eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Hiponimia/hiperonimia erlazioaren erauzketa automatikoa

Irati Ugarteburu
Tutorea: Aitor Soroa

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua

lortzeko bukaerako proiektua

2012ko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia**,** Elektronika eta Telekomunikazioak.

LABURPENA

Hizkuntzaren prozesamenduaren arloan baliabide lexiko-semantikoak oso lagungarriak dira, baina berauek sortu eta mantentzea denbora eta errekurtso asko eskatzen dituen lana da.

Lan honetan euskarazko baliabide semantikoak era automatikoan edo erdiautomatikoan elikatzeko teknikak aztertzen dira. Zehazki testutik hiponimo eta hiperonimoak erauzteko modu desberdinak azaltzen dira.

Hitz gakoak: terminologia erauzketa, testuinguru antza, patroi lexiko-semantikoak, hiponimia/hiperonimia

ABSTRACT

Lexical-semantic resources are very useful in natural language processing, but they are expensive to build and maintain.
In this work we present methods to enhance existing resources such as Euskal WordNet in an automatic or semi-automatic way. We will explain different techniques to extract hypernyms and hyponyms from corpora.

Key words: terminology extraction, context similarity, lecical-sintactical patterns, hypernymy/hyponymy

Aurkibidea

Taulen Aurkibidea

Irudien Aurkibidea

1 Proiektuaren definizioa

Ontologiak eta thesaurusak bezalako baliabide lexiko-semantikoak oso lagungarriak izan daitezke hizkuntzaren prozesamenduan aurkitu daitezkeen arazo ugariri aurre egiteko. Besteak beste, kolokazioen erauzketan (Pearce, 2001), testuen sailkapenean (Clark and Weir, 2002), galdera-erantzun sistemetan (Moldovan et al., 1999; Pasca and Harabagiu, 2001), edota testuen lotura aztertzean (Moldovan et al., 1999) erabili izan dira baliabide lexiko-semantiko hauek.

Bereziki WordNet (Fellbaum, 1998) izan da NLP ikerketan eragina izan duen baliabide lexiko-semantikoa. Hala ere, WordNeten estaldura mugatua da oraindik ere (Pennacchioti eta Pantel, 2006) eta (Hovy et al., 2009) lanetan nabarmendu den bezala. Izan ere, WordNet eta gainontzeko taxonomia gehienak eskuz sortuak dira eta ondorioz oso zailak dira eguneratuta mantentzeko, batez ere uneoro aldatzen ari diren domeinuetan. Egunero terminologia berria sortuz doa, beste hizkuntza batzuetatik hitz berriak mailegatzen dira, hitz batzuk zaharkituak gelditzen dira, eta hitz batzuen esanahia ezberdina izan daiteke domeinuaren arabera. Ondorioz, ia ezinezkoa da horrelako errekurtso bat eskuz osatu eta mantentzea.

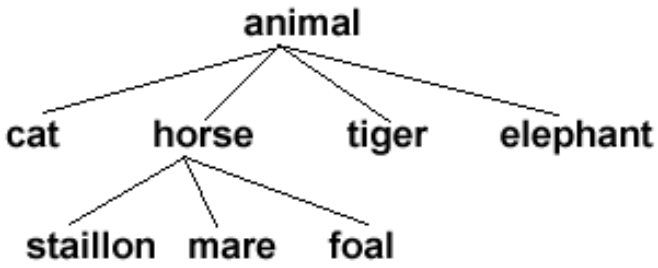
Beraz, argi dago beharrezkoa dela metodo automatiko edo erdi-automatikoak garatzea baliabide lexiko-semantikoen sorkuntza eta mantenurako.

1.1 Hiponimia eta hiperonimia

Semantika lexikalak lexikoko elementuen artean dauden erlazio lexiko-semantikoak biltzen ditu: sinonimia, antonimia, hiperonimia/hiponimia, eta beste. Erlazio lexiko-semantiko horiek sare-semantiko moduko batean adierazten dira esplizituki. Ingelesezko sare-semantikoen artean ezagunena WordNet (Fellbaum, 1998) izenekoa dugu, eta haren euskararako egokitzapenari Euskal WordNet (Pociello, 2008) deitzen diogu. Esan bezala, Euskal WordNet sare semantikoa automatikoki edo erdiautomatikoki aberastea da lan honen funtsa.

Lan honetan izenen arteko hiperonimia/hiponimia erlazio semantikoan zentratuko gara. Hizkuntzalaritzan, hiponimoa bere eremu semantikoa beste termino baten eremuaren barnean duen hitz bat da. Adibidez, kolore hitzaren hiponimoak gorri, beltz eta abar dira. Hiponimoak biltzen dituen terminoari, edo beste hitz batzuen esanahia

bere baitan gordetzen duen hitzari, hiperonimo deritzo. 1. irudian ikus daiteke hiponimo eta hiperonimoen arteko erlazio hierarkiko baten adibidea.



1. irudia: Hiponimia/hiperonimia erlazioaren adibidea

Hiperonimoak ez du bere hiponimoak ez daukan inolako tasun semantikorik baina hiponimoak badu bere hiperonimoak ez dauzkan eta berarengandik aldentzen dituen tasun semantikorik.

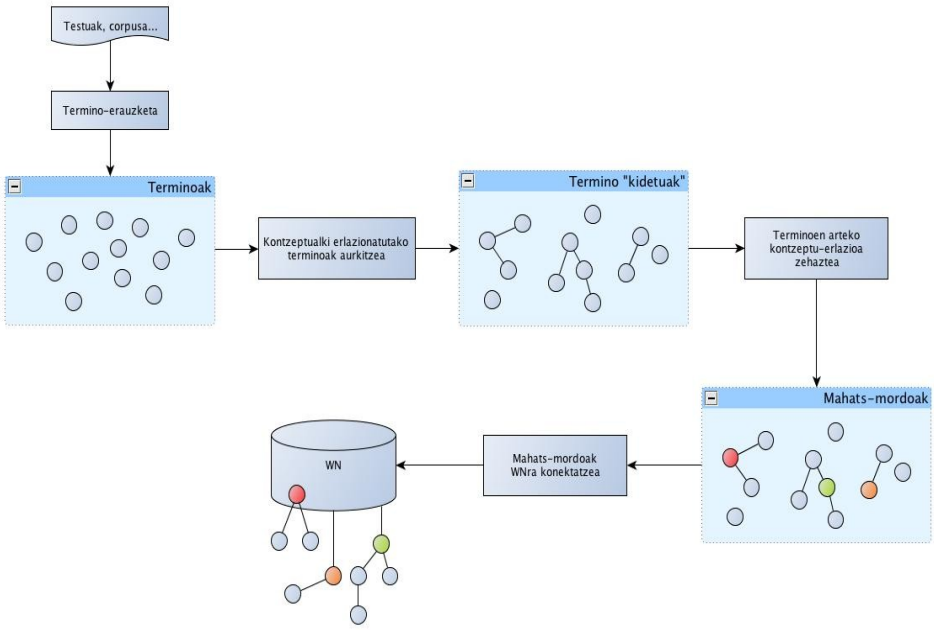
> Adibidez, altzari hiperonimoa da eta lanpara hiponimoa. Altzari hitzak sema asko ditu: [+etxeko objektua], [+erabilgarria]... Bere hiponimoak sema horiek guztiak izan beharko ditu eta zehaztasun gehiago ematen dioten beste sema batzuk [+argia emateko balio du]...
>
> Lanpara hitzaren hiperonimo altzari.
>
> Altzari hitzaren hiponimoak: lanpara, aulki, mahai...

Esan bezala, WordNetek bere baitan mota honetako erlazio semantikoak biltzen ditu, eta horiek izango ditugu aztergai lan honetan. Hala ere, etorkizunean hiponimia erlazioaz gain, sortutako sistemak beste erlazio batzuk bilatzeko ere balio izatea espero da, sinonimia esaterako.

1.2 Helburuak

1.2.1 Helburu nagusia

Lan hau proiektu orokorrago baten barnean kokatuta dago. Euskarazko testu espezializatuetako terminoekin kontzeptu-erlazioak automatikoki erauzteko tresna garatzea da lan orokorraren helburu nagusia.



2. irudia: Domeinu-ontologiak sortzeko prototipoaren prozesu-diagrama orokorra

2. irudian ikus daitekeen bezala, lehenengo pausoa corpus batetik terminoak erauztea da. Ondoren, erlazionatuta dauden terminoak bilatu eta beraien arteko kontzeptu-erlazioa zehaztu beharko da. Erlazionatutako termino horiekin mahats-mordoak sortuko dira eta, azkenik mahats mordo horiek Euskal WordNeten dagozkien tokian kokatu beharko dira. Beraz, lau pauso nagusi banatuko lirateke prozesuan:

1. Corpusetik terminoak erauztea
2. Erlazionaturik dauden terminoak aurkitzea
3. Termino horien arteko erlazio-mota hiperonimia/hiponimia den automatikoki erabakitzea, eta hala denean, hiperonimoa aurkitzea eta terminoak egituratzea (mahats-mordoak eratzea)
4. Mahats-mordoen hiperonimoa desanbiguatu eta WordNeten zintzilikatzea

Prozesu orokorra kontuan izanda, lan honetan, hiponimia/hiperonimia kontzeptu-erlazioa duten bikoteak bilatzea izango da helburu nagusia.

1.2.2 Helburu zehatzak

Hiponimia erlazioa hainbat bidetatik jorratu daiteke, eta horren arabera zehaztu dira helburu zehatzak. Nagusiki hiru bide bereizten dira: termino habiatuen tratamendua, markatzaile linguistikoak eta antzekotasun distribuzionalean oinarritutako teknikak.

Honakoak dira definitu diren helburuak:

- Teknika linguistikoen garapena:
  - Terminoen egitura morfosintaktikoan oinarrituak: termino-aldaeren eta termino habiatuen tratamendua. Termino habiatuek bere
  - Terminoen testuinguru linguistikoan oinarrituak: kontzeptu-erlazioaren euskarazko adierazle diren “markatzaileak” (sintagmatikoak, esaldi-mailakoak, diskurtsiboak) corpusak aztertuz zehaztea, eta kontzeptu-erlazioak automatikoki esleitzeko metodologia lantzea
- Teknika estatistikoen garapena:
  - Antzekotasun distribuzionalean oinarritutako teknikak garatzea. Antzeko testuingurua duten terminoak semantikoki erlazionatuak izateko probabilitatea dute (sinonimoak, hiper-hiponimoak, kideak...)
- Erlazio-erauzlea: aurreko hiru teknikak ikaste automatikoko sistema batean integratzea.

Hizkuntza Teknologien beste hainbat alorretan bezala, hemen ere metodo estatistikoen eta linguistikoen arteko lehiaren / kooperazioaren auzia dago. Azkenaldian, gero eta joera argiagoa nabari da kooperazioaren alde. Literaturan bi era horietako teknikak proposatu eta erabiltzen dira, eta gure asmoak ere ildo horretatik doaz, teknika linguistikoek eta estatistikoak integratzerantz, alegia.

2 Aurrekariak

Testutik hiponimo-hiperonimoak era automatikoan erauzteko momentuan aplikatu diren teknikak hiru multzo nagusitan banatu daitezke: hurbilpen estatistikoak, patroi lexiko-sintaktikoak darabiltzatenak, eta dependentzia zuhaitzak darabiltzatenak.

Hurbilpen estatistikoen kasuan bi ikuspegi hartzen dira kontutan, batetik hitzen antzekotasun-distribuzionala kontutan hartzen dutenak (Cohen and Widdows, 2009; Cederberg and Widdows, 2003; Poon and Domingos 2010), eta bestetik hitzak hitz-zerrendetan duten agerkidetza (Riloff and Shepherd, 1997; Roark and Charniak, 1998; Widdows and Dorow, 2002).

2.1 Patroi lexiko-sintaktikoak

Patroi lexiko-sintaktikoak definitzeko lan asko egin da Hearstek (1992) egin zuen lehen definiziotik. Patroi hauek definitzeko teknika ezberdinak erabiltzen dira, eta horien azalpena da atal honetan biltzen dena.

