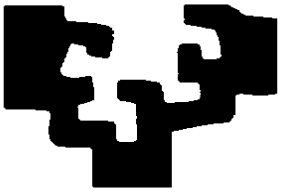
LENGOAIA ETA SISTEMA INFORMATIKOAK SAILA

eman ta zabal zazu



INFORMATIKA FAKULTATEA

KONTZEPTUEN ARTEKO ERLAZIO-IZAERAREN FORMALIZAZIOA ONTOLOGIAK ERABILIAZ: DENTSITATE KONTZEPTUALA

aplikazioak ezagutza-base lexikalen eraikuntzan, adiera-desanbiguazioan eta testu-zuzenketa automatikoan

Eneko Agirre Bengoak

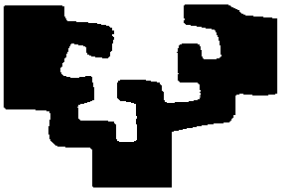
Informatikan Doktore titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

Donostia, 1998ko urria.

LENGOAIA ETA SISTEMA INFORMATIKOAK SAILA

eman ta zabal zazu



INFORMATIKA FAKULTATEA

KONTZEPTUEN ARTEKO ERLAZIO-IZAERAREN FORMALIZAZIOA ONTOLOGIAK ERABILIAZ: DENTSITATE KONTZEPTUALA

aplikazioak ezagutza-base lexikalen eraikuntzan, adiera-desanbiguazioan eta testu-zuzenketa automatikoan

Eneko Agirre Bengoak Kepa Sarasola Gabiola eta **Arantza Diaz de Ilarrazaren** zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 1998ko urria.

AURKIBIDEA

IRUDIEN AURKIBIDEA

TAULEN AURKIBIDEA

HIZTEGIA

Abarkatze-faktore. Branching factor

Adimen Artifizial. Artificial Intelligence

Agerkidetza. Co-occurrence

Antzekotasun. Similarity

Arbaso. Ancestor

Benetako-hitz errore. Real-word error

Datu urrien arazo. Sparse data problem

Dentsitate Kontzeptual (DK). Conceptual Density

Distantzia Kontzeptual. Conceptual Distance

Doitasun. Precision

Dokumentuen berreskuratze. Document retrieval

Dokumentuen sailkapen. Document clustering

Elkarren Arteko Informazio (EAI). Mutual Information

Erabaki-zerrenda. Decision list

Eraginkortasun. Efficiency

Erlazio-izaera . Relatedness

Erlazio-izaera paradigmatiko. Paradigmatic Relatedness

Erlazio-izaera sintagmatiko. Syntagmatic Relatedness

Erlazionatutako. Related

Estaldura. Coverage

Ezagutza-base lexikal (EBL). Lexical Knowledge Base

Ez-hitz errore. Non-word error

Goi-ontologia. Top ontology

Hautapen-murrizpen. Selectional restriction

Hitz isolatuen zuzenketa. Isolated Word Correction

Hitzen adiera-desanbiguazio (HAD). Word Sense Disambiguation

Hiztegi elektroniko. Machine Readable Dictionary

Hiztegi ezagutza-base (HEB). Dictionary Knowledge Base

Hurbiltasun. Proximity, closeness

Informazioaren berreskuratze. Information Retrieval

Informazio-eduki. Information content

Karaktere-ezagutze optiko. Optical character recognition

Kategoria-etiketatzaile. Part of speech tagger

Lengoaia Naturalaren Prozesamendu (LNP). Natural Language Processing (NLP)

Leuntze. Smoothing

Log-sinesgarritasun. Log-likelihood

Markov-en eredu izkutu. Hidden Markov Model

Multzokatze. Clustering

Murrizpen-Gramatika (MG). Constraint Grammar

Nabaritasun. Relevance

Nabarmentasun. Saliency

Nahaste-multzo. Confusion-set

Ondorengo. Descendant

Sendotasun. Robustness

Testuingururik gabeko gramatika hedatu. Augmented Context-Free Gramars

Thesaurus. Thesaurus

Zati erlazioa. Part-of relation

LABURDURAK

AA. Agerkidetza arruntak

AB. Agerkidetza bektoreak

AR. *Association Ratio*

BM. Brown maiztasunak

DK. Dentsitate Kontzeptual

DM. Dokumentuko maiztasunak

EAI. Elkarren Arteko Informazio

EBL. Ezagutza-Base Lexikal

EZ. Definizioko hitzen ezkontzea

H1. Izen nagusien heuristikoa

H2. Hitz laburren heuristikoa

HAD. Hitzen Adiera-Desanbiguazio

HEB. Hiztegi Ezagutza-Base

LNP. Lengoaia Naturalaren Prozesamendu

LPPL. *Le Plus Petit Larousse*

MG. Murrizpen-gramatika

OFED. Oxford French/English Dictionary

OR. Adieren ordena

SB. Etiketa semantikoen bektoreak

TS. Testuingurua kontuan hartzen duen metodoa

WN. WordNet

I. Kapitulua

PROIEKTUAREN AURKEZPENA

I.A. Motibazioa

Gizakiok era naturalean erabakitzen dugu edozein gauza zein puntutaraino erlazionatuta dauden ala ez. Zer dago ardiarekin erlazionatuago, behia, bixigua ala irratia? Halako galderei erantzuteko arazorik ez dugu izaten. Ordenadoreek aldiz, sen onaren alderdi gehienekin gertatzen den bezala, ez daukate nondik heldu galdera horri. Ez dakite zer diren ardia, bixigu edo irratia, ezta beraien arteko erlazioak zeintzuk diren. Galdera horri erantzun ahal izanez gero aplikazio interesgarri askotara hedatu ahal izango dira ordenadoreak. Gu Lengoaia Naturalaren Prozesamenduan (LNP) zentratuko gara. Askoren ustez halako galderei erantzuteko ahalmena da prozesamendu semantikoaren giltza. Ahalmen horri erlazio-izaeraren neurri deitzen diogu, hau da, bi hitzen arteko erlazioak zer indar duen emango digun neurria. Neurri hau izenentzat batez ere definitu ohi da.

Erlazio-izaera formalizatzeko aukera ezberdinak aztertu izan dira literaturan. Lan batzuetan hitzen arteko erlazio-izaera landu izan da soilik, baina beste askok adierekin lan egiten dute. Lehenbizikoek, adibidez, *banku*-ren adiera ezberdinen artean ezin dute bereizi, baina bigarrenek *banku* eta *aulki* hertsiki erlazionatuta dauden ala ez galderari, "segun" erantzungo liokete: *banku* hori esertzekoa baldin bada orduan bai, baina eraikuntza bada, diruarekin zer ikusia duena, orduan ez daude hain hertsiki erlazionatuta. Hitzen arteko erlazio-izaera baino, ene ustez, adieren artekoaren formalizazioa interesgarriagoa da.

Formalizazioak oinarri duten baliabide lexikalaren arabera ere sailka daitezke:

- idatzizko testu multzoak diren corpusak erabiltzen dituztenak
- hiztegietako informazioa erabiltzen dutenak, bereziki hitzen definizioak

- ezagutza egituratua darabiltenak, hala nola, Hiztegi Ezagutza-Base (HEB), Ezagutza-Base Lexikal (EBL) eta ontologiak.

Hiru baliabide motak aztertu ondoren ezagutza egituratuan oinarritzea iruditu zaigu zentzuzkoena. Baliabide lexikal guztiek daukate informazio interesgarria, hein handi batean bata bestearen osagarria dena. Hala ere tradizio sendoena ontologian oinarritutako neurriena da, psikologia eta adimen artifizialeko lanetan errotua. Erabaki horretan EBL zabal bat, WordNet, eskura eduki ahal izateak lagundu digu, eta ezagutza-base horren gainean inplementatu dugu erlazio-izaeraren neurria. III.A atalean baliabide ezberdinetan oinarritutako neurriak aztertuko ditugu, eta III.D atalean ontologiatan oinarritutakoak hobesteko arrazoiak azaldu. Guk aurkezten dugun izenen arteko erlazio-izaeraren neurriari Dentsitate Kontzeptual (DK) deitzen diogu, eta ontologiako kontzeptuen arteko hierarkian oinarritzen da. Nahiz eta WordNet-eko ontologia erabiliaz inplementatu, kontzeptuen hierarkia eta erlazioak dituen edozein baliabide lexikaletan aplika daiteke.

Adieren arteko erlazio-izaeraren neurria aplikazio askotarako ezinbestekoa edo gutxienez lagungarria da, hala nola, egitura sintaktikoen desanbiguazioa, hitzen adiera desanbiguazioa, ontologien eraikuntza, hautapen-murrizpenen ikasketa, ontologia ezberdinen bat egitea, ontologien ebaluazioa, informazioaren berreskuratzea, dokumentuen berreskuratze eta sailkapena, kontzeptuen multzokatzea, testu-zuzenketa automatikoa, bai eta interpretazio semantiko orokorra ere.

Erlazio-izaera sarri azaltzen zaigu Hitzen Adiera-Desanbiguazioari (HAD) lotuta, eta aplikazio hori ere erabili nahi izan dugu gure formalizazioa ebaluatzeko. Beraz, testu libreetako izenen adieren artean desanbiguatzeko Dentsitate Kontzeptuala erabili dugu. Gaur egun pil-pilean dagoen gaia da hau, guztiz irekita jarraitzen duen arazoa. Itzulpen automatikorako 60.eko hamarkadan egin ziren sistemek adiera-desanbiguazioari ezin izan zioten aurre egin, eta hori izan zen beraien porrotaren arrazoietako bat. Erlazio-izaeraren inplementazioak informazio zabala erabiltzen hasi diren heinean, hitzen adiera-desanbiguazioan emaitza hobeak lortzen joan dira. Egungo teknologiarekin aplikazio errealetan aplikatzeko moduan egon ez arren, hurbil ikusten da edozein testutako hitzen adiera-desanbiguazio azkarra eta errore-maila onargarrikoa.

HADen hurbilpen hedatuenek polisemia eta homonimia adieren zerrenda itxi batez errepresentatzen dituzte, eta informazio xumea[1] soilik erabiliaz adiera egokia hautatzeko gai direla nabarmentzen dute. Badaude honen aurrean eszeptikoak direnak. Alde batetik daude HAD LNParen beste arazoetatik isolatuta ezin tratatu daitekeela diotenak, LNP orokorrerako beharrezko

[1] Xumea diogu ezagutza intentsiboa erabiltzen dutenekin alderatzen badugu. Beste era batera esanda, oraingo metodoek ezagutza estentsiboa erabiltzen dutela esan daiteke.

ezagutza guztia ezin delako alde batera utzi. LNPan aurrera egin ahala HAD naturalki ebatziko dela uste dute. Beste aldetik daude lexikoaren izaera dinamikoa aldarrikatzen dutenak. Horientzat metonimia eta metafora bezalako prozesuak ulertu gabe ezin da adieren arteko ezberdintasunik planteatu. Beste batzuk harantzago doaz, eta adieren artean mugak jartzerik ez dagoela diote, eta adieren beraien existentzia entitate diskretu bezala zalantzan jartzen dute. Gure ustez kritika horiek ikuspuntu ezberdinak besterik ez dira, kontuan hartu behar direnak, eta ahal dela HAD sisteman integratu beharrekoak (eta hau aldi berean LNP orokorrean integratu noski), baina ezin uka daiteke bitartean emaitza interesgarriak lortzen ari direla, eta HAD tratatu nahian teknika berritzaileak garatu izan direla. Nolabait badirudi eztabaida alde praktikora eraman dela, adieren existentzia bera zalantzan jartzen duen Kilgarriff-ek berak, *Senseval*[2] deritzon lehiaketa antolatu baitu HADaren inguruan 1998. urtean.

Gure ikerkuntza-taldean idazketarako laguntza-tresnak garatzea da helburu iraunkorretako bat. Bide horretatik Xuxen euskararako testu-zuzentzaile komertziala garatu genuen. Ortografia-erroreen aurrean programa zuzentzaileek erabiltzaileari hitz zuzena eskaintzen saiatzen dira, proposamen zerrenda baten bidez. Giza-erabiltzailearen esku dago proposamen zuzena aukeratzea. Testu-editoreen kasuan nahikoa bada ere, beste aplikazio batzuetan beharrezkoa da programak berak zuzenketa egokia aukeratzea. Halako aplikazioen adibide bat karaktere-ezagutza optikoa (*optical character recognition*) da. Jakina da, paperean dauden testuak ordenadorera pasatzea nahi baditugu karaktere-ezagutza optikoek ez dutela beti asmatzen (hitz hasieratako *I* adibidez *l* bezala interpretatzen dute sarritan), eta beraz post-prozesu bat egin beharra dagoela errore horiek zuzentzeko, normalean zuzentzaile ortografikoa erabiliz eta eskuz hautatuz aukera zuzena. Taldean garatu diren tresna sintaktikoen eta Dentsitate Kontzeptualaren bidez testu-zuzenketa automatikorako bidean saiakera bat egin dugu tesi honetan. Bide batez, erlazio-izaeraren neurria beste zeregin batean probatzeko aukera eman digu horrek.

Tesi lan honen beste motibazio garrantzitsu bat baliabide lexikalen sorrera da. 80.eko hamarkadan, ordurarte sintaxian buru belarri zegoen LNParen komunitatean, baliabide lexikal zabal eta aberatsen beharra zabaldu zen. LNPrako aplikazioak kalera atera ahal izateko testu errealei aurre egin beharra zegoen, eta horretarako ezinbesteko zen lexiko zabalak edukitzea. Bestalde, ordurarte erregela konplexu eta ugariren bidez deskribatzen ziren fenomeno linguistiko askok jatorri lexikala zutela jabetzean, lexikoa hitz zerrenda laua izatetik informazio aberats eta konplexua zuen sistema izatera pasatu zen. Gauzak horrela, lantaldeak lexiko horiek eskuz eraikitzen hasi ziren. Kodetu beharreko informazio kopurua itzela da eta gizon-urte askotako ahalegina suposatzen du, proiektu erraldoi

[2] http://www.itri.bton.ac.uk/events/senseval/

gutxi batzuen esku dagoena (adibidez CYC, EDR edo WordNet). Eskuzko kodeketaren alternatiba gisa, lexikoak edukiz betetzeko laguntza automatiko edo semi-automatikoak ere bilatu izan dira, eta horrekin arreta bestelako baliabide lexikalen tratamendura zuzendu zen, corpus eta hiztegietara.

Hiztegietatik Ezagutza-Base Lexikalak (EBL) erauzi izan dira. Erauzitako informazioaren artean, semantikari dagokionean, garrantzitsuena adieren arteko hierarkiak izan dira. Tamalez lantalde gehienek hitzen arteko hierarkiak besterik ezin izan dituzte lortu, ezin izan baitute automatikoki erabaki zein zen adiera egokia. Honen salbuespena LDOCE hiztegiarekin eginiko lana da, hiztegi konkretu horretan kodetuta dagoen informazio lagungarria erabiliz automatikoki eraiki baitute adieren arteko hierarkia. Erauzitako EBL gehienak ingeleserako izan dira, eskuz eraiki izan diren ontologia eta EBL erraldoiak bezala. Horrek gainontzeko hizkuntzak LNParen aurrean posizio ahul baten uzten ditu. Bi irtenbide osagarri daude egoera horren aurrean:

- hizkuntza bakoitzerako dauden corpus eta hiztegietatik abiatuta EBLak sortzea
- ingeleserako eraiki diren EBLetaz baliatzea beste hizkuntzatarako EBLak sortzeko

Hau da, hizkuntza bakoitzerako dauden baliabideetaz profitatu, nola ez, baina baita ere ingeleserako baliabideetan dagoen ezagutza interesatzen zaigun hizkuntzara nolabait itzuli. Gure ustez erlazio-izaeraren neurriaren formalizazioak bi irtenbide horietan lagundu dezake.

Orain arte aipatu ditugun bi motibazio nagusiak, erlazio-izaeraren formalizazioa eta baliabide lexikal egituratuen eraikuntza, uztartu egiten dira. Hizkuntza baterako erlazio-izaera ezin da definitu hizkuntza horretarako baliabide lexikal egituraturik ez badago, bereziki EBL eta ontologiak. Bestalde erlazio-izaera gabe ez da erraza baliabide lexikal egituratuak sortzeko laguntza-tresna automatikoak egitea. Ingelesarentzat eraiki izan diren EBL eta ontologiak erabiliaz posiblea da ingeleserako erlazio-izaera definitzea. Horretaz baliatuz, beste hizkuntzatan dauden baliabide egituratuen erauzte automatikoa azkartu eta ingelesera lotzea posible bada, orduan ingeleserako sortu diren baliabideetan dagoen aberastasuna xurgatu ahal izango da, eta baliabide aberats horiek beste hizkuntzetarako ere erabilgarriak izango dira. Ildo horretatik bi lan nagusi burutu nahi izan ditugu. Alde batetik gure taldean ikerkuntzaren objektu izan den *Le Plus Petit Larousse* frantses hiztegiko adierak WordNet ingeles EBLko adieretara lotu ditugu. Eta bestetik, hiztegian bertan dagoen informazioaz eta WordNetera egindako loturetaz baliatuz, LPPL-tik erauzitako hierarkiak desanbiguatu ditugu.

Tesi-lan honi ekin genionean ez zegoen baliabide lexikal egituratu zabalik ingelesa ez ziren hizkuntzentzat. Hori dela eta erlazio-izaera ingeleseko adierentzat definitu dugu, eta adiera-

desanbiguazioa baita testu-zuzenketa ere ingeleseko testuen gainean egin dugu. EBLen aberasketa eta trinkotzeari dagokionean, frantsesezko hiztegi bat zegoen era sakonean landuta taldean, eta hori saiatu gara aberasten. Dena den, nahiz eta tesi honetan ezin landu izan dugun, euskara da garatutako teknika guztien azken jomuga, gure ikerkuntza taldean gertatzen den bezala. Jorratu ditugun bideak eta azterketak euskararako, edo orokorrean beste edozein hizkuntzarako, baliabide lexikal zabal baten eraikuntzarako funtsa dira.

I.B. Helburuak

Motibazio nagusiei erantzunez, erlazio-izaeraren formalizazioa eta baliabide lexikal egituratuen eraikuntza, bi helburu nagusi jarri dizkiogu tesi-lan honi:

a) teorikoa: ezagutzan oinarritutako hitz eta kontzeptuen arteko erlazio-izaera neurtzea
b) praktikoa: ingelesezkoak ez diren baliabide lexikal egituratuak aberastu eta trinkotzeko teknikak lantzea

Bi helburuak baliabide lexikalen inguruan dihardute. Helburu teorikoari dagokionez, baliabide lexikaletaz profitatzen saiatuko gara inferentzia mota bat aurrera eramateko. Helburu praktikoa baliabide lexikal aberatsagoen eraikuntzaz ari da, hiztegietatik EBLetara.

Helburu hauek gauzatzeko hiru eginkizun nagusi eraman ditugu aurrera:

1. WordNet-en oinarritutako Dentsitate Kontzeptuala diseinatu eta inplementatu.
2. Le Plus Petit Larousse frantses hiztegiko adierak WordNet-i lotu.
3. Le Plus Petit Larousse-etik erauzitako Hiztegi Ezagutza Baseko (HEB) adieren hierarkiak desanbiguatu eta trinkotu.

2. eta 3. eginkizunak aurrera eramateko beharrezkoa izan da Dentsitate Kontzeptuala erabiltzea. Aipatzekoa da, behin HEBko adieren hierarkia sendoa eraiki eta gero, posible izango dela Dentsitate Kontzeptuala zuzenean eraikitako hierarkia horren gain aplikatzea, frantseserako erlazio-izaeraren neurria lortuz. Gure hurbilpen honen atzetik honako hipotesia dago:

Ingelesezkoak ez diren EBLak sendotzeko kanpoko ezagutza behar dela, eta kanpoko ezagutza hori normalean ingelesez egon badagoenez, lotura eleanitzen bidez eskuratu daitekeela.

Goian aipatu bi helburu nagusiez gain, definitutako erlazio-izaeraren neurria beste bi arlotan aplikatu eta ebaluatu nahi izan dugu. Beraz goiko eginkizunetaz gain beste bi hauei ere aurre egin diegu:

4. Dentsitate Kontzeptualaren aplikazio, fintze eta ebaluazioa: hitzen adiera-desanbiguazioa
5. Dentsitate Kontzeptualaren aplikazio eta ebaluazioa: testu-zuzenketa automatikoa

Ingelesezko HAD burutzeko WordNet gainean inplementatutako Dentsitate Kontzeptuala besterik ez dugu erabili. Eginbehar honek berez duen interesaz gain, erlazio-izaera ebaluatzeko erabili dugu. Izan ere erlazio-izaera zuzenean ebaluatzeko metodo adosturik ez dago, eta nahiago izan dugu eginkizun praktiko eta konparagarri baten bidez ebaluatzea.

Testu-zuzenketa automatikoa aurrera eraman ahal izateko, WordNet gaineko Dentsitate Kontzeptualaz gain, ezagutza iturri ezberdinak erabili ditugu. Alde batetik sintaxiari buruzko ezagutza, eta bestetik hitzen maiztasun eta agerkidetzei dagozkien eredu estatistikoak.

I.C. Tesiaren egitura

Tesiaren helburu eta eginkizunekin kapitulu bakoitzak duen harremana, 1. taulan laburbildu dugu.

Tesiaren helburu nagusiak	Eginkizunak	Kapituluak
		I Sarrera
		II Baliabide Lexikalak
Ezagutzan oinarritutako hitz eta kontzeptuen arteko erlazio-izaera definitzea	WordNet-en oinarritutakok DK diseinatu eta inplementatu	III Erlazio-Izaera eta Dentsitate Kontzeptuala
	DK aplikazio, fintze eta ebaluazioa: hitzen adiera-desanbiguazioa	IV Hitzen Adiera-Desanbiguazioa
	DKren aplikazio eta ebaluazioa: testu-zuzenketa automatikoa	V Zuzenketa Automatikoa
Ingelesa ez diren hizkuntzetarako baliabide lexikal egituratuak aberastu eta trinkotzeko teknikak lantzea	*Le Plus Petit Larousse* frantses hiztegiko adierak WordNet-i lotu, eta *Le Plus Petit Larousse*-etik erauzitako HEBko adieren hierarkiak desanbiguatu eta trinkotu	VI Hiztegi Ezagutza-Basearen Aberasketa
		VII Ondorioak

1. taula: tesiaren egitura eta helburu nagusiak

Sarrera-kapitulu honen ondoren, **II. kapituluan** (Baliabide Lexikalak: Erabilera Praktikoak), baliabide lexikalei buruz hitz egingo dugu. Gaur egun Lengoaia Naturalaren Prozesamenduan baliabide lexikalak duten garrantzia aipatu ondoren, baliabide garrantzitsu eta ezagunenak aipatuko ditugu, arreta berezia eskainiz tesi honetan erabili ditugun corpus, hiztegi eta baliabide egituratuei. Hitzen adiera-desanbiguazioan *Semcor* eta *Brown* corpusetaz baliatuko gara emaitzak ebaluatzeko, eta testu-zuzenketan, aipatutakoez gain, *Bank of English* corpusa ere azalduko zaigu. Hiztegiei

dagokionez *Le Plus Petit Larousse* eta *Oxford French-English Dictionary* erabili ditugu. Baliabide lexikal egituratuei dagokionean, WordNet – Dentsitate Kontzeptuala inplementatzeko aukeratu dugun hierarkia – beste ontologiekin alderatuko dugu.

III. kapituluan (Erlazio-Izaera eta Dentsitate Kontzeptuala) hitz eta adierak hertsiki erlazionatuta ote dauden neurtzeko moduak aztertu eta tesi-lan honen ekarpen nagusiena den Dentsitate Kontzeptuala aurkeztuko dugu. Dentsitate Kontzeptuala azaldu aurretik, erlazio-izaeraren bestelako formalizazioak aztertuko ditugu. Dentsitate Kontzeptualaren inplementazioa azaltzerakoan, enpirikoki erabaki beharreko parametro batzuk aurkeztuko ditugu. Ondoren ontologien gainean definitutako erlazio-izaeren nagusitasuna defendituko dugu, eta ontologiatan oinarritutakoen barruan Dentsitate Kontzeptualak dauzkan abantailak. Bukatzeko kapitulu honi dagozkion ekarpenak eta etorkizuneko eginkizunak aipatuko ditugu.

IV. kapituluan (Hitzen Adiera-Desanbiguazioa Testu Errealetan) Dentsitate Kontzeptuala aplikazio praktiko batean ebaluatu nahi izan dugu, eta bide batez aplikazio horren emaitzen arabera Dentsitate Kontzeptualaren parametroak doitu. Aurreko kapituluan Dentsitate Kontzeptualaren abantaila teoriko eta praktikoak azaltzen badira ere, aplikazio praktikoetan emaitza onak ematen dituela frogatu nahi izan dugu. Hitzen Adiera-Desanbiguazioan, testu bateko hitz bat bere zein adieratan erabiltzen den erabaki behar da. Erlazio-izaeraren neurri gehienak Hitzen Adiera-Desanbiguazioan (batez ere izenen desanbiguazioan) aplikatu izan dira, eta are gehiago, askotan horretarako diseinatu izan dira espreski. Kapitulu hau aurrekarien azterketa batez hasiko da, ezagutza-iturri ezberdinen beharra azpimarratuz. Ondoren gure esperimentuaren diseinua eta desanbiguatzeko Dentsitate Kontzeptuala erabiltzen duen algoritmoa azalduko ditugu. Aurrez aldetik desanbiguatuta dagoen corpus bat erabili dugu, automatikoki neurtu ahal izateko sistemaren doitasuna. Corpus horretako 4 fitxategi zoriz aukeratu ditugu, eta 2.000 inguru izen desanbiguatu ditugu, WordNet-eko adiera egokia esleituaz. Atal berezi bat erabiliko dugu Dentsitate Kontzeptualaren parametroen eragina aztertzeko, eta parametroentzako balio hoberenak aukeratzeko. Emaitzen ebaluazioaren ondoren, beste metodoekin alderatu dugu. Ontologian oinarritutako beste bi metodo inplementatu eta aplikatu ditugu, emaitza okerragoak lortuaz. Bukatzeko, kapitulu honen ekarpen eta etorkizunerako lanak.

V. kapituluan (Zuzenketa Automatikoa) beste aplikazio praktiko bat landu dugu idazketa-erroreen zuzenketa automatikoaren inguruan. Kapitulu honetan zuzenketa-proposamenen artean zuzena automatikoki aukeratzen saiatzen den sistemaren diseinu eta inplementazioa aurkezten ditugu. Lehenbizi aurrekarien azterketa egin dugu. Ondoren sintaxi eta semantikan oinarritutako

bideragarritasun-azterketaren emaitzak aurkezten ditugu. Ondorio bezala semantikaren ekarpena ezinbestekotzat jo genuen. Gure algoritmoan errorea azaltzen den esaldi eta testuinguruari erreparatuko diogu proposamen zuzena aukeratzeko orduan. Lehenbizi egitura sintaktiko onargarria ematen ez dituzten proposamenak baztertuko ditugu. Gainontzeko proposamenen artean esanahiaren aldetik testuinguruan zentzu gehien egiten duena aukeratu ahal izateko ezagutza semantikoa erabiliko dugu. Alde batetik WordNet-en dagoen ezagutzaz baliatzeko izenen arteko Dentsitate Kontzeptuala erabili dugu, bide batez Dentsitatearen ekarpena neurtuko dugularik. Beste aldetik corpusetan oinarritutako teknikak ere erabiliko ditugu. Ebaluazioa bi corpus ezberdinen gainean egin dugu: artifizialki sortutako erroreen corpusa eta corpus naturala. Bukatzeko, beste kapituluetan legez, ekarpenak eta etorkizunerako lana azaldu ditugu.

VI. kapituluan (Hiztegi Ezagutza-Basearen Aberasketa) tesi honen beste helburua den ingelesa ez diren hizkuntzetarako EBLen aberaste eta trinkotzea jorratu dugu. Lehenbizi aurrekariak aztertuko ditugu, eta hiztegietatik erauzitako hierarkiek dauzkaten arazoak azaldu. Kontuan hartu behar da hierarkia horiek ez direla guztiz desanbiguatuta egoten. Gainera, erauzitako hierarkiak sakonera apalekoak izan eta elkarrengandik isolatuta egoten dira, eta hierarkiaren goi-mailan koherentzia arazoak gertatzen dira. Arazo hauen iturburu dira, hein batean, hierarkian gertatzen diren bigiztak eta adiera batzuek hierarkian ezin kokatu izana, erlatore berezien bidez definitzen direnak hain zuzen ere. Aurrekarietan baliabide lexikal eleanitzen arteko lotura ere begiratu dugu. Ondoren gure hurbilpena azaldu dugu.

EBLen eraikuntza sendotzea posiblea den ala ez ikusteko frantseseko *Le Plus Petit Larousse*-etik erauzitako Hiztegi-Ezagutza Basea erabili dugu. HEB hori LNPrako EBL bezala erabili ahal izateko arestian aipatutako arazoak konpondu beharko lirateke. Bi arazo horiei era bateratu batez erantzuten saiatu gara. Hasteko, bigizta eta erlatore bidezko definizioak aztertu ditugu eta horiek LPPLra lotu ahal izateko, kanpoko EBL batera lotu ditugu LPPLko adiera guztiak. Gure kasuan WordNet hautatu dugu kanpoko EBL bezala. Ondoren hierarkiak automatikoki desanbiguatu ditugu. Azkenik, LPPL-WordNet lotura erabili dugu hierarkiak elkarrekin harremanetan jartzeko, bide batez goi-mailako koherentzia falta ere konponduaz.

LPPLtik erauzitako HEBa WordNet-i lotzeko hiztegi elebidun bat erabili dugu, hizkuntzen arteko zubi bezala. Lotura automatikoki egin ahal izateko Dentsitate Kontzeptualaz baliatu gara, LPPLko adiera bakoitzari WordNet-eko kontzeptu bat (edo gehiago) esleitzeko balioko diguna. Hierarkia desanbiguatzeko hiztegian bertan dagoen ezagutzaz eta WordNet-i egindako loturez baliatu gara. Hainbat teknika independente erabili ditugu, Dentsitate Kontzeptuala barne, eta teknika horiek

konbinatu ondoren adiera egokienak aukeratu ditugu. Bukatzeko, kapitulu honetan eginiko ekarpenak eta etorkizunerako lana bildu ditugu.

Kapitulu bakoitzean tesi-lan honen ekarpenak adierazten saiatu gara. Azkeneko kapituluan horiek guztiak bildu eta etorkizunean egin daitezkeen hobekuntzak ere aipatzen ditugu.

Mamiarekin hasi aurretik, tesi honen garapenean argitaratutako artikuluen irakur gida azaldu nahi dugu. Irakur gida honetan, artikulu bakoitzak tesi honen egituran duen lekua zein den adierazten dugu, eta bide batez artikuluen zerrenda aurkeztu ere.

Kapitulua	Atala	Artikulua
III Erlazio-Izaera eta Dentsitate Kontzeptuala	B.1	(Agirre et al., 1994b) (Agirre & Rigau, 1995) (Agirre & Rigau, 1996a) (Agirre & Rigau, 1996b)
IV Hitzen Adiera-Desanbiguazioa		(Agirre & Rigau, 1995) (Agirre & Rigau, 1996a) (Agirre & Rigau, 1996b)
V Zuzenketa automatikoa	B.1 B C 	(Agirre, 93) (Agirre et al., 1994b) (Agirre et al., 1995) (Agirre et al., 1998b) (Agirre et al., 1998c)
VI Hiztegi Ezagutza-Basearen aberasketa	C.1 D	(Rigau & Agirre, 1995) (Rigau et al., 1997)

II. Kapitulua

BALIABIDE LEXIKALAK: ERABILERA PRAKTIKOAK

Sarreran aipatu dugu tesi honetako helburuen atzean baliabide lexikal egituratuen eraikuntza eta erabilera dagoela. Kapitulu honetan baliabide lexikalak lau sailetan banatuko ditugu. Sail bakoitzeko baliabide ezagun eta erabilienak II.A. atalean aipatuko ditugu. Horiek azaldu ondoren, ontologiak, Ezagutza-Base Lexikal (EBL) eta Hiztegi Ezagutza-Baseak (HEB), sail berdinean sailkatzeko arrazoiak aztertuko ditugu (II.B. atala). Bukatzeko, tesi honetan erabili ditugun baliabide lexikalen ezaugarriak azalduko dira.

II.A. Baliabide lexikal motak

Baliabide lexikalak lau sail nagusitan banatu ditugu:

1. Corpusak
1. Hiztegiak
2. Egituratuak: ezagutza-base lexikalak eta hiztegi ezagutza-baseak
3. Egituratuak: ontologiak

Sailkapen honetako ordena informazioaren elaborazio-mailaren arabera egin dugu. Corpusetan hitzei buruzko informazio gordina dago. Hiztegietan, aldiz, lexikografoek kategoria, erabilera kodeak, definizioa, adibideak, etab. biltzen dituzte. Hitzak ez ezik, hitzen adierak ere azaltzen zaizkigu. HEBetan hiztegietan dagoen informazio inplizitua esplizitu bihurtu eta hitzei buruzko informazio lexikala biltzen da. EBLetan LNPrako sistema batek ulermen eta sormena egiteko hitzei buruz behar duen informazio guztia biltzen dute. Ontologiak munduari buruzko

kontzeptualizazioak dira, munduari edo alor konkretu bati buruz jakin beharrekoak (gauza, gertakizun, arrazonamendu, eta abar, sen ona azken finean) biltzen saiatzen direnak.

II.A.1. Corpusak

Linguistikaren barruan aspalditik izan dira linguistika enpirikoa aldarrikatu dutenak. Hauentzat, linguistika ahozko edo idatzizko hizkuntzaren azterketa enpirikoan oinarritu beharko litzateke (McEnery & Wilson, 1996). Idatzizko hizkuntzaren kasuan, azterketaren subjektua idatzizko corpus batek osatzen du. Testu multzo bat corpus izateko lau baldintza jartzen diote McEnery eta Wilson-ek: lagin errepresentatiboetan oinarritua egotea, tamainaz finitua izatea eta makinek tratatzeko modukoa izatea. Corpusek, gainera, errepresentatzen duten lengoaia-aldaeraren erreferentzia estandarra izateko bokazioa eduki beharko lukete.

Corpusetan oinarritutako linguistikak kritika zorrotzak jaso zituen 50. hamarkadan, iharduera asko murriztu zelarik. 80. hamarkadatik aurrera, ordea, onarpen zabala jaso izan du. Zalantzarik gabe, ordenadoreen ahalmena eta makinaz tratatu daitezkeen testuen kopurua etengabe hazten joatea, besteak beste, daude linguistika enpirikoaren berragerpenaren atzean. Gaur egun linguistikaren alor guztietara zabaldu du bere eragina, ezagutza-baseen aberasketara eta hitzen eta kontzeptuen arteko erlazio-izaeren ikerkuntzara ere. III., IV., V. eta VI. kapituluetan ikusiko ditugu corpusetan oinarritutako tekniken adibide batzuk.

Ingeleserako erreferentzia-corpus ugari sortu izan dira. Estatubatuarrak izan ziren aitzindari, Brown deritzon corpusarekin (Francis & Kucera, 1967). Bretainia Handiko ingelesarentzat ondoren etorri zen London-LUND corpusa (Svartvik, 1990), eta orduz gero etengabe ari dira corpusak berritu, sortu eta aberasten. Corpusean berez hitzak besterik ez daude, testu gordinak, baina corpusen erabilera asko zabaltzen da informazio linguistikoa gehitzen badiegu. Informazio hori hitzen kategoria izan daiteke, edo esaldien egitura sintaktikoa (adibidez, Penn Treebank delakoa edo Birmingham-eko *Bank of English* corpusa[3], Murrizpen-Gramatiken bidez (Karlsson et al. 1995) kategoria eta egitura sintaktikoz etiketatu dena), edo informazio semantikoa (aurrerago aipatuko dugun *SemCor*[4], hitzen adierez etiketatu den Brown corpusaren azpimultzoa, Miller et al. 1993a). Euskararako Euskaltzaindiak UZEIren laguntzaz bildu izan du Egungo Euskararen Bilketa Sistematikoa, gerra ondorengo testuen laginez osatutako miloi bat hitzetako corpusa (Urkia & Sagarna, 1990). IXA taldea euskara estandarra biltzen duen corpus zabalago bat biltzen ari da.

[3] http://titania.cobuild.collins.co.uk/boe_info.html

[4] http://www.cogsci.princeton.edu/~wn/

Tesi lan honetarako erabili ditugun corpusak Brown (ikus IV. eta V. kapituluak), Semcor (IV. kapituluan) eta Bank of English (V. kapituluan) dira. Aurrerago ikusiko ditugu hauei buruzko datu gehiago.

II.A.2. Hiztegiak

Lengoaia Naturalaren Prozesamenduan, 80. hamarkadarainoko sistemetan ahaleginaren gehiengoa sintaxi-egituretara eta sintaxitik semantikarako zubietara mugatzen zen. Lexikoa arazorik gabe beteko litzatekeen hitz zerrenda soil bat besterik izango ez zela uste zen. Garai horretan konturatu ziren LNPrako sistemen hedakuntzarako arazo nagusiena lexikoa urriegia izatea zela, eta lexikoa edukiz betetzea uste baina lan neketsuagoa zela. Garai berdinean, formalismo sintaktiko berri batzuk egitura sintaktikoen pisua lexikoira pasatzen hasi ziren, lexikoaren egitura konplexuago bihurtuz.

Lexiko zabal eta konplexuen eraikuntza eskuz egitea gehiegizko lana izango zela eta, hiztegietan zegoen informazioa ustiatzen ahalegindu ziren. Hiztegi elebakarretan hitzen kategoria, azpikategoria, definizioa, erabilera adibideak, etab. aurkitu daitezke. Gainera hitzen esanahiak antolatuta daude, adieren bidez. Thesaurus izeneko hiztegietan hitzak eremu semantikoen arabera multzokatuta daude, aurretik emandako sailkapen bat jarraituz. Berrikiago, hiztegi elebidunetan dagoen informazioa ere ustiatzen hasi da, bai hizkuntza batetik besterako ordainak, baita hizkuntza bateko kolokazio edo eremu semantikoa bezalako informazioa ere.

Hiztegi elebakarren artean, bat izan da tratatua bereziki, *Longman Dictionary of Contemporary English* deritzona (LDOCE, Procter, 1978). Bertako definizioak hiztegi mugatu bat erabiliaz egin dira, ingelesa ikasten ari direnentzat pentsatua. Bestalde, aditzen azpikategorizazioari buruzko informazioa, izenen kode pragmatikoak, arlo semantikoari buruzko kode semantikoak, eta abar jasotzen ditu. Lengoaia naturalaren prozesamenduan aipatzen diren beste hiru hiztegi *The Webster's Seventh New Collegiate Dictionary* (W7, Gove, 1969), *Oxford Advanced Learner's Dictionary of Current English* (OALDCE, Hornby, 1974) eta *Collins COBUILD English Language Dictionary* (CED, Sinclair, 1987) dira. Ingelesa ez diren hizkuntzatan hiztegi gutxi tratatu izan dira. Gaztelerarako, adibidez, *Diccionario General Ilustrado de la Lengua Española* (DGILE, Alvar, 1987) da formatu elektronikora pasatu den gutxietakoa. Frantseserako *Le Plus Petit Larousse* (LPPL, Larousse, 1980) dago, gure taldean analizatu izan dena eta tesi honetan landu duguna. Euskararako, *Euskal Hiztegia* (Sarasola, 1997) dago formatu elektronikoan.

Hiztegi hauen erabilera nagusiak, bertatik informazio sintaktikoa erauztea (adibidez, ALVEY-ko lexikoa horrela eraiki zuten, Boguraev & Briscoe, 1987) eta haiekin HEB edo EBL bat eraikitzea litzateke, hurrengo atalean ikusiko dugun bezala (VI. kapitulua ere aztertuko ditugu saiakera hauek).

Beste hiztegi mota bat thesaurusak dira, sarrerak eduki semantikoaren arabera antolatuta dauzkatenak. Lengoaia naturalaren prozesamenduan *Roget's Thesaurus* (Kirkpatrick, 1987) dezente erabili izan da.

Hiztegi elebidunen artean Collins argitaletxeak ingeles-gaztelania, ingeles-frantses, ingeles-italiera, eta abar eskuragarri dauzka formatu elektronikoan. Gaztelania eta ingelesaren artean ere bada *Diccionario Vox/Harrap's Esencial Español-Inglés* (Biblograf, 1992). Tesi lan honetan *Oxford French-English Dictionary* (OFED, OUP, 1989) erabili dugu. Euskarari dagokionean Elhuyarren euskara-gaztelania hiztegia (Elhuyar, 1996), Aulestia eta White-en euskara-ingelesa hiztegia (Aulestia & White, 1992) eta Morris-en euskara-ingeles hiztegia (Morris, 1998) formatu elektronikoan dauzkagu IXA taldean.

II.A.3. Ezagutza-base lexikalak eta hiztegi ezagutza-baseak

Lengoaia naturalen prozesamendu sintaktiko eta semantikoa egin ahal izateko, lexikoiak hitz zerrenda izatetik EBL izatera pasatu dira, hitz eta adierei buruzko informazioa dutenak. EBL baten hizkuntza ulertu ahal izateko ordenadoreak hitzei buruz jakin beharreko guztia egon beharko litzateke (Yokoi, 1995). EBLen ezaugarri garrantzitsuena heredentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen dira eta (Copestake, 1990). EBLak eskuz eraiki daitezke, adibidez WordNet (Miller et al., 1993b) eta EDR (EDR, 1993), baina askotan hiztegietatik erauzten dira (Copestake, 1990; Bruce et al. 1992).

LNParen beste ikuspuntu batetik, HEBek hiztegietatik erauzitako informazioa jasotzen dute (Artola, 1993). Erauzitako informazioaren artean, hemen ere, adieren hierarkiak dira aipagarriak. HEB batetik EBL bat eratorri daiteke, hiztegitik zuzenean EBL eraiki daitekeen bezala. HEB baten enfasia hiztegiko informazioan da, inplizitu egon eta esplizitu bihurtu dena, giza erabiltzaileak edo programa batek erabiltzeko moduan. EBL baten enfasia, ordea, LNP aplikazioetarako baliagarria izatea da. Tesi honi dagokionez, erlazio-izaera definitzeko beharrezko informazioa duten heinean ez gara gehiegi arduratuko EBL edo HEB bat denentz, eta ez ditugu bereiziko.

EBL eta HEBak eraikitzeko, hiztegietatik erauzi izan den informazio semantikoa definizioen azterketatik etorri ohi da batez ere, adieren hierarkia eratuz, eta hitzen (edo adieren) arteko bestelako erlazio lexikal-semantikoak finkatuz. Lehenbizi definizioen analisi sintaktikoa egin behar

da, eta ondoren analisiaren emaitzatik erlazio lexikal-semantikoen erauzketa. Erlazio horietan azaltzen diren hitzen desanbiguazioa ere egin behar da, adieren arteko erlazioak eduki ahal izateko. Honi buruzko zehaztasun gehiago ikusiko ditugu VI. kapituluan.

Atal honetan banan-bana ikusiko ditugu eskuz eraiki diren EBL batzuk (WordNet, EuroWordNet, Item eta EDR direlakoak) eta hiztegietatik egindako erauzketa automatikoan aritu diren proiektu batzuk (Acquilex, Nounsense, MindNet eta Hiztsua). Arlo honetan egin izan diren lanak ugariak izanda, ez gara zerrenda exhaustiboa egiten saiatu, esanguratsuenak biltzen baizik. Hiztegien erauzketari buruzko lan gehiago VI. kapituluan aipatuko ditugu.

II.A.3.a) WordNet, EuroWordNet eta Item

WordNet EBLa (Miller et al. 1993b) sinonimiaren inguruan antolatuta dago. Sinonimo multzo bakoitza, *synset* deritzona, hitzen adieraz eratuta dago, eta kontzeptu bat errepresentatzen du. WordNet-eko synset-en artean erlazio lexikal anitz daude, baina batez ere hiperonimia eta meronimia dira landuta daudenak. Synset-ak berez hierarkiatan antolatzen dira, baina multzo semantiko nagusietan ere multzokatuta daude. Izenen kasuan 15 eremu semantiko bereizten dira. Kontzeptu kopuruari dagokionez, WordNet 1.5 bertsioan orotara 91.591 kontzeptu daude 126.520 hitzentzat, izenen kasuan 60.557 kontzeptu eta 87.642 izen. WordNet edozeinek eskuratu dezake Internet bidez[5], eta gaur egun oso erabilia da LNP inguruko ikerkuntzan (artikuluen zerrenda bat ikusi daiteke WordNet-eko amaraun-orrian). Guk ere WordNet erabili dugu adieren arteko erlazio-izaera definitzeko (ikus III. kapitulua). Hurrengo atalean, WordNet-i buruzko zehaztasun gehiago ikusiko ditugu.

EuroWordNet[6] (Vossen, 1997) proiektua 1996an hasi eta 1999raino luzatuko den proiektu europarra da. EBL honek WordNet-en diseinuaren antzekoa erabiltzen du, baina Europako zortzi hizkuntzatara zabaltzen da. WordNet-en baino hizkuntza barneko erlazio mota gehiago daude, batez ere kategoria ezberdinen artekoak. Oraindik edukia guztiz bete gabe dago, baina dirudienez hemen ere batez ere hiperonimia erlazioa izango da landuena. Hizkuntzaren barne-erlazioez gain, kontzeptuak WordNet-eko synset-era lotuta daude, *Inter-Lingual Index* deritzonaren bidez, hizkuntzen arteko ordainak errepresentatuz. Horretaz gain, hizkuntzatik aparteko moduluan Goi-ontologia bat (*top ontology)* eta Domeinu-ontologiak (*domain ontology*) ere badaude. Lehenbizikoak WordNet ezberdinen goi aldeko synset-ak ezaugarri semantikoen arabera sailkatzea ahalbideratzen du, eta nolabait esateko, WordNet-en eremu semantikoen papera jokatzen du, nahiz eta motibazio

[5] http://www.cogsci.princeton.edu/~wn

[6] http://www.let.uva.nl/~ewn/

linguistiko sakonagoak hartu diren kontuan. 63 kontzeptu edo ezaugarri semantikok osatzen dute goi-ontologia hau. Hizkuntza bakoitzerako edukia, dagokion taldeak eskura dauzkan baliabideez baliatuz betetzen dute. Oraingoz ez dago edukien kopuruaren berririk, baina bai talde guztien artean adostutako oinarrizko kontzeptuen zerrenda, 1024 kontzeptuz osatua dagoena.

Donostiako Informatika Fakultateko Lengoaia Naturalaren Prozesamendurako IXA taldea[7] EuroWordNet proiektura lotuta dago, kanpoko eraikitzaile bezala. Horren inguruan, eta Estatu mailako ITEM proiektuaren[8] barnean, Euskararako WordNet-a eraikitzen ari gara EuroWordNet-eko diseinua jarraituaz. EuroWordNet-eko gaztelera eta euskararako WordNet-en eraikuntza automatikoan erabiltzen ari diren teknikak, tesi lan honen VI. kapituluan ikusiko ditugunen antzekoak dira, guk frantseseko hiztegia WordNet-i lotzeko erabili ditugunei hertsiki lotuak.

II.A.3.b) EDR

Japoniako ikerkuntza-agentziak, itzulpen automatikoan lexikoaren garapenak zeukan garrantzia ikusita, *Japan Electronic Dictionary Research Institute* sortu zuen 1986 urtean, japoniera eta ingelesaren tratamendu automatikorako lexikoa eraiki zezaten (Yokoi, 1995; EDR 1993). Proiektu erraldoi honek 9 urte ondoren bere emaitzak salmentan jarri zituen. Hizkuntza bakoitzerako corpusa, agerkidetza eta esaldi analizatuen baseak bildu zituzten, eta 300.000 hitz inguruko lexikoiak eraiki. Horretaz gain lexikoi elebidunak eta 4.000.000 kontzeptu biltzen dituen EBLa ere sortu dituzte. Kontzeptuak biltzen dituen ezagutza-baseak kontzeptuen deskribapenak eta kontzeptuen arteko erlazio lexikal anitz biltzen ditu, hierarkia osatzen duen azpiklase erlazioa izanik garrantzitsuena.

II.A.3.c) Acquilex

Acquilex[9] (Briscoe et al. 1993) proiektuaren helburua hiztegi elektronikoetatik (elebakar eta elebidunak) informazio lexikala erauzteko tresna eta metodologia automatikoak garatzea zen, Europako lau hizkuntzarentzat. Hiztegietatik erauzitako informazioarekin LNPrako aplikazioetarako EBL eleanitz baten prototipoa sortu zuten. Adibidez, izenetan janariari buruzko azpimultzoa landu zuten, ingeleserako 1.000 eta beste hizkuntzatarako 300 adiera inguru tratatuaz. Hiztegi elebakarretatik egindako erauzketa automatikoaren enfasia hierarkien eraikuntzan jarri zuten, nahiz eta bestelako atributu eta erlazioak erauzten ere saiatu. Hierarkia automatikoki eraikitzeko orduan ez zuten adiera-desanbiguaziorako irizpide automatiko garbirik proposatu (Copestake, 1990).

[7] http://ixa.si.ehu.es/

[8] http://sensei.ieec.uned.es/item/

[9] http://www.cl.cam.ac.uk/Research/NL/acquilex/

II.A.3.d) NounSense

NounSense LDOCE hiztegitik erauzi den EBLaren izena da (Wilks et al. 1996; Bruce et al. 1992). NounSense-en, hitzen definizioetako testuez gain, LDOCE-n dauden kode pragmatiko eta semantikoetatik erauzitako informazioa ere dago. Enfasi handiena izenen hierarkia eraikitzean jarri da eta horretarako hiztegiko definizioak analizatu eta desanbiguatu ondoren hiperonimia erlazioa erauzten da adieren artean. Emaitza bezala 39.000 adieren arteko hierarkia lortu zuten.

II.A.3.e) MindNet

MindNet (Richardson, 1997) LDOCE eta *American Heritage Dictonary* hiztegietatik erauzitako informazioaz eraiki da. Definizioetatik hiperonimiaz aparte beste 23 erlazio ere erauzi izan dira, izen, adjektibo eta aditzentzat, adierak desanbiguatu egin direlarik. Horrek ez du berrikuntza gehiegizkorik suposatzen berez, Acquilex, NounSense, eta beste hainbat proiekuetan ere planteatu izan da eta. Baina MindNet da teknika horien aplikazio zabalez eginiko lehenbiziko EBLa, eraginkortasun handia omen duena. Emaitza 191.000 adieratako sare semantikoa da.

II.A.3.f) Hiztsua eta Anhitz

Hiztsua (Artola, 1993) *Le Plus Petit Larousse* (LPPL) hiztegitik erauzitako Hiztegi Sistema Urgazle Adimentsua da. Bere funtzionalitatearen oinarrian automatikoki sortutako HEB aberats bat dago. Hiperonimia erlazioaren inguruan antolatuta dago batez ere baina beste 14 erlazio ere erauzi ziren, izen, aditz eta adjektiboentzat. LPPL-ko adiera guztietatik, nahiz eta denak analizatu, 6.130 adiera sartu ziren HEBan. Tesi honetako IV. kapituluan, kopuru hori zabaldu eta izenen 13.740 adierentzat eratorriko dugu hierarkia desanbiguatua. Aurrerago aipatuko ditugu HEB honen ezaugarri gehiago.

Hiztsua HEB elebakarra bada, Anhitz proiektuan (Arregi, 1995) eleaniztasunaren dimentsioa eransten zaio. Horretarako frantses eta euskarazko bi HEB elebakarren arteko zubia eraikitzen da, hiztegi elebidunetan oinarrituz. Itzulpengintzarako laguntza den sistema honen prototipoak 168 hitz eta 305 kontzeptu dauzka euskarazko partean eta 541 hitz eta 1139 kontzeptu frantseseko zatian. Horien artean 556 lotura elebidun landu dira.

II.A.4. Ontologiak

Arestian esan dugun bezala, ontologiak mundu errealaren kontzeptualizazioak dira, mundu errealari buruzko inferentziak egiteko gaitasuna dutenak. Definizio lauso hau aukeratu dugu, Adimen Artifizialaren arloan definizio zehatzagoek kontrobertsia pizten baitute, eta tesi honi dagokionean

ontologien ezaugarri bat izango delako guretzat garrantzitsuena: hierarkia darabiltela bizkarrezur bezala. Ontologiak aplikazio askotarako eraiki izan dira (softwarearen berrerabilgarritasuna, medikuntzako sistema-adituak, datu-base heterogeneoen integrazioa, lengoaia naturalen sorkuntza, ulermen, itzulpen, eta abar), eta normalean eremu espezifikoetarako eraiki ohi dira. Hala ere, badira ezagutza orokorragoa biltzen saiatzen direnak ere, adibidez Mikrokosmos, Sensus, CYC, etab. Ikus ditzagun ontologiak definitzeko egin diren saiakera batzuk:

> *Ontology ia a model of the world; an ontology defines the ways in which concepts are related, their relative significance, and their dependencies. The most significant relationship between concepts in the ontology is that of "hyponym/hypernymy" which determines if a concept belongs to the class defined by another concept.* (Onyshkevich & Nirenburg, 1994)

> *Ontologies are often equated with taxonomic hierarchies of classes, class definitions, and the subsumption relation, but ontologies need not be limited to these forms. … The word "ontology" seems to generate a lot of controversy in discussions about AI. It has a long history in philosophy, in which it refers to the subject of existence. It is also often confused with epistemology, which is about knowledge and knowing. …In the context of knowledge sharing, I use the term ontology to mean a specification of a conceptualization. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy* (Gruber, 1993)[10]*:*

> *An ontology is a specification of a conceptualization. … is a logical theory whose models constrain a particular conceptualization, without exactly specifying it…. In many cases, the axioms of an ontology only express subsumption (ISA) relationships between unary predicates, but of course a more detailed axiomatization is often necessary in order to exclude unwanted interpretations* (Guarino, 1997)

[10] http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Autore guztiak daude ados ontologiak oso hetereogeneoak direla esatean, norberaren beharretara neurrira eginak. Hala ere, ontologia denek edukitzen dute kontzeptu zerrenda bat eta kontzeptu horien arteko hierarkia, klase/azpiklase erlazioak egituratuta dagoena. Hori izaten da ontologien ezaugarririk garrantzitsuenetakoa, goian aipatutako definizio guztietan azaltzen dena. Ikus ditzagun Lengoaia Naturalaren Prozesamenduarekin zerikusia duten ontologia garrantzitsuenetako hiru, eta ondoren arituko gara ontologia eta HEB/EBLen arteko berdintasun eta ezberdintasunei buruz.

II.A.4.a) Mikrokosmos

Mikrokosmos-eko[11] ontologia (Onyshkevich & Nirenburg, 1994) Ontos zeritzon ontologiatik abiatuta garatu izan da. Mikrokosmos ezagutzan oinarritutako itzulpen automatikorako proiektua da. Sistema honetan lexikoia eta ontologia bereizi egiten dira. Lehenbizikoan informazio morfologikoa, sintaktikoa, eta abar dago, baita ere semantikari buruzko zenbait informazio ere. Ontologian munduari buruzko ohiko kontzeptualizazioa dago. Lexikoko hitzen eta ontologiako kontzeptuen arteko harremana ez da sinplea. Kasu sinpleenean hitzaren adiera bati ontologiako kontzeptu bat egokituko zaio, adibidez, *dog-n1* adierari *%dog* kontzeptua dagokio. Kasu konplexuagoetan ontologiako kontzeptura dagoen loturaz gain, kontzeptuari murrizpen osagarriak gehitzen zaizkio, adibidez, *eat-v1* adierari *%ingest* kontzeptua dagokio, baina honelako murrizpenekin: aditzaren subjektua ekintzaren agentea izan eta *%animal* kontzeptu azpian egon behar du ontologian (hautapen-murrizpena). Ezagutza semantikoari dagokionean beraz, kontzeptuen hierarkia ontologian dago, eta murrizpen eta hitz-kontzeptu loturak aldiz lexikoan.

Mikrokosmos-eko autoreen arabera ontologia bat ezin da egon *adieratan* oinarrituta, praktikoa izateko kontzeptu gehiegi egongo bailirateke ontologian[12]. Bestalde, ez zaie iruditzen unitate lexikalen esanahia primitibo gutxi batzuen bidez deskonposatzea bideragarria denik, eta lexikoko sarrerak konplexuegiak egingo lituzkeela diote. Horregatik eduki semantikoa ontologia eta lexikoaren artean banatu behar dela uste dute.

Ontologiak 4.500 kontzeptu dauzka[13], eta hiperonimiaz gain beste erlazioetan ere aberatsa da, kontzeptu bakoitzak batez-beste 14 erlazio dauzka eta. Horretaz gain, lexikoan ere badaude bestelako erlazioak ere, adibidez hautapen-murrizpenak.

[11] http://crl.nmsu.edu/Research/Projects/mikro/index.html

[12] Hobbs-ek (1995) *ontological promiscuity* deitzen zion honi.

[13] Lexikoian dauden adieren kopurua ezin izan dugu inon topatu.

II.A.4.b) Sensus

Sensus[14] (Hovy & Nirenburg, 1992), itzulpen automatikorako Pangloss[15] sistemaren ontologia da. Beste ontologia eta baliabide lexikalen bategitearen bidez sortu zen: Penman Upper Model (Bateman, 1990), Ontos (Mikrokosmosen aurretikoa), LDOCE eta WordNet. Ontologiaren goi aldea, *Ontology Base* deritzona, prozesamendu linguistikorako beharrezkoak diren desberdintasunak jasotzen dituzten 400 kontzeptuz osatuta dago. Ontos eta Upper Model-eko kontzeptuak eta izenen kasuan LDOCE-ko kode semantikoak eskuz bilduz sortu zen. Ontologiako gainontzeko kontzeptuak automatikoki gehitu dituzte WordNet-etik, Knight eta Luk-en (1994) artikuluan adierazten den bezala. Guztira 70.000 kontzeptu inguru dauzka ontologiak. *Ontology Base* deritzonean erlazio aberatsak daude, baina gainontzeko kontzeptuentzat hiperonimia besterik ez dago. Ontologia honi ingelesezko 90.000 hitz, japonierazko 120.000 hitz eta gaztelerazko 40.000 hitz lotu dizkiete.

II.A.4.c) CYC

CYC 13 urte baina gehiagotan lanean aritu den proiektu erraldoia izan da (Lenat, 1995)[16], pertsonok dugun sen ona ezagutza-base batean islatu nahi izan duena. Bere helburua ez da Lengoaia Naturalaren Prozesamendua bakarrik, Adimen Artifizialean planteatu izan diren arazo latz askori erantzun nahi izan baitio. Horretarako 100.000 kontzeptu eta kontzeptuen instantziei buruzko 1.000.000 bat baieztapen sartu izan dira ezagutza basean. CYC-eko lexikoan 14.000 lema daude. Duela gutxi ezagutza-basetik 3.000 kontzeptutako ontologia erauzi dute, edonork erabili dezan (ezagutza-base osoa erabiltzeko ordaindu beharra dago). Ontologia murriztu horretan klase/azpiklase eta instantzia-erlazioak besterik ez daude.

II.B. Ontologiak eta HEB/EBLak

Tesi honetan zehar, eta batez ere erlazio-izaera aztertzen duen III. kapituluan, ontologiei buruz arituko gara era zabalean, eta HEB eta EBLak ontologia bezala ere sailkatuko ditugu. Aurreko atalean ontologiak aztertu izan ditugunean, nabaria izan da kontzeptuen arteko klase/azpiklase hierarkiak duen garrantzia. HEB eta EBLetan ere erlazio hau da ezagutza-basea egituratzen duena, nahiz eta adieren arteko hiperonimia/hiponimia deitu.

