synset*-ak ezkerraldean *artifact* marka darama, eta horixe da bere eremu semantikoa[7]:

EuroWordNet WordNet-en garapen bat den bezala, Multilingual Central Repository (MCR) ezagutza-basea EuroWordNet-en bertsio aurreratuago bat da, "MEANING: Developing Multilingual Web-Scale Language Technologies" (IST-2001-34460) proiektuarekin garatua (Atserias *et al.*, 2004). MCR WordNet eta EuroWordNet-en informazioaz baliatzen da, eta honetaz gain, informazio berria dakar. Esaterako MCRn, WordNet domeinuen hierarkia batekin aberastu dago (Magnini eta Cavagli, 2000). Honek *synset*-ak domeinu edo gaien arabera antolatzen ditu: *sports*, *restaurant*, *traffic*, eta abar. Esperimentu honetan MCRko domeinu hierarkia erabili dugu. 5. irudiko *synset*-aren eremu semantikoa *artifact* dela esan dugu, eta bere domeinua (MCRren arabera) *town planning* da. Ikus daitekeen bezala eremu semantikoak orokorragoak dira.

Azken urte hauetan hautapen-murriztapenen inguruan egindako lanari begiratuta, eta gure ikerkuntza taldean bertan ditugun tresnei erreparatuz, hautapen-murriztapenen jabekuntza EuroWordNet-en oinarritzea erabaki dugu, hau da, ikusiko dugun bezala, aukeratutako aditzak eta hauen hautapen-murriztapenak *synset* edo domeinu-eremu semantikoaren informazioa daraman bikote baten bidez adierazita etorriko zaizkigu. Eta hautapen-murriztapenen eskuratzea EuroWordNet-etik aukeratutako zortzi *synset*-entzat egingo dugu[8]:

[7] Aldaketa hauetatik hautapen-murriztapenen azterketa honi begira, guri interesatzen zaizkigunak aipatuko ditugu. Argibide gehiago txosten honekin batera entregatutako beste ikerlanean (Pociello, 2004) eta Vossen-en lanean (1998).

[8] Esperimentu honetan ingeleseko eta euskarako ordainak erabiliko ditugunez, hauek bakarrik zehaztu ditugu.

1) *00605818 play1 / jokatu2; "play games, play sports"*

2) *00610422 encounter5, meet10, play24, take on5 / jokatu3; "contend against an opponent in a sport or game"*

3) *00468052 coach2, train7 / entrenatu1; "teach and supervise, as in sports or acting"*

4) *00059698 train8 / entrenatu3; "exercise in order to prepare for an event or competition"*

5) *00630097 equalize1, get even1 / berdindu16; "compensate; make the score equal"*

6) *00630097 draw25, tie2 / berdindu15; "finish a game with an equal number of points, goals. . . "*

7) *00620486 win1 / irabazi3; "be the winner in a contest or competition; be victorious"*

8) *00620218 lose2 / galdu9; "fail to win"*

Nahiz eta ikerlana hauekin guztiekin egin den, lan honen alderdi metodologikoa pausoz pauso azaltzeko (4. eta 5. atalak) *synset* batean oinarritu gara (*00605818 play1 / jokatu2* "play games, play sports"). Haatik, txosten honetan azaltzen diren argibide eta ondorioak aditz guztien emaitzetatik abiatuta lortutakoak dira, eranskinetan guztien emaitzak ditugularik.

3.2 Esperimenturako erabili diren eskuratze-teknikak

Ikerlan honetan klase semantikoan oinarritzen diren eskuratze-teknikak erabili dira (ikus 2.2.2. atala) eta EuroWordNet baliatu dugu klase semantiko horiek adierazteko. Horrela bada, eskuratze-teknika hauek aditzen objektu/subjektuen hautapen-murriztapenak adierazteko EuroWordNet-eko klase semantikoak darabiltzate. Hala ere, teknika honen barruan aldaerak egon daitezke, gu lau eskurapen-teknika ezberdinez jardungo gara, bi multzo nagusitan banatu ditugunak hauen azalpena ulergarriagoa egin ahal izateko:

(a) *Synset* batekin adierazitako hautapen-murriztapenak.

(b) Domeinu-eremu semantiko bikote batekin adierazitako hautapen-murriztapenak.

3.2.1 *Synset* batekin adierazitako hautapen-murriztapenak

Mota honetako eskuratze-teknikek aditz baten hautapen-murriztapenak ***synset*** batez adierazten dute, *synset* hau klase bezala kontsideratzen dutelarik, hau da, *synset*-a bera eta honen hiponimo guztiak izango dira aditz horren objektu/subjektuen hautapen-murriztapena.

Objektu/Subjektuen hautapen-murriztapena klase gisa kontsideratua izango den bezala, aditzari dagokionez, ikuspuntu ezberdinetik landu daiteke, eta hori izango da sailkapen honetan eskuratze-teknikak ezberdinduko dituena.

Aditzaren hautapen-murriztapenak eskuratzean, hautapen-murriztapen 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 hautapen-murriztapenak eskuratzean aditz horrek izan ditzakeen adiera guztietan oinarritu dela, aditz horren edozein adierari dagokion hautapen-murriztapena ikas dezake ([+lehiaketa] edo [+jabetza] esate baterako).

Hautapen-murriztapenak aditzaren adiera bakarrarentzat ere lor daitezke, **aditz-adierarentzat**, alegia. Adibidez, *irabazi* aditzaren objektu hautapen-murriztapenak eskuratzerakoan, eskuratze-teknikak aditz-forma honen adiera bakarra hartuko du kontuan (adibidez, *lehiaketa irabazi* kirol adiera). Hala, eskuratze-teknika honek adiera horri bakarrik dagozkion objektuen hautapen-murriztapenak ikasiko ditu: [+lehiaketa], [+kirola], eta abar.

Aditz-forman oinarritzen den eskuratze-teknikari ***word-to-class*** (w2c) deritzo, eta aditz-adieran oinarritzen denari ***class-to-class*** (aurrerantzean c2c)[9]. Izenak adierazten duen bezala, w2c teknikak hitzetik abiatuta (aditz-formatik) klaseak diren hautapen-murriztapenak lortzen ditu; c2c-ek, aldiz, aditz-klase batetik abiatuta klaseak diren hautapen-murriztapenak lortzen ditu.

Hautapen-murriztapenak adierazteko *synset*-a darabiltzaten eskuratze-teknika hauen ezberdintasun nagusiena azaldu ondoren, hautapen-murriztapen hauek eskuratzeko jarraitzen diren irizpideak aipatuko ditugu. Nahiz eta w2c eta c2c-en ikasketa 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.

Class-to-class (c2c)

Hautapen-murriztapen mota hau zertan datzan ulertu ahal izateko, lehendabizi nola lortzen den ulertzea garrantzitsua da.

Aditz baten c2c hautapen-murriztapenak eskuratzeko, lehenengo corpusaren gainean Minipar analizatzailea edo *parser*-a (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 hurrengo agerpena, (14) adibideko hirukotean bilakatuko luke:

(14) Futbol-taldeak irabazi zuen.

(15) [Futbol-talde (IZENA), Subjektua (ERLAZIO SINTAKTIKOA), Irabazi (ADITZA)]

Ondoren, hirukote bakoitzean dauden izenak EuroWordNet-en kontsultatzen dira, horrela, aditza bera, eta aditz horrekin agertu den izen bakoitzaren adiera (bere *synset*-zenbakiarekin) desanbiguatua izan da automatikoki. Hortaz, orain hirukotea [IZENA eta bere SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKOA, ADITZA eta bere SYNSET-ZENBAKIA] motakoa izango da.

(16) [Futbol-talde/05167683 (IZENA/SYNSET-ZENBAKIA), Subjektua (ERLAZIO SINTAKTIKOA), Irabazi/00620486 (ADITZA/SYNSET-ZENBAKIA)]

SemCor corpusaren gainean ari bagara, hirukote hau corpusetik zuzenean datorkigu, corpusa bera EuroWordNet-eko *synset*-zenbakiekin eskuz etiketatua baitago.

Azkenik, hirukote bakoitzaren probabilitate kopurua kalkulatu egiten da corpusean duten maiztasunaren arabera[10]. Hirukoteak daraman kopuru hau 1 zenbakitik geroz eta

[9] Eskuratze-tekniken terminologia ingelesez mantendu dugu, hizkuntzalaritza konputazionalean horrela ezagutzen direlako. Hala ere, hauek euskaraz ***hitza-klase*** eta ***klase-klase*** bezala izenda daitezke.

[10] Argibide gehiago (Agirre eta Martínez, 2002, 2001).

gertuago egon, orduan eta ziurrago egon gaitezke hirukoteak adierazten duen harremana egokia dela.

Beraz, [IZENA/SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKOA, ADITZA/SYNSET-ZENBAKIA] motako hirukoteak dauzkagu, ondoan hautapen-murriztapen egokitasuna markatzen duen probabilitate kopuru batekin. (17) adibidean (14), (15) eta (16)ko hirukote bera dakargu, baina probabilitatea gehituta eta prozesuaren ondorioz ikus ahal izango dugun itxurarekin[11]:

(17) **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, synset-eko sinonimoak eta definizioa)

Izen mota hauek **izen klaseen** bidez datoz adierazita. Eskuratze-eredu honek corpusetik jasotzen dituen objektu/subjektuen izenak EuroWordNet kontsultatzen ditu, gerora izen horiek guztiak multzokatzen dituen klase semantikoa aukeratzeko, normalean hauen hiperonimo bat. Horrela, corpuseko izen hori orokor dezaken beste izen bat lortzen da, aditz batekin joan daitekeen izen multzo bat mugatzen duena, hain zuzen ere.

(14) adibidearekin jarraituz, ezin da ukatu *futbol-talde* izena *irabazi* aditzaren subjektua izan daitekeela, baina era berean esan dezakegu:

(18) Saskibaloi-taldeak irabazi zuen.

(19) Errealak irabazi zuen.

Esandakoaren arabera, (17) ez da eskuratze-prozesuaren azken emaitza, *futbol-talde* izenaren ordez, hau orokortzen duen hiperonimo bat agertuko zaigulako[12]:

(20) **c2c.subj**
irabazi 00620486
04771851 0.101 0.145 gizatalde "Mota bereko izaki bizidunen multzoa"

Hautapen-murriztapen 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 hautapen-murriztapenak izen klaseak izango dira.

Bestalde, eskurapen-teknika honek aditzaren adiera ere kontuan hartzen du. c2c eskuratze-teknikak lortzen dituen hautapen-murriztapenak aditzaren adiera jakin baterako dira. Beraz, EuroWordNet kontsultatzean *irabazi* aditzari 00620486 *synset*-zenbakia egokitu bazaio ("Lehiaketa baten irabazlea izan"), automatikoki eskuratutako hautapen-murriztapenak *irabazi* aditzaren adiera horrentzat bakarrik izango dira, eta inolaz ere aditzaren beste adierentzat.

[11] Azalpena ulergarriagoa izan dadin, 3.2.1 atal honetako hautapen-murriztapenen adibide, glosa eta *synset* asmatuak euskaraz jarri ditugu. Hala ere, hurrengo ataletan ingelesez aurkeztuko ditugu, esperimentuan eskuratze-tekniken emaitza guztiak ingelesez daudelako eta hauek itzultzea lan handia litzatekeelako.

[12] Kontuan izan beharrekoa da EuroWordNet hierarkia bat dela eta batzuetan ez dela horren erraza hautapen-murriztapena adierazten duen *synset* "egokia" aukeratzea, gerta litekeelako *synset* hori orokorregia izatea (hierarkian goregi egotea) eta zehatzegia izatea (hierarkian beheregi egotea). EuroWordNet-eko hiperonimo zehatzegi bat proposatuz gero, aditz baten objektu/subjektuen aukeraketa gehiago (eta batzuetan gehiegi) mugatuko genukeen; eta alderantziz, hiperonimoa orokor bat proposatuz gero (batzuetan orokorregia) aukera gehiegi zabal daiteke eta zuzenak ez diren hautapen-murriztapenak ere agertuko lirateke. Hau guztia datorren ataletan zehazkiago ikusiko dugu.

(21) adibidean *irabazi* aditzaren hautapen-murriztapenen etsenplu bat dugu, 00620486 *synset*-ari dagokion adierarekin, hots, kirol adierarekin ("Lehiaketa baten irabazlea izan").

(21) **c2c.obj**
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 hautapen-murriztapenak baliagarriak dira aditz horrentzat, bere *synset*-ean dituen sinonimo guztientzat, eta bere troponimoentzat. (20)ren kasuan, hautapen-murriztapen horiek *irabazi 0060486 synset*-ari eta honen azpian dauden beste *synset* guztiei dagozkie. Horrela bada, eskuratze-teknika honen hautapen-murriztapenak aditz-klase oso bati dagozkie. SemCor etiketatutako corpus bat izaki, eskuratze-teknika honek corpusean *irabazi 0060486 synset*-aren troponimo bat agertuko balitz, bere hiperonimoarekin (*irabazi 0060486*) erlazionatzeko gai izango da, eta klase guztiari hautapen-murriztapen berdinak egokitzen dizkio[13].

Azkenik, aipatu beharra dago, eskuratze-teknika honekin (eta besteekin) ez dela aditz bakoitzarentzat hautapen-murriztapen bakarra lortzen, aditz bakoitzak probabilitate kopuru altuenetik baxuenera ordenaturiko hautapen-murriztapen zerrenda bat izango baitu. Horrela, aditz baten objektu/subjektu argumentu gisa agertzen diren izenen zerrenda dugu probabilitate altuenetik baxuenera. Zerrenda hau oso luzea izan daiteke, eta hamar hautapen-murriztapen baino gehiagok osatzen dutenean lehenengo hamarretara bakarrik mugatzen gara lan honetan. Irizpide hau esperimentu honetako eskuratze-teknika guztiekin erabili dugu.

Word-to-class (w2c)

Eskurapen-teknika honen prozesua aurrekoaren oso antzekoa da. Lehenik, Minipar *parser*-aren bitartez [IZENA, (izena eta aditzaren arteko) ERLAZIO SINTAKTIKOA, ADITZA] hirukote modukoak ateratzen dira eta c2c eskuratze-teknikarekin bezala, bigarren pausoa EuroWordNet-en kontsulta egitea da, baina oraingo honetan, hirukoteko izenak bakarrik begiratzen dira EuroWordNet-en. Hala, izen horiek dagokien adiera edo *synset*-zenbakiarekin desanbiguatuta izango ditugu. Beraz, orain hirukotea [IZENA/SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKOA, ADITZA] motakoa izango da.

(22) [Futbol-talde (IZENA), Subjektua (ERLAZIO SINTAKTIKOA), Irabazi (ADITZA)]

SemCor corpusaren gainean ari bagara, [IZENA/SYNSET-ZENBAKIA, ERLAZIO SINTAKTIKO, ADITZA] hirukotea corpusetik zuzenean datorkigu, corpusa bera WordNet-eko *synset*-zenbakiekin eskuz etiketatua baitago.

Azkenik, hirukote bakoitzaren probabilitate kopurua kalkulatu egiten da corpusean duten maiztasunaren arabera[14]. Hirukoteak daraman kopuru hau 1 zenbakitik geroz eta gertuago egon, orduan eta ziurrago egon gaitezke hirukoteak adierazten duen harremana egokia dela. Horrela bada, hautapen-murriztapen hauek duten itxura aurrekoaren oso antzekoa da:

[13] Honen berri hurrengo atalean emango dugu.
[14] Argibide gehiago (Agirre eta Martínez, 2002, 2001).

(23) **w2c.subj** (*eskuratze-teknika eta erlazio sintaktikoa*)
irabazi (*aditza*)
05167683 0.070 futbol-talde "Futbolean jokatzen duen taldea"
(synset-zenbakia, probabilitatea, synset-eko sinonimoak eta definizioa)

Eskurapen-teknika hau c2c ereduaz desberdintzen da aditzaren adiera kontuan hartzen ez duelako. Eredu honekin **aditz-formaren** (hitzak izan ditzakeen adiera guztiak kontuan hartuta) objektu edo subjektuen hautapen-murriztapenak lortzen dira (c2c teknikan objektu edo subjektuen hautapen-murriztapenak lortzen ziren bezalaxe).

(24) adibidean *irabazi* aditzaren objektuaren zenbait hautapen-murriztapena ditugu w2c ereduaren arabera, hau da, aditz horren objektu gisa agertu diren izenen zerrenda dugu probabilitate altuenetik baxuenera ordenaturik:

(24) **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 harreman eta eskubidea"

Hautapen-murriztapen hauek *irabazi* aditzak objektu gisa har ditzakeen izen klase batzuk dira, beheko izenetatik abiatuta lortutako hiperonimoak:

(25) **partidua** irabazi (hiperonimoa: *lehiaketa*)

(26) **futbolean** irabazi (hiperonimoa: *talde-ekintza*)

(27) **dirua** irabazi (hiperonimoa: *jabego*)

Ikus daitekeen bezala, w2c eskuratze-teknika honek eskaintzen dituen hautapen-murriztapenak aditz-forma osoarentzat dira, hots, aditzaren adiera guztiei erreparatzen diotenak. Honela bada, (25) eta (26)ko perpausak *irabazi* aditzaren kirol adierari dagozkie eta (27)koa, aldiz, finantza adierari.

3.2.2 Domeinu eta eremu semantiko batekin adierazitako hautapen-murriztapenak

Hautapen-murriztapen hauek aurrekotik ezberdintzen dira. Mota honetako eskuratze-teknikek aditz baten hautapen-murriztapenak domeinu-eremu semantiko bikote batez adierazten dute, bikote hau klase bezala kontsideratzen dutelarik, hau da, bai domeinua eta bai eremu semantiko hori duten izen guztiak izango dira aditz horren objektu/subjektuen hautapen-murriztapena[15].

Synset batekin adierazitako hautapen-murriztapenen barruan w2c eta c2c eskuratze-teknikekin gertatzen zen bezala, hemen ere eskuratze-teknikak ezberdintzen dira hautapen-murriztapenak aditz-formatik edo aditz-adieratik abiatuta ikasten direlako.

Aditzaren hautapen-murriztapenak eskuratzean, hautapen-murriztapen hauek aditzaren adiera guztientzako izan badaitezke, (**aditz-formarentzat**, alegia) ***word-to-semantic-field*** (aurrerantzean w2semf)[16] eskuratze-teknikaz baliatu gara, hots, hitzetik abiatuta domeinu-eremu semantiko bikoteak lortzen dituenaz.

[15] Domeinua eta eremu semantikoei buruzko azalpenak 3.1.2 puntuan eman ditugu. Ezagutza-basean ingelesez daudenez, hala mantendu ditugu.

[16] 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*** bezala izenda daitezke.

Hautapen-murriztapenak aditzaren adiera bakarrarentzat ere lor badaitezke (**aditz-adierarentzat**, alegia), orduan, ***sense-to-semantic-field*** (aurrerantzean s2semf) eskuratze-teknikaz baliatu garela esango dugu, hau da, aditz-adieratik[17] abiatuta domeinu-eremu semantiko bikoteak lortzen dituenaz.

Har ditzagun, berriro ere, *irabazi* aditza eta (25), (26) eta (27) adibideak. Aditz honen objektu w2semf hautapen-murriztapenak aditzaren adiera guztientzat lirateke.