2.1.1 Eskuz definitutako patroiak

Patroi lexiko-sintaktikoen alorrean Hearst(1992) da aurrekari bezala hartzen dena. Hearstek hainbat patroi definitu zituen, eskuz, hiponimo-hiperonimo erlazioa zuten bikoteak testutik erauzteko:

1. NP such as {NP , NP ... , (and|or)} NP
2. NP such as {NP ,}* {(or|and)} NP
3. NP {NP}* {,} or other NP
4. NP {NP}* {,} and other NP
5. NP {,} including {NP ,}* {or|and} NP
6. NP {,} specially {NP ,}* {or|and} NP

Patroi hauek erreferentziatzat hartuta hainbat ikerketa lan egin dira geroztik (Riloff eta Jones, 1999; Berland eta Charniak, 1999; Fleischman eta Hovy, 2002; Thelen eta Riloff, 2002; Pasca et al., 2004; Kozareva et al., 2008).

(Kozareva et al., 2008) lanean Hearsten patroietako bat, azaleko patroi bat, hartzen da hiponimo-hiperonimo bikoteak lortzeko, eta grafo egitura batekin konbinatzen dute. Aukeratutako patroi hori DAP (*double-anchored pattern)* bat da:

DAP: [seedTerm1] such as [seedTerm2] and <X>

Lan honen arabera, mota honetako patroiek, klase izena eta klaseko kide bat biltzen dituztenek, zehaztasun handia izango dute oso espezifikoak direlako, beraz oso patroi fidagarriak dira. Baina emaitza zuzen asko bildu arren emaitza okerrak ere lortzen dira patroi hauekin. (Kozareva et al., 2008) lanean erabilitako algoritmoak patroiekin erlazionatutako bi ezaugarri biltzen ditu: popularitatea eta produktibitatea. Hautagai bat popularra izango da beste hitz askoren bidez aurkitu badaiteke, eta produktiboa izango da hitz horren bidez beste asko bilatu badaitezke. (Kozareva et al., 2008) artikuluan instantziek bata bestea aurkitzeko duten gaitasuna hartzen da kontutan. Inolako part-of-speech eta etiketatzerik gabe.

Hautagaien bilaketa egiteko orduan, weba darabilte, patroi lexiko sintaktikoak eragin dezaketen datu-urritasunari aurre egiteko. Lortutako hautagai bakoitza patroia berriz bete eta kide berriak lortzeko erabiliko da hautagai berririk ez dagoen arte. Hautagaiak grafoan sartzerakoan pisu bat ezartzen zaie, eta horretarako metodo ezberdinak konparatzen dituzte. Produktibitatea eta popularitateari garrantzi maila ezberdinak emanez. Lortutako emaitzak esperantzagarriak direla diote.

(Hovy et al., 2009) lanean oraindik urrunago doaz, eta (Kozareva et al., 2008) lanean erabilitako azaleko patroiak lortutako emaitzak hartu eta bigarren pauso bat gehitzen diote algoritmoari. Pauso berri horretan tarteko-kontzeptuak dira bilatzen direnak.

DAP-1: <X> such as [seedTerm1] and [seedTerm2]

DAP-1 patroia da kasu honetan aplikatzen dena, patroi bera da baina kasu honetan klasearen izena bilatzen da. Metodo honen bidez jada aurkitutako kontzeptu ugari itzultzen dira.

Ebaluazioa egiteko WordNeten kontrako konparaketa bat eta adituen konparaketa bat egin dituzte. Bi ebaluazioak garrantzitsuak direla diote. Oinarritzat hartutako (Kozareva et al., 2008) lanarekin konparatuz, alderantzizko buelta gehitzeak 5 aldiz oinarrizko termino gehiago lortzea ekartzen du, antzeko doitasun maila mantenduz. Erlazioak ebaluatzean lortutako beste ondorio bat da, WordNetek lortutako erlazioen erdia faltan dituela (Animalien esparruan 804 eta pertsonen esparruan 539).

(Hovy et al., 2009) lanaren bide berdina jarraitzen dute hainbat ikertzailek, baina kasu hauetan gehiago zentratzen dira lortutako erlazioekin ontologiak zerotik sortzean (Kozareva et al., 2010; Navigli et al., 2011). Lan hauek gure helburuko prozesuaren azken atalarekin izango lukete zerikusia (ikus 2. irudia).

2.1.2 Hutsetik sortutako patroiak

Baina patroiak eskuz etiketatzeaz gain, hutsetik erlazio semantikoak ikasteko moduak ere badaude. (Bollegala et al., 2010) lanean, esate baterako, inongo jakintzarik izan gabe entitateen arteko erlazio mota ezberdinak ikasten dituzte eta horietako bat hiponimia erlazioa da.

Aztergai dituzten testuak webetik hartzen dituzte, eta clustering bidez sailkatzen dituzte erlazio mota ezberdinak. Erlazio horietako bat hiponimia/hiperonimia erlazioa da. Azterketa egiteko momentuan erlazio semantikoaren adierazpen diadikoa hartzen dute, intentsio-izaera eta hedadurazko definizioa. Hau da, erlazio bakoitza sortzen dituen entitate bikoteek zehaztu dezakete, hedadurazko definizioa erabiliz. Edota, erlazioa berau adierazten duten patroi lexiko-sintaktikoen bidez erabiliz, intentsiozko definizioan. Entitate potentzialak identifikatzeko POS etiketatzaile bat eta izen-sintagma chunker bat darabilte. Horretarako matrizeak sortzen dituzte non patroi lexiko-sintaktikoei eta hautagai bikoteei dagozkien datuekin.

Weba aztergai dutenez, eskuz etiketaturiko daturik behar ez duen sistema bat eraiki dute. Klase-anitzeko erregresioa erabiltzen dute patroi esanguratsuak entrenatzeko.

Erlazio berriak ia ezerezetik sortzen dituen beste tresna bat KnowItAll dugu (Etzioni et al., 2005). abiapuntu gisa aurrez definitutako erlazio orokor batzuk hartzen ditu. Sistemak ikasketa erdi-gainbegiratua darabil eta sistemak berak aukeratu eta etiketatzen ditu interesatzen zaizkion adibideak. Beraz, ez du etiketatutako daturik behar. Patroi orokor batetik abiatuz, Y klaseko X hautagaiak lortuko lituzke, eta ondoren hautagai horiek erabili patroi gehiago bilatzeko. Hala ere, sistema hau hasieran definitutako patroiekiko menpekoa da.

2.2 Dependentzia Zuhaitzak

Hearsten patroiak hiperonimia adibide batzuk era egokian aurkitzeko balio dute. Hala ere, Hearstek proposatutako patroi hauek testu naturalean aurkitu daitezkeen egituretatik gutxi batzuk besterik ez dituzte biltzen. Gainera Mititeluk (2006) azaltzen

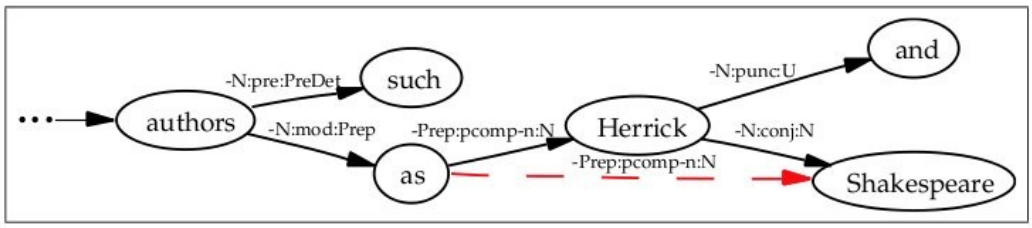
duen bezala, Hearsten patroiek hiponimo/hiperonimo bikoteak erauzteaz gain erlazionatuta ez dauden terminoak ere hartzen dituzte:

food particularly chocolate – waters particularly the reservoirs

aspirin and other drugs – rats and sometimes other creatures

Beraz, Mititeluk (2006) beharrezkotzat jotzen du sintaxiari lotutako informazioa gehitzea markatzaile linguistikoei dagokienean.

Ondorioz, (Snow et al., 2005) lanean esaldiak dependentzia-zuhaitzen bidez adierazten dituzte (ikus 3. irudia), eta era automatikoan definitzen dituzte dependentzia patroiak, egunkarien esparruko corpus batetik abiatuz. Irudiko arkuek hitzen arteko dependentzia mota ezberdinak irudikatzen dituzte, eta arku bakoitza adierazteko "ezkerreko hitzaren kategoria : dependentzia-erlazioa : eskuineko hitzaren kategoria" formatua erabiltzen dute.



3. irudia. (Snow et al., 2005) laneko MINIPAR dependentzia zuhaitza

Era honetan, hautagai bikote bat emanez gero, bikote hori agertzen den esaldietako ezaugarriak erauzi eta ea hiponimia erlazioa duten esango du sailkatzaileak.

Metodo honekin lortzen dituzten patroietako batzuk bat datoz Hearst-en patroiekin, 1.taulan ikus daitekeen bezala:

Patroi lexiko-sintaktikoa	**Snowren dependentzia**
NPX and other NPY	(and,U: PUNC :N),-N: CONJ :N, (other,A: MOD :N)
NPX or other NPY	(or,U: PUNC :N),-N: CONJ :N, (other,A: MOD :N)
NPY such as NPX	N: PCOMP - N :PREP,such as,such as,PREP : MOD :N
Such NPY as NPX	N: PCOMP - N :P REP,as,as,P REP : MOD :N,(such,PRE DET: PRE :N)
NPY including NPX	N: OBJ :V,include,include,V: I :C,dummy node,dummy node,C: REL :N
NPY , especially NPX	N: APPO :N,(especially,A: APPO - MOD :N)

1 Taula: Snowk aurkitutako Hearsten patroiak

Hearstek definitutako patroiez gain, patroi berri batzuk ere aurkitu dituzte 2. taulan ikus daitekeen bezala.

Patroi lexiko-sintaktikoa	**Snowren dependentzia**
NPY like NPX	N: PCOMP - N :P REP,like,like,P REP : MOD :N
NPY called NPX	N: DESC :V,call,call,V: VREL :N
NPX is a NPY	N: S :VBE,be,be,-VBE: PRED :N
NPX , a NPY	(appositive)N: APPO :N

2 taula: Snowk aurkitutako eta Hearstek definitu gabeko patroiak

Sailkatzaile honek aurretik erabilitakoekin alderatuz askoz emaitza hobeak ematen ditu. Baina hala ere arazo bat izaten jarraitzen du, esaldi berean agertzen diren hiponimo-hiperonimo bikoteak besterik ez ditu aurkitzen. Arazo honi aurre egiteko koordinazio erlazioa gehitzen dute beraien algoritmoan “Y Xen koordinatua izango da, baldin eta X eta Yk hiperonimo berdina partekatzen badute”. Horretarako antzekotasun distribuzionalaz eta “X, Y and Z” bezalako koordinazio patroiez baliatzen dira. Beraz, dependentzia zuhaitzen emaitzak eta koordinazioaren emaitzak konbinatzen dituzte hautagai bikote batek hiponimia erlazioa duen kalkulatzeko. Konbinazio honekin WordNeteko sailkatzaile onenarekiko %43ko F-score hobekuntza lortzen dute.

2.2.1 Dependentzia zuhaitzak eta Kernel funtzioak

Baina dependentzia zuhaitzek ematen dituzten emaitzak hobetzeko asmoz, datu hauek beste hainbat ezaugarrirekin konbinatuz sortutako sailkatzaileak ere badaude. Horien artean kokatzen dira Dependentzia Zuhaitzen Kernelak (Zelenko et al., 2003; Culotta and Sorensen, 2004; Wang et al., 2006). Kernel metodoak instantzien artean *kernel funtzio* bat konputatzen duen dentsitate estimazio ez parametrikoak dira. Kernel funtzio bat antzekotasun neurri bat izan daiteke; hipotesiaren arabera antzekoak diren erlazioek zuhaitz egitura antzekoa izango dute.