Autore batzuk ontologia eta HEB/EBLen arteko diferentziak azpimarratzen saiatu dira, eta ez ditugu guk horiek ukatuko: kontzeptu guztiak lexikalizatuak egon behar duten edo ez, hierarkiaren

[14] http://www.isi.edu/natural-language/resources/sensus.html

[15] http://www.lti.cs.cmu.edu/Research/Pangloss/

[16] http://www.cyc.com/

goi aldearen antolatzeko irizpideak linguistikoak soilik diren edo ez, eta abar. Hala ere beraien arteko mugak ez daude garbi. Adibidez, EAGLES-eko lexikoari buruzko taldeak bere behin-behineko barne-txostenean[17] WordNet ontologiatik hurbil ikusten du, nahiz eta WordNet EBL bezala aurkezten den beti:

> *WordNet can best be characterized as somewhere in between a semantic network and a conceptual ontology. The synsets are conceptual units rather than lexical semantic units. The relations are better seen as semantic inferencing schemes than as lexicalization patterns.*

Beste konparazio lan batean, *Communications of the ACM* aldizkariaren 1995.eko azaroko alean, EDR, CYC eta WordNet bata bestearekin alderatzen dira, errepresentazio oinarri eta eduki kopuruen arabera. Aipatzen den diferentzia nabariena orientazioa da: CYC-ek sen ona eta munduari buruzko inferentziak jaso nahi dituen bitartean, EDR eta WordNet inferentzia linguistikoetarako daude prestatuta. Baina diferentzia hori ez da hain garrantzitsua, EAGLES-en barne txostenean aitortzen den bezala EuroWordNet-i buruz ari direnean:

> *EuroWordNet is different from AI-ontologies such as CYC or Sensus/Pangloss in that its focus is on the linguistically-motivated relations rather the semantic inference schemes only. In this respect it provides information on the exact semantic relation between the lexicalized words and the expressions of languages (this may still be useful for making inferences as well).*

Diferentzia nagusia orientazioan dagoela uste dugu guk ere: ontologiatan munduari buruzko informazioa dugu, kontzeptuen arteko erlazioak ez dute zertan motibazio linguistikorik eduki behar. Bestalde, EBLak hizkuntzaren ulermen eta sormenerako beharrei erantzutera mugatzen saiatzen dira, baina azken finean jakina da LNP *AI-complete* dela, hau da, adimen artifizialeko arazo garrantzitsuenak, sen ona barne, ebatzi behar direla LNP osoa egin ahal izateko. Beraz, EBLetan munduari buruzko informazioa egon behar da. Adibide garbi bat hiperonimia erlazioa da. Izan ere ontologietan eta EBLetan gordetzen den informazio semantikoa gainjarri egiten da, biak egitura isolatu bezala diseinatuko balira, ezagutza bera bi aldiz errepresentatu beharko litzateke, adibidez hiperonimiari dagokiona.

Tesi-lan honi dagokionez, kontzeptu/adieren arteko erlazio-izaera landu nahi dugunez, kontzeptu/adieren arteko erlazioak dira interesatzen zaizkigunak, bereziki klase/azpiklase edo

[17] "Preliminary Recommendations on Semantic Encoding" (Interim Report, May 1998). http://www.ilc.pi.cnr.it/EAGLES96/rep2/

hiperonimo/hiponimo. Erlazio horiek ontologia eta EBLetan topatu ditzakegunez (ontologiatan kontzeptuen arteko erlazio semantiko eta pragmatiko bezala, eta EBLetan adieren arteko erlazio lexikal-semantiko bezala) ez zaigu axolako beraien jatorria. Tesi honetan garatu dugun erlazio-izaera berdin aplikatu daiteke ontologia, EBL edo HEBetan azaltzen diren erlazioetara.

Tesi lan honetan ontologia eta EBL artean bereiziko ez dugun bezala, kontzeptu eta adiera inongo diferentziarik egin gabe erabiliko ditugu, nahiz eta jakin adiera lotuago dagoela hiztegi eta EBLetara, eta kontzeptua ontologietara.

II.C. Erabili ditugun baliabide lexikalak

WordNet da zalantzarik gabe tesi honetako baliabide oinarrizkoena, III. kapituluan definitzen dugun erlazio-izaera WordNet-eko erlazioen gainean inplementatu dugu eta. WordNet aukeratzeko arrazoiak beherago ikusiko ditugu. Corpusei dagokionez, hitzen adiera desanbiguazioan (ikus IV. kapitulua) lortutako emaitzak ebaluatzeko SemCor erabili dugu. Zuzenketa automatikoari buruzko ikerketan (V. kapitulua) Brown eta Bank of English.

VI. kapituluan HEB bat aberastuko dugu, *Le Plus Petit Larousse* hiztegitik erauzi izan dena, eta horretarako WordNet-era lotuko dugu *Oxford French/English Dictionary*-ren bitartez.

Ikusi dezagun arreta handiagoz zeintzuk diren baliabide hauen ezaugarriak.

II.C.1. Brown eta Semcor

Brown deritzon corpusak (Francis & Kucera, 1967) Estatu Batuetako ingeles idatziko 1.000.000 bat hitz jasotzen ditu. Idatzizko genero ezberdinetatik laginak jasoaz burutu izan da. Jasotako genero batzuen adibideak: *press-reportage, press-editorial, learned-science* eta *humour* dira.

Semcor Brown corpusaren azpimultzo bat da, WordNet egin zuen talde berak etiketa semantikoak gehitu dizkiona (Miller et al. 1993). Brown corpuseko 186 testu daude barnean, eta hitz guztietatik –359.732–, adjektibo, izen, aditz eta adberbioak daude WordNet-en dagokien adierarekin markatuta –192.639– (ikus 2. taula). Adieraz etiketatu daudenetik 666 hitzek jaso dute adiera bat baino gehiago, bi esanahirekin erabili direlakoan. Gainontzeko guztiak adiera bakarraz daude etiketatuta.

Hitz kopurua	359.732
Adieraz etiketatuta	192.639
Adiera anitzez etiketatuta	666

2. taula: Semcor-en datu batzuk

Corpus honetan horrela azaltzen da "*The couductor said to Ritchie:*" esaldia (WordNet 1.4 bertsioaren arabera etiketatuta):

```
<s>
<stn>50</stn>
<wd>The</wd><tag>DT</tag>
<wd>conductor</wd><sn>[noun.person.1]</sn><tag>NN</tag>
<wd>said</wd><mwd>say</mwd><msn>[verb.communication.0]</msn><tag>VBD</tag>
<wd>to</wd><tag>TO</tag>
<wd>Ritchie</wd><df>person</df><sn>[noun.Tops.0]</sn><pn>person</pn><tag>NP</tag>
<wd>:</wd><tag>:</tag>
</s>
```

Etiketak SGML formatua jarraitzen dute. Hitz-formak `<wd> </wd>` artean daude, kategoria sintaktikoa `<tag> </tag>` artean ematen da, eta etiketa semantiko `<sn> eta </sn>` artean. Adibidez *conductor* hitza izen bat da (NN) eta esaldi horretan `noun.person.1` bidez errepresentatzen den adiera dagokio, hau da, `person` kode semantikoa duen lehenbiziko adiera (WordNet ikustean komentatuko dugu zer diren kode semantiko horiek). Izen berezien kasuan etiketa semantikoa izen berezi horrek ordezkatzen duen entitatearen araberakoa da, adibidez *Ritchie*-ri pertsonaren adiera bat egokitu zaio. WordNet aipatzean ikusiko dugu etiketa semantikoen esanahia.

II.C.2. Bank of English

Collins hiztegigintzako konpainiaren COBUILD proiektuaren barnean[18], ingelesaren bilakaera monitorizatzeko corpusa da, Birmingham-eko Unibertsitatearen laguntzarekin jasotzen ari dena[19]. 1996.enean corpusak 320 miloi hitz zeuzkan eta hazten darrai egunotan ere. Brown corpusa ez bezala ezin da libreki eskuratu, eta baimena eskatu behar da corpusaren zatiak ikusi ahal izateko.

II.C.3. WordNet

WordNet (Miller et al. 1993b) da zalantzarik gabe tesi honetako baliabide oinarrizkoena. III. kapituluan definitzen dugun erlazio-izaera ontologietako erlazioetan oinarritzen da, eta eskuragarri dauden ontologien artean[20] hitz kopuru aberatsena duenez (126.520), WordNet hautatu dugu erlazio-izaeraren inplementazioa gauzatzeko. Beste hautagaiak Mikrokosmos eta Sensus ziren. Lehenbizikoak erlazio aberatsak dauzka kontzeptuen artean, baina lexikoa nahiko mugatua du (hitz kopururik ez dute aipatzen, baina bai 4.500 kontzeptu dituela). Bigarrena, hein handi batean, Mikrokosmos eta WordNet bat egitetik sortu zen. Hitz kopuru interesgarria dauka (90.000), baina

[18] http://titania.cobuild.collins.co.uk/

[19] http://titania.cobuild.collins.co.uk/boe_info.html

[20] Ontologia zabal gehienak lortu ahal izateko, CYC eta EDR kasu, gogotik ordaindu behar dira. MindeNet-en kasuan ezta ordaintzen ere ezin da eskuratu. Beste ontologia batzuk barne-erabilerarako dira, eta ez daude prestatuta kanpokoak erabiltzeko (adibidez, NounSense).

era automatikoan eraiki zenez erroreak daude hierarkian. Tamalez ez da errore horren neurririk ematen (Knight & Luk, 94). Azkenik, aipatu behar da WordNet oso erabilia dela LNP inguruko ikerkuntzan eta edozeinek eskuratu dezakeela Internet bidez[21].

WordNet Estatu Batuetako ingelesarentzat eraiki den EBLa da. Diseinatzeko orduan psikolinguistikako printzipioak aplikatu nahi izan dituzte. Kategoria nagusiek (izen, aditz, adjektibo eta adberbioak) sistema erlazional separatuak eratzen dituzte. Sistema erlazional horiek sinonimo multzoa (*synset*) daukate unitate kontzeptual bezala. Hitz batek adiera anitz baditu hainbat synsetetan azalduko da, eta adiera bakarra badu synset bakarrean. Adibidez *woman*-ek lau adiera dauzka, bakoitzean sinonimo ezberdinak dituelarik:

```
1. woman, adult female
2. womanhood, woman
3. charwoman, char, cleaning woman, cleaning lady, woman
4. woman ((informal) a female person who plays a significant role
```

4. adierak ez du sinonimorik, eta beraz glosa bat ere ematen du (glosa horiek beste adierentzat ere lortu daitezke.

Synset-en artean erlazio lexikal-kontzeptualak definitu dira (ikus 3. taula). Sinonimiaz gain, izenen kasuan garrantzitsuena hiperonimia da, hierarkia eratzen duena. Adibidez, *woman*-en lau adieren hiperonimoak hauek dira:

```
woman, adult female                          => female, female person
womanhood, woman       => class, social class, socio-economic class
charwoman, char, cleaning woman, cleaning lady, woman    => cleaner
woman                                        => female, female person
```

Bestelako erlazioen artean meronimia eta antonimia ere azaltzen dira, baina ez daude hain sistematikoki landuta. Izenen artekoa ez den erlazio bakarra *ezaugarri* erlazioa da, izen eta adjektibo bat lotzen baititu. Adibidez *canary*-ren ezaugarri bat *small* izatea da. Erlazio bakoitzak bere alderantzizkoa ere badu.

WordNet-en 1.5 bertsiorako izenen datuak 3. taulan ikus daitezke. Izenek batez-beste 1,22 adiera dauzkate[22]. Erlazioei dagokienez gehienak hiperonimia eta hiponimia erlazioak dira, eta synset

[21] http://www.cogsci.princeton.edu/~wn

[22] Kontuan izan synset bat hitz bat baina gehiagoren adiera izan daitekeela, beraz 1,22 ez da izen eta synset kopuruaren arteko zatiketa soila.

bakoitzak bana dauka batez-beste. Meronimia edo holonimia erlazioak synset-en erdiak dauzkate, eta gainontzeko erlazioak askoz gutxiago azaltzen zaizkigu.

		Kopurua	Izeneko	Synset-eko
Izenak		87.671		
Synset-ak		60.631	1,22	
Erlazioak	Hiperonimia/hiponimia	122.246		2,01
	Meronimia/holonimia	35.067		0,58
	Antonimia	1.713		0,03
	Ezaugarriak	645		0,01
	Guztira	159.670		2,63

3. taula: WordNet 1.5-eko datu batzuk izenentzat

Informazio honetaz gain WordNet-eko izenen synsetak 26 eremu semantikotan sailkatuta daude. Eremu horiek 4. taulan daude. WordNet-en izen baten adiera zuzenean adieraz daiteke, edo zeharka, eremu semantiko horren arabera. Adibidez *conductor* izenaren synset bat, garraio publikoan kobratzen duen pertsonari dagokiona[23], *person* eremu semantikoari dagokio. Synset hori bi eratara adieraz daiteke, bere 3. adiera bezala, edo `[noun.person.1]` bezala, hau da, *person* eremu semantikoan *conductor*-ek duen lehenbiziko adiera bezala[24]. Eremu semantikoen artean noun.Tops berezia da, hierarkien goialdean dauden synset-ak bildu besterik ez du egiten eta.

noun.Tops	noun.feeling	noun.possession
noun.act	noun.food	noun.process
noun.animal	noun.group	noun.quantity
noun.artifact	noun.location	noun.relation
noun.attribute	noun.motive	noun.shape
noun.body	noun.object	noun.state
noun.cognition	noun.person	noun.substance
noun.communication	noun.phenomenon	noun.time
noun.event	noun.plant	

4. taula: WordNet-eko izenen kode semantikoak

II.C.4. LPPL

Le Plus Petit Larousse (Larousse, 1980) frantseseko hiztegi elebakarra da. 5. taulan ikus daitezke hiztegiaren datuak. Hiztegi honen gainean ikerkuntza ugari egin ditu gure taldeak. Lehenbizi Datu-Base Lexikal bat eratu zen hiztegiko informazio guztiarekin: sarrera, adiera zenbaki, kategoria, erabileremu, definizio eta adibide. Definizioen gainean egindako analisi sintaktikotik hainbat erlazio lexikal-semantiko erauzi ziren. Izenen kasuan honako erlazio hauek erauzi ziren: sinonimia eta antonimia, hiperonimia, meronimia, gabezia, erreferentziazkoa, eratorpena eta kasu-erlazioak.

[23] Synset hori horrela adierazten da WordNet-eko interfazean: *conductor -- (the person who collects fares on a public conveyance)*

[24] Eremu semantiko berdineko beste adieraren synseta: *conductor, music director, director1 -- (the person who leads a musical group)*

	Guztira	izenak
Sarrerak	15.953	10.506
Adierak	22.899	13.740
Hiztegiko hitzak (guztira)	97.778	66.323
Definizioen luzera (batez-beste)	3,27	3,82

5. taula: LPPL-ko datuak

Erlazio lexikal-kontzeptual horiekin Hiztegi-Ezagutza Base bat eratu zen, sare semantiko baten itxura zuena.

II.C.5. OFED

Oxford French-English Dictionary (OUP, 1989) neurri ertaineko hiztegi elebiduna da. Hiztegi honen frantses-ingeles zatia bakarrik dugu makinak irakurtzeko formatuan. Hiztegiari buruzko datuak 6. taulan ikus ditzakegu. Hiztegiak izenentzat 13.030 sarrera dauzka. Sarrera bakoitzak jatorrizko hitzarentzat adiera bakarra edo gehiago eduki ditzake. Halako adiera elebidun bakoitzari azpisarrera deituko diogu lan honetan. Adibidez *maintien* izenaren sarrera bi azpisarreratan bana daiteke:

> *maintien n.m. (attitude) bearing; (conservation) maintenance*

> *maintien 1: n.m. (attitude) bearing*
> *maintien 2: n.m. (conservation) maintenance*

Hiztegi elebidunak 16.917 halako azpisarrera dauzka izenentzat. Beste ikuspegi batetik ikusita, 13.030 izen frantses eta 11.969 izen ingeles daude hiztegian (ikus 6. taula).

	Sarrera kop.	Azpisarrera kop.
Guztira	21.322	31.502
Izenak	13.030	16.917
Ingeles izenak	11.969	–

6. taula: OFED hiztegi elebiduneko datuak

Azpisarreren barruan hainbat eremu azaldu daitezke: kategoria (derrigorrez, adibidez izen maskulinoa, *n.m.*), eremu semantikoa (aukerakoa, 20 eremutako bat izan daiteke, adibidez beheragoko adibideko *comm.*, komertziala), frantsesez dagoen argibidea (aukerakoa, adibidez goiko *attitude* eta *conservtion*, edo beheko *ressources*), eta azkenik derrigorrezkoa den ingelesezko itzulpen-hitza edo hitz-zerrenda. Eremu semantikoa eta frantsesezko argibidea azpisarrera horretako itzulpenaren argibideak dira, testuinguru edo erabilpenari buruzko oharrak, hiztegiaren erabiltzaileari itzulpena hautatzean laguntzeko.

folie 1: n.f. madness
provision 1: n.f. supply, store
trésor 2: n.m (ressources) (comm.) finances

III. Kapitulua

ERLAZIO-IZAERA ETA DENTSITATE KONTZEPTUALA

Kapitulu honen helburu nagusia ezagutzan oinarritutako kontzeptuen arteko erlazio-izaera definitzea da, eta horretarako WordNet-en oinarritutako Dentsitate Kontzeptuala diseinatu eta inplementatu dugu. Lehenbizi erlazio-izaera zer den azalduko dugu, eta literaturan azaldu diren lan garrantzitsuenak gainbegiratuko ditugu, lan bakoitzak erabili izan duen baliabide lexikalaren arabera sailkatuta. Hurrengo atalean, ontologiatan oinarritzen den Dentsitate Kontzeptuala azalduko dugu, bere aurrekaria den Distantzia Kontzeptualarekin batera. Ondoren WordNet-erako egin dugun inplementazioa agertzen da. III.D. atalean Dentsitate Kontzeptualaren ezaugarrien ebaluazioa egin, eta gainontzeko proposamenekin alderatuko dugu. Bukatzeko, kapitulu honen ekarpen nagusiak eta etorkizunerako lana labur aipatuko ditugu.

III.A. Sarrera eta aurrekariak

Kapitulu honen helburuan sakondu aurretik, lan honetan erabiliko dugun terminologia finkatuko dugu, antzekotasunaren literaturan gertatzen den nahastea argitu nahian. Bi ideia nagusi dira ardatz, maiz nahasten direnak: **antzekotasuna** (ingelesezko *similarity*) eta **erlazio-izaera** (*relatedness*). Lehenbizikoa bi objektuk elkarren antza dutela adierazteko erabiltzen da, adibidez goilare eta sardexka. Bigarrena bi objektu horien artean nolabaiteko erlazioa (lotura, harremana) badagoela esateko, adibidez sardexka eta txuleta artekoa. Antzekoak diren bi gauza erlazionatuta egongo dira (adibideko goilare eta sardexka), noski, baina alderantziz ez da gertatzen: erlazionatutako bi gauzek ez daukate zertan antzeko izan behar (adibideko sardexka eta txuleta). Literaturan antzekotasun hitza da zabalduena, sarritan erlazio-izaera beharko lukeen tokian erabilia. Gure ustez orokorrean erlazio-izaeraz hitz egin daiteke, eta antzekotasuna azpimultzo bat izango litzateke. Antzekotasun

eta erlazio-izaerari ontologien inguruko lan batzuetan **distantzia semantikoa** (*semantic distance*) kontrajarri ohi zaio: antzekotasun handiko bi kontzepturen artean distantzia semantiko txikia egongo litzateke. Antzekotasuna eta distantzia semantikoa bata bestearen alderantzizkoa dira, eta beraz distantzia semantikoa definitzea ez da beharrezkoa. Tesi honetan, hala ere, distantzia semantikoa ez baina **distantzia kontzeptuala** erlazio-izaeraren neurri eta inplementazio konkretu bezala bai azalduko zaigula.

Kontua da erlazio-izaera hori askoren ustetan lengoaia naturala ulertzeko giltzetako bat dela. Ulermenerako giltza edo ez, LNParen aplikazio konkretu askotan erabiltzen da erlazio-izaeraren inplementazioren bat edo beste: zuzenketa automatikoan (ikus V. kapitulua), informazioaren berreskuratzean (*Information Retrieval*), dokumentuen berreskuratze eta sailkapenean (*Document indexing and retrieval*) (Sussna, 1993), multzokatzean (*clustering*) (Schütze 1992a; 1992b), desanbiguazioan (anbiguetate sintaktikoan –adibidez, ingelesezko *prepositional phrase attachment* arazoan (Resnik, 1993)– edo hitzen adiera-desanbiguazioan, ikus IV. kapitulua), ontologien eraikuntzan (taxonomiak eraikitzean –ikus VI.A atala– , hautapen-murrizpenak ikastean (Grishman & Sterling, 1994), ontologiak biltzean (Knight & Luk, 1994; Utiyama & Hasida, 1997), edo ontologien ebaluazioan (Rada et al., 1989)) eta baita ere interpretazio semantikoan (EDR, 1993).

Arrazoi honengatik erlazio-izaerari buruzko literatura zabala da. Artikulu gutxi batzuetan bera da helburua edo aipatu egiten da, baina artikulu asko aplikazio konkretu bati buruzkoak dira eta ez da erlazio-izaerari buruz esplizituki hitz egiten, nahiz eta inplizituki erlazio-izaeraren neurriren bat definitu. Gai honen inguruko literatura aztertzean tesian aipatuko ditugun artikulu ugari azalduko zaizkigu, nola edo hala erlazio-izaeraren neurriren bat erabiltzen dute eta.

Artikulu eta lanak sailkatzea ez da makaleko lana, bai kopurua bera itzela delako, bai hurbilpen oso ezberdinak daudelako. Nolabait esateko badirudi ikertalde bakoitzak bere erlazio-izaeraren formalizazioa bilatu duela. Formula guztiek dauzkate ahuleziak eta aldeko ezaugarriak, alor hau heldutasunera heldu ez denaren seinale, ziur aski. Baina ulergarria da, bestalde, kontuan hartzen badugu aplikazio askotarako beharrezkoa izanda, ikertalde bakoitzak abiapuntu ezberdinetik heldu diola erlazio-izaerari. Denak sakonean aztertzea, beraz, tesi lan honen helburutik kanpo dago, baina bai arrakastatsuenak eta lan honetatik hurbilago daudenak sailkatu eta aztertzen saiatuko garela. Sailkapena egituratzeko irizpide nagusi bat erabili dugu, erabilitako baliabidearena: ontologia, hiztegi elektronikoa, corpusa edo horien konbinazioren bat.

Beste kontzeptu batzuk ere erabiliko ditugu lanak sailkatu ahal izateko. Hasteko, hurrengo bereizketa egingo dugu bi hitz edo bi kontzepturen arteko erlazioen inguruan:

1. **Erlazio paradigmatikoa**: linguistikoki, esaldi batean hitz bat beste baten ordez trukatu daitekeenean. Kontzeptualki, munduaren ontologia jakin baten arabera, mota berdineko kontzeptuak direnean. Hau da antzekotasun bezala ulertu daitekeena, antzeko kontzeptuak klase berean egoten baitira sailkatuta.
2. **Erlazio sintagmatikoa**: linguistikoki, bi hitz kate mintzatu berean azaltzen direnean. Koordenatu pare batez esan daiteke erlazio paradigmatikoa bertikala baldin bada, sintagmatikoa horizontala dela (UZEI, 1982). Kontzeptualki, mota ezberdineko kontzeptuak izanda ere horien artean erlazioa dagoenean. Hau da berez erlazio-izaera. Testuinguruari dagokionez are gehiago bereizi ohi dira:
    - **Erlazio sintagmatiko lokala**: kolokazioak dira honen adibide bat, adibidez 'on egin', edo argumentu egituran azaltzen dena, adibidez 'urdaiazpikoa jan'. Halakoetan bi hitzak esaldian gertu egon ohi dira, gutxi gora behera.

    - **Erlazio sintagmatiko globala**: erlazionatutako hitzak ez dira zertan hurbil azaldu behar, ezta esaldi berean ere. Topiko edo mintzagaiarekin lotutako erlazioak azaltzen zaizkigu hemen: adibidez sukaldaritza balizko topikoari lotutako urdaiazpiko, lapiko, sardexka, sukalde, etab. Mintzagaiak erlazionatzen dituela esan daiteke beraz.

      Kasu batzuetan ez da eskatzen bi hitzak kate mintzatu berean azaltzea, baizik eta ezaugarri (sintaktiko edo semantiko) amankomuneko kate mintzatuetan. Honela, bi hitz antzeko testuinguruetan maiz azaltzen badira, beraien artean **zeharkako erlazio sintagmatiko globala** dagoela esan daiteke.

Bereizketak lauso samarra iruditu arren, aurrerago ikusiko dugu lan praktiko ugaritan ondo nabarituko dela bata edo bestearen erabilera, formalizazio bakoitzak aukeratu duen hurbilpenaren arabera.

Beste bereizketa bat hitzen arteko erlazio-izaera eta kontzeptuen arteko erlazio-izaera bereizten dituena da. Bigarrena da gehien interesatzen zaiguna, hau da, erlazio linguistikoa baino kontzeptuala. Adieraren garrantziaz jabetzeko zera dio Hirst-ek (1987: 5 or.): "*Any practical NLU system must be able to disambiguate words with multiple meanings, and the method used to do this must necessarily work with the methods of semantic interpretation and knowledge representation used in the system*". Ontologia eta Hiztegi Ezagutza Baseak (HEB) ere kontzeptuen inguruan antolatu ohi dira, adibidez WordNet-en: "*The most ambitious feature of WordNet, however, is its attempt to organize lexical information in terms of word meanings, rather than word forms*" (Miller et al., 1993b: Sarrerako 3 or.). HEB eta EBLetan badaude

salbuespenak, sistema batzuetan ezin izan baitute adiera-desanbiguazioa aurrera eraman, baina adieraren inguruan antolatu beharra aitortzen dute, Richardson kasu (1997: 113 or.): "*In the future, this approach may be much more viable with a sense disambiguated Lexical Knowledge Base, which is work currently in progress.*".

Hitzen edo adieren artekoa izanda ere, bi erlazio-izaerak hertsiki erlazionatuta daude. Linguistikoki antzekoak diren hitzak, beraien adieraren batean kontzeptualki antzekoak ere izango dira, eta alderantziz, antzekoak diren bi kontzepturentzat ahoskatzeko balio duten hitzak ere antzekoak izango dira.

Orain artekoa kontuan izanik, erlazio-izaerari buruzko lanen azterketan sei ezaugarriri egingo diegu arreta berezia:

- erabiltzen den baliabidearen ingurukoa: hiztegi, ontologia, corpus edo nahasketa.
- erlazio-izaera paradigmatiko edo sintagmatikoa (global/lokala) den.
- hitz edo kontzeptuen arteko erlazio-izaera den.
- testu zabalekin ebaluatua, hitz gutxi batzuekin ebaluatua edo ebaluatu gabeko proposamena den.
- izenekin soilik ebaluatua edo kategoria guztiekin ebaluatuta dagoen.
- lortutako emaitzak: emaitzarik ez, kaxkarrak, onak edo oso onak, azaldutako doitasunaren arabera.

Esan dugun bezala, erlazio-izaeraren formalizazioen ebaluazioa ez da erraza. Batzuetan zuzenean pertsonen iritziz *ad hoc* osatutako zerrendekin parekatuaz egiten da, baina gehienetan adiera-desanbiguazioan, informazioaren berreskuratzean edo beste aplikazio espezifiko batean lortutako emaitzen bidez zeharka egiten da. Lehenbizikoak duen arazoa zera da, pertsona ezberdinek osatzen dituzten zerrendak ez direla guztiz bat etortzen, eta zerrenda eraikitzeko irizpide garbirik ez dagoela. Erlazio-izaera erabiliaz automatikoki eraikitako zerrendak pertsonek eraikitakoekin parekatzean, zerrendak antzekoak badira orduan formalizazioa ontzat hartzen da, baina ez da kontuan hartzen bat ez datozen hitzak zerikusirik duten edo guztiz erratuak dauden.

Aurrekarien azterketari ekingo digu orain. Goian aipatutako ezaugarrietan arreta berezia jarriko dugunez, sistema edo lan esanguratsuenak komentatu ondoren lerro batean laburbilduko dugu ezaugarri bakoitza zertan den lan konkretu horrentzat.

III.A.1. Ontologian oinarritutako aurrekariak

Ontologia (ikus II. kapituluan zer den ontologia lan honi dagokionean) oinarritzat hartuz gero, bi objekturen arteko erlazio-izaera ontologia bertan dagoen informaziotik erauzi daiteke. Psikologiaren eremutik sortu zen antzekotasunaren lehenbiziko axiomatizazioan Tversky-k (1977) zera zioen: "*A new set-theoretical approach to similarity is developed in which objects are represented as collections of features, and similarity is described as a feature-matching process*". Ezaugarrietan oinarritutako errepresentazio eredua erabiltzen zuen beraz. Bere neurria zeregin ezberdinetara aplikatzen du, hala nola, hizkien antzekotasuna, aurpegien antzekotasuna eta nazioen antzekotasuna. Ebaluazioan bere axiomatizazioa antzekotasunari buruzko giza iritziarekin alderatu zuen.

Adimen Artifizialean ordea, garai horretan sare semantikoak ziren errepresentazio eredu ohikoenak, eta antzekotasuna batez ere *spreading activation* izeneko tekniken bidez landu zen (Quillian, 1968; Collins & Loftus, 1975). Collins eta Loftus-en arabera "*The conceptual network is organized along the lines of semantic similarity. The more properties two concepts have in common, the more links there are between the two nodes via these properties and the more closely related are the conccepts*"[25]. Ez zuten beraien eredua zuzenean inplementatu, baina psikolinguistikako esperimentuen araberako datuekin bat omen zetorren.

Spreading activation horren inplementazioa errazte aldera, Rada-ren taldeak (Rada et al., 1989) lan ugari egin zituen sare semantikoen ebaluazio eta fusionatzearen inguruan. Beraiek aurkezten duten erlazio-izaeraren neurriari Distantzia Semantikoa (*Semantic Distance*) deritzo:"… *[in spreading activation] semantic relatedness is based on an aggregate of the interconnections between the concepts. This is different from semantic distance which is equal to the minimal path length between two concepts*". Are gehiago, kontuan hartuaz sare semantikoak egituratzen dituen erlazio pribilegiatu bat egon badagoela – klase-azpiklase edo *is-a* erlazioa – erlazio mota guztiak erabili ordez azpiklase erlazioa nahikoa dela diote: "*we hypohesize that […] is strong enough for the length of is-a paths to be used as a measure of semantic relatedness*". Proposatzen duen distantziaren formulan (1. ekuazioa) A eta B kontzeptuen arteko distantzia bi kontzeptuen arteko *is-a*[26] erlazioz osatutako bide motzenaren luzera da.

$$\mathrm{dist}(A,B) = \min_{p \in \mathrm{path}(A,B)} \mathrm{length}(p) \qquad (1)$$

[25] Ikusten den bezala antzekotasun eta erlazio-izaera kontzeptuak nahastu egiten dira hemen ere.

[26] Tesi honetarako beharrezkoa ez denez, ez dugu ezberdinduko *is-a*, klase/azpiklase edo hiperonimo/hiponimo erlazioen artean.

Distantziaren neurria txikia litzateke hertsiki erlazionatutako bi kontzepturentzat, eta alderantziz. Ez dute ebaluaziorik adierazi. Bere sinpletasunean nahikoa erabili izan da, adibidez ontologia ezberdinak biltzeko (Knight & Luk, 1994; Utiyama & Hasida, 1997).

[27]Ontologia/paradigmatikoa/kontzeptuak/gutxi/izenak/emaitzarik ez

Sussna-k (1993) Rada-ren taldearen ideia landu eta WordNet ezagutza-baseari aplikatzen dio, dokumentuak indexatzeko adiera-desanbiguazioa beharrezkoa dela eta. Ezagutza-baseko kontzeptuak adierak dira kasu honetan, eta azpiklase erlazioaz gain WordNet-ek dauzkan beste guztiak ere erabiltzea proposatzen du. Erlazio bakoitzak antzekotasun pisu bat edukiko du (2. ekuazioko[28] $w_r(x,y)$), antzekoagoak baitira adibidez sinonimo-erlazio bidez lotutako kontzeptuak, zati (*part-of*) erlazioaz lotutakoak baino (ikusi baita ere Tversky 1977). Sare semantikoan elkarren ondoan dauden bi kontzepturen arteko distantzia ($w(x,y)$ 2. ekuazioan) bi kontzeptu horien artean dagoen erlazio guztien pisuen batura izango da. Aipatzekoa da, gero eta sakonago egon kontzeptuak hierarkian gero eta distantzia txikiagoa aitortzen diela (hori da *d* zatitzailea).

$$w(x, y) = \sum_{r \in Wordnet-relation} \frac{w_r(x, y)}{d} \qquad (2)$$

Edozein bi kontzepturen arteko distantzia, beraz, lotzen dituztenen bide guztien artean pisu txikiena duenak emango digu (3. ekuazioa).

$$\text{dist}(x, y) = \min_{(x, x_1, \ldots, x_n, y) \in \text{path}(x, y)} \sum_{i=0}^{n} w(x_i, x_{i+1}) \qquad (3)$$

$$\text{non } x = x_0 \text{ eta } y = x_{n+1}$$

Sussna-k ere ez du ebaluazio zuzenik egiten, zeharka adiera-desanbiguazio esperimentuetan lortutako emaitzen arabera baizik.

Ontologia/paradigmatikoa/kontzeptuak/zabala/izenak/emaitza onak

Mahesh-ek eta (Mahesh et al., 1996; 1997) Mikrokosmos ontologia abiapuntu bezala hartu eta adiera-desanbiguazio lan baterako *spreading activation* itsu mutuan aritzen dela diote: " … *spreading activation … does not make use of available knowledge*." Beraien planteamenduan esaldiaren analisi

[27] Hauek dira lehen aipatutako ezaugarrien balioak Radaren taldearen lanarentzat.

[28] Sussna-ren laneko w_r hemen azaltzen dena baino konplexuagoa da, baina berak aitortzen duen bezala "*the particular weights used* [w_r] *may not make that much difference*".

semantikoan ateratako argumentu-egitura errespetatu egin behar da adieren arteko bideak bilatzean. Beste modu batetara ikusita erlazio-izaerak aditz edo adjektiboaren hautapen-murrizpenei hoberen egokitzea neurtzen du, hautapen-murrizpen kontzeptualak ontologian errepresentatuaz, eta kontzeptuen arteko hurbiltasun paradigmatikoa ere erabiliaz. Mikrokosmos-en estaldura urria dela eta, ez dute aurkezten ebaluaziorik.

Ontologia/paradigmatikoa eta sintagmatiko lokala/kontzeptuak/proposamena/ izen-aditz/emaitzarik ez

III.A.2. Hiztegi elektronikoetan oinarritutako neurriak

Hiztegietan ez dago kontzepturik, adierak dira azaltzen direnak. Adiera horiek hala ere lexikografoak egindako kontzeptualizazioei erantzuten diete, eta hein handi batean ontologia bateko kontzeptuekin parekatu daitezke. Nola neurtu adiera horien arteko erlazio-izaera? Ontologia lanetan ez bezala, hemen ez dago psikologia edo ezagutzan oinarritutako formalizaziorik, praktikoak diren hurbilpenak baizik.

Erlazio-izaera motari buruz, hemen zeharkako erlazio sintagmatiko globalak erabiltzen direla esan dezakegu. Bi adiera lotuta dauden jakiteko adierak azaltzen diren testuingurua aztertzen da (hiztegiaren kasuan adieraren definizioa bera), eta testuinguru horiek antzekoak badira orduan adierak erlazionatuta egongo dira. Atzean dagoen hipotesia zera da, erlazionatutako adierak hitz antzekoez definitu izango direla.

Lesk-ek (1986) adiera-desanbiguaziorako aplikazioan zuzenean aplikatu zuen hipotesi hori: bi adieren arteko erlazio-izaeraren neurria beraien definizioetan agertzen diren amankomuneko hitzen kopurua da. Zenbat eta hitz gehiago egon bi definizioetan, hainbat eta estuago erlazionatuta egongo dira bi adierak. Bere intuizioa emankorra izan da orain ikusiko dugun bezala, baina bere horretan ahula da oso, definizioa idaztean hautatutako hitzen menpe baitago. Metodo honen ebaluazioa adiera desanbiguazio lan baten bidez egiten du. Cowie-ren taldeak (Cowie et al., 1992; Wilks et al., 1996) metodo bera proposatzen du, baina kontzeptu multzo zabalen arteko erlazio-izaera neurtzean eraginkortasuna hobetzeko *simulated annealing* delakoa erabiliaz.

Hiztegi/sintagmatiko globala/kontzeptuak/zabala/ izenak/emaitza kaxkarrak

Véronis eta Ide-k (1990) hurbilpen berari heltzen diote, baina hedatu egiten dute definizio zirkular bat erabiliz. Bi adieren arteko erlazioaren neurria definitzeko erabili diren hitzen arteko erlazioaren neurrien baturak emango du. Hau da, orain ez da beharrezkoa hitz berak agertzea bi adieren definizioan, nahikoa da erlazionatutako hitzak erabiltzea. Eta noiz daude bi hitz erlazionatuta? Beraien adierak erlazionatuta daudenean. Hipotesi hau eraginkorra den edo ez frogatzeko sare neuronal erraldoi bat eraiki zuten hiztegiko definizioetako hitzak erabiliz, definiendum eta

definizioko hitzen arteko loturak gehituz[29], eta adiera-desanbiguazio lan batean probatu zuten (ebaluazio sistematikorik ez dute). Hurbilpen antzekoa darabilte Kozima eta Furugorik ere (1993), baina eraginkortasun-arazoak konpontzeko kasu honetan bektore-eredu batera itzultzen dute informazioa (Kozima & Ito, 1995), orain ikusiko dugun eredura (ikusi (Niwa & Nitta, 1994) ere). Hauek ebaluazioa pertsonen iritziz eraikitako antzekotasun-zerrendekin parekatuaz egiten dute.

Hiztegi/sintagmatiko globala/hitzak/gutxi/izenak/emaitzarik ez

Lesk-en metodoak beste hedapen bat jaso zuen, hiztegietako definizioetako agerkidetzez osatutako bektore-ereduekin. Wilks-ek eta (Wilks et al., 1990; 1996) LDOCE hiztegitik (ikus II.A.2 atala) hitzen agerkidetzak jaso zituzten. LDOCE-k murriztutako hiztegi bat (2781 hitz dituena) erabiltzen du definizioak idazteko, eta beraz agerkidetzak murriztutako hitz horietara mugatzen dira. Lan horien arabera bi hitzen arteko agerkidetza definizio berean azaltzen direnean ematen da. Hitz bakoitzaren agerkidetzak kodetzeko bektore bat erabiltzen dute (4. formulako $\vec{v}_w$). Bektore horretan hiztegi murriztuko hitz bakoitzeko (4. ekuazioko N da hiztegi murriztu horren tamaina) balio bat egongo da (v_i^w), hitzen arteko agerkidetzaren indarra errepresentatzen duena. Horretarako sei formula ezberdin proposatzen dituzte, denak hitzen eta agerkidetzaren maiztasunetan oinarritutakoak. 5. ekuazioan, adibidez, bektoreko balio bezala agerkidetzaren maiztasun gordinak azaltzen dira, hau da, w eta z_i hitzak elkarrekin agertzen direneko maiztasuna.

$$\vec{v}_w = \left(v_0^w, \cdots, v_N^w\right) \tag{4}$$

$$v_i^w = f_{w,z_i} \tag{5}$$

Bi hitzen arteko erlazio-izaera kalkulatzeko bektore horien arteko erlazioa matematikoki kalkulatu daiteke, adibidez angeluaren kosinua erabiliaz (ikusi 6. ekuazioa, baina beste hiru formula ere proposatzen dituzte). Wilks eta harantzago doaz ordea, eta adieren arteko erlazio-izaeraren neurria eduki ahal izateko, hiztegiko adiera bakoitzarentzat bektore bat eratzen dute bere definizioan agertzen diren hitzen bektoreak batuaz (7. ekuazioa). Horrela bi bektoreren arteko erlazioaren neurri matematikoak bi adieren arteko erlazio-izaera ematen digu (nahikoa da 6. ekuazioan w eta z hitzak baino adierak izatea, 7. ekuazioko bektorez errepresentatuak).

[29] Definizioak ez zeuden lematizatuta, ezta analizatuta ere.

$$\mathrm{sim}(w,z) = \cos(\vec{v}_w, \vec{v}_z) = \frac{\sum_{k=1}^{N}(v_k^a v_k^b)}{\sqrt{\sum_{k=1}^{N} v_k^a \sum_{k=1}^{N} v_k^b}} \quad (6)$$

$$\vec{v}_a = \sum_{w \in \mathrm{def}(a)} \vec{v}_w \quad (7)$$

Metodo honek definizioetako hitzen gainjartzea zuzenean neurtu ordez, hitz horien bektoreak erabiltzen ditu beraz. Ebaluazioa ez da oso sakona, *bank* hitzaren agerpen batzuk soilik desanbiguatuaz egin zuten eta.

Hiztegi/sintagmatiko globala/kontzeptuak/gutxi/ izenak/emaitza onak

Richardson-ek (1997) hartzen duen hurbilpenak ontologiakoentzat ikusitakoen antza dauka. Izan ere bi hiztegitako (*LDOCE* eta *W7*, ikus II.A.2 atala) definizioak sintaktikoki analizatu eta erlazio semantikoak erauzten ditu, sare semantiko bat eraikiaz. Erlazio bakoitzak maiztasunetan oinarritutako pisu bat du. Sare semantiko honetan definizioetako hitzak desanbiguatu gabe daudenez ezinezkoa zaio bi adieren arteko erlazio-izaera neurtzea. Horren ordez hitzen arteko erlazio-izaera lantzen du, baina bi hitz lotzen dituen bidean errorea egon daitekeenez (adiera ezberdinekoak balira tarteko hitzak) oso bide motzak erabiltzen ditu, gehienez bi definiziotako hitzak erabili ahal direlarik. Ideia beraz zera da, bi hitz hertsiki erlazionatuta egongo dira beraien artean erlazio-bide asko badaude. Erlazio-bide guztiak ez dira esanguratsuak ordea, eta erlazio-bide mota bakoitzaren erabilgarritasuna neurtzeko beharrean aurkitzen da. Hori egiteko metodo enpiriko bat erabiltzen du, thesaurus bateko 50.000 hitz pare eta erlaziorik ez duten beste 50.000 hitz pare erabiliz. Ebaluazioa thesaurus hori bera erabiliz egiten du, pisu horiek kalkulatzeko hitzak kontuan hartu gabe, noski.

Hiztegi/paradigmatikoa eta sintagmatiko globala/hitzak/zabal/ izenak/emaitza onak

III.A.3. Corpusetan oinarritutako alternatibak

Corpusen erabilera bultzatzen dutenak maiz aipatzen dute "*you shall know a word for the company it keeps*" (Firth, 1957), hau da, hitzen ezaugarri eta esanahia hitz hori azaltzen den testuinguruak emango du, edo hobeto esanda, azaldu izan den testuinguru guztien analisiak. Horretan oinarrituta, honako hipotesia zabaldu da: bi hitz hertsiki erlazionatuta egongo dira antzeko testuinguruetan azaltzen badira. Hitzen arteko erlazio-izaera aztertzeko, hitz horiek azaltzen diren testuinguruak konparatzea besterik ez da behar. Erlazio sintagmatiko global eta lokala kontuan hartuko diren edo ez testuinguruaren definizioaren arabera egongo da: erlazio sintagmatiko lokala aztertu nahi bada,

erlazio sintaktiko zuzena duten hitzak erabiliko dira. Erlazio sintagmatiko globalaren kasuan leiho zabalak definitzen dira, ±50 hitzetakoak, baina ordena kontuan hartu gabe eta izen, adjektibo eta aditzetaz soilik baliatuaz.

Corpusean oinarritzen diren teknikak adieren arteko erlazioetara hedatu ahal izateko corpuseko hitzak beraien adieraz etiketatu behar dira, entrenamendu-corpus bat eratuaz. Hau da hain zuzen ere corpusetan oinarritutako tekniken arazo bat, eskuzko desanbiguazio zabal baten beharra.

Neurri sinple eta arrakastatsuenetako bat informazioaren teorian oinarritutako Elkarren Arteko Informazioa da (EAI, *mutual information*, Church & Hanks, 1990; Gale et al. 1992; 1993). Horren arabera beti elkarrekin azaltzen badira bi hitzen arteko erlazio-izaera indartsua izango da, eta ahula, aldiz, ez direnean inoiz testuinguru berean azaltzen. Church eta Hanks-en arabera, testuingurua hitzen agerpenaren inguruan dauden 100 hitzetako leihoak erabiltzen dira. Horrela *v* eta *w* hitzen EAI kalkulatzeko hitz bakoitza agertzeko eta biak batera agertzeko probabilitateak eduki behar dira kontuan (8. ekuazioa).

$$\mathrm{EAI}(v, w) = \log \frac{\Pr(v, w)}{\Pr(v)\Pr(w)} \tag{8}$$

Probabilitate horiek estimatzeko modu errazena agerpenak kontatu (ikusi *f* 9. ekuazioan) eta *N* hitz kopuru totalaz zatitzea da (aukera gehienezko estimazioa – *maximum likelihood estimate* – deritzon teknika).

$$\Pr(x) \cong \frac{f_x}{N} \tag{9}$$

Corpus/sintagmatiko globala/hitzak[30] */gutxi/ izenak/emaitza onak/datu urrien arazoa*[31]

EAI aplikazio askotan erabili izan da, eta beraren inguruko literaturan gehien aipatzen den arazoa estimazioarena da. Gainontzeko teknika estatistikoek ere arazo honi aurre egin beharko diote, izan ere, hitz gutxi batzuk oso maiz azaltzen dira testuetan, baina hitz gehienak oso-oso urritan (Zipf-en legearen arabera). Horregatik deitzen zaio arazo honi datu urrien arazoa (*sparse data problem*). Zein da inoiz ikusi gabeko hitz bat azaltzeko probabilitatea? Edo corpus osoan birritan azaldu diren bi hitz

[30] Hitzen arteko erlazio-izaera bezala jartzen dugu, adieratara hedatzea ez delako naturala, eskuzko etiketatzea beharko lukeelako.

[31] Corpusetan oinarritutako alternatibetan garrantzitsua denez, datu urrien arazoei buruzko iruzkinak gehitu dizkiegu beste ezaugarrieri.

batera aurkitzeko probabilitatea elkarrekin azaldu ez direnean? Garbi dago ez dela 0. Honi aurre egiteko teknikei leuntze-teknika (*smoothing*) deritze.

Schütze-ek (1992a; 1992b; 1998) hitzen arteko gertakidetzak kontuan hartzeko beste modu bat erabili zuen, bektore bezala kodetu eta bektoreen arteko angelua erabili hitzen gertutasuna neurtzeko (ikus Wilks-en metodoa aurreko III.A.2. atalean). Adieren arteko erlazio-izaerara hedatu ahal izateko, hitz baten testuinguruak multzokatu egiten ditu, horretarako testuinguruko bektore guztiak batu eta multzokatze teknika aplikatuz. Horrela hitz jakin baten agerpenak giza-aditu batek adieren arabera sailkatu ditzake, testuinguru jakin batzuei adiera bat esleituz. Corpus osoko hitzaren agerpenak banan-banan etiketatzea baino ahalegin gutxiago beharko litzateke, autorearen arabera.

Corpus/sintagmatiko globala eta lokala/hitzak/zabal/izenak/emaitza onak/datu urrien arazorik ez

EAIak eta Schütze-ren bektoreek testuingurutik hitzen agerpena besterik ez dute kontuan hartzen, baina bada gehiagorik, egitura sintaktikoa esate baterako. Era sinpleenean hitzaren aldamenean dauden hitzen kategoriak hartu daitezke kontuan, baina egitura landuagoan argumentu egiturak (aditz-objektu, izen-adjektibo, etab.) ere erabili daitezke. Halakoei ezaugarri deitzen zaie, eta beraz hitz baten testuinguru sintaktikoa corpusetatik erautsitako ezaugarriez (hau da, hitzaren agerpenetan bere inguruan azaltzen diren kategoria zein argumentu egiturak) errepresentatzen da. Horrelako ezaugarriak adiera desanbiguaziorako erabiltzen dira zuzenean (ikusi IV.A.4 atala), baina erlazio-izaera formalizatzeko zeharka erabili behar da: testuinguru beretsuetan azaltzen diren hitzak erlazionatuta daude. Hala egiten du Grefenstette-ek (1992; 1996) hitzen arteko erlazio-izaeraren neurria halako pista sintaktikoen gainean definitzerakoan. Ebaluazio interesgarria egiten du, Richardson-en antzekoa, thesaurus-ak erabiliz erlazio-izaeraren estandar bezala.

Corpus/sintagmatiko globala/hitzak/gutxi/ izenak/emaitza onak/datu urrien arazoa

Lan batzuetan zuzenean jotzen da erlazio sintaktiko berezi baten azterketara. Hala da aditzaren hautapen-murrizpenaren inguruan egiten diren azterketa gehienetan, adibidez Grishman eta Sterling-en lanean (1994) edo Lee-ren tesian (1997). Hauek corpus zabaletatik aditz-objektu pareak atera eta aditz bakoitzak hobesten duen izenen klasea topatzen saiatzen dira, aditz eta izenen arteko erlazio-izaera definituaz.

Corpus/sintagmatiko globala/hitz/gutxi/izenak/emaitza onak/datu urrien arazoa

III.A.4. Ontologia eta corpusen arteko konbinazioak

Hainbat lanek aurreko hurbilpenak hedatu beharra dagoela diote. Arrazoi anitz aipatzen dute. Nagusiena, goiko teknika guztiek hiztegia egitura semantiko gabeko zerrenda bezala tratatzen dutela da. Hitzak eta kontzeptuak klaseetan eratu ohi dira, eta hitz baten ezaugarri semantiko asko bere

klasearenak dira. Horretara, zertarako gorde informazioa hitzez-hitz, zati handi bat klasearen ezaugarria baldin bada? Bestalde, hurbilpen estatistikoaren aldekoen buruhauste nagusiak (**datu urrien arazoa** eta **eskuzko desanbiguazioaren beharra**) gutxitu litezke hitzak sailkatuta edukiz gero. Adibidez, jan aditzaren objektu tipikoak adierazteko hobe da *gauza-jangarri* klasea erabiltzea, banan-banan *otarteko*, *urdaiazpiko*, *legatz*, *sagar*, etab. zerrendatzea baino. Gainera, nahiz eta kiwi ez azaldu inoiz corpusean *jan*-en objektu bezala, *gauza-jangarri* bezala sailkatuta badago gai izango gara *kiwi* eta *jan*-en arteko erlazioa asmatzeko.

Lan batzuek klase horiek corpusetik bertatik erauzten dituzte (adibidez lehenago aipatutako Schütze-ren lanak) baina horrek askotan errore-zama bat sartzen du. Horren aurrean askok thesaurus edo ontologietara jotzea proposatzen dute, klaseen definizio intuitibo eta zuzenen bila. Yarowsky-k (1992), adibidez, kategoria bezala Roget thesaurusak emandakoak erabiltzen ditu, adiera-desanbiguazio lan batean. Roget thesaurusean (ikus II.A.2 atala) klase bakoitzeko hitz zerrenda bat dator. Klase bakoitza agertzen den testuinguru tipikoak zeintzuk diren jakiteko klaseko hitzak agertzen diren 100 hitzetako leihoak jasotzen ditu Grolier entziklopediatik. Testuinguru horien arabera, kategoria bakoitzerako hitz adierazgarrienak[32] biltzen ditu, nabarmentasun (*saliency*) izeneko neurri estatistikoaren arabera (ikus 10. ekuazioa).

$$\text{saliency}(w) = \log \frac{\Pr(w \mid c)}{\Pr(w)} \qquad (10)$$

Lan honetan ez da esplizituki azaltzen hitzen arteko erlazio-izaera, hitzak Roget-eko klaseekin etiketatzeko metodoa baizik. Hala ere, corpusen inguruko beste lanetan bezala, neurri hauetatik posible liteke hitz edo adieren arteko erlazio-izaera kalkulatzea. Basili-k eta (Basili et al. 1995; 1997; Cucchiarelli & Velardi 1997) ere antzeko neurria erabiltzen dute ontologietarako informazioa erauzi nahian.

Ontologia+corpus/sintagmatiko globala/kontzeptuak/zabal/izenak/emaitza oso onak/datu urrien arazorik ez

Resnik-ek (1993a; 1993b; 1995; 1997), aldiz, beste era batera konbinatzen du corpuseko eta ontologiako informazioa. Bi adieren arteko erlazio-izaera neurtzeko ontologian amankomunean duten arbaso hurbilena bilatzen du, baina distantzia neurtu ordez, klase horren informazio-edukia (*information content*) kalkulatzen du (ikusi 11 formula, non v eta w izenak diren, eta c izen horiek biltzen dituen klase txikiena).

[32] Yarowsky-ren hitzetan "*words that are likely to co-occur with the members of the category*".

$$\text{antzekotasuna}(v, w) = -\log \Pr(c) \tag{11}$$

Klasearen probabilitatea klase horren kide diren hitzek corpusean duten maiztasunetatik estimatu daiteke:

$$\Pr(c) \cong \frac{\sum_{w \in c} f_w}{N} \tag{12}$$

Neurri hauek aditz eta izenen adieren arteko erlazioaren indarra neurtzeko soilik erabili izan ditu, aditzen hautapen-murrizpenak eskuratzeko bidean. Ebaluazio zeharkakoa egin zuen, izenen adiera-desanbiguazioan eta aditzen hautapen-murrizpenen bilaketan. Li eta Abe-k (1995; 1996) ere hurbilpen hau erabiliko dute hautapen-murrizpenen indukzioan eta izenen multzokatze automatikoan.

Ontologia+corpus/paradigmatikoa/kontzeptuak/gutxi/izenak/emaitza onak/datu urrien arazorik ez

Schütze-ren aurreko lanean (1992a; 1992b) aurkeztu zen hitzen arteko erlazio-izaera WordNet-eko informazio hierarkikoarekin konbinatzen saiatzen dira Hearst eta Schütze (1993). Helburu bezala hierarkiaren bidez erlaziorik ez duten kontzeptuak erlazionatzea jartzen dute, adibidez pilota eta frontoia, baina azkenean egiten dutena zera da, WordNet-eko izenen sysnset guztiak 726 kategoriatan banatu eta horien arteko erlazioak landu. Emaitzen ebaluazio sistematikorik ez dago, eta autoreek beraiek aitortzen dute erlazio gutxi lotu dituztela. Lortutako sare kontzeptual berrirako ez du erlazio-izaera neurri berririk ematen.

Ontologia+corpus/sintagmatiko globala eta lokala/kontzeptuak/gutxi/izenak/emaitza onak/datu urrien arazorik ez

Ontologiak baino hiztegiak erabiltzen dituzte Karov eta Edelmann-ek (1996; 1998) hitz baten adiera bati lotuta dauden testuinguruak (esaldiak kasu honetan) lortzeko. Erlazio-izaeran zirkularitate bat dagoela iruditzen zaie: hitzak erlazionatuta daude esaldi beretsuetan azaltzen badira, eta esaldiak erlazionatuta egongo dira erlazionatutako hitzak badituzte. Zirkularitate hori puskatzeko algoritmo iteratibo bat erabiltzen dute, euren esanetan konbergentziara heltzen dena. Beraien hurbilpenaren abantaila bat datu gutxiagorekin entrenatzeko gai direla izango litzateke.

Hiztegia+corpus/sintagmatiko globala/kontzeptuak/gutxi/izenak/emaitza onak/datu urrien arazorik ez

III.B. Dentsitate Kontzeptuala

Ontologian oinarritutako erlazio-izaera formalizatzeko gure proposamena aurkeztuko dugu atal honetan. Formalizazio horrek honako baldintza hauek edukitzea nahi dugu:

1. Ontologiatan oinarritutakoa.
2. Adieren arteko neurria: ontologiako kontzeptuei erreferentzia egingo diena
3. Erlazio paradigmatiko eta sintagmatikoetako informazioa erabiliko duena
4. Kategoria irekietako[33] hitzekin lan egingo duena
5. Eraginkorra izatea, testu zabalekin lan egin ahal izateko bezalakoa.

Lehenbiziko bi baldintzak lotuta daude, ontologia erabiltzen denean kontzeptuen arteko erlazioak berez landuta daude eta. Ontologian erlazio paradigmatiko eta sintagmatikoek egon beharko dute, erlazio-izaera zenbaiterainokoa den erabakitzean ahal den informazio gehien izan dezagun. Ontologia erabiltzearen beste abantaila informaziorik ikasteko beharrik ez dagoela da, hau da, ez da beharrezkoa aurrez eskuz ezer desanbiguatzea. Azkenik izen, adjektibo eta aditzekin lan egiteko balio behar du, eta testu errealekin lan egitea nahi dugu, ez ordea dozena eskas hitz konkreturekin.

Baldintza horiek betetzen saiatuko diren bi formula aurkeztuko ditugu. Lehenbizi 2 kontzepturen arteko neurria ematen duena, eta ondoren edozein kontzeptu multzorako neurria.

III.B.1. Bi kontzepturen artekoa: Distantzia

Rada (Rada et al., 1989) eta bereziki Sussna-ren (1993) lana hartu dugu abiapuntu bezala. Lan horien arabera erlazio-izaera ontologiako kontzeptuen arteko Distantzia Kontzeptualaren[34] bidez kalkula daiteke[35]. Sussna-k egindako ikerketaren arabera bi faktorek daukate zerikusia Distantzia Kontzeptuala kalkulatzean: bi kontzeptuen arteko erlazio-bidearen luzera (bide luzeagoa den heinean distantzia handiagoa) eta bide horretan dauden kontzeptuen sakonera (sakonean dauden kontzeptuen artean distantzia txikia). Horren arabera ondoko formula proposatu genuen (Agirre et al., 1994b):

[33] Horrela izendatu ohi dira izen, aditz eta adjektiboak.

[34] Kapitulu honen hasieran aipatu dugu distantzia semantikoa, erlazio-izaera bera definitzen ari ginenean. Distantzia semantikoa formalizatu gabe egonik, guk ontologia batera lotu dugu, eta horregatik deitzen diogu distantzia kontzeptuala.

[35] Erlazio-izaera eta Distantzia Kontzeptuala alderantzizkoak dira: hertsiki erlazionatuta dauden kontzeptuen artean distantzia kotzeptuala 0ren hurrena da, eta erlaziorik ez duten bi kontzepturen arteko distantzia kontzeptuala ∞-rantz hurbiltzen da.

$$\text{Dist}(a,b) = \min_{p \in \text{bide}(a,b)} \sum_{c_i \in p} \frac{1}{\text{sakonera}(c_i)}$$
$$\text{non } a = c_0 \text{ eta } b = c_n \qquad (13)$$

Bi kontzepturen (13. ekuazioko *a* eta *b*) arteko Distantzia Kontzeptuala bide (*p*) motzenak emango digu, luzera modu berezi batean kalkulatuz gero: bideko kontzeptu bakoitzarengatik hierarkian duen sakoneraren alderantzizkoa gehituko dugu. Honek islatzen duena zera da, zenbat eta gertuago eta sakonago egon kontzeptuak ontologian, orduan eta Distantzia Kontzeptual txikiagoa egongo da bien artean (Agirre et al. 1994b).

III.B.2. N kontzepturen artekoa: Dentsitatea

Distantzia hau baliagarria da bere horretan aplikazio askotan, baina bi kontzepturen distantzia N kontzeptutara orokortu nahi badugu leherketa konbinatorio bat sortzen da. Binakako distantzia erabiliz N kontzepturen arteko distantzia neurtzeko modua pare posible guztien distantziak batzea da (Sussna, 1993). Zortzi kontzepturen arteko distantzia kalkulatzeko, adibidez, zortziren binakako konbinazio guztiak, 28, eduki behar dira kontutan[36]. Esaldi bateko hitzen arteko distantzia neurtu nahiko bagenu gauzak okertu egiten dira, hitzen anbiguetatea dela medio. Demagun esaldiak 8 hitz dauzkala, eta bakoitzak 3 adiera, adiera guztien arteko binakako distantziak kalkulatu behar izanez gero hitzen arteko binakako pare guztiak (28 berriz ere) adiera konbinazio guztientzat (3^2) probatu beharko dira: guztira 252. Orokorrean N hitz badaude, batez beste M adiera dituztenak $\binom{N}{2} \times M^2 = \frac{N \times (N-1)}{2} \times M^2$ aldiz neurtu beharko dugu bi kontzepturen arteko distantzia.