(28) **w2semf.obj** (*eskuratze-teknika eta erlazio sintaktikoa*)
irabazi(*aditza*)
obj economy-possession 33
obj sport-event 28
(erlazio sintaktikoa, domeinu-eremu semantiko bikotea eta probabilitatea)

(28)ko hautapen-murriztapenak (24)koen berdinak dira baina lehenengoak domeinu-eremu semantiko bikoteekin adieraziak, eta bigarrenak *synset*-ekin adieraziak. Formatu aldetik, ordea, bi hurbilpen hauek ezberdinak direnez, (28)n w2semf eskuratze-teknikaren emaitzak dakarren informazioa adierazi dugu.

Aditz horren kirol adieran oinarrituz gero (*irabazi 00620486*), s2semf eskuratze-teknikak aditz horren kirol domeinuarekin harremanetan dauden objektuen hautapen-murriztapenak bakarrik ikasiko lituzke:

(29) **s2semf.obj**(*eskuratze-teknika eta erlazio sintaktikoa*)
irabazi 00620486(*aditza eta bere synset-zenbakia*)
obj play-act 33
obj sport-event 28
(erlazio sintaktikoa, domeinu-eremu semantiko bikotea eta probabilitatea)

(29)ko hautapen-murriztapenak (21)ekoen berdinak dira baina lehenengoak domeinu-eremu semantiko bikoteekin adieraziak, eta bigarrenak *synset*-ekin adieraziak. Oraingoan ere, s2semf eskuratze-teknikaren emaitzek duten formatuaren berri adibidean azaldu dugu.

Atal honen hasieran esan bezala, bikote hauek klaseak dira: *sport* domeinua eta *event* 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 eta s2semf-ekin ere. Lehenengo, corpusaren gainean Minipar analizatzailea edo *parser*-a (Lin, 1993) erabili behar da, aditz horren corpuseko agerpen bakoitza [IZENA, (izena eta aditzaren arteko) ERLAZIO SINTAKTIKOA, ADITZA] hirukote modukoetan adierazteko. Ondoren, hirukote bakoitzean dauden izenen EuroWordNet-eko eremu semantikoak eta

[17] c2c eta s2semf ezberdintzen dira, aditzaren izaeran. Lehenengoak aditzaren *synset*-eko sinonimoak eta troponimoak kontuan hartzen ditu; eta bigarrenak, aditzaren *synset*-eko sinonimoak bakarrik.

domeinuak kontsultatzen dira. Hortaz, orain hirukotea [IZENA eta bere DOMEINUA/EREMU SEMANTIKOA, ERLAZIO SINTAKTIKOA, ADITZA] motakoa izango da. Adibidez, (14)ko corpuseko *irabazi* aditzaren agerpena, (30) adibideko hirukotean bilakatuko luke:

(30) [Futbol-talde/football/group (IZENA/DOMEINUA/EREMU SEMANTIKOA), Subjektua (ERLAZIO SINTAKTIKOA), Irabazi (ADITZA)]

Hautapen-murriztapena aditzaren adiera bakarrarentzat lortzen denean, hirukote hau aditzaren *synset*-arekin zehaztuta dator.

Azkenik, hirukote bakoitzaren pisua kalkulatzen da corpusean duten maiztasunaren arabera[18]. Hirukoteak daraman pisua geroz eta handiagoa izan, orduan eta fidagarritasun handiagoa. Azkeneko emaitza (28) eta (29)koen itxurakoak dira.

[18] Argibide gehiago (Agirre eta Martínez, 2002, 2001).

4 INGELESEKO HAUTAPEN-MURRIZTAPENEN AZTERKETA

Atal honetan corpus eta teknika desberdinak erabiliaz ingeleserako eskuratutako hautapen-murriztapenak aztertu eta ebaluatuko ditugu. Gogoan izan, hautapen-murriztapenen eskuratzea EuroWordNet-etik aukeratutako zortzi *synset*-entzat egingo dugula:

(1) *00605818 play1 / jokatu2; "play games, play sports"*[1]

(2) *00610422 encounter5, meet10, play24, take on5 / jokatu3; "contend against an opponent in a sport or game"*

(3) *00468052 coach2, train7 / entrenatu1; "teach and supervise, as in sports or acting"*

(4) *00059698 train8 / entrenatu3; "exercise in order to prepare for an event or competition"*

(5) *00630097 equalize1, get even1 / berdindu16; "compensate; make the score equal"*

(6) *00630097 draw25, tie2 / berdindu15; "finish a game with an equal number of points, goals. . . "*

(7) *00620486 win1 / irabazi3; "be the winner in a contest or competition; be victorious"*

(8) *00620218 lose2 / galdu9; "fail to win"*

Synset hauetako ingeleseko ordainetan oinarritu gara esperimentuaren lehenengo atal honetan. Ingeleseko aditz hauekin guztiekin erabilitako metodologia ulergarriago egitearren, zortzi *synset* hauetatik *play 00605818 synset*-ean oinarrituko gara.

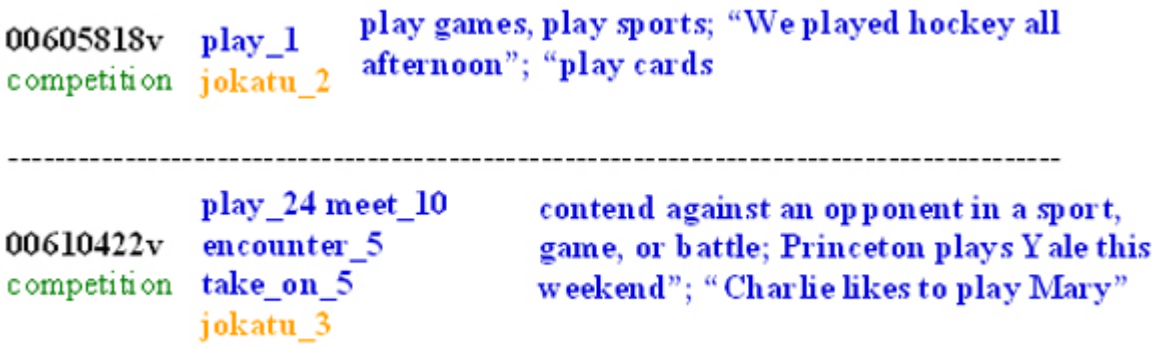
4.1 Aditzen aukeraketa

Aurrerago aipatu izan dugun bezala (ikus 3. atala), esperimentu 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, EuroWordNet-era jo dugu, eta adiera guzti horietatik 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 domeinuari begiratuta.

[1] EuroWordNet-en *synset*-ek zenbaki bat daramate (00605818), baita *synset* barruko ordainek ere (*play1*). Lehenengoa *synset* osoari dagokio, osatzen duten ordainak barne. Bigarrenak hitzaren adiera zehazten du, hau da, hitz polisemikoen adierak zenbakituak datoz. Hala ere, biekin gauza bera adieraz daiteke: *play1*-ek EuroWordNet-eko *play* hitzaren lehenengo adiera adierazten du; eta *play 00605818*k, *play* hitzak 00605818 *synset*-eko adiera duela, hots, *play1*.

Har dezagun *jokatu* aditza. EuroWordNet-en begiratu ondoren, kirolarekin harremanetan dauden bi *synset* ditu; batak, *zerbaitetan jokatu* adierazten du (*jokatu 00605818*) eta besteak, *-ren aurka jokatu* (*jokatu 00610422*). 6. irudian ikus daitekeen bezala ezberdintasun hau glosan eta eremu semantikoan[2] adierazita dator.



6. irudia: *jokatu* aditzaren bi kirol *synset*-ak.

Domeinuari erreparatuz (ikus 7. irudia), bi *synset* hauek *sport* domeinuaren labela daramate[3]. Dena den, *synset*-ek domeinu bat baino gehiago izan dezakete, bi *synset* hauen kasuan ikus daitekeen bezala[4]. Ikusiko dugun bezala, honek hautapen-murriztapenetan ondorioak izango ditu.

Behin euskarako aditz-forma horren kirol adierak mugatu ditugula, EuroWordNet-eko *synset* horietan dauden ingeleseko ordainak hartuko ditugu ingeleseko hautapen-murriztapenen azterketa eta ebaluazioa egiteko. Gure kasuan, *jokatu 00605818 synset*-aren azterketan murgilduko gara, beraz, hemendik aurrera, bere ingeleseko ordaina (*play 00605818*) hartuko dugu oinarri gisa[5].

[2] Eremu semantikoa *competition* labelak adierazten du interfazean.

[3] EuroWordNet-eko terminologia ingelesez dagoenez, guk ere horrelaxe mantenduko dugu.

[4] *Play* eta *Sport* domeinuak 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.

[5] *Jokatu 00610422*ren kasuan, bere ingeleseko ordainak lau dira (*encounter, meet, play, take on*), hau da, kontzeptu hori adierazteko ingelesez sinonimo horiek erabil daitezke. Ikerlan honetan *synset* berean dauden ingeleseko ordain guztien hautapen-murriztapenak aztertu ditugu.

Synset-eko hitza(k)	*Kategoria*	*Synset-zenbakia*	*Domeinua*	*Domeinua*
jokatu, jokoan jardun	Aditza	00605818	play	**sport**
jokatu	Aditza	00610422	play	**sport**

7. irudia: *jokatu* aditzaren kirol *synset*-ak eta beraien domeinuak EuroWordNet-en.

4.2 Ingeleseko urre-patroiak (*goldstandard*)

Eskuratze-teknika desberdinen hautapen-murriztapenak ebaluatzeko, *synset* bakoitzeko urre-patroi batzuk zehaztu dira, kasu honetan *play 00605818*rentzat.

Bestalde, urre-patroiak, eskuratze-teknika bakoitzaren ereduan sortuko dira, hau da, guk sortutako urre-patroiek teknika hauen emaitzek hartzen duten itxura hartuko dute. Hala, alde batetik, hautapen-murriztapenak adierazteko *synset*-ean oinarritzen diren teknikak ditugu (w2c eta c2c), eta bestetik, domeinu-eremu semantikoetan oinarritzen direnak (w2semf eta s2semf). Gure urre-patroiak ere bi azpimultzo hauetan banatu ditugu; patroi batzuk *synset* bidez adieraziko ditugu w2c eta c2c tekniketatik lortutako hautapen-murriztapenak ebaluatzeko, eta beste patroiak domeinu-eremu semantiko bikoteen bidez definituko ditugu, w2semf tekniketatik lortutako hautapen-murriztapenak ebaluatu ahal izateko.

Hortaz, argi dago urre-patroi hauek proposatu ahal izateko EuroWordNet erabili behar izan dugula. Honezaz gain, esperimentuan erabilitako corpusetan ere oinarritu gara. Corpus hauetatik hartutako esaldietatik, aztertu beharreko aditz-adiera bakoitzaren jokaera linguistikoa orokortzen saiatu gara, gerora, orokortasun horiek (hautapen-murriztapenak, alegia) EuroWordNet-eko *synset* eta domeinu-eremu semantiko batzuen bidez adierazteko. Azken finean, makinak eskuratze-tekniken bidez egin beharko lukeena egiten saiatu gara.

(31)n ditugu *play 00605818* aditzaren urre-patroiak eta (32)n patroien adibideak[6]:

(31) **play 00605818 OBJEKTUAK**
w2c, c2c:
00240760 sport, athletics “an active diversion requiring physical exertion and competition”
00254052 game “a contest with rules to determine a winner”
04771851 contest, competition “an occasion on which a winner is selected from among two...”
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 individual, someone, somebody, mortal, human soul “a human being”
00017008 group, grouping “any number of entities (members) considered as a unit”
s2semf, w2semf:
person-person
factotum[7]-group

[6]Eskuratze-teknikek ematen dituzten emaitzak ingelesez daude, EuroWordNet-eko informazioa ingelesez dagoelako, hau da, EuroWordNet-eko euskarri informatikoa ingelesez dago, ingelesez ez dagoen bakarra beste hizkuntzetako ordainak eta glosak dira. Euskarako glosak oraindik ez daude guztiz itzulita, horregatik, ingelesekoetan oinarritzen gara.

[7]Adiera batek domeinurik ez duenean *factotum* labelarekin adierazten da.

(32) **OBJEKTUAK**
(a) John played **football**.
(b) John played **a match**.
(c) John played **five minutes**.
(d) John played **a game**.

SUBJEKTUAK
(a) **John** played football.
(b) **The football-team** played a match.

Kontuan izan beharrekoa da EuroWordNet hierarkia bat dela eta batzuetan ez dela horren erraza hautapen-murriztapena adierazten duen *synset* "egokia" aukeratzea, gerta litekeelako *synset* hori orokorregia izatea (hierarkian goregi egotea) eta zehatzegia izatea (hierarkian beheregi egotea). Esate baterako, *play* aditzarentzat {*contest, competition*}[8] hautapen-murriztapena proposatu ordez, EuroWordNet-eko bere hiponimoa (*match "a formal contest in which two or more persons or teams compete"*) proposatuz gero, aditz horren objektuen aukeraketa gehiago mugatuko genukeen, eta {*contest, competition*} bezalakoak ezingo genituzke zuzentzat jo. Alderantziz, {*contest, competition*} hautapen-murriztapenaren ordez, bere hiperonimoa *social event* (*"an event characteristic of persons forming groups"*) proposatu izan bagenu, aukera gehiegi zabalduko genukeen eta zuzenak ez diren hautapen-murriztapenak ere agertuko lirateke (adibidez, *play 00605818* aditzak *social event* horren hiponimoa den *ballet* hautapen-murriztapena onartuko luke).

Arazo hau bera areagotua egiten da domeinu-eremu semantiko bikoteen bidez adierazitako hautapen-murriztapenak ebaluatzean. Domeinu-eremu semantiko bikote hauek *synset*-ak baino orokorragoak dira. Adibidez, *Errealak partidua jokatu zuen* perpausean, subjektuaren hautapen-murriztapena *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).

Hautapen-murriztapenak adierazteko arazo hau dela eta, hauek ebaluatzeko maila desberdineko labelak erabili ditugu:

(a) **Zuzena:** Urre-patroiarekin bat datorrenean.

(b) **Onargarria:** Urre-patroiaren hiperonimoa edo hiponimoa denean. Domeinu-eremu semantiko bikoteen bidez adierazitako hautapen-murriztapen kasuan, onargarri bezala kontsideratu ditugu urre-patroia baino orokorrago edota zehatzago direnak.

(c) **Okerra:** Urre-patroiarekin bat ez datorrenean eta EuroWordNet-eko hierarkian ere loturarik ez dutenean.

Label hauek ez digute inolako arazorik eman *synset*-ekin adierazitako hautapen-murriztapenak ebaluatzerakoan. Haatik, domeinu-eremu semantiko bikoteekin adierazitakoak ebaluatzeko, batzuetan onargarriak ala okerrak diren erabakitzeko zailtasunak izan ditugu.

[8] *Synset* berean ordain bat baino gehiago agertzen direnean, azalpenetan *synset*-a adierazteko bi ordainak giltzen artean adieraziko ditugu.

Esate baterako, *play 00605818 synset*-ak [+gizaki] motako subjektuak har ditzake; *synset*-ekin adierazita, *00004865 person, individual, human "a human being"* hautapen-murriztapena litzateke, eta domeinu-eremu semantiko bikoteekin adierazita, *person-person*. Eskuratze-tekniken emaitzetan hauexek agertuz gero, *play 00605818*ren urre-patroietan definituak 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[9] 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*, *play 00605818*ren kirol adierarekin zerikusia duenez[10], onargarritzat joko genuke, eta *transport-person* eta *administration-person*, berriz, okertzat (*play 00605818*ren adierarekin bateragarriak ez direlako: *?Administrators played football*). Hala ere, datuak eta corpusak aztertuz, konturatu gara hauek *Brazilians*, *cyclist* eta bezalako agerpenetatik datozela, eta *play 00605818*rekin onargarriak direla (*Brazilians played football*). Hala, lehenengo begiratuan okerrak diruditen hautapen-murriztapen hauek kontuan hartzeko eta onargarri bezala ebaluatzeko, irizpide bat finkatu dugu: domeinu-eremu semantiko bikote batekin adierazitako hautapen-murriztapenak (*transport-person*, esaterako), urre-patroiko (*person-person*) eremu semantiko bera badu (*person*), orduan hautapen-murriztapen hori onargarritzat hartuko dugu.

Ikus daitekeen bezala, domeinu-eremu semantiko bikoteekin *synset*-ekin baino arazo gehiago sortu zaizkigu, eta horren ondorioz, irizpide batzuk zehazteko beharra sumatu dugu.

4.3 Hautapen-murriztapenen azterketa

Orain arte jarraitutako pausoak laburbilduko ditugu:

(a) Euskarako *jokatu* aditz-formatik abiatu gara eta honek dituen kirol adierak (*synset*-ak) bilatu ditugu EuroWordNet-en (*jokatu 00605818* eta *jokatu 00610422*).

(b) *Synset* hauek kirol adiera dutela egiaztatzeko beraien domeinua *sport* dela egiaztatu dugu.

(c) *Synset* bat hartu dugu –gure kasuan *jokatu 00605818* eta bere ingeleseko ordaina hartu dugu (*play 00605818*)– aditz-adiera honen hautapen-murriztapenak ingeleseko corpusetatik lortzeko.

(d) Eskuratze-tekniken emaitzak ebaluatu ahal izateko, ingeleseko corpusetan oinarrituta aditz-adiera horrek hartzen dituen hautapen-murriztapenen urre-patroiak sortu ditugu landutako eskuratze-teknika mota guztientzako.

[9] 3.1.5 atalean azaldu dugun bezala, EuroWordNet-ek (eta MCRk) domeinu hierarkia bat du, non domeinuak hiperonimia/hiponimiaren arabera antolatuak dauden.

[10] EuroWordNet-eko domeinu hierarkian *basketball* domeinua *sport* domeinuaren hiponimoa da.

Puntu honetara iristerakoan, eskuratze-teknika mota bakoitzaren emaitza ebaluatzeko gai izango gara. Hurrengo ataletan lan honen azalpenari ekingo diogu, eta horretarako, azalpena corpusen arabera antolatu dugu. Horrela, 4.3.1. atalean SemCor corpusetik eskuratutako hautapen-murriztapenen azterketa dugu, 4.3.2. atalean BNCtik eskuratutakoena, eta azkenik, 4.3.3. atalean EFEtik eskuratutakoena.

4.3.1 SemCor-etik ikasitako hautapen-murriztapenen azterketa

Corpus honetan c2c, w2c eta s2semf eskuratze-teknikak erabili dira. Hauekin irizpide metodologiko berdintsuak baliatu ditugun arren, beraien artean bada berezitasunik.

c2c SemCor-etik

c2c eskuratze-teknikak lortzen dituen objektu edo subjektuen hautapen-murriztapenak aditzaren adiera jakin baterako dira: *play 00605818*. Eskuratze-teknika honetan hautapen-murriztapenak aditz-adiera horrentzat baliagarri diren neurrian, *synset*-ean dituen sinonimoentzat, eta bere troponimoentzat ere baliagarri da, esan dugun bezala (ikus 3.2.1 atala).