(Zelenko et al., 2003) artikuluan entitateen arteko erlazio mota ezberdinak erauzten dituzte, horien artean hiponimia erlazioa. SVM metodoan oinarritutako sailkapena proposatzen dute, eta ondorioz, ezaugarri bektoreak nahi adina ezaugarri sartzen dituzte. Lana burutzeko azaleko parserrak erabiltzen dituzte kernel funtzioetan. Erlazioak definitzeko patroiak dagoeneko erauzitako erlazio batzuetatik lortzen dira, azaleko parserren bidez. Algoritmo honetan oinarritzen da (Culotta and Sorensen, 2004) lana ere, baina kasu honetan aurkeztutako hurbilpenak esaldien errepresentazio aberatsagoa,

framework orokorragoa, eta kernelen sakabanatzeari aurre egiteko kernel konposatuak erabiltzea proposatzen dute. Lan honetan erlazio potentzialak sortzeko esaldi berean gertatzen diren entitate pare bakoitza hartzen da bien artean ahal den dependentzia-zuhaitzik txikiena sortzen dute. Era honetan ezaugarri lokalak nabarmendu eta zarata murrizten dituzte. Dependentzia-zuhaitza sortzeko parseatutako esaldia erregela batzuen bidez eraldatzen da. Ondoren, zuhaitzaren nodo bakoitza ezaugarri bektore bat bezala adierazten dute, nodo informazio gehigarria erantsiz: hitza bera, POS, POS orokorra, chunk-tag, entitate mota, entitate maila, WordNeteko hiperonimoak eta erlazioko argumentuak.

(Wang et al., 2006) lanean ere bide berdina jarraitzen dute, kasu honetan biltzen dituzten ezaugarriak honakoak dira: bi entitateei eta inguruko hitzei dagozkienak, POS Tag-ei dagozkienak, hautagaien arteko posizio erlatiboa, entitateak dauden chunken ezaugarriak, dependentzia zuhaitza, eta WordNeten oinarritutako ezaugarri semantikoak. Artikulu honetan ezaugarri bakoitzak sailkapenaren emaitzetan eragiten duen hobekuntza neurtzen dute, eta ondorio bezala azaleko analisia eta analisi sintaktiko sakona erabiltzearen artean dagoen antzekotasuna nabarmentzen dute. Baita WordNet bezalako baliabide semantiko batek ematen duen laguntza ere.

2.3 Patroi lexikoak edo dependentzia patroiak?

Patroi lexiko-sintaktikoek zein dependentzia-zuhaitzeko beraien alde onak eta txarrak dituzte. Printzipioz badirudi dependentzia patroiak, informazio gehiago izanda, hobeak izan behar direla erauzketa egiteko orduan. Baina dependentzia patroiak lortzeko testu-prozesagailu garatuagoak behar dira eta ez da nahikoa azaleko analizatzaile bat izanda. Lehenengo metodoak hitzen lema eta POSak behar ditu, bigarren metodoak, aldiz, hitzen arteko erlazio sintaktikoak. (Sang and Hoffman, 2009) lanean hiponimia erlazioak erauzteko bi metodoetatik egokiena zein den ikertzen dute. Froga hau egin ahal izateko egunkarietan oinarritutako corpus orokor bat eta Wikipedia darabiltzate. Sang eta Hoffmanek (2009) lortutako emaitzen arabera, ez dago ezberdintasun nabarmenik bi metodoen artean; patroi lexikoak eta dependentzia zuhaitzak konparatuta, gainontzeko informazio gehigarririk gabe, emaitzetan ez dago alde nabarmenik.

2.4 Hurbilpen estatistikoak

Hiponimo-hiperonimo bikoteak erauzteko hainbat hurbilpen estatistiko ere proposatu dira, batez ere, testuinguru-antza eta koordinazioa hartzen dituzte kontutan. Curranek (2005) lanean ondorioztatzen duen bezala, distribuzionalki antzekoak diren hitzak semantikoki antzekoak izan arren eta kasu askotan klase bereko kideak diren arren, metodo hau bakarrik erabiltzeak doitasun maila oso baxua ematen du. Hala ere, beste metodo batzuen lagungarri gisa, patroi lexiko-sintaktikoek lortutako emaitzak osatzeko adibidez, emaitza onak ematen ditu (Cederberg and Widdows, 2003; Giovannetti et al., 2008).

2.4.1 Testuinguru-antza

Testuinguru-antza neurtzeko neurriak, LSA esate baterako, dokumentuen eta hauetan dauden terminoen analisi bat egiteko erabiltzen dira. Termino bakoitzaren testuinguruari dagozkion datuak bektore batean biltzen dira. Ondoren, bi terminoren testuinguru-antza jakiteko bi bektoreen antzekotasuna neurtzen da. Era honetan, corpusean dauden bikote bakoitzeko beraien distantzia-distribuzionalaren neurria lortzen da, eta neurri horren arabera ranking bat sortu daiteke. Ranking horretan antzekoenak diren hitzak semantikoki erlazionatuta daudenak izango dira.

(Cederberg and Widdows, 2003) lanean, esate baterako, patroi lexiko-sintaktikoak erabiltzen dituzte hiponimo-hiperonimo hautagaiak lortzeko. Lortutako emaitzen doitasuna hobetzeko LSA neurria erabiliz hiponimo-hiperonimo hautagaiak ranking batean biltzen dituzte, egokiak ez direnak baztertu ahal izateko. Emaitzarik onenak aztertuz gero, doitasuna %18 hobetzea lortzen dute.

(Giovannetti et al., 2008) lanean ere patroi lexiko-sintaktikoak eta testuinguru-antza erabiltzen dituzte, kasu honetan bi metodoez baliatzen dira hiponimo-hiperonimo bikote hautagaiak corpusean bilatzeko. Ondoren, hautagai hauek patroi lexiko-sintaktiko fidagarri batzuen bidez webean kontsultatzen dituzte, eta itzulitako bilaketen arabera onartu edo baztertu egiten dituzte. Metodo honekin bost patroi fidagarri ebaluatu dituzte hiponimia erlazioak erauzteko, eta lortutako zehaztasunak oso onak dira (%82tik gora hiperonimo zuzenak bilatzen), hala ere bikote asko ez dira patroi horien bidez aurkitzen.

Ritter et al. (2009) ere Hearsten patroiak betetzen dituzten hautagaietatik hasten dira hiponimo-hiperonimo erlazioak erauzten. Patroi lexiko sintaktikoez gain, ezaugarri

gehiago eransten dituzte eta SVM metodoa erabiltzen dute sailkatzeko. Jarraian, emaitzak hobetu asmoz, testuinguru-antza darabilen beste sailkatzaile bat egin eta bien emaitzak uztartuz doitasuna 0.8tik 0.82ra igotzen dute, eta estaldura 0.65etik 0.71ra.

Hiponimia erlazioa erauzteko beste teknika bat proposatzen da (Alfonseca eta Manandhar, 2002) lanean. Kasu honetan corpus bat eta WordNet dituzte abiapuntu hiponimia erlazio berriak erauzteko. WordNeteko synset orokorrenetik hasi eta synseta bera eta bere hiponimoak aztertzen dituzte. Testuan testuinguru-antzaren bidez hautagaiak hartu, eta *H* hautagaiak *S* synsetarekiko duen antzekotasuna bere hiponimoek dutena baino handiagoa bada, orduan *H* hautagaia *S*-ren hiponimoa izango da.

2.4.2 Koordinazioa

Baina hala ere, patroi lexiko-sintaktikoek hiponimo-hiperonimo kopuru mugatua biltzen dute, eta hautagai gehiago lortzeko modu egokia izan daiteke izenen koordinazioa. Azkotan co-hiponimoak diren hitzak biltzen baitira egitura horietan, hurrengo adibidean ikus daitekeen bezala:

> Antzinatik ezagutzen ziren **beruna , burdina , eztainua , kobrea , merkurioa , urrea eta zilarra.**
> Ba , hiru mota desberdinetan : **egoera solido , likido eta gaseosotan.**

Esate baterako, izenen koordinazio informazioa eta grafoetan oinarritutako clustering metodo bat darabiltzate Cederberg eta Widdowsek (2003). Izan ere, hitz-zerrendetan batera agertzen diren hitzak, semantikoki antzekoak dira askotan, (Roark and Charniak, 1998) lanean adierazten duten bezala. Mota honetako informazioa lexiko erauzketa automatikoan erabili izan da (Riloff and Sheperd, 1997; Widdows and Dorow, 2002).

Prozesua honakoa da: gaixotasun motak lortu nahi baditugu esate baterako, tifoide hitza har daiteke abiapuntutzat, eta tifoide hitzarekin zerrendetan agertzen diren hitzak lortu. (Widdows and Dorow, 2002) hitz bakarretik hasita clusterrak eraikitzeko algoritmo bat garatu zuten, eta aukeratutako kategorietan gehitutako kide berriak %82ko zehaztasuna zuten.

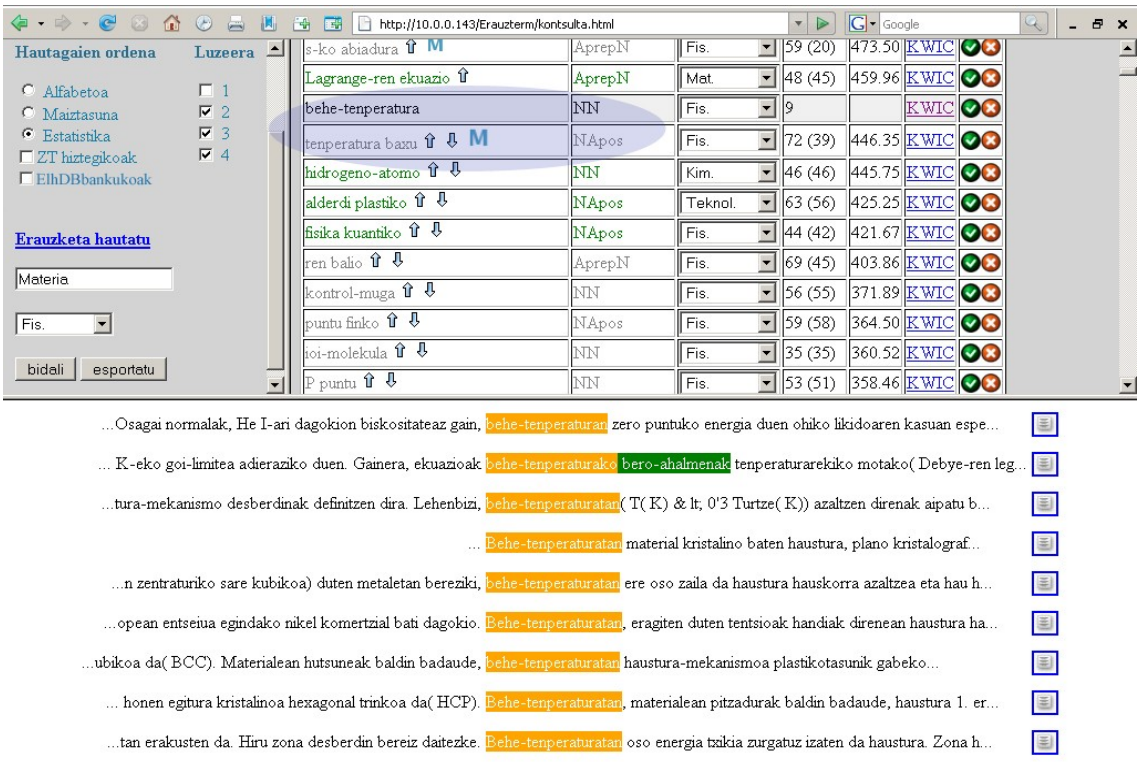
Kasu honetan, patroien bidez lortutako hiponimo-hiperonimo bikoteak hedatzen dituzte, eta 45etatik abiatuz 459 lortzen dituzte. Horietatik %46 dira zuzenak. Widdows eta Dorowren (2002) lanarekin alderatuz, emaitza kaskarra da, baina hasieran hartutako hitzetan dago aldea, Widdows eta Dorowren lanean hasierako hitzak tentuz aukeratuta daude eta.

Azkenik, bi metodoak konbinatzen dituzte: patroi lexiko-sintaktikoen bidez lortutako emaitzak koordinazio bidez hedatu, eta ondoren emaitzak LSA neurriarekin birfintzen dituzte. Hasierako pauso batean errorea %33 jaistea lortzea lortu dute.

Aipatu bezala, (Snow et al., 2005) lanean ere koordinazio informazioa erabiltzen da emaitzak hobetzeko. Zehazki, "X, Y and Z" bezalako koordinazio patroiez baliatzen dira. Beraz, dependentzia zuhaitzen emaitzak eta koordinazioaren emaitzak konbinatzen dituzte hautagai bikote batek hiponimia erlazioa duen kalkulatzeko.