Bestalde, kontzeptu multzoen arteko konparazioak zaildu egiten dira. Demagun A eta B multzo bakoitzean bi kontzeptu dugula. Horrela posible da esatea A-ko bi kontzeptuak B-ko biak baino elkarrengandik gertuago daudela. Pare ezberdinen arteko distantziak konpara daitezke. Baina, A multzoari beste kontzeptu bat gehituz gero distantzia handitu egingo da, eta ezinezkoa da A berri honen distantzia B-renarekin alderatzea, kontzeptu kopuru ezberdinaren distantzia neurtzen ari garelako.

Hori dela eta, lehen aipatutakoez gain, beste baldintza pare bat gehituko diogu gure neurriari:

[36] $\binom{8}{2} = \frac{8 \times 7}{2} = 28$

6. N kontzepturen arteko neurria izatea
7. Kontzeptu kopuru ezberdineko multzoen gertutasunak konparagarriak izatea.

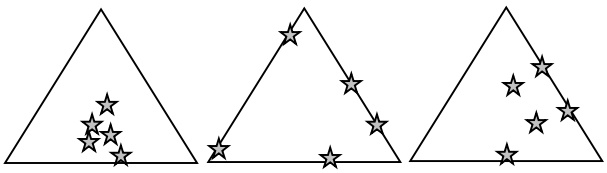
Lehenbiziko baldintzara bueltatuz gero, neurria errealitatean aplikatzeko ontologia bat aukeratu beharra dago. Tamalez, gaur egun, ontologia zabal eta libre eskuragarri gutxi daude, hiztegi aldetik zabala dena bakarra: WordNet (ikusi honi buruzko eztabaida II.C atalean). WordNet-en ezaugarri batek eragina edukiko du ezarritako beste baldintza batetan, erlazio paradigmatikoa baita gehienbat landua dagoena. WordNet aukeratzeak eragin digu hirugarren baldintza, gogoz kontra, murriztu behar izatera:

1. WordNet ontologian oinarritutakoa
3. Erlazio paradigmatikoetako informazioa erabiltzen duena

Autore batzuen ustez, erlazio paradigmatikora mugatzea ez da hain murrizpen gogorra: "*we hypohesize that … is strong enough for the length of is-a paths to be used as a measure of semantic relatedness*" (Rada et al., 1989). Erlazio hierarkikoak soilik erabiltzeak, gainera, efizientzia aldetik sekulako hobekuntza ekarriko digu, gero ikusiko dugun bezala.

N kontzepturen arteko neurria garatzea ez da hain gauza naturala. Orain arte nahiko garbi zegoen bi kontzepturen arteko bidearen luzera dela gure formulazioaren muina, eta sakonera ere kontuan hartu beharra dagoela. N kontzepturen arteko neurriak ordea beste jite bat hartu behar du, eta distantzia baino dentsitatea izango da kontuan hartu beharrekoa: bideen luzera baino azpizuhaitzetan dauzkagun kontzeptuen kopuruak. Harira joan aurretik, aurrerantzean nahasteak saihesteko terminologia kontua: kontzeptu-multzo batean erlazio-izaera neurtu nahi dugunean, multzoko kontzeptuei **arrasto** deituko diegu, azpizuhaitzeko beste kontzeptuekin ez nahasteko.

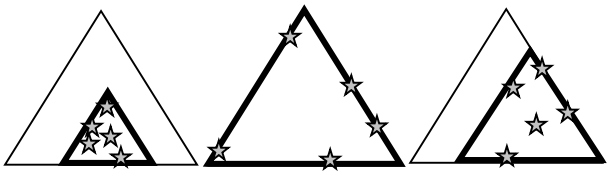
Neurri honen muina galdera honetatik dator: zenbat arrasto behar dira ontologiako azpizuhaitz batean, azpizuhaitz hori arrastoz asea edo betea dagoela esateko? Edo beste modu batera esanda, bi azpizuhaitz konparatzean nola neurtu dezakegu zein den beteago dagoena?



1. irudia: azpizuhaitz bera hiru arrasto multzo ezberdinekin.

1. irudian ontologiaren zati bera hiru arrasto multzo ezberdinekin azaltzen da. Hiru kasuetan arrastoen arteko gertutasuna berdina dela esango genuke? Ez. Badirudi erlazio-izaera handia izan beharko litzatekeela ezkerrekoarentzat, txikia erdikoarentzat eta tartekoa eskuinekoarentzat (edo dentsitateari buruz bagabiltza, dentsitate handiena ezkerrekoak eta txikiena erdikoak). Binakako Distantzia Kontzeptuala erabiliko bagenu emaitza bera jasoko genuke, hau da, ezkerreko arrastoen artean bide motzak daude, eta erdikoaren artean bide luzeak.

Bideak alde batera utziz, hiruen arteko ezberdintasun bat zera da, zein den 5 arrastoak estaltzen dituen azpizuhaitz minimoa, 2. irudian azaltzen den bezala.



2. irudia: arrasto multzoak estaltzen dituzten azpizuhaitz minimoak (marra lodiagoz).

Azpizuhaitz horiek kontuan hartuz nahiko garbi azaltzen da arrastoen arteko erlazio-izaerak azpizuhaitz minimoaren arteko tamainarekin[37] erlazio zuzena duela: Dentsitate handienekoak tamaina txikiena du (ezkerrekoak), eta dentsitate gutxienekoak tamaina handiena (eskuinekoak). Hemendik soma daiteke Dentsitate deituko dugun hori arrasto kopuruaren eta azpizuhaitz minimo horren tamainaren arteko erlazioa dela. Lehenbiziko hurbilpen batean, adibidez, *a* arrasto estaltzen dituen *Z* azpizuhaitzaren Dentsitatearen neurrirako 14. ekuazioa dugu, hau da, arrasto kopurua (*a*) zati zuhaitzaren tamaina (zuhaitzaren azalera ere deituko duguna).

$$\text{dentsitate}(Z,a)=\frac{a}{\text{azalera}(Z)} \tag{14}$$

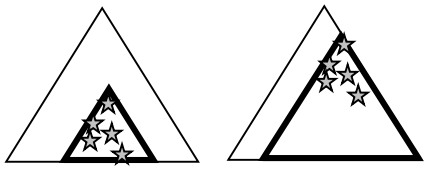
Eta zein izango da arrasto multzo baten Dentsitatea? Arrasto multzoa (demagun A dela) estaltzen duen azpizuhaitz minimoaren Dentsitatea, edo beste era batera esanda A multzoko arrastoak estaltzen dituzten azpizuhaitz guztietatik, Dentsitate maximoa lortzen duenaren Dentsitatea, 15. ekuazioan[38] azaltzen den bezala.

[37] Honako hirurak gauza bera adierazten dute: azpizuhaitz baten adabegi kopurua, tamaina eta azalera.

[38] Z azpizuhaitzak A estaltzen duela adierazteko $A\cap Z=A$ erabiltzen da, eta A multzoan dagoen arrasto kopurua adierazteko bere kardinala $|A|$.

$$\text{dentsitate}(A) = \max_{Z,\ \text{non}\ Z \cap A = A} \text{dentsitate}(Z, |A|) \qquad (15)$$

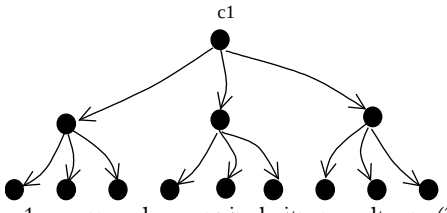
14. ekuaziora itzuliz, Distantzia Kontzeptualaren ezaugarri nagusiak biltzen ditu: gertutasuna eta sakonera. Zenbat eta gertuago egon, orduan eta txikiagoa izango baita arrastoak estaltzen dituen azpizuhaitz minimoaren azalera. Sakonerarekin beste hainbeste: arrastoak sakonago egonda azpizuhaitz minimoa txikiagoa izango da eta. Esandakoaren adibideak aurki daitezke 2. eta 3. irudietan: bietan Dentsitate handieneko arrasto multzoak ezkerrekoak dira.



3. irudia: arrasto multzoak estaltzen dituzten azpizuhaitz minimoak (marra lodiagoz).

14. ekuazioko neurri honek, ordea, arazo asko ditu. Hauek aztertu aurretik zuhaitzen topologiari buruzko neurri batzuk eta beraien arteko erlazioa definituko ditugu: azpizuhaitzaren altuera (h_Z), zuhaitzeko kontzeptuek batez beste duten ume kopurua (μ_Z, adarkatze faktorea ere deitua – *branching factor*), eta azpizuhaitzaren azalera, azpizuhaitzak dituen kontzeptu kopuruak ematen duena. Hiru neurri hauen arteko erlazioa 16. ekuazioak jasotzen du. Neurri hauen adibidea 4. irudian azaltzen den azpizuhaitz erregularrak ematen digu.

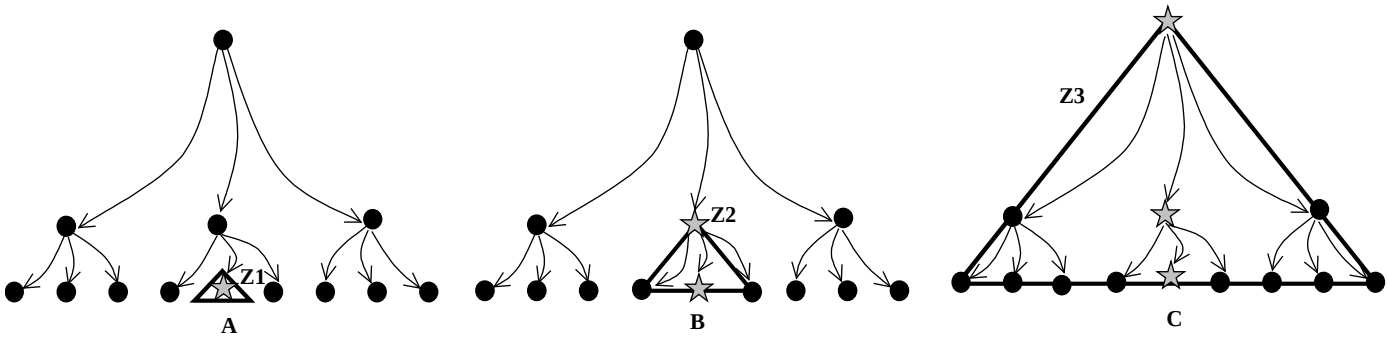
$$\text{azalera}(Z) = \text{kontzeptu\_kop}(Z) = \sum_{i=0}^{h_Z - 1} (\mu_Z)^i \qquad (16)$$



4. irudia: c1-en erroa duen azpizuhaitzaren altuera (3 maila), batezbesteko ume kopurua (3), eta azalera edo kontzeptu kopurua ($13=3^0+3^1+3^2$).

14. ekuazioak dituen arazoak 7. ezaugarritik datoz, hau da, kontzeptu kopuru ezberdineko multzoen arteko gertutasunak konparatu nahi izateagatik. Aztertu ahal izateko, demagun hiru kontzeptu

multzoren Dentsitatea neurtu nahi dugula (A, B eta C multzoak): batek arrasto bakarra, besteak bi eta azkenak hiru dituena, 5. irudian azaltzen den bezala. Arrasto multzo bakoitza estaltzen duen azpizuhaitza hiruki bezala marraztua dago.



5. irudia: hiru arrasto multzo azpizuhaitz berean. Kontzeptuak ● bidez adierazita daude, eta arrastoak ☆.

Intuitiboki zer esango genuke? B multzoko kontzeptuak C multzokoak baino estuago daudela erlazionatuta? Edo bi multzoek erlazio-izaera neurri berdina beharko luketela? Guk formalizatu nahi dugun erlazio-izaerarentzat garbi dago B eta C multzoko kontzeptuen artean gertutasun berdina dagoela. 14. ekuazioak, aldiz, bestela esaten digu:

dentsitate(A) = dentsitate(Z1,1) = 1/1 = 1

dentsitate(B) = dentsitate(Z2,2) = 2/4 = 0,5

dentsitate(C) = dentsitate(Z3,3) = 3/13 = 0,23

Gure ustez hiru arrasto multzo horien Dentsitatea 1 izan beharko litzateke, eta horretarako arrastoak kontatu baino, bestelako erreferentzia bat behar dugu: azalera eta arrastoen kopuruen arteko erlazioa ez da nahiko, altuera ere hartu beharko dugu kontuan. Adibidez, 5. irudian Z1 azpizuhaitzaren altuera 1 da eta arrasto bat du, Z2-ren altuera 2 da eta 2 arrasto ditu, eta Z3-ren altuera 3 izanda 3 arrasto dauzka, hiru kasuetan batezbesteko ume kopuruak berdinak direlarik.

Beste modu batera ikusita, nolako pisua eman beharko litzaioke arrasto bakoitzari 5. irudiko arrasto multzoen Dentsitatea 1 izan zedin? Galdera honi erantzun aurretik, idatz dezagun 14. ekuazioa beste modu batera, azaleraren ordez 16. ekuazioko formula jarriko dugu (ikus 17. ekuazioa), zatikizuna arrasto kopuruaren funtzio ezezagun bezala utziz –f(*a*).

$$\text{dentsitate}(Z,a) = \frac{\text{f}(a)}{\text{azalera}(Z)} = \frac{\text{f}(a)}{\sum_{i=0}^{h_Z - 1} (\mu_Z)^i} \tag{17}$$

Demagun 3 zuhaitz horientzat Dentsitate bera lortu nahi dugula, eta gainera horien Dentsitatea 1 izatea nahi dugula. Bilatzen dugun erlazioa altuera eta arrasto kopuruaren artekoa izan behar denez, eta altuera zatitzailearen batukarian azaltzen denez, 17. ekuazioaren zatikizunean zatitzailearen formula bera jarriko dugu, baina altuera dagoen lekuan arrasto kopurua jarriaz (18. ekuazioa)

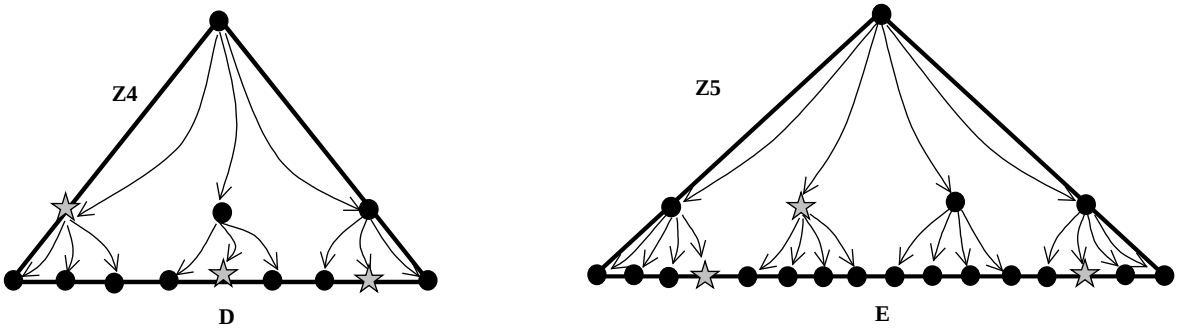
$$\text{dentsitate}(Z,a) = \frac{\sum_{i=0}^{a-1} (\mu_Z)^i}{\sum_{i=0}^{h_Z - 1} (\mu_Z)^i} = \frac{\sum_{i=0}^{a-1} (\mu_Z)^i}{\text{azalera}(Z)} \tag{18}$$

18. ekuazioko zatitzaileak azpizuhaitzaren azalera adierazten du, eta zatikizunak *a* arrastoko eta batezbesteko ume kopuru bereko zuhaitz erregularrak Dentsitatea 1 izateko eduki beharko lukeen azalera. Beste era batera esanda, zatikizunak 1 Dentsitatea eta μ_Z batezbesteko ume kopurua dituen zuhaitz erregularra errepresentatzen du, altuera eta arrasto kopurua berdinak dituena. Horrela islatzen da arrasto eta azpizuhaitzaren azaleraren arteko erlazioa.

14. ekuazioak bazeukan beste arazo bat arrasto multzo ezberdinen artean konparatzean, topologiarekin zerikusia duena. Ezaguna da ontologiaren zatiek topologia ezberdina eduki ohi dutela; alderdi batzuk kontzeptuz aberatsak direla, eta beste batzuk pobreagoak. Alderdi aberatsetan batezbesteko ume kopurua handia izango da, eta alderdi pobreetan txikia. Demagun bi kontzeptu multzo ditugula, biak hiru arrastokoak (6. irudiko D eta E), baina ontologiaren eremu ezberdinetan daudenak. Halakoetan, distantzia berera daude D-ko kontzeptuak eta E-ko kontzeptuak, baina 14. ekuazioaren arabera Dentsitate ezberdina izango dute:

dentsitate(D) = dentsitate(Z4,3) = 3/13 = 0,23

dentsitate(E) = dentsitate(Z5,3) = 3/21 = 0,14



6. irudia: Dentsitatea 1 duten neurri ezberdineko bi azpizuhaitz

18. ekuazioa erabiliaz, aldiz, orain arte erakutsi ditugun kontzeptu multzo guztietan Dentsitatea 1 da, guk nahi genuen bezala[39]:

dentsitate(A)=dentsitate(Z1,1) = 1/1 = 1

dentsitate(B)=dentsitate(Z2,2) = (1+3)/4 = 1

dentsitate(C)=dentsitate(Z3,3) = (1+3+9)/13 = 1

dentsitate(D)=dentsitate(Z4,3) = (1+3+9)/13 = 1

dentsitate(E)=dentsitate(Z5,3) = (1+4+16)/21 = 1

Kontzeptu multzo baten dentsitate kontzeptuala, beraz, 15. eta 18. ekuazioen bidez definituko dugu tesi lan honetan.

III.C. Inplementazioa

Dentsitate Kontzeptuala WordNet-eko hiperonimia erlazioa erabiliaz inplementatu dugu. Distantzia Kontzeptuala bai WordNet eta bai LPPL Hiztegi-Ezagutza Baserako inplementatu dugu. Tesi-lan honetan Dentsitateari buruz arituko garenez, ez dugu azalduko Distantziaren inplementaziorik. Inplementazio bera azaldu aurretik, parametroei buruz arituko gara.

III.C.1. Dentsitate Kontzeptualaren aldaerak

Dentsitate Kontzeptuala inplementatzean parametro eta aldaera batzuk ikertzea interesgarria izan daitekeela ikusi dugu. Horien artean egokiena zein izango den aldez aurretik erabakitzea zaila denez, enpirikoki aplikazio batean lortutako emaitzen arabera egitea erabaki genuen. Aplikazioa hitzen adiera-desanbiguazioa da. Atal honetan parametro eta aldaerak aurkeztuko ditugu, eta esperimentuen emaitzen berri IV.C.2 atalean emango dugu.

[39] Gogoratu azpizuhaitz guztietarako μ_Z 3 dela, Z5-entzat ezik, honentzat μ_Z 4 baita.

III.C.1.a) *Parametroa:* α

Dentsitate Kontzeptualaren formulak arazo txiki bat dauka: azpizuhaitz baten azpian dagoen arrasto kopurua oso handia denean, 18. formulako zatikizuna gehiegi handitu daiteke. Izan ere, Dentsitatea 1 izan dadin altuera eta arrasto kopurua berdina izatea eskatzen dugu, baina erabaki hau erabat arbitrarioa da. Dentsitatea 1 izateko altuera eta arrasto kopuruaren arteko erlazioa aldatzeko, α parametroa gehitu genion formulari, enpirikoki aztertu eta balioa bilatu dioguna. Parametrodun formula 19. ekuazioan azaltzen zaigu.

$$\text{dentsitate}(Z,a) = \frac{\sum_{i=0}^{a-1} (\mu_Z)^{i^\alpha}}{\text{azalera}(Z)} \qquad (19)$$

III.C.1.b) *Nola kalkulatu* μ*:* μ_Z *eta* μ_{WN}

Dentsitate Kontzeptuala kalkulatzerakoan zuhaitzaren topologia (μ_Z batezbesteko ume kopuruaren bidez islatzen duguna) kontuan hartzea garrantzizkoa da. Egikaritzapen-garaian konputatzea garestia izan daiteke, eta ontologia hierarkikoa izanda komenigarriagoa dirudi aldez aurretik azpizuhaitz posible bakoitzarentzat konputatua edukitzea. Horrekin batera azpizuhaitz bakoitzaren azalera ere gorde daiteke. Dentsitatea kalkulatzean nahikoa litzateke azpizuhaitzari dagozkion μ_Z eta azalera taula batetik atzitzea.

Aurrerago ikusi dugu azpizuhaitz baten batezbesteko ume kopurua, azpizuhaitzaren altuera (h_Z) eta azalera (*azalera(Z)*, adabegi kopurua) erlazionatzen dituen ekuazioa (ikus 16. ekuazioa). 7. irudian azaltzen da altuera (H) eta azalera (A) emanda batezbesteko ume kopurua (μ) kalkulatzeko erabili dugun programazio linealeko pseudokodezko algoritmoa. Parametro bezala, emaitzari eskatzen zaion doitasuna (d) eman beharra dago.

```
Sarrera:        H altuera, A azalera
Irteera:        μ batezbesteko ume kopurua
Parametroa:     d doitasuna
Aurrebaldintza: A>H

   baldin 1 <= A < H
      orduan  μ := 1 - 1/a
      bestela μ := a^(1/n)
   ambaldin
   bigizta
      s := μ^n;
      e := (μ*(s-A) + A - 1)/(H*s - A);
      μ := μ - e;
   harik eta |e/μ| < d ambigizta
```

7. irudia: μ_Z konputatzeko algoritmoa

Bestalde, μ_Z lokala erabili ordez, WordNet ontologia osoarentzat kalkulatutako batezbesteko ume kopurua erabiliko bagenu (μ_{WN}), Dentsitateak okerrago egingo lukeela espero daiteke. Hau horrela den edo ez neurtzeko aipatutako esperimentuak egin ditugu, IV.C.2 atalean azalduko ditugunak.

III.C.1.c) WordNet-eko beste erlazioak: meronimia

Dentsitate Kontzeptualak hiperonimia besterik ez du erabiltzen. Hala ere WordNet-eko izenen artean badaude beste erlazio hierarkikoak, meronimikoak (ikus II.C.3 atala). Printzipioz, are eta erlazio mota gehiago hartu kontutan, are eta emaitza hobeagoak espero daitezke. Meronimia erlazioa erabiliaz emaitza hobeagoak lortu diren edo ez enpirikoki aztertu dugu (ikus III.C.1.c) atala). Dentsitate Kontzeptualaren formulari dagokionez, ez dugu ezer aldatu meronimia kontuan hartzeko. Azpizuhaitzen azalera kalkulatzean, edo adiera bat bestearen azpian dagoen erabakitzean, ez dugu bereiziko hiperonimia edo meronimia erlazio artean.

III.C.2. WordNet-en gaineko inplementazioa

WordNet-erako egindako inplementazioan erlazio hierarkikoak besterik erabiltzen ez ditugunez, horretaz baliatzen den algoritmo eraginkorra diseinatu dugu.

Dentsitatea neurtzerakoan, adiera multzo bat (AM) ematen digute, arrasto ere deitu ditugunak. Adiera horientzat WordNet-en azpimultzoa den hierarkia eraikitzen dugu, arrastoen hiperonimo kateak jarraituaz. Hierarkia horretan egongo dira kontuan hartu behar ditugun azpizuhaitz guztiak. Izan ere, azpian arrasto bat ez badu azpizuhaitz batek, horren Dentsitatea 0 izango da. 8. irudian azaltzen da hierarkia hori eraikitzen duen algoritmoa. Adiera (arrasto) multzo bat emanda, kontuan hartu beharreko azpizuhaitz guztiak dauzkan hierarkia (`H`) bueltatzen digu. `H` aldagaia egitura bat da: `hipo` eremuan hierarkiago adabegi bakoitzaren hiponimoa gordetzen da, `arrasto_kopurua` eremuan adabegi bakoitzaren azpian dagoen arrasto kopurua, eta `azpizuhaitzak` eremuan

adabegi guztien zerrenda. 8. irudiko algoritmoa sinplifikazio bat da, adiera bakoitzarentzat hiperonimo bakarra suposatzen baitu (zuhaitz egitura izango balitz bezala). Hori ez da beti horrela WordNet-en. Hori konpontzeko `eman_hiperonimo_katea` funtzioak kate bat baino gehiago itzuliko luke, zerrenda bat eduki beharko luke.

```
FUNTZIOA: Eraiki_hierarkia(AM)
   Sarrera:  AM arrasto multzoa
   Irteera:  H hierarkia

      bigizta A barne AM bakoitzeko
         hiper_katea := eman_hiperonimo_katea(A) ;
         hipo := A ;
         bigizta h barne hiper_katea bakoitzeko
            H.hipo[h] = hipo ;
            H.arrasto_kopurua[h] ++ ;
            hipo := h ;
            sartu(h,H.azpizuhaitzak) ;
         ambigizta
      ambigizta
      bueltatu(H)
```

8. irudia: Dentsitate Kontzeptuala neurtu behar den arrastoen hiperonimoekin hierarkia eraikitzea

19. ekuazioaren inplementazioa 9. irudian dago. Arrasto kopuru jakin bat duen azpizuhaitz baten Dentsitatea kalkulatzen du, α parametroaren arabera. Funtzioaren argumentuak azpizuhaitza bera eta horren azpian dagoen arrasto kopurua dira. Dentsitatea kalkulatu ahal izateko azpizuhaitz horren azalera (`Z.azalera`) eta batezbesteko hiponimo kopurua ($Z.\mu$) jakin behar ditu (aldez aurretik kalkulatu ditugunak, ikus III.C.1.b) atala).

```
FUNTZIOA:   DK(Z,A)
   Sarrera:    Z  azpizuhaitza
               A  arrasto kopurua
   Irteera:    DK dentsitate kontzeptuala
   Parametroa: α
   Datuak:     Z.azalera
               Z.μ

      d1 := 0
      i := 0
      bitartean i < A
         d1 := d1 + Z.μ ^ (i^α)
      ambitartean
      DK := d1/Z.azalera
      bueltatu(DK)
```

9. irudia:Dentsitate Kontzeptuala kalkulatzeko algoritmoa

Azkenik, edozein arrasto multzo baten Dentsitate Kontzeptuala jakiteko, 15. ekuazioa jarraituz, arrasto multzo hori estaltzen duten azpizuhaitzen artean Dentsitate Kontzeptual altuena zeinek

duen kalkulatu beharko dugu. 10. irudiko algoritmoak hori bera egiten du. Arrasto guztiak estaltzen dituzten azpizuhaitzetatik (`H.azpizuhaitzak`) Dentsitate altuenekoaren Dentsitatea itzultzen du.

```
FUNTZIOA:   DK(AM)
   Sarrera:    AM arrasto multzoa
   Irteera:    DK dentsitate kontzeptuala

      DK := 0 ;
      H := Eraiki_hierarkia(AM) ;
      bigizta Z barne H.azpizuhaitzak bakoitzerako
         d := DK(Z,H.arrasto_kopurua[Z]) ;
         baldin d > DK orduan DK := d ;
      ambigizta
      bueltatu(DK)
```

10. irudia: adiera multzo baten Dentsitatea

III.D. Ebaluazioa eta besteekiko alderaketa

Dentsitate Kontzeptuala eta Distantzia Kontzeptuala tesi honetan definituta bezala (15. eta 18. ekuazioak) ez ditugu zuzenean ebaluatu, hau da, ez ditugu erlazionatutako hitz multzoen zerrendekin probatu jakiteko ea ekuazioetako erlazio-izaeraren neurria eta giza-sena bat datozen, gorago aipatutako arrazoiengatik (ikusi III.A atalean ebaluazioari buruzko gogoeta). Ebaluazioa Dentsitatea erabili den aplikazio bakoitzaren arabera egingo da, beste sistemek lortutako emaitzekin alderatuaz (ikusi IV.D atala bereziki, baina baita ere VI eta I.A.1)

Atal honetan emaitzen ebaluazioa baino ezaugarrien alderaketa egingo dugu beste sistemekiko, helburua honako baieztapen hau arrazoitzea izanda:

> *Nahiz eta zeregin batzuetan emaitza onenak lortu ez , bai oinarri teorikoaren aldetik baita zeregin ezberdinetarako prestatuta egoteagatik ere, ontologian oinarritutako erlazio-izaeraren formalizazioak hobeak dira, eta ontologietan oinarritutakoen artean Dentsitatea orokorragoa, eraginkorragoa eta emaitza onenak dituena da.*

Baieztapen honetako bi oinarriak, ontologian oinarritutako tekniken nagusitasuna eta ontologian oinarritutakoen artean Dentsitatearen abantailak, aztertuko ditugu orain. Hurrengo kapituluan (IV) aplikazio konkretu baten lortutako emaitzetan oinarrituta alderatuko dugu Dentsitate Kontzeptuala beste lanekin.

III.D.1. Ontologiatan oinarritutako tekniken nagusitasunaren inguruan

Lehenbiziko aztergaia ontologian oinarritutakoen nagusitasuna izango da, beraz. Aurrekarien atalean ikusi bezala ontologian oinarritutako neurriek psikologia eta adimen artifizialean egindako

ikerketetan dute erroa, eta lan horiek dira erlazio-izaera berez aztertzen dituzten bakarrak, aplikazio konkretuetatik abstraituz.

Hiztegietako neurriak nahiko *ad hoc* dira. Corpusetarako teknika berak erabiltzen dira maiz (Wilks-enak kasu), baina badute corpusetako teknikek ez duten abantaila bat: hiztegietan kontzeptuak azaltzen dira, hitzaren adierak, eta hitz baten adieran azaltzen den informazioak adiera (kontzeptua) karakterizatzeko balio lezake. Hori da hain zuzen ere Lesk, Cowie, Véronis, Kozima eta Niwa-ren taldeen hurbilpenaren funtsa: adierei buruzko informazioa erabili erlazio-izaera formalizatzeko (Lesk, 1986; Cowie et al., 1992; Wilks et al., 1996; Véronis & Ide, 1990; Kozima & Furugori; Niwa & Nitta, 1994). Karov eta Edelmann-ek ere (1996; 1998) halatsu egiten dute, baina corpusa eta hiztegiko adierak lotzeko metodo bat planteatzen dute. Ez dugu esango hiztegian erlazio-izaerari buruzko informaziorik ez dagoenik, alderantziz, baina informazio hori era gordinean dago, egituratu gabe. Eta hori da hain zuzen ere Microsoft-eko taldearen ekarpena (Richardson, 1997), erlazio-izaera hiztegitik erauzitako Hiztegi-Ezagutza Base egituratu baten oinarrituta formalizatzea, eta ez zuzenean hiztegiko informazio gordina. Guk ere hiztegien ekarpena hor ikusten dugu, erlazio anitz erauzi ahal izateko potentziala duten gordailu bezala. Richardson-en lanean (1997) ez bezala, horrek ontologiak eta hiztegiak lotzea eskatzen du. Adierak eta kontzeptuak lotu behar dira, eta erlazioak hitzen artekoak baizik adiera/kontzeptuen artekoak izan behar dute. VI kapituluan helduko diogu gai horri, HEB batean adiera desanbiguazioa egin eta kanpoko ontologia bati lotzeari.

Corpusetako lanak dira zalantza gabe erlazio-izaeraren aplikazioetan emaitza onenak lortu dituztenak. Lengoaia Naturalaren Prozesamenduan asko ari dira hedatzen horrelakoak, eta nahiz eta batez ere lan enpirikoak izan, corpusen erabileraren inguruan ere eratzen ari da halako marko teoriko bat. Hala ere, kontzeptuen erlazio-izaera lantzean arazo garrantzitsuekin topatu ohi dira. Lehenbizikoa **adieraren definizio zuzenik ez** egotea da, ez dago kontzeptuenganako loturarik inon. Lan batzuek, horrela izanda, hitzen arteko erlazio-izaera besterik ez dute definitzen (Grefenstette, 1992; 1996; Grishman & Sterling, 1994; Lee, 1997; Golding & Schaves, 1996). Teoriaren aldetik kezkagarria bada, alderdi praktikoan ere arazoak ekartzen ditu, adieratara hedatu ahal izateko **eskuzko etiketatze semantikoa** eskatzen baitu (Church & Hanks, 1990; Hearst, 1991)[40]. Eskuzko etiketatzeak planteatzen duen arazo nagusiena denbora eta eskulan kopuruarena da, baina ez hori bakarrik, adieren mugak lausoak izaten baitira sarritan, eta giza-etiketatzaileen arteko ezadostasun maila nahiko altua da (%32koa Jorgensen-en arabera (1990)).

[40] Gale-ek eta (Gale et al. 1992; 1993; Yarowsky, 1993) adierak beste testuinguru baten definitzen dituzte, testu paraleloetan itzulpen ezberdinaren arabera. Horrela, aplikazio mugatu batentzat – itzulpen kontuetan – eskuzko desanbiguazioaren arazoa ekiditzen dute. Hala ere hau ezin izan dute orokortu beste adiera edo kontzeptuen definizioetara, eta arazo teorikoak hor dirau.

Corpusen inguruko hasierako proposamenei egindako hobekuntzak alor honetan izan dira batez ere: nola lortu eskuzko desanbiguaziotik alde egitea eta adierak euskarri trinkoago bati lotzea. Schütze-k (1992a; 1992b) hitzen agerpenak automatikoki multzokatzen ditu. Hearst berak eta Schütze-k (1993) WordNet-eko kategoriak multzokatu eta corpusetako hitzen agerpenak multzo horiei lotzen dizkie. Yarowsky-k (1992) adierak thesaurus bateko etiketa semantikoz bereizten ditu, baita ere automatikoki. Yarowsky berak, aurreragoko lanean (1994; 1995) beste hurbilpen bat hartu eta giza-anotazio lana errotik gutxituko duen algoritmoa plazaratzen du. Lan hauek guztiak norabide interesgarria edukita ere, ez dira heltzen adierei oinarri sendo bat ematera, eta batez ere ez dute lortzen hitzen agerpenak ontologiako kontzeptuei lotzea. Leacock-en azken lanean (Leacock et al., 1998) WordNet-en erabilera planteatzen dute eskuzko lana gutxitzeko. Ahalegin handiak eginda ere gaur egun eskuz etiketatutako adibide kopuru oso urria ikusita ez dirudi oraingoz corpusetan oinarritutako teknikak hitz gutxi batzuetatik harantzago joan ahal izango direnik.

Corpusetan oinarritutako lanek ere badute beste arazo bat, jadanik aipatutako **datu urrien arazoa** ikus III.A.3 atala). Arazo hori hitzak egitate isolatuak bezala aztertzetik dator, klase edo multzoak kontuan hartu gabe. Honek, nahiz eta paradoxiko iruditu, beste arazo bat ere sortzen du, **datu gehiegizkoen arazoa** deitu daitekeena. Alde batetik, datu urrien arazoa arintzeko ahal den corpus zabalenak erabiltzea komeni da. Bestetik, hitzen agerpen guztiak hartu behar dira kontuan erlazio-izaera behar den bezala aztertzeko. Hori dela eta hiztegiko hitz bakoitzarentzat bildu beharreko informazioa oso zabala da, eta hitz guztiena batzen badugu izugarria (ikus arazo honen adibide bat testu-zuzenketa automatikoan, V.D.3 atalean). Ziur aski hau da corpusen inguruko sistemak hitz multzo txikiekin[41] ebaluatu izatearen arrazoi nagusia. Bi arazo hauei erantzuten die Resnik-en lanak (1993a; 1993b; 1995; 1997), WordNet-eko klaseei corpusetan duten maiztasunari buruzko informazioa gehituaz, eta aditz eta objektuen arteko erlazio-izaera hitzez-hitz egin beharrean ontologiako kontzeptu eta klaseen arabera egiten den.

Hiztegietarako esan dugun bezala, corpusak ere informazio biltegi erraldoiak dira, baina bertan dagoen altxor hori, erabilgarria izanda ere, era egituratu batera erauzi beharra dago. Corpusetatik, beste inondik baino hobeto ziur aski, erraz atera daiteke urdaiazpiko eta sardexka hertsiki erlazionatuta daudela, baina ez litzateke nahikoa izan behar lotura horrek 0,967 indarra duela, erlazio beraren izaera ere lortu beharko litzateke. Horrela balitz, informazio esanguratsuena ontologiatan bildu zitekeen, modu konpaktuago batean, inferentzia mota ezberdinetarako integratuaz. Esandakoaren adibide bat da gorago aipatutako Resnik-en lana, aditzen hautapen-

[41] Yarowsky-k (1992) adibidez 8 hitzen gainean egiten du.

murrizpena WordNet-eko klaseen arabera deskribatzen baitu, hitzez-hitzeko informazioa laburbilduaz.

Ontologietan oinarritutako neurriak dira, beraz, teoria sendoena dutenak. Gainera adiera zer den garbi dago definitua, ontologiako kontzeptuen erreferentzien bidez. Ontologien arazoa, hala ere, eduki arazoa da. Ontologiaren diseinuan ezaugarri eta erlazio aberatsak egonda ere, beharrezko kontzeptu guztietan erlazio eta ezaugarri horiei balio bat ematea ez da makaleko lana. Eta ontologiak hiztegiaren zati garrantzitsua estaltzea ere beharrezkoa da. II. kapituluan aipatu dugu hau guztia, eta ikusi nola ontologia guztiek hiztegiaren estaldura arazoak dauzkaten. Hiztegi aldetik aberatsenetakoa WordNet da, baina honek erlazio gutxi dauzka landuta. Ontologietan oinarritutako erlazio-izaeren arazoa hori da, hain zuzen ere, ontologian dagoen informazioa besterik ezin dutela erabili (ikus horri buruzko iruzkinak III.F. atalean).

III.D.2. Dentsitatea eta ontologiatan oinarritutako beste teknikak

Tversky eta Quillian-en lanak interesgarriak izanda ere, erlazio-izaeraren inplementazio eraginkorra egiterakoan alde batera utzi izan dira. *Spreading activation* bitartez erlazio-izaera kalkulatzeko sare semantikoko[42] adabegi guztiak bisitatu behar dira, ez behin, baizik eta hainbat aldiz.

Radaren taldeak, sare semantikoen antolaketa kontuan hartuaz, beste erlazioak alde batera utzi eta erlazio paradigmatikoa soilik erabiltzea planteatu zuen, eraginkortasuna nabarmenki hobetuaz. Sussna izan zen lehenbizikoa Distantzia Kontzeptuala WordNet-en inplementatzen, bi kontzepturen arteko bideak erabiliaz. Erlazio paradigmatikoak soilik ez, eskura zituen meronimikoak ere erabili zituen, hobekuntza apala lortuaz bere esperimentuetan. Nahiz eta inplementazioak eraginkortasun arazorik ez eduki bi kontzepturen arteko distantzia bilatzeko (hierarkiaren batezbesteko sakonera bezainbat erlazio esploratu behar dira soilik, hau da ordena konstantea duen algoritmo batez –O(kte)– kalkulatu ahal da), lehen ikusi dugun bezala (III.B.2 atala) batez-beste M adiera dituzten N hitzen arteko distantziak kalkulatzeko $\frac{N\times(N-1)}{2}\times M^2$ aldiz bilatu behar da bidea. Honek $O(N^2)$ konplexutasuneko algoritmoa eskatzen du. Kontuan izanik autore batzuek 100 hitzetako leihoak darabilzkitela (adibidez adiera-desanbiguazioa egitean), arazo hau larria bihurtzen da.

Dentsitate Kontzeptualak, aldiz, beharrezko hitz guztien adieren arteko Dentsitatea behin kalkulatzen du, $N\times M$ adierak behin tratatuaz eta beraz konplexutasun apalagoko algoritmo bat onartuaz.

III.B.2 atalean aipatu bezala, binaka neurtzearen arazoa ez da praktikoa bakarrik, teorikoki ez dago oso argi N kontzepturen arteko binakako distantziak batzeak zer esan nahi duen, eta gainera ez da ageri modu errazik kopuru ezberdineko kontzeptu multzoen arteko distantziak konparatu ahal izateko. Dentsitateak aldiz edozein tamainako kontzeptu multzoen gertutasuna era natural batean neurtzeko neurria ematen du.

Dentsitatearen azterketarekin bukatzeko (emaitzen araberako ebaluazioa IV.C.3, V.D.3, VI.C.2 eta VI.D.9 ataletan jorratuko dugu), gogora ditzagun aldez aurretik jarri genizkion baldintzak:

1. Ontologiatan oinarritutakoa.
2. Adieren arteko neurria: ontologiako kontzeptuei erreferentzia egingo diena
3. Erlazio paradigmatiko eta sintagmatikoetako informazioa erabiliko duena
4. Kategoria irekietako hitzekin lan egingo duena
5. Eraginkorra izatea, testu zabalekin lan egin ahal izateko bezalakoa.
6. N kontzepturen arteko neurria izatea
7. Kontzeptu kopuru ezberdineko multzoen gertutasunak konparagarriak izatea.

Ezaugarri desiragarri hauetatik ikusi dugu Dentsitateak 1, 2, 5, 6 eta 7 betetzen dituela. 4. ezaugarriari dagokionean, Dentsitate Kontzeptuala izenekin besterik ez dugu probatu (ikusi IV., V.eta VI. kapituluak), baina ez dago eragozpenik beste kategoriatara hedatzeko. Ikustekoa da, noski, izenekin bezain emaitza onak lortu ahal izango direnik.

3. ezaugarriari buruz, II. kapituluan eta III.B.2 atalean ikusi izan dugu nola, gaur egun, ez dagoen hedadura zabaleko ontologiarik eskuragarri WordNet ez denik. WordNet-en arabera diseinatu da beraz Dentsitate Kontzeptuala, eta hala izan da erlazio hiperonimiko eta meronimikoak bakarrik erabiltzen dituela. Hau da, erlazio sintagmatikoetaz ez da baliatzen.

III.E. Ekarpena

Kapitulu honen helburu nagusia ezagutzan oinarritutako kontzeptuen arteko erlazio-izaera definitzea da, eta horretarako WordNet-en oinarritutako Dentsitate Kontzeptuala diseinatu eta inplementatu dugu.

Lehenbizi erlazio-izaera eta antzekotasunaren hainbat formalizazio aztertu ditugu, ezaugarri batzuen arabera. Hiztegi eta corpusetan oinarritutakoak interesgarriak izanda ere, oinarri teoriko sendoa dutenak ontologiatan oinarritutakoak direla arrazoitu dugu. Corpusen kasuan, emaitza oso onak

[42] Tesi honi dagokionean, sare semantikoak ontologia mota berezi bat bezala hartu ditzakegu.

lortuta ere adieraren definizio sendorik ez dagoela ikusi dugu, eta eskuzko desanbiguazioa ezinbestekoa dela sistemak adierak ezagutu ditzan. Gainera datu urrien eta datu gehiegizkoen arazoei aurre egin beharrean daude.

Hiztegi eta corpusetatik informazioa erauzteko beharra aitortzen dugu, eta garatutako erlazio-izaera espezifikoak horretan oso lagungarriak izango dira, baina defendatzen dugun ideia ontologia aberastearena da, ontologia aberats horretan lan egingo duen erlazio-izaera egokia definituaz. Laburbilduz ontologiek honako ezaugarriak eskaintzen dizkigute, besteen aurrean:

- Oinarri teoriko sendoa
- Adieren definizio sendoa
- Ez dute eskuzko desanbiguazio beharrik, ez eta datu urrien edo gehiegizkoen arazorik.

Ontologietan oinarritutako hauek eraginkortasun-arazoak dituzte. Gainera erlazio-izaeraren neurri guztiak bi kontzepturen artekoak, eta ez gehiago, izaten dira. Guk definitu dugun Dentsitate Kontzeptualak edozein kopurutako hitz multzoen erlazio-izaera kalkulatzeko gai da, konplexutasun apalagoarekin. Laburbilduz ezaugarri hauek dauzka Dentsitate Kontzeptualak:

1. Ontologiatan oinarritutakoa da.
2. Adieren arteko neurria da: ontologiako kontzeptuei erreferentzia egiten die.
3. Erlazio paradigmatikoetako informazioa erabiltzen du (hiperonimia eta meronimia).
4. Izenekin lan egiten du (aditzetarako ere egokia izan daiteke).
5. Eraginkorra da, testu zabalekin lan egin ahal izateko adinakoa.
6. N kontzepturen arteko neurria da.
7. Kontzeptu kopuru ezberdineko multzoen gertutasunak konparagarriak dira.

Dentsitate Kontzeptuala WordNet-en gainean inplementatu dugu, II. kapituluan arrazoitu bezala.

Erlazio-izaeraren neurri honek ez du eskatzen aurretiko inongo prestakuntzarik eta zeregin oso ezberdinetan aritu daiteke lanean, hurrengo kapituluetan ikusiko dugun bezala:

- Hitzen Adiera-Desanbiguazioa (IV. kapitulua)
- Testuen Zuzenketa Automatikoa (V. kapitulua)

- Ingelesa ez diren baliabide lexikal egituratuen eraikuntza sendotzeko (VI. kapitulua). Gure erabilera bikoitza da:
    - *Le Plus Petit Larousse* frantses hiztegiko adierak WordNet-i lotu
    - *Le Plus Petit Larousse*-etik erauzitako HEBko adieren hierarkiak desanbiguatu eta trinkotzea

III.F. Etorkizunerako lana

Dentsitate Kontzeptuala hobetzeko hiru alor nagusi hauek ikusten ditugu:

1. Darabilen informazioari dagokiona: erlazio sintagmatikoak dituen ontologia bat lortu edo WordNet erlazio sintagmatikoez aberastu.
2. Formulari dagokiona: Dentsitate Kontzeptualaren formula aldatu, bestelako erlazioak kontuan har ditzan.
3. Inplementazioari dagokiona: Inplementazioa azkartu.

Dentsitate Kontzeptualaren euskarri den WordNet ontologiak erlazio paradigmatikoak besterik ez dituenez, Dentsitate Kontzeptuala kalkulatzean ez da azaltzen erlazio sintagmatikorik. Informazio hori oso baliagarria izan daiteke erlazio-izaera neurtu ahal izateko: *baloia* eta *oina*, adibidez, oso urruti daude bata bestearengandik erlazio paradigmatikoak bakarrik erabiltzen baditugu, baina argi dago erlazio estua dagoela bien artean, beraien artean erlazio funtzional bat dagoelako eta ondorioz testuinguru berdinetan maiz azaltzen direlako. Arestian aipatu dugun bezala, hau da ontologia eta EBLek daukaten muga garrantzitsuetako bat, oso zaila baita halako informazioa eskuratzea.

Saiakerak egin dira, halere. Adibidez, ikus hautapen-murrizpenak corpusetatik ikasteari buruzko lanak (Grishman & Sterling, 1994; Ribas, 1995; Resnik, 1997). Corpusetatik adiera bakoitzak dauzkan kolokazioak ikasiz gero ere lagungarriak dira adieren arteko gertutasuna kalkulatzeko (Yarowsky, 1993; 1995). Hiztegi elektronikoetako definizioetatik agente, objektu eta antzeko erlazioak erauzi daitezke ere (Artola, 1993; Richardson, 1997). Topikoari buruzko informazioa ere interesgarria izan daiteke, Roget's tesaurusean edo LDOCE hiztegian azaltzen diren bezalakoak. Aurrekarien kapituluan azaldu ditugun informazio iturri guztiak izan daitezke baliagarriak.

Beraz, erlazio sintagmatiko horiek corpusetatik edo hiztegietatik erauzi daitezke, horien analisiaren bidez. Baliabide jakin batetatik erauzitako erlazioak baliabide horretako kontzeptu eta adieren artekoak izango direnez, WordNet aberasteko baliabide horiek bat egin beharko lirateke WordNet-ekin. Horrela WordNet-en integratuz joango lirateke corpus eta hiztegietatik erauzitako

informazioa, edo bestelako ontologietan dauden erlazioak. Baliabide egituratuen bat egiteari buruz V. kapituluan arituko gara.

Nahiz eta erlazio paradigmatiko eta sintagmatikoak dituen ontologia eskura eduki, bai WordNet aberastuaz lortutakoa edo zuzenean eskuratutakoa, Dentsitate Kontzeptuala, hemen definitu dugun bezala, ez da gai informazio berri horretaz baliatzeko. Izan ere Dentsitate Kontzeptuala hierarkientzat dago pentsatuta, eta bestelako erlazioak integratzeko hedatu egin beharko zen. Erlazio sintagmatikoak erabiltzen dituzten erlazio-izaeren artean, (Agirre et al. 1994b)-ek LPPL-tik erauzitako erlazio paradigmatiko eta sintagmatikoak erabiltzen dituen Distantzia Kontzeptualerako proposamena egiten du. (Mahesh et al. 1997)-ek Mikrokosmos ontologiarekin eta Richardson-ek (1997) hiztegietatik erauzitako HEBarekin ere egiten dute beraien proposamen propioa.

Azkenik, nahiz eta Dentsitate Kontzeptualaren algoritmoa konplexutasun gehiegizkoa ez izan, inplementazio azkarragoa lor daitekeela uste dugu. Horren arrazoietako bat LISP lengoaiaz inplementatuta egotea da, eta bestea WordNet-eko informazioaren atzipena ez dagoela optimizatuta. Egun, C++ lengoaiaz inplementatutako bertsio bat lantzen ari gara, UNED-eko Elektrizitate eta Elektronika saileko ikerkuntza taldearekin batera, ITEM[43] proiektuaren barruan. Bertsio hau ingeniaritza linguistikorako GATE[44] ingurunearen barruan (Cunningham et al. 1997) integratuta egongo da laster.

[43] http://sensei.ieec.uned.es/item/

[44] http://www.dcs.shef.ac.uk/research/groups/nlp/gate/

IV. Kapitulua
HITZEN ADIERA-DESANBIGUAZIOA TESTU ERREALETAN

Kapitulu honetan Dentsitate Kontzeptualaren ebaluazio praktikoa egin nahi izan da, aplikazio bezala Hitzen Adiera-Desanbiguazioa (HAD) erabiliz. Horretaz gain Dentsitate Kontzeptualaren parametro batzuk finkatzea ere nahi izan dugu. Hasteko hitzen adiera-desanbiguazioari buruzko sarrera egingo dugu eta aurrekariak aztertu. IV.B. atalean esperimentuaren diseinua azalduko dugu. Ondoren Dentsitate Kontzeptuala erabiltzen duen algoritmo desanbiguatzailea aurkeztuko dugu, eta ebaluazioari buruz ihardun aurretik Dentsitatearen parametroak doituko ditugu. IV.D. atalean beste metodoen emaitzekin konparatuko dugu gurea, eta bukatzeko kapitulu honetako ekarpenak aipatuko ditugu.

IV.A. Sarrera eta aurrekariak

HADaren garrantziari buruz honakoa zioen Hirst-ek (1987: 5 or.): "*Any practical Natural Language Understanding system must be able to disambiguate words with multiple meanings, and the method used to do this must necessarily work with the methods of semantic interpretation and knowledge representation used in the system*.". Duela gutxi argitaratutako HADaren egungo egoeraren azterketan Ide eta Véronis (1998) ere uste berekoak dira: "*Sense disambiguation is an intermediate task, which is not an end in itself, but rather necessary at one level or another to accomplish most natural language processing tasks. It is obviously essential for language understanding applications,* …". Lengoaia naturalaren ulermenerako soilik ez ordea, HADaren ekarpena oinarrizkoa da beste aplikazio askok eraginkortasuna lortu dezaten, hala nola, itzulpen

automatikoa, informazioaren berreskuratzea, dokumentuen berreskuratze eta sailkapena, analisi sintaktikoan bertan, testu eta mintzairaren prozesamenduan, etab.

Azken urte hauetan HAD bigarren maila batetatik Lengoaia Naturalaren Prozesamenduaren lehentasunezko arazo izatera pasatu da berriz ere. HADaren berezko zailtasuna Bar-Hillel-en (1960) itzulpen automatikoari buruzko tratatu ezagunean puntu nagusia zen. Bere argumentuek ALPAC (1966) txostenaren oinarria izan ziren, hain zuzen ere 60. hamarkadan itzulpen automatikoaren finantzaketaren beherakada ekarri omen zuena. Adimen artifizialean oinarritutako desanbiguazio sistemak aro horretan hasi ziren zabaltzen, baina beti ere arrakasta apalarekin. Orduko desanbiguatzaileak hitz anbiguo aldrebesekin frogatu ohi ziren, esaldi gutxi batzuetako agerpenak aztertuaz. Azken hamarkada honetan, ordenadorez atzitu daitekeen testu kopuru zabalak bultzada eman die datuetan oinarritutako teknikei, emaitza deigarriak lortuaz testu errealetan. Hori dela eta HADak inoiz baino arreta gehiago jaso izan du (adibidez, HADari buruzko 171 artikulu dauzkagu jasota, horietatik 161 90. hamarkadan argitaratuak), eta gaur egungo lengoaia naturalaren prozesamenduak duen arazo nagusienetako bezala aipatu izaten da.

Adimen artifizialaren inguruko ikerlariek, baita LNPko buru gehienek ere, HAD arazoa *AI-complete* dela uste dute, hau da, konpondu ahal izateko lehenbizi adimen artifizialeko arazo gaitz guztiak, sen ona eta ezagutza entziklopedikoaren errepresentazioa barne, ebatzi beharko lirateke. Gaur egun, ordea, eta aurreko baieztapenari arrazoia kendu gabe ere, HAD teknologia heltzen ari dela usten dutenak ugaltzen doaz, eta testu libreetako hitz gehienentzat adiera egokiena topatzea eskura dugula aditzera eman nahi dute, nahiz eta modu ez-perfektu batez izan, hainbat aplikaziotan erabilgarria izateko modura.

HADak hartu duen garrantzia dela eta, honen inguruko lanak asko ugaldu dira. Horien azterketari ekin baino lehen HADaren karakterizazio bat egingo dugu. Bi pauso nagusi ezberdindu ohi dira (Ide & Véronis, 1998):

1. hitzek dauzkaten **adieren zehaztapena**.
2. hitz bakoitzaren agerpenari **adiera bat esleitzeko metodoa**.

Adieraren definizio zehatza zein den Aristotelerengandik hasi eta egun arte erabaki gabeko eztabaidaren gunea da. Hori horrela izanda, autore batzuek (ikusi adibidez Kilgarriff, 1997a) adierak zerrendatu beharko liratekeen ere zalantzan jartzen dute, eta beraien kritikek filosofia, psikologia eta linguistikan erro sendoak dituzte. Kritikak kritika izanda ere, **adiera-zerrendak** dira nagusi

ikerkuntza arlo honetan, eta lan gehienak aurrez emandako adieratan oinarritzen dira, hurrengo eratakoak barne:

- adieren zerrenda (hiztegietan aurkitu daitekeenaren antzekoa)
- ezaugarri multzoa (adibidez, corpusetatik ateratako testuinguru-ezaugarriak)
- beste hizkuntza baterako itzulpenen zerrenda

Hitz baten agerpena desanbiguatzean hainbat ezagutza-iturritara jotzen da, baina nagusienak horrela sailkatu ditzakegu:

- desanbiguatu behar den hitzaren **agerpenaren testuingurua**: testua, diskurtsoa, informazio extra-linguistikoa, etab.
- kanpoko ezagutza iturriak: **baliabide lexikal edo entzilopedikoak**, eskuz sortutako ontologiak, etab.

Desanbiguazio lan orok zera eskatzen du, hitzaren agerpenaren testuingurua kanpoko ezagutza iturriko informazioarekin edo lehenago corpusetan desanbiguatutako hitzaren beste agerpenetatik eratorritako informazioarekin ezkontzea Bata edo bestea aukeratzeak sortzen du HADan dauden bi familia nagusien arteko bereizketa: **ezagutzan oinarritutako HAD** edo **datuetan oinarritutako HAD**. Adierari buruzko informazioa agerpenarekin ezkontzean asoziazio metodoak erabiltzen dira batez ere, aurreko kapituluan aipatutako erlazio-izaerak hain zuzen ere (ikusi III.A atala)

Erlazio-izaeraren garrantzia HADan bistakoa da, erlazio-izaeraren formalizazio askoren motibazioa HAD bera da eta. Aurrekarien azterketan ikusiko dugun bezala, HADaren hurbilpen askotan erlazio-izaera hutsa erabiltzen da adiera aukeratzeko, nahiz eta lan teorikoenek beste teknika eta informazio iturriak beharrezkoak direla planteatu, orain ikusiko dugun bezala. Beraz, ezagutzan oinarritutako HAD lanek ontologia zein hiztegiz baliatutako erlazio-izaera darabilte, eta datuetan oinarritutako HAD sistemak corpusetan errotutako erlazio-izaera neurriak. Beheko IV.A.1. atalean azalduko dugu nola erabili daitekeen erlazio-izaera HADrako.

Aurrekarien azterketa hasi aurretik hitz bi anbiguetate lexikalaren inguruan. Hiru mota bereizi izan dira anbiguetate lexikalean: polisemia (adieren esanahia erlazionatuta dago), homonimia (adieren artean ez dago erlaziorik) eta kategoriazko anbiguetatea (adierak kategoria ezberdinekoak dira). Kategoriazko anbiguetatea garrantzitsua izanda ere, albo batera utzi ohi da, ezagutza sintaktiko hutsez erraz ebatzi ohi da eta. Homonimia eta polisemia inguruan ez da normalean bereizketarik

egiten, ez bada sistema batzuk homonimia mailako anbiguetatea bakarrik kontuan hartzen dutelako. Hirst-ek (1987) hala zioen polisemia eta homonimiaren arteko bereizketaren inguruan: "*The semantic objects we will be using are discrete entities, and if a word maps to more than one such entity, it will generally (but not always) be a matter of indifference how closely related those two entities are.*". Gainera polisemia, homonimia eta metaforen artean ez dago muga garbirik; alde batetik polisemia eta homonimia arteko ezberdintasuna oso erlatiboa da, subjektiboa, eta bestalde urteak pasa ahala metafora lexikalizatu egin daiteke, homonimia edo polisemia emanaz. Metaforaren tratamendua, dena den, lan honen esparrutik kanpo dago.

Aurrekarien azterketa, aurreko kapituluan bezala, ezagutza iturriaren arabera antolatu dugu: ontologiak, hiztegiak, corpusak eta konbinazioak. HADaren inguruko azterketa bibliografiko sakona egitea gehiegizkoa litzateke hemen, eta aurreko kapituluan aurkeztutako lanak azalduko zaizkigu hemen bereziki. Lanetako gehienek ezagutza iturri bakarra lantzen dute, halakoarekin desanbiguazio onargarria lortu daitekeelakoan. Horiek aztertu baino lehen aipatu ditzagun HAD ikuspuntu orokorrago batetik aztertu izan dituzten oraintsuko lan bi.

IV.A.1. Beharrezko diren ezagutza iturriak

Adimen Artifizialean oinarritutako hurbilpenak dira, zalantza gabe, desanbiguazioan parte hartzen duten faktoreak sakon aztertu eta desanbiguatzeko beharrezkoak diren informazio iturriak bereizi dituztenak. Hirst-ek (1987) adibidez diskurtsoaren testuingurua eta esaldiko bertako pistak kontuan hartu beharrekoak zirela uste zuen[45]. Alde batetik diskurtsoaren testuingurua finkatzea lortuz gero (testuaren domeinu eta topikoa) hitzaren adiera bakarra izan daiteke egokia testuinguru horrentzat. Bestetik, esaldian bertan dauden pistak nahikoak izan daitezke adiera bereizteko, hara nola esaldiko hitzen adieren arteko erlazio-izaera (Quillian 1968), pista sintaktikoak, edo hautapen-murrizpenak. Dena den, badaude hainbat kasu goi mailako inferentzia (sen ona) eskatzen dutenak adiera hautatu ahal izateko.