Eskuratze-teknika honen emaitza ebaluatzeko garaian hurrengo urratsak jarraitu ditugu:

(a) **Hautapen-murriztapen bakoitzaren iturria ezagutu:** Hautapen-murriztapenak lortzeko corpusaren agerpen zehatzetan oinarritzen garenez (zehazkiago esanda aditzak hartzen dituen izenetan[11]), gure lehenengo lana corpuseko iturria zein den jakitea da. Hala, eskuratze-teknikaren lana oinarritik ebaluatu dezakegu, gerta baitaiteke corpuseko izenari okerreko hautapen-murriztapena egokitzea (geroago ikusiko dugun bezala). Horretarako, corpusean aditz horrekin subjektu edo objektu gisa agertu diren izenen zerrenda oso baliagarria litzaiguke, eta horixe bera eskaintzen digute w2w eta s2s[12] deituriko eskuzko lanerako baliabideak (w2c eta c2c teknikentzat, hurrenez hurren). Corpusetik agerpen horiek eskuz ateratzen jardun ordez, hauen bidez automatikoki ematen zaizkigu fitxategi batean (gero ikusiko ditugu hauen adibideak).[13]

(b) **Izena corpuseko testuinguruan kokatu:** Behin aditzaren agerpen zehatzak ezagutzen ditugula, corpusean hauen testuingurua bilatzen dugu, hauek guztiak aztertzen ari garen kirol aditzarekin bateragarriak diren ala ez eskuz egiaztatzeko.

(c) **Hautapen-murriztapenen ebaluazioa:** Eskuratze-tekniken hautapen-murriztapenen eta hauen corpuseko iturria aurrean izanda, ebaluazio egiten has gaitezke.

Adibidez, (33)n *play 00605818* aditzaren c2c objektu/subjektuen hautapen-murriztapenak aurkezten ditugu[14]. Gogoratu eskuratze-tekniken emaitzak izenen *synset*-zerrendak direla (gehienez hamarrekoak), eta beraien formatua hurrengoa dela:

[11] Ikus 3.2.1 atala.

[12] Terminologia ingelesez mantendu dugu, hizkuntzalaritza konputazionalean horrela ezagutzen direlako. Hala ere, hauek euskaraz ***hitza-hitza*** eta ***adiera-adiera*** bezala izenda daitezke.

[13] Hitzean oinarritzen den eskuratze-teknikaren antza handia dute (ikus 2.2.2. atala), baina hauek corpuseko agerpenak zuzenean hartzen ditu, inolako probabilitaterik eskaini gabe. Ez dira eskuratze-teknikak, hizkuntzalariaren lana errazten duten baliabideak baizik.

[14] Hautapen-murriztapen batzuk zertxobait eraldatu ditugu. Glosa luzeegiak laburtu egin ditugu, eta ordain ugari zituzten *synset* horiei ordain batzuk ere kendu dizkiegu.

c2c.obj (*eskuratze-teknika eta erlazio sintaktikoa*)
play 00605818 (*aditza eta bere synset-zenbakia*)
00004865 0.085 contest, competition "an occassion on which a winner. . . "
(synset-zenbakia, probabilitatea, synset-eko sinonimoak eta definizioa)

(33) **c2c.obj**
play 00605818
00228990 0.215 activity "any specific activity or pursuit"
00004865 0.117 person, individual, human "a human being"
00017008 0.102 group, grouping "any number of entities considered as a unit"
00009469 0.071 object, physical object "a physical (tangible and visible) entity"
04771851 0.035 contest, competition "an occasion on which a winner is selected from. . . "
03875944 0.029 interest, involvement "a sense of concern with curiosity about someone. . . "
08162378 0.014 cost "the total spent for goods [. . .] including money and time and labor"
01691640 0.011 horse "solid-hoofed herbivorous quadruped domesticated. . . "

c2c.subj
play 00605818
00017008 0.517 group, grouping "any number of entities (members) considered as a unit"
00004865 0.507 person, individual, human "a human being"
00009469 0.079 object, physical object "a physical (tangible and visible) entity"
08413915 0.032 digit "one of the elements that collectively form a system of numbers"
03953834 0.032 idea, thought "the content of cognition"

Hautapen-murriztapen hauen ebaluazio linguistikoa egin baino lehen, bakoitzaren izen zehatza zein izan den begiratu dugu, eta horretarako s2s erabili dugu. (34)n *play 00605818*rekin corpusean agertu diren izenak ematen ditugu, hauek s2s fitxategiko datuak dira[15]:

(34) **play00605818.s2s.obj**
play 00605818
ball 02103632
basketball 00270464
card 02245777
football 00263159
game 00254326
game 00256308
golf 00261291
group 00017008
person 00004865
pinball 00256739
rightfield 02836043

[15]Aditz eta izenen *synset*-ak zehaztuta datoz corpusetik, SemCor semantikoki etiketatua dagoen corpusa izaki, hori zuzenean lor baitaiteke. Hori dela eta s2s deritzogu, bai aditzaren adiera eta bai honekin agertu den izenaren adiera zehaztuta daudelako. Honek iturria ezagutzeko eta ebaluatzeko lana asko errazten du. Etiketatu gabeko corpusetatik zerrenda hau ez lituzke izenen *synset*-ak zehaztuta; corpusean aditz batekin agertu diren izenen zerrenda soila litzateke, hots, w2w deritzoguna – aditzaren adiera eta honekin agertu den izenaren adiera zehaztugabe daude.

play00605818.s2s.subj
play 00605818
group 00017008
line 05351374
mate 06390424
nine 08416391
person 00004865
young man 05971919

Behin aditzaren agerpenak ezagutzen ditugula, corpusean hauen testuingurua bilatzen dugu, *play 00605818* adiera horrekin erabili daitezkeela egiaztatzeko[16]. Testuinguru hauen (35)en ematen ditugu eta letra beltzez markaturik datoz hautapen-murriztapenaren iturri izan diren izenak - (34)n agertzen direnak, alegia[17].

(35) **Objektuen testuinguruak:**
a) I'll also play **a provisional ball** and get a ruling.
b) He played **basketball** there while working toward a law degree.
c) Nelson played **magnificent football**.
d) Skorich [...] played **football** at Cincinnati University.
e) ...

Subjektuen testuinguruak:
a) **The Mustangs** don't play this week.
b) **Our interior line and out linebackers** played exceptionally well.
c) **Nine of the league's teams** play in baseball parks and therefore...
d) For **a serious young man** who plays golf with a serious intensity.
e) ...

Testuinguru hauek guztiak *play 00605818* aditzarekin bateragarriak dira. s2s-eko datuei eta EuroWordNet-eko hiperonimoei esker, hurrengo pausoan corpuseko agerpen hauek zer hautapen-murriztapenekin islatzen diren azter dezakegu. Izen hauek zer hautapen-murriztapenetan bilakatu den jakiteko, bere hiperonimoari erreparatu behar zaio. Esate baterako, (34)ko play00605818.s2s.subj fitxategian dugun *mate* izena (*"a fellow member of a team"*), [+pertsona] hautapen-murriztapen gisa gauzatu da, EuroWordNet-en bere hiperonimoa {*person, individual, human*} *synset*-a delako.

(33)ko hautapen-murriztapenak (36) errepikatzen ditugu, eta (34)ko izenetatik abiatutako hautapen-murriztapenak letra lodiz adierazi ditugu, dagokien corpuseko agerpenak (izenak) ere zehaztuz:

[16]SemCor corpus etiketatua denez, oso erraz aurki daitezke corpusean *play 00605818*rekin agertu diren izen hauek eta beraien testuingurua.
[17](35)en agerpen batzuen testuinguruak bakarrik idatzi ditugu.

(36) **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 a unit"

PLAY: *The Owls*

00009469 0.071 object, physical object "a physical (tangible and visible) entity"

PLAY: *ball, card, rightfield*

04771851 0.035 contest, competition "an occasion on which a winner is..."

PLAY: *game2*

03875944 0.029 interest, involvement "a sense of concern with curiosity about someone..."

08162378 0.014 cost "the total spent for goods [...] including money and time and labor"

01691640 0.011 horse "solid-hoofed herbivorous quadruped domesticated..."

c2c.subj

play 00605818

00017008 0.517 group, grouping "any number of entities considered as a unit"

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"

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"

Letra lodiz markatu gabe hautapen-murriztapen ugari geratu dira. Gogoratu beharra dago c2c eskuratze-teknika aditz *synset* horren hautapen-murriztapenak eskuratzeaz gain, bere troponimoenak ere eskuratzen dituela. SemCor etiketatutako corpus bat izaki, eskuratze-teknika honek corpusean *play 00605818 synset*-aren troponimo bat agertuko balitz, bere hiperonimoarekin (*play 00605818*) erlazionatzeko gai izango da, eta klase guztiari hautapen-murriztapen berdinak egokitzen dizkio. Hortaz, pentsa daiteke iturria zehaztu gabe geratu diren horiek, *play 00605818*ren troponimoetatik datozela. Horretarako, 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 00605818*rekin orain arte jarraitu dugun metodologia bera erabiliko dugu troponimo hauekin ere.

Lehenengo troponimoak eta beraien domeinuak ezagutu behar ditugu (ikus 8. irudia). Ondoren, s2s-hype erabilita troponimoen agerpenak corpusean zehaztu eta hauen testuinguruak aztertu behar ditugu kirol adiera dutela egiaztatzeko:

Synset-eko hitza(k)	*Kategoria*	*Synset-zenbakia*	*Domeinua*	*Domeinua*
start	Aditza	00607112	play	**sport**
field	Aditza	00611046	play	**sport**
bet on	Aditza	00646526	baseball	**sport**
stake	Aditza	00646526	play	**sport**
parlay	Aditza	00646865	play	**sport**

8. irudia: *play 00605818 synset*-aren troponimoak eta bere domeinuak EuskalWordNet-en.

(37) ***play 00605818*ren troponimoak eta hauen corpuseko objektu/subjektuen agerpenak:**

a) start 00607112 "play in the starting line-up, in team sports"
OBJEKTUA: **mate** 06390424 "a fellow member of a team"
SUBJEKTUA: **person** 00004865 "human being"
AGERPENA: **Haddix SUBJ** will start against [...] Cardinal **matesOBJ**.
b) field 00611046 "play as a fielder, in baseball or cricket"
OBJEKTUA: **team** 05166149 "a cooperative unit"
SUBJEKTUA: **group** 00017008 "any number of entities [...] considered as a unit"
AGERPENA: **the Orioles SUBJ** fielded possibly their strongest **team OBJ** of...
c) bet on 00646526 place a bet on; "Which horse are you backing?"
OBJEKTUA: **pony** 01698920 "an informal term for a racehorse"
SUBJEKTUA: **person** 00004865 "human being"
AGERPENA: **Berry SUBJ** got elected on his advocacy of betting on **the ponies OBJ**.
d) stake 00646526 place a bet on; "I'm betting on the new horse"
OBJEKTUA: **career** 00341672 "the particular occupation for which you are trained"
SUBJEKTUA: **person** 00004865 "human being"
AGERPENA: **I SUBJ**'m willing stake my political **career OBJ** on it.
e) parlay 00646865 "stake winnings from one bet on a subsequent wager"
OBJEKTUA: **earnings** 08148137 "something that remunerates"
SUBJEKTUA: **person** 00004865 "human being"
AGERPENA: **He SUBJ** parlayed his **earnings OBJ** [...].

*Play 00605818*ren troponimo batzuek (*bet on*, *parlay* eta *stake*) ez dirudite kirolarekin harreman handirik dutenik, apustua domeinua baitute. Hala ere, aditz eta testuinguru hauek zuzentzat jo ditugu, berez, EuroWordNet-en aditz hauek *play 00605818*ren troponimoak direlako[18].

Eta azkenik, izen hauen *synset* hiperonimoari begiratuta, zer hautapen-murriztapenetan bilakatu diren jakin dezakegu. (38)n letra lodiz markatu ditugu corpuseko izenetatik eratorritako hautapen-murriztapenak eta beraien azpian zerrendatuak datoz corpuseko agerpenak (bai *play 00605818*renak eta bai honen troponimoenak):

[18] Adiera ezberdintasun hau domeinuekin ere adierazten da. *Play 00605818*ren troponimoek, *sport* domeinuaz gain, *play* domeinua dute. *Play* eta *Sport* domeinuak 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.

(38) **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 a unit"

PLAY: *The Owls*

FIELD: *team*

00009469 0.071 object, physical object "a physical (tangible and visible) 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 someone..."

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 a unit"

PLAY: *The Mustangs, Texans, line...*

FIELD: *The Oriols*

00004865 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; the main thing you are thinking about"

Horrela bada, troponimoak kontuan izanda, ia hautapen-murriztapen guztien iturria lor dezakegu, hau da, uler dezakegu makinak zer pauso jarraitu dituen hautapen-murriztapen horiek eskuratzeko. Dena den, oraindik geratu dira hautapen-murriztapen batzuk iturria 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 honekin egin duen eskuratzeaz?

Ebaluazio honekin hasi baino lehen, ekar dezagun gogora hasieratik eskuratze-teknika mota hauentzako proposatutako urre-patroiak, hauekin parekatu behar baitugu c2c hautapen-murriztapen hauek:

(39) **play 00605818 OBJEKTUAK**
w2c, c2c:
00240760 sport, athletics "an active diversion requiring physical exertion and competition"
04771851 contest, competition "an occasion on which a winner is selected from among..."
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 individual, someone, somebody, mortal, human soul "a human being"
00017008 group, grouping "any number of entities (members) considered as a unit"

(39)n oinarrituta, (40)n letra lodiz markatu ditugu zuzenak iruditu zaizkigun hautapen-murriztapenak, beste guztiak okertzat jo ditugu:

(40) **c2c.obj**
play 00605818
00228990 0.215 activity "any specific activity or pursuit" 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 which..." 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 labela daraman bakarra *activity* objektu hautapen-murriztapena da, eta hauxe da probabilitate-neurri handieneko hautapen-murriztapena (0.215), berez, eskuratze-teknikak egokitzat proposatzen duena. *Synset* hau *football, basketball* eta abarren hiperonimoa da, baina tartean badaude hautapen-murriztapen gisa egokiagoak direnak, urre-patroian proposaturiko {*sport, athletics*} adibidez. Hizkuntzalaritzari begira, *activity* klase semantikoa ezin da beti izan *play 00605818*ren 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 hautapen-murriztapena da, eta hau probabilitate-neurriaren zerrendan ez da lehenengoetakoa (bosgarrena da). Beste hautapen-murriztapen guztien iturria ez zen aditz-adiera honentzat egokia. Esate baterako, *person* hautapen-murriztapena ez dagokio *play 00605818*ri baizik eta *play 00610422*ri. Azken *synset* honek objektu gisa [+pertsona] tasuna daramatenak hartzen ditu bere EuroWordNet-eko glosan adierazten den bezala (*contest against an opponent*). Zergatik azaltzen dira *play 00610422*ren hautapen-murriztapenak *play 00605818*koekin nahastuta? SemCor-en etiketatze-erroreak daudelako, eta horren adibide *play 00605818* eta *play 00610422*ren arteko nahasketa delako, hau da, *play* kirol adierarekin agertzen denean SemCor-en hau *play 00605818* bezala etiketatu dute. Hortaz, SemCor-eko *play*

*00605818*ko hautapen-murriztapenetan *play 00610422*renak ere azaldu dira. 4.4 atalean azalduko ditugu errore hauen arrazoia sakonkiago.

Okerrak diren *object* eta *digit* hautapen-murriztapenen azalpena 4.4 atalean dago.

Azkenik, esan beharra dago troponimoetatik etorritako hautapen-murriztapen gehienak okerrak direla. Zuzenak direnak troponimo gabe lortu dira. *Play 00605818*ren kasuan *bet on*, *parlay* eta *stake* bezalako troponimoak ditu, hots, apustua domeinuarekin zerikusia dutenak. Honenbestez, *play* domeinua dute, *sport*-ekin batera. Beste domeinuak indar gehiago duela dirudi eta honek hautapen-murriztapenetan eragina izan du. Hauen hautapen-murriztapenak *play 00605818*renekin zeharo ezberdinak dira, esate baterako, aditz hauen objektu arruntenetako bat dirua izango da (*cost* hautapen-murriztapenetan). *Horse* hautapen-murriztapena, adibidez, *bet on a pony* testuingurutik dator. Beraz, ez dirudi aditz batek eta bere troponimoek hautapen-murriztapen berdinak dituztenik (behintzat EuroWordNet hierarkian oinarritzen bagara).

w2c SemCor-etik

3.2.1. atalean adierazi dugun bezala, eredu honekin aditz-formaren (hitzak izan ditzakeen adiera guztiak kontuan hartuta) objektu edo subjektu hautapen-murriztapenak lortzen dira. Beraz, gure kasuan, hautapen-murriztapen 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 adiera guztiak kontuan hartu, adiera guzti horien artean guk arreta berezia kirol adierari emango diogu, horrela, adiera hori bakarrik eskuratzen dituzten teknikekin erkatzeko gai izango gara.

(41)en *play* aditz-formaren w2c objektu/subjektuen hautapen-murriztapenak ditugu[19]:

(41) **w2c.obj**
play
00228990 0.148 activity “any specific activity or pursuit”
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 as a unit”
00018599 0.029 communication “something that is communicated between people or groups”
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 which a winner is selected from...”

w2c.subj
play
00004865 0.308 person, individual, human “a human being”
00017008 0.125 group, grouping “any number of entities (members) considered as a unit”
00009469 0.059 object, physical object “a physical (tangible and visible) entity”
00012670 0.043 abstraction “a general concept formed by extracting common features from...”
06467898 0.029 physical phenomenon “a natural phenomenon involving the physics...”
08522741 0.016 situation, state of affairs “the general state of things; the combination of...”
08125923 0.011 community “common ownership”
00012878 0.008 cognition knowledge “the psychological result of perception and learning...”

Hautapen-murriztapen hauen ebaluazioa egin baino lehen bakoitzaren iturria ezagutzen saiatu gara, eta berriro ere, s2s-ko datuak erabili ditugu. Beraz, (41)eko hauta-

[19]Gogoratu lehenengo hamar hautapen-murriztapenak bakarrik hartzen ditugula, probabilitate-neurri handienekoak, alegia. Berez, zerrenda luzeagoa izan baitaiteke.

pen-murriztapenen iturria jakiteko, s2s fitxategiko datuetara jo behar dugu, hauek (42)n ikusgarri ditugularik. Eskuratze-teknika honek (w2c) aditzaren adiera guztiak kontuan hartzen dituenez, s2s fitxategian EuroWordNet-en *play* aditzaren *synset* guztiek hartu dituzten izenak dauzkagu, eta izenen ondoan dagozkien EuroWordNet-eko *synset*-zenbakia. Letra lodiz idatzi ditugu *play* kirol adierari dagozkionak.