2.5 Eta euskararentzat zer egin da?

Euskaraz termino erauzketarekin erlazionako lanak egin diren arren, ez dago terminoen arteko hiponimia/hiperonimia erlazioa erauzten duen sistemarik.



4. irudia. Erauztermen interfazea

Erauzterm (Alegria et al., 2005) testutik terminoak erauzten dituen tresna da (ikus 4. irudia). Termino bakunak eta hitz anitzeko terminoak erauzten ditu, baina ez du terminoen artean erlaziorik finkatzen.

Terminoen arteko erlazioak antolatzeko Euskal WordNet (Pociello et al., 2008) euskarazko ezagutza base lexikala dugu. Ezagutza-base hau eskuz aberastu eta eguneratzen da, horrek suposatzen duen lanarekin eta denborarekin.

(Lersundi, 2005) lanean hiztegien definizioetatik abiatuz ezagutza-base bat sortzeko beharrezkoak diren erlazioak erauzten dira. Hiztegietako definizioen kasuan esaldien egiturak nahiko zurrunak dira. Gure kasuan, ordea, erauzketa corpusetik egiteko teknika ezberdinak landu beharko dira.

Elhuyar Fundazioan eta IXA taldean terminoen arteko erlazioekin hainbat esperimentu egin badira ere, oraindik ez da sortu corpusetatik terminoen arteko erlazioak erauzi eta ezagutza-base lexikal bat hutsetik sortu edo dagoeneko badagoen bat aberasten lagunduko duen hizkuntza-produkturik.

3 Metodologia

Ataza honen barnean atal esperimentalean erabili ditugun baliabideak azaltzen dira.

3.1 Corpusak

Esan bezala, proiektu honen helburua WordNet era automatikoan hedatzea da, honek duen estaldura handitu ahal izateko. WordNeten atal espezializatuak osatzea da zehazki egin nahi dena eta horretarako Zientzia eta teknologiaren Corpusa erabili da.

Zientzia eta Teknologiaren Corpusa (ZTC) zientzia eta teknologiaren alorreko euskarazko testu-bilduma egituratu eta etiketatua da, eta alor horietako euskararen erabilera ikertzeko baliabidea izatea du helburu nagusia. Corpus berezi edo espezializatua da, eta UPV/EHUko IXA taldeak eta Elhuyar Fundazioak elkarlanean eratu dute. Corpusaren tamaina 8,5 milioi hitzekoa da.

ZT corpusa hainbat eremu ezberdinetan banatuta dago: zientzia zehatzak, materiaren eta energiaren zientziak, lurraren zientziak, biziaren zientziak, teknologia, orokorra, bestelakoak.

Corpus etiketatua da, bai testuaren egiturari eta formatuari dagokionez, bai linguistikoki. Etiketatze linguistikoa egiteko, euskara automatikoki prozesatzeko teknologia aurreratua erabili da (IXA taldearen *Eustagger* etiketatzailea). Testuko hitz bakoitzaren lema eta kategoria/azpikategoria etiketatu dira.

3.2 Perl programazio lengoaia

Proiektu hau burutzeko erabilitako lengoaia Perl(Practical Extraction and Report Language) da. Programazio lengoaia hau Larry Wallek sortu zuen UNIX sistemetako administrazio lanak sinplifikatzeko asmoz. Hala ere, gaur egun helburu orokorretarako erabiltzen den lengoaia bihurtu da Perl. Gainera, gaur egun sistema eragile ezberdinetara egokitua izan da, besteak beste, MacOS, Windows edo Amiga.

Proiektu honetarako Perl aukeratu izanaren arrazoi nagusia, lengoaiak testuak eta fitxategiak prozesatzeko eskaintzen duen erosotasuna da.

Ia edozein plataformatara egokitua izan da scriptetan oinarritutako lengoaia hau. Dituen osagai garrantzitsuenetako bat adierazpen erregularrak dira, hauekin lana egiteko eskaintzen duen kudeaketa ona dela eta.

Ez da dotorea izateko egindako lengoai bat, praktikoa izateko egindakoa baizik. Hori dela eta, script dotoreak izatea nahi bada, kode txukun bat idatzi nahi bada, kontutan izan beharko da programatzeko modua.

Perlek duen abantaila eta aldi berean desabantaila programazio filosofiarik ez ezartzea da. Ezin daiteke esan objektuetara bideratutako lengoaia, lengoaia estrukturala edo modularra denik, nahiz eta paradigma guzti hauek onartzen dituen.

3.3 Weka

Lema-lema bikote batek hiponimia/hiperonimia erlazioa duen zehazteko, sailkatzaile bat egin dugu Weka (Witten eta Frank, 2005) tresnaren bidez. Datu-meatzaritzako atazetarako erabiltzen diren ikasketa automatikoko algoritmoen multzoa da Weka. Edozein plataformatan (Windows, Linux, Mac Os...) erabil daiteke eta software librea da.

Wekak ikasketa automatikoa burutzeko aukera ezberdinak ematen ditu: sailkatzaile ugari, ezaugarrien trataera, interfaze grafikoa...

Wekarekin lana egiteko datuak arff formatura bihurtu beharra dago.

Aipatu bezala, wekak sailkatzaile ezberdinak eskaintzen ditu eta zazpi multzotan daude banatuta:

- Bayes: Bayes funtzioaren aldaera ezberdinak erabiltzen dira sailkapena egiteko.
- Functions: Funtzio ezberdinetan oinarritzen diren sailkatzaileak daude multzo honetan.
- Lazy: Instantzietan oinarritzen diren sailkatzaileak biltzen dira
- Meta: Sailkatzaile ezberdinen konbinazioak biltzen ditu multzo honek.
- Rules: Erregeletan oinarritutako sailkatzaileak biltzen dira multzo honetan.
- Trees: Zuhaitzetan oinarritzen diren algoritmoak daude atal honetan.
- Misc: Hiperpipes, VFI eta SerializedClassifier daude.

Lan honetan, sailkatze-algoritmo ezberdinak aztertu ditugu, emaitza onenak zeintzuk ematen dituzten ikusteko. Aukeratutako sailkatze algoritmoak honakoak dira:

- JRIP erregeletan oinarritutako sailkatzailea

- Naive Bayes sailkatzaile baiesiarra
- J48 erabaki zuhaitza C4.5 zuhaitzaren inplementazioa da.
- SMO oinarri bektoreetan (SVM) oinarritzen den sailkatzailea
- IB1 Knn alkoritmoan oinarritutako sailkatzailea da, non K-ren bakioa 1 den. Funtsean ikasketako instantzietatik hurbilen dagoena darabil sailkapenerako.
- Adaboost sailkatzaileak hainbat sailkatzaile konbinatzen ditu. Sailkatzaile berri bat erabiltzen du iterazio bakoitzean eta aurrekoan gaizki sailkatutako adibideak gehiago lantzen ditu, pisuak egokituz.
- VFI (Voting Feature Intervals) algoritmoak ezaugarri tarteak sortzen ditu. Tarte horietako bakoitzak ezaugarriari dagokion balio multzo bat adierazten du.

4 Gure hurbilpena

Atal honetan egin diren esperimentazioak zeintzuk izan diren zehazten da. Helburua hiponimo-hiperonimo bikoteak erauzteko sistema bat da. Beraz, lehen pausoa testuetan erlazioak aurkitu eta markatzaile linguistiko batzuk definitzea izan da. Ondoren, markatzaile linguistiko horien bidez bikote hautagaiak lortuko dira, eta markatzaileen jokaera aztertuko da. Jarraian azaltzen diren puntuetan prozesu hau era sakonagoan azaltzen da.

4.1 Termino habiaratuak

Ataza honetan, terminologia-erauzketa automatikoan termino-habiaratzearen prozesamenduaren bidez eskuratzen den informazioa (sintagma-azpisintagma bikoteak) aztertu da hiperonimia/hiponimia erlazioa ezartzeko. Horretarako, Erauzterm (Gurrutxaga et al., 2005) termino-erauzlea erabili da. Erauzterm egokitua izan da haren irteeran termino-aldaeren eta habiaratutako terminoen arteko kontzeptu-erlazioen informazioa emateko

Erauztermek erauzitako termino bikoteen adibide batzuk 3. taulan aurkitu daitezke.

Sintagma nagusia (hiponimo hautagaia)	**Azpisintagma (hiperonimo hautagaia)**
Eremu magnetiko	eremu
Erreakzio kimiko	erreakzio
Alderdi plastiko	alderdi
grabitate-zentro	grabitate
Aurkako noranzko	noranzko
Uhin bidaiari	uhin
Milioi argi-urte	argi-urte

3 taula. Erauztermek erauzitako termino habiatuak

Erauztermek ematen dituen termino habiatuen hiperonimo-hiponimo doitasuna neurtu da. Horretarako, corpusetako sintagma-azpisintagma izen-bikoteak atera ditugu, eta sintagma-azpisintagma bikote horiek hiperonimo-hiponimo bikoteak diren ala ez egiaztatu dugu. Azterketa hau ZT corpuseko Materia eta Energiaren Zientziak atala eta Elektrizitatea eta Elektronika atala erabiliz burutu da.

4.2 Markatzaile linguistikoak

Markatzaile linguistikoak aztertzea da beste bide bat. Horretarako lehenik eta behin corpusa aztertu eta markatzaile horiek definitu beharko dira. Jarraian, markatzaileak baliozkoak direla frogatu beharko da, hauekin lortutako lema-lema bikoteen doitasuna aztertuz. Era honetan markatzaile egokiak zeintzuk diren jakingo da, eta zeintzuk baztertu behar diren ere bai.

4.2.1 Markatzaile linguistikoak definitu

Patroiak definitzeko lehen pausoa, Corpusetik esaldi berean gertatzen diren lema-lema bikoteak erauztea da. Ondoren Euskal WordNetekin alderatuz, hiponimia/hiperonimia erlazioa duten bikoteak aurkitu dira.

Euskal WordNeten estaldura txikia dela eta, WNTerm-en(Pociello et al., 2008) gaur egun landuta dauden eta oraindik WordNeten gehitu gabe dauden terminoak ere hartu dira kontutan azterketa honetan. WordNeten hedapen honekin hasierako bikote aztergai gehiago lortzea izan da asmoa.

Hasiera batean, patroiak definitzeko asmoz aztertutako esaldi batzuen adibideak aurkezten dira jarraian:

1. Marrazki bikote hauek **estereoskopio** izeneko **gailu** berezi baten bidez begiratzen dira.
    - Hiperonimoa: gailu
    - Hiponimoa: estereoskopio
    - Markatzaile linguistikoa: A izeneko B
2. Beste **zenbaki** (**ezaugarri**) garrantzitsu bat spin momentu angeluarra da.
    - Hiperonimoa: ezaugarri
    - Hiponimoa: zenbaki
    - Markatzaile linguistikoa: A (B)
3. ... bitamina eta uraz eta liseritu ezin diren **zuntz** moduko beste **substantzia** batzuez.
    - Hiperonimoa: zuntz
    - Hiponimoa: substantzia
    - Markatzaile linguistikoa: A moduko beste B
4. Egoera-ekuazio guztietan masa / erradio erlazio handienak erabiltzen dituztenek bakarrik onartzen dute 2000 hertzeko **biraketa-abiadura** eta, horien artean ere, bere masa zulo beltzen mugatik oso hurbil duten pultsarrek lor lezakete aipatutako **abiadura**.
    - Hiperonimoa: abiadura
    - Hiponimoa: biraketa-abiadura
    - Markatzaile linguistikoa: ?

Aurreko adibideetan markatzaile ezberdinak ikus ditzakegu ZT corpusean aurkitutako esaldietan. Kasu batzuetan markatzaileak linguistikoak aurkitzea ez da begi-bistakoa, informazio sintaktikoa beharrezkoa da (4. adibidea). Beste batzuetan ordea, argi eta garbi nabarmentzen dira markatzaile linguistikoak (1, 2 eta 3 adibideak), eta horiek dira guk bilatu nahi ditugunak.