McRoy-k (1992), oraintsuagoko lanean, ezagutzan oinarritutako sistema batek kontuan hartu beharko lituzkeen ezaugarri eta ezagutza motak zerrendatzen ditu:

- hitzaren morfologia
- testuinguruari egokitzen zaion hitzaren kategoria
- domeinu edo maiztasunaren arabera, zein adiera diren egokiagoak

- ea hitza kolokazioren bateko parte den, adibidez *soda cracker* edo *take action*
- testuinguruak adieraren bat nahiago duen: testuinguruko beste adierekin topiko, egoera edo kategoria semantikoari dagokionez harremanik duen
- ea adierek dauzkaten baldintza sintaktikoak testuinguruan betetzen diren
- ea hautapen-murrizpenak betetzen diren
- ea adiera diskurtsoan indarrean dagoen zerbaiti lotuta dagoen

Guzti horietatik ordea, garrantzitsu edo emankorrenak hauek direla deritzo: adieren kategoria bera (kategoriazko anbiguetatea ebazteko), morfologia (adibidez *agreement*-en *agree*-ren adieretatik 3 bakarrik dira egokiak eratorpen horretan), kolokazio eta hitzen arteko asoziazioak (adibidez, *bank/money* elkarrekin azaltzea edo *increase in* orduan *in* hori normalean ekintzaren pazienteari dagokio eta ez akzioaren leku edo norabidea). Hautapen-murrizpenak ere garrantzitsuak direla dio, baina bigarren mailan.

Hirst eta McRoy-k aipatutako ezagutza iturriak laburbilduz gero:

1. adieraren agerpenaren kategoria
2. morfologia
3. pista sintaktikoak eta kolokazioak maneiatzeko mekanismoa
4. hautapen-murrizpenak betetzen diren erabakitzeko mekanismoa
5. inguruan dauden hitzen arteko harremanak bilatzeko mekanismoa
6. testuinguruaren ezagutza (topiko eta domeinua)
7. inferentzia orokorra, azken irtenbide bezala

Esan bezala, kategoriazko anbiguetatea ebatzita dago, gaur egun dauden kategoria etiketatzaileen eraginkortasunari esker. Egile gutxik aitortzen dute, bestalde, morfologiaren garrantzia, aplikazio gehienetan hitzaren barruko egitura ez delako interesgarria, nahikoa lan dute hitz osoaren adiera erabakitzen. Literaturako lanetan nekez topatzen dira inferentzia orokorraren ekarpenari buruzkoak ere. Onartu izaten da desanbiguazio osoa lortzeko beharrezkoa dela, baina ezaguna da ordenadoreen sen ona ez dela oraingoz inondik inora ageri. Gehien azaltzen diren ezagutza iturriak, beraz, 3.etik 6.era doazenak dira.

[45] "*For an NLU system to be able to disambiguate words, it is necessary that it use both the discourse context in which the word occurs and local cues within the sentence itself.*" (Hirst, 1987; 6. or.)

Hirst eta McRoy-ren aipaturiko lanetan ez bezala, ordea, gehienetan ezagutza iturri horiek ez dira modu bereizi baten azaltzen. Izan ere ezagutza iturri horiek ez daude eskuragarri, eta beraien erauzketa eta eraikuntza LNParen arazo estuenetako bat da. Horregatik beharbada, lan gehienetan ezagutza iturri bakarra erabili ohi izaten da, corpus, hiztegi edo ontologietan oinarritutako erlazio-izaeraren neurriren bat (ikusi 7. taula).

3.	pista sintaktikoak eta kolokazioak: ..... adierei buruzko ezagutza sintaktikoa
4.	hautapen-murrizpenak: .....erlazio sintagmatiko lokala, paradigmatikoaz konbinatuaz
5.	hitzen arteko harremanak: ..... erlazio sintagmatiko lokal, global eta paradigmatikoa
6.	topiko eta domeinua:..... erlazio sintagmatiko global eta paradigmatikoa

7. taula: desanbiguatzeko beharrezko ezagutza eta erlazio-izaeraren arteko harremana

Erlazio-izaeraren bidez modelatzen ez den ezagutza mota bakarra pista sintaktikoak eta kolokazioak dira. Horien garrantzia esperimentalki frogatu izan da, eta emaitza hoberenak lortu dituzten sistemak erlazio-izaerekin integratu izan dituzte (ikusi IV.A.4. atala).

IV.A.2. Ontologiatan oinarritutako HAD

Arestian aipatutako bi lanak (Hirst, 1987; McRoy, 1992) Adimen Artifizialean oinarritutako hurbilpenen adibide tipikoak dira. Alde batetik adierak zer diren definitzeko ontologiako kontzeptuetara jotzen dute, eta hala hitz bat anbiguotzat joko da ontologiako kontzeptu bati baino gehiagori egiten badio erreferentzia. Adierak, beraz, adiera-zerrenda baten bidez definitzen dira. Bestalde, nahiz eta potentzialki anbiguetate korapilatsuak ebazteko prestatuak egon, praktikan, ontologian dagoen informazio murritza dela eta, adibide gutxi batzuekin besterik ez dira probatu izan. Bi lan horietan ebaluazioa era abstraktuan egin izan da, inongo emaitza enpirikorik azaldu gabe.

Sussna-k (1993) bai ebaluatzen duela ontologiatan oinarritutako bere sistema, informazioa iturri bakarrera murriztearen truke: WordNet-eko ezagutza paradigmatikoa. Distantzia Kontzeptualaren bidez lortzen dituen emaitzak ez ditu zuzenean konparatzeko moduan ematen, baina 8. taulan azaltzen den doitasuna kalkulatu diogu, gutxi gora behera. IV.D atalean lasaiago ebaluatuko dugu, gure hurbilpenaren antzekoa da eta. WordNet erabiltzen duten lan gehiago ere badaude, eta horietako batzuk IV.A.5. atalean ikusiko ditugu, ezagutza iturri gehiago konbinatzen baitituzte.

Mahesh-ek eta ontologian bertan errepresentatzen dute hautapen-murrizpena. Izenen adierak desanbiguatzeko hautapen-murrizpenak eta ontologiako kontzeptuen adieren arteko hurbiltasuna

nahiko direla defendatzen dute, baina nahiz eta ondo arrazonatu, ezin dute zenbakizko daturik eman, beraien Mikrokosmos ezagutza-basearen estaldura urria delako.

Aurreko kapituluan aipatu ziren atal honetako lan batzuk ez ditugu hemen aipatu, HADan erabili ez direlako. Berdin gertatuko da beste sailetan ere.

	Erlazioak			Adierak		Ebaluzioa			
	Par	Lok	Gl	Jatorria	Granularitatea:	Kop.	Kat.	Est.	Doi.
Hirst, 1987	X	X	X	Ontologia	Polisemia	-	-	-	-
McRoy, 1992	X	X	X	Ontologia	Polisemia	-	-	-	-
Sussna, 1993	X			WordNet	Polisemia	~1000	Izena	?	~%47
Mahesh et al., 1997	X	X		Mikrokosmos	Polisemia	-	-	-	-

8. taula: ontologian oinarritutako lanen sinopsia[46]

IV.A.3. Hiztegietan oinarritutako HAD

Hiztegietan oinarritutako sistemetan adiera zerrendak erabiltzen dira adierak definitzeko, noski. Erlazio sintagmatiko globalean oinarritutako erlazio-izaera erabiltzen da. Lesk-en hasierako proposamenaren hedadura ezberdinek antzeko emaitzak lortzen dituzte: %50 baino gutxiago polisemia mailan eta %70 inguru homonimia mailan (ikusi 9. taula). Aipatu beharra dago Véronis eta Ide-ren kasuan (baita Niwa eta Nitta-renean ere) emaitzak ematerakoan erabilitako irizpideak ez daudela batere garbi.

	Ezagutza			Adierak		Ebaluzioa			
	Par	Lok	Gl	Jatorria[47]	Granularitatea	Kop.[48]	Kat.	Est.	Doi.
Lesk, 1986			X	W7 OALDCE CED	Polisemia	2 testu	denak	?	%50-%70
Cowie et al., 1992			X	LDOCE	Polisemia	50 esaldi	denak	?	%47
					Homografia				%72
Véronis & Ide, 1990			X	CED	?	?	?	?	%72
Niwa & Nitta, 1994			X	CED	Domeinua (2 adiera)	9x20 hitz	izenak	?	~%75
Wilks et al., 1990			X	LDOCE	Polisemia	1x197 hitz	izenak	100%	%45
					Homografia				%90

9. taula: hiztegietan oinarritutako lanen sinopsia

[46] Taulako eremuen esanahia:
Erlazioak: paradigmatiko, sintagmatiko lokala eta sintagmatiko globala.
Adierak: adieren jatorria eta granularitatea, bereizketa xehea (polisemia) edo zabala (homonimia, homografia edo domeinua).
Ebaluazioa: zenbat hitzekin egin den, hitzen kategoria, emaitzen estaldura eta doitasuna (~ ikurrak gutxi gora bera esan nahi du).

[47] Laburduren esanahia: CED *Collins COBUILD English Language Dictionary* (Sinclair, 1987), W7 *Webster's Seventh New Collegiate Edition* (Gove, 1969), OALDCE *Oxford Advanced Learner's Dictionary of Current English* (Hornby, 1974), LDOCE *Longman's Dictionary of Contemporary English* (Procter, 1978).

[48] 9x20 hitz azaltzen denean, 9 hitz aukeratu eta horietako bakoitzaren 20 agerpen desanbiguatu direla adierazi nahi da.

IV.A.4. Corpusetan oinarritutakoak

Adierak espezifikatzeko adiera-zerrendak erabiltzen dira batez ere, eta horrek corpusetan adieren etiketak eskuz jarri beharra suposatzen du. Tamalez, adieraz etiketatutako corpusak oso urriak dira, eta beraz metodo hauen etsai amorratuena eskuzko desanbiguazioaren beharra da. 10. taulan zutabe bat gehitu dugu honen inguruko beharra argitzeko.

Lan gehienetan arazo horri ebazpenak bilatzen saiatzen dira. Beherago aurkeztuko ditugun Hearst (1991) eta Yarowsky (1995), adibidez, hitz etiketatu urrietatik ikasten ahalegintzen dira, eskuzko lana gutxitu ahal izateko. Schütze-k beste bide bat bilatzen du, eta lehenbizi hitzaren agerpenak automatikoki multzokatu, eta gero adiera-etiketa jartzen die, multzo guztiari batera. Gale, Church eta Yarowsky-k, bestalde, adiera-zerrenden hurbilpena alde batera utzi eta adierak beste era batera espezifikatzea proposatzen dute: hitz batek adiera ezberdinak ditu beste hizkuntza baten itzulpen ezberdinak baditu. Honen abantaila zera da, corpus elebidunetatik automatikoki etiketatu daitezkeela hitzen adierak (itzulpen ezberdinak) eskuzko lana erabat saihestuz. Tamalez, gaur egun corpus elebidunak oso urriak eta gai espezifikoei buruzkoak dira.

Erabiltzen den informazioaren inguruan, bi korronte nagusi egon dira: alde batetik pista sintaktiko eta kolokazioak soilik erabiltzen dituztenak (Hearst, 1991), eta bestetik erlazio sintagmatikoan oinarritutako erlazio-izaera soilik erabili izan dituztenak (Gale et al., 1992; 1993; Schütze 1992a; 1992b; ikus III.A.3 atala). Lehenbizikoen kasuan, hitzaren adiera bakoitza agertzen den testuinguru sintaktikoa aztertu ohi da, eta bigarrenean bai esaldi eta bai diskurtsoan zein hitzekin azaldu ohi den. Inplementatzeko orduan, lehenbizikoan adieraren inguruko ±2 zabalerako leihoan dauden hitz eta kategoriak hartzen dituzte kontuan (kolokazio eta pista sintaktikoak bildu nahian), orden eta posizioari buruzko informazioa kontuan hartuaz, eta bigarrenenean leiho zabalagoetan (±50) azaltzen diren izen, adjektibo eta aditzak, ordena eta posizioari buruzko informazio gabe (erlazio-izaera sintagmatiko lokal eta globala neurtu nahian). Emaitzei dagokionez, bata edo besteaz soilik baliatzen direnak %90 doitasunaren inguruan dabiltza (ikusi 10. taula).

Hearst-ek (1991), adibidez, , analisi sintaktiko oso azaleko batetik abiatu (kategoria-etiketak eta sintagma sinpleen mugak) eta adieraren testuinguruan dauden pista sintaktikoak (hitzaren ezker/eskuinean kategoria zehatz bat egotea, ezker/eskuinean preposizio jakin bat egotea, etab.) aztertzen ditu. Adiera bakoitzarentzat eskuz etiketatuko corpusetik ezaugarri sintaktiko horiek erauzi, eta hitz bat desanbiguatzerakoan ezaugarri horiek hitzaren testuinguruarekin konparatu eta adiera egokiena aukeratzen du. Eskuzko etiketatze lana aurrezte aldera algoritmo lagungarri bat ere aurkezten du.

Nahiz eta ikertzaile askok bere hurbilpenaren onurak defendatu beste batzuek bien beharra aitortzen dute[49]. Horrek estatistika alorreko arazo teorikoak planteatzen ditu, iturri ezberdineko ebidentzia ez independenteak (erlazio-izaera sintagmatikoari dagozkionak, pista sintaktikoak eta kolokazioak) konbinatzeari dagokionean.

Yarowsky-k (1993; 1994; 1995) erlazio sintagmatikoan oinarritutako erlazio-izaera (Gale et al. 1992; 1993) hedatu eta Hearst-en antzeko pista sintaktiko eta kolokazioak ere kontuan hartzen ditu[50], emaitzak hobetuaz. Bi ezagutza iturriak integratzean ez ditu ebidentzia guztiak konbinatuko, eta ebidentzia indartsuena bakarrik aukeratuko du (erabaki-zerrenda edo *decision list* direlakoak erabiliz). 1995ko lanean eskuzko lana gutxitu ahal izateko metodo iteratibo bat ere azaltzen du. Ebidentziak konbinatzeko beste modu bat erabiliaz Towell eta Voorhees-ek (1998) erabaki zerrendak baino sare neuronalak darabiltzate, emaitza onekin. Izenetarako lana adjektibo eta aditzetara zabaltzen dute, baina tamalez kategoria bakoitzeko hitz bakarra probatuz.

Aipatu behar da corpusen inguruko lan gehienetan, hemen aipatutakoak barne, ez direla testu bateko hitz guztiak desanbiguatzen ahalegintzen, eta ebaluazioa hautatuko hitz gutxi batzuen agerpenak erabiliaz egiten dute. Bestalde, bi adiera besterik ez direla bereizten, bata bestearengandik oso ezberdinak eta maiz topiko ezberdinetakoak. Salbuespen bakarra Towell eta Voorhees-en lana da, WordNet-eko adierak erabiltzen dituzte eta. Probatutako hiru hitzen adiera kopurua kontuan hartuta (sei, hiru eta lau) oso emaitza onak lortzen dituztela esan daiteke[51].

Niwa eta Nitta (1994) aipatzen ditugu berriro hemen, corpusen kookurrentzietan oinarritutako bektoreen bidez lortutako emaitzak hiztegietako emaitzekin alderatzen dituzte eta (konparatu 9. eta 10. tauletako doitasunak).

---

[49] Hala dio Schütze k (1992[a]): "*The disambiguation algorithm presented doesn't use any information that is encoded in the order of words and ignores morphology and function words. … Future research has to be done on how the method can be extended to include a wider range of linguistic phenomena*".

[50] Adieren testuinguru sintaktikoaren eredu bezala ezker/eskuinera dauden hitz eta kategoriak soilik hartuko ditu kontuan soilik.

[51] Hala ere ez dituzte WordNet-en dauden adiera guztiak erabiltzen, izan ere erabilitako hitzentzat WordNet-ek 27, 13 eta 29 adiera dauzka eta. Horregatik azaltzen da granularitateari buruzko zutabean ~ ikurra.

	Ezagutza mota				Adierak		Ebaluazioa				Lana eskuz
	Par	Lok	Gl	Sx	Jatorria:	Granularitatea	Kop.	Kat.	Est	Doi	
Gale et al., 1993		X	X		Itzulpena	Domeinua (2 adiera)	6	Izenak	%100	>%90	ez
Schütze, 1992		X	X		Multzokatzea +eskuz	~Homonimia	4x200	Izenak	%100	>%90	dezente
Hearst, 1991				X	Eskuz	Homografo (2 adiera)	4x30	Izenak	%100	~%90	dezente
Yarowsky 1995		X	X	X	Eskuz, itzulpena	~Homonimia (2 adiera)	10x4000	Izenak	%100	~%97	gutxi
Towell & Voorhess, 1998		X	X	X	Eskuz, WordNet	~Polisemia	3x350	Izen, adj. eta aditzak	%100	~%86	dezente
Niwa & Nitta, 1994		X	X		Eskuz, Roget's	Domeinua (2 adiera)	9x20	Izenak	?	~%85	asko

10. taula: corpusetan oinarritutako lanen sinopsia

IV.A.5. Konbinatutako HAD

Atal honetako sistemak corpusetan oinarritutako hedaduratatik datoz batez ere, eta eskuzko desanbiguazioa eta datu urrien arazoa gutxitzea dute helburu, bide batez hitzen adierak definitzeko helduleku sendoago bat ezarriaz. Aurreko kapituluko III.A.4 atalean aurkeztutako lanetaz aparte, Leacock, Chodorow eta Miller-ek (1998) pista sintaktikoak eta erlazio sintagmatikoa ere erabiltzen dituen sistema aurkezten dute. Lan honen berrikuntza, ordea, eskuzko desanbiguazioa saihesteko sistema da: adiera baten adibideak lortzeko, WordNet erabiliaz adiera horren sinonimoak aztertu eta horietako bat momosemikoa bada, orduan sinonimo hori azaltzen den testuinguruak adiera beraren testuinguru bezala jotzen dira.

Emaitzei dagokionez, alde nabariena adiera bereizketa zabal edo finen artean dago berriz ere. Bereizketa zabala egiten dutenen artean %90eko emaitzak lortzera heltzen dira. WordNet-ek egiten dituen bezalako adiera bereizketa finentzat desanbiguatzean, ordea, %40 inguruko doitasuna aipatzen du Resnik-ek. Tartean leudeke Leacock-ek eta lortutako emaitzak (%80 inguru), Towell eta Voorhees-en (1998) kasuan bezala ez baitute adiera WordNet-en xehetasun osoz bereizten.

	Ezagutza				Adierak		Esperimetua				Lana eskuz
	Par	Lok	Gl	Sx	Jatorria	Granularitatea	Kop.	Kat.	Est.	Doi.	
Yarowsky 1992			X		Roget's	Domeinua (2 adiera)	12 x asko	Izenak	%100	~%92	ez
Resnik 1997,	X				WordNet	Polisemia	Asko	Izenak	?	~%40	ez
Hearst & Schütze, 1993			X		WordNet	Domeinua	-	-	-	-	ez
Karov & Edelmann 1996			X		Eskuz	Domeinua (2 adiera)	4x125	Izenak	%100	>%90	ez
Leacock et al., 1998		X	X	X	WordNet	~Polisemia	14 x asko	Denak	%100	~%80	ez

11. taula: konbinatutako lanen sinopsia

IV.B. Ebaluaziorako esperimentuaren diseinua

Azken urteotako salbuespen batzuk kenduta, orain arteko lan gehienek (eta corpusean oinarritutakoen kasuan, guztiek) hitz kopuru mugatu batekin lan egin izan dute. Hori dela eta ez da inolaz frogatu corpusetan oinarritutako adiera desanbiguatzaileak, emaitza onenak eskaintzen dituztenak izanda ere, hitz konkretu batzuk desanbiguatzetik testu orokorrak desanbiguatzera pasatu daitezkeenik. Aipatzekoa da, baita ere, emaitza arrakastatsuenek bi adiera oso bereizi artean besterik ez dutela desanbiguatu izan. Adieren granularitatea hain ezberdina izatean, oso zaila da sistema ezberdinen arteko emaitzak konparatzea.

Arrazoi horiengatik, sistemen artean konparatu ahal izateko, esperimentua horrela diseinatu genuen:

1. Ausaz aukeratutako testu osoak desanbiguatu.
2. Domeinu publikoan dauden testu etiketatuak erabili.
3. Domeinu publikoan dagoen adiera espezifikazioa erabili.

Azken bi puntuak betetzen dituen corpusa aipatu dugu jada: SemCor (ikusi II. atala). Semcor domeinu publikoan dago, eta WordNet-eko adieraz dago etiketatuta. WordNet erabiliko da adiera espezifikazioentzat, eta testu sorta osoak etiketatu beharko dira. HADan WordNet erabiltzearen kontrakoak ere badaude, egiten diren adiera bereizketak xeheegiak direla eta. Hori dela eta, adieraren bi maila definituko ditugu: WordNet-eko adiera bera, eta adieren bereizketa zabalagoa egin ahal izateko, adieren etiketa semantikoa (ikusi IV.C.3 ebaluazioaren atala).

Literaturako lan gehienenetan bezala, izenekin egingo dugu ebaluazioa, baina gure kasuan testuan agertzen diren izen guztiekin. SemCor-eko lau fitxategi aukeratu genituen ausaz (ikusi 12. taula). Fitxategian testu ezberdinetatik hartutako puskak egon daitezke. Fitxategi bakoitza genero ezberdin batekoa suertatu zen: br-a01 delakoa “Press:Reportage” bezala zegoen sailkatua, br-b20 “Press:Editorial”, br-j09 “Learned:Science” eta azkenik br-r05 “Humour” bezala. Fitxategietako izenen %11 ez zegoen WordNet-en. WordNet-en topatutako izenetatik %32 adiera bakarrekoa zen.

SemCor-eko fitxategiak WordNet-eko adieraz etiketatuta daude, eta hortaz automatikoki ebaluatu daiteke desanbiguatzean lortutako emaitza zein den, desanbiguatzailearen erabakia SemCor-en dagoenarekin konparatuaz. Ebaluazioa horrela egiteak adiera zuzen bakarra onartzea dakar, hau da, nahiz eta sistemak aukeratutako adiera eta SemCor-ekoa ia berdinak izan, ebaluazioari dagokionez txartzat joko da.

testuak	hitzak	izenak	WNen dauden izenak	izen monosemikoak
br-a01	2079	564	464	149
br-b20	2153	453	377	128
br-j09	2495	620	586	205
br-r05	2407	457	431	120
guztira	9134	2094	1858 (%89)	602 (32%)

12. taula: esperimentuko testuen datuak

IV.C. HAD Dentsitate Kontzeptuala erabiliaz

Gure hurbilpenean garbi geneukan ontologia eta erlazio-izaera orokor baten gainean oinarritu beharra zegoela, eta WordNet izango zela gure erreferentzia ontologikoa (guzti honen justifikaziorako jo III.D atalera). Hurbilpen honek adieraren definizio sendo bati heltzen dio, eta corpusetan oinarritutako tekniken arazorik ez du, hala nola eskuzko desanbiguazioaren beharrik edo datu urrien arazorik.

HADaren lan teorikoen ildoa jarraituz (ikus kapitulu honetako IV.A.1 atala), ez dugu uste ezagutza mota bakarra nahikoa denik desanbiguazio zorrotza lortzeko, ezta WordNet-en daudenik beharrezko erlazio mota guztiak[52]. Bestalde, WordNet-en oinarritutako Dentsitate Kontzeptuala desanbiguaziorako erabilgarria dela frogatu nahi dugu, eta adiera-desanbiguazio sistema arrakastatsu baten oinarrizko osagaia, erlazio-izaera paradigmatikoa formalizatzen duena.

Kapitulu honetan aztertu nahi dugun hipotesia beraz, zera da: WordNet-eko ezagutza paradigmatikoa baliagarria dela adiera-desanbiguazioan eta Dentsitate Kontzeptuala beste erlazio-izaera paradigmatikoen formalizazioak baino hobeto baliatzen dela horretaz.

IV.C.1. Algoritmoa

Adiera-desanbiguatzailearen sarrera SemCor-eko bertako testuak direnez, lehenbizi testu horien garbiketa egin behar da: lematizatu, izenak ez direnak bota eta WordNet-en ez dauden izenak baztertu. Bai lema eta bai kategoria jakiteko SemCor-en bertan dagoen informazioa erabiltzen da. 11. irudian azaltzen da esaldi baten adibide bat. Esaldi horretatik WordNet-eko izenak ez direnak ezabatu eta lemak bakarrik utziz gero, irudiaren azpiko bost izenak gelditzen dira, algoritmoak desanbiguatu beharko dituenak hain zuzen ere.

[52] Berdina uste dugu, aurreko kapituluan aipatu bezala, corpus eta hiztegietatik erauzi daitekeen informazioari buruz.

```
The jury(2) praised the administration(3) and operation(8) of the Atlanta
Police_Department(1), the Fulton_Tax_Commissioner_'s_Office, the Bellwood and
Alpharetta prison_farms(1), Grady_Hospital and the Fulton_Health_Department.

<s>
<wd>jury</wd><sn>[noun.group.0]</sn><tag>NN</tag>
<wd>administration</wd><sn>[noun.act.0]</sn><tag>NN</tag>
<wd>operation</wd><sn>[noun.state.0]</sn><tag>NN</tag>
<wd>Police_Department</wd><sn>[noun.group.0]</sn><tag>NN</tag>
<wd>prison_farms</wd><mwd>prison_farm</mwd><msn>[noun.artifact.0]</msn>↵
      <tag>NN</tag>
</s>
```

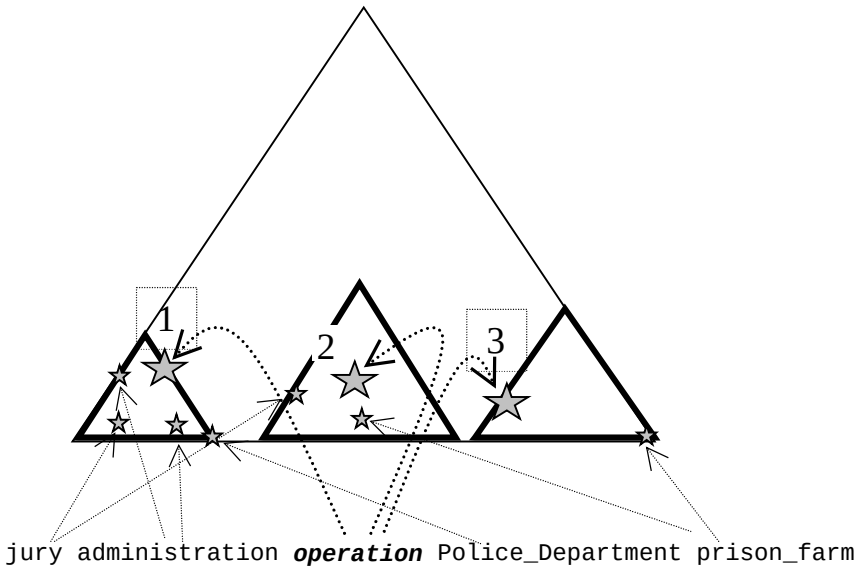
```
jury administration operation Police_Department prison_farm
```

11. irudia: SemCor formatua eta algoritmoaren sarrera

Hitzak desanbiguatzean III. kapituluan definitutako erlazio-izaera erabiliko dugu, Dentsitate Kontzeptuala. Demagun 11. irudiko *operation* hitza desanbiguatu nahi dugula, bere testuinguruari gehien lotzen zaion adiera aukeratuaz. Horretarako *operation*-en adiera bakoitzeko, bere testuinguruan dauden izenentzat (hobe esanda, izen horien adierentzat, arrastoentzat[53]) Dentsitate Kontzeptuala kalkulatu, eta Dentsitate handiena lortzen duen adiera aukeratuko dugu. Hobeto esanda, azpizuhaitz bakoitzak duen Dentsitatea kalkulatzen da, eta desanbiguatu behar den hitzaren adiera bat duen Dentsitate handieneko azpizuhaitza aukeratuko da.

Adibidez, demagun 11. irudiko 5 izenen adierak 12. irudiko izartxoak direla. *Operation* hitzak 3 adiera ditu, 1, 2 eta 3. Adiera horiek WordNet-en hierarkian (hiruki zabalena) kokatu eta hierarkiako azpizuhaitz guztien Dentsitatea kalkulatu ondoren, *operation*-en adiera bat duten zuhaitzen artean Dentsitate handienekoak 12. irudian lodiz azaltzen diren hirukiak direla ateratzen zaigu. Itxuraz Dentsitate handienekoa ezkerreko azpizuhaitzak duenez *operation*-en lehenbiziko adiera aukeratuko luke algoritmoak. Benetako algoritmoa, orain ikusiko dugun bezala, zertxobait konplexuagoa da.

[53] Arrasto deitzen genien beren arteko erlazio-izaera neurtu nahi genuen adierei (ikus III.B.2 atala).



12. irudia: izen baten desanbiguazioa Dentsitate Kontzeptuala erabiliaz. Adieren kopuru eta kokapena asmatutakoak dira.

Programak egiten duen lehenbiziko gauza leiho bat definitzea da: leiho horren erdian dagoen izena da desanbiguatu beharrekoa, eta gainontzeko izenak testuingurua. Behin leiho-zabalera jakin bat emanda, programak leihoa ezkerretik eskuinera mugituko du, izen bat mugimendu bakoitzean. Erdiko izen hori desanbiguatzeko algoritmoaren pseudokodea 13. irudian azaltzen da.

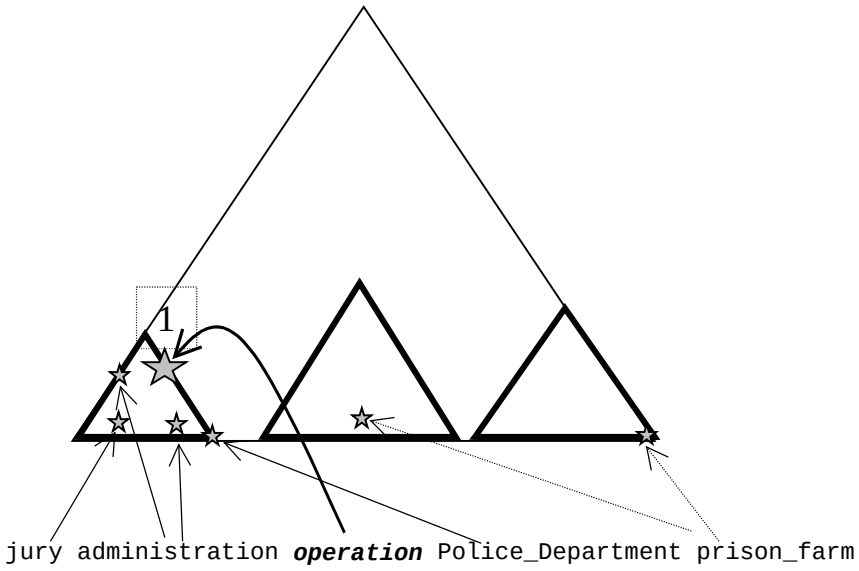
```
(1)   hierarkia := proiekzioa (hierarkia_osoa, izena, testuingurua)
      bigizta
(2)      hierarkia := konputatu_DK(hierarkia)
(3)      azpizuhaitza := DK_handieneko_azpizuhaitza(hierarkia)
         baldin azpizuhaitza = hutsa orduan atera bigiztatik
(4)      hierarkia := markatu_desanbiguatuak(hierarkia,azpizuhaitza)
      ambigizta
(5)   adiera := aukeratu_adiera(izena, hierarkia)
```

13. irudia: izen bat desanbiguatzeko algoritmoa.

Lehenbiziko pausuan, testuinguruko izenen eta desanbiguatu nahi den izenaren adiera eta hiperonimoekin hierarkia bat eraikitzen da (WordNet-en azpimultzo bat). Ondoren, 2. pausuan, Dentsitate Kontzeptuala konputatzen da WordNet-eko azpizuhaitz bakoitzarentzat, bakoitzak duen arrasto kopuruaren arabera (20. ekuazioko a_Z, ikus III. kapituluko 19. ekuazioa ere). Horretarako nahikoa da eraiki berri den hierarkiako azpizuhaitzen Dentsitatea kalkulatzea, WordNet-eko gainontzeko azpizuhaitzek, arrastorik ez dutenez, Dentsitatea 0 izango dute eta. Dentsitate handieneko zuhaitza 3. pausuan aukeratzen da. 4.ean azpizuhaitz horretan zeuden adierak landutzat hartzen dira, Dentsitate Kontzeptual handiena beraien artekoa da eta. Adiera horien hitzak landuta daudenez, hitz horien beste adierak baztertu eta hierarkiatik ezabatu egiten dira.

$$\text{dentsitate}(Z, A) = \text{dentsitate}(Z, a_Z) \quad \text{non } a_Z = |Z \cap A| \tag{20}$$

12. irudiko azpizuhaitzetan dentsoena ezkerrekoa balitz, orduan azpizuhaitz horretako adierak hautatu dira eta dagozkien hitzak landutzat hartzen dira, beraien gainontzeko adierak baztertuaz (ikus 14. irudia): adibideko *jury*, *administration*, *police_department* eta *operation* bera hain zuzen ere (gezi sendoez markatuta daudenak). Horietatik hiru guztiz desanbiguatuta daude, adiera bakarra baitute azpizuhaitz dentsoenean, baina *operation*-ek bi adiera ditu. Landu gabe dagoen hitz bakarra *prison_farm* da.



14. irudia: izen baten desanbiguazioa Dentsitate Kontzeptuala erabiliaz. Adieren kopuru eta kokapena asmatutakoak dira.

Adierak azpizuhaitz horretan ez dituztela eta, hitz batzuk oraindik landu gabe egon daitezke, eta beraz hierarkiako Dentsitate Kontzeptual handieneko hurrengo azpizuhaitza aukeratzeko beste bigizta bat behar da. Hitz guztiak landu direnean (edo ezin denean ezer gehiago landu, adibidez 14. irudiko *prison_farm*-en kasuan) orduan bigiztatik atera eta 5. pausuan desanbiguatu nahi genuen hitzari dagokion emaitza bueltatuko da. Hiru aukera daude emaitzari dagokionez:

1. Adiera bat aukeratzea, azpizuhaitzean adiera bakarra zegoen eta.
2. Adiera anitz aukeratzea, azpizuhaitzean hitzaren adiera bat baino gehiago zegoen eta.
3. Adierarik ez aukeratzea, adiera guztiak isolatuta zeudelako.

14. irudiko hitzentzat honako litzateke emaitza: *jury*, *operation* eta *police_department* guztiz desanbiguatu dira, adiera bakarra baitute. Landuta baino desanbiguatu gabe gelditzen da *administration*, bi adieraz, eta landu gabe *prison_farms*, bi adierarekin ere.

Ikusten den bezala leihoaren erdiko izena desanbiguatzerakoan leihoko gainontzeko izenak ere desanbiguatzen dira. Izan ere algoritmo hau testua zatika desanbiguatzeko diseinatua dago. Tamalez SemCor-en ez dago paragrafo edo antzeko zatiketen adierazlerik, eta horregatik hautatu genuen hitzak banan-bana desanbiguatzea esperimentu honetan.

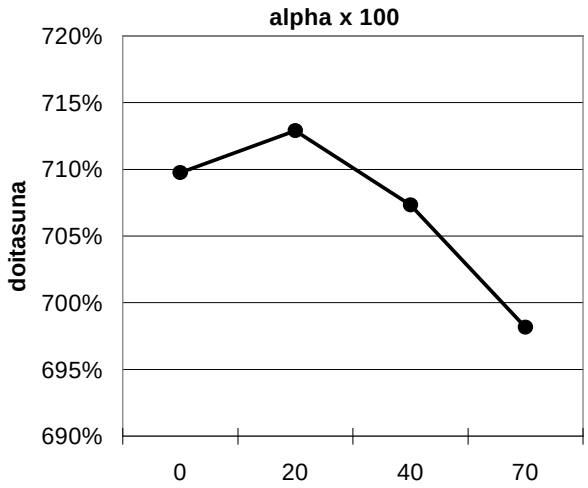
IV.C.2. Dentsitate Kontzeptualaren aldaeren ebaluazioa

Esperimentuen emaitzak lau fitxategietan lortutakoaren batezbestekoa kalkulatuz ematen dira. Desanbiguazioa ez denean erabatekoa, hau da, adiera bat baino gehiago aukeratu direnean, desanbiguatu ez balu bezala hartuko dugu. Bi neurri erabiliko ditugu ebaluazioan: doitasuna (desanbiguatutako izenetatik ondo daudenen ehunekoa) eta estaldura (izen guztietatik zenbat desanbiguatu izan diren, ehunekotan), beti ere hitz polisemikoentzat kalkulatuta. Taula gehienetan emaitzak leihoaren zabaleraren arabera eman ohi dira, zabalera izenen arabera kalkulatuta egonik.

IV.C.2.a) Parametroa: α

Leihoaren tamaina handitzen denean, azpizuhaitz jakin baten azpian dagoen arrasto kopurua azpizuhaitzaren altuera baino dezente handiagoa izan daiteke. Horrelakotan III. kapituluko 18. formulako zatikizuna gehiegi handitu daiteke. Efektu hori leuntzeko parametro bat gehitu genion formulari, enpirikoki aztertu eta balioa bilatu dioguna (ikus III. kapituluko 19. ekuazioa).

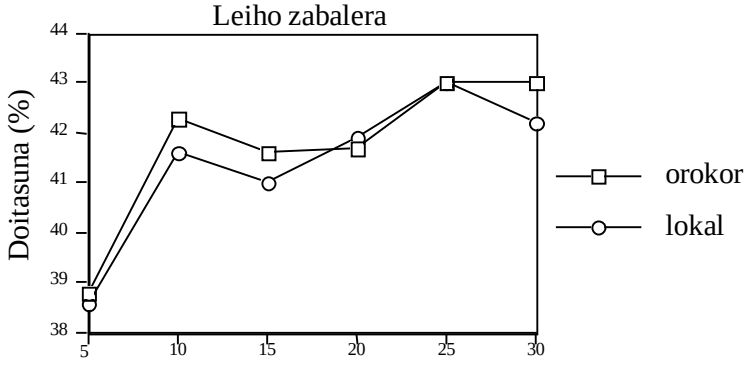
Parametroaren balio onena bilatzeko esperimentu sorta zabala egin genuen, hainbat leiho-zabalera eta testuren gainean, α-ren 0tik 1erako aukerak probatuz. 15. irudian egindako 20 esperimentutan batutako doitasunak azaltzen dira. Bai grafiko honetan, eta baita egindako beste esperimentu batzuetan ere, garbi ikusten da desanbiguazio emaitza hoberenak α 0,2 baliotik hurbil dagoenean lortzen direla, nahiz eta aldea txikia izan: 15. irudian, adibidez 0,2 eta 0,4 artean dagoen aldea 20 esperimentuetan %5 ingurukoa da, baina esperimentu bakoitzerako batez-beste %0,25koa da bakarrik. Beraz, aurrerantzean azaltzen diren esperimentu guztietan α-k 0,2 balioa izango du.



15. irudia: α parametroaren balioen araberako doitasuna.

IV.C.2.b) Nola kalkulatu μ_Z

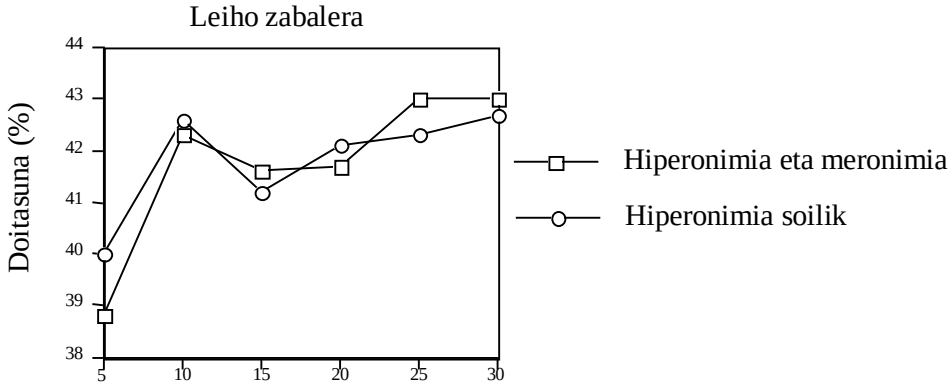
Aurreko kapituluan azaldu bezala umeen batezbestekoa azpizuhaitz bakoitzerako erabili ordez (μ_Z lokala), WordNet ontologia guztirako batezbestekoa (μ_{WN} orokorra) erabili daiteke. Honen eragina zein den aztertzeko hainbat esperimentu egin genituen leiho zabalera ezberdinak kontuan hartuaz (ikusi 16. irudia) eta μ_Z lokala erabiliaz doitasuna ozta-ozta hobetzen dela ondorioztatu genuen. Kontuan hartuz μ_{WN} orokorra erabiltzea eraginkorragoa dela, berau erabili genuen gainontzeko esperimentuetan.



16. irudia: μ_Z lokala edo μ_{WN} orokorra

IV.C.2.c) WordNet-eko beste erlazioak: meronimia

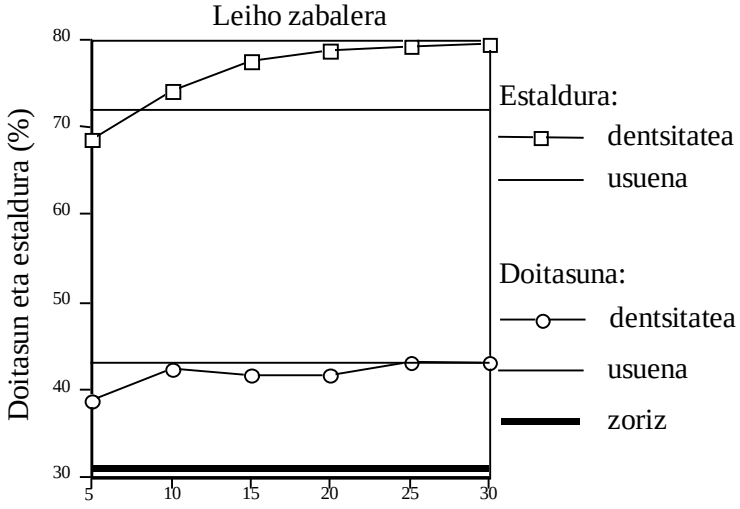
Enpirikoki aztertu dugu Dentsitate Kontzeptualak meronimia erlazioak ere erabiltzean desanbiguatzailearen doitasunean eraginik daukan edo ez. Emaitzen arabera (17. irudia), doitasuna antzekoa dela ondorioztatu daiteke, baina estaldura %3 altxatzen da, eta beraz meronimia erlazioak erabiltzea erabaki genuen.



17. irudia: meronimia erabiltzearen eragina

IV.C.3. Ebaluazioa

Orain arte faktore jakin batzuk aztertu ditugu, onuragarriak diren edo ez erabakitzeko. Orain emaitza orokorrak aztertuko ditugu. 18. irudian azaltzen den bezala, testuinguruaren leihoa zabaltzen den heinean estaldura handitu egiten da. %80 inguruan egonkortzen da, 20 izen baino leiho zabalagoetarako hobekuntza gutxi jasoz. Doitasuna, bestalde, %43raino igotzen da leihoa zabaldu ahala.



18. irudia: doitasuna eta estaldura

Irudian bi *baseline* azaltzen dira: zorizkoa eta SemCor-eko izenen adiera usuenak aukeratzekoa. Lehenbiziko kasuan doitasuna (%30 inguru) analitikoki kalkulatu genuen, testuetako hitz polisemikoen adiera kopuruak erabiliz. Ondoren enpirikoki baieztatu genuen, 10 aldiz egikarituaz adierak ausaz aukeratzen zuen programa. Estaldura, noski, %100eko litzateke. Adiera usuenak aukeratzeko beharrezkoa da eskuz desanbiguatutako materiala edukitzea. Adieren maiztasunak kontatzeko SemCor bera erabili genuen, aukeratutako lau testuak kontuan hartu barik. Doitasuna Dentsitate Kontzeptualaren berdina da, baina estaldura %8 apalagoa.

Leiho hoberenarentzako datuak 13. taulan azaltzen dira. Oraingoz adiera mailako doitasunari buruz hitz egin dugu, aurrerago ikusiko dugu fitxategi mailako doitasunaren esanahia. Irudietan azaltzen diren datuak izen polisemikoentzat dira bakarrik, baina adiera bakarreko izenak ere kontuan hartzen baditugu doitasuna %64,5era eta estaldura %86,2ra iristen dira.

leihoa=30		Estaldura	Doitasuna
polisemikoak	adiera	%79,6	%43,0
	fitxategi		%53,9
guztira	adiera	%86,2	%64,5
	fitxategi		%71,2

13. taula: leiho hoberenarentzako datuak

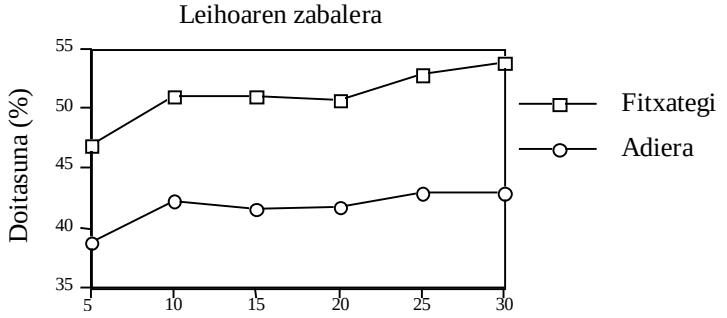
Ebaluazio orokorraz gain, emaitzak beste ikuspuntu batzuetatik ere aurkeztu ditugu.

IV.C.3.a) Desanbiguazio maila: adiera edo fitxategia

WordNet-ek adierak lexikografoen fitxategietan multzokatzen ditu. Izenen kasuan fitxategi horiek domeinuaren inguruko irizpidez egituratu dira, eta 25 daude. Gure algoritmoak izenari adiera egokia bilatzeaz gain fitxategia ere esleitzen dio. Gure ustez granularitate maila biak dira interesgarriak, WordNet adiera bereizketak xeheegiak direla esan izan baita, eta fitxategi mailako bereizketa zabalagoa da.

Homografo edo domeinu mailan beste algoritmoek eman izan dituzten bereizketak baino xeheagoa da hala ere, adibide batez ikusiko dugun bezala. Yarowsky-k (1992) *bass* izenarentzat bi adiera bereizten zituen: *MUSIC* bezala etiketatzen zuen bata, eta *ANIMAL* bezala bestea. WordNet-en aldiz 9 adiera bereizten dira. Musikari buruzko 6 adierak 4 fitxategitan daude banatuta: *ARTIFACT*, *ATTRIBUTE*, *COMMUNICATION* eta *PERSON*. Animaliei buruzko 3 adierak aldiz bi fitxategitan azaltzen dira: *ANIMAL* eta *FOOD*. Yarowsky-k 2 adiera bereizten zituen lekuan, WordNet-en fitxategi mailan 6 leudeke, eta adiera mailan 9.

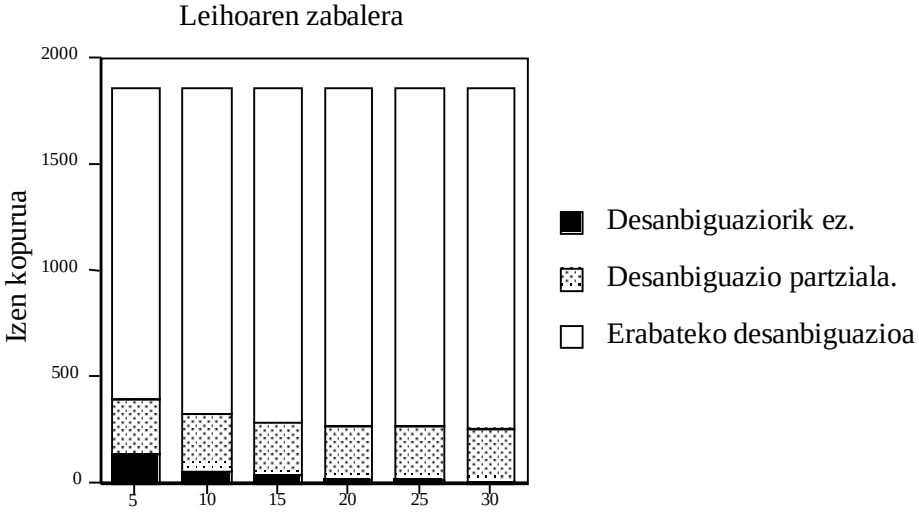
Fitxategi eta adiera mailako emaitzak konparatzeko jo 13. taulara eta 19. irudira.



19. irudia: adiera eta fitxategi mailako emaitzak

IV.C.3.b) Desanbiguazio partziala

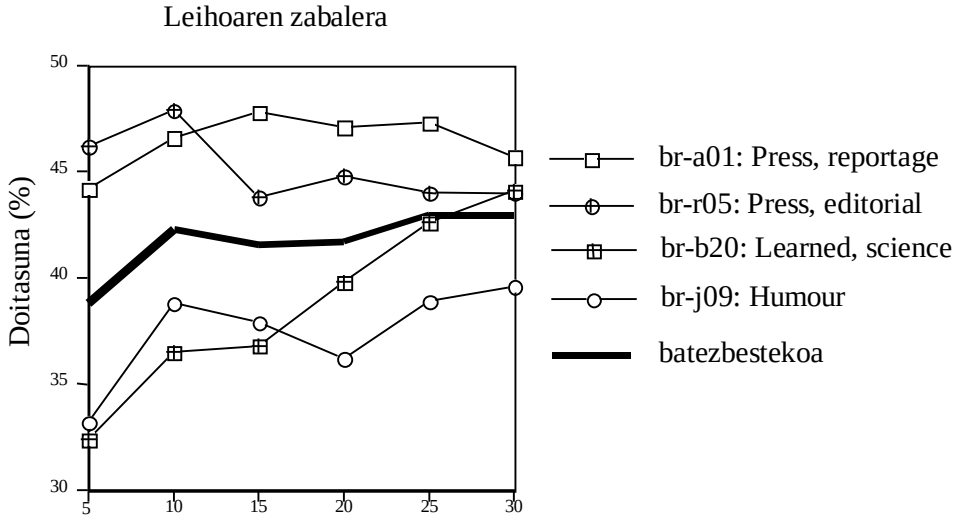
Aipatu izan dugun bezala gure algoritmoak adiera bakarra aukeratzeaz gain badauzka beste bi aukera: adiera multzo bat aukeratzea edo bat ere ez aukeratzea. Adiera multzo bat aukeratu izan duenean, orain arteko emaitza guztietan desanbiguatu izan ez balu bezala hartu dugu. 20. irudian ikusten den bezala, algoritmoaren estaldura %80koa dela esan dugunean, ez ditugu hartu kontuan partzialki desanbiguatutakoak. Leiho zabalenaren kasuan partzialki desanbiguatutakoak kontuan hartu izanez gero, %100eko estaldura izango genuke.



20. irudia: desanbiguazio partziala

IV.C.3.c) Testuinguruaren zabaleraren eragina

Lehenago azaldutako datuetan desanbiguatu izan diren lau fitxategietarako lortutako emaitzen batezbestekoa erabili da. Horrela, testuingurua zabaldu ahala doitasunaren emaitzak hobetzen direla ikusi dugu. Hala ere, fitxategi horien egitura eta topikoa hain ezberdinak izanda, ez genuen espero ezaugarri hori guztientzat beteko zenik. Eta hala gertatzen da, 21. irudian ikusten den bezala.



21. irudia: testuinguruaren zabaleraren eragina testu fitxategietan

Testu-fitxategi bakoitzaren portaera ezberdinaren arrazoia, testuen domeinu ezberdinaz gain, SemCor-en akats batean ere egon daiteke. Izan ere fitxategiak esaldi segida bezala daude egituratuak, titulu, atal, paragrafo edo testuaren iturriaren aldaketa adierazi barik. Izenak diskurtso berean gertatzen diren jakin gabe pilatzen du testuingurua algoritmoak. Horrek argi lezake elkarrizketa oso motzez osatutako br-r05 fitxategian doitasun hoberena 10 izeneko zabalerarekin lortzea, edo prentsako editorialak dauzkan br-b20 fitxategian testuinguru txikiekin emaitza askoz okerragoak lortzea.

IV.D. Konparazioa beste metodoekin

Dentsitate Kontzeptualaren emaitzak beste lanekin konparatzea zaila da. Emaitzan izugarri eragiten duten faktoreak ezberdinak izaten dira lan batetik bestera: adiera bereizketaren iturri eta granularitatea, zenbat hitzekin probatu izan den (hitz multzotxoa edo testu errealetako hitz guztiak), zein kategoriekin probatu den, ontzat emateko irizpideak (emaitza partzialak, adiera bat baino gehiago ontzat emateko aukera), ebaluaziorako neurri ezberdinak, etab. Bestalde desanbiguatzaile batzuk *standalone* sistema (desanbiguatzeko beharrezko guztia dakiena) bezala aurkezten diren bitartean, beste algoritmo batzuk, gurea barne, ezagutza iturri osagarriekin integratu beharko liratekeen azpisistema bezala planteatu izan dira.

Faktore horiek kontuan eduki gabe, ezin da emaitzen arteko konparazio hutsa egin. Gure sistemak zailtasun handieneko ebaluazioari egiten dio aurre: adiera bereizketa xeheak, testu errealetako izen guztiak, emaitza partzialak baztertu eta adiera bakarra ontzat eman. Horregatik gure sistemaren doitasuna (%43) ezin da, adibidez, Yarowsky-ren 1995.ekoaren parean jarri (%97).

Sistemen arteko konparazioa, beraz, ez dugu horrela egingo. Konparaziorako hautatu dugun bidea zera izan da, WordNet-eko ezagutza erabili izan duten edo erabiltzera egokitu daitezkeen sistemak gure testuen gainean lanean jarri eta emaitzak konparatzea. Horretaz gain WordNet erabili izan duten bestelako lan batzuk ere laburki aipatuko ditugu.

Gure lanarengandik hurbilen dagoena Sussna-rena (1993) da (ikusi III.A.1 atala eta kapitulu honetako IV.A.2 atala). III. kapituluko III.D atalean ere aipatu ditugu Dentsitate Kontzeptualak Distantzia Kontzeptualarekin alderatuta dauzkan abantailak. Sussna-k, gure antzera, domeinu publikoko corpus bateko testu batzuetako izenak desanbiguatu zituen. Guk ez bezala, adiera bat baino gehiago onartzen zituen, eta adiera egokia topatzen ez zuenean ebaluaziotik baztertu egiten zuen. Binakako Distantzia Kontzeptuala erabiltzeak dakarren leherketa konbinatorioa saihesteko testuinguruaren leihoa 10 izenetara mugatu beharra zeukan[54], edo testuingurua zabaldu nahi izanez gero, behin izen bat desanbiguatu ondoren hurrengo izenen testuinguruak aukeratutako adiera hori erabili behar zuten testuinguruan (adierak geldiaraztea deitu zion honi, *freezing*). Hori dela eta, adiera bat gaizki aukeratuz gero, hurrengo izenetarako erabakia okertu zitekeen. Adiera bakarra aukeratzera ere derrigortzen du honek, eta horretarako, adiera bat baino gehiago egokiak direnean, bat ausaz aukeratu beharra dauka.

Sussna-ren arabera emaitza hoberenak ematen zituen algoritmoa inplementatu dugu, garrantzi gutxiko faktoreak alde batera utzita[55]. Emaitzak konparatu ahal izateko Dentsitate Kontzeptualak adiera bat baino gehiago hautatzen duenean ausaz aukeratzera behartu dugu. 14. taulan ikus daitekeen bezala, Dentsitateak doitasun hobeagoa lortzen du. Sussna-k bere lanean aurkeztutako emaitzetan izen monosemikoak kontuan hartuz gero %63,4ko emaitzak azaltzen ditu, hemengoak baina %10 hobeagoak. Izen guztientzat erabaki beharra eta adiera on bakarra existitzeak eragina eduki dute, ziur aski, bere artikulu eta gure esperimentuaren arteko aldean.

		Estaldura	Doitasuna[56]
Sussna, 1993	Adiera	%100	%52,3
	Fitxategia		%64,5
Dentsitatea (leihoa=30)	Adiera	%100	%60,1
	Fitxategia		%70,1

14. taula: Sussna (1993) eta Dentsitatea

[54] Bere algoritmoa br-r05-eko lehenbiziko 10 izenekin erabiltzean, 200.000 adiera pareren arteko distantzia kalkulatu izan behar genuen.

[55] Hasierako 10 hitzen adierak batera aukeratu, eta hortik aurrera adierak geldiarazten ditu 41 izenetako leihoak erabiliaz. Meronimia-erlazioak ere erabili dira,eta erlazio guztiak pisu bera eduki dute.

[56] Taula honetan doitasuna izen guztientzat ematen da, hau da, adiera bakarrekoak ere kontuan hartuta.

Aipatu beharra dago gureari lotutako lan bat, (Resnik, 1997) (ikus III.A.4 atala, eta kapitulu honetako IV.A.5 atala). Erlazio-izaera erlazionatutako izen multzoen gainean probatu zuen, printzipioz errazagoa dirudien lana, WordNet-eko adiera xeheak erabiliaz. Berak gutxi gora bera %40ko doitasuna aipatzen du, gure %43 baino %3 apalagoa.

Eskuzko lanik behar ez duen lan arrakastatsu bat Yarowsky-k (1992) egindakoa da. Ontologia bat (*Roget's thesaurus*) eta corpus bat (*Grollier's Enciclopedia*) konbinatu zituen, corpus bereko 12 izenei ontologiako klase marka jartzeko asmoarekin. Bere algoritmoa (ikusi III.A.4 atala ere) inplementatzeko ontologiako klase bezala WordNet-eko lexikografoen fitxategiak erabili ditugu. Bere algoritmoaren emaitzak eta fitxategi mailan Dentsitateak lortutakoak 15. taulan azaltzen dira. Dentsitatearen doitasuna %7 garaiagoa da. Nahiz eta zorizko hautaketaren bitartez Dentsitatearen estaldura %100era igo (ikusi 14. taula) Dentsitatearen doitasunak hobeagoa izaten jarraitzen du (%70,2).

	Estaldura	Doitasuna
Yarowsky, 1992	%100,0	%64,0
Dentsitatea	%86,2	%71,2

15. taula: Yarowsky (1992) eta Dentsitatea

WordNet-en oinarritutako adiera bereizketa darabilten bi lan (Leacock et al., 1998) eta (Towell & Voorhees, 1998) dira, baina adiera bereizketa ez da hain xehea[57]. Biak daude batez ere corpusetan oinarrituta; bigarrenak bakarrik darabil WordNet-eko ezagutza, nahiz eta sinonimia erlaziora bakarrik mugatu.

IV.E. Ekarpena

Kapitulu honetan WordNet-eko ezagutza paradigmatikoa darabilen Dentsitate Kontzeptualean oinarritutako desanbiguatzailea eraiki eta probatu dugu. Emaitzen arabera Dentsitate Kontzeptuala HADrako erabilgarria dela frogatu dugu, eta WordNet-eko ezagutzaz erlazio-izaera paradigmatikoen beste formalizazioak baino hobeto baliatzen dela erakutsi ere bai. Hurbilpen honek adieraren definizio sendo bati heltzen dio, eta corpusetan oinarritutako tekniken arazorik ez du, hala nola eskuzko desanbiguazioaren beharrik edo datu urrien arazorik.

Gure sistemaren onurak:

- Ontologia bateko adieratara lotzen ditu testu errealetako izenak.

[57] Beraien lanean erabilitako izen bakarrarentzat (*line*) 6 adiera bereizten dituzte. WordNet-ek bereizten dituen adierak 27 dira.

- Oinarri teoriko sendoak dituen erlazio-izaera erabiltzen du.
- Emaitza onak, nahiz eta testu zailetan probatu dugun.
- Edozein domeinutan erabil daiteke inongo egokitzapen beharrik gabe.
- Konplexutasun aldetik erakargarria.
- Ez du eskuzko desanbiguazioaren beharrik.
- Ez du datu urrien arazorik.

HADaren literaturan azaltzen diren esperimentuekin alderatzean gure esperimentuak arazoaren alde zailenari egin dio aurre: adiera bereizketa xeheak, testu errealetako izen guztiak, emaitza partzialak baztertu eta adiera bakarra ontzat eman. Enpirikoki frogatu dugu Sussna (1993) eta Yarowsky (1992) baina emaitza hobeagoak lortzen dituela testu sail berdinean, ausaz aukeratu izan diren lau testu zabal erabiliaz (guztira 10.000 hitz). Testuak ez ziren inolaz ere errazak desanbiguatzeko. Adibidez, horietako bat humorezko elkarrizketa motzez osatua zegoen. Hala ere, WordNet-eko adiera finetarako desanbiguatzean %64ko doitasuna lortzen dugu, eta fitxategi-mailan desanbiguatzen badugu %71koa. Estaldura oso zabala da, testuetako izenen %86 desanbiguatzen dugu eta.