(42) **play.s2s.obj**

play 00008435
    theme 04314223
play 00008579
    host 06200482
    role 00399406
play 00605818
    ball 02103632
    basketball 00270464
    card 02245777
    football 00263159
    game 00254326
    game 00256308
    golf 00261291
    group 00017008
    person 00004865
    pinball 00256739
    rightfield 02836043
play 00986444
    performance 04487114
    song 04567799
    tune 04555665
play 00986807
    group 00017008
    musician 06219943
    performance 04487114
    recital 04488642

play.s2s.subj

play 00008435
    east 05409719
play 00605818
    group 00017008
    line 05351374
    mate 06390424
    nine 08416391
    person 00004865
    young man 05971919
play 00983496
    actor 05919271
    person 00004865
play 00986444
    motif 04556729
    person 00004865
play 01369076
    child 05996700

Hala eta guztiz ere, w2c eskuratze-teknika honekin zaila da lotzea hautapen-murriztapen bakoitza bere iturriarekin, ez baitakigu hautapen-murriztapen hori zer adierari dagokion. Esaterako, (43)n begiratzen badugu, *play 00605818*ren subjektua izateko probabilitate handiena duen hautapen-murriztapena, {*person, individual, human*} *synset*-ak adierazten duena da, [+pertsona] alegia. Hortaz, badakigu *play 00605818*k 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.

SemCor-eko s2s izen-zerrendari esker, hautapen-murriztapen bakoitzaren iturria zehazteko gai izan gaitezke. (42)ko zerrenda guztiaren hiperonimoak begiratuta zer hautapen-murriztapenetan bilakatu diren jakin dezakegu. Baina, lan honek gure esperimentuari ez lioke abantaila handirik ekarriko, eta gainera, honetatik oso erabilera konputazional mugatua lortuko genuke. Itzulpen automatikoan edo adiera desanbiguazioan, adibidez, w2c ez litzateke oso erabilgarria izango, aditz-forma baten aurrean ezingo genukeelako honen hautapen-murriztapenetatik bere adiera mugatu. Horregatik adiera batean oinarritzearen garrantzia.

Hautapen-murriztapen hauetan adiera guztiak nahasturik daudenez, ezinezkoa zaigu aditz-adiera baten hautapen-murriztapenak ebaluatzea, aditz horren adiera posible guztiak kontuan hartuta daudelako. Horregatik, w2c motako hautapen-murriztapenak aztertzerakoan, *play 00605818*rekin zerikusirik duten hautapen-murriztapenak ezberdintzen saiatu gara, gerora *play 00605818*rekin egindako beste eskuratze-tekniken emaitzekin bat datozen ikusteko. Hala, (41)eko hautapen-murriztapenak (44)n errepikatu ditugu, eta letra lodiz markatu ditugu gure ustez *play* aditzaren kirol adieraren objektu/subjektuak izan daitezkeen hautapen-murriztapenak, (43)ko 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 hautapen-murriztapenak izan daitezkeen heinean, zuzenak izan daitezkeelako. Bestalde, hautapen-murriztapenen azpian SemCor-eko *play 00605818*rekin agertu diren izenak zerrendatuak datoz.

(43) **play 00605818 OBJEKTUAK**
w2c, c2c:
00240760 sport, athletics “an active diversion requiring physical exertion and competition”
00254052 game “a contest with rules to determine a winner”
04771851 contest, competition “an occasion on which a winner is selected from among. . . ”
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”

(44) **w2c.obj**
play
00228990 0.148 activity "any specific activity or pursuit" 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 as a unit"
00018599 0.029 communication "something that is communicated between people or groups"
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 which..." ZUZENA
PLAY: *game2*

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 features from..."
06467898 0.029 physical phenomenon "a natural phenomenon involving the physics of..."
08522741 0.016 situation, state of affairs "the general state of things; the combination of..."
08125923 0.011 community "common ownership"
00012878 0.008 cognition knowledge "the psychological result of perception and learning..."

Ikus daitekeen bezala, urre-patroiko hautapen-murriztapen gehienak azaldu egiten dira. Subjektuen kasuan ez da harritzekoa, beste adieren subjektuek ere hautapen-murriztapen horiek onar baititzakete. Arrazoi horregatik daude probabilitate altueneko postuetan. Objektuen artean, kirolari bakarrik dagokion hautapen-murriztapena {*contest, competition*} da, eskuratze-tekniken proposamenean azkena, probabilitate baxuenarekin agertu dena, alegia. Nahiz eta *play 00605818*k ekintza bat har dezakeen objektu gisa (*activity*-k jasotzen dituen *football, basketball,* eta abar), beste adieretan ere hautapen-murriztapen hau ager daiteke.

s2semf SemCor-etik

Eskuratze-teknika honek aditzaren adiera bakoitzarentzat hautapen-murriztapenak domeinu-eremu semantiko bikoteekin adierazten ditu. Honek orain arte erabilitako metodologia baldintzatzen du, ezin jakin baitezakegu hauen agerpen zehatzak zeintzuk diren. 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, hautapen-murriztapen desberdin hauek izen beretik abiatutakoak izan daitezke. Hala, gehinetan ezinezkoa zaigu hautapen-murriztapen hauen iturria zein den jakitea.

Bestalde, bikote hauek adierazten dutena ulertzea ez da begibistakoa. Domeinuak eta eremu semantikoen informazioa *synset*-ena baino orokorragoa da eta gehienetan EuroWordNet-era jo behar dugu hauen azpian zer dagoen ulertu ahal izateko. Beraz, ezin

dugu eskuratze-teknika honen ebaluazio sakon bat egin, baina datuak aurrean izanda[20], subjektiboki bada ere, horietatik zuzenak zein diren aipatu dezakegu.

Ebaluazioarekin hasi baino lehen, komeni da gogora ekartzea batetik, eskuratze-tekniken emaitzak izenen domeinu-eremu semantiko bikoteen zerrendak direla (gehienez hamarrekoak), eta beraien formatua hurrengoa dela:

s2semf.obj(*eskuratze-teknika eta erlazio sintaktikoa*)
play 00605818 (*aditza eta bere synset-zenbakia*)
obj sport-event 28
(erlazio sintaktikoa, domeinu-eremu semantiko bikotea eta probabilitatea)

Eta bestetik, zeintzuk diren eskuratze-teknika mota honentzat proposatutako urrepatroiak:

(45) **play 00605818 OBJEKTUAK**
s2semf, w2semf:
sport-event
time period-time
sport-act
play-act

play 00605818 SUBJEKTUAK
s2semf, w2semf:
person-person
factotum-group

(46)n letra lodiz markatu ditugu zuzenak/onargarriak iruditu zaizkigun hautapen-murriztapenak:

(46) **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 1ONARGARRIA
subj factotum-group 1 ZUZENA
subj factotum-Tops 1 ONARGARRIA
subj person-person 1 ZUZENA
subj biology-Tops 0.5
subj anthropology-Tops 1

[20](34)ko datuak erabili ditugu.

Objektu hautapen-murriztapenetako *play-act, sport-act* urre-patroietan daudenez ez dugu inolako zalantzarik zuzen bezala ebaluatzeko. Hauen zehaztapen gisa har daitezke *golf-act* eta *basketball-act*, are gehiago, domeinuen hierarkian *golf* eta *basketball*, *sport* domeinuaren jasota baitaude. Arrazoi horregatik onargarri bezala hartu ditugu, urre-patroia baino zehatzagoak direlako. Urre-patroiko beste bi objektuen hautapen-murriztapenak ez dira s2semf hautapen-murriztapen 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-artifact*-en 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 EuroWordNet-en *artifact* da, eta arrazoi horregatik agertu da hautapen-murriztapen hori.

Beste hautapen-murriztapen bat *play ball* 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*. EuroWordNet-en kontsultatuz gero, *ball synset* ugaritan dago baina horietako batek ere ez du ekintza-adiera hori[21]. Beraz, makinak horren ordezko bat hartu du, *artifact* adiera duena, hain zuzen ere.

Subjektuei dagokionez, s2semf eskuratze-teknikak urre-patroian proposaturiko bi hautapen-murriztapenak lortu ditu. Horietaz gain, onargarri bezala ebaluatu ditugun *sport-person* eta *factotum-Tops* ere baditu. Lehenengoa, *person-person* horren zehaztapena da, eta honen iturria *mate* izenaren agerpena izan daiteke, honen domeinua *sport* delako. Hala ere, errepikatu beharra dago hautapen-murriztapen hauen iturria zehaztea ez dela lan batere erraza. Bigarrena, oso hautapen-murriztapen orokorra da[22] eta honen iturria edozer izan daiteke.

Probabilitate altueneko subjektu *number-quantity* hautapen-murriztapena ez da zuzena, baina honek c2c eskuratze-teknikako *digit* hautapen-murriztapenenarekin zerikusia duela uste dugu (azalpena 4.4 atalean).

4.3.2 BNCtik ikasitako hautapen-murriztapenen azterketa

Corpus honetan c2c eta w2c eskuratze-teknikak erabili dira. Erabilitako irizpide metodologikoa orain artekoaren ezberdina izan da. BNC corpusa ez dago adierekin etiketatua, ezta domeinuka antolatuta ere. Honek guztiak hautapen-murriztapenak nondik datozen zehaztea ezinezkoa egiten du. SemCor-eko eskuratze-teknika batzuk aztertzerakoan, s2s (eta s2s-hype) fitxategiak genituen non aditzaren adierak (*synset*-zenbakia) zehaztuak zeuden eta baita izenenak ere. BNC 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[23]:

[21] Kontuan izan beharrekoa da, WordNet eta EuroWordNet etengabe eguneratzen dauden ezagutza-baseak direla, eta batzuetan horrelako hutsuneak aurki daitezkeela.

[22] Bikote honek ia edozer adieraz dezake, *factotum*-ekin domeinurik ez duten hitzak adierazten direlako, eta *Tops* eremuak EuroWordNet-eko hierarkian oso goian dauden *synset*-ak jasotzen dituelako. Beraz, oso orokorra den kontzeptu baten aurrean gaude.

[23] Oso zerrenda luzea da, eta hemen objektu bezala sailkatu dituenen lagin bat bakarrik dago. Zerrenda osoa eranskinetan dago.

(47) **bnc.w2w.kirola.obj**
play
After Wentworth
Afterwards
Alain
Albert Hall
Albrecht
Alfred
All Blacks
Allcock
Although
American
Americans
And
Anderlecht
Andy Lloyd
Anglicised
Argentina
Arsenal
As
At
Australian
Australian Open
Austria
B
BB
Bach
Bach Brandenburg Concerto
Back
Baliol
Ballesteros
Baresi
Because
Becker
Bet
Billy
Blackeyes
Boswell
Botvinnik
Bountiful
Brazil
Brownie Hansen
...

Mila hitzetik gora osatutako zerrendak dira, eta izugarrizko eskuzko lana litzateke bakoitzaren testuinguruak aztertu eta kirolaren domeinuari dagozkioenak baztertzea, gero horren arabera beraien EuroWordNet-eko *synset* eta hiperonimo posibleak zehazteko.

Arrazoi horregatik, eta datu enpirikoetan oinarritu gabe, zuzenean BNC gainean aplikatutako eskuratze-teknika hauen hautapen-murriztapenak gure urre-patroiekin erkatu ditugu.

w2c BNCtik

Teknika honekin *play*ren adiera guztien objektu edo subjektuen hautapen-murriztapenak lortzen dira. (48)n ikus ditzakegu:

(48) **w2c.obj**
play
00228990 0.082 activity "any specific activity or pursuit"
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 and learning..."
04771851 0.009 contest, competition "an occasion on which a winner is selected from..."
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"
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 people who work together"
04313427 0.02 message, content, subject "what a communication that is about something..."
00016649 0.01 act, human action, human activity "something that people do or..."
00018966 0.01 measure, quantity "how much there is of something that you can measure"
00014314 0.01 location "a point or extent in space"
00012878 0.01 cognition, knowledge "the psychological result of perception and learning..."

Subjektu hautapen-murriztapenetako lehenengoa eta hirugarrena, ingeleseko *he* izenordainari dagozkie. Aurreprozesu lanetan ez zirenez izenordainak markatu, *parser*-ak ez ditu detektatzen, eta gainera, EuroWordNet-en izenordainik ez dagoenez, makinak *he* izenordainaren idazkera antzekoa duten beste bi *synset*-ekin parekatu du. Arrazoi horregatik dira probabilitate handiena dituzten hautapen-murriztapenak. Honi buruz, 4.4 atalean mintzatuko gara.

Eskuratze-teknika honen hautapen-murriztapenak gure urre-patroiekin erkatu ditugu (ikus (49)), kirol adierari egokitu dakiozkenak nabarmentzeko –letra lodiz (50)en. Urre-patroia bera edo antzekoa (hiperonimo edo hiponimo bat adibidez) denean zuzen bezala kontsideratu dugu; baina bat ez datozenak ez ditugu okertzat hartu, hauek berez, beste aditz-adiera baten hautapen-murriztapenak izan daitezkeen heinean, zuzenak izan daitezkeelako.

(49) **Play 00605818 OBJEKTUAK**
w2c, c2c:
00240760 sport, athletics "an active diversion requiring physical exertion and competition"
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"

(50) **w2c.obj**
play
00228990 0.082 activity "any specific activity or pursuit" 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 and learning..."
04771851 0.009 contest, competition "an occasion on which..." 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 something..."
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 you can measure"
00014314 0.01 location "a point or extent in space"
00012878 0.01 cognition, knowledge "the psychological result of perception and learning..."

Ikus daitekeen bezala, urre-patroiko hautapen-murriztapen gehienak azaldu egiten dira. Objektuen artean, kirolari bakarrik dagokion hautapen-murriztapena {*contest, competition*} da. Onargarri labela daraman hautapen-murriztapena (*activity*) urre-patroiko {*sport, athletics*}-en hiperonimoa da. Nahiz eta *play 00605818*k ekintza bat har dezakeen objektu gisa (*activity*-k jasotzen dituen *football, basketball* eta abar), beste adieretan ere hautapen-murriztapen hau ager daiteke (*He played Hamlet* perpausean, adibidez), eta horregatik du probabilitate-neurri altuena.

Subjektuen kasuan, {*organisation, organization*} onargarri bezala dago, {*group, grouping*} *synset*-aren hiponimo bat delako, talde mota zehatzagoa, alegia. Zuzentzat hartu dugun bakarra (eta probabilitate-neurri altuenetakoa duena) *person* hautapen-murriztapena izan daiteke. Hau baino probabilitate-neurri handiagoa *helium*-ek du, goraxeago aipatutako arrazoiengatik.

Bestalde, *location* bezalako subjektu hautapen-murriztapenak agertzen direnean, eta w2w fitxategietan begiratuta, leku izen berezietatik etor daitezkeen (*Argentina, Madrid...*)

susmoa dugu. Horrelakoekin corpusean kirol taldeak adierazi nahi dira eta EuroWordNet-en leku-izen berezi bezala daude. Hori dela eta, *location* bezalako hautapen-murriztapenak ditugu *play* aditzarekin.

Beraz, kirol adierari dagokion hautapen-murriztapen bakarra {*contest, competition*} dela dirudi.

c2c BNCtik

Eskuratze-teknika honek lortzen dituen objektu edo subjektuen hautapen-murriztapenak *play 00605818* adierarako dira. Bestalde, gogoan izan, objektu/subjektuen hautapen-murriztapenak izen klase gisa adierazten dituela: corpuseko izenetatik abiatuta, horien hiperonimoak darabiltza. Gainera, c2c eskuratze-teknikak, aditza ere klase bezala ulertzen du, hau da, lortutako hautapen-murriztapenak baliagarriak dira aditz horrentzat, bere *synset*-eko sinonimoentzat eta bere troponimoentzat (ikus 4.3.1. atala).

(51)en oinarrituta, (52)n letra lodiz markatu ditugu zuzenak iruditu zaizkigun hautapen-murriztapenak, beste guztiak okerrak dira:

(51) **play 00605818 OBJEKTUAK**
w2c, c2c:
00240760 sport, athletics "an active diversion requiring physical exertion and competition"
04771851 contest, competition "an occasion on which a winner is selected from among..."
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 individual, someone, somebody, mortal, human soul "a human being"
00017008 group, grouping "any number of entities (members) considered as a unit"

(52) **c2c.obj**
play 00605818
09065837 0.006 period, amount of time "an indefinite length of time" ZUZENA
08813320 0.004 helium "a very light colorless element that..."
or parties 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 which..." 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 hautapen-murriztapenetan denborazkoak harrapatu ditu bakarrik, bata zuzena (zerrendatik probabilitate-neurri handiena duena gainera) eta bestea onargarria (aurrekoaren hiponimo bat). Eta subjektuaren hautapen-murriztapenetan ez du bat bera ere harrapatu. Berriro ere, aipatu behar dugu, subjektuaren hautapen-murriztapenetako *helium synset*-a ingeleseko *he* izenordainari dagokiola, eta hauxe dela subjektu hautapen-murriztapenen artean probabilitate-neurri altuena duena.

Horrela bada, eskuratze-teknika honen emaitzak ez dira batere onak izan. Corpusarengatik izan daiteke (etiketatua ez egotea, kirol domeinua bakarrik ez izatea...), baina hala ere, harritzekoa da subjektuetan hautapen-murriztapen zuzen bat bera ere ez lortzea, subjektuen hautapen-murriztapenen ikasketan aukerak askoz gutxiago izanik (aditzen objektuak mota askotakoak izan daitezke, aditzen subjektuak, aldiz, normalean [+pertsona] dira). Objektuekin ere harritzekoa da kirol domeinuan arruntak diren {*contest, competition*} edo {*sport, athletics*} objektu hautapen-murriztapenen ordez denborazkoak bakarrik eskuratu izana.

Bestalde, troponimoen eraginak zerikusirik duela pentsa dezakegu, baina SemCor ez bezala, BNC etiketatu gabeko corpusa denez, oso zaila egiten zaigu hipotesi hori zehatzmehatz egiaztatzea.

4.3.3 EFEtik ikasitako hautapen-murriztapenen azterketa

EFE domeinuka antolatutako corpusa da, eta guk kirol domeinuari dagokion atala erabili dugu esperimentu honetarako. Corpus honetan w2semf eskuratze-teknika aplikatu dugu. Teknika honek ikasten dituen hautapen-murriztapenak aditz-formarentzat dira, aditzaren adiera guztientzat, alegia. Gogoratu, probabilitate kopuru altuenetik baxuenera ordenaturiko domeinu-eremu semantiko bikoteak direla.

BNCren antzera, corpus hau ez dago semantikoki etiketatuta, eta horrek hautapen-murriztapenen iturria zehaztea zaildu egiten du. Corpus honek ere w2w fitxategi bat du, bertan EFE corpuseko kirol domeinuan *play* aditz-formarekin agertu diren hitzen zerrenda dago, hauen maiztasunaren arabera ordenaturik[24]:

(53) **w2w.play.kirola.obj**

play
- 103 game
- 75 match
- 30 which
- 21 team
- 14 host
- 13 soccer
- 10 role
- 8 Wednesday
- 7 tournament
- 7 season
- 7 man
- 7 Cup
- 6 who
- 6 two
- 6 Sunday
- 5 fan

[24]Oso zerrenda luzea da, eta hemen objektu bezala sailkatu dituenen lagin bat bakarrik dago. Zerrenda osoa eranskinetan dago.

5 defense
5 Juniors
4 year
4 sport
4 series
4 one
4 half
4 Thursday
4 Saturday
4 Bolivar
3 weekend
3 week
3 time
3 ...