Behin esaldi berean erlazioa duten lema-lema bikoteak identifikatu ondoren, hiperonimo-hiponimo erlaziorik adierazten ez duten esaldiak eskuz baztertzea da hurrengo pausoa. Hiperonimo-hiponimo erlazioa duten bikoteak hartuta, beraien testuingurua atera, eta markatzaile linguistiko posibleak bilatu daitezke. Lan hau eskuz egin da, ZT Corpuseko Materia eta Energiaren Zientziak atala erabiliz.

ZT Corpuseko Materia eta Energiaren Zientziak atalak, guztira 615.000 hitz ditu. Corpus osoa teknikoa izanik, zientziaren eta teknologia alor ezberdinak jorratzen dituzten gainontzeko ataletan ere egitura antzekoak errepikatuko direla aurreikusten da.

Aurkitutako markatzaile linguistikoak definitzeko hitzen POS datuak erabili dira (lema eta kategoria izan dira oinarrizkoenak). Aurrekarietan aipatutako hainbat metodok erlazionatutako bikoteak aurkitzeko analisi sintaktikoa ere erabiltzen dute. Hala ere, euskararen kasurako analizatzaile sintaktiko eskuragarririk ez dagoenez, informazio hori gehitzea etorkizunerako pauso bezala utziko da, eta proiektu oraingoz ez da kontutan izango.

4.2.2 Patroiak betetzen dituzten hautagaiak erauzi

Behin markatzaile linguistikoak definituta ditugunean, markatzaile horiek testuan bilatu, eta patroiak betetzen dituzten lema-lema bikoteak bilatzeko ordua da.

Horretarako corpusa esaldiz-esaldi aztertuko duen Perl script bat garatu da. Script horrek alde batetik markatzaile linguistikoak jasotzen ditu, eta bestetik etiketatutako corpus bat. Ondoren esaldi guztiak aztertzen ditu markatzaile horietakoren bat duten ikusteko. Esaldi bakoitza behin bakarrik pasatuz gero exekuzio denbora azkarragoa izango da, eta corpus tamainak kontutan izanda atal hau garrantzitsua izan daiteke.

Atal honen bukaeran markatzaileren bat duten esaldiak eskuratuko dira, eta markatzaile hori betetzen duten lema-lema bikoteak ere bai. Beraz, hiponimo-hiperonimo hautagaiak eta dagozkien markatzaileak erauziko dira esaldi bakoitzetik. Beti ere, markatzaile linguistikoren bat duten esaldietatik lortuko dira datu hauek eta markatzailerik ez dutenak baztertu egingo dira.

4.2.3 Patroien jokaera aztertu

Testuan aurkitutako markatzaile guztiak aztertuta ditugu eta hiponimo-hiperonimo hautagaiak erauzita.

Beraz, hiponimo-hiperonimo hautagai bikote bakoitzeko betetzen diren markatzaileak pilatu ditzakegu, eta horien arabera bikote horrek hiponimo-hiperonimo erlazioa duen esan ahal izango dugu. Era honetan, aztergai dugun corpuserako markatzaile esanguratsuenak zeintzuk diren ondorioztatzeko aukera dago. Gainera bikote bakoitzak markatzaile ezberdinak betetzeko duten maiztasunak emaitzetan ondoriorik izan dezakeen ere aztertzeko aukera izan dezakegu. Hau da, maiztasunaren datua garrantzitsua den jakiteko, edo markatzaileen presentziarekin nahikoa den erabakitzeko.

Definitutako markatzaileen erabilgarritasuna aztertzea da, beraz, puntu honen helburu nagusia.

4.3 Teknika estatistikoak

Hiperonimia/hiponimia erlazioa hobeto rankingeatzen duten teknikak eta antzekotasun distribuzionaleko neurriak konparatiboki ebaluatzea da atal honen helburua.

Horretarako, LLR neurriaren arabera esanguratsuenak diren 70 termino bakunak prozesatu dira, eta horiekin antzekotasun distribuzional handiena duten 100 hitzak hiponimo edo hiperonimoak diren aztertu da. Hautagaien arteko antzekotasuna neurtzeko LSA neurria erabili da.

LLR (likelihood ratio) testa, bi modeloren artetik egokiena zein den aukeratzeko erabiltzen den test estatistiko bat da. LLR neurriak datu bat modelo horietako batean edo bestean egoteko aukera zein den esaten digu. Gure kasuan corpus espezializatu baten eta corpus orokor baten artean egiten da konparaketa. Beraz, LLR neurri handiena duten hitzak termino espezializatuagoak izango dira, corpus zientifikoari lotuagoak daudenak. Neurri honen bidez corpusaren arloari dagozkion hitzak aukeratzea lortuko dugu.

Ondoren LSA neurriaren bidez lortutako termino horien antzekoenak erauzten dira. LSA hizkuntzaren prozesamenduan erabiltzen den teknika bat da. Dokumentuen eta hauetan dauden terminoen analisi bat egiteko erabiltzen da teknika hau. Termino bakoitzaren testuinguruari dagozkion datuak bektore batean biltzen dira. Ondoren, bi

terminoren testuinguru-antza neurtzeko bi bektoreen antzekotasuna erabiltzen da. Beraz, lema-bikote bakoitzeko hau kalkulatzen da eta ranking bat sortu.

Testuinguru-antzaren azterketa egiteko rankingean lotutako bikote denak eskuz behatu dira banan-banan. Hiponimia/hiperonimia erlazioa duten bikoteak eta ez dutenak banandu dira.

4.4 Markatzaileak eta teknika estatistikoak

Hizkuntza-teknologien beste hainbat alorretan bezala, hemen ere metodo estatistikoen eta linguistikoen arteko lehiaren / kooperazioaren auzia dago. Azkenaldian, gero eta joera argiagoa nabari da kooperazioaren alde. Literaturan bi era horietako teknikak proposatu eta erabiltzen dira, eta erraz aurkitu daitezke antzekotasun distribuzionala eta markatzaile linguistikoak konbinatzen dituzten sailkatzaileak. (Cederberg and Widdows, 2003; Giovannetti et al., 2008) lanetan patroi lexiko-sintaktikoak erabiliz lortutako emaitzak hobetzea lortzen dute sailkatzaileari testuinguru antza neurtzen duten balioak gehituta. Beraz, markatzaileez gain antzekotasun distribuzionalari dagozkion datuak ere gehitu dira, hauek emaitzetan izan dezaketen eragina zenbaterainokoa den ikusteko.

5 Emaitzak

Atal honetan testutik hiperonimo-hiponimo bikoteak erauzteko egindako esperimentuetan lortutako emaitzak azaltzen dira. Aipatu bezala hiru bide ezberdin aztertu dira: termino habiatuak, markatzaile linguistikoak eta testuinguru-antza. Beraz, bide bakoitzean lortutako emaitzak aipatzen dira jarraian.

5.1 Termino habiaratuen azterketa

Termino habiaratuak aztertzeko Erauztermekin (Alegria et al., 2005) erauzitako terminoak hartu dira. Aurretik esan bezala, kasu honetan ZT corpuseko *Materiaren eta energiaren zientziak* atala erabili da.

Lortutako emaitzak automatikoki eta eskuz aztertu dira:

- **Automatikoki:** Metodo estatistikoak erabilita, eta erreferentzia gisa, Euskal WordNet eta WNTERM ezagutza-baseak hartuz, hiponimo-hiperonimo hautagaien artean erlaziorik dagoen definitu da. Bilaketa automatiko horretan WordNeteko 3 maila erabili dira erlazionatutako terminoak bilatzeko.
- **Eskuz:** Erlazionatutako izen bikoteak ea hiperonimo-hiponimo bikoteak diren ala ez eskuz egiaztatuta.

Atal honetan ditugun Erauztermeko irteeren informazioa honakoa izango da:

Sintagma nagusia; Eremu magnetiko

Azpisintagma: Eremu

- Sintagma nagusiaren LLR: Sintagma nagusiak(Eremu magnetiko) corpusaren alorreko termino espezializatua izateko duen aukera LLR neurriaren arabera.
- Sintagma nagusiaren maiztasuna: “Eremu magnetiko” sintagma zenbat aldiz agertzen den aztergai dugun corpusean.
- Azpisintagmaren LLR: Azpisintagmak (Eremu) corpusaren alorreko termino espezializatua izateko duen aukera LLR neurriaren arabera.

  Azpisintagmaren maiztasuna: “Eremu” azpisintagma zenbat aldiz agertzen den aztergai dugun corpusean.

Eskuz eta automatikoki lortutako habiaratutako terminoen emaitzak ebaluatu ditugu bide hau hiponimia/hiperonimia erlazioa lortzeko egokia den ikusteko. Eskuzko ebaluazioa eta ebaluazio automatikoa egiteko sintagma-azpisintagma bikoteetako kide bakoitzaren Likelihood-ratio (LLR) neurrian eta maiztasunean oinarritutako rankingak egin dira, eta ranking horietako lehen 1000 bikoteak aztertu dira (ikus 4. eta 5. taulak).

Rankingak LLR eta maiztasunen arabera egin dira, termino espezifikoa izateko aukera asko dituen sintagmak eta sarritan errepikatuta agertzen diren sintagmak emaitza hobeak emango dituelako. 4. eta 5. tauletan bildutako emaitzak, ZT corpuseko alor ezberdinetarako sintagma-azpisintagma bikoteentzat sortutako rankingetan dauden hiponimo-hiperonimo bikoteen ehunekoa adierazten dute.

	WordNetekin emaitza (%)	**Eskuzko emaitza (%)**
Sintagmaren maiztasuna	2,1	19,7
Sintagmaren LLR	2,1	23,3
Azpisintagmaren maiztasuna	1,1	12
Azpisintagmaren LLR	0,7	7,3

4 taula: Materia eta Energiaren Zientziak ataleko sintagma-azpisintagma probabilitate-neurketa ehunekotan

	WordNetekin emaitza (%)	**Eskuzko emaitza (%)**
Sintagmaren maiztasuna	0,8	14,7
Sintagmaren LLR	0,7	13,9
Azpisintagmaren maiztasuna	0,6	7
Azpisintagmaren LLR	0,2	6,2

5 taula: Elektrizitatea eta Elektronika ataleko sintagma-azpisintagma probabilitate-neurketa ehunekotan

Emaitzetan ikus daitekeenez, sintagma-azpisintagma bikoteen azterketan argi gelditzen da WordNeten estaldura hobetu beharra dagoela. Eta hori da lan honen helburua hain zuzen ere.

Hala ere, eskuzko emaitzak ere hobetu beharra dute. Argi dago, rankingak egiteko orduan sintagma osoari dagozkion datuak hartu behar direla kontuan, baina bere horretan hartuta emaitzak ez direla nahiko onak. 4. eta 5. tauletako emaitza horiek birfintzeko sintagmen egiturak mugatu ditugu; hau da, izen-izen (NN) eta preposizio-izen (AprepN) egiturak hartuta, eskuzko azterketa errepikatu dugu birfinketa horiek egiteko (ikus 6. eta 7. taulak).