IV.F. Etorkizunerako lana

Egindako esperimentuetan baziren hobetu zitezkeen alor batzuk.

- Diskurtso-egituraren araberako testu zatiak batera desanbiguatu. SemCor corpusak, tamalez, ez du paragrafo markarik, eta ezinezkoa izan da testuak diskurtsoaren egituraren arabera zatitzea. Horrela egin izanez gero hitzak bakarka desanbiguatu ordez testu zati oso bat batera desanbiguatu zitekeen, eraginkortasun hobeagoa lortuaz. Gainera doitasuna ere hobetuko litzateke, zerikusirik ez daukaten testu zatiak alde batera utziko ziren eta.
- Dentsitatearen neurri eta adiera-aukeraketaren artean koerlaziorik ote dagoen ikertzea interesgarria izango litzateke. Koerlazio bat balego Dentsitatearen balio batetik behera daudenak desanbiguatu gabe utzi eta doitasuna hobetuko litzateke (estaldura gutxitzearen truke).

Desanbiguazioaren emaitzei dagokionean, emaitza hobeagoak lortzeko informazio iturri berriak gehitzea ezinbestekoa da. Gure ustez Dentsitate Kontzeptualak WordNet-ek duen ezagumendu paradigmatikoa ezin hobeto ustiatzen du, baina horrekin ez da nahikoa. Kapitulu honetako IV.A.1

atalean ikusi dugun bezala bestelako ezagutza iturriak ere erabili behar dira. Bi sailetan bereiziko genuke ezagumendu hori:

- Dentsitate Kontzeptualean integratu beharreko ezagumendua, WordNet aberastuaz lortuko zena (ikus III. kapituluko etorkizunerako lanari buruzko atala).
- Desanbiguazioan erabilgarriak diren bestelako ezagutza. Honen adibideak dira, adibidez, adieren maiztasunak, bai orokorrean edo desanbiguatzen ari garen testuan, ea adiera bat kolokazio modura azaltzen den, adiera bakoitzaren inguruan dagoen egitura sintaktikoari buruzko informazioa (pista sintaktikoak, ikus Leacock et al. 1998), eta abar.

Horrela adiera-desanbiguaziorako sistema osoago bat eraikiko genuke, Dentsitate Kontzeptualaren bidez informazio lexikal-semantikoa kodetzen duena eta hau beste ezagutzarekin konbinatzeko gai dena.

Tesi hau idazten ari garen bitartean, SENSEVAL txapelketa[58] gertatzen ari da. Mundu mailan adiera desanbiguatzen duten sistema hoberenek parte hartzen dute bertan. Yarowsky-ren (1995) lana, Dentsitate Kontzeptuala eta hiztegietan oinarritutako bestelako erlazio-izaeraren neurriak (VI. kapituluan ere azalduko zaizkigunak) integratzen saiatzen ari gara txapelketa horretarako.

Aurreko kapituluan aipatu dugun bezala, adiera-desanbiguaziorako algoritmoaren inplementazio azkarrago bat lantzen ari gara, UNED-eko Elektrizitate eta Elektronika saileko ikerkuntza taldearekin batera, ITEM[59] proiektuaren barruan. Bertsio hau ingeniaritza linguistikorako GATE[60] ingurunearen barruan (Cunningham et al. 1997) integratuta egongo da laster.

[58] http://www.itri.bton.ac.uk/events/senseval/cfp2.html

[59] http://sensei.ieec.uned.es/item/

[60] http://www.dcs.shef.ac.uk/research/groups/nlp/gate/

V. Kapitulua

TESTU-ZUZENKETA AUTOMATIKOA

Kapitulu honetan Dentsitate Kontzeptuala beste alor baten aplikatzen saiatuko gara, testu-zuzenketa automatikoan. Gure ikerkuntza taldeak testu-zuzenketan egindako ikerkuntzan tradizioa badu, eta zuzenketa proposamenak automatikoki aukeratzea zein puntutaraino posible den edo ez aztertu nahi izan dugu. Horretarako ez dugu Dentsitate Kontzeptuala bakarrik erabiliko, ezagutza sintaktikoa ere guztiz beharrezkoa baita. Lehenbizi, V.A. atalean, testu-zuzenketa alorraren sarrera txiki bat eta aurrekarien azterketa egin ditugu. V.B. atalean esperimentua bera egin baino lehenagoko aurre-azterketaren berri eman dugu. Ondorengo atalean esperimentuan erabilitako teknika guztiak labur azaldu, eta, V.D. atalean, esperimentuaren diseinua, emaitzak eta ebaluazioa aipatzen dira. Azkenik kapitulu honetako ekarpenen laburpena eta etorkizunerako lanak aurkezten dira.

V.A. Sarrera eta aurrekariak

Testuetako idazketa-erroreen ordenadore bidezko zuzenketa oraindik ikertzen ari den alorra da. Arazo honen ebazpen idealean, testuan egindako errore guztiak programa batek automatikoki zuzenduko dizkiola espero du erabiltzaileak. Gaur egun ordea, testu-prozesadoreetan (Word, 1997; Ispell, 1993; Aduriz et al. 1997) aurkitzen duguna zuzenketarako laguntza besterik ez da izaten:

- Sakatze edo ortografia erroreak detektatuz. Adibidez:

```
Lehio* bat apurtu dut.
Ukenduaren uzainak* erlea aldendu zuen.
Araso* hau konpontzeko eskatu dut.
```

- Errore horren ebazpen posibleen zerrenda bat emanaz. Adibidez:

```
lehio*: lehia, lesio, leiho
uzaina*: zaina, usaina, uhaina
araso*: eraso, arazo, arasa, arbaso
```

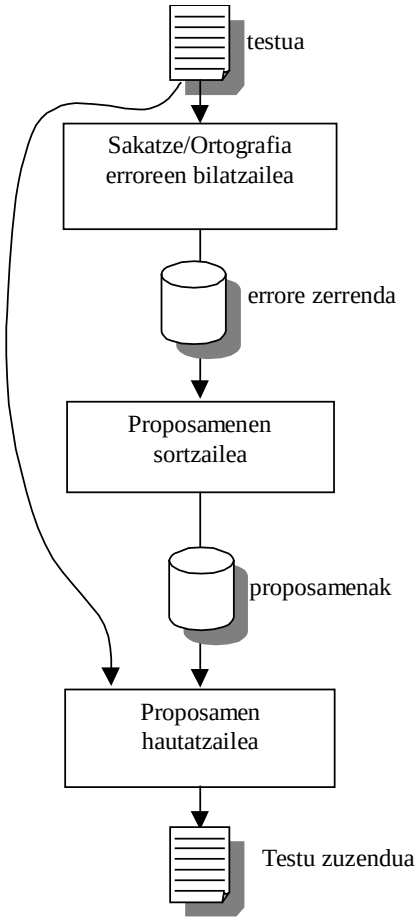
Sistema hauek muga garbi batzuk eduki ohi dituzte (Kukich, 1992):

- ortografia erroreek hitz posible bat sortzen badute, ezin da detektatu. Adibidez, `usaina` idatzi nahi eta sakatze errore baten ondorioz `uhaina` idazten dugunean. Errore mota honi *benetako hitza* errore deituko diogu (*real-word error*), eta kontrakoari, hau da, errorearen erruz idatzitako hitzak ez denean existitzen, *ez-hitza* errore deituko diogu (*non-word error*).

- Proposamen bakarra ezin eman izatea. Hau da zuzenketa guztiz automatikorako (giza-zuzentzailearen parte hartze gabekoa) oztopo garrantzitsuena.

- Errore gramatikalak topatzeko zailtasunak. Adibidez konmuztadura erroreak: `nik izan naiz`. Nahiz eta gramatikaren alor honetan laguntzen saiatu, gaur egungo testu-prozesadoreen laguntza mugatua da oso, alarma faltsu gehiegi eta ezer gutxi zuzentzen dutelako. Testu-prozesadoreak batez ere ez-hitza erroreen zuzenketara mugatzen dira.

Kapitulu honetan azalduko den hobekuntza **proposamen bakarraren** ildotik doa, hau da, sakatze edo ortografia erroreak detektatu ondoren proposatzen diren zuzenketetatik bakarra aukeratzea. 22. irudian zuzentzaile baten eskema ikus daiteke. Guri interesatzen zaigun modulua proposamen-hautatzailearena da: errorearen testuingurua aztertu eta horren arabera testuinguruari hoberen lotzen zaion proposamena aukeratzen du. Esan bezala, beste bi moduluentzat ondo garatutako teknologia badago (Word, 1997; Ispell, 1993; Aduriz et al. 1997), baina proposamen sortzaileek ez diote erreparatzen testuinguruari eta beraz proposamen egokia aukeratzeaK –zuzenketa automatikoaK– irekita dagoen arazo bat izaten jarraitzen du (Kukich, 1992).



22. irudia: proposatutako sistemaren eskema

V.A.1. Aplikazioak eta zuzenketa automatikoaren beharra.

Testuetan aurkitzen diren erroreen iturria, ordea, ez da beti giza errakuntza. Gaur egun testua jatorri ezberdinetatik eskura daiteke: eskanerrak, arkatz optikoetan oinarritutako interfazeak, ahotsaren ezagutzarako gailuak, edo aipatu den giza-erabiltzaileak teklatuetan sakatuta. Jatorri horren arabera tratamendu ezberdina jasoko du testuak: karaktere-ezagutze optikoa (*Optical Character Recognition*) delakoa, idazkeraren ezagutza, ahotsaren tratamendua edo testu-prozesadorea. Errore baten aurrean, hala ere, antzera jokatu beharko dute: errorea dela detektatu, errore horri ebazpen proposamen zerrenda bat proposatu eta ahal dela proposamen bakarra lehenetsi. Sistema horietan guztietan ezin da orokorrean errore zuzenketari buruz hitz egin, batzuentzat egokiagoa baita hitz-ezagutzeaz hitz egitea. Hala ere azken urte hauetan bi arazo mota hauentzat ebazpide amankomunak planteatzen ari direnez, badago bien elkarketa bat (Kukich, 1992). Kapitulu honetan zehar errore-zuzenketari buruz hitz egingo dugu, baina teknika gehienak beste eremuetara ere heda daitezke.

Proposamen bakarrak ahalbideratzen duen zuzenketa automatikoa egin ahal izateko, gaur egun sistema hauek bi eratara jokatzen dute:

1. Hiztegia mugatuz, proposamenen zerrenda motzagoa izan dadin (adibidez, ahotsaren tratamenduan) [61].
2. Erabiltzailearen esku utziz, azken erabakia har dezan (testu-prozesadoreak).

Badira halere zuzenketa automatikoa beharko luketen aplikazioak, adibidez testuak ahoskatzen dituzten sistemak. Hauen kasuan beharrezkoa da, nahiz eta ahoskatu behar duten testuan erroreak egon, testu osoa ahoskatzea. Horretarako inoren laguntza gabe proposamen zuzena bilatu behar da, hiztegia mugatu gabe. Testuekin lan egiten duten edo gizakiarekin elkarrekintza duten sistemen sendotasuna eta komunikatzeko gaitasuna ere nabarmenki hobetuko litzateke. Zuzenketa automatikoak aurrerabide galanta suposatuko luke testu eta programen edizioan, ordenadorez lagundutako argitaratzean, hizkuntzen irakaskuntzan, ordenadorez lagundutako tutoreetan, datu-baseekin elkarrekintzan eta baita ahotsaren erabilera plantea daitekeen beste aplikazioetan ere (Kukich, 1992).

V.A.2. Aurrekariak

Zuzenketarako proposamenen artean zuzena aukeratzea horrela zehaztu daiteke: errorea proposamen bakoitzarekin ordezkatzean sortzen diren esaldi posibleen artean "hoberena" aukeratzea (Mays et al. 1991). Esaldi hoberena zein den erabakitzeko ezagutza-iturri ezberdinetara jo beharko dugu. Ezagutza horien inguruan egingo dugunez bibliografiaren azterketa, ikus dezagun (Mays et al. 1991) eta (Kukich, 1992) lanetan LNParen prozesamendu klasikoaren arabera eginiko iturrien banaketa zein den:

a) Erroreen iturriei buruzko ezagutza
b) Sintaxiari buruzkoa
c) Semantikari buruzkoa

V.A.2.a) Erroreen iturriei buruzko ezagutza

Lehenbizikoa bakarrik erabiltzen denean hitz isolatuen zuzenketa deritzonean (*isolated word correction*), erroreen iturburuak eta erroreak sortzen dituzten faktoreak aztertzen dira (sakatze erroreak, entzumen sistemen erroreak, eskanerren erroreak etab.). Kernighan-ek eta (Kernighan et al. 1990) horrelako sistema bat aurkezten dute. Ebaluazioa bi proposamen dauzkaten erroreentzat egiten da, %87ko doitasuna lortuz. Garai berdinean Kukich-ek (1990) proposamen kopurua edozein izanda ere lan egiten duen sistema aurkezten du, baina hiztegi murriztua behar duena (521 eta 1872 hitz

[61] Ahotsaren tratamenduan entzundako esaldiarentzat interpretazio bakarra behar izaten da. Soinutik abiatuta hitza ezagutzean aukera bat baino gehiago egoten da, eta aukera horien artean bakarra utzi ahal izateko ezagutzen diren hitzen zerrenda mugatzen da.

dituzten bi hiztegirekin egiten ditu saiakerak). Honek, noski, lan honen erabilera erreala zalantzan jarriko luke. Kukich-en emaitzak %75eko doitasunaren inguruan dabiltza. Lan berdinean, errorearen testuinguruaren berri eduki gabe gizakiak zein puntutaraino lor zezakeen hitz zuzena aukeratzea %66 eta %87 artean neurtu zuen. Bere esanetan horrek adieraziko luke ezagutza iturri hau erabiltzen duen sistema baten gehienezko doitasuna. Beranduago Ingels-ek (1997) bere tesian Kernighan-en eta (Kernighan et al. 1990) lana moldatuko du. Egindako esperimentuak bi testu sortari lotuta daude: datu-base baten galdeketa sistemako elkarrizketak eta ordenadoreen eskuliburu bat. Sistema testu horietarako trebatzen du. Lehenbiziko testuan %74ko doitasuna lortzen du, eta bigarrenean, hiztegi zabalagoa daukanean, %54ko doitasuna. Kernighan eta Ingeles-en lan hauen sintaxi-hedapenak ondoren ere aztertuko ditugu.

V.A.2.b) Sintaxia

Sintaxia erabiliko lukeen sistema batek esaldian errorea proposamen bakoitzaz ordezkatzean gelditzen den egitura sintaktikoaren egokitasuna egiaztatuko luke. Egitura onargarriei dagozkien proposamenak bakarrik aukeratuko lirateke, edo egokitasun sintaktikoaren neurri bat erabiliz gero, proposamenak egokitasun-neurri horren arabera ordenatuko lirateke. Funtsean, semantikarekin ere antzekoa gertatuko litzateke, egokitasun kontzeptu ezberdinak erabiliaz.

Sakonago aztertu aurretik iturri bakoitza, LNParen bi eskola nagusietako ukitua ikusiko dugu hemen ere: tradizionala edo sinbolikoa, eta corpusetan oinarritutako estatistikoa. Ebazpen estatistikoak hedatuagoak daude literaturan, gure ustez arrazoi nagusi batengatik: testu errealekin lan egiteko hobeto prestatuta daudelako, estaldura eta sendotasun aldetik batez ere. Aurrera jarraitu aurretik esan beharra dago hurbilpen estatistikoetan ezagutza sintaktiko eta semantikoaren arteko ezberdintasuna lausotu egiten dela sarritan, hitzetan oinarritutako maiztasunak erabiltzen dituzte eta.

Sintaxiari dagokionez nahiz eta esaldien analisi osoa desiragarria litzatekeela aitortu (Kukich, 1992), kategoria etiketa soilekin lan egiten da nagusiki, esaldien egitura sakonago aztertu gabe. Salbuespena Vosse (1994) da, Testuingururik Gabeko Gramatika Hedatuetan (*Augmented Context-Free Gramars*) oinarritutako analisi sintaktiko osoa proposatzen baitu. Nahiz eta bere lana batez ere errore morfo-sintaktikoak zuzentzera zuzendu, ez-hitza erroreen kasuan proposamen bakarra aukeratzen ere saiatzen da. Bere lanean egiten den ebaluazio kuantitatiboa nahasgarria da oso, baina ez-hitza erroreekin lortzen duen doitasuna testu errealetan %60aren inguruan legoke. Testu errealen aurrean doitasuna asko jaisten dela eta, analizatzaile sintaktikoak halakoekin arazoak dauzkala aitortzen du.

Estatistikan oinarritutako lanetan ezagutza sintaktikoa hitz jakin baten inguru hurbilean dauden hitz eta kategoria multzoen kontaketara mugatzen da: kategoria-bigrama eta -trigramak, hitz-bigrama eta -trigramak, eta horien arteko konbinazioak (Gale & Church, 1990; Mays et al. 1991; Golding & Schabes, 1996). Lehenbiziko lanean testuingururik gabe lan egiten zuen sistemari (Kernighan et al. 1990) kategoria-bigrama eta -trigramak gehitzen dizkiote, AP Newswire corpus zabaletik (10^6 hitz) jasotakoak, eta zuzentzailearen doitasuna %87tik %89,7ra igotzen da. (Mays et al. 1991; Golding & Schabes, 1996) lanetan ez-hitzen zuzenketa automatikoa baino harantzago doaz, benetako hitza erroreak detektatu eta zuzendu nahi baitituzte. Golding eta Schabes-en kasuan ingelesez maiz gertatu ohi diren bi hitzen arteko nahasketak aztertzen dituzte (18 bikote guztira), adibidez `wheather/whether` edo `dairy/diary`. Sistemak `dairy` edo `diary` aurkitzen duenean, bere testuingurua aztertu eta errore bat dela erabaki dezake, beste hitzaz ordezkatuz. Ezagutza sintaktikoa bi eratara erabiltzen da: alde batetik kategoria-trigramak erabiliz kategoria egokiena esleituko duen etiketatzailea dago (Mays et al. 1991), eta bestetik hitz/kategoriaz osatutako bigrama eta trigrama nahastuak (Yarowsky, 1994). Etiketatzailea kategoria ezberdina duten hitzen artean erabakia hartzeko gai da (`weather` izena, `whether` konjuntzioa), baina kategoria berdina dutenean ez (`dairy` eta `diary` izenak dira, esneki eta agenda esanahia dutenak). Azkeneko bikotearentzat honako hitz/kategoria trigramak jasotzen du `diary`-ren aldeko ebidentzia:

```
in POSS-DET _
```

Hau da, `dairy/diary`-ren aurrean `in` preposizioa eta edutezko determinantea daudenean `diary` hobetsiko da. Sistema honetan ezagutza sintaktiko eta semantikoa ez daude bereizita, biak era nahasian erabiltzen baitira, eta beraz emaitzak beherago ikusiko ditugu.

Ingels-en (1996; 1997) lanean lehenago aipatutako erroreen iturburuaren eredua eta ezagutza sintaktikoaren eredua konbinatzen dira. Bigarrena kategoria-bigrametan oinarritutako Markoven Eredu Ezkutu (*Hidden Markov Model*) batez egiten du. Ezagutza sintaktikoa erabilitako bi testuei hertsiki lotzen zaie, testu bakoitzarentzat trebatzen baitu bere sistema. Testu berri baten aurrean sistema berriz trebatu beharko litzateke. Emaitzei dagokionez, ezagutza sintaktikoak nabarmen hobetzen ditu emaitzak ez-hitz erroreen kasuan, %89tan ondo zuzentzen baitu lehenbiziko esperimentuan (ezagutza sintaktiko gabe %74) eta %83 bigarrengoan (gabe %54).

V.A.2.c) Semantika

Ezagutza semantikoa erabiltzen duen sistema sinboliko inplementaturik ez dugu ezagutzen, sistema estatistikoak bai ordea. Hauetan ezagutza semantikoa errepresentatzeko agerkidetzak (*cooccurrence*) erabili ohi dira, hau da, testuinguruko N hitzeko leiho batean hitzak zenbat aldiz azaltzen diren

kontatzean. Hala egiten dute Golding eta Schabes-ek (1996) sintaxiari buruzko ezagutza (Yarowsky, 1994)-n proposatutakoarekin hedatuz. Goragoko adibidearekin jarraituz honako hau izan liteke `dairy`-ren aldeko ebidentzia semantiko bat:

```
milk ± 10 hitzetako leiho barruan
```

Ezagutza sintaktiko eta semantikoa konbinatuaz 18 bikoterentzat egindako esperimentuetan %70etik %98.9ra doazen doitasunak lortzen dituzte.

Gorago bi aldiz aipatu dugu Yarowsky-ren (1994) lana. Berez ez da testu-zuzenketari buruzkoa, Frantsesera eta espainierako hitzetan azentu egokia jartzeari buruzkoa baizik. Hala ere berak erabilitako ezagutza iturri eta algoritmoak Golding-ek aplikatu zituen (Golding, 1995; Golding & Schabes, 1996) testu-zuzenketan. Informazioa jasotzeko prozedura bera da bietan, aurrerago ikusiko dugun bezala. Ezberdintasun nagusia informazioa konbinatzeko metodoan datza. Yarowsky-k erabakia hartzeko (`dairy/diary`) ebidentzia indartsuena erabiltzen du. Golding-ek, ordea, ebidentzia guztiak konbinatu nahi ditu, emaitza hobeagoak atera nahian. Ebidentziak estatistikoki (Bayes-en erregela medio) konbinatu ahal izateko beraien arteko dependentzia ezabatu beharra dago, eta beraz heuristiko batzuk erabiltzen dira horretarako, nahiko *ad hoc* direnak. Yarowsky-ren metodoa sinpleagoa da, eraginkorragoa konputazioan, eta Golding-en (1995) artikuluan azaltzen denez konbinazioarekin doitasunaren irabazia apala litzateke (batez-beste %81,3tik %82,9ra).

Ezagutza iturrien konbinazioari buruz, beste testuinguru batean Ménèzo-ren taldeak (Ménèzo et al. 1996; Genthial et al. 1994) adimen artifizial banatuan oinarritutako ortografia eta gramatika zuzentzaile bat proposatzen du, ezagutza iturri ezberdinak konbinatzen dituenak. Sistema hau LNP tradizionalaren baitan koka dezakegu, baina nahiz eta sistema interesgarria iruditu, ez du emaitza kuantitatiborik aurkezten.

V.B. Sintaxian eta semantikan oinarritutako zuzenketaren bideragarritasuna

Semantikak proposamen bakarra aukeratzeko orduan egin lezakeen ekarpena neurtzeko bi aurre-azterketa egin genituen, bata euskararako eta bestea frantseserako. Lehenbizikoan, ezagutza sintaktikoarekin bakarrik zuzenketa egokia ezin dela proposatu arrazoitzen da, bigarrenean, IDHS-ren (Agirre et al. 1997) ezagutza-basean dagoen informazioak eta Dentsitate Kontzeptualak oinarri ona eskaintzen dutela.

V.B.1. Euskararen azterketa

Zuzenketarako sistema diseinatu aurretik, sintaxi eta semantikaren ekarpena neurtu nahi genuen. Horretarako euskaraz egindako erroreen corpusa bildu eta ortografia-erroreen azterketari ekin genion (Agirre, 1993). Azterketa honetan ingurune ideal bat aurresuposatu zen, hau da, analisi sintaktiko sendo, oso eta sakona, baita ezagutza semantiko osoa ere. Analisi sintaktikoa pertsona batek simulatu zuen. Ezagutza semantikoa aplikatzeko oinarri izateaz gain, proposamenak baztertzeko bere gaitasuna ere neurtu zen. Adibidez:

```
Laguntza behar dut arazo hau kopontzeko*.

kopontzeko* ⇒ konpontzeko, kopontzako

Laguntza behar dut arazo hau konpontzeko ONDO
Laguntza behar dut arazo hau kopontzako  GAIZKI
```

Adibideko erroreak bi proposamen ditu, lehenbizikoa aditza eta bigarrena izena (kopa izenaren plural hurbila). Lehenbiziko proposamena sintaxi aldetik onargarria da, ez horrela bigarrena.

Semantikari dagokionean ezagutza semantiko tradizionala ere hartu genuen kontuan: hautapen-murrizpenak alde batetik eta erlazio-izera lexikal-kontzeptuala bestetik.

Hautapen-murrizpenak aditz eta adjektiboen argumentuek betebeharrezko ezaugarri lexikalak dira. Adibidez:

```
konpondu [agente:pertsona, objektu:gailu edo faktore kognitibo]

apurtu [agente: animali, objektu: izaki fisiko]
```

Horrek esan nahiko luke konpondu ekintzaren agentea pertsonaren bat izan beharko litzatekeela, eta ekintzaren objektua gailu edo faktore kognitiboren bat. Ezagutza hori erabiliz, errorea argumentu batean ematen denean[62], argumentu horren hautapen-murrizpena betetzen ez dituzten proposamenak errefusatuko dira. Ikusi ditzagun pare bat adibide:

```
Araso* hau konpontzeko eskatu dut.
araso* ⇒ eraso, arazo, arasa, arbaso
       [gailu edo faktore kognitibo] ⇒ arazo, arasa

Lehioa* apurtu da.
lehio* ⇒ lehia, lesio, leiho
       [izaki fisiko] ⇒ leiho
```

[62] Hautapen-murrizpenak erabili ahal izateko analisi sintaktikoa behar da, gutxienez aditzaren argumentu nagusienak ezagutuko dituena.

Konpontzen aditzaren subjektua gailu edo faktore kognitibo bat izan behar da, eta *araso*-ren proposamenetatik bi bakarrik betetzen dute murrizpena, arazo eta arasak. Beste adibidean, apurturen objektua izaki fisiko bat izan behar da, beraz *lehioa*-ren proposamenetatik lehia eta lesio bazter daitezke, leiho bakarrik utziz. Hautapen-murrizpenak aplikatu ahal izateko, analisi sintaktikoak aditzaren argumentuak (agente, objektu, eta abar) zeintzuk diren eman beharko digu.

Erlazio-izaera lexikal-kontzeptuala Dentsitate Kontzeptualaren bidez kalkula daiteke (ikusi III. eta IV. kapituluak). Proposamen posible eta testuinguruko izenen arteko Dentsitatea neurtu, eta Dentsitate maximoa duen izena izango litzateke testuinguru horrentzat hurbilena. Adibideko esaldian, *uzain*-en proposamenen eta testuinguruan dauden izenen (ukendu eta erle) arteko Dentsitatea neurtu eta usain-ek edukiko luke Dentsitate maximoa.

```
Ukenduaren uzainak* erlea aldendu zuen
uzainak* ⇒ zainak, usainak
```

$$\max_{x \in \{\text{usain,zain}\}} \text{dentsitatea}(\{\text{ukendu,erle}\}, x) \Rightarrow \text{usain}$$

Testuak Irale programako ikasle batzuei jasoak dira (Maritxalar & Ilarraza, 1996). Euskararen ezagutza maila ertaina denez ortografia-erroreetan oparoa da, baita sintaxi-erroreetan ere. Guztira 48 testu dira, 8.290 hitz. Horietatik Xuxen zuzentzaile ortografikoak 1.022 okertzat jo zituen (nahiz eta 102 ez izan benetako erroreak, adibidez izen propio batzuk), 520 errore proposamenik gabe gelditu ziren, eta 95entzat zuzenketa egokia ez zegoen proposamenen artean[63]. Horrek 305 errore uzten ditu zuzentzeko moduan, baina 123k proposamen bakarra dutenez tratamendua behar dutenak 182 dira (ikusi 16. taula).

[63] Xuxen-en alde esan beharra dago errore batzuk oso gaitzak zirela, eta bestalde azterketa hau egin zenetik (1993) Xuxenen bertsio berriak atera direla.

Testu kopurua	48	
Hitzak guztira	8.290	
Hitz okerrak (Xuxen)	1.022	
Zuzenak	102	
Proposamenik gabekoak	520	
Proposamen zuzenik ez	95	
Proposamenekin	305	
Proposamenak	305	
Proposamen bakarra	123	%40
Proposamen anitz	182	%60
Proposamen anitz	182	
Sintaxia	128	%70
Semantika	54	%30
Semantika	54	
Proposamen bakarra	34	%63
Proposamen anitz	11	%20
Ez du ezer egiten	9	%16

16. taula: euskararako azterketaren emaitzak

Ezagumendu sintaktikoa proposamen bakarra aukeratzeko gai zatekeen erroreen %70ean, baina %30ean proposamen bat baino gehiago utziko lituzke semantikaren esku. Lehenago aipatutako hautapen-murrizpen eta Dentsitate Kontzeptuala erabiliz errore horien %63rako proposamen bakarra aukeratzeko gai izango litzateke, nahiz eta proposamen batzuk ezabatu 2 edo 3 proposamen geldituko lirateke %20an, eta %16an ez da gai proposamenik aukeratzeko.

Semantikari esparru zabala azaltzen zaio

Aurre-azterketa honetatik sintaxiak bere kabuz gehienez erroreen %70a konponduko lukeela erakusten du, nahiz eta analizatzaile sintaktiko perfektua simulatu, eta gogoan izanda euskarak informazio morfosintaktiko asko eskaintzen duela beste hizkuntzekin alderatuz gero. Beraz semantikari esparru zabala geratzen zaio, gutxienez erroreen %30a, eta errore horietatik %63 konpontzera hel daiteke. Sintaxi eta semantika perfektua suposatuz ere errore gutxi batzuk zuzentzeke gelditzen dira. Horietarako beste informazio-iturri batzuk, hala nola, munduaren ezagutza, pragmatika, eta abar hartu beharko lirateke kontuan. Hona hemen azkeneko kasuaren adibide bat:

```
Astoa handia izanez, eta erlia* txikia izanez, …
erlia* ⇒ eria, erlea, erbia, erdia
```

Erlazio-izaera semantikoak eria eta erdia baztertuko lituzke, baina ez dago arrazoirik erlea edo erbia aukeratzeko.

V.B.2. LPPL-ren HEBaren egokitasunaren azterketa

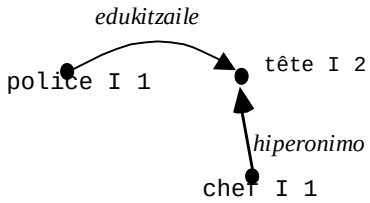
Ezagutza semantikoa beharrezkoa bada, non egongo da jasota ezagutza hori? Zein errepresentazio eredu erabiliko dugu? Ezagutza hori ez badago jasota nondik eta nola eskura dezakegu?

Orain arte bi ezagutza semantiko aipatu ditugu: hautapen-murrizpenak eta erlazio-izaera lexikal-kontzeptuala. Bigarrenerako ikusi dugu Dentsitate Kontzeptuala formalizazio egokia dela, eta nola inplementa daitekeen WordNet ezagutza-basearen gainean (ikus III.C.2 atala). Azter dezagun orain adibide baten bidez nola heda daitekeen Dentsitate Kontzeptuala HEBan dauden hautapen-murrizpenei lotutako informazioa ere integratzeko:

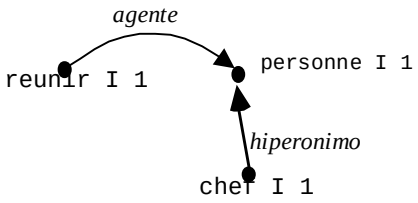
```
Le cheé* de la police ha reunit vingt hommes sur la place du village.

cheé* ⇒ chef, cher, chez, chié, chéri, chic
```

Proposamen eta testuinguruaren artean hainbat erlazio topa daitezke LPPL-ren ezagutza-basean, bai Dentsitate Kontzeptuala aplikatzeko (ikus 23. irudia), bai hautapen-murrizpenak ebazteko (ikus 24. irudia). Bi irudietan azaltzen diren erlazioak LPPL-ko HEBan jada existitzen dira. Aipatzeko da, WordNet-en ez bezala, HEB honetan erlazio paradigmatikoez gain bestelako erlazio asko ere badaudela (adibidez 23. irudiko *edukitzaile*, edo 24. irudiko *agente*, ikus II. kapitulua ere), eta Dentsitate Kontzeptuala horietaz balia zitekeen.



23. irudia: chef eta police-en kontzeptuen arteko erlazioa



24. irudia: reunir-en hautapen-murrizpena chef-ek nola bete dezakeen

Ondorioz, LPPL-tik erauzitako Hiztegi Ezagutza-Baseak zuzenketarako beharrezkoa den ezagutza semantiko hori eskaintzen duela ikusi dugu –honi buruzko argibide gehiagotarako ikus (Agirre et al. 94; Agirre et al. 95)–.

V.B.3. Bideragarritasun-azterketaren ondorioak

Orain arte azaldutakoaren arabera ondorio hauetara iritsi gara:

- Ezagutza sintaktikoa ez da nahikoa (ikus V.B.1 atala).
- Ezagutza semantikoaren erabilera beharrezkoa eta posiblea da, frantseserako eraiki den LPPL-ren ezagutza-baseak erakusten duen bezala (ikus V.B.2 atala).

Ondorioz esperimentu erreal bat prestatzeko arrazoiak badaude. Bestalde, ezagutza iturri hauekin esperimentu errealista bat diseinatzeko orduan hainbat muga agertu zaizkigu:

- Frantseserako HEBaren estaldurak testu librearekin lan egitea galarazten du.
- Hautapen-murrizpenak ez daude jasota ez HEBan[64], ez WordNet-en, ez eta eskuragarri dauden beste ezagutza-base orokorretan ere.
- Analisi sintaktiko osoa egingo digun sistemarik ez dugu eskura. Horrelako sistema beharrezkoa da hautapen-murrizpenak dagozkien hitzei aplikatu ahal izateko.

Muga hauek bultzatu gintuzten esperimentu erreala ingeleserako prestatzen, estaldura zabaleko WordNet gainean Dentsitate Kontzeptuala erabiliz, baina hautapen-murrizpenik eta erlazio ez-paradigmatikorik gabe. Bestalde, corpusetan oinarritutako hainbat teknikek ere nolabaiteko ezagutza semantikoa isladatzen dutenez (Yarowsky, 1994; Golding & Schabes, 1996), horrelakoak ingeleserako corpus zabal batetik eskuratzea ere bideragarritzat jo genuen.

V.C. Erabilitako teknikak

Semantika kontuan hartzen duten teknikez gain, ezagutza sintaktikorako Murrizpen-gramatika erabili da, eta eskura egon zitezkeen beste heuristikoak ere ez dira baztertu. Tamalez hitz isolatuen zuzenketarako eredurik ezin izan genuen eskuratu. Hauek dira erabili ditugun teknikan.

V.C.1. Murrizpen-gramatika (MG)

Murrizpen-gramatika testu librearen analisi sintaktikorako sortu zen, sendotasun eta estaldura zabala helburu. Hainbat hizkuntzatan aplikatu bada, euskara barne, ingeleserako kategoria etiketatzaile bezala eduki du arrakasta handiena (Karlsson et al. 1995). Guri dagokigunean, Murrizpen-Gramatika izango da proposamenak aukeratzeko izango dugun ezagutza sintaktikoa.

V.C.2. Dentsitate Kontzeptuala (DK)

Ezagutza semantikoa WordNet gainean lan egingo duen Dentsitate Kontzeptualak emango digu (ikus III.B atala). Honen arabera, proposamen guztiak izenak direnean, beraien artean

[64] HEBan hautapen-murrizpenak erauzteko beharrezko ezagutza egon badago, aurreko ataleko adibideak erakusten dutenez, baina ez dira oraindik fisikoki erauzi.

testuinguruarekiko Dentsitate altuena duena aukeratuko genuke. Dentsitate Kontzeptuala adieren arteko neurria izanik, Dentsitate altueneko adiera duen izena aukeratzen da. Dentsitatea neurtzeko orduan inguruko 60 izen hartu dira testuinguru bezala, hitzen adiera desanbiguazioan emaitza onenak lortu diren berak.

V.C.3. Maiztasuna (BM eta DM)

Bi maiztasun jaso dira: hitzen maiztasun orokorra, Iparramerikako ingeleserako estandar bilakatu den Brown corpusetik hartua (Francis & Kucera, 1967), eta dokumentuan bertan dauden hitzen maiztasuna. Lehenbizikoari BM eta bigarrenari DM deitu diegu.

V.C.4. Testuingurua kontuan hartzen duten metodo estatistikoak (TS)

Yarowsky-ren lana (1994) hartu genuen oinarritzat (ikus lan honi buruzko oharrak IV.A.4 atalean). Bere lanean ez bezala gurean ezinezkoa da aldez aurretik proposamenen multzoa mugatzea, errorea edozein hitzetan azaldu daiteke eta (ikus V.A.2 atala). Hori dela eta maiztasun gordinak fitxategietan gorde eta proposamenak zuzentzeko orduan kalkulatuko dira beharrezko neurriak.

Maiztasun informazioa bildu ahal izateko, lehenbizi Brown corpusa tokenizatu eta ondoren bigramak, trigramak eta leiho jakin baten[65] agertzen diren hitz pareak fitxategi batzuetan gorde dira, bakoitzaren agerpen kopurua zenbatuta dagoela. Gure kasuan ez kategoriak ezta lemak ere ez genituen erabili, hitz-formak soilik. Neurri estatistiko bezala log-sinesgarritasuna (*log-likelihood*) erabili dugu, Yarowsky-ren moduan. Log-sinesgarritasuna erabiltzearen abantaila, erabakia topatutako ebidentzia indartsuenaren arabera hartzea da. Beste neurriekin ebidentzia guztiak konbinatu behar dira prozesu garestiago batean, prozesu guztia motelduz abantaila gehiegirik atera gabe (ikusi V.A.2 atala).

V.C.5. Bestelako heuristikoak (H1 eta H2)

Esperimentuak egin ahala heuristiko sinple batzuen beharra ikusi genuen. Alde batetik sarritan proposamenen artean izen nagusiak agertzen zirela igarri genuen, nahiz eta errorearen lehenbiziko hizkia xehea izan. Halakoetan izen nagusiak ziren proposamenak arazo gabe ezaba zitezkeen (H1 heuristikoa).

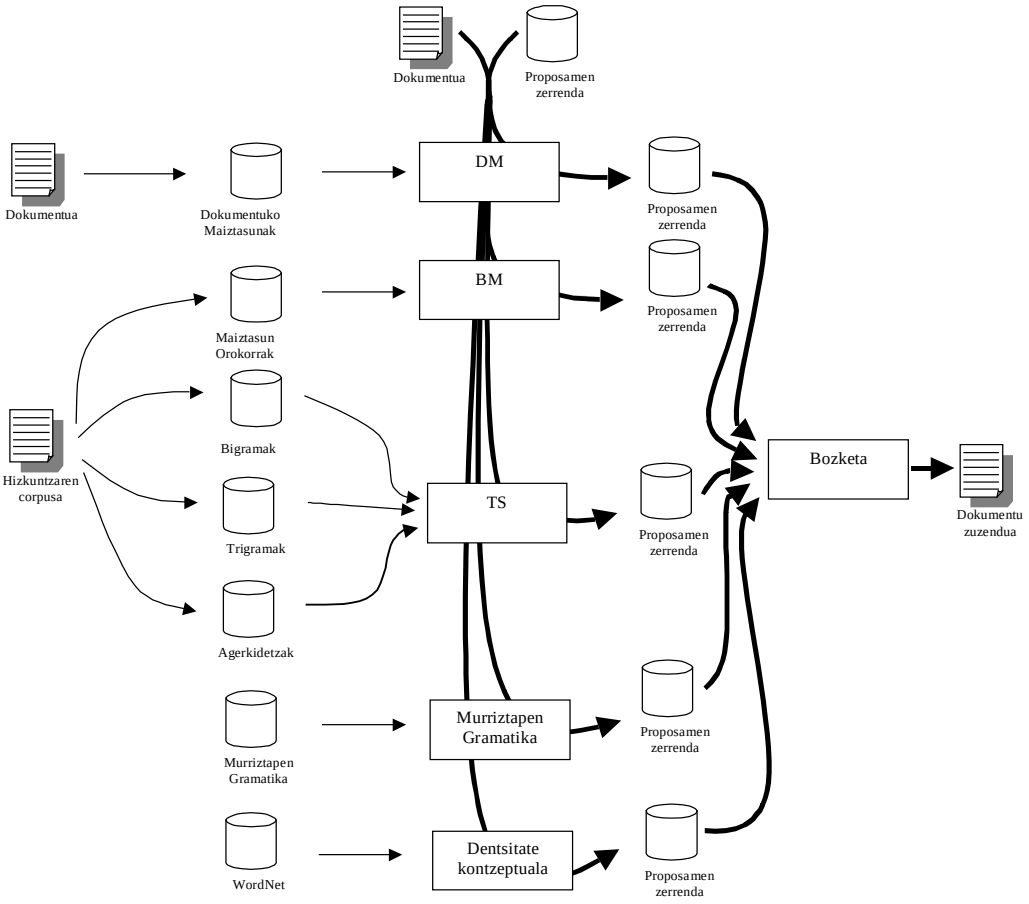
Beste aldetik, ingelesez hiruzpalau letratako erroreetarako proposamenak ugariegiak suertatzen dira. Sistemaren doitasuna errore labur horietarako erantzunik eman gabe hobetuko zelakoan geunden (H2 heuristikoa).

---

[65] . Leihoaren zabalera 40 hitzetakoa izan da.

V.C.6. Konbinazioa: bozketa

Lau metodo eta bi heuristiko sinple ikusi ditugu. Teknika heterogeneoak direnez ez da erreza aurrikustea zein izango den denak konbinatzeko sistemarik hoberena. Hori dela eta 5 teknikentzat VI.D.7 atalean erabilitako sistema bera erabiliko dugu: bozketa. H1 heuristikoak ez duenez kale egiten beti aplikatu izan dugu, H2 ez ordea. Sistema osoaren eskema 25. irudian ikusi daiteke.



25. irudia: proposamenaren hautapenerako ezagutza iturriak eta konbinatzeko sistema

V.D. Ingeleserako esperimentuak

Teknika bakoitzak pisu ezberdina eduki dezakeenez bozkatzerako orduan, teknika eta pisuen aukeraketa esperimentua egiten genueneko testuari lotua ez egotea lortu behar genuen. Horretarako konbinazio guztiak lehenbiziko corpus baten gainean probatu eta onenak aukeratu genituen. Aukeratutako konbinazio horiek beste corpus ezberdinean probatu ziren, emaitzak baieztatzeko.

V.D.1. Aukeratutako corpusak: sortutako erroreak eta benetako erroreak

Bi corpus aukeratu genituen esperimenturako. Alde batetik Brown corpuseko dokumentu batzuetan errore artifizialak sortu genituen ausaz, eta bestetik benetako erroreak zeuzkan testuak bildu

genituen. Lehenbizikoan erroreen zuzenketa zein den automatikoki jakin dezakegunez, nahi bezain handia egin dezakegu, eta beraz berebizikoa da saiakera ezberdin asko egiteko. Bigarrenak aldiz, benetako testu baten aurrean espero genezakeen emaitzaren berri emango liguke.

Lehenbizikorako 8 bertsio egin genituen. Damerau-ren (1964) legeak jarraituz batez-beste 20 hitzetan behin errore bat sortzen duen *antispell* programa (lan honetarako espreski garatua) 8 aldiz egikaritu genuen aldez aurretik aukeratutako Brown corpuseko lagin baterako. Corpus honi *artifiziala* deituko diogu. Bozketa-saiakera ezberdinak egin ahal izateko, aipatu bezala, corpus hau bitan zatitu zen: zati baten lau testu eta bestean beste lauak (ikusi datuak 17. taulan).

Benetako erroreak dituen corpusa, *benetako* corpusa, *Bank of English* delako corpusetik aldizkarien testuak bilduz jaso genuen. Hauetarako bai, eskuz erabaki izan behar genuen zein zen errore bakoitzaren zuzenketa.

	1. erdia	2. erdia	benetakoa
Hitzak	47.584	47.584	39.733
Topatutako erroreak	1.354	1.403	519
Proposamendun erroreak	1.354	1.403	369
Ispell-en proposamenak	7.242	8.083	1.257
Proposamen anitzeko erroreak	810	852	158
Erroredun hitz luzeak (H2)	968	980	331
Hauentzako proposamenak (H2)	2.245	2.313	807
Proposamen anitzekoak (H2)	430	425	124

17. taula: errore corpusen datuak.
Lehenengo bi zutabeak corpus artifizialari dagozkio.

Bi corpus hauek *ispell* izeneko ingeleserako zuzentzailetik pasatu genituen. Corpus artifizialerako arazo gabe topatu zituen erroreak eta sortu proposamenak. Benetako corpusean arazoak izan zituen, 150 hitzentzat ez baitzuen proposamenik sortu (gehienak izen bereziak edo hitz arrotzak). Guztira 1.354, 1.403 eta 369 errore zeuden corpus bakoitzean, zegozkien proposamenekin. Proposamen bakarrekoak kenduta, 810, 852 eta 158 errore geratu zitzaizkigun tratatzeko. 17. taulan H2 heuristikoa aplikatuz gero, hau da, errore laburrak kontuan hartuko ez bagenitu, gelditzen den errore kopurua ere azaltzen da. Honek ematen digu egin beharreko lanaren neurria. Adibidez, benetako testuan erroreen erdiak baino gehiago proposamen bakarra dauka, eta proposamen anitzeko erroreek batez beste 6,62 proposamen dituzte, errore luzeak soilik kontuan hartuz gero, aldiz, 4,84.

V.D.2. Emaitzak

Ebaluaziorako ondoko hiru neurriak erabili ditugu:

- Estaldura: proposamenen bat aukeratzen den aldi kopurua, hau da, erantzunik ez dagoenean ezik.
- Doitasuna: proposamen-aukeraketa egin denean, zuzena zenbat aldiz geratu den.
- Aukeratutako proposamenak: batez-beste zenbat proposamen aukeratu izan diren errore bakoitzeko.

V.D.2.a) Konbinazio hoberenen bilaketa

Esan bezala, corpus artifizialaren erdian egin genituen lehenbiziko saioak. 18. taulan azaltzen dira teknika bakoitzerako lortutako emaitzak, hala nola konbinazio arrakastatsuenen emaitzak, beti ere proposamen anitz dituzten erroreentzat, hau da, proposamen bakarrekoak kontuan eduki gabe. Horiek kontuan edukiko balira emaitzak hobeak lirateke, noski, baina okerrago adieraziko lukete teknika bakoitzaren eraginkortasuna. Teknika isolatuen artean DM eta TSek lortu zituzten emaitza aipagarrienak (taulan ilun azaltzen direnak), denak ere ausaz lortuko liratekeen emaitzak erraz gaindituz. DKak lortzen ditu emaitza apalenak, estaldura aldetik batez ere, kasuen %8an bakarrik baita erantzuteko gai. Aurrerago aztertuko dugu hori.

Konbinazioen artean gutxi batzuk erakusten ditugu, hoberenak ilun azaltzen diren MG1+DM1+TS1 eta MG1+DM1+TS2 dira, hau da, Murrizpen-Gramatikak botu bat, Dokumentuko Maiztasuna botu bat eta Testuinguruko Estatistikek botu bat edo bi jasotzen dituzteneko konbinazioak, hurrenez hurren. Doitasun gorena beraz %86koa da, estaldura ia osoaz, eta 1.12 proposamen utziaz batez beste.

Hitz luzeak bakarrik zuzentzen ahaleginduz gero (H2) orduan estaldura ia erdira jaisten da, doitasun kasu onenean (MG1+DM1+TS1+H2) %92raino heltzen delarik 1,11 proposamen utziaz.

	%Estal.	%Doi.	#prop
Oinarrizko teknikak			
Ausazkoak	100,00	23,70	1,00
MG	99,75	78,09	3,23
DK	8,27	75,28	1,01
BM	93,70	76,94	1,00
DM	84,20	81,96	1,03
TS	94,48	84,94	1,02
ausazkoak+H2	52,70	36,05	1,00
MG+H2	52,57	90,68	2,58
BM+H2	48,04	81,38	1,00
DM+H2	38,48	89,49	1,03
TS+H2	47,79	89,77	1,02
Konbinazioak			
MG1+DM2	99,88	83,93	1,28
MG1+DM1+BM1	99,88	81,83	1,04
MG1+DM1+BM1+DK1	99,88	81,83	1,04
MG1+DM1+TS1	99,88	86,45	1,12
MG1+DM1+TS2	99,88	85,45	1,07
MG1+DM2+H2	52,70	91,86	1,43
MG1+DM1+BM1+H2	52,70	88,14	1,06
MG1+DM1+BM1+DK+H2	52,70	87,91	1,05
MG1+DM1+TS1+H2	52,70	92,12	1,11
MG1+DM1+TS2+H2	52,70	90,32	1,09

18. taula: proposamen anitz duten erroreetarako emaitzak (1. erdia).

	%Estal	%Doi	#prop
Oinarrizko teknikak			
Ausazkoak	100,0	23,71	1,00
DM	84,04	81,42	1,03
TS	95,39	84,90	1,02
Ausazkoa+H2	50,12	34,35	1,00
DM+H2	36,32	87,66	1,04
TS+H2	46,93	86,35	1,02
Konbinazioak			
MG1+DM2	99,41	83,59	1,31
MG1+DM1+BM1	99,41	79,81	1,05
MG1+DM1+BM1+DK1	99,41	80,05	1,05
MG1+DM1+TS1	99,53	86,14	1,15
MG1+DM1+TS2	99,53	85,26	1,07
MG1+DM2+H2	50,12	90,12	1,50
MG1+DM1+BM1+H2	50,12	84,24	1,06
MG1+DM1+BM1+DK+H2	50,12	84,47	1,06
MG1+DM1+TS1+H2	50,12	88,70	1,16
MG1+DM1+TS2+H2	50,12	86,59	1,07

19. taula: proposamen anitz duten erroreetarako emaitzak (2. erdia).

V.D.2.b) Konbinazio hoberenen egiaztapena

Behin konbinazio hoberenak zeintzuk izan zitezkeen zehaztu eta gero, corpus artifizialaren bigarren erdiarekin probatu genituen. 19. taulan ikusten den bezala emaitzak mantendu ziren, nahiz eta kuantitatiboki zertxobait jaitsi orokorrean. Honek frogatzen du konbinazio hoberenak orokorrak direla, ez daudela corpusari lotuta.

V.D.2.c) Benetako erroreen corpusa

Artifizialki sortutako erroreek emaitzak baldintza zitzaketelakoan, benetako erroreen corpusaren gainean probatu genituen konbinazio hoberenak, bai eta oinarrizko teknikak ere[66]. 20. taulako emaitzek adierazten duten bezala, konbinazio hoberenak mantendu egiten dira, baina doitasun eta proposamen kopuruak dezente okertu. Horrela puntako doitasuna H2 erabili barik %78koa da (8 puntu gutxiago) 1,56 proposamen utziz, eta H2 erabilita %81 (11 puntu gutxiago) eta 1,53.

Oinarrizko teknika guztien beherakada ikusita ere, bereziki aipagarria da Dokumentuen Maiztasunarena (50 puntu gutxiago estalduran, 20 puntu gutxiago doitasunean, ikusi 18. eta 20. taula). Hau ez da harritzekoa, izan ere benetako corpuseko dokumentuak motzak dira oso, batez beste 50 hitzekoak.

[66] Dentsitate Kontzeptualarekin ez ginen saiatu, bere estaldura (%8) eta benetako errore kopurua (158) txikiak izanik, ondorioak ateratzeko lagin txikiegia zelakoan.

Jaitsiera orokorraren arrazoiak erroreen izaeran bertan egon daitezke, benetako erroreak artifizialak baino zailagoak suertatu direlako edo. Bestalde aipatu beharra dago dialektoen arteko gatazka baten aurrean ere egon gaitezkeela, Brown corpusa Estatu Batuetako ingelesez eta *Bank of English* Bretainia Handikoaz daude eta.

	%Estal.	%Doi.	#prop.
Oinarrizko teknikak			
Ausazkoak	100,00	29,75	1,00
MG	98,10	62,58	2,45
DM	30,38	62,50	1,13
BM	96,20	54,61	1,00
TS	93,21	74,16	1,05
Ausazkoak+H2	76,54	34,52	1,00
MG+H2	75,93	73,98	2,52
DM+H2	12,35	75,00	1,05
BM+H2	72,84	60,17	1,00
TS+H2	67,28	75,36	1,03
Konbinazioak			
MG1+DM2	100,00	70,25	1,99
MG1+DM1+BM1	100,00	55,06	1,04
MG1+DM1+TS1	100,00	78,51	1,56
MG1+DM1+TS2	100,00	75,94	1,09
MG1+DM2+H2	76,24	75,81	2,15
MG1+DM1+BM1+H2	76,54	59,68	1,05
MG1+DM1+TS1+H2	76,54	81,58	1,53
MG1+DM1+TS2+H2	76,54	78,11	1,08

20. taula: proposamen anitz duten erroreetarako emaitzak (benetako corpusa)

	%Estal.	%Doi.	#prop.
Oinarrizko teknikak			
Ausazkoak	100,00	69,92	1,00
MG	99,19	84,15	1,61
DM	70,19	93,05	1,02
BM	98,37	80,99	1,00
TS	97,02	89,10	1,02
Ausazkoak+H2	89,70	75,47	1,00
MG+H2	89,43	90,30	1,57
DM+H2	61,52	97,80	1,00
BM+H2	88,08	85,54	1,00
TS+H2	85,64	91,50	1,01
Konbinazioak			
MG1+DM2	100,00	87,26	1,42
MG1+DM1+BM1	100,00	80,76	1,02
MG1+DM1+TS1	100,00	90,80	1,24
MG1+DM1+TS2	100,00	89,70	1,04
MG1+DM2+H2	89,70	90,94	1,43
MG1+DM1+BM1+H2	89,70	84,89	1,02
MG1+DM1+TS1+H2	89,70	93,10	1,20
MG1+DM1+TS2+H2	89,70	91,80	1,03

21. taula: emaitza orokorrak (benetako corpusa).

V.D.3. Ebaluazioa

Oinarrizko teknikei dagokionez, emaitzek ondoko ondorioetara garamatzate:

- Murrizpen-Gramatika ez da espero zitekeen bezain bikaina, errore dezente sortzen baitu erroredun testuetan aplikatu dugunean (%62ko doitasuna bakarrik benetako corpusean).
- Dentsitate Kontzeptualaren emaitzak apalak dira. Proposamen guztiak izenak izan behar direnez, erroreen %8an bakarrik aplikatu ahal izan da.
- Brown Maiztasuna ez da batere eraginkorra.
- Dokumentuko Maiztasunak emaitza onak lortzen ditu, testuak motzegiak ez badira (benetako corpusaren kasuan gertatu den bezala).
- Testuinguru-Estatistikak dira zalantzarik gabe emaitza hoberenak lortzen dituztenak, estaldura, doitasun eta proposamen kopuruari dagokionean.

Teknika ezberdinen bozketa bidezko konbinazioari esker Testuinguru-Estatistiken emaitzak hobetzea lortzen da, doitasun gorena eta erabateko estaldura lortuz. Horretarako MG, DM eta TS teknikak konbinatzea nahikoa da. Besteen laguntzak ez du emaitza ezertan hobetzen. H2 heuristikoa erabilgarria da doitasuna altxatzeko, baina estaldura %76ra jaisten da.

Sistemaren irteera zein izango litzatekeen neurtzeko proposamen bakarreko erroreak ere kontuan hartu behar dira (ikusi 21. taula). Datu horien arabera bi irteera planteatu daitezke:

- Erabateko estaldura, %90eko doitasuna, eta proposamen bakarra 25 erroretik 24tan (MG1+DM1+TS2).
- Doitasun gorena, %93, baina %90eko estaldura eta proposamen bakarra 5etik 4tan (MG1+DM1+TS1+H2).

Tekniken ekarpenaren azterketa egitean, metodo tradizionalen aurrean, metodo estatistikoak portaera hobeagoa eduki duela ikusi dugu, bai doitasun eta bai estaldura aldetik ere. Murrizpen-Gramatikaren bidez proposamenak baztertzean erantsitako errorea %38koa bada, Dentsitate Kontzeptualak doitasun hobea lortzen du, %75ekoa corpus artifizialean, baina estaldura oso apalaz. Sintaxi, semantika eta kolokazioei buruzko informazioa jasotzen duten Testuinguru-Estatistikek erraz gainditzen dute beste bi horien konbinazioa, zalantzarik gabe.

Dentsitate Kontzeptuala eta Testuinguru-Estatistiken emaitzen arteko aldea argitzeko, lehenbizi DK proposamen guztiak izenak direnean bakarrik aplika daitekeela aipatu behar da. Doitasunari dagokionez DKak ez du testuinguruan agertzen diren izenen taxonomiari buruzko informazioa besterik. TSek ordea corpusetik erauzitako informazio inplizitu aberatsa daukate, bai kategoria ezberdinen artekoa, baita taxonomikoa ez den ezagutzari buruzkoa ere.

Testuinguru-Estatistiken emaitzak, hala ere, Yarowsky-k (1994) kontatzen dituenak baino dezente apalagoak dira. Nahiz eta guk kategoria eta lemak ez erabili, arrazoi nagusiena berak maiz gertatzen diren nahaste-multzo (*confusion-set*) gutxi batzuentzat ebaluatzen duela izan daiteke. Gure kasuan aurrikusi ezin daitezkeen erroreentzat proposamena aukeratu beharra dago, nahiz eta proposamenen horientzako maiztasunak oso baxuak izan (III.D.1 atalean azaldu zaigun datu urrien arazoa). Honek, noski, erabakiaren fidagarritasuna erabat mugatzen du. Arazo honek ez dauka ebazpen errazik. Proposamen batzuek maiztasun baxua dutenez, pentsatu daiteke corpus handiagoak erabilita horien maiztasuna igoko dela. Praktikak erakusten digu hori ez dela zehazki horrela, hitz berriak azalduko zaizkigu eta. Hiztegia ez da zerrenda mugatu bat, are eta corpus

handiagoak bildu, orduan eta hiztegi zabalagoak beharko ditugu. Horrela diote Church eta Gale (1990) beraiek:

> *One might think that the sparse data problem could be solved by collecting larger corpora, but ironically, the problem gets worse as we look at more data. The vocabulary is not fixed: both N - size of corpus - and V - size of vocabulary - grow as we look at more data. The rate of growth is still a matter of debate, but the evidence shows that* $V > O(\sqrt{N})$*, and therefore, the sparse data problems only get worse as we look at more and more data.*

Arazo honetaz gain, zuzenketa edozein errorerentzat egin nahi izatean errepresentazio- eta erabilgarritasun-arazoa ere azaleratzen dira. Alde batetik datu gordinak gorde behar direnez (horietako asko ezertarako ere balioko ez dutenak) fitxategi erraldoiak sortzen dira[67] (datu gehiegizkoen arazoa, ikus III.D.1), erabiltzeko motelak. Bestetik, datu gordinak izanda, erabili nahi diren bakoitzean prozesatu beharra dago, behin eta berriz errore bakoitzerako.

Datu horietatik abiaturik maila jasoagoko errepresentazioa sortu beharko litzateke, erabilgarriagoak izateko, eta arrazonamendu ezberdinen euskarri izan ahal izateko. Ezagutza-baseak aberasteko erabiliko balira, adibidez, jatorri ezberdinetako ezagutza integratuko litzateke, eta erlazio mota berriak erantsi. Honek erlazio mota eta kopuru zabaleko ezagutza-base belaunaldi berri bat ekarriko luke, automatikoki erauzitako datuek osatuta. Aipatu dugun bezala, Dentsitate Kontzeptualaren ahulezia ez dago formulan, kontzeptuan. Bere gabeziak lotutako ezagutza-basearen gabezien ispilu dira. LPPL HEBra lotzen badugu erlazio mota aberatsa ustia dezake, baina lexiko eta erlazio kopuru aldetik estaldura eskasa edukiko du. WordNet erabiltzen badugu, lexiko aberatsa edukiko du eskura, eta hiperonimia erlazioa era sakonean landuta, baina beste erlazio motarik ez dago. Corpusetako informazioarekin aberastutako ezagutza-basea izanez gero, Dentsitate Kontzeptualak gaur egun Testuinguru Estatistikek duten ondasun gordin hori guztia ustiatu ahal izango luke.