Hirurehun hitzetik gorako zerrendak dira, eta izugarrizko eskuzko lana litzateke bakoitzaren testuinguruak aztertu eta kirolaren domeinuari dagozkionak baztertzea, gero horren arabera beraien EuroWordNet-eko *synset*, eremu semantiko eta domeinu posibleak zehazteko.

Honekin batera, corpus honekin erabili dugun w2semf eskuratze-teknikak ematen dituen hautapen-murriztapenek ez dute laguntzen hautapen-murriztapenen iturria bilatzeko garaian. Batetik, ez direlako ulerterrazak, hau da, domeinuak eta eremu semantikoen informazioa *synset*-ena baino orokorragoa da, eta gehienetan EuroWordNet-era jo behar dugu hauen azpian zer *synset* jasotzen diren jakiteko. Bestetik, hitz berak domeinu eta eremu semantiko bat baino gehiago har ditzakeelako (4.3.1 atalean ikusi dugun bezala). Honezaz gain, EFE gainean erabilitako eskuratze-teknikak aditz-forma osoan funtsaturikoa da.

Arrazoi hauek guztiengatik, eta datu enpirikoetan oinarritu gabe, zuzenean EFE gainean aplikatutako eskuratze-teknika hauen hautapen-murriztapenak gure urre-patroiekin erkatu ditugu.

w2semf EFEtik

(54)n eskuratze-teknika honentzat proposatu ditugun urre-patroiak daude eta (55)en *play* aditzaren w2semf objektu/subjektu hautapen-murriztapenak ditugu (letra lodiz gure ustez *play 00605818* aditzari dagozkionak):

(54) **play OBJEKTUAK**
w2semf:
sport-event
time period-time
sport-act
play-act

play SUBJEKTUAK
w2semf:
person-person
factotum-group

(55) **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 ONARGARRIA
subj chemistry-substance 52.66
subj sport-group 44.01 ONARGARRIA
subj zoology-group 40.5 ONARGARRIA
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 bera edo antzekoa (domeinu edo eremu semantiko orokorrago edo zehatzago bat adibidez) denean zuzen bezala kontsideratu dugu (esaterako, *sport-group*); baina bat ez datozenak ez ditugu okertzat hartu, hauek berez, beste aditz-adiera baten hautapen-murriztapenak izan daitezkeen heinean, zuzenak izan daitezkeelako. Hautapen-murriztapen batzuk zalantzan jar daitezke. *Sport-group*-en kasuan ez dago dudarik kirol adierarekin zerikusia duenik; *administration-group*-en kasuan, nahiz eta lehenengo begiratuan okerra zela iruditu, w2w zerrendak eta corpusak aztertuz, konturatu ginen *Colombians, Brazilians* eta abar bezalako agerpenetatik zetozela. Izen hauen domeinuak EuroWordNet-en *administration* da. Horregatik dugu *administration-group* bezalako hautapen-murriztapen bat. Hau horrela izanda, onargarritzat jo dugu. Eskuratze-teknika honek izen bereziak x batez adierazten ditu.

Eskuratze-teknika honek izen bereziak x batez adierazten ditu.

Aditzaren adiera guztiak kontuan hartzen dituen eskuratze-teknika izateko, kirolari dagozkion hautapen-murriztapen ugari daude. Urre-patroiko objektu hautapen-murriztapen guztiak daude eta oso probabilitate-neurri altuekin, gainera. Dirudienez, eta aditz-forman oinarritutako beste eskuratze-tekniken emaitzekin erkatuz gero, kirol domeinuan oinarritutako corpus baten gainean aritzeak badu eraginik, corpuseko domeinuak aditzaren beste adierak baztertu egiten dituelako.

Orain arteko eskuratze-teknikekin aipatu ditugun erroreak ikus daitezke w2semf honetan ere (gero 4.4 atalean azalduko ditugunak). Esate baterako, ingeleseko *he* eta *helium*-en arteko nahasketa subjektu hautapen-murriztapenetan *chemistry-substance* eta *physics-substance* bezala ageri da. Beste adibide bat, leku-izen bereziak (*Argentina, Madrid...*) – *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.

4.4 Erroreen azterketa

Eskurapenean 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, horiek guztien berri emateko, atal hau sortu dugu.

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.

4.4.1 Etiketatze-erroreak

Errore mota hau SemCor corpusean bakarrik gertatu da, hau baita erabili dugun corpus etiketatu bakarra. Eskuz etiketatutako corpusa izan arren, etiketatze-erroreak gertatu ohi dira. Esate baterako, SemCor-en *play 00605818* eta *play 00610422* (ikus 6. irudiko glosak) ez dituzte bereizi, hau da, *play* aditzaren agerpen guztiak *play 00605818 synset*-arekin etiketatuak daude. Hortaz, (56) bezalako perpausak, nahiz eta berez *play 00610422*ren adibide bat izan, *play 00605818*rekin etiketatua dago.

(56) SMU will play **the Owls** at Rice Stadium in Houston.

Nahasketa horrek objektuaren hautapen-murriztapenetan ondorioak izan ditu. *Play 00605818*ren objektuen arten *person* eta *group* ageri zaizkigu, [+gizaki] tasuna daramatenak, hain zuzen ere. Objektu mota hauek *play 00610422*ren hautapen-murriztapenak izan beharko lukete.

Etiketatze-erroreak ez dira aditzekin bakarrik gertatzen, izenenekin ere gertatzen dira.

(57) Our interior **line** and out linebackers played exceptionally well.

(58) For a serious **young man** who plays golf with a serious intensity.

(57)ren kasuan *line* izena *linebacker*-en (futbol jokalaria) laburdura bat da, eta *a formation of people* (pertsonen errenka) adierarekin etiketatua dago.

(58)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 hautapen-murriztapenetan ez dute eragin handirik izan. Beraien hiperonimoak *group* eta *person* direnez, makinak hautapen-murriztapen horietan bilakatu ditu; urre-patroian zuzentzat definitu ditugunak.

4.4.2 Falta diren adierak

Hautapen-murriztapenak EuroWordNet-en oinarrituta adierazi ditugu (corpuseko izenen *synset*-en hiperonimoak edota domeinu eta eremu semantikoak erabilita). Gerta liteke EuroWordNet-en adiera-inbentarioan baten bat ez egotea. Esate baterako, *ball football, basketball...* bezala uler daiteke ingelesez, ekintza bat bezala, alegia:

(59) I play football/basketball/ball...

EuroWordNet-en kontsultatuz gero, *ball synset* ugaritan dago baina horietako batek ere ez du ekintza-adiera hori. SemCor etiketatzerakoan, antzekoena izan zitekeen beste *synset* batekin etiketatu behar izan zuten.

(60) 02103632 ball "round object that is hit or thrown or kicked in games"

Makinak corpusean *ball* izena *02103632* bezala topatzen badu *play 00605818*ren objektu gisa, honen hautapen-murriztapena eskuratzeko zuzenean hiperonimora joko du, eta {*sport, recreation*}-en (edo *sport-act* domeinu-eremu semantikoaren) ordez, *object* (*artifact* eremu semantikoa) objektu hautapen-murriztapena lortzen du.

EFE eta BNCn, etiketaturik ez dauden corpusetan, antzeko prozesua gertatzen da. Makinak corpusean *ball* izena topatzen duenean *play 00605818*ren objektu gisa, eta honen hautapen-murriztapena eskuratu behar duenean, EuroWordNet-etik *ball* ekintza adierazten duen horren ordezko bat hartzen du, *object* adiera duena hain zuzen ere. Hala, honen hiperonimotik abiatuta *object* (edo *artifact* eremu semantikoa) objektu hautapen-murriztapena lortzen du, berez dagokion {*sport, recreation*}-en (edo *sport-act* domeinu-eremu semantikoaren) ordez.

Antzeko beste adibide bat, leku-izen bereziak dira (*Argentina, Madrid, ...*). EuroWordNet-en leku izan berezi bezala bakarrik daude, baina corpusean hauekin kirol taldeak adierazi nahi dira. Hori dela eta, *location* edo *geography-location* bezalako hautapen-murriztapenak ditugu *play 00605818* aditzarekin.

4.4.3 Anbiguotasuna

Gure ustez, hau izan daiteke hautapen-murriztapenen eskuratzean gehienetan gerta daitekeen fenomenoa, etiketatu gabeko corpusen gainean aritzean, noski. Baina, errore hau antzematen zailenetakoa da.

Corpuseko izenek adiera bat baino gehiago izan dezakete, eta etiketatu gabe daudenean, eskuratze-teknikak adiera horietako bat aukeratu behar du EuroWordNet-etik. Gerta daiteke dagokion adiera aukeratzea, eta ondorioz, zuzena ez den hautapen-murriztapena sortzea. Esate baterako, ingeleseko *game* izenak bost adiera ditu EuroWordNet-en:

(a) *00254052 game1 "a contest with rules to determine a winner"*

(b) *00254326 game2 "a single play of a game; the game lasted 2 hours"*

(c) *00256308 game3 "an amusement or pastime"*

(d) *01485683 game4 "animal hunted for food or sport"*

(e) *00341531 game5 "informal terms for your occupation; he's in the plumbing game"*

Kirol adierak lehenengo biak izan daitezke (a eta b), baina horietaz gain beste adiera batzuk ditu. 4.3.4 atalean aztertutako hautapen-murriztapenen artean *zoology-group* eta *zoology-animal* bezalakoak genituen, eta okerrak bezala ebaluatu ditugu. Horien atzean anbiguotasunaren arazoa dago, makinak *game* izena *game4* bezala desanbiguatu du (animalia bezala, alegia), eta ondorioz, *synset* horren hautapen-murriztapenak lortu dira.

4.4.4 *Parser*-ak eragindako erroreak

3.2.1. atalean ikusi dugun bezala, aditz baten hautapen-murriztapenak eskuratzeko, lehenengo corpusaren gainean Minipar analizatzailea edo *parser*-a (Lin, 1993) erabili behar da. *Parser*-ak errore batzuk izan ditzake, eta ondorioz, honek hautapen-murriztapenetan eragina izan du. Honen adibide argi bat da *play 00605818*ren (61)eko subjektuaren hautapen-murriztapena; (62)n honi dagokion (SemCor) iturria dugu:

(61) 08413915 0.032 digit "one of the elements that collectively form a system of numbers"

(62) **Nine of the league's teams** play in baseball parks and therefore...

Subjektuaren burua ez da *nine*, baizik eta *teams*, baina *parser*-ak digitua hartu du burutzat, eta horregatik dugu honen hiperonimoa subjektu hautapen-murriztapen gisa.

4.4.5 Izen berezien ezagutza eta anaforaren ebazpena

Bi errore hauek eragotziko lirakete 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 EuroWordNet-eko lotura egin daitekeelarik.

Anaforak berarekin informazio linguistiko asko darama, baina hau ezin da eskuratu baldin eta corpus batean etiketaturik ez dauden. Aipatu dugu subjektuaren hautapen-murriztapen batzuetan agertutako *helium* (elementu kimikoa) eta *he* (hebrear alfabetoaren bosgarren letra), ingeleseko *he* izenordainari dagozkiola. EuroWordNet-en ez daudenez izenordainik, makinak izenordain hori idazkera antzekoa duten beste bi *synset*-ekin parekatu du. Hortik, hautapen-murriztapen okerrak izatea. Anafora automatikoki landua izanez gero, horrelakorik ez genukeen izango, eta anaforaren aurrekariaren informazioa jaso ahal izango genuen.

4.5 Emaitzen azterketa

*Play 00605818*n 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 hautapen-murriztapenak eman dituzten ere aztertu dugu. Ebaluazio honen laburpenaren berri 1. taulan ematen dugu, hau da, corpus bakoitzean erabili den eskuratze-teknika bakoitzetik *play 00605818*ren zenbat objektu/subjektuen hautapen-murriztapen diren zuzenak (urre-patroiarekin bat datozenak), zenbat diren onargarriak (urre-patroiaren hiperonimo edo hiponimoak direnak) eta urre-patroikoetatik zenbat ez diren eskuratu (eskuratu gabe bezala izendatu ditugunak)[25]. Datu hauek kopuru zehatzak erabiliaz adierazi ditugu; esaterako, eskuratze-teknika bakoitzaren objektu/subjektuen hautapen-murriztapenetatik (gehienez hamar) zenbat diren zuzenak eta onargarriak zenbakitu ditugu; eta baita eskurapen-teknika bakoitzarentzat proposatutako urre-patroietatik zenbat geratu diren eskuratu gabe ere. Horrelako taula bat egin dugu esperimentu honetan erabilitako kirol aditz bakoitzarentzat,

[25] Domeinu-eremu semantiko bikoteen ebaluazioan erabilitako irizpide nagusia 4.2 atalean aipatu dugu. Honekin batera, eskuratu gabeak diren ala ez neurtzeko, beste irizpide batzuk finkatu ditugu: batetik, zuzen/onargarri bezala ebaluatutako hautapen-murriztapen batekin, bi urre-patroi eskuratu daitezke. Adibidez, *play 00605818*ren 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* hautapen-murriztapenarekin. Eta bestetik, alderantziz ere gerta daiteke, onargarritzat jo dugun hautapen-murriztapena eskuratu gabea bezala ebaluatzea; esate baterako, izen-bereziak (x baten bidez adieraziak datozenak), pronominalak (*pro* baten bidez adieraziak datozenak), eta *factotum-Tops* bikotea.

hots, EuroWordNet-etik aukeratutako zortzi *synset*-entzat (ikus 4. atalaren sarrera)[26]. 1. taularen antzeko eredua jarraituta, ingeleseko aditz guztiak kontuan hartuta lortu diren emaitzak ditugu 2. taulan, oraingoan ehunekotan adierazi ditugularik. Eskuratu gabeen zerrendan datu azpimarragarriena %0 zenbakira hurbiltzen dena, honek eskuratze-teknikak urre-patroiko hautapen-murriztapen guztiak lortu dituela esan nahi duelako. Zuzen eta onargarrien zerrendan, aldiz, datu nabarmenenak %100era gerturatzen direnak dira, eskuratze-teknikak eskuratutako hautapen-murriztapen guztiak zuzenak/onargarriak direla adierazten duelako. Bestalde, emaitzek adierazten dutena ulerterrazagoa egitearren, zuzenak/onargarriak kopuruen batura ere adierazi dugu eta taulan *Batura z/o* bezala izendatu dugu. Taula hauek aurrean izanda, hurrengo atalean, hauetatik ondoriozta ditzakegun emaitzak komentatuko ditugu.

		Objektua			Subjektua		
Iturria	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	0	4tik 3	0	0	2tik 2
EFE (kirola)	w2semf	10etik 4	10etik 1	0	0	10etik 4	2tik 1

1. taula: Corpus ezberdinetatik *play 00605818*rentzat eskuratutako hautapen-murriztapenen emaitzak.

		Objektuak				Subjektuak			
Iturria	Teknika	Zuzena	Onargarria	Batura z/o	Eskuratu gabe	Zuzena	Onargarria	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	%61,6	%68,6	%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 (kirola)	w2semf	%14,1	%16,4	%30,5	%40,9	%2,7	%38,4	%41,1	%36,3

2. taula: Kirol aditz guztientzat corpus eta eskuratze-teknika ezberdinak erabiliaz, lortutako emaitzak.

[26]Taula hauek guztiak eranskinetan daude ikusgarri.

4.5.1 SemCor-etik eskuratutako hautapen-murriztapenak

Corpus honetatik hiru hautapen-murriztapen mota jaso ditugu:

(a) **w2c:** Eskuratze-teknika honek aditz-forma kontuan hartzen duenez, zehazten zaila da zer hautapen-murriztapen diren kirolaren domeinuari dagozkionak. Urre-patroiarekin bat etorri direnak kontsideratu ditugu domeinu horretakoak eta urre-patroietatik gutxi geratzen dira eskuratu gabe.

(b) **c2c:** Emaitzak aurrekoaren (w2c-en) antzekoak badira ere, eta kontuan izanda etiketatutako corpusa dela, ez dira espero genituen emaitzak, c2c hautapen-murriztapen gehienak okerrak baitira. Honen arrazoia corpuseko etiketatze-erroreetan, *parser*aren analisi okerrean, eta corpusean agertu diren baina EuroWordNet-en ez dauden adieretan egon daiteke. urre-patroietatik bat 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 00605818*ren kasuan, honek *bet on*, *parlay* eta *stake* bezalako troponimoak ditu, hots, apustua domeinuarekin zerikusia dutenak. Hauen hautapen-murriztapenak *play 00605818*rekin zeharo ezberdinak dira, esate baterako, aditz hauen objektu arruntenetako bat dirua izango da (*cost* hautapen-murriztapenetan). Beraz, ez dirudi aditz batek eta bere troponimoek hautapen-murriztapen berdinak dituztenik (behintzat EuroWordNet hierarkian oinarritzen bagara). Bestalde, aipagarria da eskuratze-teknika honek subjektuekin eman dituen emaitza onak, eskuratu gabe %3,5a bakarrik utzi baitu. HOnen arrazoia corpus etiketatu bat dela da, hau da, entitateak landuta eta etiketatuta daude, eta eskuratze-teknikak ez ditu desanbiguatu behar.

(c) **s2semf:** Hautapen-murriztapen hauek domeinu-eremu semantiko bikoteekin definitua datorrenez, eta hitzak domeinu edo eremu semantiko bat baino gehiago izan ditzakeenez, batzuetan zaila da zehazten zein hautapen-murriztapen diren hauen iturria, eta ondorioz, ezinezkoa zaigu zuzenak diren ala ez jakitea. Hori dela eta, eskuratze-teknika honen ebaluazio subjektiboago bat egin dugu. 1. taulako emaitzei erreparatuz, aurreko biak baino hautapen-murriztapen hobexeagoak lortzen dituela esan genezake. 2. taulan, aditz guztiak kontuan hartuta, ezberdintasuna ez da horrenbestekoa: zuzen eta onargarrien batura altua (%57 eta %68,6) da baina baita eskuratu gabeena ere (%64,2 eta %60).

4.5.2 BNCtik eskuratutako hautapen-murriztapenak

Etiketatu gabeko corpus honen gainean w2c eta c2c eskuratze-teknikak erabili ditugu berriro ere. Beraz, eskuratze-teknika hauek bi corpus ezberdinetan erabili dira (SemCoren eta BNCn), emaitza zeharo ezberdinak emanik.

(a) **w2c:** Teknika honen hautapen-murriztapenak, aditzaren adiera guztietan oinarritzen denez, zer adierari dagokien asmatzen oso zaila da, baita hauen iturria aurkitzea ere. Honenbestez, BNCren gainean aplikatuta hautapen-murriztapen batzuk lortu ditu (eskuratu gabeen kopuru txikiena honek du), baina hauek SemCor-en gainean lortutakoak baino kalitate baxuagoa dute.

(b) **c2c:** Teknika honek espero baino emaitza okerragoak eman ditu, *play 00605818*ren hautapen-murriztapen bakarra asmatu baitu, eta beste aditz guztiekin ere hala-moduzko emaitzak izan ditu (ikus 2. taula). Corpusarengatik izan daiteke (etiketatua ez egotea, kirol domeinuarena bakarrik ez izatea...). 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.

4.5.3 EFEtik eskuratutako hautapen-murriztapenak

Kirol domeinuko eta etiketatu gabeko corpus honetan w2semf eskuratze-teknika erabili da.