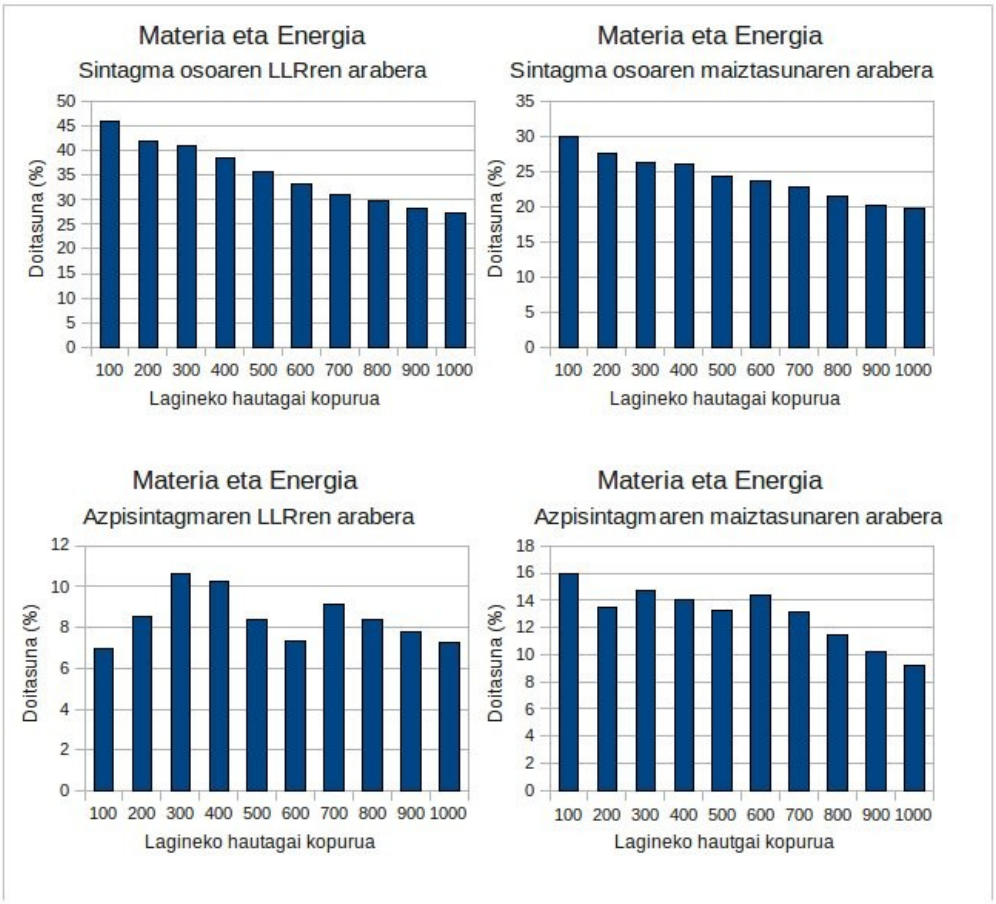
	Datu guztiak (%)	**AprepN eta NN egiturak (%)**
Sintagmaren maiztasuna	19,7	19,7
Sintagmaren LLR	23,3	27,4
Azpisintagmaren maiztasuna	12	9,2
Azpisintagmaren LLR	7,3	7,3

6 taula: Materia eta Energiaren Zientziak ataleko sintagma-azpisintagma probabilitate-neurketa

	Datu guztiak (%)	**AprepN eta NN egiturak (%)**
Sintagmaren maiztasuna	14,7	17,7
Sintagmaren LLR	13,9	19,7
Azpisintagmaren maiztasuna	7	5,6
Azpisintagmaren LLR	6,2	6,2

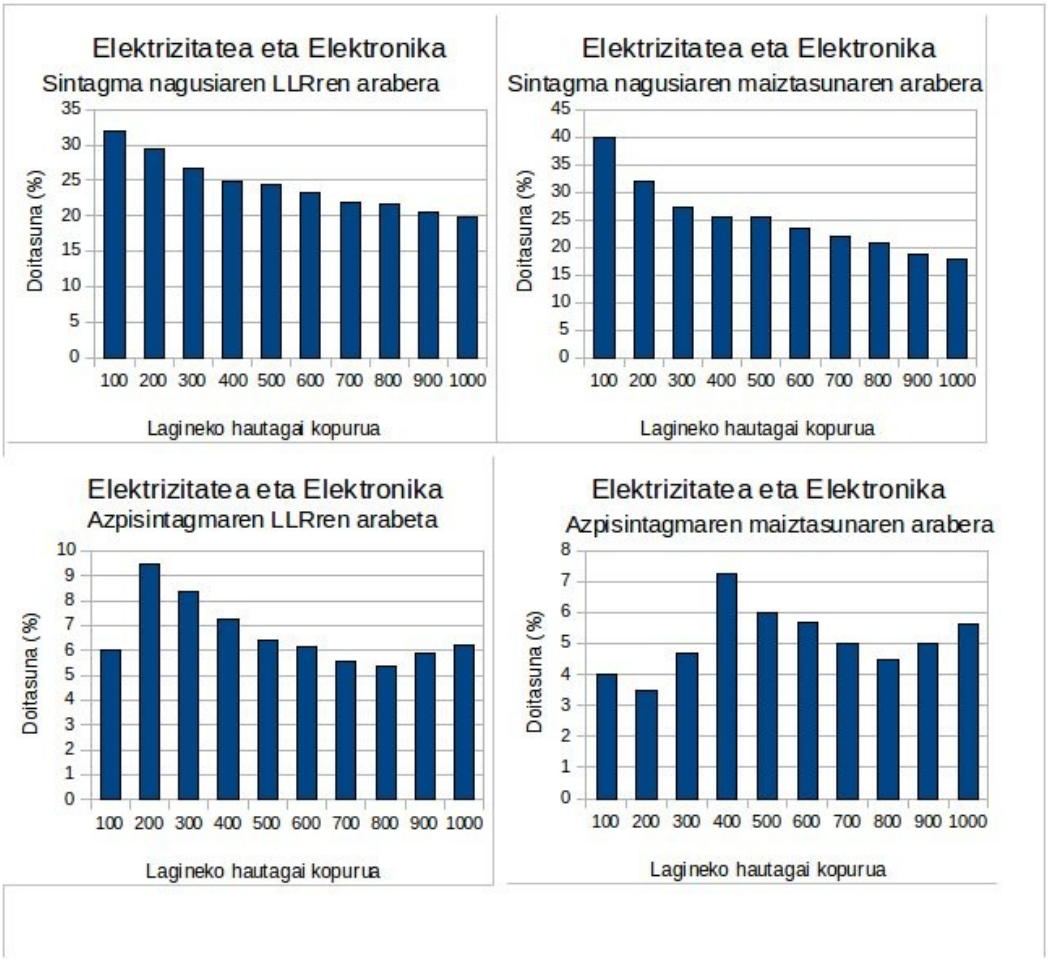
7 taula: Elektrizitatea eta Elektronika ataleko sintagma-azpisintagma probabilitate-neurketa

Bestalde, hartutako laginaren tamainak izan dezakeen garrantzia ere aztertu dugu, eta tamaina handitu ahala aurkitutako hiponimo-hiperonimo bikote kopuruak nolako joera duen ikusi daiteke. 5. irudian Materia eta Energiaren Zientziak atalean lortutako emaitzak ikus daitezke, eta 6. irudian Elektrizitatea eta Elektronika alorrean lortutakoak.



5. irudia. Materia eta Energiaren Zientziak atalean termino habiatuek hiponimo-hiperonimo izateko aukeraren bilakaera lagin tamainaren arabera

5 eta 6 irudietan ikus daitekeenez, sintagma nagusiaren (hiponimo hautagaiaren) maiztasuna edota LLR neurria aztertuz gero, lagin tamaina zenbat eta txikiagoa izan hiponimo-hiperonimo ehunekoa orduan eta handiagoa da. Azpisintagmaren (hiperonimo hautagaiaren) arabera emaitzak aztertuz gero, ez dago erlazio zuzenik hartutako tamainaren eta erlazio kopuruaren artean. Emaitza onenak sintagma nagusiaren LLR neurriaren arabera lortzen direnak dira beti ere. Beraz, LLR handia duten bikoteen kasuan lortutako hiponimo-hiperonimo bikote kopurua esanguratsua izan daiteke.



6. irudia. Elektrizitatea eta Elektronika atalean termino habiatuek hiponimo-hiperonimo izateko aukeraren bilakaera lagin tamainaren arabera

Argi dago termino habiaratuei dagokienean lortutako emaitzak lagungarriak direla WordNet aberastu ahal izateko. Batez ere, eskuzko azterketan aurkitutako termino asko ez daudelako WordNeten. Hala ere, elikatze prozesu hau guztiz automatikoa egiteko informazio gehiago behar da, eta froga gehiago egin beharko dira.

5.2 Markatzaile linguistikoak

Markatzaile linguistikoei dagokienean, aipatu bezala, lehenik eta behin markatzaile linguistikoak identifikatu beharra izan da. Ondoren, markatzaile horiekin hiponimo/hiperonimo hautagaiak bilatu dira eta azkenik lortutako emaitzak ebaluatu dira.

5.2.1 Hiponimia-markatzaileen definizioa

Hasierako pauso honetan patroi zehatzak eta orokorrak bildu dira. Hasiera batean patroi posible guztiak definitu eta aurrerago baliagarriak ez direnak baztertzeko asmoz.

Patroia	Adibidea
A izeneko B	Molekulak atomo izeneko zati txikiagoz osatuta daude
Ari B deitzen zaio	Ur molekulen arteko lotura horri kohesio deitzen zaio.
A esaten zaio Bri	Gasak igortzen duen argi horri izpi katodiko esaten zaio.
A deritzo Bri	Sistemaren erresonantzia deritzo fenomeno horri.
A B da	Soinuen fisika oso gai konplexua da.
A (B)	Ziklo horretan zehar , gizakiak energia ateratzen ikasi du : erregai solidoak (egurra) , likidoak (alkohola) eta gasa (biogasa).
A : B1, B2	Ostadarrak jarraikako kolore zerrenda hau erakusten du : gorria , laranja , horia , berdea , urdina , anila eta morea.
hainbat A erabili zituzten, B1, B2	Urreaz gain , hainbat metal erabili zituzten , aluminioa , burdina eta beruna esaterako.
A, B,	John Fleming-ek asmatu zuen lehenengo irrati hodia , diodoa , 1904an.
A, hots, B	Adierazpen horretan eta hurrengoetan unitate atomikoak erabili dira , hots , elektroiaren karga , masa eta Planck-en konstantea unitatetzat hartu dira.
A, hau da, B	Nahiz eta egoera batzuetan metal puruak erabili, gehienetan aleazioak erabiliko dira , hau da, metal ezberdinen arteko nahasketak.
A eta, orohar, B	Gaur egunean ba dakigu etere beharrik ez dagoela, eta argia eta, orohar, erradiazio elektromagnetikoa hutsean ere hedatzen dela.
A eta, zehazki, B	Oso garrantzitsua da ulertzea eta azaltzea zientzia eta, zehazki, fisika nonahi dago eta.
A, hala nola B	Hipotesi horren nahiko erraz azaltzen zituen argiaren zenbait fenomeno , hala nola islapena , errefrakzioa eta garai hartan aurkitu berria zen errefrakzio bikoitza.
A, B esaterako	Urreaz gain , hainbat metal erabili zituzten , aluminioa , burdina eta beruna esaterako.
A moduko B	Makinen ezaugarri naturalak garatzeko bidea irekia dago , eta noizbait filmetan agertzen diren androideen moduko robotak egingo dira lan horiek aurrera eginez gero.
A bezalako B	balantza eta termometroa bezalako neurgailuak kontuz erabiltzea.
A antzeko B	Erakutsi ere egiten dute uhin-funtzioa ez dela eremu erreal bat eta bere bat-bateko aldaketa (proiekzioa) ez dela eremu baten aldaketa antzeko prozesu fisiko bat.
A eta beste hainbat B	Sistemaren energia , karga elektrikoa , abiadura eta beste hainbat ezaugarri nolakoak diren galde dezakegu , eta erantzun guztiak uhin-funtzioan dauErwin Schr dinger.
A, Bz aparte	Lurretik gertuen dagoen izarra , Eguzkiaz aparte , Zentauroko Alfa izenekoa da.
A guztiak B izan ezik	Bi aukera besterik ez daude : hiru quarkek barioia osatzea ala quark batek eta antiquark batek mesoia osatzea (gogoratu zatiki baten antizatikiak haren ezaugarri guztiak masa izan ezik aurkakoak dituela)
A edo B	Zenbait iruzkin egingo ditugu lege hauei buruz , norainoko legeak diren ikusteko eta dauzkaten zenbait hutsune edo akats azpimarratzeko.
A eta B	Arazoak eta oztopoak aukera bilakatzen eta nire helburuei eusten irakatsi didate pertsona horiek.

8 taula: markatzaileak lortzeko esaldien adibideak, ZT corpusetik erauzita

Lortutako markatzaile linguistikoak bilatzeko erabilitako esaldien adibide batzuk biltzen dira hurrengo taulan 8. taulan.

Aurkitako hiponimo-hiperonimo bikoteen testuinguruak aztertuz gero argi ikus daiteke lortutako egiturak ingeleserako aurkeztutakoen antzekoak direla. Adibidetzat hartutako esaldiok jarraitzen dituzten markatzaileak Hearstek definitu eta Snowet al.-ek (2005) osatutako patroiekin parekatu ditzakegu. Kasu batzuetan berdinketa horietan markatzaile bat baino gehiago aurkitu daitezke euskararen kasurako, markatzailearen hitz-gakoak sinonimoengatik aldatu baitaitezke (ikus 9. taula).