V.E. Ekarpena

Kapitulu honetan Dentsitate Kontzeptuala aplikazio erreal batean probatu dugu. Alde batetik zuzenketa automatikoa gaur egungo teknologiaren eskura dagoela frogatu dugu, eta bestetik Dentsitate Kontzeptualaren ekarpena apala izan dela ikusi dugu.

[67] Brown corpuseko hitz formentzat: bigramen fitxategiak 10 Mega, trigramenak 41 Mega, zabalera 8 duen lehioko agerkidetzak 42 Mega eta 40rako leihoak 168 Mega. Yarowsky-k (1994) gomendatzen duen leiho zabalera 100ekoa da, eta hitz-formez gain kategoria eta lemak ere barne egon beharko lukete.

Testu-zuzenketa automatikoa egiten duen sistema diseinatu eta eraiki dugu, ez-hitz motako sakatze erroreentzat proposamen egokia aukeratzen dena. Hasierako azterketa batean, ikasleen testu batzuk bildu eta baliabide osoak suposatuz ere, sintaxia soilik proposamen bakarra aukeratzeko gai ez dela ondorioztatu genuen. Semantikaren ekarpena ezinbesteko ikusi genuen, beraz, erlazio-izaera lexikal-semantikoaren eta hautapen-murrizpenen bidez gauzatuko zena. LPPL-ko HEB frantseserako zuzenketa automatikoa egiteko baliabide egokia litzatekeela ikusi dugu, baina lexikoaren estalduraren aldetik arazoak direla eta, esperimentu errealista batetarako ez zegoela prest iritzi diogu. Baliabide zabalagoen bila ingeleserako WordNet aukeratu dugu, baina kasu honetan ez du eskaintzen hautapen-murrizpenei buruzko informaziorik, eta erlazio-izaera kontzeptuala erlazio paradigmatikoetara mugatzen da.

Ondoren, ingelesaren zuzenketa automatikorako sistema aurkeztu da, *ispell* zuzentzaileak topatzen dituen erroreentzat (*ispell*-ek proposamenak sortzeko oso zehatza dela erakutsi digu) proposamen bakarra aukeratzen saiatzen dena. Sistema honek ezagutza-mota ezberdinak konbinatzen ditu: sintaktikoa (Murrizpen-Gramatikak), semantikoa (Dentsitate Kontzeptuala), hitzen maiztasunak, testuinguru-estatistikak eta heuristiko espezifikoak. Murrizpen-Gramatika, Dokumentuko Maiztasun eta Testuinguru-Estatistikei esker, gai da 25 erroretatik 24etan proposamen bakarra aukeratzeko (bestela bi proposamen) %90eko doitasunarekin, eta errore **guztientzat** erantzuten du. Emaitza hauek frogatzen dute zuzenketa automatikoa egingarria izan daitekeela.

Dentsitate Kontzeptualaren ekarpena eskasa izan da. Hasteko beharrezkoa da proposamen guztiak izenak izatea, eta hori oso gutxitan gertatzen da (erroreen %8 inguru errore artifizialak dituen corpusean). Lagin txiki horrekin, fidagarritasun gutxiko datua izanda ere, %75eko doitasuna lortu da. Ausazkoak lortzen duen %23aren ondoan hobekuntza nabarmena. Kapitulu honetan azaldu dugunez, doitasun eta estaldura hauen arrazoia ez da DKarena berez, erabilitako WordNet ezagutza-basearen gabeziena baizik. Bestalde Testuinguru-Estatistikak erabiltzea erabilgarritasun- eta biltegitze-arazo larriak ditu, eta LNPko beste atazetara egokitzeko orduan ez da berrerabilgarria. Hala ere Testuinguru-Estatistikak biltzen duen datu-nahaspilan erlazio baliagarri asko ezkutatzen dira, eta erlazio horien bidez ontologiak (adibidez, WordNet) aberastuz gero aurrerapauso ederra emango zitekeen ontologien eraikuntzan, eta Dentsitate Kontzeptualak gaur egun Testuinguru Estatistikek duten ondasun gordin hori guztia ustiatu ahal izango luke.

V.F. Etorkizunerako lana

Esperimentua diseinatzeko orduan ez genuen kontuan hartu ikasteko corpusa (Brown) eta probatzekoa (Bank of English) dialekto ezberdinekoak zirenik. Ziurra da arazo honek maiztasun

orokorrak erabiltzen dituen heuristikoaren eta Testuinguru-Estatistikak erabiltzen dituenaren emaitzak kaltetu dituela. Komenigarriena Bank of English corpuseko bertako datuetatik ikastea izango litzateke, baina tamalez datu horiek eskuratzeko murrizpen gogorrak daude. Murrizpen hauen ondorioz ere errore errealen corpusak oso testuinguru txikia zeukan errorearen inguruan. Horrek modu erabakiorrean kaltetu du Dokumentuko Maiztasunen teknika, bestela oso indartsua zena. Arazo horiek konpondu ondoren doitasuna nabari hobetuko delakoan gaude.

Sistemak duen zehaztasun mugatuak ez dezan galarazi zuzenketa automatikoa, beharrezkoa da erabilitako ezagutza fintzea. Murrizpen-Gramatika, adibidez, erroreak dituzten testuetara hobeto egokitu daiteke, guk erabili dugun bertsioa ez baitzegoen horretarako diseinatuta.

Dentsitate Kontzeptualak emaitza hobeak lortu ahal izateko beharrezkoa litzateke erlazio paradigmatikoak ez direnak eta hautapen-murrizpenak kodetzea ezagutza-basean, kasu honetan WordNet-en. Horrelako erlazioak erauzteko bi iturri ikusten ditugu:

- Hiztegietako *differentiaren* azterketatik erauzi, hurrengo kapituluan aipatu dugun legez.
- Corpusen azterketatik. Izan ere Testuinguru-Estatistikek darabilten informazio gordinean, modu inplizituan bada ere, erlazio eta hautapen-murrizpenak ere ezkutatzen dira.

WordNet bezalako ezagutza-basea bi iturri horiekin aberastuz gero, ezagutza-base lexiko ahaltsuagoa litzateke, eta zuzenketa automatikorako beharrezkoa den ezagutza semantikoaren zati nagusia edukiko luke, Dentsitate Kontzeptualak ustiatuko lukeena. Bestalde, Testuinguru-Estatistiketako informazio hori ezagutza-basean integratuz gero, biltegitze- eta berrerabilgarritasun-arazoak hobetuko lirateke, eta inplizituki adierazita zeuden erlazio ugari esplizituki errepresentatuko lirateke ezagutza-basean, LNPko beste eginbeharretarako prest.

Bukatzeko, III. kapituluan aipatu dugun Dentsitate Kontzeptualaren ezaugarrietako bat kolokan gelditu zaigu. Nahiz eta Dentsitate Kontzeptualaren ezaugarrien artean hitzen arteko erlazio-izaera neurtzeko baliagarria izatea jarri, kapitulu honetako emaitzek zalantzan jartzen dute modu egokian erabili ote dugun. Esperimentu honetarako erabili dugun algoritmoan proposamenen adieren arteko Dentsitate handiena zuen hitza hautatu dugu, baina bestelako konbinazioak ere probatu beharko genituzke, adibidez, hitz bakoitzaren adiera guztien Dentsitatea batu eta batura handiena duen hitza aukeratu proposamen egokia bezala.

VI. Kapitulua
HIZTEGI EZAGUTZA-BASEAREN ABERASKETA

Jadanik azaldu zaigu, aurreko kapituluetan, erlazio-izaera neurri on bat edukitzeko eta Hitzen Adiera-Desanbiguazioa (HAD) egiteko, zein garrantzitsua den baliabide lexikal egituratu aberatsak edukitzea. Hori izango da hain zuzen ere kapitulu honen gaia. Aurre egingo diogun eginkizuna frantses hiztegi bateko adierak WordNet-i lotzea eta hiztegi horretatik erauzitako HEBko hierarkiak desanbiguatu eta sendotzea izango da. Lehenbiziko sailean gaia kokatuko dugu, aurrekariak aztertuz eta gure hurbilpena aurkeztuz. VI.B. atalean hierarkien eraikuntzak dauzkan arazoei buruz ihardungo gara, ziklo eta erlatore bidezko definizioak nola tratatu ditugun azalduz. Hurrengo atalean LPPL-tik erauzitako HEBko adiera eta WordNet-eko kontzeptuen arteko lotura nola egin dugun aurkeztuko dugu. Erabilitako algoritmoak eta lortutako emaitzak ere eztabaidatuko ditugu ebaluazioa egin aurretik. VI.D. atalean hierarkiak aberastu eta trinkotzeko berebiziko garrantzia duen genus-desanbiguazioari buruz arituko gara, emaitzak eztabaidatu eta ebaluazioa ere aurkeztuz. Ondoren, VI.E. Atalean, HEBaren hierarkiak sendotzeko beste prozesua azaldu eta ebaluatuko dugu, hierarkien goiko geruzaren osatzeari buruzkoa. Bukatzeko kapitulu honetan egindako ekarpenak eta etorkizunerako lanak agertzen dira.

VI.A. Aurrekariak eta planteamendua

Kapitulu honetan erlazio-izaera paradigmatikoa beste eremu batean aplikatzen saiatuko gara, hiztegietako ezagutzaren erauzketan hain zuzen ere. Tesi honen helburu nagusietako bati, ingelesa ez diren baliabide lexikal egituratuen eraikuntza sendotzeko teknikak lantzeari, erantzuten dio. 80ko hamarkadan, ordurarteko sistemek lexiko irrigarriak erabiltzen zituztela eta, laborategietako jostailuetatik bizitza errealeko lexikoak eraikitzera pasatzeko beharra nagusitu zen (Boguraev &

Briscoe, 1989a). Lexikoen eraikuntzak, ordea, pentsatzen zena baino giza-ahalegin handiagoa eskatzen zuela ohartuaz, bide automatikoen erabilerara jo zen. Non bilatu ezagutza lexikala, ordea? Hiztegietan, noski. Garai horretan hasi ziren hedatzen hiztegietatik abiatuta Ezagutza-Base Lexikalak (EBL) eta Hiztegi Ezagutza-Baseak (HEB) eraikitzeko ahaleginak, gaur egun ere jarraitzen dutenak. Eremu zabal honetan ez dugu sakonduko hemen, baina hiztegietatik erauzitako informazioari egotzi izan zaizkion bi muga bai aztertuko ditugula, kapitulu honetako gaia izango dira eta:

1. EBL eta HEBen bizkarrezurra diren hierarkiak desanbiguatu gabe egotea
2. Hierarkia horiek txikiak, kalitate gutxikoak eta goi mailan koherentzia gutxikoak izatea

Guk landu dugun hiztegia *Le Plus Petit Larouss* (LPPL) hiztegia da, II. kapituluan aurkeztu duguna. LPPL hiztegiaren gainean egindako lanen ebaluazioak –ikus II. kapitulua eta LPPL-ren inguruan egin dugun lanari buruzko artikuluak (Artola, 1993; Agirre et al., 1994a; 1994c; 1994d; 1997)– erauzitako HEBaren egungo egoeraren ahulezia batzuk planteatu zituen, eta lau alorretan hobetu beharra aitortu genuen:

1. HEBaren hierarkia trinkotzea
2. Adierazpidearen aberasketa
3. Argumentu tipiko eta hautapen-murrizpenen erauzketa
4. Erlazio berriak inferitzeko erregelak

Tesi honetan lehenbiziko puntua landuko dugu. Alde batetik genus gehienak desanbiguatu gabe daude, eta beraz kontzeptuen taxonomian anbiguotasun handia dago. Beste aldetik hierarkia txiki anitz daude, bata bestearekin lotu gabe daudenak.

Hierarkia trinkotzeko egin dugun ahalegina aztertu aurretik, gai honetako aurrekariak aipatuko ditugu.

VI.A.1. Hierarkia-eraikuntza

LNPrako lexikoiak eraikitzeko orduan hiztegiak oinarri sendoa zirela, 80ko hamarkadan bultzada hartu zuen ideia da. Bultzada hori, hasiera batean Amsler-ek (1981) eman zuen, eta ondoren bi hiztegiren azterketa sakonetan gorpuztu zen: *The Webster's Seventh New Collegiate Dictionary* (W7, Gove, 1969) eta *Logmans Dictionary of Contemporary English* (LDOCE, Procter, 1978). Batzuen ahaleginak sintaxiari buruzko informazioaren erauztera aplikatu baziren ere (Boguraev & Briscoe, 1987), hiztegien analisiaren ekarpen nagusia hiztegietan gordeta zegoen egitura semantikoa

eskuratzeko saiakerak izan dira (batzuek aipatzeagatik, Michiels & Nöel, 1982; Calzolari, 1983; 1984; Chodorow et al., 1985; 1988; Markowitz, 1986; Binot & Jensen, 1987; Byrd et al., 1987; Byrd, 1990; Cohen & Loiselle, 1988; Vossen 1989; Vossen & Serail, 1990; Boguraev & Briscoe, 1989; Wilks et al., 1990; Briscoe et al, 1990; Castellón, 1992; Artola, 1993; Richardson, 1997; Rigau, 1998).

Informazio semantiko ezkutu hori erauzterakoan, definizio-esaldiak idazteko modu finko samarrak egotea izan zen abiapuntua. Horietako garrantzitsuena Aristotelerengandik ezaguna den *genus* eta *differentia specifica* bidezko definizioa. Definitzen ari garen kontzeptuaren mota edo klasea ematen digu genusak, eta differentiak mota horretako beste kontzeptuetatik bereizteko ezaugarriak. Idazteko modu finko horien arabera, definizioen sailkapena hiru multzotan egin daiteke[68] (Artola, 1993):

- Sinonimo bidezko definizioak
- Genus eta differentia motakoak
- Oinarri-sarrera batekiko erlazio sintaktiko bidezkoak

Definizio mota bakoitzetik erlazio ezberdinak erauzten dira. Sinonimia eta hiperonimia/hiponimia[69] dira garrantzitsuenak. Hirugarren definizio motan, lexikografoak sarrera definitzeko erlazio sintaktiko bat aukeratu du, guk erlatore-berezi izendatu duguna. Erlazio hori meronimia, instrumentala, etab. izan daiteke (Nakamura & Nagao, 1988; Bruce et al., 1992; Artola, 1993).

Proiektu askoren helburu nagusiena genusaren erauzketa automatikoa edo erdiautomatikoa izan da (Chodorow et al., 1985; 1988; Tsurumaru et al., 1986). Baina hiperonimo-erauzketa baino harantzago joan eta diferentia lantzen duten lanak ere badaude (Ahlswede, 1989; Wilks et al., 1990; Michiels & Nöel, 1982; Castellón, 1992; Artola, 1993; Agirre et al., 1994d; Richardson, 1997).

Genusaren garrantzia hiperonimiaren bidez mota-hierarkiak eraikitzeko balio izatetik datorkio batez ere (Amsler, 1981; Vossen & Serail, 1990). Hierarkiak garrantzitsuak dira, EBLen hezurdura izateagatik, egitura formala eman eta erredundantzia ekiditen dutenak, ezaugarrien herentziaren bitartez (Cohen & Loiselle, 1988; Briscoe et al., 1990).

---

[68] Beste sailkapenak ere badira, baina deitzeko moduaz gain, denak datoz bat gutxi gora bera. Vossen-ek (1989) Amsler-en (1981) antzera sinonimoei *cross-reference* bidezkoak, genus motakoei *non-complex genus*, eta hirugarren motakoak bitan sailkatzen zituen: *complex genus* eta *derivational*. *Complex genus* horiei beste autoreek izen ezberdinak eman dizkiete: *function words* (Nakamura & Nagao, 1988), *empty heads* (Chodorow et al., 1985), *disturbed heads* (Bruce et al., 1992) edo erlatore berezi (Artola, 1993).

[69] Lexikografian hala deitu ohi zaio adimen artifizialean klase/azpiklase edo IS-A bezala ezagutzen den erlazioari.

Hiztegietatik erauzitako hierarkiei buruzko kexuak, sakonera gutxikoak izatea (Chodorow et al., 1988; Wilks et al. 1990) eta hierarkia bakarra baino mihiztatutako hierarkiez (*tangled hierarchy)* osatuta egotea izan ohi dira (Calzolari, 1983). Wilks-ek, adibidez, hierarkien mihiztadura eta sakonera apala batez ere hitzen arteko hierarkiak eraikitzeagatik sortzen dela dio. Hitz horiek, genusak, desanbiguatuz gero, kontzeptuen arteko hierarkia edukiko genuke, mihiztadura gehiena desagertuaz eta sakonera-arazoak konpontzeko bidea emanaz (Bruce et al., 1992). Genusa desanbiguatzearen beharra autore gehienek aitortzen dute (Richardson, 1997; Ide eta Véronis, 1994). Adibidez zera diote Ide eta Véronis-ek: "*the undisambiguated hierarchy is unusable because, following the path upwards from saucepan, we find that saucepan can be a kind of leaf, which is clearly erroneous*"..

Baina nahiz eta genusa desanbiguatu, hierarkiei buruzko beste kritikak bere horretan darraite. Horietako nagusiena kontzeptu orokorrak definitzeko lexikografoen irizpide eza izan ohi da (Ide & Véronis, 1994). Horrek hierarkietan bigiztak sortu ohi ditu (lehenagotik ere jakina zena, ikus adibidez Amsler 1981; Chodorow et al. 1985; Vossen & Serail, 1990). Bestalde adiera asko hierarkiaren goialdean gelditu ohi dira, hierarkien arteko homogeneotasun faltak sortuaz. Orokorrean, kritikak hierarkietako goi aldeko adieretan zentratzen dira. Bestalde, hierarkiak hain lauak izatearen arrazoia ia eduki semantikorik ez duten genusen erabileratik dator, ekintza edo modu bezalakoak alegia. Azkenik, eta genusak alde batera utzita, erlatore berezien bidez egindako definizioak hierarkiatik kanpo gelditu ohi dira, edo bestela, adabegi isolatu batetik zintzilik (Bruce et al., 1992).

VI.A.2. Genusen adiera-desanbiguazioa

Esan bezala, hierarkiak erabili eta ezagutza-baseetan antolatzeko, beharrezkoa da genusaren desanbiguazioari ekitea. Aipatutako lan gehienetan, desanbiguazio hori eskuz egin beharko litzatekeela aitortzen dute, nahiz eta kasu batzuetan heuristiko ahul eta estaldura eskasekoak planteatu (Amsler, 1981; Chodorow et al. 1985; Vossen & Serail, 1990). Copestake-ek (1990) LDOCE hiztegiko kode semantikoetan oinarritutako teknika aipatzen du, baina ez du espezifikatzen ezta probatzen ere. Richardson-ek (1997) Microsoft-eko lantaldeak horretan diharduela dio, baina ez du inongo erreferentziarik ematen metodoaren edo egungo egoeraren inguruan.

Mexiko Berriko Unibertsitatean hainbat lan egin zituzten LDOCE-ko genusak automatikoki desanbiguatzearen helburuarekin (Bruce & Guthrie, 1991; Bruce et al., 1992). 1992.eko lanean LDOCE-ko adieren maiztasunak, kode semantikoak eta kode pragmatikoak erabili zituzten. Gainera, algoritmoak genus usu eta oso anbiguo batzuetarako (10 baino gutxiago) sistematikoki kale

egiten zuela ikusirik, genus horientzat adiera egokia eskuz aukeratu zuten. Horrelako teknika sinpleekin %90eko doitasuna lortu zuten.

HADan erabilitako teknikak (ikusi IV. kapitulua), batez ere hiztegietan oinarritutako erlazio-izaera erabiltzen dutenak, erabat aplikagarriak izanda ere, ez dira espero zitekeen bezainbat erabili izan. Arrazoia, beharbada, Wilks-ek eta hain heuristiko sinpleekin lortutako arrakasta izango litzateke.

VI.A.3. Hierarkia-trinkotzea

Aipatu ditugu arestian hiztegietatik erauzitako hierarkiei egiten zaizkien kritikak:

1. Hierarkietako bigiztak
2. Hierarkien sakonera apala eta goi mailako homogeneotasun falta
3. Erlatore bereziak hierarkian integratzeko arazoak

Halako arazoen aurrean, Ide eta Véronis-ek (1994) hiztegi ezberdinen arteko hierarkiak lotzea aurrerapauso bat izan zitekeela planteatzen dute, gehiegi sakondu gabe. Literaturan arazo hauen irtenbideak ez dira aipatu, norberak bere modura moldatuko balu bezala, edo arazoekin bizitzen ohitu izan balitz bezala. Salbuespen bat Mexiko Berriko taldeak (Bruce et al., 1992) eginiko lana da. Lan horretan, genusak desanbiguatzeaz gain hierarkiaren eraikuntzan azaltzen diren arazo horiei irtenbide integratu bat bilatzen zaie. Horretarako LDOCE-k dauzkan kode semantikoez baliatzen dira. Kode horiek hierarkia bezala antolatu eta beren ontologiako primitibo semantiko bezala hartzen dituzte. Ohiko hierarkia eraikuntzatik kanpo gelditzen diren kasuetan, hau da, genusak eduki semantikorik ez duenean, zikloak apurtzerakoan eta erlatore bidezko definizioak lotzean, adieren kode semantikoa erabiltzen da primitibo semantikoetara lotzeko. Bestalde, ordura arte loturarik gabe zeuden hierarkiak ere erlaziona daitezke, hierarkia horien erroa primitibo semantikoetara lotuaz. Primitibo semantikoen hierarkiak aterki baten funtzioa egiten du.

VI.A.4. Iturri lexikal eleanitzen arteko lotura

Hiztegi elebakarrek jaso duten arretaren aldamenean, hiztegi elebidunek izan duten ahanztura harrigarria dela esan liteke. Hiztegi elebidunen erabilerari buruzko artikuluak ez dira azaldu 90. hamarkadan ondo sartu arte, salbuespen gutxi batzuk kenduta (Byrd, 1990; Rizk, 1989). Ikerlarien interes falta baino arrazoi praktikoak izan dira gehienetan errudunak, hiztegi elebidun horiek eskuratzeko zailtasunak eta tipografia tratatu beharra alegia.

Hiztegi elebidunen ustiapenean, itzulpenaren inguruko informazioa bera baino, askotan, sarrerei buruzko bestelako informazioa izan da helburua. Kolokazio eta kode semantikoak erauzi izan dira

batez ere (Heylen et al., 1993; Helmreich et al., 1993), eta berriki hauekin sare semantiko elebakarra eraiki daitekeela erakutsi du Fontenelle-ek (1997). Lan gehienek, halere, itzulpen automatikorako lexikoen eraikuntzan laguntzeko erabili izan dituzte (Byrd, 1990; Helmreich et al, 1993; Knight & Luk, 1994; Klavans & Tzoukermann, 1995). Beste erabilera sofistikatuago batean, berriz, itzulpenaren aukeraketa automatikorako ere erabili izan dira (Rizk, 1989; Michiels, 1996).

Iturri eleanitzen arteko loturari dagokionez, ordea, lan gutxi dago. Helmreich-ek eta (1993) gaztelerazko hitzak LDOCE, WordNet eta PENMAN Upper Model baliabideak integratzen dituen PANGLOSS ontologiako kontzeptuei lotu nahi dizkiete, baina eskuz egin beharrean aurkitzen dira, nahiz eta automatizatzeko bidea aurreikusi. Bide hori Knight eta Luk-ek (1994) jorratuko dute, Collins-eko gaztelera-ingelesa hiztegiaz baliatu eta gaztelerako hitzak PANGLOSS-eko kontzeptuei lotzen dizkiete eta. Horretarako hiru heuristiko erabiltzen dituzte:

- Itzulpenean bi hitz edo gehiago badaude, eta hitz horiek WordNet-en sinonimoak edo guraso beraren semeak badira, lotu kontzeptu horiei.
- Itzulpen bakarra izanda, itzulpen hori monosemikoa bada, lotu kontzeptu horri.
- Itzulpena bakarra izanda, polisemikoa bada, erabili elebiduneko eta LDOCE-ko kode semantikoen arteko ezkontza adiera bat aukeratzeko.

Emaitzen aldetik ez dute asko esaten, ebaluaziorik ez baitute egin, baina 50.000 lotura proposatzen ditu beraien sistemak.

Garai berean antzeko ahalegin bat egiten dute Okumura eta Hovy-k (1994), kasu honetan japoniera-ingeles hiztegi bat erabiliaz ontologia berera lotzeko. Lehenbizi lotura posibleak lau multzotan sailkatzen dituzte, hitz-kontzeptu arteko erlazio posibleen arabera: hitz 1 kontzeptu 1, hitz 1 N kontzeptu, N hitz kontzeptu 1, N hitz M kontzeptu. Lehenengo eta hirugarren kasuan loturak besterik gabe egiten dituzte. Bigarren eta laugarren kasurako goiko heuristikoez gain aditzen kasuan azpikategorizazio eta argumentuen arteko ezkontza ere erabiltzen dute. Ebaluazioa nahiko iluna da, %100eko doitasuna adierazten baitute loturen %27rako, baina ez dute ematen besteetarako daturik, ez eta doitasun altuko kasuak automatikoki bereiz litezkeen edo ez azaldu ere. 15.000 hitzentzat aplikatzen dute algoritmoa.

Tesi-lan honen egilearekin batera argitaratutako artikulu batean (Rigau & Agirre, 1995), Rigauk ere bikoteen erabilera planteatzen du gaztelerako hitzak WordNet-era lotzeko, eta konbinazio posible bakoitzarentzat doitasun-neurriak hartzen ditu, frogatuaz bikoteak erabilgarriak direla zeregin

honetan. Lan hau, EuroWordNet proiektuaren barruan (Vossen, 1996; Vossen et al. 1997), zabaldu egingo da (Atserias et al., 1997), tesi-lan honetan aurkezten diren tekniken antzekoak gehituaz eta, ondoren, IV. kapituluan azaldutakoaren antzera bozkatuaz. Laginen bitartez, %85etik gorako doitasuna lortzen duten bikoteak bilatu, eta horiek bakarrik onartzen ditu, gaztelerazko 10.000 inguru izen lotuaz WordNet-era.

Aipatu beharra dago lan guztiek hitzak lotzen dituztela beste hizkuntza bateko ontologiara. Lan honetan ordea, adiera edo kontzeptuak lotzen saiatuko gara, orain ikusiko dugun bezala.

VI.A.5. Gure hurbilpena: LPPL hiztegi ezagutza-basearen aberasketa

Hierarkiei egindako kritika LPPL-tik erauzitako sare semantikoari ere aplika dakioke. Artolak (1993), lehen aipatu bezala, definizioen analisia egin eta hiru definizio mota bereizi zituen izenentzat: sinonimoak, genus eta differentia motakoak, eta erlatore bidezkoak[70]. Ez zuen genus eta erlatore berezien analisi soila egin, definizioaren gainontzeko informazioa eskuratzen ere saiatu zen, erlazio aberatsak dituen sare semantikoa edo HEBa eraikiaz[71]. Sare semantiko honen bizkarrezurra hiperonimia/hiponimia erlazioaz osatutako hierarkiak dira, eta horiek sendotzen saiatuko gara lan honetan.

	Adiera kop.
Isolatuak	2190
Hierarkiatan	2743
Gainean	840
Tartean	86
Hostoak	1817
Guztira	4933

Neurria	Hierarkia kop.
1	2190
2-9	1008
10-24	25
25-49	6
>49	1

Sakonera	Hierarkia kop.
1	2190
2	784
3	47
4	9

22. taula: LPPL HEBko izenen adieren kokapena hierarkiatan (ezkerrean), eta hierarkien neurri eta sakonerak.

LPPL-n dauden adieretatik, HEBan oinarrizkoenak besterik ez ziren kargatu hasiera batean (Artola, 1993). Izenen kasuan, LPPL-ko 13.740 adieretatik 4.933 sartu ziren HEBan. Horien genus asko desanbiguatu gabe zeudenez, hierarkiak oso txikiak dira eta adieren ia erdia isolatuta dago, inongo hierarkiatan egon gabe (ikus 22. taulako datuak).

Hierarkia horiek trinkotu ahal izateko egin beharrekoak hiru ataletan bana daitezke:

[70] Erlatore berezien artean sailkatu zituen *action* bezalako eduki semantikorik ez duten genusak ere.

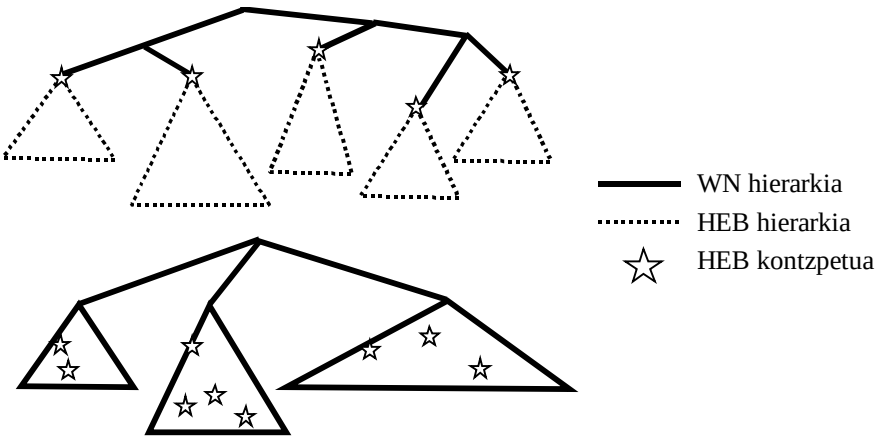
[71] Hemendik aurrera HEBari buruz hitz egitean LPPL-tik erauzitako HEBari buruz arituko gara.

1. Hierarkiaren eraikuntza osatu: bigiztak askatu eta definizio erlazionalak integratu.
2. Genusen adiera-desanbiguazioa.
3. Hierarkiaren gainaren berrantolaketa.

Erabiliko ditugun bitartekoei dagokionez azpimarragarria da lehenengo aldiz LPPL-n bertan ez dagoen ezagutza sartuko dela. HEBak orain arte bestelako ezaugarria izan du: bertan dagoen informazio semantiko guztia hiztegitik bertatik erauzi izan da. HEBaren eraikuntzan egindako prozesuak inplizitu zegoena esplizitu bihurtzen saiatu ziren. Orain aldiz, hierarkiak elkarren artean erlazionatu ahal izateko kanpoko ezagutza gehitzea ezinbestekoa izango da, bai eskuz bai automatikoki.

Bestalde, genusa desanbiguatzerakoan III. kapituluan aipatutako teknikak erabili nahi baditugu baliabide egokien faltan aurkitzen gara. Badaude teknikak hiztegiko bertako ezagutza erabiltzen dutenak (ikus III.A.2 edo VI.A.2 atalak), baina gure kasuan ez dugu ez kode semantiko edo pragmatikorik eta LPPL-ko definizioak motzak dira (3,82 hitz batez-beste). Hiztegitik kanpoko baliabideetara jotzean, bestalde, HAD erabili izan dugun Dentsitatea erabili ahal izateko ontologiaren bat beharko genuke. Frantseseko ontologia zabalik existitzen ez denez, WordNet erabiltzen saiatuko gara hiztegi elebidun bat zubi bezala erabiliaz.

WordNet ontologia desanbiguaziorako oinarri bezala erabiltzeak, bestalde, beste onura bat dakar, LPPL-ko hierarkiak lotzeko eta bigizta eta erlatore bereziak integratzeko bidea izan daiteke eta. Bi aukera aurreikusi ditugu: bata, arestian azaldutako Wilks-ek eta egindako lanaren antzera (Bruce et al., 1992), WordNet-eko gaineko aldea aterki bezala erabiltzea izango litzateke, LPPL-ko hierarkien erroak lotzeko WordNet-eko goi-kontzeptu eta erlazioak hartuz. Bestea, adierak banan-bana WordNet-i lotuz gero, LPPL-ko hierarkia alde batera utzi eta guztiz WordNet-eko hierarkia gureganatzea litzateke. Hurbilpen hauek 26. irudian azaldu ditugu.

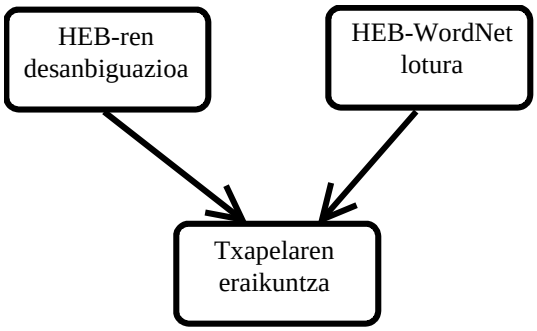


26. irudia: LPPL-ko hierarkiak trinkotzeko bi modu

LPPL-tik HEBa eratzean eduki dugun filosofiari atxikiaz (Artola, 1993), LPPL-ko egitura eta informazioa hobetsiko dugu, eta ondorioz lehenbiziko hurbilpena, *txapela kontzeptual* ere deitu duguna, aukeratu dugu. Txapela hori aurrera eramateko bi eginkizun dira beharrezkoak:

1. HEB-WordNet lotura: LPPL-ko adierak WordNet-eko kontzeptuei lotzea.
2. Txapelaren eraikuntza hierarkien gaineko adieren lotura erabiliaz.

Jadanik somatu daiteke HEBaren desanbiguazio, HEB-WordNet lotura eta txapelaren eraikuntzaren artean elkarrekintza konplexuak gertatzen direla, batak bestea egin ahal izateko informazio baliagarria eduki dezake eta. Elkarrekintza horien eragina aurrez asmatzea oso zaila denez, hipotesi sinpleenetik abiatu gara, etorkizunerako utziaz eredu elaboratuagoak. Hipotesi sinpleenean, HEB-WordNet lotura eta HEBaren desanbiguazioa independenteak dira, eta bata bestea kontuan hartu gabe eraman daitezke aurrera. Txapela eraikitzeko bai, garbi dago aurrekoen emaitzak kontuan hartu behar direla. Horrela irudikatuko dugu gure hurbilpena:



27. irudia: Prozesuen arteko dependentziak (hipotesia)

Hierarkiaren eraikuntza osatzea albo batera utzi dugu aurreko eskeman, izan ere informazioa erabili bai baina beste prozesuei ez die eskaintzen pisuzko informaziorik. Hurrengo ataletan joango gara banan-bana aztertzen eginbehar hauek.

VI.B. Hierarkiaren eraikuntza

LPPL-ren HEBan 10.506 izen daude, 13.740 adiera dituztenak. LPPL-ko hitzek duten adiera kopurua 23. taulan azaltzen zaigu. Adieren definizioen analisiaren emaitza bezala, definizioak horrela sailkatu ditugu:

- Definizio sinonimikoak
- *Genus et differentia* motako definizioak
- Definizio erlazionalak

Artolak LPPL-ko definizio guztiak automatikoki analizatu zituen (Artola, 1993; Agirre et al., 1994a). Analisi horren emaitzaren arabera gai izan gara definizioen %92, gutxi gora behera, sailkatzeko (ikus 24. taula). Sinonimo edo genusa ez denean LPPL-ko sarrera ezin izango dugu desanbiguatu, noski.

Adiera hitzeko	Izen kopurua
1	7639
2	1904
3	451
4	148
5	46
>5	18
Guztira	10.506

23. taula: LPPL-ko sarreren adiera kopurua

		%
Sinonimikoak	2836	20,6
Genus+differentia	7961	57,9
Erlazionalak	1773	12,9
Sailkatu ezinak	1085	7,9
Sinonimo/Genusa LPPL-tik at	85	0,6
Guztira	13740	100,0

24. taula: definizioen sailkapena

Definizio sinonimikoen kasuan, sinonimo bat baino gehiago erabil daitezke, eta hala da: 582 definiziotan 2 sinonimo erabiltzen dira, eta 18 definiziotan 3.

Bi arazoz arduratuko gara atal honetan, genus hierarkietan azaltzen diren bigiztak eta definizio erlazionalak hierarkian integratzeko dauden zailtasunak.

VI.B.1. Bigiztak

Genus eta sinonimoak oraindik desanbiguatu gabe daudenez, ezin da bigiztak benetakoak diren edo ez ikusi. Adibidez, nahiz eta *balle I 1*-en genusa *pelote* izan, eta *pelote I 3*-rena *balle* izan (beheko adibidean ikusten den bezala), genus horiek desanbiguatu arte ezin izango dugu jakin adieren hierarkian bigizta bat dagoen edo ez.

balle I 1 : petite pelote ronde pour joue
balle I 2 : projectile des armes à feu
balle I 3 : gros paquet
balle II 1 : enveloppe du grain des céréales

pelote I 1 : boule de fil roule
pelote I 2 : coussinet pour piquer les épingles
pelote I 3 : balle pour jouer

Bada-ezpada, definizioen artean bigizta posibleak bilatu genituen, hau da, genus eta sinonimo kateak jarraituaz gertatu zitezkeen bigiztak. Horrelako 59 balizko bigizta topatu ditugu, adibidez *balle I 1* eta *pelote I 3* artean.

Bigizta potentzialak aztertzean, konturatu gara lehenbizi desanbiguatu eta gero bigiztak bilatuz gero 59 bigizta potentzial horiek benetakoak zirela. Hipotesi hori aztertutako bigizta guztiekin betetzen

zenez, bigizta-bilatzaileak bigizta horiek desanbiguatu ditu. Goiko adibidean ere, bigizta egon dadin *balle I 1*-en genusa den *pelote*, bere hirugarren adieran erabilita egon behar da, eta *pelote I 3*-ren *balle* lehenbiziko adieran. Hau da, bigizta egotekotan *balle I 1* eta *pelote I 3* artean izan beharko litzateke, eta hala da.

Behin bigiztak desanbiguatuta, arazoa apurtzea da, hau da, bigiztako zein adiera dagoen goian eta zein behean. Irizpide bezala genus bezala orokorrena aukeratzea komeni da, hori izan dadin bestearen hiperonimoa. Genus orokorragoa zein den erabakitzeko, desanbiguatu gabeko hierarkiak eginez gero azpian adiera gehien edukitzea neurtu dugu, hau da, genus bezala zenbat aldiz azaltzen den. Goiko adibidean, *balle* 5 aldiz azaltzen da LPPL-ko definizioen genus bezala, eta *pelote* behin bakarrik, eta beraz *balle* hartu dugu *pelote*-ren hiperonimo bezala. Ondoren hiperonimo bezala dagoena non kokatu erabaki beharra dago, eta horretarako WordNet-i egindako loturaz baliatuko gara, VI.C.2 eta VI.E.1 ataletan ikusiko dugun bezala.

VI.B.2. Definizio erlazionalen integrazioa hierarkian

Definizio erlazionalak mota askotakoak izan daitezke, erlazio baten bidez definitzen direnak, edo genusak eduki semantikorik eduki ez eta definizioaren funtsa differentiak daukanean (Artola, 1993). Hiztegiko beste adieraren batekin hiperonimia ez den erlazio batez lotzen direnak hierarkiatik kanpo gelditzen dira. Genus hutsa dutenak oso zailak gertatzen dira desanbiguatzen, oso izen orokorrak izaten dira eta.

Artolak egindako tratamenduaren arabera, erlatore berezi bakoitzari hiztegiko adiera bat egokitzen zitzaion eta horrela sarrera adiera horren hiponimo bezala sartzen da hierarkian. Hori eginda ere askotan ez da nahikoa, hiperonimo horiek oso orokorrak izan eta informazio gutxi ematen dute eta. Beharrezkoa zaigu beste zerbaiti lotzea, erlatore bereziaren arabera:

1. Erlatoreari hiperonimia erlazioa dagokioenean, noski, posiblea da erlazionatutako hitza genustzat hartzea, eta VI.C.2 eta VI.D ataletan ikusiko dugun bezala desanbiguatu eta WordNet-era lozea. Horrelako erlatoreak dira *espèce de, genre de* eta *sorte de*. Adibidez:
   bolet I 1 : espèce de champignon

2. Erlatoreari meronimia erlazioa dagokionean, WordNet-eko meronimia erlazioa erabil daiteke erlazionatutako hitza desanbiguatu eta WordNet-era lotzeko. Sarrera eta erlazionatutako hitzaren adieraren baten artean erlazio meronimikoa badago WordNet-en, adiera hori aukeratu

eta erabilitako WordNet-eko kontzeptuetara lotuko ditugu[72]. Erlatoreak: *membre de* eta *élément de*. Adibidez:

aristocrate I 1 : membre de l'aristocratie

3. Erlatoreari kontzeptu bat dagokionean, horri lotu hiperonimo bidez. Adibidez *qui* erlatorea erabiltzen denean sarrera pertsona dela ziurta daiteke, eta beraz *personne*-ren lehenbiziko adieraren hiponimo bezala jar daiteke sarrera hori. Adibidez:

agitateur I 1 : qui excite à la révolte

Horrelakoak dira baita ere *état de, art de, action de, faculté de, manière de, qualité de, caractère de* eta *manque de*. Erlatore hauetako gehienek kontzeptu bera adierazten dute definizio guztietan, eta beraz zuzenean esleitu daiteke adiera bakarra, *qui*-rekin egin dugun bezala. Beste erlatore batzuek, ordea, ez dute beti kontzeptu bakarra adierazten. Adibidez, *état de* erlatoreaz definitutako sarrera atributua (*état*-en lehenbiziko adiera) edo egoera (*état*-en bigarren adiera) izan daiteke. Horrelakoetan, bi aukerak irekita utzi eta 1. puntuan bezala tratatuko dira. Adibidez:

âpreté I 1 : état de ce qui est âpre

4. Erlatore batzuen kasuan, ez dira beti erabili izan erlatore bezala, hau da, batzuetan genus arruntak dira. Horren adibidea da, adibidez, *partie de* erlatorea, batzuetan meronimia adierazteko erabiltzen dena, baina salbuespen batzuetan festa bat deskribatzeko (*partie*-ren 4. adiera). Halakoak dira *partie de*, *pièce de*, *ensemble de*, *réunion de* eta *groupe de*. Bi aukerak aztertu eta fidagarritasun hoberena lortzen duena aukeratuko da. Erlazioaren aukera 2. puntuan bezala aztertuko da, eta kontzeptuaren aukera Distantzia Kontzeptualaren bidez, VI.D.6 atalean azaldu bezala.

Laburbilduz erlatoreen kasuan tratamendua honakoa izan da:

1. Erlazio hiperonimikoa denean, erlazionatutako hitza genus bezala jarri. Ondoren beste genusak bezala tratatuko dugu genus hori.
2. Erlazio meronimikoa denean, erlazionatutako hitzaren eta sarreraren artean WordNet-eko erlazio meronimikoa bilatu.

[72] Ez dugu algoritmoa gehiago zehaztuko, baina funtsean VI.C.2.b) eta VI.D.6 ataletako algoritmoen antzekoa da.

3. Erlatoreari kontzeptu bat egokitzen zaionean, zuzenean jarri hiperonimo bezala adiera hori eta WordNet-en dagokion synset-a ere (LPPL-WordNet lotura ahalbideratzeko), edo adiera bat baino gehiago daudenean sarrerarekiko Distantzia Kontzeptuala erabili..
4. Erlazio edo kontzeptu izateko aukera dagoenean, biak probatu, eta indar gehien lortzen duena aukeratu.

Behin erlatoreak modu honetan tratatu eta gero, horrela definituta zeuden 1773 definizioetatik %78 desanbiguatu eta %63 WordNet-i lotu dira. 40 adierako lagina aztertu eta bai desanbiguazioa eta bai WordNet-eko lotura %90eko doitasunarekin egin dela ikusi dugu..

VI.C. HEB-WordNet lotura: iturri lexikal eleanitzen arteko lotura

Ezagutza-baseen artean zubiak ezartzea erabilgarri suertatzen da oso. LNP orokorrerako sistema batek behar duen ezagutza ez da normalean iturri bakarrean egoten, horretarako espreski eskuz sortutako ezagutza-base bat egin ez bada behintzat. Halakoetan ere beti da aberasgarria beste ezagutza-iturrietarako zubiak sortzea. Aurrerago aipatu dugun bezala HEBko adierak WordNet-eko kontzeptuetara lotu nahi ditugu.

Hizkuntza ezberdinetako hitzez ari gara, edo hobeto esanda hizkuntza ezberdinetako kontzeptuez. Hitzak hiztegi elebidunen bidez daude lotuta, beraz frantses-ingeles hiztegia erabiliko dugu. Baina aurrerago esan bezala hitzak ez dira nahiko, beraien adieren arteko loturak interesatzen zaizkigu eta. Azken finean interesatzen zaiguna HEBko frantses hitzen adierak WordNet-eko ingeles hitzen adierei lotzea.

Modu ezberdinetara egin daiteke hau guztia:

1. lehenbizi hiztegi elebidun eta WordNet-en artean loturak ezarri eta horren emaitza erabili HEB-WordNet lotura gauzatzeko. Kasu honetan elebidun “aberastu” bat –informazio gehigarria duena– erabiliaz gauzatuko litzateke
2. zuzenean HEB-WordNet zubiak eraiki elebidun “gordina” medio.

Aukera onena zein izango den ezin aurretik esan, eta beraz banan-bana aztertu dugu bakoitza. Lehenengo bidea jorratuz gero, HEB-WordNet loturaz gain hiztegi elebidun “aberastu” bat lor dezakegu.

VI.C.1. Elebiduna-WordNet lotura

Atal honetan frantsesezko hitzak WordNet-eko adierei lotzen saiatuko gara, ahal denean WordNet-eko kontzeptu bakarrari.

VI.C.1.a) Hiztegi elebiduna

Frantses-ingeles hiztegiak –*Oxford French-English Dictionary* edo OFED (OUP, 1989), ikus II. kapitulua– 21.322 sarrera dauzka. Sarrera bakoitzak jatorrizko hitzarentzat adiera bakarra edo gehiago eduki ditzake. Elebiduneko adiera bakoitzari azpisarrera deituko diogu atal honetan. Adibidez *maintien* izenaren sarrera bi azpisarreratan bana daiteke:

> *maintien n.m. (attitude) bearing; (conservation) maintenance*

> *maintien 1: n.m. (attitude) bearing*
> *maintien 2: n.m. (conservation) maintenance*

Hiztegi elebidunak 31.502 halako azpisarrera dauzka, horietako 16.917 izenei dagozkielarik. Lan honetan, esan bezala, izenetan zentratuko gara.

Azpisarreren barruan hainbat eremu azal daitezke: kategoria (derrigorrez), eremu semantikoa (aukerakoa, 20 eremutako bat izan daiteke, adibidez beheragoko adibideko *comm.*, komertziala), frantsesez dagoen argibidea (aukerakoa, adibidez goiko *attitude* eta *conservation*, edo beheko *ressources*), eta azkenik derrigorrezkoa den ingelesezko itzulpen-hitza edo hitz-zerrenda. Eremu semantikoa eta frantsesezko argibidea azpisarrera horretako itzulpena ulertzeko laguntza dira, testuinguru edo erabilpenari buruzko oharrak, hiztegiaren erabiltzaileari itzulpena hautatzean laguntzeko.

> *folie 1: n.f. madness*
> *provision 1: n.f. supply, store*
> *trésor 2: n.m (ressources) (comm.) finances*

Itzulpeneko ingelesezko hitzak anbiguoak izan daitezke, WordNet-en adiera bat baino gehiago izan dezakete eta. Frantsesezko sarreraren elebiduneko adierei, azpisarrerei, dagozkien WordNet-eko adierak zeintzuk diren jakin ahal izateko, desanbiguatzeko, algoritmoak (hiztegiaren erabiltzaileak bezala) testuinguruko informazioa behar du. Ez badugu testuinguruko inongo informaziorik, eta itzulpena anbiguoa bada, orduan ez da posible WordNet-eko adiera egokia topatzea. Itzulpena desanbiguatzen saia gaitezkeen kasuak honako hauek dira:

1. itzulpeneko hitzak adiera bakarra du WordNet-en
2. itzulpena hitz-zerrenda batez ematen da

3. itzulpena frantsesezko argibide batez lagunduta dator
4. itzulpena eremu semantiko batez lagunduta dator

Aurreko atalean azaltzen diren adibideen kasuan, *folie*-ren itzulpena polisemikoa da WordNet-en, eta beraz ezin da desanbiguatu; *provision*-ek bi itzulpen dauzka, eta beraz batabestearen testuinguru izan daitezke (2. kasua); *trésor*-ek itzulpen monosemikoa dauka, eta gainera frantsesezko argibide eta eremu semantikoaz lagunduta dator (1., 2. eta 3. kasua).

Hiztegi elebiduneko izenen azpisarrerak goiko kasuen arabera sailkatu ditugu. 25. taulan azaltzen den bezala, izenen azpisarreren %52rentzat ez dugu zer eginik, itzulpena WordNet-en ez dagoelako –%24– edo itzulpen bakarra eduki eta adiera anitz dituelako –%28–. Honen arrazoiak beherago aztertuko ditugu. 26. taulan trata daitekeen %48aren sailkapena egin dugu, kasuen arabera. Azpisarrera batzuk aldi berean egon daitezke kasu batean baino gehiagotan sailkatuta.

Itzulpena ez dago WordNet-en	4.081	%24
Itzulpen bakarra, adiera anitz	4.761	%28
1., 2., 3. edo 4. kasuak	8.075	%48
Guztira	16.917	%100

25. taula: izenen azpisarreren sailkapena (1)

1. kasua: adiera bakarra	5.039	%30
2. kasua: itzulpen anitz	630	%4
3. kasua: argibidea	2.954	%17
4. kasua: eremu semantikoa	1.067	%6

26. taula: izenen azpisarreren sailkapena (2)

WordNet-ek itzulpenen %76 bakarrik estaltzea kezkagarria zen. Arazo ezberdinek sortzen dute estaldura urri hau: itzulpena pluralean egotea, itzulpena izen-sintagma bat izatea, parentesiak, etab. Aniztasun berdina somatu genuen frantsesezko argibideetan ere. Halako itzulpen eta argibideei konplexu deitu diegu, tratamendu berezitua behar dute eta. Tratamendu hori, analisi sintaktiko zabal eta sendo baten faltan, ahal izan genuen bezala egin genuen, lematizazio eta heuristikoen bitartez. Itzulpen konplexuen tratamendua ebaluatzeko 50 azpisarrerako lagina hartu eta %88tan informazio zuzena ateratzeko gai garela ikusi dugu. Ez gara hemen gehiago arituko honetaz (ikus Rigau eta Agirre (1995) xehetasun gehiagotarako). Tratamenduaren emaitzen ondoren, 27. eta 28. tauletako sailkapena gelditzen zaigu.

Itzulpena ez dago WordNet-en	891	%5
Itzulpen bakarra, adiera anitz	6.440	%38
1., 2., 3. edo 4. Kasuak	9.586	%57
Guztira	16.917	%100

27. taula: izenen azpisarreren sailkapena (1')

1. kasua: adiera bakarra	5.119	%30
2. kasua: itzulpen anitz	958	%6
3. kasua: argibidea	3.702	%22
4. kasua: eremu semantikoa	1.365	%8

28. taula: izenen azpisarreren sailkapena (2')

Beraz, WordNet-en ez dauden itzulpenen kopurua %5era jaitsi da, eta trata ditzakegun azpisarreren kopurua %57ra igo. Horietatik 5.119k itzulpen monosemikoa dute, eta beraz 4.467 –%27–

azpisarreratan testuingurua erabili beharko dugula itzulpena desanbiguatzeko (2., 3. edo 4. kasuetan daudenak, kontuan izan azpisarrera bat kasu batean baino gehiagotan egon daitekeela).

VI.C.1.b) Emaitzak

Ideia sinplea da, HADan bezala, testuinguruari hobeto dagokion itzulpenaren adiera aukeratzea Dentsitate Kontzeptuala[73] erabiliaz. Hiztegi elebidunen kasuan testuingurua urria da, baina itzulpenarekin erlazio estuagoa izan ohi duena. Hiru testuinguru motetatik, ez dago oso garbi nola erabili eremu semantikoa, Dentsitate Kontzeptualak erlazio paradigmatikoez soilik baliatzen da eta. Horregatik oraingoz alde batera utzi dugu.

Frantsesezko argibidea duten hitzen kasuetan, argibidean azaltzen diren hitzak hiztegi elebidunean bertan begiratu eta horien ingelesezko ordainak dira testuinguru bezala erabiltzen direnak. Argibide hauek erabiliaz espero daitekeen desanbiguazioaren doitasuna kalkulatzeko esperimentu txiki bat egin genuen ia 60 azpisarrera erabiliaz. 29. taulan azaltzen den bezala, Dentsitatea erabiliaz %67ko doitasuna lortzen da. Doitasun kaxkar honen arrazoia argibideen ahultasunean bertan egon daiteke. Pista konplexuen tratamenduak eta hiztegi elebiduna erabili beharrak ere ez du laguntzen. Doitasun hori jasotzeko azterketa egin eta proba batzuk egin ondoren ikusi genuen 5 adiera baino gehiagokoak baztertuz gero doitasuna %83,3ra igotzen zela, estalduraren kaltetan.

	Estaldura %	Doitasuna %
zorizkoa	100	44,8
Dentsitatea	72,9	67,4
<6 adiera	50,8	83,3

29. taula: frantsesezko argibideetarako estimazioa

Itzulpen anitz daudenean, bakoitzak besteen testuinguru-papera joka dezake. Dentsitatea aplikatzean itzulpen guztiak batera desanbiguatzen dira. Espero daitekeen doitasuna estimatzeko halako 30 azpisarrera hartu genituen, eta 30. taulan azaltzen den bezala, doitasun eta estaldura oso onak lortzen dira. Oraingoan 5 adiera baino gehiagokoak baztertzen dituen heuristikoak ez du doitasuna apenas altxatzen.

	Estaldura %	Doitasuna %
zorizkoa	100	44,8
Dentsitatea	93,3	89,3
<6 adiera	73,3	90,9

30. taula: itzulpen anitzetarako estimazioa

[73] IV. kapituluko 20. ekuazioa, α=0,2 izanda, μ_{WN} eta erlazio meronimimoak erabiliaz.

Kasu bakoitzerako estimazioak egin ondoren, desanbiguatzeko algoritmoa izenen azpisarrera guztietarako egikaritu genuen. Eremu semantikoa alde batera utzi dugunez, 1., 2. eta 3. kasuetako azpisarrerak bakarrik desanbiguatu ahal izango ditugu. Azpisarrera bakoitza algoritmoak kasuz kasu aztertzen du: itzulpenak adiera bakarra badauka hori aukeratuko da, itzulpen anitz baditu Dentsitatea soilik erabiliko da, eta bestela, frantsesezko argibideak baditu, Dentsitatea eta heuristikoa erabiliko dira. Algoritmoa egikaritu ondoren azpisarrera guztien %43a lotzea lortu zen (ikus 31. taula).

	Azpisarrera kopurua		Doitasuna
Loturarik ez	9.676	%57	-
Lotura	7.241	%43	%95
1. kasua: adiera 1	5.119	%30	%99
2. kasua: itzulpen anitz	723	%4	%89
3. kasua: frantsesezko argibideak	1.399	%9	%83
Guztira	16.917		

31. taula: Elebidun-WN, lotutako azpisarrerak

Lotutako azpisarrera gehienak adiera bakarrari esker egin izan dira. Adiera bakarrekoetan, itzulpena konplexua denean %88ko doitasuna espero daiteke, eta bestela %100ekoa. Itzulpen konplexudunak 80 besterik ez direnez (ikus 26. eta 28. taulen arteko aldea), batez-beste adiera bakarrekoen doitasuna %99,8 da. Itzulpen anitz ditugunean %89koa eta frantsesezko argibideak dauzkatenentzat %83koa direnez, egindako 7.241 loturentzat batezbesteko doitasuna %95ekoa da.

VI.C.2. HEB-WordNet lotura

Atal honetan LPPL hiztegiko adierak WordNet-eko adieretara lotzen saiatuko gara. Horretarako erabiliko ditugun ezagutza-iturriak honakoak dira:

1. LPPL-ko adieraren definizioko hitzak, bereziki genusa eta sinonimoa[74].
2. Hiztegi elebiduna: bai bere horretan edo aurreko atalean WordNet-i lotu diogun bertsioan. Hiztegi elebiduna izango da LPPL-ko hitz frantses eta WordNet-eko ingelesezko adieren arteko zubia.
3. WordNet-eko adieren arteko Dentsitate Kontzeptuala

Hurbilpenaren filosofiaren atzean LPPL-ko genusen adiera-desanbiguazioan erabilgarria izatea dago. Horretarako komeni zaigu ahal den bezainbeste adiera WordNet-i lotuta edukitzea, baita definizio bakoitzean azaltzen den genusa lotzea ere. Estaldura ahal den zabalena eta errore ahal den

[74] Atal honetan, bai eta VI.D atalean ere, genusei buruz arituko gara, genus edo sinonimo bidezko definizioa den arduratu gabe. Izan ere genus eta sinonimo bidezko definizioak berdin tratatu ditugu.

txikiena nahi dugunez, ez zaigu arduratuko adiera bati WordNet-eko adiera anitz lotzen badizkiogu, betiere adiera zuzena horien artean badago.

Lotura egiteko algoritmo ezberdinak erabil daitezke, eta horietako batzuk ikertu ditugu. Hemen emaitza nabarmenenak azalduko ditugu. Hasteko, Dentsitatea erabili gabe, heuristiko pare baten eraginkortasuna probatuko dugu. Ondoren hiztegi elebiduna eta Dentsitatea erabiliz saiatuko gara, hiztegi elebidun hutsarekin lehenbizi, eta WordNet-i lotutako elebidunarekin gero. Bukatzeko hiztegi guztiari konbinazio hoberena aplikatuko diogu. Algoritmo horiek aztertu aurretik laginari buruz jardungo gara.

Algoritmo bakoitzaren eraginkortasuna neurtzeko 27 adierako lagina erabili dugu. 27 adiera horien definizioetan 31 genus eta sinonimo azaltzen dira. Lagina zabalagoa zen, baina genus gabekoak eta elebidunean aurkitzen ez zirenak alde batera utzi ditugu. Hiztegi elebidunak daukan beste arazo bat adieraren esanahiari dagokion itzulpena azaltzen ez denean gertatzen da. Ezinezkoa da automatikoki igartzea hori gertatzen ari dela. Eskuz, lagineko 7 kasutan gertatzen dela ikusi dugu. Algoritmoen emaitzak bi eratara emango ditugu, elebiduneko zulo horiek kontuan hartuta eta kontuan hartu gabe. Bestalde, inplementazio-arrazoiak medio, definizio baten genus edo sinonimo bat baino gehiago daudenean bakoitza aparte lotzen da, nahiz eta gero desanbiguatzerakoan informazio hori kontuan izan. Horregatik 31 genusentzat ematen dira emaitzak. Definiendum batek, beraz, bi lotura jaso ditzake, genus bakoitzetik bat.

VI.C.2.a) Hiperonimia eta beste heuristikoak

Dentsitate Kontzeptualaz gain, hiru heuristiko ere probatu ditugu. Lehenbizikoan, definiendumak hiztegi elebidunean ordain bakarra badu, eta ordain horrek WordNet-en adiera bakarra baldin badu, orduan zuzenean lot dakioke WordNet-eko adiera hori definiendumari. Heuristiko hau algoritmo guztietan erabili da.

Bigarren heuristikoari dagokionez, definiendum eta genusen itzulpenak aztertzean, batzuetan ingelesezko hitz bera bietan azaltzen zela ohartu ginen. Halakoetan, hitz hori desanbiguatu ahal izateko testuingururik azaltzen ez zenez, ingelesezko hitz horren adiera guztiak esleitzen genizkien definiendum eta genusari. Azpian ikus daitekeen bezala, adibidez, *partie*-ren 4. adieraren genusa *jeu* da, eta bai *partie* eta bai *jeu game* bezala itzul daitezke ingelesera. Hori horrela izanda *game*-ek WordNet-en dituen adiera guztiak esleituko zaizkio *partie I 4* adierari, eta gainera jakingo dugu *partie*-ren elebiduneko 2. adiera eta *jeu*-ren elebiduneko 1. adiera erabili direla definizio horretan.

partie I 4 : jeu , divertissement en commun

partie 1: part
partie 2: (cartes, sport) game
partie 3: (jurid.) party

jeu 1: game
jeu 2: (amusement) play
jeu 3: (au casino etc.) gambling
jeu 4: (théâtre) acting
jeu 5: (série) set
jeu 6: (de lumière, ressort) play

	Estaldura	Doitasuna
Itzulpen komuna	%16	%100

Bestalde, genusa definiendumaren hiperonimoa denez frantsesez, beraien itzulpenentzat WordNet-en gauza bera gertatzen den edo ez azter daiteke. Horrela, genusaren itzulpenetako baten adieraren bat definiendumaren ordezko baten hiperonimoa bada WordNet-en, adiera horiek esleituko dizkiegu frantseseko definiendum eta genusari.