(a) **w2semf:** Nahiz eta hautapen-murriztapen hauek aditzaren adiera guztientzat izan, hauek emaitza onak lortu dituzte. SemCor-eko w2c eta c2c-ekin alderatuz, w2semf-en zuzen/onargarrien batura zertxobait txikiagoa da (%30,5 eta %41,1), baina kontuan izanda etiketatu gabeko corpusa dela, azpimarratu beharreko emaitzak dira. Corpusaren domeinuak (kirola) beste adierak baztertzen lagundu duela dirudi.

4.5.4 Hautapen-murriztapenen erkaketa

1. eta 2. tauletatik abiatuta, batetik eskuratze-teknika erkatuko ditugu, eta bestetik corpusak.

Eskuratze-teknikaren arabera

(a) **w2c eta c2c:** Emaitzei erreparatuz, w2c-ekin eskuratutako hautapen-murriztapenak *play 00605818*ren argumentuekin hobeto egokitzen dira, c2c-ekin eskuratutakoak baino. Hala ere, hauek ez dute informazio gehiegirik ematen, hautapen-murriztapen hauek aditz-formarentzat baitira.

(b) **w2semf/s2semf eta c2c/w2c:** s2semf eta w2semf-en hautapen-murriztapenak zailak dira beste biekin erkatzeko. SemCor-eko corpusean s2semf-ek beste bi eskuratze-tekniken emaitzak baino hobexeagoak eskaintzen dizkigu, baina, esan dugun bezala, eskuratu gabekoen ehunekoa oso altua da (%64,2 eta %60). Bestalde, EFEko corpusaren gainean, kontuan izanda etiketatu gabeko corpusa dela, w2semf hautapen-murriztapenak dira eskuratze-tekniketatik hoberenak. Baliteke, corpusari esker izatea, EFE corpusa kirol domeinuari baitagokio. Hala ere, w2c-ekin gertatzen den antzera, hautapen-murriztapen hauek ez dute informazio gehiegirik eskaintzen, aditz-formarentzat baitira.

Corpusaren arabera

(a) **BNC eta SemCor corpusen erkaketa:** SemCor-en gainean erabilitako w2c eta c2c eskuratze-teknikek, BNCn baino emaitza hobeak lortu dituzte. Hortaz, kalitate aldetik, SemCor-eko hautapen-murriztapenak hobeak dira. Hala ere, desberdintasun handiagoa espero genuen, SemCor semantikoki etiketatutako corpusa dela kontuan hartuz. Honen arrazoia corpusen tamaina da, hau da, SemCor corpus txikia da BNCkin parekatuta, eta hori dela eta:

- SemCor-en aditz bakoitzeko agerpen gutxiago daude eta ondorioz, eskuratze-teknikek ezin dituzte hautapen-murriztapen batzuk ikasi, hau da, urre-patroi batzuk eskuratu gabe geratzen dira.

- BNCn eskuratze-teknikak agerpen gehiagotan oinarritu daitezke, eta horrela, urre-patroi gehiago eskuratzen dira. Dena den, BNC etiketatu gabeko corpusa izaki, hautapen-murriztapen hauen kalitatea ez da SemCor-ekoa bezain ona.

(b) **EFE:** Corpus honetatik lortu dira emaitzarik onenak. Baliteke, corpusari esker izatea, EFE corpusa kirol domeinuari bakarrik baitagokio. Domeinu jakin batekin lan eginda, aditzaren adiera eta bere hautapen-murriztapenena corpusaren domeinutik lortu daitekeela deritzogu. Dena den, hau gehiago aztertu beharrekoa litzateke, kasuistika handia egon baitaiteke. Aditz batzuk domeinu batekiko harreman gehiago dute beste batzuk baino. Horren adierazgarri, esperimentu honetako ingeleseko *meet* eta *equalize* aditzekin lortutako emaitzak dira[27]. Nahiz eta EFE kirol corpusera mugatu, badirudi aditz hauen beste adierek indar edo erabilera handiagoa dutela. Beraz, ikusteko dago domeinua aditz jakin batzuekin bakarrik erabiltzea baliagarria den ala aditz guztietara orokortu daitekeen.

4.5.5 Domeinu-eremu semantiko bikoteen ebaluazioa

Domeinu-eremu semantiko bikote batekin adierazitako hautapen-murriztapenen ebaluazioan zenbait arazo izan ditugu. Behin baino gehiagotan adierazi dugu mota honetako hautapen-murriztapenak ez direla ulerterrazak. Batetik, domeinuen eta eremu semantikoen informazioa *synset*-ena baino orokorragoa delako, eta gehienetan EuroWordNet-era jo behar dugulako hauen azpian zer *synset* jasotzen diren jakiteko. Eta bestetik, hitz berak domeinu eta eremu semantiko bat baino gehiago har ditzakeelako (4.3.1 atalean ikusi dugun bezala).

Honi gehitu behar zaio, batzuetan zaila dela hautapen-murriztapena adierazten duen *synset* edo domeinu-eremu semantiko bikote "egokia" aukeratzea, gerta litekeelako *synset* edo domeinu-eremu semantiko bikote hori orokorregia izatea (hierarkian goregi egotea) eta zehatzegia izatea (hierarkian beheregi egotea)[28]. Eta arazo honi aurre egiteagatik, hautapen-murriztapenak ebaluatzeko garaian, maila desberdineko labelak erabili ditugu:

(a) **Zuzena:** Urre-patroiarekin bat datorrenean.

(b) **Onargarria:** Urre-patroiaren hiperonimoa edo hiponimoa denean. Domeinu-eremu semantiko bikoteen bidez adierazitako hautapen-murriztapen kasuan, onargarri bezala kontsideratu ditugu urre-patroia baino orokorrago edota zehatzago direnak.

(c) **Okerra:** Urre-patroiarekin bat ez datorrenean eta EuroWordNet-eko hierarkian ere loturarik ez dutenean.

Label hauek ez digute inolako arazorik eman *synset*-ekin adierazitako hautapen-murriztapenak ebaluatzerakoan. Haatik, domeinu-eremu semantiko bikoteekin adierazitako hautapen-murriztapenak ebaluatzeko, bikote bat onargarria ala okerra den erabakitzeko zailtasunak izan ditugu. Eta horretarako, irizpide nagusi bat finkatu dugu (ikus 4.2 atala).

[27] Eranskinetan aditz guztien emaitzak daude.
[28] Azalpen gehiagorako ikus 4.2 atala.

Orain arteko emaitzak irizpide honen arabera lortutakoak dira, eta domeinu-eremu semantikoak erabiltzen dituzten eskuratze-tekniken emaitzak beste eskuratze-tekniketakoekin duten aldea ikusita, 4.2 atalean zehaztutako irizpidea malguegia ez ote zen bururatu zaigu. Horrela, berriro egin dugu ebaluazioa, irizpide zorrotzago batean oinarrituz, eta emaitzak nahiko ezberdinak izan dira (ikus 3. taula).

Bi irizpideen artean aldaketa txiki bat bakarrik egon da, domeinu-eremu semantiko bikoteei dagokiena: bikote bat onargarritzat hartuko dugu, urre-patroia baino orokorrago edota zehatzago bada, **eta domeinuko beste izen gehienak aditz horren argumentu izan badaitezke**. Ikus dezagun adibide bat: aurreko irizpidearen arabera, *administration-person* adibidez, onargarritzat hartzen genuen, urre-patroiaren (*person-person*) aldaera bat zelako. Oraingo irizpidearekin, ordea, okerra da. Irizpide berriaren arabera, zuzentzat hartuko ditugu, urre-patroia baino orokorrago edota zehatzago diren hautapen-murriztapenak, baldin eta domeinuko beste izen gehienak aditz horren argumentu izan daitezkeen. Kasu honetan *administration* domeinuaren azpian EuroWordNet-eko *chairman*, *administrator*, *chancellor* eta abar bezalakoak daude sailkatuak; eta hauek ezin dute *play 00605818*ren hautapen-murriztapenak 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 behar dugu[29].

		Objektuak				Subjektuak			
Iturria	Teknika	Zuzena	Onargarria	Batura z/o	Eskuratu gabe	Zuzena	Onargarria	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 (kirola)	w2semf	%14,1	%10	%24,1	%45,4	%2,7	%21,8	%24,5	%41

3. taula: Kirol aditz guztientzat corpus eta eskuratze-teknika ezberdinak erabiliaz, lortutako emaitzak. Ebaluatzeko irizpide zorrotzagoa erabilita.

3. taulan ikus daitekeen bezala, domeinu-eremu semantikoetan oinarritzen diren eskuratze-tekniken (w2semf eta s2semf) kalitatea asko jaitsi da, batez ere sibjektuen hautapen-murriztapenena, eta honek 4.5 atalean azaldutako ondorioetan eragina du. Domeinu-eremu semantikoetan oinarritutakoen hautapen-murriztapen zuzen eta onargarrien kopurua desente txikitu da eta eskuratu gebeena handitu. SemCor eta BNCen gainean erabilitako teknikak (c2c eta w2c, hurrenez hurren) dira hautapen-murriztapen gutxien eskuratu gabe utzi dituztenak: objektuen hautapen-murriztapenetan BNCko w2c (%15,9) eta SemCor-eko w2c (%29,5) dira emaitzarik onenak, eta subjektuen hautapen-murriztapenetan SemCor-eko c2c (%3,5) eta BNCko w2c (%13,6). Datu hauek hasierako susmoekin bat datoz:

[29]Eranskinetako ebaluazioa azkeneko irizpide honen arabera egina dago.

(a) SemCor corpus etiketatua izanda, besteak baino emaitza hobeagoak izan behar zituela (hala ere, espero baino emaitza kaxkarragoak lortu dira.)

(b) BNC corpus handiena izaki, eskuratu gabe oso hautapen-murriztapen gutxi geratu behar zirela.

Hortaz, domeinu-eremu semantiko bikoteekin adierazitako hautapen-murriztapenen emaitzak oso aldakorrak dira ebaluatzeko irizpideen arabera. Objektibotasun gabezia eta *synset*-ekin parekatzeko duten zailtasuna kontuan izanda, esperimentu honetatik abiatuta aurrerantzean egingo diren beste lanetan, domeinu-eremu semantiko bikoteekin adierazitako hautapen-murriztapenak alde batera utziko dira.

5 EUSKARAKO HAUTAPEN-MURRIZTAPENEN AZTERKETA

4. atalean ikusi ditugun hautapen-murriztapenak, euskararentzat baliagarriak izan daitezkeen aztertuko dugu atal honetan. Hau da, ingeleseko hautapen-murriztapenak eskuratzeko erabili diren teknika ezberdinak aurkeztu eta ebaluatu ondoren, orain hauen aplikazioa eleanitza izan daitekeen ala ez aztertu nahi dugu (azterketaren bukaeran, ikusiko dugu benetan aplikazio eleanitza posible dela).

Horrela, batetik, ingeleseko zortzi *synset* horientzat ikasitako hautapen-murriztapenak *synset* horietako euskarako ordainentzat berrerabiliko ditugu, euskararentzat erabilgarriak diren ala ez ikusteko. Berrerabilpenerako ez dira eskuratze-teknika guztietako hautapen-murriztapenak hartu, esperimentu hau hastapenekoa izaki, honen emaitzak ikusteko lagin bat erabiltzearekin nahikoa dela iruditu zaigu. Ingelesetik euskarara zuzenean itzuli behar genituen hautapen-murriztapenak aukeratzerakoan bi irizpide hauetan funtsatu gara:

(a) **SemCor-etik ikasitako hautapen-murriztapenak izatea, eta gainera, aditz-adiera bakarrari egokitzea.** Horrela, EuroWordNet baliatuta, zuzenean itzul ditzakegu euskarara bai ingeleseko corpuseko hitzak (*synset*-ekin etiketatutakoak), eta bai hautapen-murriztapenak (*synset*-ekin adieraziak); EuroWordNet-eko *synset*-a abiapuntu izanda, hortik zuzenean beraien euskarako ordainera pasa gaitezkeelako eta horrek itzulpen lana errazten duelako. SemCor da erabili dugun corpus etiketatu bakarra, eta honen gainean aditza-adiera hautapenak ikasteko, c2c eta s2semf eskuratze-teknikak aplikatu dira.

(b) **Domeinu corpus bateko hautapen-murriztapenak erabiltzea (gure kasuan, EFE)**. Honetatik lortutako hautapen-murriztapenak beste corpus orekatuetakoekin parekatzea interesgarria iruditzen zaigulako, batez ere, ingeleserako emaitzak onenak hemendik lortutakoak direla ikusi ondoren[1]. EFE gainean w2semf eskuratze-teknika erabili dugu.

Hala, guztira, ingeleseko c2c, s2semf eta w2semf hautapen-murriztapenak berrerabili ditugu euskararako.

Eta bestetik, w2semf eskuratze-teknika euskarako corpus batean erabili dugu. Eskuratze-teknika hau aukeratu dugu, momentuan inplementatzeko sinpleena zelako eta honekin ingeleserako emaitza onak[2] lortu direlako. Horrela, honen ingeleserako eta euskararako

[1]Irizpideen alderaketa esperimentu hau guztia amaitu ondoren egindakoa da. Horregatik, euskararen azterketa hasi behar genuenean, ariketa honetarako w2semf iruditu zitzaigun eskuratze-teknikarik egokiena, oraindik domeinu-eremu semantiko bikoteak ebaluatzeko irizpide bakarra genuelako. Honezaz gain, euskarako hautapen-murriztapenen ebaluazioa irizpide honen arabera egin dugu (ikus 4.2 atala). Hala ere, ingelesekoekin bezala, irizpide zorrotzagoa erabilita lortutako emaitzen berri ere emango dugu amaieran (5.6.4 atalean).

[2]Ikus aurreko oin-oharra.

emaitzak baliatuz, euskarari zer 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 ikasteko aukera ematen digu. Hortaz, euskarako hautapen-murriztapenak kirol domeinuan oinarritutako corpusetik lortu ditugu. Hala ere, kirol domeinuarekin erabilitako eskuratze-teknika bera erabili dugu corpus osoaren gainean, hau da, domeinurik zehaztu gabe. Hala, eskuratze-teknika bera erabili da kirol domeinua duen corpus baten gainean eta domeinurik gabeko corpus batean. Emaitzek domeinuaren eragina zenbaterainokoa izan daitekeen aztertzen ahalbidetuko digu.

Aurrera jarraitu baino lehen, azpimarratu behar da, ingeleseko hautapen-murriztapenak euskarara itzultzen, argumentuekin zenbait zailtasun sumatu ditugula, hots, ingeleseko argumentuak ezin dira zuzenean euskara itzuli. Hala nola, ingeleseko subjektuak eta objektuak nominatiboan eta akusatiboan agertzen badira ere, hurrenez hurren, euskaraz ez da horrela gertatzen. Euskarako subjektuak ergatibo edo absolutibo markak eraman ditzake, eta objektuak absolutiboaz gain, beste hainbat kasu-marka ere (inesiboa adibidez)[3].

Euskarako hautapen-murriztapen hauen guztien azalpenerako, ingelesekoekin bezala, *00605818 play1 / jokatu2; "play games, play sports" synset*-eko euskarako ordainean (*jokatu 00605818*n) oinarrituko gara.

5.1 Euskarako urre-patroiak

Eskuratze-teknika desberdinen hautapen-murriztapenak ebaluatzeko, EuroWordNet-etik aukeratutako kirol aditz bakoitzeko urre-patroi batzuk zehaztu dira, kasu honetan *jokatu 00605818*rentzat. Bestalde, urre-patroiak eskuratze-teknika bakoitzaren ereduan sortuko dira, hau da, guk sortutako urre-patroiek teknika hauen emaitzek hartzen duten itxura hartuko dute. Hala, euskarako azterketan, alde batetik, hautapen-murriztapenak adierazteko *synset*-ean oinarritzen den teknika dugu (c2c), eta bestetik, domeinu-eremu semantikoetan oinarritzen direnak (w2semf eta s2semf).

Euskarako *jokatu 00605818*ren urre-patroiak proposatu ahal izateko corpusetan oinarritu gara. Corpusetik hartutako perpausetatik, aztertu beharreko aditz-adiera bakoitzaren jokaera linguistikoa orokortzen saiatu gara, gerora, orokortasun horiek (hautapen-murriztapenak, alegia) EuroWordNet-eko *synset* eta domeinu-eremu semantiko batzuen bidez adierazteko. Corpuseko izen bat hautapen-murriztapen batean orokortzeko, gehienetan izen horrek EuroWordNet-en duen hiperonimoetara jo dugu[4]. Euskaldunon Egunkaria etiketatu gabeko corpusa denez, lehendabizi *jokatu* aditzaren *jokatu 00605818* kirol adiera besteetatik (*zuzen jokatu, pastorala jokatu,* eta abar bezalakoetatik) desberdindu behar genuen.

Corpusean ikusitakoaren arabera, *jokatu 00605818* aditzak *lehiaketa, txapelketa* eta abar bezalako objektuak hartzen ditu, orain arte hautapen-murriztapenetan {*contest, competition*} bezala agertzen dena[5]:

[3]Hurrengo atalean aztertuko ditugu honi buruzko adibide batzuk.

[4]Askotan aipatu dugu, orokortzen duen hiperonimo egokia bilatzea nahiko lan arriskutsua dela, hiperonimo batzuekin aditza har ditzakeen argumentuen esparrua gehiegi zabaldu edo murriztu daitekeelako.

[5]04771851 *synset*-ean {*contest, competition*} izenak daude, eta *synset* bereko euskarako ordainak *lehiaketa, txapelketa* dira. Orain arte hautapen-murriztapenak ingelesez eman ditugu, eskuratze-tekniken emai-

(63) **OBJEKTUA**
(a) Sidneyko **Joko Olinpikoak** jokatuko baitira irailaren 14tik urriaren 1a bitartean.
(b) Aste Santuan jokatuko da **Euskal Herriko txapelketa**.
(c) Klub Arteko **Munduko Txapelketa** jokatuko da Brasilen.
(d) **Euskadiko Kopako finalerdia** jokatuko du Zarautzen.

Joko Olinpikoak eta *finalerdia* izenak {*contest, competition*} *synset*-aren hiponimoak dira, beraz, hiperonimoaz baliatu gara *jokatu 00605818*ren objektuak orokortu ahal izateko.

Subjektuen kasuan, taldeak eta pertsonak izan dira nagusi:

(64) **SUBJEKTUA**
Taldea:
(a) **Realak** datorrean asteazkenean jokatu behar duten partidua...
(b) **Kataluniako Eskubaloi Selekzioa** jokatu gabe geratu zen...
(c) Adiskidantzazko partidu gehiago jokatuko ditu **Bidasoak**.
(d) Bestalde, hilak 14ean, hiruko torneoa jokatuko du **Bidasoak** egunean Bermeon.

(65) **SUBJEKTUA**
Pertsona:
(a) Gutxienez bi partidu egongo da **Rider** jokatu gabe.
(b) **Agirresarobe - Iriatek** jokatuko dute.
(c) **Iruk** jokatuko du hasieratik.
(d) **Dmitri Khokhlov errusiarrak** hasieratik jokatutako partidu nagusia.