Hala ere, guk aurkitutako testuinguru batzuk taulatik kanpo gelditzen dira “A edo B” markatzailea esate baterako (8. taula). Markatzaile hauek oso orokorrak diren arren, eta beste hainbat erlazio definitu ditzaketela argi eta garbi ikusi arren, kontutan hartzeak emaitzetan eragina izan dezakeenez, markatzaile orokor hauen jokaera ere aztertu dugu.

Hearsten patroiak	**Eskuz aztertutako esaldietan aurkitutako patroiak**
NPX and other NPY	A eta beste B A eta beste hainbat B
NPX or other NPY	A edo beste B
NPY such as NPX	A moduko B A bezalako B
NPY , especially NPX	A eta, orohar, B A eta, zehazki, B
NPY like NPX	A antzeko B A gisako B
NPY called NPX	A izeneko B Ari B deitzen zaio
NPX is a NPY	A B da
NPX , a NPY	A, B,

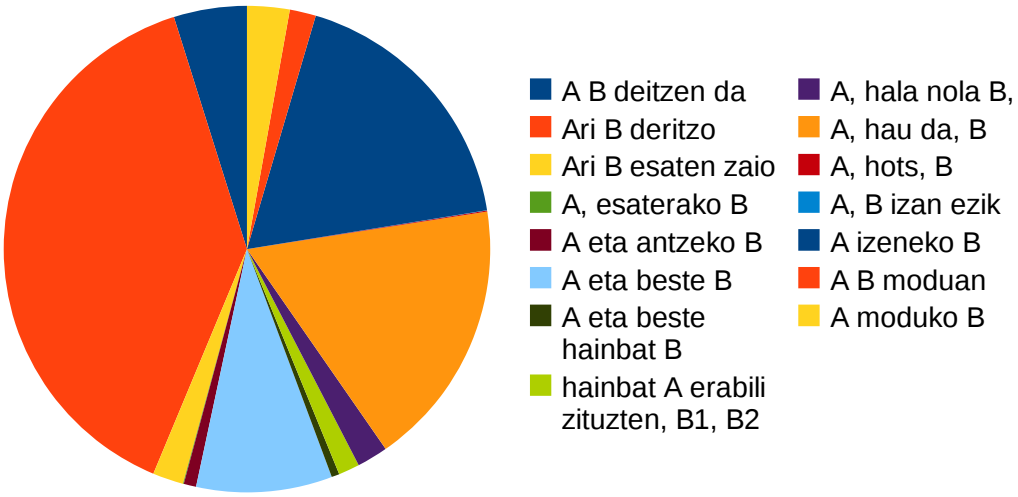
9 taula: Hearsten patroien eta euskararako aurkitutako patroi hautagaien antzekotasuna

Markatzaile hauekin ZT corpusa prozesatu ostean lortutako esaldien datuak ikus daitezke ondorengo taulan (10. taula).

Markatzaile linguistiko denak erabiliz lortutako esaldi kopurua	594563 esaldi
Markatzaile ez-orokorrak erabiliz lortutako esaldi kopurua	118162 esaldi
Lortu diren hautagai bikoteak	57286 bikote

10 taula. Markatzaile linguistikoekin ZT Corpusetik lortutako hautagai kopuruak

Aurreikustekoa den bezala, patroi orokorrenak dira gehien agertzen direnak. 10. taulan ikus daitekeenez aurkitutako esaldien bostena inguru dira benetan hiponimia/hiperonimia markatzaile linguistikoak betetzen dituztenak. Horiek aztertuz gero, markatzaileek ZT corpusean duten banaketa 7. irudian agertzen dena da.



7. irudia. Markatzaile linguistikoen banaketa ZT Corpusean

"Ari B deritzo", "A izeneko B" eta "A,hau da, B" dira gehien agertzen diren markatzaileak, eta gainontzeko balioak gutxiago agertzen dira. Hala ere, corpusaren tamaina handituz gero, edo beste era bateko corpusa erabiliz gero patroiek duten banaketa ezberdina izango da ziurrenik.

5.2.2 Markatzaileen maiztasuna eta presentzia

Patroien jokaera orokorra aztertu ondoren, egin den lehen gauza hiponimo/hiperonimo hautagaiak aztertzea izan da. Lortutako hautagai guztiak eskuz sailkatzea lan handia denez, lagin bat hartu da froga egiteko eta eskuz aztertu da zeintzuk duten hiponimia erlazioa.

Esperimentuetarako erabilitako laginak 830 lema-lema bikote ditu, erdiek hiponimia/hiperonimia erlazioa dute, eta beste erdiek ez. Hautagaiek erlazioa duten edo ez definitzeko lagin bat hartu eta eskuz aztertu dira hautagai guztiak, banan-banan.

Lagin orekatu bat erabili izanak badu bere arrazoia. Izan ere, bikote-erauzleak emandako emaitzetan erlaziorik ez duten bikote ugari lortzen dira, eta ausazko banaketa duen lagin bat hartuz gero erlaziorik ez duten bikoteek pisu handiegia hartzen dute sailkapenean.

Jarraian, patroien egokitasuna aztertzeko, hautagai bikote bakoitzaren agerpen maiztasuna eta patroietako bakoitza betetzen duen aldi kopurua kontutan izan dira sailkatzaile bat sortzeko. Horretarako hasieran aipatutako weka tresna erabili da, eta sailkatzaile mota ezberdinen emaitzak konparatu dira egokiena zein izan daitekeen aztertzeko.

Wekari pasatako balioak, beraz, honakoak izan dira, hautagai bikote bakoitzeko:

- Patroien betetze maiztasuna (corpusean aurkitutako balio bakoitzeko atributu bat)
- Bikoteko hitzak esaldi berean agertzeko duten maiztasuna
- Bikoteko hautagai bakoitzak bere aldetik corpusean duen maiztasuna.

	Erlazionatuen doitasuna	**Erlazionatuen estaldura**	**Ez-erlazionatuen doitasuna**	**Ez-erlazionatuen estaldura**	**Zehaztasuna**
Naïve Bayes	0,53	0,88	0,64	0,33	0,55
SMO	0,52	0,8	0,58	0,27	0,54
Ibk (k=1)	0,58	0,59	0,58	0,58	0,58
AdaBoostM1	0,59	0,67	0,62	0,54	0,6
VFI	0,68	0,41	0,58	0,8	0,61
JRip	0,61	0,59	0,6	0,62	0,6
J48 tree	0,65	0,55	0,61	0,71	0,63

11 taula: Markatzaile linguistikoen sailkapenean lortutako emaitzak, maiztasunari dagozkion datuak erabiliz.

Froga maiztasunaren ordez presentzia erabilita ere errepikatu da. Izan ere, (Cimiano et al. 2004, Ryu et al. 2006) lanetan aipatzen den bezala, patroien maiztasuna oso txikia da corpusean zehar. Beraz, presentzia edo maiztasuna erabiltzeak emaitzetan izan dezakeen eragina aztertu nahi izan da.

	Erlazionatuen doitasuna	**Erlazionatuen estaldura**	**Ez-erlazionatuen doitasuna**	**Ez-erlazionatuen estaldura**	**Zehaztasuna**
Naïve Bayes	0,6	0,75	0,66	0,49	0,62
SMO	0,67	0,52	0,61	0,75	0,63
Ibk (k=1)	0,61	0,59	0,6	0,61	0,6
AdaBoostM1	0,6	0,66	0,62	0,55	0,61
VFI	0,66	0,46	0,58	0,76	0,61
JRip	0,6	0,6	0,6	0,6	0,6
J48 tree	0,64	0,62	0,63	0,66	0,64

12 taula: Markatzaile linguistikoen sailkapenean lortutako emaitzak, presentziari dagozkion datuak erabiliz.

11. eta 12. taulak konparatuz gero maiztasuna zein presentzia erabiliz lortutako emaitzak oso antzekoak dira. Arrazoia, markatzaileek duten maiztasun eta presentzia txikia izan daiteke. Hau da, hiponimo-hiperonimo hautagai bakoitzak patroi bera behin eta berriz errepikatzea ez da askotan gertatzen, eta bikote berdinak patroi ezberdin ugari errepikatzea ere ez.

5.2.2.1 Markatzaile orokorrek emaitzak hobetzen dituzte?

Markatzaileak definitu ditugun momentuan, Hearstek definitutako markatzaile zehatzekin parekatu daitezkeen patroiak genituen alde batetik, eta patroi orokorrak bestetik. Patroi orokor horiek mantendu egin dira, nahiz eta jakin badakigun ez dutela hiponimia/hiperonimia erlazioa soilik definitzen. 13. taulan ikus daitezke patroi orokorrak erabili gabe presentzia datuekin lortutako emaitzak.

	Erlazionatuen doitasuna	**Erlazionatuen estaldura**	**Ez-erlazionatuen doitasuna**	**Ez-erlazionatuen estaldura**	**Zehaztasuna**
Naïve Bayes	0,57	0,63	0,58	0,51	0,56
SMO	0,76	0,21	0,54	0,93	0,57
Ibk (k=1)	0,59	0,61	0,6	0,58	0,6
AdaBoostM1	0,56	0,68	0,6	0,48	0,58
Vfi	0,81	0,15	0,53	0,96	0,56
JRip	0,56	0,58	0,57	0,55	0,56
J48 tree	0,54	0,75	0,59	0,36	0,57

13 taula. Markatzaile orokorrak gabe presentzia datuak erabiliz lortutako emaitzak.

12. eta 13. taulak alderatuz, sailkatzaile denekin lortutako datuetan argi ikusten da patroi orokorren presentzia kontuan hartzeak emaitzak hobetzen dituela. Gainera, alde hori batez ere hiponimia/hiperonimia erlazioa duten bikoteen estalduran ikus daiteke, beraz markatzaile orokorrak ere kontuan izan beharko dira. (Mititelu, 2006) lanean adierazten den bezala “X, Y and Z” bezalako patroiak kontuan hartzekoak baitira hiperonimia erlazioak erauzteko momentuan.

5.2.2.2 Zeintzuk dira atributu esanguratsuenak?

“A edo B” bezalako markatzaile orokorrak garrantzitsuak direla ikusi dugu aurreko atalean, baina Wekako sailkatzaileen arabera esanguratsuenak diren atributuak jakitea interesgarria litzateke. Era honetan markatzaile garrantzitsuenak zeintzuk diren jakingo baitugu. Hauek dira atributu esanguratsuenak CfsSubsetEval eta ExhaustiveSearch atributu-aukeratze algoritmoen arabera:

- A gisako B
- A moduko B
- hainbat A ADI, B
- A: B1, B2
- A eta antzeko B
- A edo beste B

Beraz, hautagai-bikoteak sailkatzeko momentuan esanguratsuenak diren patroiak horiek dira. Kontua da, hainbat lanetan (Mititelu, 2006) aipatzen den bezala, patroi batzuek hiponimo-hiperonimo bikoteez gain beste erlazio batzuk bildu ditzakete. Askotan, instantziak izan daitezke.

5.3 Estatistiketan oinarritutako metodoak

Testuinguru-antza erabiliz, hiperonimia/hiponimia erlazioa hobeto rankingeatzen duten teknikak ebaluatu dira. Horretarako, LLR neurriaren arabera esanguratsuenak diren 70 termino bakunak prozesatu dira, eta horiekin antzekotasun distribuzional handiena duten 100 hitzak hiponimo edo hiperonimoak diren aztertu da:

- Automatikoki: Euskal WordNet eta WNTERM ezagutza-baseak erabilita.

- Eskuz: Erlazionatutako izen bikoteak hizkuntzalari batek eskuz egiaztatuta ea hiperonimo-hiponimo bikoteak diren ala ez.

Aztertutako hitzen emaitzak ondorengo taulan (14.taula) daude bilduta. Taulan, alde batetik, hartutako hitz esanguratsuen kopuruak emaitzetan eragiten duen ikusi nahi da. Bestalde, hitz bakoitzeko testuinguru antzekoena duten hitzen kopurua handitu ahala, hiponimo-hiperonimo kopurua gutxituz doan edo ez ikusi nahi da. Hau da, hiponimo-hiperonimoak rankingaren lehen postuetan dauden edo ez da aztertu nahi dena.