	Estaldura	Doitasuna
Hiperonimia	%42	%85

VI.C.2.b) Dentsitate Kontzeptuala hiztegi elebiduna erabiliaz

Hiztegi elebidun gordina erabiliaz, posible da definiendum, genus eta definizioko gainontzeko hitzen ordainak lortzea. Definiendumari dagokion WordNet-eko adiera desanbiguatzeko Dentsitate Kontzeptuala erabili dugu, testuinguru bezala definizioko hitzen ordainak erabiliaz. Definizioko hitzen artean genusak garrantzi gehiago duenez erlazio paradigmatikoari dagokionez, Dentsitatea kalkulatzeko orduan nabaritasun-pisu (*nb*) bat esleituko diegu definizioko itzulpenen adierei: genusaren edo definiendumaren adierak badira *nb*=1, eta definizioko beste hitzenak badira *nb*=0,1 (ikus 21. ekuazioa, non *A*-n definiendum, genus eta definizioko gainontzeko hitzen itzulpenen WordNet-eko adierak dauden, eta *Z* WordNet-eko edozein zuhaitz, eta ikus IV. kapituluko 20. ekuazioa ere). Hainbat proba egin ondoren emaitza hoberenak pisu horientzat lortu ziren.

$$\text{dentsitate}(Z, A) = \text{dentsitate}(Z, a_Z) \quad \text{non } a_Z = \sum_{c \in A \cap Z} nb_c \tag{21}$$

	Estaldura	Doitasuna
Dentsitatea	%87	%74

Definizioko hitzek testuinguru gutxi eskaintzen dutenez (LPPL-ko izenen definizioek batez-beste 3,82 hitz dauzkate), testuinguru hori zabaltzea pentsatu dugu. LPPL-n hitzak erlazionatuta daude: bata bestearen genus bezala, bata bestearen sinonimo bezala, erlatore berezi baten bidez lotuta edo

definizio berdinean azaltzen direla eta. Definiendum-ari WordNet-eko zein adiera dagokion erabaki ahal izateko, genusarekin erlazionatutako hitz horiek pista eman dezakete, nahiz eta ez jakin genusaren zein adiera den benetan definiendumari lotua dagoena. Alderantziz ere, genusari dagokion WordNet-eko adiera zein den erabakitzean ere, definiendumarekin erlazionatutako hitzek laguntza eskain dezakete. Testuinguru hori guztia erabiltzeko orduan aukera bat baino gehiago genituen, eta esperimentu batzuen ondoren onena Dentsitatea birritan kalkulatzea zela ikusi genuen: genusa desanbiguatzeko definiendumaren testuingurua erabili eta definienduma desanbiguatzeko genusaren testuingurua erabili, bietan definizioko gainontzeko hitzak ere erabiliaz. Lehen bezala nabaritasun-pisua erabili dugu: *nb*=1 testuinguruko genus edo sinonimo bezala erlazionatutako hitzei, eta nb=0,1 erlatore bidez edo definizioan egoteagatik erlazionatutako hitzei.

	Estaldura	Doitasuna
Dentsitatea (erlazioekin)	%97	%63

Doitasuna erlazionatutako hitzak erabili gabe baino okerragoa da, baina estaldura %100era hurbiltzen da. Orain lehen baino datu gehiago dauzkagu, baina sistemak ez ditu guztiz aprobetxatzen. Hurrengo atalean (ikus VI.D) azalduko dugun bezala, estaldura garrantzitsua izango da LPPL-ren genus desanbiguazioa aurrera eramateko, eta horren alde egin genuen aukera. Teknika honen emaitzak hobetzen saiatuko gara hurrengo ataletan.

VI.C.2.c) Dentsitate Kontzeptuala elebiduna-WordNet lotura erabiliaz

Atal honetan hiztegi elebidun landu gabea erabili ordez, aurreko atalean WordNet-ekin lotura duen bertsioaz profitatzen saiatu gara. Aurreko atalean frantseseko hitz bati dagozkion WordNet-eko adierak lortzeko, lehenbizi elebidunean itzulpena lortu eta ondoren itzulpen horien adierak begiratzen genituen. Orain, ordea, elebidun landuan zuzenean atzi daitezke WordNet-eko adierak. Gogoratu bertsio honetan adiera batzuk desanbiguatuak izan direla, beste batzuk baztertuaz. Esperimentuan erlazionatutako hitzak ere erabili ditugu, eta doitasuna nabari igo da.

	Estaldura	Doitasuna
Dentsitate (erlazioekin eta elebidun aberastuarekin)	%97	%70

VI.C.2.d) Konbinazioa

Orain arte aipatutako heuristikoak eta Dentsitatea, ordena jakin batean egikaritzea erabaki dugu:

1. Itzulpen bakarra eta adiera bakarrekoa badu definiendumak, lotu horrekin.

2. Bestela, definiendum eta genusak itzulpen bera badute, horren WordNet-eko adiera guztiekin lotu.
3. Bestela lotura hiperonimikoak erabili.
4. Bestela erabili Dentsitatea VI.C.2.c) atalean bezala.

	Estaldura	Doitasuna
Konbinazioa	%97	%77

Bestalde, bai elebidun eta bai WordNet begiratzean sortzen ziren arazo batzuk ere tratatu izan dira: lematizazioa, frantseseko hiztegiko sarrera konplexuak, etab. Arazo horiek ebazten saiatu ondoren doitasuna are gehiago igotzen da, eta estaldura %100era heltzen da. Emaitzak aztertzean konturatu gara LPPL-ko adiera batzuk ez daudela elebidunean, hau da, frantses hitzen adiera guztiak ez daudela behar den bezala itzulita elebidunean. Halako adiera bat lotzean emaitza okerra lortu dugu ziurrenik. Laginean behar bezala itzulita dauden adierak uzten baditugu soilik –31 definizioetatik 24– doitasuna %88ra igotzen da.

	Estaldura	Doitasuna
Konbinazioa + tratamendua	%100	%82
” + itzulpen zuzenekoak soilik	%100	%88

Laginerako emaitzak ikusita, azkeneko algoritmo hau LPPL hiztegi osoarekin egikaritu genuen. Estaldurari buruzko datuak, eta loturaren iturriari buruzko datuak 32. taulan azaltzen dira. Aipatu beharra dago laginean estaldura %100 zenean, hiztegi guztirako %64,8 dela. Laginean definizio erlazionalak, genus gabekoak eta hiztegi elebidunean topatu ezin zirenak alde batera utzi genituen, baina hiztegi osoa tratatzean, horrelakoak (adieren %35,1) ezin WordNet-era lotu, noski.

Adiera kopurua	13740	
Loturarik ez	4832	%35,1
Definizio erlazionalak	1462	%10,5
Genusik ez	1085	%7,9
Elebidunak kale	2285	%16,6
Lotura	8908	%64,8
Lotura bakarra	4976	%36,2
Elebiduneko adiera bakarra	1828	%13,3
Elebiduneko adiera anitz	2104	%15,3

32. taula: LPPL-WN emaitza orokorrak

Definizio erlazionaletako batzuek –280– lotura lortu dute, adiera bakarreko itzulpena zuten eta. Kasu gehienetan bai definienduma eta bai genusa lotu dira, baina lotutako adieren %6,21ean definiendumak ez du loturarik eta %18,48an genusa da loturarik ez duena. Loturen jatorriari buruzko datuak 33. taulan daude jasota.

1. adiera bakarra 2. sinonimia	%7,18
3. hiperonimia	%42,37
4. Dentsitatea	%50,45

33. taula: loturen jatorria

VI.C.2.e) Nabarmentasunean oinarritutako hedadura

Goiko metodoen konbinazioarekin, hiztegiko izenen %64 WordNet-eko adierei lotzeko gauza gara. IV.C.3.a) atalean aipatu dugun bezala, WordNet-eko adiera jakinda WordNet-eko etiketa semantikoa ere ezagut dezakegu. Etiketa semantiko horiek oso interesgarriak dira, adiera baten eremu semantikoa adierazten digute eta (ikus II. kapitulua). Gainera, oraindik lotu gabe dagoen hiztegiko adiera baten etiketa semantikoa jakinez gero, errazagoa da hari dagokion WordNet-eko adiera (bat edo gehiago) lortzea. Adibidez, orain ikusiko dugun teknikaren bidez, posiblea da *adulte I 1* adierari *noun.person* etiketa semantikoa esleitzea (ikusi behean), pertsonei buruzko definizioa delakoan. *Adulte*-k itzulpen bakarra du, *adult*, eta honek bi adiera ditu WordNet-en, pertsonei buruzkoa bata eta animaliei buruzkoa bestea, eta horrela *adulte I 1* adierari WordNet-eko *adult/1* adiera dagokiola ondorioztatu ahal izango dugu.

adulte I 1 : arrivé à l'âge d'homme ⇒ noun.person

adulte: adult

adult 1: <noun.person> adult, grownup: a fully developed person from maturity onward
adult 2: <noun.animal> adult: any mature animal

Orain arte lotutako adierei esker, etiketa semantiko jakin bati dagozkion frantseseko adierak eta definizioak bil ditzakegu, etiketa semantiko bakoitzarekin erlazionatutako hitz multzoa lortuz. Behin hitz multzo horiek bilduta Yarowsky-ren (1992) teknika erabil dezakegu etiketa semantiko bakoitzaren hitz nabarmenenak ezagutzeko, eta hitz nabarmen horiek erabilita oraindik lotu gabe dauden definizioak etiketatzeko (ikus III.A.4 atala).

Goian aipatutako prozesu horri esker aurreko atalean lotu gabe gelditu diren definizioak lotzeko gai izan gaitezke. Metodo honen doitasuna estimatu ahal izateko lagin bat hartu dugu, etiketatu gabeko 40 adieraz osatua, eta %70eko doitasuna neurtu dugu. Nabarmentasun neurri handiagoa zuten kasuetan emaitza hobeagoak lortzen diren edo ez neurtu nahi genuen. Nabarmentasunaren balio minimo bat exigituz gero, doitasuna hobetzea lortzen da estalduraren kaltetan (ikus 34. taula).

Nabarmentasun minimoa	doitasuna	estaldura
0	%70	%100
1	%78	%88
2	%81	%65

34. taula: nabarmentasunaren bidezko hedadura (lagina)

Hiztegi osora aplikatzean estaldura ez da hala ere %100era heltzen. Definizio batean, gerta daiteke hitzetako bat ere ez egotea etiketa semantiko baten nabarmentasun-zerrendetan. Horrela denean ez dago adiera horri etiketa semantikoa esleitzeko oinarririk. WordNet-eko adierak esleitzeari dagokionez, definiendum edo genuserako elebidunaren bidez topatutako WordNet-eko adiera guztietatik, etiketa semantikoa betetzen dutenak bakarrik aukeratuko dira. Elebidunean dagoen estaldura eskasa dela eta, nahiz eta etiketa semantikoa eduki, ezin izan dira beti definiendum eta genusak WordNet-eko adierei lotu, baina bai bata edo bestea.

VI.C.2.f) Emaitzak

Behin nabarmentasunaren bidezko loturak eta erlatore berezien bidezkoak gehitu ondoren, hiztegiko adieren %87 lotzea lortu dugu, 35. taulan azaltzen den bezala. Aurreko ataletan azaldutako lagin ezberdinetarako emaitzak kontuan izanda, erlatore bidezko loturentzat %90eko doitasuna espero dezakegu, Dentsitate bidez lortutakoentzat %82, eta nabarmentasuna erabiltzen dutenetarako %70. Hiru neurrien batezbestekoa eginez gero WordNet-en loturentzat espero daitekeen doitasuna %80koa da.

WN-era lotu gabe	1824	%13
WN-era lotuta	11916	%87
Erlatore	1114	%8
Dentsitate	8615	%62
Nabarmentasun	2187	%16
Guztira	13740	

35. taula: LPPL-WN loturaren emaitzak

VI.C.3. Ebaluazioa

Hiztegi elebiduneko hitzak WordNet-eko adierekin lotzean lortu zen estaldura nahiko apala da (%43), nahiz eta doitasun altua lortu (%95). Estaldura zabalagoa lortzeko hiztegi elebiduneko eremu semantikoak erabiltzeko sistema bat pentsatu beharko litzateke. Bestalde bikoteen azterketa egin zitekeen, doitasun apalagoa edukita ere lotura asko egitea posible egiten baitu (Okumura & Hovy, 1994 ; Rigau & Agirre, 1995; Atserias et al. 1997). Guk erabilitako hiztegi elebiduna aberatsagoa eta zabalagoa izango balitz ere (aipatutako lanetan erabilitakoen antzera) lotura gehiago lortuko

genituzke, bai frantsesezko hitz gehiago egongo liratekeelako, baita itzulpen anitzen bidez desanbiguatzeko aukera gehiago egongo liratekeelako ere.

Elebidun zabalagoarekin HEB-WordNet emaitzak ere hobetuko lirateke. Alde batetik estaldura zabalagoa lortuko litzateke, bai hiperonimo eta bai Dentsitatearentzat bereziki (doitasun orokorra hobetuaz), eta bestetik LPPL-ko adieraren baterako itzulpenik ez egotea errore-iturri denez, elebidun zabalagoarekin halako gutxiago gertatuko lirateke. Dentsitatearen bidez lortutako LPPL-WordNet loturetan, adibidez, doitasuna %82tik %88ra igoko litzateke, lehen aipatu dugun bezala.

Bestalde, elebiduna-WordNet lotura erabiliaz, nahiz eta elebidunaren %43 besterik ez eduki lotuta, Dentsitatearen doitasuna %63tik %70era igotzen da. Beraz, merezi du lehenbizi elebiduna lotzea eta lotura hori erabiltzea HEBa lotzeko. Elebiduna-WordNet loturan lehen aipatutako bikoteak erabiliz gero, HEB-WordNet loturan ere emaitza hobeak lortuko liratekeela espero dugu. Dena den, esandako lanetan bikoteak hizkuntza bateko hitzen eta ingelesezko ontologiaren arteko loturak egiteko erabili dira, ez adiera edo kontzeptuen eta ingelesezko ontologiaren artean, guk nahi dugun bezala.

Bikoteen erabilerari buruz aipatu behar da (Atserias et al, 1997) lanean elebiduneko adierak ez direla kontuan hartzen, Okumura eta Hovy-renean ez bezala (1994). Elebiduneko adierak kontuan hartzeak elebiduneko argibideak eta itzulpen anitzak erabiltzea ahalbideratzen du. Bestalde, elebidunean dagoen informazioa (kolokazio, kode semantiko, frantsesezko argibidea, etab. (Fontenelle, 1997)) WordNet bera aberasteko erabil daiteke, eta HEB-WordNet lotura egitean, itzulpenerako erabili den elebiduneko adiera zein izan den gordetzen denez, LPPL-ko adierak ere aberats daitezke. Lotura honen inplikazio guztiak ez ditugu sakonean aztertu. HEB-WordNet lotura medio, noski, batek duen informazioa erabiliz bestea aberats daiteke, VI.E atalean ikusiko dugun bezala..

VI.D. HEBko kontzeptuen desanbiguatze lexikala

Atal honetan LPPL-ko definizioetan azaltzen diren genus, sinonimo eta erlatore berezien bidez lotutako hitzen desanbiguazioari buruz arituko gara. Erlatore bereziek jaso duten tratamendua aipatu izan dugu jadanik VI.B.2 atalean, baina erlatore batzuek hiperonimo erlazioa islatzen zutenez, horientzat genusaren papera jokatzen zuen hitza topatu eta genus bezala tratatu dugu. Bestalde, genus eta sinonimoen desanbiguazioa era berdinean tratatu dugu, VI.C atalean bezala. Izan ere sinonimo asko ez dira halakoak, hiperonimoak baizik, hau da, diferentzia gabeko genusak dira. Gainera, sinonimoak izanda ere, Dentsitatearen bidezko desanbiguazioa aproposa da.

HADaren kapituluan ez bezala, orain gure helburua ez da Dentsitatearen egokitasuna frogatzea bakarrik. Beraz, Dentsitate Kontzeptualaz gain, bestelako heuristikoak ere erabiltzea erabaki dugu, heuristiko guztiak bozketa bitartez konbinatuz. Desanbiguaziorako metodoak bi multzotan bana daitezke: barne-ezagutza soilik darabiltenak eta kanpokoa ere badarabiltenak. Lehenbiziko taldean leudeke lehen adiera, agerkidetza eta bektoreetan oinarritutako teknikak, LPPL hiztegiko informazioan, definizioetan, oinarritutakoak. Bigarrenean LPPL-WordNet loturatik erauzitako etiketa semantikoez osatutako bektoreak eta Dentsitate Kontzeptuala ditugu, LPPL-ko informazioaz gain WordNet ere erabiltzen dute eta. Hurrengo ataletan banan-bana azalduko dugu teknika bakoitza, eta ondoren konbinatzeko modua eta lortutako emaitzak ikusiko ditugu.

VI.D.1. Adieren ordena (OR[75])

Heuristiko honek adierak garrantziaren arabera ordenatuta daudela suposatzen du, hau da, adiera erabilienak arraroagoak baino gehiagotan emango direla. Lehenbiziko adierari 1 esleituko dio, bigarrenari 0,9, etab.

VI.D.2. Definizioko hitzen ezkontzea (EZ)

Lesk-ek (1986) erabilitako teknika bera aplikatzen dugu hemen. Genusaren adiera bakoitzeko definizioetako hitzak definiendumaren definiziokko hitzekin[76] konparatu, eta hitz berdin[77] bakoitzeko puntuazioa gehitu egiten da. Hitz gehien amankomunean dauzkan adiera da aukeratuko lukeena, horri 1 pisua emanaz, eta besteei normalizatutako balioa.

VI.D.3. Agerkidetza arruntak (AA)

Bi hitz definizio berean gertatzen direnean agerkidetza bikotea dugula esango dugu. LPPL-tik horrelako agerkidetza bikoteak atera genituen. Agerkidetza bikote baten indarra neurri ezberdinez isla daiteke: maiztasun gordina, Elkarren Arteko Informazioa (EAI, ikus III. atala) edo *Association Ratio* (AR, Resnik, 1992) deritzana. Gure kasuan, emaitza hoberenak agerkidetza arrunten kasuan AR erabiliaz lortu ditugu.

Hitzen ezkontzarekin bezala, definiendum eta genusaren adiera bakoitzaren definizioa konparatuko ditugu hemen ere, baina ez zuzenean ea hitzak berdinak diren begiratuaz, zeharka baizik. Definizio eta genusaren adiera baten arteko erlazioaren indarra (22. ekuazioko $GA(D,G_i)$) neurtzeko,

[75] Teknika bakoitzari laburdura bat emango diogu, tauletan erabiltzeko.

[76] Hitzak berdinak direla erabakitzeko lemak erabili ditugu, ez hitz-formak. Bestalde funtzio-hitzak (izen, adjektibo eta beste kategoria irekietakoak ez diren hitzak, adibidez *et, le,* etab.) ez dira kontuan hartu, gainontzeko tekniketan bezala.

[77] $\mathrm{AR}(v,w) = \Pr(v,w) \cdot \mathrm{EAI}(v,w) = \Pr(v,w) \cdot \log \dfrac{\Pr(v,w)}{\Pr(v) \cdot \Pr(w)}$

definiendum eta genusen definizioetako hitzen arteko agerkidetzak aztertuko ditugu, beren pisuak batuaz (gp-k bi hitzen arteko agerkidetza indarra adierazten du).

$$\mathrm{GA}(D, G_i) = \sum_{w_j \in D \wedge w_k \in G_i} \mathrm{gp}(w_j, w_k) \tag{22}$$

VI.D.4. Agerkidetza bektoreak (AB)

Wilks-ek eta (1990) agerkidetza bektoreak erabiltzea proposatu zuten agerkidetza arrunten ordez. III.A.2 atalean azaldu dugun bezala, hitz batekin azaltzen den agerkidetza guztien bidez osatzen da hitzaren agerkidetza bektorea. Bai definiendum eta bai genusaren adieraren bektorea eraikitzeko beraien definizioetan dauden hitzen bektoreak batu besterik ez dago. Genusaren adieretatik definiendumaren bektoretik gertuen dagoena aukeratuko dugu, bektoreen arteko hurbiltasun-neurriren bat erabiliaz (ikusi 23. ekuazioa).

$$\mathrm{GB}(D, G_i) = hurbil(\vec{b}_D, \vec{b}_{G_i}) \tag{23}$$

Hainbat proba egin genituen hurbiltasun-funtzio ezberdinak erabiliaz (distantzia euklidearra, kosinua edo bektoreen arteko puntu-biderkaketa *–dot product–*), baita agerkidetzaren indarraren neurri ezberdinekin (maiztasun gordina, EAI eta AR) ere. Emaitza onenak EAI-z osatutako bektoreen arteko kosinuak eman zituen.

VI.D.5. Etiketa semantikoen bektoreak (SB)

Aurreko atal batean (VI.C.2) ikusi dugu nola lotu LPPL-ko definizioak WordNet-eko adiera eta etiketa semantikoei. Etiketa semantikoei dagokionez, definizio batek pisu bat lortzen du etiketa bakoitzeko (ikus VI.C.2.e). Etiketa bakoitzeko pisu horiek bektore moduan antola daitezke, eta behin bektore moduan antolatuta, aurreko atalean bezala, bektoreen arteko hurbiltasuna neurtu. Hurbiltasun neurri ezberdinak probatu ondoren, kosinuarenak eman zituen emaitza onenak.

VI.D.6. Distantzia Kontzeptuala erabiliaz (DK)

LPPL-ko definizioak WordNet-eko adierei lotu izan ditugun bezala (ikus VI.C.2.b), posiblea da hiztegi elebiduna eta WordNet erabiltzea ere definiendum eta genusaren adiera bakoitzaren arteko erlazio-izaera bilatzeko. Erlazio-izaera hori bilatzeko orduan erabili beharreko testuingurua finkatu behar da. Kasu sinpleenean, definiendum eta genusaren adierako definizioaren genusaren arteko Dentsitatea neur dezakegu.

Horretaz gain definizioetako gainontzeko hitzak ere erabiltzea badago, hala nola erlazionatutako hitzak (VI.C.2 atalean bezala). Aukera ezberdinak eta nabaritasun-pisu eskema ezberdinak ere probatu ondoren, ikusi dugu bai aukera sinpleena edo bai bitxiena hartuta ere ez dagoela hobekuntza handirik. Gainera, bi kontzepturen arteko hurbiltasun kontzeptuala neurtu behar denez soilik, Dentsitatearen ordez Distantzia Kontzeptuala (ikus III. kapituluko 13. ekuazioa) erabiltzea erabaki dugu.

VI.D.7. Heuristikoen arteko bozketa

Heuristiko guztien emaitzak konbinatzeko orduan, ikasketa automatikoan (*machine-learning*) hainbesteko arrakasta lortu duen sistema sinplea erabili dugu: bozketa. Dietterich-ek (1997) dioen bezala, "*while it may appear that more intelligent voting schemes should do better, the experience in the forecasting literature has been that simple, unweighted voting is very robust*". Beraz, heuristikoen bozketa sinplea erabili dugu, orain azalduko dugun bezala.

Aipatu dugu jadanik heuristiko bakoitzak genusaren adiera bakoitzari pisu bat esleituko diola. Pisu hori 1ekoa izango da egokien bezala jo den adierarentzat eta 0tik 1era doan balio batekoa gainontzekoentzat. Tarteko pisu hori lortzeko heuristikoen berezko balioak, 24. ekuazioko balio(D,G_i) –D definiendumaren eta G_i genusaren i-garren adieraren arteko erlazio-izaeraren neurria–, normalizatu egiten dira. Horretarako balio hori genus horren beste adierentzat lortutako balio maximoaz zatitu egiten da. 24. ekuazioan azaltzen da edozein heuristikok genusaren adiera batentzat (G_i) emandako bozaren pisua.

$$\text{pisu}(D,G_i) = \frac{\text{balio}(D,G_i)}{\max_{G_j}(\text{balio}(D,G_j))} \qquad (24)$$

Boza emateko arauei dagokionez, heuristiko batzuek (adibidez Distantzia Kontzeptualak) adiera guztiei pisuren bat ematen diote, beharrezko informazioa eskura baldin badute behintzat. Adiera bati buruz heuristikoak ezin badu ezer erabaki (adibidearekin jarraituaz, hiztegi elebidunean zulo bat badago), ezin gara arriskatu beste adiera aukeratzera. Beste heuristiko batzuentzat (adibidez definizioko hitzen ezkontzea), ordea, nahiko da adiera baten ezagutza edukitzea adiera hori aukeratzeko. Beraz bitan sailkatu dira heuristikoak: agerkidetza bektoreak, etiketa semantikoen bektoreak eta Distantzia Kontzeptuala alde batetik, eta adieren ordena, definizioko hitzen ezkontzea eta agerkidetza arruntak bestetik. Lehenbiziko multzokoetan, genusaren adieraren batentzat ezin bada bozik eman, orduan ez da bozik emango genus horren gainontzeko adierentzat. Bigarren multzoko heuristikoek, ordea, beti emango dute boza.

Genusaren adiera bakoitzerako heuristiko bakoitzak ematen duen bozaren pisua batzen da, eta pisu handiena lortzen duen genusaren adiera izango da hautatua.

VI.D.8. Emaitzak

Laginaren datuak 36. taulan aurkezten ditugu. 115 genus desanbiguatzen saiatu gara. %3tan genusa bilatzen zuen programak kale egin du. Batezbesteko adiera kopurua genus bakoitzeko 2,29 da, nahiz eta %36an adiera bakarra eduki. Lagina eskuz desanbiguatu dugu aurrez, eta egokia zenean adiera bat baino gehiago ontzat eman ditugu. Batez-beste 20 genusetatik batek dauzka bi adiera onargarri.

Lagina	115
Genus zuzena topatu	111 (%97)
Genusak adiera bakarra	40 (%36)
Adierak genus bakoitzeko	2,29
” (polisemikoentzat)	3,02
Adiera zuzenak genus bakoitzeko	1,05
” (polisemikoentzat)	1,06

36. taula: laginaren datuak

Heuristiko bakoitzak eta bozketak genus polisemikoentzat lortzen dituzten emaitzak 37. taulan azaltzen dira. Lehenbiziko zutabean zorizko hautaketak[78] lortuko zituzkeen emaitzak gehitu ditugu. Heuristikoek, banan-banan hartuta, emaitza kaxkarrak lortzen dituztela dirudi, baina beti zorizko emaitzen gainetik. Doitasuna eta estaldura kontuan hartuz gero, adieraren ordenari buruzko heuristikoa da hoberena. Doitasunari dagokionean Distantzia Kontzeptuala azpimarratu beharko litzateke, onena izateagatik, eta agerkidetza arruntak, okerrena izateagatik. Harrigarria izan daiteke, sinplea den heinean, definizioko hitzen ezkontzak lortu duen doitasuna. Heuristiko gehienek adiera bakar bat aukeratzen dute normalean, Distantzia Kontzeptualak izan ezik. Honek batez-beste 1,25 adiera aukeratzen ditu. Estaldurari dagokionez, heuristiko gehienak kasu gutxitan aplikatu ahal izan dira.

	Zorizkoa	OR	EZ	AA	AB	SB	DK	Bozketa
Doitasuna	%36	%66	%66	%44	%61	%57	%76	%73
Estaldura	%100	%100	%12	%25	%36	%19	%66	%100

37. taula: genus polisemikoentzat lortutako emaitzak

Heuristikoen bozketak emaitza onenak lortzen ditu zalantzarik gabe: doitasun ia hoberena eta estaldura osoa. Nahiz eta Distantzia Kontzeptualaren doitasunarekin alderatzean iruditu 3 puntu gutxiago lortzen dituela, bozketak adiera bakarra hautatzen du, distantziak ez bezala. Bozketak,

[78] Zorizko emaitzak analitikoki kalkulatu ditugu, genus polisemikoak dituzten adiera kopuruak erabiliaz.

beraz, heuristiko guztien emaitzak gainditzen ditu. Genus monosemikoak ere kontuan hartzen baditugu bozketaren doitasuna %82ra igoko litzateke (ikus 38. taula).

	Zorizkoa	Bozketa
Doitasuna	%59	%82
Estaldura	%100	%100

38. taula: genusentzat (monosemikoak barne) lortutako emaitzak

Heuristiko guztien bozketak emaitza onenak lortuta ere, zalantzan jar daiteke heuristiko guztiak beharrezkoak direnik. Heuristiko bakoitzaren ekarpena neurtzeko bozketa errepikatu genuen, baina heuristiko bat ezabatuaz kasu bakoitzean. 39. taulan azaltzen den bezala, edozein heuristiko kenduz gero doitasuna gutxienez 4 puntu erortzen da. Gale-k eta (Gale et al., 1993) ez dute uste hitz polisemikoentzat doitasuna %75 baino gutxiagokoa duten adiera-desanbiguatzaileak kontuan hartu behar direnik. Gure emaitzak, ordea, doitasuna %44 bezain baxua duen agerkidetza arrunta erabili gabe bozketaren emaitza 6 puntu erortzen dela adierazten du (39. taulako -AA zutabea). Beraz, heuristiko kaxkarrenek ere besteek ez duten ezagutzaz laguntzen dute emaitza hobetzen. Hau ados dago Dietterich-ek (1997) esandakoekin, heuristikoen konbinazioak heuristiko isolatuak baino doiagoak izan daitezkela esaten duenean. Horretarako baldintza bakarra, Dietterich-en ustez, heuristikoak elkarrekin ados ez egotea litzateke, hau da, bata bestearekin ahal den independienteena izatea, ezagutza ezberdina erabiltzea.

	Bozketa	-OR	-EZ	-AA	-AB	-SB	-DK
Doitasuna	%82	%75	%73	%76	%77	%77	%78
Estaldura	%100	%99	%100	%100	%100	%100	%100

39. taula: heuristikoen ekarpena, genus monosemikoak barne

VI.D.9. Ebaluazioa

Bigizta eta erlatore berezien tratamendua (ikus VI.B atala) eta genus-desanbiguazioa integratuz gero, genusa daukaten LPPL-ko adieren %97 desanbiguatu dugu (adiera guztien %88, ikus 40. taula), hiperonimiaren bidez hierarkiatan antolatuaz. Taula berean azaltzen da adierak zein metodoren bidez desanbiguatu izan diren. Genusik topatu ez zaien adierak ere hierarkiatan integratu gabe gelditu dira, noski. Desanbiguatuetatik, erlatore bidez egin direnentzat doitasuna %90ekoa da, monosemikoentzat %100ekoa eta bozketa bidez egindakoentzat %73koa. Batezbeste %84ko doitasuna edukiko genuke desanbiguatutako loturentzat.

Genus gabe	1251	%9
Desanbiguatu gabe	368	%3
Desanbiguatuta	12137	%88
Erlatore	1378	%10
Genus monosemikoa	4089	%30
Genusa (bestela)	6670	%48
Guztira	13740	

40. taula: genus-desanbiguazioaren emaitza orokorrak

Hurrengo atalean hitz egingo dugu desanbiguatu ondoren lortzen diren hierarkiez, baina lehendabizi aldera dezagun beste lanekin.

Genusen desanbiguazioan egin den lan ezagunenak (Bruce et al., 1992) LDOCE hiztegian kodetuta dauden informazio semantiko eta pragmatiko bereziak erabiltzen ditu. Gainera, adierak maiztasunaren arabera ordenatuta daude, lehenbiziko adiera usuena izanez. Ingelesa ez diren hizkuntzetarako ordea, ez da ohikoa halako informazio aberatsa edukitzea hiztegian. Lan honetan aurkeztu diren teknikak guztiz orokorrak dira, eta edozein hiztegi elebakarretarako balio dute, definizioetako testua besterik ez du erabiltzen eta. Hala ere, LDOCE-rako lortutako emaitzen parekoak lortzen ditugu, nahiz eta LPPL-ko definizioak askoz laburragoak izan. LDOCE-ko kodeketa bereziak erabiltzen duen metodo automatikoaren bidez %80ko doitasuna lortzen dute, guk aldiz %84. Beraiek genus usuenak aztertu eta horientzat adiera jakin bat emanaz doitasun hori %90era jaso zuten.

VI.E. HEBaren goiko geruzaren osatzea

Aurreko ataletako informazioa erabiliaz, hierarkiak beren artean lotzen saiatuko gara orain. Lehenbizi orain arteko lanarekin eraiki daitezkeen hierarkiak aztertuko ditugu.

VI.E.1. Hierarkien eraikuntza

Genusen desanbiguaziotik lortutako adierei, erlatore bidezkoen emaitzak gehitu behar zaizkie eta bigizten tratamendua aplikatu (ikus VI.B atala). Orain arte ez bezala, sinonimoei eta genusei tratamendu ezberdina emango diegu, genusak soilik sartzen dira berez hierarkia eraikuntzan, hiperonimia erlazioa adierazten baitute. Genus edo erlatore bidezko hiperonimia erlazioak lotzen baditugu, hierarkiak osatuko ditugu. Behin hierarkiak eraiki eta gero, LPPL-ko 13.740 adieretatik 10.241 adiera integratu ditugu 710 hierarkiatan, eta 3499 adiera isolatuta gelditu dira. Horrek esan nahi du, isolatutako adierez gain, hierarkien 710 erroak ere gelditu direla hiperonimo gabe. 41. taulan azaltzen da lotu gabeko adiera horien jatorria, eta nabarmena da sinonimoak direla ugarienak.

	Erroak	Isolatuak
Genusak	39	80
Erlatoreak	46	123
Sinonimoak	504	2.331
Analisi gabe	121	965
Guztira	710	3.499

41. taula: erro eta adiera isolatuen jatorria

Sinonimo horietatik asko desanbiguatuta daude, eta nola edo hala hierarkiatan integratzea badago. Lehenbiziko hurbilpen bezala, desanbiguatuta daudenean, beren sinonimoa hierarkiaren baten barne dagoen konprobatu, eta hala bada sinonimo horren senide bezala koka daitezke, beti ere zikloak sortzen ez direla ziurtatuaz. Hala egin ondoren hierarkia batzuk fusionatu eta isolatuta zeuden adiera batzuk hierarkietara lotzen dira, 527 hierarkia eta 2.258 adiera isolatu utziz.

	Erroak	Isolatuak
Genusak	39	80
Erlatoreak	46	123
Sinonimoak	321	1.090
Analisi gabe	121	965
Guztira	527	2.258

42. taula: erro eta adiera isolatuen jatorria, sinonimo batzuek tratatu ondoren

Oraindik ere sinonimoak dira lotu gabeko gehienak. 265 adiera izan ezik, beste guztiak desanbiguatuta daude, baina halere adiera hauek zikloak sor ditzakete. Esan dugunaren adibide bat *paysage* eta *vue*-ren bi adieren artean pasatzen da, *vue paysage*-en genusa izanda bere 5. adierara desanbiguatu dugu, eta *paysage vue*-ren sinonimoa izanda bere 1. adierara, hemen ikus daitekeen bezala.

paysage I 1 : ***vue*** *d'ensemble d'un site*
vue I 5 : ***paysage***

Sinonimoentzako tratamendu sofistikatuagoa pentsatu beharko litzateke, baina oraingoz zikloak sortzen dituzten sinonimoak ez ditugu integratuko hierarkian. Lortutako hierarkien neurri eta sakoneren datuak 43. eta 44. tauletan daude.

Neurria	Hierarkia kopurua	Portzentai metatua (%)
100-3293	13	0,5
50-99	15	1,0
25-49	22	1,8
10-24	67	4,2
2-9	410	18,9
1	2258	100,0

43. taula: hierarkien adiera kopuruak

Sakonera	Hierarkia kopurua	Portzentai metatua (%)
10	1	0,0
9	1	0,1
8	2	0,1
7	4	0,3
6	5	0,5
5	24	1,3
4	50	3,1
3	110	7,1
2	330	18,9
1	2258	100,0

44. taula: hierarkien adiera kopuruak

VI.E.2. "Txapelaren" inplementazioa

Goiko geruza (txapela) horren diseinua VI.A.5 atalean aurkeztu duguna da, hierarkien erroak WordNet-eko kontzeptuei lotzea. Bestalde, adiera batzuk isolatuta gelditu dira, inongo lotura gabe, eta horiek ere lotzen ahaleginduko gara. Adiera bat hiperonimo gabe gelditzeko lau arrazoi egon daitezke:

1. definizioan genusik aurkitu ez izana (beharbada ez zeukalako)
2. genusa eduki bai, baina ezin desanbiguatu izatea
3. bigiztak apurtzean zintzilik geratzea
4. erlatore bidezko definizioetan, lotzen saiatu baina ezin izatea

Txapelaren bidez, beraz, lau arazo horiei erantzungo diegu modu integratu eta natural batean. VI.C.2 atalean ikusi dugu nola lotu HEBko adierak WordNet-i. Lotura horren emaitza bezala, kasu onenean, WordNet-eko kontzeptu bat izango dugu. Bestela, posible da kontzeptu bat baino gehiagotara lotzea. Bestalde, kode semantikoa ere esleitzen saiatu izan gara, eta hemen ere bi aukerak daude, hau da, kode semantiko bakarra edo gehiago edukitzea. Kontzeptuak lotzeko orduan honek aukera asko irekitzen dizkigu. Lan honetan hartu dugun hurbilpena sinpleena izan da, horren ekarpena ebaluatu eta aurrerago tratamendu konplexuagoak diseinatzeko. Lotzeko algoritmoa beraz horrelakoa izango da:

1. hierarkien erroak eta adiera isolatuak bildu
2. WordNet-eko kontzeptu bakarra badaukate esleituta, lotu kontzeptu horri
3. WordNet-eko kode semantiko bakarra badaukate esleituta, lotu kode semantiko horren kontzeptu adierazgarri bati
4. kontzeptu anitz eta kode semantiko anitz badituzte, utzi oraingoz

Txapelaren eraikuntzan hierarkien arteko erlazio konplexuak sor daitezke, hau da, bi hierarkia WordNet-eko kontzeptu berera lotuta egon daitezke, edo bata bestearen azpian dauden kontzeptuetara, etab. Lehen bezala, ez gara hemen ur handitan sartuko, eta ez dugu hierarkien loturen arteko harremanik kontuan hartuko.

VI.E.3. Ebaluazioa

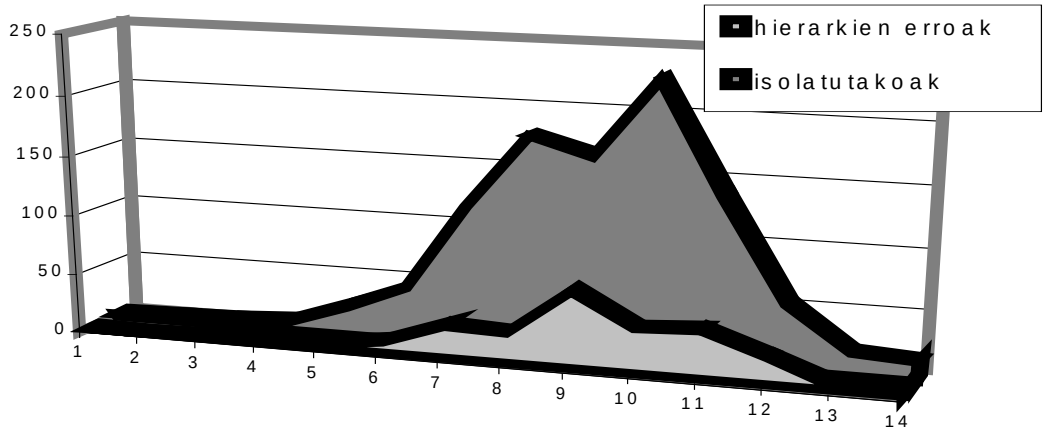
Lehenbizi hierarkien erroak eta adiera isolatuak bildu ditugu, eta horien loturak aztertu (ikus 45. taula). Aipatzekoa da hierarkien %15 bakarrik gelditu dela automatikoki lotzeko aukerarik gabe.

	Hierarkiako erroak		Isolatutakoak	
Kontzeptu bakarra	236	%45	991	%44
Kode semantiko bakarra	116	%22	353	%16
Anbiguoak	95	%18	145	%06
Lotura gabe	80	%15	769	%34
Guztira	527		2258	

45. taula: hierarkia eta adiera isolatuen loturak WordNet-era

Behin kontzeptu edo kode semantiko bakarra duten adierak lotu eta gero, hierarkiak eta isolatutako adierak WordNet-ko hierarkian integratzen dira. Lotura anbiguoa dutenentzat tratamendu bat egin beharko litzateke. Lotura honen fidagarritasuna aztertzeko, kontzeptu bakarrera lotuta daudenen lagin bat hartu dugu, 50 loturakoa, eta loturaren doitasuna %64koa bakarrik dela atera zaigu. Doitasun hau, LPPL-WordNet loturarako ditugun neurriak baino dezente apalagoa da (%80, ikus VI.C.2.f), ziur aski orain lotzen ari garen adierak, hierarkiatako erroak barne egonda, batezbestekoa baino orokor eta anbiguoagoak direlako. Eskuz aztertu ondoren, gehienak definizioan genus edo sinonimorik ez dutenak direla ikusi dugu, eta horrek LPPL-WordNet loturaren doitasun galtzean ere zer ikusia izan dezake.

WordNet-eko kontzeptu bakarrera lotu ditugun erro eta adiera isolatuei dagokionean, 28. irudian azaltzen da WordNet-eko zein sakoneratan kokatu diren. Ikusten den bezala, gehienak WordNet-eko alderdi sakonetan kokatu dira, 7 eta 11ko sakoneren artean. Isolatuentzat normala izan daiteke, baina hierarkien erroen kasuan adiera orokorragoak izatean, WordNet-eko goi aldean kokatzea espero zitekeen.



28. irudia: hierarkietako erroen eta adiera isolatuen kokapenaren sakonera WordNet-en

Aipatu ditugun datu hauek, loturaren doitasuna eta sakonera, hierarkiaren osatze guztiz automatikoa zalantzan jartzen dute. Loturak eskuz errepasatu beharko lirateke, hierarkiak WordNet-en dagozkion tokian kokatzeko. Dena den, 117 hierarkia zabalenek (10 adiera baino gehiago dituztenak) 10.127 adiera estaltzen dituzte, eta beraz, eskuzko lan gutxirekin hierarkia garrantzitsuenak lot daitezke.

Beharbada lotura finegia egiten saiatu gara, WordNet-eko kontzeptu mailan, eta egokiagoa litzateke lotura kode semantiko mailan egitea, asmatzeko aukera hobeagoak edukiko genituzke eta. Horretarako VI.C.2.e) atalean azaltzen diren kode semantikoak erabili zitezkeen. Hala egiten dute (Bruce et al., 1993) lanean, LDOCE-ko kode semantikoak erabiltzen baitituzte hierarkiak eta adiera isolatuak lotzeko.

VI.F. Ekarpenak

Kapitulu honetan ingelesa ez diren baliabide lexikal egituratuen eraikuntza sendotzeko teknikak landu ditugu. Izan ere, LPPL-tik erauzitako hierarkiak (Artola, 1993) ez dira libratzen hiztegietatik erauzitako hierarkiei egin izan zaizkien kritiketatik:

1. Hierarkia hitzen artekoa izatea, adieren artekoa izan ordez.
2. Hierarkietako bigiztak.
3. Erlatore bereziak hierarkian integratzeko arazoak.
4. Hierarkien sakonera apala eta goi mailako homogeneotasun falta.

Arazo horiek ebazteko bidean kanpoko ontologia baten beharra ikusi dugu, hierarkien goi-mailak antolatuko dituena eta hierarkia solteak lotzeko balioko duena. Bestalde, ontologia hori bigiztak hausteko eta erlatore bidezko definizioak hierarkian integratzeko ere erabili dugu. Planteatu diren 2., 3. eta 4. arazoak modu orokorrean konpontzeko metodoa aurkeztu dugu. Ontologia bezala WordNet erabili dugu, eta LPPL-ko adierak WordNet-era lotzeko tresna nagusi bezala Dentsitate Kontzeptuala erabili dugu.

1. arazoari dagokionean, behin LPPL-ko adierak WordNet-era lotu eta gero, genusak desanbiguatu ditugu, LPPL bertako informazioa eta LPPL-WordNet loturako informazioaz baliatuaz. Lortutako hierarkiak ebaluatu ditugu, bere horretan, eta gero LPPL-WordNet loturaz profitatuz, hierarkia guztiak lotu ditugu "txapela" kontzeptualaren bidez.

Egin diren lau zereginetarako garatu diren metodoak berritzaileak dira:

1. Bigizta eta erlatore bidezko definizioak tratatzekoa
2. LPPL-ko adierak WordNet-era lotzekoa
3. LPPL-ko genusak desanbiguatzekoa
4. Hierarkiak lotzekoa

VI.F.1. Bigizta eta erlatoreen tratamendua

Bigiztak puskatu eta hierarkian integratzeko modua aurkeztu dugu, LPPL-WordNet loturaz baliatzen dena. Aurkeztutako metodoari esker bigizta guztiak puskatzeko gai izan gara. Erlatoreen tratamenduari dagokionez, erlatore bidezko definizioen %78a hiperonimo desanbiguatu bati lotzea lortu dugu (LPPL-ko hierarkietan sartuz), eta %63 WordNet-eko adiera bati lotu. Erlatoreen kasuan, bai desanbiguazio eta bai WordNet-eko loturaren doitasuna %90era heldu dira. Emaitza hauei esker, bigiztak normal integratuko dira hierarkiatan, eta erlatore bidezko definizio gehienak edo hierarkian integratuta edo WordNet-i lotuta egongo dira. Geroago, hierarkiak lotzeko tratamenduari esker, WordNet-i bakarrik lotutako adiera horiek beste hierarkiekin ere lotu ahal izan ditugu.

VI.F.2. Kontzeptuen arteko lotura eleanitzak

LPPL-ko adierak WordNet-eko kontzeptuei lotzeko metodoa (elebiduna-WordNet) aurkeztu eta ebaluatu dugu. Metodo honi esker hizkuntza ezberdinetako baliabide lexikal egituratuak lotu daitezke kontzeptu/adiera mailan. Metodo honen emaitzak are hobeagoak izango lirateke hizkuntza bereko baliabideak lotuko balira, adibidez LDOCE eta WordNet lotuko balira.

Lehenbizi frantses-ingeles hiztegi elebidun bateko adierak WordNet-eko kontzeptuei lotu dizkiegu, horretarako Dentsitate Kontzeptuala soilik erabiliaz. Metodo horren bidez izenen adieren %43 lotu dugu %95eko doitasunarekin. Nahiz eta lan honetan garrantzi gehiegirik ez eman lotura hauei, ez bada LPPL-WordNet loturarako laguntza bezala, mota honetako loturak oso garrantzitsuak dira hizkuntza arrotzak ontologia jakin bati lotzeko. Izan ere guk garatutako metodo baino apalagoak erabili izan dira helburu horrekin, bai Sensus ontologiara gaztelerako hitzak lotzeko (Okumura & Hovy, 1994), bai EuroWordNet proiektuaren barruan gaztelerazko WordNet eraikitzeko (Rigau & Agirre, 1995; Atserias et al. 1997). Lan horietan tesi honetako metodoa aplikatuz gero, beraien doitasunak hobetuko liratekeela uste dugu.

LPPL-WordNet loturei dagokionean, elebiduna-WordNet loturek emaitzak hobetzeko balio izan du. Lotura horretaz gain, Dentsitate Kontzeptuala, hiperonimia erlazioak, heuristiko sinple batzuk eta nabarmentasunean oinarritutako hedadura erabili ditugu, erlatore berezien bidezko tratamendua barne. Horrela LPPL-ko izenen adieren %87 WordNet-era lotzea lortu dugu, batezbesteko %80ko doitasunarekin. Bai Dentsitate Kontzeptuala eta hiperonimia erlazioak WordNet-eko lotura paradigmatikoetan oinarritzen dira. *Nabarmentasun* bidezko metodoa, hiztegiko informazioaz eta WordNet-eko kode semantikoez baliatzen da, neurri estatistikoak erabiliz.

VI.F.3. Genus-desanbiguazioa

Gaurdaino hiztegietako genusen desanbiguazioa hierarkia zabal eta erabilgarriak automatikoki sortzeko arazo garrantzitsuena zela uste izan da. Arazo hau LDOCE hiztegirako bakarrik ebatzi izan da arrakastaz (%80ko doitasuna era guztiz automatikoan, %90 eskuzko laguntzari esker), hiztegiak adierentzat dauzkan kode pragmatiko eta semantikoei esker. Lan honetan erakutsi dugu genusen desanbiguazioa ez dagoela LDOCE-ra soilik mugatuta, bestelako lanetan ere bideragarria izan daitekeela, eta tesi honetarako garatutako metodoak LPPL hiztegirako %82ko doitasuna lortzen du. Beste edozein hiztegitara aplikatzeko balio du metodoak, eta hala frogatu izan da gaztelerarako DGILE (ikus II. kapitulua) hiztegian eginiko esperimentuetan, pareko doitasuna –%83– lortu izan baitugu (Rigau et al. 1997).

VI.F.4. Hiztegietatik erauzitako hierarkien lotzea

Doitasun handiko prozedura automatiko baten bidez erauzitako hierarkiak, nahiz eta bigizta eta erlatore bereziak behar bezala tratatu, badauzkate arazoak: hierarkia asko txikiak izan eta gainera solte daude, elkarrekin inongo loturarik eduki gabe. Gainera ezaguna da hiztegietatik erauzitako hierarkiek goi aldean duten egitura ez dela oso egokia. Bi arazo horiei automatikoki erantzuteko prozedura planteatu dugu, WordNet-era eginiko loturetaz baliatzen dena. Prozedura horretan

hierarkietako erroak WordNet-era lotzen ditugu, horrela WordNet-en goiko geruzak ematen du koherentzia eta gainera hierarkia solte guztiak WordNet-en bidez lotuta gelditzen dira.

Planteatutako metodoa orokorra da, eta hiztegietatik erauzitako hierarkiak edozein ontologiatara lotzeko balioko luke, gehien interesatzen zaigun goi-maila hautatzeko aukera emanaz.

VI.G. Etorkizunerako lanak

Ekarpenetan egin dugun bezala, hemen ere atalka aztertuko ditugu aurrerantzean egin daitezkeen hobekuntza eta lanak. Lehenbizi lotura eleanitzak aipatuko ditugu. Ondoren genus desanbiguazioari buruz eta hierarkia isolatuak lotzeko metodoari buruz ihardungo dugu.. VI.G.4. Atalean lotura eleanitz eta genus desanbiguazioaren artean dagoen dependentzia funtzionala aipatuko dugu, izan ere, batak bestearen emaitza hobetu dezake, edo agian era integratu batean konpondu daitezke biak. Bukatzeko, ezagutza-base lexikalen sendotzeari buruzko lan orokorrak aipatuko ditugu.

VI.G.1. Kontzeptuen arteko lotura eleanitzak

Lotura eleanitzen hobekuntzari dagokionez, ez dago zalantzarik elebidun zabalago bat erabiltzeak emaitzak hobetuko lituzkeela. Alde batetik estaldura zabalagoa lortuko litzateke (eta elebiduna-WordNet-en kasuan estaldura hobeagoak zuzenean LPPL-WordNet loturaren doitasunaren hobekuntza dakar). Bestetik, LPPL-ko adieraren baterako itzulpenik ez egotea errore-iturri denez, elebidun zabalagoarekin halako gutxiago gertatuko lirateke, doitasuna ere hobetuz.

Elebiduna-WordNet loturaren estaldura jasotzeko beste modu bat frantsez-hitz/ingeles-hitz bikoteetan oinarritutako heuristikoak dira (Okumura & Hovy, 1994; Rigau & Agirre, 1995; Atserias et al. 1997). EuroWordNet proiektuan halako heuristikoak arrakastaz erabiltzen ari dira gaztelerarako WordNet-a eraikitzeko. Hala ere hitz bikote hauek badute murrizpenik, ez baitira hiztegi elebiduneko adierak kontuan hartzen. Horrek arazoak sortu ditzake: adibidez frantsesez monosemikoa den hitzak bi itzulpen ezberdin dituenean, eta itzulpen horiek kontzeptu ezberdinei dagozkienean. Beti ere azalduko zaizkigu, metodoa nola nahikoa izanda ere, hizkuntzen arteko *mismatch*-ak (Arregi 1995).

Elebiduneko adierak erabiltzeari esker, WordNet eta LPPL hiztegi elebidunetan dagoen informazio zabalarekin (Fontenelle, 1997) aberastu zitekeen. Arlo hau jorratzea interesgarria litzateke.

Gaur egun, EuroWordNet eta ITEM proiektuei lotuta, Euskararako WordNet-a ere eraikitzen ari gara, kapitulu honetan aurkeztutako teknikak eta arestian aipatutako hitz bikoteak euskara-ingelesa hiztegi elebidunari aplikatuaz. ITEM proiektuan gaztelerarako WordNet ere eraikitzen ari den

heinean, hiztegi elebidunen kateak erabiliaz (euskara-gaztelera, euskara-ingelesa eta gaztelera-ingelesa) estaldura eta doitasuna hobetuko direlakoan gaude.

Kapitulu honetan garatutako metodoak baliabide lexikal egituratuak lotzeko balio duenez, ontologia eta EBLren bat egitean eragin handia izan dezake. Baliabide batek besteak duen ezagutza xurgatu eta ontologia aberatsagoak eraikitzen joateko bide egokia dirudi honek, *ANSI Ad Hoc Ontology Standards Group*[79] komiteak proposatzen duen bidea jarraituz (Hovy, 1997a; 1997b).

VI.G.2. Genus-desanbiguazioa

Nahiz eta lortu ditugun emaitzak oso onak izan, badago desanbiguazioaren doitasuna jasotzeko metodorik. (Rigau et al. 1998) artikuluan, kapitulu honetan azaldutako metodoa DGILE hiztegiari aplikatu ondoren (Rigau et al., 1997), genusak multzokatu egin genituen, WordNet-era lotzean lortu den kode semantikoaren arabera. Kode semantiko bakoitzerako genus usuenak bakarrik aukeratuaz doitasuna altxatu egiten da, estalduraren golkora. Hobekuntza hau LPPL-rekin egitea ere otu zitzaigun, noski, baina LPPL-ren neurri txikia dela eta genus usuenak ez ziren oso maiz gertatzen, eta ez genuen emaitzak hobetzerik lortu.

Bartzelonako UPC-ko lexikografia konputazionalean aritzen den taldeak eta gureak azterketa paraleloak egin ditugu, LPPL hiztegi txikiarentzat gurean, eta DGILE hiztegi zabalagoarentzat haienean. Hiztegien azterketa paralelo honetatik hiztegi txikietarako metodoak handietan ere balio izan digutela ondorioztatu daiteke. Gainera, hiztegi handietatik hierarkia zabalago eta interesgarriagoak jasotzen dira, eta orain aipatu dugun bezala, eta hobekuntzarako aukera gehiago ematen dituzte.

Bozketaren emaitzei dagokionean, bozketaren emaitzen azterketak oraindik hobekuntzarako tokia baduela uste dugu. Azterketa txiki batean ikusi genuen boza ematen dutenen artean gutxienez bosten adostasuna eskatuz gero %95eko doitasuna lortu genezakeela, baina estalduraren kaltetan (%18). Honek doitasun handiz egindako loturak identifikatzeko metodo bat eman lezake.

Bestalde, desanbiguazioa egitean definizioan bertan zegoen informazioa erabili dugu, baina hierarkien arteko desanbiguazioa ere planteatu daiteke, hau da, genus bat desanbiguatzerakoan, ziurtzat dauzkagun definiendumaren hiponimoak eta genusaren adiera bakoitzaren beste hiponimo eta hiperonimoak ere kontuan har ditzakegu.

Ildo berean, behin "txapela" eginda ere, errazagoa izan daiteke genus desanbiguazioa egitea.

[79] http://ksl-web.stanford.edu/onto-std/

VI.G.3. Hiztegietatik erauzitako hierarkien lotzea

Tesi-lan honetan hierarkien eraikuntzan sinonimia erlazioa ez dugu kontuan hartu. Autore gehienek hiztegietatik erauzitako sinonimia erlazioari ez diote jaramonik egin, baina LPPL-ren kasuan sinonimo bidezko definizioak ugariak dira (adiera guztien %20). Artolak (1993) bere HEBan sinonimo ziren adieretan erlazioak sinonimo batetik bestera kopiatzen zituen, eta horren eragina neurtzea interesgarri deritzogu. Bestelako hurbilpenak, adibidez WordNet-en sinonimo diren adiera guztiak kontzeptu bakarra osatzen dute, ere aztertu beharko lirateke.

Nahiz eta hierarkiak lotzeko metodoak etorkizuna duela erakutsi dugun, ez dugu zehatz-mehatz ebaluatu lortzen den hierarkiaren kalitatea. Halakoak ebaluatzeko ez dago gaur egun irizpide finkorik, ez bada hiponimo/hiperonimo lotura bakoitzaren doitasuna, jadanik eman duguna (%82). Neurri hori oso mugatua izanda, aplikazio batetarako –adibidez, informazioaren erauzketa– erabilgarria den edo ez izan daiteke irizpide interesgarria. Bestalde ezin da ahaztu *ANSI ad hoc Ontology Standards Group* delakoa, arestian aipatu duguna, bere zereginen artean ontologiak ebaluatzeko irizpideak lantzen ari dela, oraingoz emaitza gabe.

VI.G.4. Sorgin-gurpila

Gure hurbilpena aurkeztean (ikus VI.A.5 atala), aipatu izan dugu kapitulu honetan azaltzen diren hiru prozeduren artean, LPPL-WordNet lotura, LPPL-ko genusen desanbiguazioa eta LPPL-ren txapelaren eraikuntza, elkarrekintza konplexuak gerta zitezkeela. Horien azterketari ezin genion eutsi lehenbizi hiru prozedurak ez bagenituen era independentean egiten, edo hobeto esanda, LPPL-WordNet lotura bere kabuz egin, LPPL-ko genusen desanbiguazioa bere kabuz egin (LPPL-WordNet lotura erabiliaz ere), eta txapela aurreko bien emaitzen gainean egin.

Hiru prozesuen arteko elkarrekintza hobeto aztertu beharko dugu. Orain arte egindakoaren informazioa hor dago beste modu batzuetara konbinatzeko. Adibidez, behin LPPL-ko hierarkiak desanbiguatu eta txapelaren bidez lotu ondoren, LPPL-WordNet lotura egiteko informazio gehiago dugu, hierarkiak lotzen ari baikara, eta suposatu daiteke emaitza hobeagoak lortu ditzakegula. Bestalde jadanik aipatu dugu, behin txapela eginda, errazagoa izan daitekeela genus desanbiguazioa egitea. Lotura elebidun hobeekin, bai txapela eta bai genus desanbiguazioa ere hobetu daitezke, eta abar. Honekin prozesu iteratibo bat plantea daiteke.

Bestelako hurbilpen interesgarri bat sare neuronalen eskutik etorri daiteke. Kapitulu honetan deskribatutako emaitza guztiak –LPPL-WordNet lotura, LPPL-ko hiponimo/hiperonimo erlazioak, WordNet beraren hierarkia– sare neuronal bateko arku bezala errepresenta daitezke. Sare neuronalari energia-funtzio egoki bat esleituz gero, ezagunak diren teknikak aplikatu daitezke

arkuen konbinazio ezin hobea bilatu dezan. Horrela aldi berean erabakiko luke zein den adiera bakoitzerako WordNet-lotura hoberena eta hiperonimo hoberena.