Ingeleseko *play 00605818*k ez bezala, euskarako *jokatu 00605818* aditzak ez ditu *futbol, golf* eta abar bezalako objektuak hartzen, ez behintzat absolutibo kasuan. Berez, *jokatu 00605818*k argumentu bezala onartzen ditu, baina beste kasu batekin: inesiboarekin.

(66) **OBJEKTUA**
Inesiboa:
(a) **FutboleaN** jokatzen badakitela erakutsi zuten Miguel Angel Lotinaren jokalariek.
(b) Beno, banekien han dena ezberdina zela, **futboleaN** ere han jokatuta bainengoen.
(c) Rafa Alkortak eta Patxi [...] **golfeaN** jokatuko duela dio irribartsu.

Euskarako subjektu eta objektuen argumentuak ergatiboa eta absolutiboarekin agertzeaz gain, beste kasu-marka batzuekin ere ager daitezkeela ikusita, euskarako hautapen-murriztapenen eskuratzea funtzio gramatikaletan oinarritu ordez —ingeleserako egin dugun bezala—, **kasu-marketan oinarrituta** egitea erabaki dugu. Hala, ergatiboen, absoltuiboen, inesiboen... hautapen-murriztapenei buruz jardungo gara.

(67)n ditugu *jokatu 00605818* aditzaren c2c-rako urre-patroiak eta (68)n w2semf eta s2semf teknikentzako:

tzak hizkuntza horretan ematen direlako. Euskaraz ere, eskuratze-tekniken emaitzak ingelesez daudenez ere, bere horretan mantenduko ditugu.

(67) **jokatu 00605818 ABSOLUTIBOA**
c2c:
04771851 contest, competition "an occasion on which a winner is selected from among. . . "
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 competition"

(68) **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, euskaraz ez ditugu objektu/subjektuen hautapen-murriztapenak eskuratu, kasuan oinarritzen diren hautapen-murriztapenak baizik. Ingeleseko eta euskarako "funtzio-kasu" desoreka hau aditz bat baino gehiagorekin gertatu zaigu, esate baterako, *play 00610422*rekin (ikus 6. irudia): ingeleseko *Princeton plays Yale*, euskaraz, *Princetonek Yaleren aurka jokatzen du* itzuliko genuke. Ingeleseko objektua (*Yale*), euskaraz *-en kontra* postposizioarekin adierazten dugu. Horregatik, *play 00610422*ren hautapen-murriztapenak ikasterakoan, ingelesean bezala objektu eta subjektuen hautapen-murriztapenak lortu ordez, euskararako ergatiboaren eta *-en kontra* postposizioaren hautapen-murriztapenetan oinarritu gara.

Desoreka honek ingeleseko hautapen-murriztapenak euskarara itzultzeko zailtasunak sortu ditu, hau da, ingeleseko argumentuak ezin dira zuzenean euskara itzuli, ingelesez funtzio gramatikalei dagozkien hautapen-murriztapenak baitauzkagu eta euskaraz kasumarkei dagozkienak. Hortaz, ingeleseko argumentuak ezin dira zuzenean euskarara itzuli. Horregatik, hautapen-murriztapenen erkaketa egin ahal izateko, bi hizkuntzetako aditzen argumentuak parekatu behar izan ditugu, lehendabizi, aditz bakoitzaren izaera sintaktiko-semantikoa definituz. Oro har, esan dezakegu ingeleseko subjektu hautapen-murriztapenak euskarako ergatibo eta absolutibo hautapen-murriztapenak izango direla — aditz iragankor eta iragangaitzei dagozkienak, hurrenez hurren—, eta ingeleseko objektu hautapen-murriztapenak euskarako absolutiboa. Hala ere, aditz bakoitzaren izaera sintaktiko-semantikoa kontuan izanda objektuen artean bestelako kasu-markak ere egon daitezke, esate baterako, inesiboa.

Bestalde, ingeleseko hautapen-murriztapenekin bezala, urre-patroi hauen arabera hautapen-murriztapenak ebaluatzeko maila desberdinak daude:

(a) **Zuzena:** Urre-patroiarekin bat datorrenean.

(b) **Onargarria:** Urre-patroiaren hiperonimoa edo hiponimoa denean. Domeinu-eremu semantiko bikoteen bidez adierazitako hautapen-murriztapen kasuan, onargarri bezala kontsideratu ditugu urre-patroia baino orokorrago edota zehatzago direnak.

(c) **Okerra:** Urre-patroiarekin bat ez datorrenean eta EuroWordNet-eko hierarkian ere loturarik ez dutenean.

Euskararako ikasitako hautapen-murriztapenak domeinu-eremu semantiko bikoteetan oinarrituak dira, eta hauen ebaluazioa irizpide batzuen arabera egin dugu; ingelesekoekin erabilitako berdinak direnez ez ditugu errepikatuko (ikus 4.5 atala)[6].

5.2 w2semf Euskaldunon Egunkariatik

Eskuratze-teknika hau 3.2.2 atalean azaldu dugu. Aditzaren hautapen-murriztapenak erauzten dituen eskuratze-teknika da eta hautapen-murriztapenak domeinu-eremu semantiko bikote batez adieraziak datoz, bikote hau klase bezala kontsideratzen delarik. Teknika hau corpus osoan (domeinuak kontuan hartu gabe) eta kirol domeinuari bakarrik dagokion zatian erabili da, emaitzek domeinuaren eragina zenbaterainokoa den erakutsiko digutelarik.

Nahiz eta ingeleserako eskuratze-teknika bera erabili, euskararako aldaketa bat egin behar izan zaio: orain ez dira objektu eta subjektu hautapen-murriztapenak, inesibo, absolutibo eta ergatiboen hautapen-murriztapenak dira, hots, kasuei (eta ez funtzioei) dagozkie.

Abiapuntuko metodologia orain arte erabilitakoaren parekoa izan arren (hautapen-murriztapenaren iturria eta corpuseko testuinguruak bilatu, hautapen-murriztapena bera ebaluatzen hasi baino lehen), aurrerago gertatu zaigun bezala (3.2.2), eskuratze-teknika honekin zaila da iturria zein den zehaztea. Batetik, hautapen-murriztapenak aditz-formarentzat direlako eta hauen iturria aztertzeko agerpenak bana-banan berrikusi beharko genituzkeelako. Bestetik, hautapen-murriztapenak adierazteko domeinu-eremu semantiko bikoteak erabiltzen dituen eskuratze-teknika izaki, eredu honen informaziotik iturrira iristeko, nahitaez EuroWordNet-era jo behar dugu domeinu eta eremu semantiko bakoitzaren azpian zer *synset* jasotzen den 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[7]:

[6]Kontuan izan, irizpideen alderaketa esperimentu hau guztia amaitu ondoren egindakoa dela eta orain ikusiko ditugun datuak aztertzeko, domeinu-eremu semantiko bikoteak ebaluatzeko irizpide bakarra genuela, 4.5 atalekoa hain zuzen ere.

[7]Oso zerrenda luzea da, eta hemen lagin bat bakarrik dago. Zerrenda osoa eranskinetan dago.

(69) **w2w.jokatu.corpus_osoa**
jokatu
abs partidu 28
abs partida 26
abs x 19
abs final 12
abs bider 3
abs paper 3
abs uefa 3
abs izan 3
abs jende 3
...
erg pro 204
erg x 9
erg batzuk 7
erg eugi 4
erg 0 3
erg guzti 3
erg beloki 3
...
ine x 23
ine postu 7
ine 0 4
ine talde 4
ine eraso 4
ine zati 3
ine igande 3
ine futbol 2
ine etxe 2
ine adar 2
ine maila 2

Oso zerrenda luzeak dira, eta lan handia litzateke bakoitzaren testuinguruak aztertu eta kirolaren domeinuari dagozkionak baztertzea, gero horren arabera beraien EuroWordNet-eko *synset*, eremu semantiko eta domeinu posibleak zehazteko.

Arrazoi hauek guztiengatik, eta datu enpirikoetan oinarritu gabe, zuzenean EFE gainean aplikatutako eskuratze-teknika hauen hautapen-murriztapenak gure urre-patroiekin (ikus (68)) erkatu ditugu.

(70)n *jokatu* aditzaren w2semf absolutibo (abs), inesibo (ine) eta ergatibo (erg) kirol domeinuko corpuseko hautapen-murriztapenak ditugu (letra lodiz gure ustez *jokatu 00605818* aditzari dagozkienak)[8].

Ebaluazioarekin hasi baino lehen, komeni da gogora ekartzea batetik, eskuratze-tekniken emaitzak izenen domeinu-eremu semantiko bikoteen zerrendak direla (gehienez hamarrekoak), eta beraien formatua hurrengoa dela:

[8]Hautapen-murriztapenen zerrenda oso luzea izan daiteke, eta aditz baten hautapen-murriztapenak hamar baino gehiago direnean, lehenengo hamarrak (probabilitate-neurri handienekoak) bakarrik aztertu ditugu.

w2semf.obj(*eskuratze-teknika eta erlazio sintaktikoa*)
play (*aditza*)
abs sport-event 28
(erlazio sintaktikoa, domeinu-eremu semantiko bikotea eta probabilitatea)

Bestalde, esan beharra dago eskuratze-teknika honek izen bereziak x batez adierazten ditu, *pro* batez anafora pronominalak eta *0* batez elipsiak.

(70) **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 ONARGARRIA
erg geography-person 1 ONARGARRIA
erg administration-person 1 ONARGARRIA
erg basketball-person 1 ONARGARRIA
erg time period-time 0.6
erg cycling-person 0.25 ONARGARRIA

(71)en corpus osoa erabilita lortutako hautapen-murriztapenak ditugu:

(71) **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 ONARGARRIA
erg basketball-person 1 ONARGARRIA

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 hautapen-murriztapenak izan daitezkeen heinean, zuzenak izan daitezkeelako.

Aditzaren adiera guztiak kontuan hartzen dituen eskuratze-teknika izateko, kirolari dagozkion hautapen-murriztapen ugari daude. Urre-patroiko objektuen hautapen-murriztapen guztiak daude eta nahiko probabilitate-neurri altuekin, gainera.

Corpus osoko eta kirol domeinuko hautapen-murriztapenak erkatuz gero, ez dago horrenbesteko alderik bata eta bestearen artean; desberdintasun nabarmenena inesibo hautapen-murriztapenek erakusten dute. Kirol domeinutik ikasitako inesiboaren hautapen-murriztapenetan urre-patroian proposaturiko hautapen-murriztapen guztiak daude: *sport-act, play-act* eta corpus osotik ikasitakoetan hauek baino orokorragoa den *factotum-act* bakarrik dago. Bestalde, kirol domeinuko corpuseko inesiboen hautapen-murriztapenetan, deigarria da *sport-act, play-act* hautapen-murriztapenak 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 00605818*ren argumentuak ez 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 hautapen-murriztapenetako (corpus osoko eta kirol domeinukoak) probabilitate-neurri handienak izen bereziek (*x*) eta anafora pronominalek (*pro*) dute. Esan beharra dago, *transport/administration/geography-person* hautapen-murriztapenekin zalantzak izan ditugula. Nahiz eta lehenengo begiratuan okerrak iruditu, w2w zerrendak — ikus (69)— eta corpusak aztertuz, konturatu ginen hauek ondorengo agerpenetatik zetozela:

(72) a) **Italiarrek** bi jokalari gutxiagorekin jokatu dute.
b) 5 kilometroko erlojupekoa jokatu dute **txirrindulariek**.

Italiar izenaren domeinuak EuroWordNet-en *administration* eta *geography* dira; eta *txirrindulari* izenarena, *transport*. Horregatik ditugu *geography-person*, *administration-person* eta *transport-person* bezalako hautapen-murriztapenak. Haatik, *politics-person* ez dugu hauekin batera onargarri gisa ebaluatu, ergatiboko w2w zerrenda aztertuta errore bat dela ikusi dugulako; w2w zerrendako ergatiboen artean, *politics* domeinua har dezakeen bakarra *defentsa* baita:

(73) a) **Defentsak** ondo jokatu zuen.

Testuingurua zuzena da eta perpauseko *defentsa* izenaren domeinua *sport* da. Hortaz, honen hautapen-murriztapena *sport-person* izan beharko litzateke. Nondik lortu da *politics-person* hautapen-murriztapena? Izen horrek EuroWordNet-en hamar *synset* inguru ditu, eta horietako bat *politics* domeinuari dagokio. Beraz, anbiguotasun errore bat egon da.

Badirudi ingeleseko eskuratze-teknikekin aipatu ditugun errore batzuk euskarako w2semf honekin ere gertatzen direla.

5.3 SemCor-eko c2c euskarara itzulita

4.3.1 atalean azaldutako c2c objektu/subjektuen hautapen-murriztapenak (74)ren errepikatzen ditugu (bakarrik zuzentzat eta onargarritzat jo ditugunak, beraien ebaluazio eta guzti), euskarako *jokatu 00605818* aditzarentzat ere baliagarriak diren egiaztatzeko. Buruan izan, c2c eskuratze-teknikak lortzen dituen objektu edo subjektuen hautapen-murriztapenak aditzaren adiera jakin baterako direla. Beraz, gure kasuan, hautapen-murriztapen hauekin *play 00605818* aditza bakarrik izan beharko dugu kontuan. Hautapen-murriztapen hauek euskaratzerakoan, beraz, *jokatu 00605818* aditza-adierarentzat bakarrik izango dira.

(74) **c2c.obj**
play 00605818
00228990 0.215 activity "any specific activity or pursuit" **ONARGARRIA**
04771851 0.035 contest, competition "an occasion on which a winner..." **ZUZENA**

c2c.subj
play 00605818
00017008 0.517 group, grouping "any number of entities considered as a unit" **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, hautapen-murriztapenen erkaketa egin ahal izateko, bi hizkuntzetako argumentuak parekatu behar izan ditugu: ingeleseko subjektu hautapen-murriztapenak euskarako ergatibo hautapen-murriztapenak izango dira, eta ingeleseko objektu hautapen-murriztapenak euskarako absolutibo eta inesibo hautapen-murriztapenak izango dira[9].

[9]Jakina, parekatze hau aditzaren izaera sintaktiko-semantikoaren araberakoa da.

(75) **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 competition"

SUBJEKTUA:
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"

Euskarako *jokatu 00605818*rentzat proposaturiko urre-patroiak (ikus (76)), ingeleseko hautapen-murriztapenekin guztiz bateragarriak dira:

(76) **c2c.obj**
jokatu 00605818
00228990 0.215 activity "any specific activity or pursuit" **ONARGARRIA**
04771851 0.035 contest, competition "an occasion on which a winner is..." **ZUZENA**

c2c.subj
jokatu 00605818
00017008 0.517 group, grouping "any number of entities considered as a unit" **ZUZENA**
00004865 0.507 person, individual, human "a human being" **ZUZENA**

5.4 SemCor-eko s2semf euskarara itzulita

4.3.1 atalean azaldutako s2semf objektu/subjektu hautapen-murriztapenak (77)n errepikatzen ditugu (bakarrik zuzentzat eta onargarritzat jo ditugunak, beraien ebaluzaio eta guzti), euskarako *jokatu 00605818* aditzarentzat ere baliagarriak diren egiaztatzeko.

Eskuratze-teknika honek aditzaren adiera bakoitzarentzat hautapen-murriztapenak domeinu-eremu semantiko bikoteekin adierazten ditu.

(77) **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 1ONARGARRIA
subj factotum-group 1 ZUZENA
subj factotum-Tops 1 ONARGARRIA
subj person-person 1 ZUZENA

Euskarako *jokatu 00605818*rentzat proposaturiko urre-patroiak (ikus (78)), ingeleseko hautapen-murriztapenekin guztiz bateragarriak dira (ikus (79)):

(78) **OBJEKTUA:**
jokatu 00605818 ABSOLUTIBOA
sport-event
time period-time

jokatu 00605818 INESIBOA
sport-act
play-act

SUBJEKTUA:
jokatu 00605818 ERGATIBOA
factotum-group
person-person

(79) **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 1ZUZENA
subj factotum-group 1 ZUZENA
subj factotum-Tops 1 ONARGARRIA
subj person-person 1 ZUZENA

5.5 EFEko w2semf euskarara itzulita

4.3.2 atalean azaldutako w2semf objektu/subjektu hautapen-murriztapenak (ebaluazio eta guzti) (80)n errepikatzen 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 esperimentu honetarako. Corpus honetan w2semf eskuratze-teknika aplikatu dugu, euskarako hautapen-murriztapenak ikasteko erabili duguna. Teknika honek ikasten dituen hautapen-murriztapenak aditz-formarentzat dira, aditzaren adiera guztientzat, alegia. Gainera, probabilitate kopuru altuenetik baxuenera ordenaturiko domeinu-eremu semantiko bikoteak dira.

(80) **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 administration-group 168.64 ONARGARRIA
subj sport-group 44.01 ONARGARRIA
subj zoology-group 40.5 ONARGARRIA

Euskarako *jokatu 00605818*rentzat proposaturiko urre-patroiak (ikus (81)), ingeleseko hautapen-murriztapenekin guztiz bateragarriak (ikus (82)) dira:

(81) **OBJEKTUA:**
jokatu 00605818 ABSOLUTIBOA
sport-event
time period-time

jokatu 00605818 INESIBOA
sport-act
play-act

SUBJEKTUA:
jokatu 00605818 ERGATIBOA
factotum-group
person-person

(82) **w2semf.jokatu.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 sport-event 23.514 ZUZENA
obj sport-act 23.038 ZUZENA

w2semf.jokatu.kirola.subj
subj x 372 ZUZENA
subj administration-group 168.64
subj sport-group 44.01 ONARGARRIA
subj zoology-group 40.5 ONARGARRIA

Corpusa	Hautapen-murriztapenak	Kasua	Zuzena	Onargarria	Eskuratu gabea
Egunkaria osoa	w2semf	abs	10etik 2	0	0
		ine	0	10etik 1	0
		erg	10etik 1	10etik 6	2tik 1
Egunkaria kirola	w2semf	abs	10etik 1	0	2tik 1
		ine	10etik 2	10etik 1	0
		erg	0	10etik 7	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

4. taula: Euskararako eskuratutako eta ingelesetik itzulitako *jokatu 00605818*ren hautapen-murriztapenen emaitzak.

5.6 Emaitzen azterketa

4. taulak laburbiltzen du euskararako *jokatu 00605818*rentzat ikasitako edo itzulitako hautapen-murriztapenen emaitzen kalitatea. Taulan corpus bakoitzean erabili den eskuratze-teknika bakoitzetik zenbat objektu/subjektuen edo absolutibo/ergatibo/inesiboen hautapen-murriztapen diren zuzenak (urre-patroiarekin bat datozenak), zenbat diren onargarriak (urre-patroiaren hiperonimo edo hiponimo bat direnak) eta urre-patroikoetatik zenbat ez diren eskuratu (eskuratu gabeak deitu duguna). Datu hauek kopuru zehatzak erabiliaz adierazi ditugu; esaterako, eskuratze-teknika bakoitzaren objektu/subjektuen hautapen-murriztapenetatik (gehienez hamar) zenbat diren zuzenak eta onargarriak zenbakitu ditugu; eta baita eskurapen-teknika bakoitzarentzat proposatutako urre-patroietatik zenbat geratu diren eskuratu gabe ere. Horrelako taula bat egin dugu esperimentu honetan erabilitako kirol aditz bakoitzarentzat, hots, EuroWordNet-etik aukeratutako zortzi *synset*-entzat (ikus 4. atalaren sarrera)[10]. 5. taulan euskararako aditz guztientzat ikasitako edo itzulitako hautapen-murriztapenen emaitzak ditugu oro har, oraingoan ehunekotan adierazi ditugularik[11]. Taula honetan zuzenen eta onargarrien kopuruak batu ditugu eta baita taulan adierazi ere (*Batura z/o* zutabean).