Konparaketa hori WordNet erabilita eta eskuz egin da, testuinguru-antzaren datua erabiltzeak WordNeten eragingo lukeen aberastea ikusteko. 14. taulan aurki daitezke ranking tamaina bakoitzerako erabilitako hitz antzekoenetatik, batez beste zenbat hiponimo-hiperonimo bikote dauden.

	Rankingeko lehen 20 postuak	**Rankingeko lehen 40 postuak**	**Rankingeko lehen 60 postuak**	**Rankingeko lehen 80 postuak**	**Rankingeko lehen 100 postuak**
50 hitz WordNet + WNterm	0,37	0,68	0,91	1,19	1,34
70 hitz WordNet + WNterm	0,38	0,63	0,96	1,31	1,46
50 hitz eskuz	0,97	1,56	2,22	2,63	2,97
70 hitz eskuz	1,17	1,73	2,56	3,08	3,42

14 taula: Antzekotasun distribuzionalen rankingen hiponimia asmatze-tasa ehunekotan(%)

14.taulako emaitzetan ikus daitekeenez, hitzen antzekotasun distribuzionala aztertuz gero, hitz horrekin lotura duten hiponimo-hiperonimoak lortu daitezke. Hala ere, hiponimia/hiperonimia erlazioa duten hitzez gain, beste mota bateko erlazioak dituzten hitzak ere lortzen dira (adibidez sinonimia edo meronimia), eta horregatik da ehunekoa txikia

Bikote hautagaia	**Erlazioa**
atomo-elektroi atomo-neutroi	Meronimia
Errotazio-biraketa	Sinonimia
hidrogeno-oxigeno	Co-hiponimia

15 taula. Antzekotasun distribuzionalaren bidez lortutako bikoteen adibideak

Beraz, erlazio semantiko ezberdinak bereizteko gai izateko, beste hurbilpen baten beharra ikusten da. Testuinguru-antza soilik erabiltzeak erlazio mota ezberdinak dituzten bikoteak lortzeko balio duelako, baina ez zein erlazio mota duten identifikatzeko.

5.4 Markatzaileak eta testuinguru-antza

Hainbat lanetan adierazi duten bezala, markatzaile linguistikoez gain banaketa distribuzionalari buruzko informazioa gehitzeak emaitzen hobekuntza dakar, ingelesaren kasurako behintzat. Euskararen kasurako ere hala den aztertu dugu.

Informazio hau gehituta, aurreko puntuak egindako kalkuluak errepikatu dira, maiztasunari buruzko datuekin (16. taula) eta baita presentziari buruzko datuekin ere (17.taula).

	Erlazionatuen doitasuna	**Erlazionatuen estaldura**	**Ez-erlazionatuen doitasuna**	**Ez-erlazionatuen estaldura**	**Zehaztasuna**
Naïve Bayes	0,55	0,9	0,66	0,68	0,64
SMO	0,59	0,72	0,57	0,48	0,65
Ibk (k=1)	0,61	0,62	0,61	0,61	0,61
AdaBoostM1	0,62	0,61	0,62	0,64	0,62
VFI	0,67	0,57	0,58	0,79	0,63
JRip	0,63	0,62	0,62	0,63	0,62
J48 tree	0,67	0,59	0,61	0,73	0,65

16 taula: Markatzaile linguistikoen sailkapenean lortutako emaitzak, maiztasunari dagozkion datuak eta testuinguru-antza erabiliz.

	Erlazionatuen doitasuna	**Erlazionatuen estaldura**	**Ez-erlazionatuen doitasuna**	**Ez-erlazionatuen estaldura**	**Zehaztasuna**
Naïve Bayes	0,62	0,75	0,66	0,51	0,63
SMO	0,68	0,57	0,61	0,75	0,64
Ibk (k=1)	0,6	0,61	0,61	0,6	0,61
AdaBoostM1	0,64	0,57	0,61	0,69	0,62
Vfi	0,65	0,46	0,58	0,76	0,6
JRip	0,64	0,63	0,64	0,63	0,64
J48 tree	0,66	0,52	0,61	0,79	0,65

17 taula: Markatzaile linguistikoen sailkapenean lortutako emaitzak, presentziari dagozkion datuak eta testuinguru-antza erabiliz.

Testuinguru-antza erabili gabe lortutako emaitzak (11. eta 12. taulak) eta testuinguru-antzarekin lortutakoak alderatuz gero (16. eta 17. taulak), ikus daiteke bigarren hauetan emaitzak hobeak direla. Hala ere, aldaketa ez da handia, eta frogak errepikatu beharko lirateke testuinguru-antza beste eraren batean adierazita, edota patroi lexiko-semantikoak erabili beharrean informazio sintaktikoa erabilita.

6 Ondorioak eta etorkizuneko lanak

Lan honetan testutik hiponimia/hiperonimia erlazioa erauzteko dauden teknika desberdinak frogatu dira euskarazko corpus tekniko batean. Egindako frogak euskararako dauden baliabideen araberakoak dira, eta terminoen arteko erlazio-semantikoen erauzketan lehen pauso bat besterik ez dira.

Dokumentu honetan ikusi den bezala markatzaile linguistikoak hiponimia erlazioa erauzteko baliabide egokia izan daitezke. Hala ere, markatzaile linguistiko bakoitzaren errepikapenak oso urriak dira corpusetan zehar. Honen arrazoia corpusa txikiegia izatea izan daiteke, baina baita mota honetako patroi lexiko sintaktikoen erabilera urria ere, ingelesaren kasuan bezala (Cimiano et al. 2004, Ryu et al. 2006). Gainera, gure corpusa espezializatua izanik, aurreikustekoa da corpus orokorragoetan baino hiponimia-hiperonimia markatzaile gehiago egongo direla. Hala ere, eskuz eta automatikoki egindako emaitzen azterketan argi gelditu da WordNeten ez dauden erlazio ugari eskuratu daitezkeela bide honetatik.

Oraindik ere lan asko dago egiteko Euskal WordNet era automatikoan edo erdi automatikoan aberastu ahal izateko. Dokumentu honetan azaldutako markatzaile linguistikoak abiapuntu egokia dira, baina gehiago landu beharra dago emaitza onak lortu ahal izateko. Patroiak findu egin behar dira.

Emaitzak ikusita argi dago markatzaile linguistikoak datu estatistikoekin konbinatuz lortutako emaitzak hobeak direla. Hala ere, erlazionatutako lanetan erabiltzen dituzten dependentzia zuhaitzak erabiltzea interesgarria litzateke.

Patroiak eskuz definitzeko momentuan ikusitako markatzaile posible ugari oso konplexuak dira sintaxia kontutan hartu gabe hiponimia erlazioa erauzteko. Beraz, markatzaile linguistikoak definitzeko patroi lexiko-semantikoak baino metodo konplexuagoak erabili beharko dira. Aukera egoki bat dependentzietan oinarritutako ikasketa automatikoa izan daiteke (Snow et al. 2006).

Etorkizuneko frogetan koordinazioari buruz beste hizkuntzetarako garatuta dauden teknikak euskarara egokitzea ere interesgarria izango litzateke (2.4.2 atalean azalduta). Bide honetatik co-hiponimia erlazioa garatuko litzateke, eta hiponimo-hiperonimo gehiago lortzeko aukera egongo litzateke.

Bestalde, hiponimia/hiperonimia erlazioaz gain WordNet elikatzeko beste hainbat erlazio ere erauz daitezke testuetatik. Sinonimia, adibidez, interesgarria izan daiteke, hiponimoak aurkitzeko momentuan hainbat sinonimo ere ikusi ditugulako. Testuinguru-antzaren bidez lortutako emaitzetan, esate baterako, sinonimo ugari aurkitu ditugu.

7 Bibliografia

Akiba, T., and Sakai, T. (2011). Japanese Hyponymy Extraction based on a Term Similarity Graph. ディジタル図書館 41–46.

Alegria, I., Gurrutxaga, A., Saralegi, X., and Ugartetxea, S. (2005). Erauzterm: euskarazko terminoak erauzteko tresna erdiautomatikoa. Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera Zientifiko-teknikoa.

Alfonseca, E., and Manandhar, S. (2002). Extending a lexical ontology by a combination of distributional semantics signatures. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web 281–293.

Ayşe, Ş., Zeynep, O., and İlknur, P. (2011). Extraction of semantic word relations in Turkish from dictionary definitions. ACL HLT 2011 11.

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., and Etzioni, O. (2009). Open information extraction for the web. University of Washington.

Bollegala, D.T., Matsuo, Y., and Ishizuka, M. (2010). Relational duality: Unsupervised extraction of semantic relations between entities on the web. In Proceedings of the 19th International Conference on World Wide Web, pp. 151–160.

Bunescu, R.C., and Mooney, R.J. (2005). A shortest path dependency kernel for relation extraction. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 724–731.

Cederberg, S., and Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4, pp. 111–118.

Culotta, A., and Sorensen, J. (2004). Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 423.

Curran, J.R. (2005). Supersense tagging of unknown nouns using semantic similarity. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 26–33.

Etzioni, O., Banko, M., Soderland, S., and Weld, D.S. (2008). Open information extraction from the web. Communications of the ACM *51*, 68–74.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence *165*, 91–134.

Giovannetti, E., Marchi, S., and Montemagni, S. (2008a). Combining statistical techniques and lexico-syntactic patterns for semantic relations extraction from text. In Proc. of the 5th Workshop on Semantic Web Applications and Perspectives, p.

Giovannetti, E., Marchi, S., and Montemagni, S. (2008b). Combining statistical techniques and lexico-syntactic patterns for semantic relations extraction from text. In Proc. of the 5th Workshop on Semantic Web Applications and Perspectives, p.

Giovannetti, E., Marchi, S., and Montemagni, S. (2008c). Combining statistical techniques and lexico-syntactic patterns for semantic relations extraction from text. In Proc. of the 5th Workshop on Semantic Web Applications and Perspectives, p.

GuoDong, Z., Jian, S., Jie, Z., and Min, Z. (2005). Exploring various knowledge in relation extraction. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 427–434.

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th Conference on Computational linguistics-Volume 2, pp. 539–545.

Hearst, M.A. (1998). Automated discovery of WordNet relations. WordNet: An Electronic Lexical Database 131–151.

Hovy, E., Kozareva, Z., and Riloff, E. (2009). Toward completeness in concept extraction and classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp. 948–957.

Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, p. 22.

Kozareva, Z., and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1110–1118.

Kozareva, Z., Riloff, E., and Hovy, E. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. Proceedings of ACL-08: HLT 1048–1056.

Lersundi M. 2005. "Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak". Euskal Filologia Saila.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 491–498.

Mititelu, V.B. (2006). Automatic extraction of patterns displaying hyponym-hypernym co-occurrence from corpora. In First Central European Student Conference in Linguistics, p.

Navigli, R., and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1318–1327.

Navigli, R., Velardi, P., and Faralli, S. (2011). A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch. In Twenty-Second International Joint Conference on Artificial Intelligence, p.

Pociello, E. (2008). Euskararen ezagutza-base lexikala: Euskal WordNet. Doktoretza-tesia, Euskal Filologia Saila (UPV/EHU). Leioa.

Pociello, E., Gurrutxaga, A., Agirre, E., Aldezabal, I., and Rigau, G. (2008). WNTERM: Enriching the MCR with a terminological dictionary.

Ponzetto, S.P., and Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In Proceedings of the National Conference on Artificial Intelligence, p. 1440.

Riloff, E., and Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 117–124.

Ritter, A., Soderland, S., and Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. In Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read, pp. 88–93.

Sang, E.T.., and Hofmann, K. (2009). Lexical patterns or dependency patterns: which is better for hypernym extraction? In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 174–182.

Snow, R., Jurafsky, D., and Ng, A.Y. (2004). Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems 17.

Wang, T., Li, Y., Bontcheva, K., Cunningham, H., and Wang, J. (2006). Automatic extraction of hierarchical relations from text. The Semantic Web: Research and Applications 215–229.

Widdows, D., and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In Proceedings of the 19th International Conference on Computational linguistics-Volume 1, pp. 1–7.

Yang, H., and Callan, J. (2009). A metric-based framework for automatic taxonomy induction. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pp. 271–279.

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. (2007). TextRunner: open information extraction on the web. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 25–26.

Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel methods for relation extraction. The Journal of Machine Learning Research *3*, 1083–1106.