Eskuratu gabeen zerrendan datu azpimarragarriena %0 zenbakira hurbiltzen dena, honek eskuratze-teknikak urre-patroiko hautapen-murriztapen guztiak lortu dituela esan nahi duelako. Zuzen eta onargarrien zerrendan, aldiz, datu nabarmenenak %100era gerturatzen direnak dira, eskuratze-teknikak eskuratutako hautapen-murriztapen guztiak zuzenak/onargarriak direla adierazten duelako.

Taula hauek aurrean izanda, hurrengo atalean, hauetatik ondoriozta ditzakegun emaitzak komentatuko ditugu.

[10]Taula hauek guztiak eranskinetan daude ikusgarri.

[11]Taula orokorrean kasu absolutibo eta ergatiboaren datuak bakarrik adierazi ditugu, aditz guztiekin agertu zaizkigunak, hain zuzen ere.

Corpusa	Hautapen-murriztapenak	Kasua	Zuzena	Onargarria	Batura z/o	Eskuratu gabea
Egunkaria osoa	w2semf	abs	%25,7	%31,4	%57,1	%3,5
		erg	%3,7	%76,2	%79,9	%75
Egunkaria kirola	w2semf	abs	%25,7	%37,1	%62,8	%3,5
		erg	%2,8	%76,2	%79	%75
SemCor	c2c	obj	%6,9	%26,4	%33,3	%44
		subj	%38	%7,1	%45,1	%3,5
SemCor	s2semf	obj	%16	%47	%63	%64,2
		subj	%7	%61,6	%68,6	%60
EFE kirola	w2semf	obj	%14,1	%16,4	%30,5	%40,9
		subj	%2,7	%38,4	%41,1	%36,6

5. taula: Euskararako eskuratutako eta ingelesetik itzulitako hautapen-murriztapenen emaitzak.

5.6.1 Euskaldunon Egunkariatik eskuratutako hautapen-murriztapenak

Euskaldunon Egunkariatik ikasitako objektuen (edo euskarazko, absolutiboen) hautapen-murriztapenak ingelesekoen baino hobexeagoak dira, beraien urre-patroi gehienak eskuratu direlako (%3,5). Dena den, datu hau aztertu beharrekoa da, susmoa baitugu euskarako objektua beste kasu-markekin adierazita datorrenean, emaitzak ez direla horren onak. Baliteke, honen arrazoia hauek inplizituki adieraziak datozela izatea, hau da, irakurleak testua ulertzeko beraien beharrik ez duenez testuan argumentu hauek ez azaltzea, eskuratu gabeko urre-patroien kopurua handituz[12].

Hala ere, Egunkariatik eskuratutako hautapen-murriztapen asko onargarriak diren arren, subjektuen kasuan, gehinak (%75) eskuratu gabe geratu dira. Zergatia ez dugu sakonki aztertu baina susmoa dugu hurrengo arrazoiek zerikusia dutela: euskarako corpusaren tamaina txikiagoa delako eta ingeleseko *parser*-a euskarakoa baino hobea delako. Bestalde, aurreprozesuan entitateak ez lantzea izan du eraginik. Ergatiboen hautapen-murriztapenetako gehienak izen-bereziak (x) edo pronominalak (*pro*) dira. Hauek onargarritzat jo ditugun arren, ezin dira urre-patroiekin parekatu, eta ondorioz, ezin ditugu eskuratuak bezala hartu. Arrazoi horregatik, euskarako hautapen-murriztapenetan eskuratu gabeen kopurua asko handitu da.

Bestalde, ingeleseko hautapen-murriztapenekin gertatu ez den bezala, euskaraz corpusa domeinu zehatz batean egoteak ez du aditzaren adiera desanbiguatzen. Corpus osoko eta kirol domeinuko euskarako hautapen-murriztapenen emaitzak oso antzekoak dira. Are gehiago, kasu askotan, kirol corpusean eta corpus osoan, hautapen-murriztapenak berdinberdinak dira, hots, 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 zela pentsatzen dugu.

[12]Honi buruz 5.2 atalean mintzatu gara.

5.6.2 SemCor-etik eskuratutako hautapen-murriztapenak

Corpus honetatik bi hautapen-murriztapen mota jaso ditugu: c2c eta s2semf. Bi eskurapen-teknikek ikasitako hautapen-murriztapenak euskararentzat baliagarriak dira (hautapen-murriztapen zuzenak eta onargarrietaz ari gara). Hortaz, eleaniztasunaren hipotesia egiaztatu egiten da. Hala ere, itzulpena egiterakoan, kontuan izan beharrekoa da bi hizkuntzetan argumentuak ez direla kasu berdinekin 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 etiketatutako corpusa dela, emaitza hobeak espero genituen. Corpusaren tamaina (hau erabilitako corpus txikiena dugu) eta etiketatze-erroreak izan daitezke zergatiak.

5.6.3 EFEtik eskuratutako hautapen-murriztapenak

Corpus honetatik hautapen-murriztapen mota bakarra erabili dugu: w2semf. Bai ingelesez eta bai euskaraz, hauek emaitza onak lortu dituzte. SemCor-eko c2c-ekin alderatuz, w2semf-en zuzen/onargarrien batura handiagoa da, eta gainera, kontuan izanda etiketatu gabeko corpusa dela, azpimarratu beharreko emaitzak dira. Corpusaren domeinuak (kirola) beste adierak baztertzen lagundu duela dirudi[13].

Oro har, emaitzei erreparatuz, w2semf teknikak eskaintzen dizkigu emaitzarik onenak, bai ingelesez eta bai euskaraz, batez ere, objektuei dagokienak. SemCor-eko c2c eskuratze-teknikaren subjektuen hautapen-murriztapenak azpimarragarriak dira. Hala, teknika hauen arteko ebakidura eginez gero, bai subjektu eta bai objektuentzat emaitza hobeak lortuko genituzke.

Amaitzeko esan dezakegu, ingeleserako hautapen-murriztapenak euskara itzul daitezkeela. Beraz, esan dezakegu, *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, eta argumentuen tasunak parekatzeko garaian ezberdintasun hauek kontuan izan beharrekoak dira.

5.6.4 Domeinu-eremu semantiko bikoteen ebaluazioa

Ingeleseko emaitzen azterketan (4.4.5 atala) esan dugun bezala, domeinu-eremu semantiko bikote batekin adierazitako hautapen-murriztapenen ebaluazioan zenbait arazo izan ditugu, eta hauek eragoztearren irizpide batzuen arabera egin dugu ebaluazioa, ondoren laburbildu ditugunak:

(a) Domeinu-eremu semantiko bikote batekin adierazitako hautapen-murriztapenak (*transport-person*, esaterako), urre-patroiko (*person-person*) eremu semantiko bera badu (*person*), orduan hautapen-murriztapen hori onargarritzat hartu dugu.

(b) Izen-bereziak (x baten bidez adieraziak datozenak), pronominalak (*pro* baten bidea adieraziak datozenak), eta *factotum-Tops* bikotea, nahiz eta onargarritzat jo, eskuratu gabe bezala kontsideratu ditugu.

[13] Euskarazko aditzen agerpen gehienak kirol domeinuari dagokion corpus-atalean bakarrik azaldu dira.

Corpusa	Hautapen-murriztapenak	Kasua	Zuzena	Onargarria	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

6. taula: Euskararako eskuratutako eta ingelesetik itzulitako hautapen-murriztapenen emaitzak. Ebaluatzeko irizpide zorrotzagoa erabilita.

(c) Zuzen/onargarri bezala ebaluatutako hautapen-murriztapen batekin, bi urre-patroi eskuratu daitezke, baldin eta eremu semantikoa bera duten: *factotum-act* hautapen-murriztapenarekin *play-act* eta *sport-act* urre-patroiak eskuratzen dira, adibidez.

Orain arteko emaitzak irizpide hauen arabera lortutakoak dira. Ingeleseko hautapen-murriztapenekin egin dugun bezala, irizpide zorrotzago bat erabilita ebaluazioa errepikatu dugu euskarako azterketan ere. Irizpide aldaketa hau domeinu-eremu semantiko bikoteei dagokie: bikote bat onargarritzat hartuko dugu, urre-patroia baino orokorrago edota zehatzago bada, **eta domeinuko beste izen gehienak aditz horren argumentu izan badaitezke**[14]. Beheko taulan, irizpide zorrotzago batzuetan oinarrituz lortutako emaitzak ditugu (ikus 6. taula).

6. taulan ikus daitekeen bezala, domeinu-eremu semantikoetan oinarritzen diren eskuratze-tekniken (w2semf eta s2semf) kalitatea asko jaitsi da, baina hala eta guztiz ere, euskarako emaitzak nahiko onak dira. Dena den, 5.6.1 atalean aipatu dugun bezala, susmoa dugu objektua beste kasu-markekin adierazita datorrenean emaitzak ez direla horren onak.

[14] Azalpen gehiagorako jo 4.4.5 atalera.

6 ONDORIOAK

Txosten honetan azaldu dugun ikerlanak bi helburu nagusi zituen:

(a) Hainbat eskuratze-teknika erabiliaz ingeleseko eta euskarako corpus ezberdinetatik ikasitako hautapen-murriztapenak aztertu eta konparatu.

(b) Hautapen-murriztapenak eleanitzak izan daitezkeen aztertu.

Esperimentu honen emaitzak behin-behinekoak dira, aditz-adiera batzuk bakarrik aztertu baititugu, eta eskuratze-teknika guztiak ezin izan direlako corpus guztien gainean erabili. Ingeleseko hautapen-murriztapenetatik hurrengoa ondorioztatu dugu:

(a) **Corpus bakoitzak bere idiosinkrasia du eta hori emaitzetan islatzen da.** SemCor eta BNCn eskuratze-teknika berak erabili dira, eta SemCor-etik ikasitakoak BNCkoak baino hobeak dira. Hala ere, SemCor eskuz etiketatutako corpusa izaki, honetatik eskuratutako hautapen-murriztapenak hobeak espero ziren. Corpus txikiagoa izatea, etiketatze-erroreak eta corpuseko adiera batzuk EuroWordNet-en ez egotea izan daitezke honen arrazoiak. Azkenik, EFE corpusa domeinu zehatz batekin erabiltzea, emaitza onak eman ditu; honetatik *play* aditz-formaren bi kirol adierak (*play 00605818* eta *play 00610422*) lortzeko gai izan gara.

(b) **c2c eskuratze-teknikak ez dira w2c-enak baino hobeak**. Lehenengoan (c2c), aditza klase bezala kontsideratzeak (troponimoaz baliatuz) ez dirudi emaitza hobeagoak lortzen laguntzen duenik. Eskuratze-teknika hau oinarri egokia iruditzen zitzaigun hautapen-murriztapenen ikasketa eleanitza egiteko, eta hizkuntza bateko hautapen-murriztapenak zuzenean beste batera itzultzeko. Emaitza ikusita, bide honetatik jarraitu aurretik, honek ikerkuntza gehiago behar duela argi dago. Bigarrenean (w2c), hautapen-murriztapenen kalitatea nahiko ona izan arren, hauek aditzaren adiera guztientzat dira, eta oso erabilera konputazional mugatua dute. Eskuratze-teknika hau domeinu bateko corpusean erabilita emango lituzkeen emaitzak ikustea interesgarriak izan daiteke.

(c) **Domeinu-eremu semantiko bikoteekin adierazitako hautapen-murriztapenak 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 onak lortu dituzte.

(d) **Domeinu batean oinarritutako eskuratze-teknikek hautapen-murriztapen hobeak ikasi dituzte, eta domeinuaren arabera aditz horren adiera mugatu daiteke**. Hala ere, beste aditzekin frogatu beharko litzateke; dirudienez, aditz batzuk domeinu batekin beste batzuekin baino lotura gehiago izan baitezakete (ikus 4.6.2 atala).

(e) **Izenen anbiguotasuna arazo bat da**. Ikusi ditugu *game* eta *defentsa* bezalako izenekin gertatu diren nahasketak. Beraien EuroWordNet-eko *synset* edo domeinu-eremu semantiko egokia hartu ordez, tresnak beste *synset* edo domeinu-eremu semantiko bat aukeratu du, eta ondorioz, hautapen-murriztapen okerra lortu du.

(f) **Erroreen azterketatik ondoriozta dezakegu, prozesaketa linguistiko hobeago batekin, hautapen-murriztapen hobeak lortuko genituzkeela**, hau da, *parser*-ean aurkitutako erroreak konponduz gero, eta anafora eta izen berezien tratamendua landuz gero, okerrak ziren hautapen-murriztapen asko eragotziko genituzkeela uste dugu.

Ingeleseko eta euskarako hautapen-murriztapenen konparaketari dagokionez:

(a) **Euskarako hautapen-murriztapenen kalitatea ingelesekoena baino zertxobait handiagoa da.** Baliteke, argumentuak kasu-marketan banatu izana eraginik izatea eta susmoa dugu euskarako objektua beste kasu-markekin adierazita datorrenean, emaitzak ez direla horren onak.

(b) **Ingeleseko aditzen hautapen-murriztapenak euskara zuzenean itzul daitezke**, hala ere, gerta daiteke ingeleseko objektua euskarako kasu ezberdinekin agertzea (inesiboan adibidez).

Oro har, domeinuetaz baliatuz gero, aditz-adieraren hautapen-murriztapen hobeak lortuko ditugu. Emaitzek erakusten dute ere hautapen-murriztapenak hizkuntza batetik bestera itzul daitezkeela, horrela, baliabide gehiago dituen hizkuntzaz baliatu gaitezke euskararen ikasketa automatikorako. Dena den, hizkuntzen argumentuen ezaugarri linguistikoak batzuetan ez datoz bat eta parekatu egin behar dira.

Etorkizuneko lanari begira, eta honako hau hastapeneko lan bat izaki, badaude gehiago lantzeko hainbat puntu. Hasteko, kirolaren domeinuaz gain beste domeinu batzuetako aditzak ere aztertu nahiko genituzke (finantzaren domeinua 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; hauek SemCor-en oinarrituta izandako emaitzak kaskarrak ikusita, eskuratze-teknika hauek berriro erabili baino lehen, antzemandako erroreak gainditzea komeni da (*parser*-aren akatsak konpondu, anafora eta izen berezien tratamendua egin, aditz klaseetan troponimia kontuan ez hartu, eta abar). Hurrengo saiakeretan, domeinu-eremu semantiko bikoteekin adierazitako hautapen-murriztapenak alde batera utziko dira. Hauek lortutako emaitzak oso aldakorrak dira ebaluatzeko irizpideen arabera. Gainera, ebaluatzeko izandako arazoetaz jabetu gara, baita *synset*-ekin parekatzeko duten zailtasunetaz ere. Horiengatik guztiengatik, beste eskuratze-tekniketan funtsatzea erabaki dugu. Bestalde, ingeleserako eta euskararako ikasitako hautapen-murriztapenen ebakidura eginez gero, errore ugari desagertuko direla uste dugu, eta hipotesi hau egiaztatu nahiko genuke. Azkenik, euskararako hautapen-murriztapen mota gehiago ikasi nahi ditugu, w2semf eskuratze-teknikatik lortutakoetaz gain. Hasiera batean, w2c eta c2c teknikekin hastea pentsatu dugu. Horrela, euskarako datu gehiago izango dugu ingelesekoekin erkatzeko.

Irudien zerrenda

Bibliografia

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

Agirre E., Atserias J., McCarthy D., Real F., Rigau G., eta Rodriguez H. MEANING: Developing multilingual web-scale language technologies. Working paper 5.2a. Technical report, 2003.

Agirre E., Atutxa A., Gojenola K., eta Sarasola K. Exploring portability of syntactic information from English to Basque. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC)*, Lisbon, Portugal, 2004.

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

Agirre E. eta Lersundi M. Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. In *Proceedings of the Annual SEPLN meeting*, 2001.

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

Agirre E. eta Martínez D. Integrating selectional preferences in WordNet. In *Proceedings of First International WordNet Conference*, Mysore, India, 2002.

Aldezabal I. *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa. Levin-en (1993) lana oinarri hartuta eta metodo informatikoak baliatuz.* PhD thesis, UPV-EHU, 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. In *Actas del XVII Congreso de la SEPLN Universidade de Jaén*, Jaén, Spain, 2001.

Arriola J. *Euskal Hiztegia-ren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira. Aditz-adibideen analisia Murriztapen-gramatika baliatuz, azpikategorizazioaren bidean.* PhD thesis, UPV-EHU, 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. In *Proceedings of EACL'99, Linguistically Interpreted Corpora*, Bergen, Norway, 1999.

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

Church K., Gale W., Hanks P., eta Hindle D. Using statistics in lexical analysis. In *Lexical Acquisition: Exploting On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

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

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

Fernández M. *Sintaxis y cognición: introducción al conocimiento, el procesamiento y los déficits sintácticos*. Editorial Síntesis, 1995.

Grishman R. eta Sterling J. Acquisition of selectional patterns. Nantes, France, 1992. COLLING-92.

Hindle D. Noun classification from predicate-argument structures. In *Prooceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 327–329, 1990.

Hindle D. eta Rooth M. Structural ambiguity and lexical relations. In *Prooceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 229–236, 1991.

Hornby A. *The Advanced Learner's Dictionary*. Oxford University Press, Oxford, 1991.

Katz J. eta Fodor J. The structure of a semantic theory. In Katz J. eta Fodor J., editors, *The Structure of Language*. Pretince Hall, 1964.

Lenat D.B. eta Guha R.V. *Building Large Knowledge-Based Systems*. Addison Wesley, 1990.

Lewandowski T. *Diccionario de la Lingüística*. Cátedra, 1992.

Lin D. Principle based parsing without overgeneration. In *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 1993.

Magnini B. eta Cavagli G. Integrating subject field codes into Wordnet. In *Proceedings of the Second International Conference on Langiage Resources and Evaluation (LREC)*, Athens, Greece, 2000.

McCarthy D. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.

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

Montemagni S. Extracting typical subjects and objects of verbs from mono- and bi-lingual dictionaries. Technical report, ESPRIT BRA-7315 Acquilex-II, 1994.

Pereira F., Tisgby N., eta Lee L. Distributional clustering of English words. In *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–19, 1993.

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

Resnik P. A class-based approach to lexical discovery. In *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*. PhD thesis, University of Pennsylvania, 1993.

Ribas F. *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomyu*. PhD thesis, Universitat Politècnica de Catalunya, 1995.

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

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

Wilks Y. Preference semantics. In Keenan E., editor, *The Formal Semantics of Natural Language*. Cambridge University Press, 1973.

Wilks Y. An intelligent analyzer and understander of English. In *CACM*, pages 264–274. 1975.