VI.G.5. Bestelakoak

EBLen sorkuntza eta aberasketa automatikoa landu dugun arren, ez ditugu horren ikuspuntu guztiak landu. Hutsune nagusietako bat definizioen *differentia*-tik erauzitako informazioa (Artola, 1993) landu ez dugula izan da. V.B atalean LPPL-ko definizioen *differentia-tik* erauzitako erlazioen adibideak ikusi ditugu. *Differentia-ren* erabilera betidanik interesgarritzat jo izan da eta lan berriek (ikus adibidez, Richardson, 1997) arnas berria eman diote bertatik erauzitako informazioaren erabilgarritasunari. Bestalde, adieren adibideen analisiak ere informazio interesgarria eman dezakeela uste dugu, adiera azaltzen den testuinguruei buruzko informazio interesgarria ematen dute eta.

Euskarari dagokionean, ezin dugu aipatu gabe utzi Euskal Hiztegiaren gainean gure taldean egiten ari den lana. Lan horren helburua euskararako ezagutza-base zabala sortzea, informazio semantikoan aberatsa. Horretarako hiztegiaren egituraren azterketa egin da eta TEI gidalerroak jarraitzen dituen kodeketara itzuli dugu (Arriola et al. 1995; 1996a; 1996b). Jadanik bukatu dugu izenen definizioetako erlatore berezi eta genusen bilaketa (Agirre et al. 1998), eta gaur egun aditz eta adjektiboen analisia, adieren adibideen analisia, eta WordNet-en lotura lantzen ari gara. Hurrengo pausoan, kapitulu honetan azaldutako metodoen bidez, izen, aditz eta adjektiboen hierarkien eraikuntzari ekingo diogu, taldeko kide baten tesiaren barruan, atal honetan planteatu ditugun hobekuntzak aplikatzen saiatuz. Bestalde, Euskal Hiztegiko definizioetan erabiltzen den hizkuntzaren azterketa ere martxan dago, oraingo genus eta erlatore bidezko bilaketa bezala. Etorkizun laburrean. taldean garatzen ari diren gramatiken bidez, *differentia*-ren azterketari ekingo diogu, taldeko beste kide baten tesia izango den lanean.

Azkenik, interesgarritzat jotzen dugu hierarkia eleanitzen eraikuntza automatikoa. Hizkuntza ezberdinetako baliabideak lotzen goazen heinean, hierarkia eleanitzak osatzen ari gara. Inplizituki, ontologia ezberdinetako informazio (erdi-)automatikoki hizkuntza batetik bestera xurgatzea posiblea den edo ez aztertzen ari gara, eta aldi berean hierarkia unibertsalak eraiki daitezkeen edo ez aztertzen. Aldi berean, hierarkia ezberdinetako informazioa bateragarria den edo ez, goi mailak automatikoki lotzea komeni ote den, eta antzeko galdera ugari sortzen zaizkigu. Galdera hauek baliabide lexikal egituratuen eraikuntzatik munduaren eta hizkuntzen ereduen azterketara garamatzate.

VII.Kapitulua

ONDORIOAK

VII.A. Sarrera

Lan honen ekarpen nagusiak bi dira:

1. Erlazio-izaeraren formalizazioa: Dentsitate Kontzeptuala
2. Hiztegietatik erauzitako hierarkiak trinkotzeko metodoa

Lan honetan izenen adieren arteko erlazio-izaeraren neurri bat formalizatu dugu: Dentsitate Kontzeptuala. Neurri hau ontologietan oinarritzen da, eta beraz LNPan erabiltzen den informazioa berrerabiltzen du. Edozein ontologiara aplikatu daiteke, ez du behar inongo aurre-prestaketarik, eta ontologiak estaltzen dituen domeinu guztietan lan egiteko gauza da. Dentsitate Kontzeptualaren inplementazio osoa WordNet gainean egin dugu.

Gure formalizazioa interesgarriagoa dela defendatu dugu, bai beste baliabide lexikaletan oinarrituta dauden neurrien aurrean (corpus eta hiztegi), baita ontologiatan oinarritzen diren bestelako neurrien aurrean ere. Nagusitasun hori praktikan erakusten saiatu gara:

- Hitzen Adiera-Desanbiguazioan (IV. kapitulua)
- Testuen Zuzenketa Automatikoa (V. kapitulua)

Hitzen adiera-desanbiguazioan emaitza onak lortzen ditu, nahiz eta nahiko zaila izan beste sistemekin konparatzea. Hobeto konparatu ahal izateko ontologian oinarritutako beste bi sistema inplementatu, eta Dentsitate Kontzeptualak beraien emaitzak gainditzen dituela ikusi dugu. Zuzenketa automatikoari dagokionez emaitzak bestelakoak izan dira. Dentsitate Kontzeptuala izenei besterik aplikatu ezin denez, zuzenketa proposamen guztiak izenak direnean bakarrik aplikatu

ahal izan dugu, eta beraz, nahiz eta erabili diren ebaluazio corpusak zabalak izan, Dentsitate Kontzeptualak oso gutxitan hartu du parte. Aurkeztu dugun zuzenketa automatikorako sistemak beste ezagutza-iturrietara ere jo du.

Bestalde, hiztegietatik erauzitako hierarkiak sendotzeko metodo bat aurkeztu dugu. Metodo honek Dentsitate Kontzeptuala eta landutako hiztegiko bertako ezagutza ere erabiltzen ditu. Maila praktikoan bi hobekuntza nabari burutu ditugu:

- *Le Plus Petit Larousse* frantses hiztegiko adierak WordNet-i lotu
- *Le Plus Petit Larousse*-etik erauzitako HEBko adieren hierarkiak desanbiguatu eta trinkotu

Lehenbizikoari esker, eraikitako hierarkia horien gabezia batzuk konpon ditzakegu, WordNet-eko hierarkia erabiliz goi-ontologia bezala: definizio erlazionalak lotzeko, bigiztak ebazteko, hierarkia isolatuak elkarren artean lotzeko eta hierarkiei goi-maila koherente bat emateko. Aurkeztu dugun metodoa edozein hiztegitatik erauzitako hierarkiak desanbiguatu eta trinkotzeko erabil daiteke. Bestalde, baliabide lexikalak ezkontzeko ere balio duenez, baliabide heterogeneoak, hizkuntza berekoak edo ez, bat egiteko erabil daiteke: ontologiak EBLetara, EBLak EBLetara eta abar. Honek perspektiba berriak irekitzen ditu baliabide lexikalen aberasketan, ezagutzan pobre den hizkuntza batek ingeleserako eraikitako ezagutza xurgatu dezake eta. Betiere, kontu izanez ekartzen den ezagutza hizkuntza horretarako baliagarria den edo ez, noski. Hitzen adiera-desanbiguazioa ere, hein handi batean, baliabide lexikalen lotzea bezala ikus daiteke, corpuseko hitzak ontologia bateko adiera/kontzeptuetara lotzen baitira. Ikuspegi honek ontologiak aberasteko bide berriak irekitzen ditu.

Etorkizunerako lanari dagokionez, behar handia ikusten dugu ontologia zabal eta aberatsak sortzeko. Izan ere, Dentsitate Kontzeptualaren bidez WordNet-ek duen informazioaz besterik ezin gara baliatu, hau da, batez ere erlazio paradigmatikoak. Nahiz eta horrela ere emaitza onak lortu aplikatu dugun zereginetan, garbi dago erlazio sintagmatikoak ere beharrezkoak direla, adibidez, hitzen adiera-desanbiguazioa hobetzeko, baina bereziki zuzenketa automatikoan Dentsitate Kontzeptualaren ekarpena zabaltzeko.

HAD egiteko, baina LNParen bestelako arazo lexikal-semantikoei irtenbide sendoa emateko ere, corpus, hiztegi eta ontologien arteko koordinazio estua behar dela uste dugu. VI. kapituluan ontologia, EBLak eta HEBak lotu eta bat egitea posible dela agertu dugu. Horrelako integrazioak WordNet aberasteko balio dezake, baina, hala ere, ez litzateke nahikoa izango HADrako beharrezko ezagutza guztia biltzeko. Adibidez, hiztegi zabal baterako hautapen-murrizpenen

zerrendarik ez dago eskuragarri inon. Holakoen ikasketa bultzatzeko hiztegiko definizioen analisia eta erabilera indartu behar da (adibidez, VI. kapituluan aipatutako teknikak erabiliaz), ondoren, definizioko hitzen adierak desanbiguatuta daudenean, ontologietan integratu ahal izateko ezagutza hori. Berdintsu gertatzen da corpusekin. V. kapituluan ikusi dugu corpusetan oinarritutako neurri estatistikoek ere hitzen arteko erlazio-izaeraren neurria ondo islatzen dutela, eta ezagutza hori ontologietan integratzeko moduan kodetu beharko litzatekeela. Hitzen adiera-desanbiguazioaren bidez, hain zuzen ere, posible izan beharko litzateke hitzen arteko erlazio horiek adieren arteko bihurtzea, eta horrela ontologiari lotu. Dentsitate Kontzeptuala modu egoki batera hedatuz, ontologia mota berri horietako erlazioez profitatuko litzateke, errepresentazio trinko eta eraginkor baten bidez, iturri ezberdin askotako ezagutza erabiliaz erlazio-izaera kalkulatu ahal izateko.

Azter ditzagun era zabalago batean kapitulu bakoitzean eskaini ditugun ekarpen nagusiak, eta ondoren etorkizunerako lanak ikusiko ditugu.

VII.B. Ekarpenak

VII.B.1. Erlazio-izaeraren neurria definitu: Dentsitate Kontzeptuala (III. kapitulua)

Dentsitate Kontzeptuala diseinatu eta inplementatu dugu, ontologietan oinarrituaz erlazio-izaera formalizatzeko. Dentsitate Kontzeptuala ontologiako erlazio paradigmatikoetaz –hiperonimia eta meronimia– baliatzen da, eta izenekin lan egiten du oraingoz, nahiz eta aditzetarako ere egokia izan daitekeen.

Ontologietan oinarritzen diren gainontzeko formalizazioen ezaugarriak dauzka. Oinarri teoriko sendoa du, adimen artifizial eta psikolinguistikan ezagutzaren errepresentaziorako eredu nagusiak baitira ontologiak. Adieren arteko neurria eskaintzen digu, adiera horien definizio sendoa eskainiz, adierak ontologiako kontzeptuei lotuta daude eta. Bestalde ez du eskuzko desanbiguazio beharrik, ez eta datu urrien edo gehiegizkoen arazorik. Ezaugarri hauek dira ontologietan oinarritutako neurriak, corpus eta hiztegietan oinarritutakoekin alderatuz gero, dituzten abantailak.

Ontologietan oinarritutako neurriak, aldiz, eraginkortasun-arazoak eduki ohi dituzte. Gainera erlazio-izaeraren neurriak bi kontzepturen artekoak, eta ez gehiago, izaten dira. Dentsitate Kontzeptualak ez dauzka murrizpen horiek, eta beraz, ontologietan oinarritzen diren beste neurriak gainditzen ditu. Edozein kontzeptu kopururen erlazio-izaera neurtu dezake, gainera kopuru ezberdineko multzoen neurriak konparatzeko modua eskainiaz. Testu zabalekin lan egiteko bezain eraginkorra da.

VII.B.2. DKaren aplikazioa: hitzen adiera-desanbiguazioa (IV kapitulua)

WordNet-eko ezagutza paradigmatikoa darabilen Dentsitate Kontzeptualean oinarritutako desanbiguatzailea eraiki eta probatu dugu. Dentsitate Kontzeptualaren ezaugarriei esker ontologiako adieren arabera desanbiguatzeko gai den sistema eraiki dugu, testu errealetako izenak denbora mugatuan desanbiguatzeko gai dena. Edozein testutara aplikatu daiteke, inongo egokitzapenen beharrik gabe.

Esperimentuko emaitzen arabera, Dentsitate Kontzeptuala HADrako erabilgarria dela frogatu dugu, eta WordNet-eko ezagutzaz erlazio-izaera paradigmatikoen beste formalizazioak –Sussna (1993) eta Yarowsky (1992)– baino hobeto baliatzen dela erakutsi ere bai.

HADaren literaturan azaltzen diren esperimentuekin alderatzean gure esperimentuak arazoaren alde zailenari egin dio aurre: adiera bereizketa xeheak, domeinu ezberdinetako testu errealak, testuko izen guztiak, emaitza partzialak baztertuaz eta adiera bakarra ontzat emanaz. Testuak (guztira 10.000 hitz) ez ziren inolaz ere errazak desanbiguatzeko. Hala ere, WordNet-eko adiera finetarako desanbiguatzean %64ko doitasuna lortzen dugu, eta fitxategi-mailan desanbiguatzen badugu %71koa. Estaldura oso zabala da, testuetako izenen %86 desanbiguatzen dugu eta.

VII.B.3. DKaren aplikazioa: zuzenketa automatikoa (V. kapitulua)

Testu-zuzenketa automatikoa egiten duen sistema diseinatu eta eraiki dugu, ez-hitz motako sakatze-erroreentzat proposamen egokia aukeratzen duena. Alde batetik zuzenketa automatikoa gaur egungo teknologiaren eskura dagoela frogatu dugu, eta bestetik Dentsitate Kontzeptualaren ekarpena apala izan dela ikusi dugu.

Sistema honek ezagutza-mota ezberdinak konbinatzen ditu: sintaktikoa (Murrizpen-Gramatikak), semantikoa (Dentsitate Kontzeptuala), hitzen maiztasunak, testuinguru-estatistikak eta heuristiko espezifikoak. Murrizpen-Gramatika, Dokumentuko Maiztasun eta Testuinguru-Estatistikei esker, gai da 25 erroretatik 24etan proposamen bakarra aukeratzeko (bestela bi proposamen) %90eko doitasunarekin, eta errore **guztientzat** erantzuten du. Emaitza hauek frogatzen dute zuzenketa automatikoa egingarria izan daitekeela.

Dentsitate Kontzeptualaren estaldura erroreen %8koa izan da soilik, proposamen guztiak izenak direnean bakarrik aplikatzen da eta. Lagin txiki horrekin fidagarritasun gutxiko datua izanda ere, %75eko doitasuna lortu da. Doitasun apal honen arrazoia ez da DKarena berez, erabilitako WordNet ezagutza-basearen gabeziena baizik, III. kapituluan arrazonatu dugun bezala.

VII.B.4. Baliabide lexikalak sendotu (VI. kapitulua)

Kapitulu honetako sarreran aipatu ditugu hiztegietatik erauzitako hierarkiak dituzten arazoak, eta *Le Plus Petit Larousse*-etik erauzitako hierarkiak ere (Artola, 1993) ez dira horretatik libratzen. Arazo horiek ebazteko bidean kanpoko ontologia baten beharra ikusi dugu, hierarkien goi-mailak antolatuko dituena eta hierarkia solteak lotzeko balioko duena. Bestalde, ontologia hori bigiztak konpondu eta erlatore bidezko definizioak hierarkian integratzeko ere erabili dugu. Kanpoko ontologia hori izan da ere hierarkiako hitzak desanbiguatzeko giltza. Lau zereginetan banatu dugu hierarkiak aberastu eta trinkotzeko metodoa:

VII.B.4.a) Bigizta eta erlatoreen tratamendua

Bigiztak puskatu eta hierarkian integratzeko modua aurkeztu dugu, LPPL-WordNet loturaz baliatzen dena. Aurkeztutako metodoari esker bigizta guztiak puskatzeko gai izan gara. Erlatoreen tratamenduari dagokionez, erlatore bidezko definizioen %78a hiperonimo desanbiguatu bati lotzea lortu dugu (LPPL-ko hierarkietan sartuz), eta %63 WordNet-eko adiera bati lotu. Erlatoreen kasuan, bai desanbiguazio eta bai WordNet-eko loturaren doitasuna %90era heldu dira. Emaitza hauei esker, bigiztak normal integratuko dira hierarkiatan, eta erlatore bidezko definizio gehienak edo hierarkian integratuta edo WordNet-i lotuta egongo dira. Geroago, hierarkiak lotzeko tratamenduari esker, WordNet-i bakarrik lotutako adiera horiek beste hierarkietara lotu ahal izan ditugu.

VII.B.4.b) Hizkuntza ezberdinetako baliabideen lotura kontzeptu mailan

Lehenbizi frantses-ingeles hiztegi elebidun bateko adierak WordNet-eko kontzeptuei lotu dizkiegu (elebiduna-WordNet), horretarako Dentsitate Kontzeptuala soilik erabiliaz. Metodo horren bidez izenen adieren %43 lotu dugu %95eko doitasunarekin. Mota honetako loturak oso garrantzitsuak dira hizkuntza arrotzak ontologia jakin bati lotzeko. Izan ere guk garatutako metodo baino apalagoak erabili izan dira helburu horrekin, bai Sensus ontologiara gaztelerako hitzak lotzeko (Okumura & Hovy, 1994), bai EuroWordNet proiektuaren barruan gaztelerazko WordNet eraikitzeko (Rigau & Agirre, 1995; Atserias et al. 1997). Lan horietan tesi honetako metodoa aplikatuz gero, beraien doitasunak hobetuko liratekeela uste dugu.

LPPL-ko adierak WordNet-eko kontzeptuei lotzeko metodoari dagokionean (LPPL-WordNet), elebiduna-WordNet loturek emaitzak hobetzeko balio izan dute. Lotura horietaz gain, Dentsitate Kontzeptuala, hiperonimia erlazioak, heuristiko sinple batzuk eta nabarmentasunean oinarritutako hedadura erabili ditugu, erlatore berezien bidezko tratamendua barne. Horrela LPPL-ko izenen adieren %87 WordNet-era lotzea lortu dugu, batezbesteko %80ko doitasunarekin. Bai Dentsitate

Kontzeptuala eta hiperonimia erlazioak WordNet-eko lotura paradigmatikoetan oinarritzen dira. Nabarmentasun bidezko metodoa, hiztegiko informazioaz eta WordNet-eko kode semantikoez baliatzen da, neurri estatistikoak erabiliz.

VII.B.4.c) Genus-desanbiguazioa

Lan honetan erakutsi dugu genusen desanbiguazioa ez dagoela LDOCE-ra soilik mugatuta., eta tesi honetarako garatutako metodoak LPPL hiztegirako %82ko doitasuna lortzen du. Beste edozein hiztegitara aplikatzeko balio du metodoak, eta hala frogatu izan da gaztelerarako DGILE (ikus II. kapitulua) hiztegian eginiko esperimentuetan, pareko doitasuna –%83– lortu izan baitugu (Rigau et al. 1997).

VII.B.4.d) Hiztegietatik erauzitako hierarkien lotzea

Automatikoki sortutako hierarkiek badauzkate arazoak: gehienak txikiak dira eta gainera solte daude, elkarrekin inongo loturarik eduki gabe. Gainera ezaguna da hiztegietatik erauzitako hierarkiek goi aldean duten egitura ez dela oso egokia. Bi arazo horiei automatikoki erantzuteko prozedura planteatu dugu, WordNet-era eginiko loturetaz baliatzen dena. Prozedura horretan hierarkietako erroak WordNet-era lotzen ditugu, horrela WordNet-en goiko geruzak ematen du koherentzia eta gainera hierarkia solte guztiak WordNet-en bidez lotuta gelditzen dira. Planteatutako metodoa orokorra da, eta hiztegietatik erauzitako hierarkiak edozein ontologiatara lotzeko balioko luke, gehien interesatzen zaigun goi maila hautatzeko aukera emanaz.

VII.C. Etorkizunerako lana

VII.C.1. Dentsitate Kontzeptualaren hobekuntza (III. kapitulua)

Dentsitate Kontzeptuala hobetzeko hiru alor nagusi hauek ikusten ditugu:

- Darabilen informazioari dagokiona: erlazio sintagmatiko eta hautapen-murrizpenak dituen ontologia bat lortu edo WordNet bera halakoez aberastu. Tamalez gaur egun informazio hori ez dago zuzenean eskuragarri, baina hiztegi eta corpusetatik automatikoki eskuratzeko metodoak ikertzen ari dira. VI. kapituluan aipatu dugu, adibidez, hiztegietako *differentia-ren* azterketatik erauztea posible dela. Zuzenketa automatikorako kapituluan ere ikusi dugu corpusetatik erauzitako Testuinguru-Estatistikek darabilten informazio gordinean, modu inplizituan bada ere, erlazio sintagmatiko eta hautapen-murrizpenak ezkutatzen direla. Baliabide lexikalen integrazioari (VI. kapitulua) eta adiera-desanbiguazioari (IV. kapitulua) esker, informazio hori WordNet ontologian integratu ahal izango litzateke.

- Formulari dagokiona: Dentsitate Kontzeptualaren formula aldatu, paradigmatikoak ez diren erlazioak kontuan har ditzan. V.B.2 atalean labur azaldu dugu nola integratu zitezkeen erlazio sintagmatikoak Dentsitate Kontzeptualean, (Agirre et al. 1994b) lanak LPPL-tik erauzitako erlazio paradigmatiko eta sintagmatikoak Distantzia Kontzeptualaren bidez erabiltzeko proposamenaren ildotik doana.

- Inplementazioa azkartu: nahiz eta Dentsitate Kontzeptualaren algoritmoa konplexutasun gehiegizkoa ez izan, egungo inplementazioa baina azkarrago bat lor daitekeela uste dugu. Horren arrazoietako bat LISP lengoaiaz inplementatuta egotea da, eta bestea WordNet-eko informazioaren atzipena ez dagoela optimizatuta. Egun, C++ lengoaiaz inplementatutako bertsio bat lantzen ari gara, UNED-eko Elektrizitate eta Elektronika saileko ikerkuntza taldearekin batera, ITEM[80] proiektuaren barruan. Bertsio hau ingeniaritza linguistikorako GATE[81] ingurunearen barruan (Cunningham et al. 1997) integratuta egongo da laster, hitzen adiera-desanbiguaziorako moduluaren barruan.Inplementazioa azkartu.

VII.C.2. Hitzen adiera-desanbiguazioa (IV. kapitulua)

Egindako esperimentuetan baziren hobetu zitezkeen alor batzuk:

- Diskurtso-egituraren araberako testu zatiak batera desanbiguatu. Horrela egin izanez gero hitzak bakarka desanbiguatu ordez testu zati oso bat batera desanbiguatu zitekeen, eraginkortasun hobeagoa lortuaz. Gainera doitasuna ere hobetuko litzateke, zerikusirik ez duten testu zatiak alde batera utziko ziren eta.

- Dentsitatearen neurri eta adiera-aukeraketaren artean koerlaziorik ote dagoen ikertzea interesgarria izango litzateke. Koerlazioa balego Dentsitatearen balio batetik behera daudenak desanbiguatu gabe utzi eta doitasuna hobetuko litzateke (estaldura gutxitzearen truke).

HADrako sistema ahaltsuago bat egin nahi bada, erlazio-izaeraz gain desanbiguazioan erabilgarriak diren bestelako informazio iturriak (IV.A.1 atalean aipatu bezala) ere erabiltzea beharrezkoa da. Honen adibideak dira, adieren maiztasunak, bai orokorrean edo desanbiguatzen ari garen testuan, adiera beti kolokazio modura azaltzen ote den, adiera bakoitzaren inguruan dagoen egitura sintaktikoari buruzko informazioa, eta abar. Horrela adiera-desanbiguaziorako sistema osoago bat eraikiko genuke, Dentsitate Kontzeptualaren bidez informazio lexikal-semantikoa kodetzen duena, eta hau bestelako ezagutzarekin konbinatzeko gai dena.

[80] http://sensei.ieec.uned.es/item/

[81] http://www.dcs.shef.ac.uk/research/groups/nlp/gate/

Tesi hau idazten ari garen bitartean, SENSEVAL txapelketa[82] gertatzen ari da. Mundu mailan adiera desanbiguatzen duten sistemek parte hartzen dute. Txapelketa horretarako, Yarowsky-ren (1995) lana, Dentsitate Kontzeptuala eta hiztegietan oinarritutako bestelako erlazio-izaeraren neurriak (VI. kapituluan azaldu ditugunak) integratzen saiatzen ari gara.

VII.C.3. Zuzenketa utomatikoa (V. kapitulua)

Esperimentua diseinatzeko orduan ez genuen kontuan hartu ikasteko corpusa (Brown) eta probatzekoa (Bank of English) dialekto ezberdinekoak zirenik. Ziurra da arazo honek maiztasun orokorrak erabiltzen dituen heuristikoaren eta Testuinguru-Estatistikak erabiltzen dituenaren emaitzak kaltetu dituela. Komenigarriena Bank of English corpuseko bertako datuetatik ikastea izango litzateke, baina tamalez datu horiek eskuratzeko murrizpen gogorrak daude. Murrizpen hauen ondorioz errore errealen corpusak oso testuinguru txikia zeukan errorearen inguruan. Horrek modu erabakiorrean kaltetu du Dokumentuko Maiztasunen teknika, bestela oso indartsua zena. Arazo horiek konpondu ondoren doitasuna nabari hobetuko delakoan gaude.

Zuzenketa automatikoaren doitasuna hobetzeko beharrezkoa da erabilitako ezagutza fintzea. Murrizpen-Gramatika, adibidez, erroreak dituzten testuetara hobeto egokitu daiteke, guk erabili dugun bertsioa ez baitzegoen horretarako diseinatua. Dentsitate Kontzeptuala ere, arestian aipatu bezala aberastuz gero, emaitza hobeagoak espero daitezke. Bestalde, WordNet-en aberasketak Dentsitatea kategoria ezberdineko kontzeptuetara zabalduko luke, eta horrela Dentsitateak zuzenketan zuen estaldura areagotu.

Bukatzeko, kapitulu honetako emaitzek ez dute berretsi III. kapituluan aipatu dugun Dentsitate Kontzeptualaren ezaugarrietako bat: hitzen arteko erlazio-izaera neurtzeko baliagarria izatea. Zuzenketa automatikorako erabili dugun algoritmoan adieren arteko Dentsitate handiena zuten proposamena hautatu dugu, baina bestelako konbinazioak ere probatu beharko genituzke, adibidez, proposamen bakoitzaren adiera guztien Dentsitatea batu eta batura handiena duen hitza aukeratu proposamen egoki bezala.

VII.C.4. Baliabide lexikalak areago sendotu (VI. kapitulua)

VII.C.4.a) Kontzeptuen arteko lotura eleanitzak

Hiztegi elebidun zabalagoak erabiliz gero estaldura eta doitasun hobeak lortuko lirateke LPPL-WordNet loturan: alde batetik elebiduna-WordNet zabalagoa edukiko genukeelako, eta bestetik,

[82] http://www.itri.bton.ac.uk/events/senseval/cfp2.html

LPPL-ko adieraren baterako itzulpenik ez egotea errore-iturri denez, elebidun zabalagoarekin halako erroreak gutxituko lituzkeelako.

Elebiduna-WordNet loturaren estaldura jasotzeko beste modu bat frantsez-hitz/ingeles-hitz bikoteetan oinarritutako heuristikoak dira (Okumura & Hovy, 1994; Rigau & Agirre, 1995; Atserias et al. 1997). EuroWordNet proiektuan halako heuristikoak arrakastaz erabiltzen ari dira gaztelerarako WordNet-a eraikitzeko. Hala ere hitz bikote hauek badute murrizpenik, ez baitira hiztegi elebiduneko adierak kontuan hartzen, eta horrek arazoak sortu ditzake.

Adiera elebidunak erabiltzeari esker, WordNet eta LPPL hiztegi elebidunetan dagoen informazio zabalarekin (Fontenelle, 1997) aberastu zitekeen. Arlo hau jorratzea interesgarria iruditzen zaigu.

Gaur egun, EuroWordNet eta ITEM proiektuei lotuta, Euskararako WordNet-a ere eraikitzen ari gara, VI. kapituluan aurkeztutako teknikak eta arestian aipatutako hitz bikoteak euskara-ingelesa hiztegi elebidunari aplikatuaz. ITEM proiektuan gaztelerarako WordNet ere eraikitzen ari den heinean, hiztegi elebidunen kateak erabiliaz (euskara-gaztelera, euskara-ingelesa eta gaztelera-ingelesa) estaldura eta doitasuna hobetuko direlakoan gaude.

Atal honetarako garatutako metodoak baliabide lexikal egituratuak lotzeko balio duenez, ontologia eta EBLen bat egitean eragin handia izan dezake. Baliabide batek besteak duen ezagutza xurgatu eta ontologia aberatsagoak eraikitzen joateko bide egokia dirudi honek, *ANSI Ad Hoc Ontology Standards Group*[83] komiteak proposatzen duen bidea jarraituz (Hovy, 1997a; 1997b).

VII.C.4.b) Genus-desanbiguazioa

Nahiz eta lortu ditugun emaitzak oso onak izan, badago desanbiguazioaren doitasuna altxatzeko metodorik. Berriki (Rigau et al. 1998) artikuluan azaldu dugun bezala, VI. kapituluan azaldutako metodoa DGILE hiztegiari aplikatu ondoren, genusak multzokatu egin genituen, WordNet-era lotzean lortu den kode semantikoaren arabera. Kode semantiko bakoitzerako genus usuenak bakarrik aukeratuaz doitasuna altxatu egiten da, estalduraren golkora. Hobekuntza hau LPPL-rekin egitea ere otu zitzaigun, noski, baina LPPL-ren neurri txikia dela eta genus usuenak ez ziren oso maiz gertatzen, eta ez genuen emaitzak hobetzerik lortu.

Bartzelonako UPC-n lexikografia konputazionalean aritzen den taldearekin batera egindako ikerkuntzak hiztegi txikietarako metodoak handietan ere balio izan digutela ondorioztatu du. Hiztegi

[83] http://ksl-web.stanford.edu/onto-std/

handietatik hierarkia zabalago eta interesgarriagoak jasotzen dira, eta hobekuntzarako aukera gehiago eskaintze dute.

Bozketaren emaitzei dagokionean, bozkatzeko modu sofistikatuagoak erabiltzea aztertzea interesgarria dela uste dugu. Azterketa txiki batean ikusi genuen boza ematen dutenen artean gutxienez bosten adostasuna eskatuz gero %95eko doitasuna lortu genezakeela, baina estalduraren kaltetan (%18).

Bestalde, desanbiguazioa egitean definizioan bertan zegoen informazioa erabili dugu, baina hierarkien arteko desanbiguazioa ere planteatu daiteke, hau da, genus bat desanbiguatzerakoan, ziurtzat dauzkagun definiendumaren hiponimoak eta genusaren adiera bakoitzaren beste hiponimo eta hiperonimoak ere kontuan har ditzakegu.

Ildo berean, behin "txapela" eginda ere, errazagoa izan daiteke genus desanbiguazioa egitea.

VII.C.4.c) Hiztegietatik erauzitako hierarkien lotzea

Lan honetan, hierarkien eraikuntzan sinonimia erlazioa ez dugu kontuan hartu. Autore gehienek hiztegietatik erauzitako sinonimia erlazioari ez diote jaramonik egin, baina LPPL-ren kasuan sinonimo bidezko definizioak ugariak dira (adiera guztien %20). Artolak (1993) bere HEBan sinonimo ziren adieretan erlazioak sinonimo batetik bestera kopiatzen zituen, eta horren eragina neurtzea interesgarria deritzogu. Bestelako hurbilpenak ere aztertu beharko lirateke: adibidez, WordNet-en sinonimo diren adiera guztiak kontzeptu bakarra osatzen dute.

Nahiz eta hierarkiak lotzeko metodoak etorkizuna duela erakutsi dugun, ez dugu zehatz-mehatz ebaluatu lortzen den hierarkiaren kalitatea. Halakoak ebaluatzeko ez dago gaur egun irizpide finkorik, ez bada hiponimo/hiperonimo lotura bakoitzaren doitasuna, jadanik eman duguna (%82). Neurri hori oso mugatua izanda, aplikazio batetarako –adibidez, informazioaren erauzketa– erabilgarria den edo ez izan daiteke irizpide interesgarria. Bestalde ezin da ahaztu *ANSI ad hoc Ontology Standards Group* delakoa, arestian aipatu duguna, bere zereginen artean ontologiak ebaluatzeko irizpideak lantzen ari dela, oraingoz emaitza gabe.

VII.C.4.d) Sorgin-gurpila

LPPL-WordNet lotura, LPPL-ko genusen desanbiguazioa eta LPPL-ren txapelaren eraikuntzaren artean elkarrekintza konplexuak gerta daitezke. Tesi lan honetan bata bestearen ondoren egin izan dira, baina hiru prozesuen arteko elkarrekintza hobeto aztertu beharko litzateke. Behin LPPL-ko hierarkiak desanbiguatu eta txapelaren bidez lotu ondoren, LPPL-WordNet lotura egiteko

informazio gehiago dugu, hierarkiak lotzen ari baikara, eta suposatu daiteke emaitza hobeagoak lortu ditzakegula. Bestalde jadanik aipatu dugu, behin txapela eginda, errazagoa izan daitekeela genus desanbiguazioa egitea. Lotura elebidun hobeekin, bai txapela eta bai genus desanbiguazioa ere hobetu daitezke, eta abar. Prozesu iteratibo bat plantea daiteke.

Bestelako hurbilpen interesgarri bat sare neuronalen eskutik etorri daiteke. Kapitulu honetan deskribatutako emaitza guztiak –LPPL-WordNet lotura, LPPL-ko hiponimo/ hiperonimo erlazioak, WordNet beraren hierarkia– sare neuronal bateko arku bezala errepresenta daitezke. Sare neuronalari energia funtzio egoki bat esleituz gero, ezagunak diren teknikak aplikatu daitezke arkuen konbinazio ezin hobea bilatu dezan. Horrela aldi berean erabakiko luke zein den adiera bakoitzerako WordNet lotura hoberena eta hiperonimo hoberena.

VII.C.4.e) Bestelakoak

EBLen sorkuntza eta aberasketa automatikoa landu dugun arren, ez ditugu horren ikuspuntu guztiak landu. Hutsune nagusietako bat definizioen ***differentia*-tik erauzitako informazioa** (Artola, 1993) landu ez dugula izan da. Differentia-ren erabilera betidanik interesgarritzat jo izan da eta egungo lanek (ikus adibidez, Richardson, 1997) arnas berria eman diote bertatik erauzitako informazioaren erabilgarritasunari. Bestalde, adieren adibideen analisia bultzatu beharko litzatekela uste dugu, adiera azaltzen den testuinguruei buruzko informazio interesgarria ematen dute eta.

Hierarkia eleanitzen eraikuntza automatikoa tesi-lan honetatik gertu dagoen alorra da. Hizkuntza ezberdinetako baliabideak lotzen goazen heinean, hierarkia eleanitzak osatzen ari gara. Inplizituki, ontologia ezberdinetako informazio (erdi-) automatikoki hizkuntza batetik bestera xurgatzea posiblea den edo ez ikertzen ari gara, eta aldi berean hierarkia unibertsalak eraiki daitezkeen edo ez aztertzen. Aldi berean, hierarkia ezberdinetako informazioa bateragarria den edo ez, goi mailak automatiko lotzea komeni ote den, eta antzeko galdera ugari sortzen zaizkigu. Galdera hauek baliabide lexikal egituratuen eraikuntzatik munduaren eta hizkuntzen ereduen azterketara garamatzate.

Euskarari dagokionean, ezin dugu aipatu gabe utzi **Euskal Hiztegiaren** gainean gure taldean egiten ari den lana. Lan horren helburua euskararako Ezagutza-Base zabala sortzea, informazio semantikoan aberatsa. Horretarako hiztegiaren egituraren azterketa egin da eta TEI gidalerroak jarraitzen dituen kodeketara itzuli dugu (Arriola et al. 1995; 1996a; 1996b). Jadanik bukatu dugu izenen definizioetako erlatore berezi eta genusen bilaketa (Agirre et al. 1998a), eta gaur egun aditz eta adjektiboen analisia, adieren adibideen analisia, eta WordNet-en lotura lantzen ari gara. Hurrengo pausoan, VI. kapituluan azaldutako metodoen bidez, izen, aditz eta adjektiboen

hierarkien eraikuntzari ekingo diogu, taldeko kide baten tesiaren barruan, atal honetan planteatu ditugun hobekuntzak aplikatzen saiatuz. Bestalde, Euskal Hiztegiko definizioetan erabiltzen den hizkuntzaren azterketa ere martxan dago, oraingo genus eta erlatore bidezko bilaketa bezala. Etorkizun laburrean taldean garatzen ari diren gramatiken bidez, *differentia*-ren azterketari ekingo diogu, taldeko kide baten tesia izango den lanean.

BIBLIOGRAFIA

Aduriz, I., Alegria, I., Artola, X., Ezeiza, N., Sarasola, K. and Urkia, M. 1997. A Spelling Corrector for Basque Based on morphology, in *Literary and Linguistic Computing*, vol. 12, no. 1. Oxford University Press (Oxford, England).

Agirre, E. 1993. Contribución de la Información Léxico-Sémantica en la Automatizción de la Corrección de Errores, in *Workshop sobre Lexicografía Computacional*. Unpublished paper (Donostia, Basque Country).

Agirre, E. and Rigau, G. 1995. A proposal for Word Sense Disambiguation using Conceptual Distance, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Agirre, E. and Rigau, G. 1996a. Word Sense Disambiguation using Conceptual Density, in *Proc. of COLING* (Copenhagen, Denmark).

Agirre, E. and Rigau, G. 1996b. An Experiment on Word Sense Disambiguation of the Brown Corpus using WordNet, in *MCCS-96-291*. Computing Research Laboratory (Las Cruces, New Mexico).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994a. A methodology for the extraction of semantic knowledge from dictionaries using phrasal patterns, in *Proc. of IBERAMIA. IV Congreso Iberoamericano de Inteligencia Artificial*, pp. 263-270. McGraw-Hill (Caracas, Venezuela).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994b. Conceptual Distance and Automatic Spelling Correction, in *Proc. of the Workshop on Computational Linguistics for Speech and Handwriting Recognition* (Leeds, England).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994c. Intelligent Dictionary Help Systems, in Brekke, M.; Andersen. I.; Dahl, T. and Myking, J. (eds.) *Applications and Implications of current LSP Research*. Fakbokforlaget (Norway).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. 1994d. Lexical Knowledge Representation in an Intelligent Dictionary Help System, in *Proc. of COLING* (Kyoto, Japan).

Agirre, E., Arregi, X., Artola, X., Díaz De Ilarraza, A., Sarasola K. 1995. Lexical-Semantic Information and Automatic Correction of Spelling Errors, in K. Korta & J. M. Larrazabal (eds.) *Semantics And Pragmatics Of Natural Language: Logical And Computational Aspects*, no. 1. Ilcli Series (Donostia, Basque Contry).

Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K. and Soroa, A. 1997. Constructing an Intelligent Dictionary Help System, in *Natural Language Engineering*. Cambridge University Press (Cambridge, England).

Agirre, E., Ansa, O., Arregi, X., Arriola, J.M., Díaz de Ilarraza, A., Lersundi, M., Soroa, A. and Urizar, R. 1998a. Extracción de relaciones semánticas mediante gramáticas de restricciones, in *Proc. of Sociedad Española para el Procesamiento del Lenguaje Natural* (Alicante, Spain).

Agirre, E., Gojenola, K., Sarasola, K. and Voutilainen, A. 1998b. Towards a Single Proposal in Spelling Correction, in *Proc. of the joint COLING and ACL meeting*.

Agirre, E., Gojenola, K., Sarasola, K. and Voutilainen, A. 1998c. Towards a Single Proposal in Spelling Correction, in *UPV/EHU-LSI TR 8-98*. UPV-EHU (Donostia, Basque Country).

Ahlswede, T.E. 1989. New technique for identifying relational structures in dictionary definitions, in U. Zernik (eds.) *Proc. of the 1st Intl. Lexical Acquisition Workshop*.

ALPAC 1966. Language and Machine: Computers in Translation and Linguistics. National Research Council (Washington, USA).

Alshawi, H. 1989. Analysing dictionary definitions, in B. Boguraev, T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*, pp. 153-169. Longman (New York, USA).

Alvar, M. (ed.) 1987. Diccionario General Ilustrado de la Lengua Española. Biblograf (Barcelona, Catalonia).

Amsler, R. A. 1981. Taxonomy for English Noun and Verbs, in *Proc. of the 19th Annual Meeting of the Association for Computational Linguistics*, pp. 133-138.

Arregi, X. 1995. Anhitz: itzulpenean laguntzeko hiztegi-sistema eleanitza, in *Ph.D. thesis*. UPV-EHU (Donostia, Basque Country).

Arriola, J.M and Soroa, A. 1996. Lexical Information Extraction for Basque, in *Student Conference in Computational Linguistics* (Montreal, Canada).

Arriola, J.M., Artola X., Soroa A 1995. Análisis automático del diccionario Hauta-Lanerako Euskal Hiztegia, in *Procesamiento del Lenguaje Natural*, no. 17, pp. 173-181. SEPLN (Bilbo, Basque Country).

Arriola, J.M., Artola X., Soroa A. 1996. Automatic extraction of lexical information from an ordinary dictionary, in *Proc. of EURALEX* (Göteborg, Sweden).

Artola, X. 1993. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza, in *Ph.D. thesis*. UPV-EHU (Donostia, Basque Country).

Atserias, J., Climent, S., Farreres, X., Rigau, G. and Rodríguez, H. 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Tchark, Bulgary).

Aulestia, G. and White, L. 1992. Euskara-ingelesa hiztegia. Elkar (Donostia, Basque Country).

Bar-Hillel, Y. 1960. Automatic Translation of Languages, in F. Alt, A. Donald Booth, and R.E. Meagher (eds.) *Advances in Computers*. Academic Press (New York, USA).

Basili, R., Della Rocca, M., Pazienza, M.T. and Velardi, P. 1995. Contexts and categories: tuning a general purpose classification to sublanguages, in *Proceedings of the Conference on Recent Advances on Natural Language Processing* (Tzigov Chark, Bulgary).

Basili, R. Della Rocca, M. and Pazienza, M.T. 1997. Towards a Bootstrapping Framework for Corpus Semantic Tagging, in *Proc. of the ACL-SIGLEX Workshop on Tagging text with Lexical Semantics: Why, What and How* (Washington, USA).

Bateman, J.A. 1990. Upper modeling: organizing knowledge for natural language processing, in *Proc. of 5th Intl. Workshop on Natural Language Generation* (Pittsburgh, USA).

Biblograf 1992. Diccionario Vox/Harrap's Esencial Español-Inglés. Biblograf (Barcelona, Catalonia).

Binot, J.L. and Jensen, K. 1987. A semantic expert using an online standard dictionary, in *Proc. of IJCAI*.

Bisson, G. 1995. Why and How to Define a Similarity Measure for Object-Based Representation Systems, in N.J.I. Mars (eds.) *Towards Very Large Knowledge Bases*. IOS Press.

Boguraev, B. and Briscoe, T. 1987. Large Lexicons for Natural Language Processing: Utilising the Grammar Coding System of LDOCE, in *Computational Linguistics*, vol. 13, no. 3-4.

Boguraev, B. and Briscoe, T. (eds.) 1989. Computational Lexicography for Natural Language Processing. Longman (New York, USA).

Briscoe, T., Copestake, A. and Boguraev, B. 1990. Enjoy the paper: lexical semantics via lexicology, in *Proc. of COLING*.

Briscoe, T., de Paiva, V. and Copestake, A. 1993. Inheritance, Defaults, and the Lexicon. Cambridge University Press (Cambridge, England).

Bruce, R. and Guthrie, L. 1991. Building a Noun Taxonomy from a Machine Readable Dictionary, in *MCCS-91-207*. Computing Research Laboratory (Las Cruces, New Mexico).

Bruce, R., Wilks, Y., Guthrie, L., Slator, B. and Dunning, T. 1992. NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour, in *MCCS-92-246*. Computing Research Laboratory (Las Cruces, New Mexico).

Byrd, R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S. and Rizk, O.A. 1987. Tools and Methods for Computational Lexicology, in *Computational Linguistics*, vol. 13, no. 2-4.

Byrd, R.J. 1990. Computational Lexicology for Building On-Line Dictionaries: the Wordsmith Experience, in L. Fignoni and C. Peters (eds.) *Computational Lexicology and Lexicography*. Giardini (Pisa, Italy).

Calzolari, N. 1983. Semantic links and the dictionary, in *Proc. of the Intl. Conference on Computers and the Humanities*.

Castellón, I. 1992. Lexicografia Computacional: Adquisición Automática de Conocimiento Léxico, in *Ph.D. thesis*. Universitat de Barcelona (Barcelona, Catalonia).

Chen, H., Lynch, K.J., Basu, K. and Ng, T.D. 1993. Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval, in *IEEE Expert*.

Chodorow, M.S., Byrd, R.J. and Heidorn, G.E. 1985. Extracting semantic hierarchies from large on-line dictionary, in *Proc. of the 23rd Annual Meeting of the Association for Computational Linguistics*.

Chodorow, M.S., Ravin, Y. and Sachar, H.E. 1988. A tool for investigating the synonymy relation in a sense desambiguated thesaurus, in *Proc. of the Conference on Applied Natural Language Processing* (Austin, USA).

Church, K. W., Hanks, P. 1990. Word Association Norms, Mutual Information, and Lexicography, in *Computational Linguistics*, vol. 16, no. 1.

Cohen, P. and Loiselle, C. 1988. Beyond ISA: Structures for Plausible Inference in Semantic Networks, in *Proc. of AAAI*.

Collins, A. M and Loftus, E. F. 1975. A Spreading-Activation Theory of Semantic processing, in *Psychological Review*, vol. 82, no. 6, pp. 407-428.

Copestake, A. 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary, in *Proc. of 1st Intl. Workshop on Inheritance in NLP* (Tilburg, Netherlands).

Cowie, J., Guthrie, J., and Guthrie, L. 1992. Lexical Disambiguation Using Simulated Annealing, in *Proc. of COLING* (Nantes, France), pp. 359-365.

Cucchiarelii, A. and Velardi, P. 1997. Automatic Selection of Class Labels from a Thesaurus for an Effective Tagging of Corpora, in *Proc. of the 5th Conference on Applied Natural Language Processing*, pp. 380-387.

Cunningham, H., Humphreys, K., Wilks, Y. and Gaizauskas, R. 1997. Software Infrastructure for Natural Language Processing, in *Proceedings of the Fifth Conference on Applied Natural Language Processing*.

Damerau, F.A. 1964. A technique for computer detection and correction of spelling errors, in *Information Processing and Management*, vol. 7, pp. 171-176.

Dietterich, T.G. 1997. Machine Learning Research: Four Current Directions, in *AI magazine*, vol. 18, no. 4, pp. 97-136.

EDR 1993. Electronic Dictionary Technical Guide, in *TR-042*. Electronic Dictionary Research Institute (Tokyo, Japan).

Elhuyar 1996. Elhuyar euskara-gaztelania hiztegia. Elhuyar K.E. (Usurbil, Basque Country).

Firth, J. 1956. A synopsis of linguistic theory 1930-1950, in M. Palmer (eds.) *Selected papers of J.R. Firth*. Longmans (London, England).

Fontenelle, T. 1997. Using a Bilingual Dictionary to Create Semantic Networks, in *International Journal of Lexicography*, vol. 10, no. 4.

Francis, S. and Kucera, H. 1967. Computing Analysis of Present-Day American english. Brown University Press.

Gale, W., Church, K. 1990. Poor Estimates of Context are Worse than none, in *Proc. of Compstat* (Dubrovnik, Yugoslavia). Springer-Verlag (New York, USA).

Gale, W., Church, K., Yarowsky, D. 1992. Work on Statistical Methods for Word Sense Disambiguation, in *Proc. of the AAAI Fall Symposium: Probabilistic Approaches to Natural Language Processing*.

Gale, W. A., Church, K.W. and Yarowsky, D. 1993. A Method for Disambiguating Word Senses in a Large Corpus, in *Computing and the Humanities*, no. 26, pp. 415-439.

Genthial, D., Courtin, J., Ménèzo, J. 1994. Towards a More User-Friendly Correction, in *Proc. of the Annual Meeting of the Association for Computational Linguistics* (Kyoto, Japan).

Golding, A. and Schaves, Y. 1996. Combining trigram-based and feature-based methods for context-sensitive spelling correction, in *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics* (Santa Cruz, USA).

Golding, A. R. 1995. A Bayesian hybrid method for context-sensitive spelling correction, in *Proc. of the 3rd Workshop on Very Large Corpora* (Cambridge, USA), pp. 39-53.

Gove, P.B. (ed.) 1969. The Webster's Seventh New Collegiate Dictionary. Merrian-Webster (Springfiled, Massachusets).

Grefenstette, G. 1992. Finding Semantic Similarity in Raw Text: the Deese Antonyms, in *Proc. of the AAAI Fall Symposium: Probabilistic Approaches to Natural Language Processing*.

Grefenstette, G. 1996. Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches, in Boguraev & Pustejovsky (eds.) *Corpus Processing for Lexical Acquisition*, ch. 11, pp. 213-225. MIT Press (Cambridge, Massachusetts).

Grishman, R. and Sterling, J. 1994. Generalizing Automatically Generated Selectional Patterns, in *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Gruber, T.R. 1993. Towards Principles for the Design of Ontologies for Knowledge Sharing, in *Proc. of the Intl. Workshop on Formal Ontology* (Padova, Italy). also as Technical Report KSL 93-04 (Stanford University, USA).

Guarino, N. 1997. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration, in Pazienza, M.T. (ed.) *Information Extraction*. Springer (Berlin, Germany).

Hearst, M., Schütze, H. 1993. Customizing a Lexicon to Better Suit a Computational Task, in *Proc. of the Workshop on Extracting Lexical Knowledge*.

Hearst, M. 1991. Toward Noun Homonym Disambiguation Using Local Context in Large Text Corpora, in *Proc. of the 7th Annual Conference of the UW Centre for the New OED and Text Research* (Waterloo, Canada).

Helmreich, S., Guthrie, L. and Wilks, Y. 1993. The use of machine readable dictionaries in the Pangloss project, in *Proc. of the AAAI Spring Symposium on Buildings Lexicons for Machine Translation*. AAAI Press.

Heylen, D., Maxwell, K.G. and Armstrong-Warwick, S. 1993. Collocations, Dictionaries and MT, in *Proc. of the AAAI Spring Symposium on Buildings Lexicons for Machine Translation*. AAAI Press.

Hirst G. 1987. Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press (Cambridge, England).

Hobbs, J. 1985. Ontological Promiscuity, in *Proc. of the 23rd Annual Meeting of the Association for Computational Linguistics*.

Hornby, A.S. (ed.) 1974. Oxford Advanced Learner's Dictionary of Current English. Oxford University Press (Oxford, England).

Hovy, E. and Nirenburg, S. 1992. Approximating an Interlingua in a Principled Way, in *Proceedings of the DARPA Speech and Natural Language Workshop* (Arden House, NY.).

Hovy, E. 1997a. Constructing and Using Large Ontologies, in *Unpublished Presentation on the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Ressources for NLP Applications* (Madrid, Spain).

Hovy, E. 1997b. A Standard for Large Ontologies, in *NSF Workshop on R&D Opportunities in the Government* (Waxhington, USA).

Ide, N. and Véronis, J. 1998. Introduction to the Special Issue on Word Sense Desambiguation: The State of the Art, in *Computational Linguistics*, vol. 24, no. 1.

Ide, N. and Véronis, J. 1994. Extracting Knowledge Bases From Machine-Readable Dictionaries: Have We Wasted Our Time?, in K. Fuchi and T. Yokoi (eds.) *Knowledge Building and Knowledge Sharing*. Ohmsha, Ltd. and IOS Press.

Ingels, P. 1996. Connected Text Recognition Using Layered HMMs and Token Passing, in K. Oflazer and H. Somers (eds.) *Proc. of the 2nd Conference on New Methods in Language Processing*, pp. 121-132.

Ingels, P. 1997. A Robust Text Processing Technique Applied to Lexical Error Recovery, in *Ph.D. thesis*. Department of Conputer and Information Science (Linköping, Sweden).

Ispell 1993. International Ispell Version 3.1.00.

Jones, M. P. and Martin, J. H. 1997. Contextual Spelling Correction Using Latent Semantic Analysis, in *Proc. of the Conference on Applied Natural Language Processing*, pp. 166-173.

Karlsson, F., Voutilainen, A., Heikkila, J. and Anttila, A. 1995. Constrait Grammar: a Language Independent System for Parsing Unrstricted Text. Mouton de Gruyter.

Karov, Y. and Edelman, S. 1996. Learning Similaity-Based Word Sense Disambiguation From Sparse Data, in *Proc. of the 6th Workshop on Very Large Corpora* (Copenhagen, Denmark).

Karov, Y. and Edelman, S. 1998. Similarity-based Word Sense Disambiguation, in *Computational Linguistics*, vol. 24, no. 1.

Kernighan, M., Church, K., Gale, W. 1990. A Spelling Program Based on a Noisy Channel Model, in *Proc. of COLING*.

Kilgarriff, A. 1997a. I don't believe in word senses, in *Computing and the Humanities*, no. 2.

Kilgarriff, A. 1997b. Evaluating Word Sense Disambiguation Programs: Progress Report, in *ITRI-97-11 Technical Report*. University of Brighton.

Kirkpatrick, B. 1987. Roget's Thesaurus. Longman (Harlow, England).

Klavans, J. and Tzoukermann, E. 1995. Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons, in *Machine Translation*, vol. 10, no. 3.

Knight, K. and Luk, S. 1994. Building a Large-Scale Knowledge Base for Machine Translation, in *Proc. of AAAI*.

Kozima, H. and Furugori, T. 1993. Similarity between Words Computed by Spreading Activation on an English Dictionary, in *Proc. of the 6th Conference of the European Chapter of the Association for Computational Linguistics*.

Kozima, H. and Ito, A. 1995. Context-Sensitive Measurement of Word Distance by Adaptive Scaling of a Semantic Space, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Kukich, K. 1990. A Comparison of Some Novel and Traditional Lexical Distance Metrics for Spelling Correction, in *Proc. of INNC* (Paris, France).

Kukich, K. 1992. Techniques for Automatically Correcting Words in Text, in *ACM Computing Surveys*, vol. 24, no. 4, pp. 377-439.

Larousse 1980. Le plus petit Larousse. Larousse (Paris, France).

Leacock, C., Chodorow, M. and Miller, G.A. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification, in *Computational Linguistics*, vol. 24, no. 2.

Lee, J.L. 1997. Similarity-Based Approaches to Natural Language Processing, in *Ph.D. thesis*. Harvard University Technical Report TR-11-97 (Cambridge, Massachusetts).

Lenat, D.B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure, in *Communications of the ACM*, vol. 38, no. 11.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, in *Proc. of the 1986 SIGDOC conference*. ACM (New York, USA).

Li, H. and Abe, N. 1995. Generalizing Case Frames Using a Thesaurus and the MDL Principle, in *Proc. of Recent Advances on Natural Language Processing*.

Li, H. and Abe, N. 1996. Learning Dependencies between Case Frame Slots, in *Proc. of the 13th Conference on Machine Learning*.

Mahesh, K., Nirenburg, S., Cowie, J. and Farwell, D. 1996. An Assesment of Cyc for Natural Language Processing, in *MCCS-96-302*. Computing Research Laboratory (Las Cruce, USA).

Mahesh, K., Nirenburg, S. and Beale, S. 1997. If You Have It, Flaunt It: Using Full Ontological Knowledge for Word Sense Disambiguation, in *Proc. of the Conference on Recent Advances in Natural Language Processing* (Tzigov Chark, Bulgary).

Maritxalar, M. and Díaz de Ilarraza, A. 1996. Hizkuntza baten ikaskuntza-prozesuan zeharreko tartehizkuntz osaketa, in *UPV/EHU-LSI TR 7-96*. EHUko Lengoaiak eta Sistema Informatikoak Saila (Donostia, Basque Country).

Markowitz, J. 1986. Semantically significant patterns in dictionary definitions, in *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Mays, E., Damerau, F., Mercer, R. 1991. Context Based Spelling Correction, in *Information Processing and Management*, vol. 27, no. 5.

McEnery, T. and Wilson, A. 1996. Corpus Linguistics. Edinburgh University Press.

McRoy, S. 1992. Using Multiple Knowledge Sources for Word Sense Discrimination, in *Computational Linguistics*, vol. 18, no. 1.

Menezo, J., Genthial D., and Courtin J. 1996. Reconnaisances pluri-lexicales dans CELINE, un système multi-agents de détection et correction des erreurs, in *Proc. of the Conference on NLP+IA* (Moncton, Canada).

Michiels, A. and Nöel, J. 1982. Approaches to thesaurus production, in *Proc. of COLING*.

Michiels, A. 1996. An experiment in translation selection and word sense discrimination, in *http://engdep1.philo.ulg.be/michiels/wdts.htm*.

Miller, G., Leacock, C., Tengi, R. and Bunker, T. 1993a. A Semantic Concordance, in *Proc. of ARPA Workshop on Human Language Technology*.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. Miller, K. and Tengi, R. 1993b. Five Papers on WordNet, in *CSL Report 43*. Cognitive Science Laboratory, Princeton University.

Morris, M. 1998. Ingelesa-euskara hiztegia. Eusenor (Donostia, Basque Country).

Nakamura, J., Nagao, M. 1988. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation, in *Proc. of COLING* (Budapest, Hungary).

Niwa, Y., Nitta, Y. 1994. Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, in *Proc. of COLING* (Kyoto, Japan).

Okumura, A. and Hovy, E. 1994. Lexicon-to-Ontology Concept Association Using a Bilingual Dictionary, in *Proc. of the 1st AMTA Conference*.

Onyshkevych, B. and Nirenburg, S. 1994. The Lexicon in the Scheme of KBMT Things, in *MCCS-94-277*. Computing Research Laboratory (Las Cruces, New Mexico).

OUP 1974. Oxford French-English Dictionary. Oxford University Press (Oxford, England).

Procter, P. (ed.) 1978. Longman Dictionary of Contemporary English. Longman (London).

Quillian, M. R. 1968. Semantic Memory, in *Ph.D. thesis*. Carnegie Institute of Technology.

Rada, R., Mili, H., Bicknell, E., Blettner, M. 1989. Development and Application of a Metric on Semantic Nets, in *IEEE Transactions on systems, man, and cybernetics*, vol. 19, no. 1.

Resnik, P. 1992. WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery, in *Proc. of AAAI*.

Resnik, P. 1993a. Semantic Classes and Syntactic Ambiguity, in *Proc. of the ARPA Workshop on Human Language Technology* (Princeton, USA).

Resnik, P. 1993b. Selection and Information: A Class-Based Approach to Lexical Relationships, in *Ph.D. thesis*. University of Pennsylvania.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in *Proc. of IJCAI*.

Resnik, P. 1997. Selectional Preference and Sense Disambiguation, in *Proc. of the ACL-SIGLEX Workshop on Tagging text with Lexical Semantics: Why, What and How* (Washington, USA).

Ribas, F. 1995. On Learning More Appropriate Selectional Restrictions, in *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*.

Richardson, S.D. 1997. Determining Similarity and Inferring Relations in a Lexical Knowledge Base, in *Ph.D. thesis*. The City University of New York.

Rigau, G. and Agirre, E. 1995. Disambiguating bilingual nominal entries against WordNet, in *Workshop On The Computational Lexicon - ESSLLI* (Barcelona, Catalonia).

Rigau, G., Rodríguez, H. and Turmo, J. 1995. Automatically Extracting Translation Links Using a Wide Coverage Semantic Taxonomy, in *Proc. of the 15th Intl. Conference on Artificial Intelligence* (Montpellier, France).

Rigau, G., Atserias, J. and Agirre, E. 1997. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation, in *Proc. of ACL/EACL* (Madrid, Spain).

Rigau, G., Rodriguez, H. and Agirre, E. 1998. Building Accurate Semantic Taxonomies from Monolingual MRDs, in *Proc. of the joint COLING and ACL meeting* (Montreal, Quebec).

Rigau, G. 1998. Automatic Acquisition of Lexical Knowledge from Machine Readable Dictionaries, in *Ph.D. thesis*. Polytechnic University of Catalonia (Barcelona, Catalonia).

Rizk, O. 1989. Sense Disambigution of Word Translations in Bilingual Dictionaries: Trying to Solve the Mapping Problem Automatically, in *RC 14666*. IBM Research Division, T.J. Watson Research Center (New York, USA).

Sarasola, I. 1997. Euskal Hiztegia. Gipuzkoako Kutxa (Donostia, Basque Country).

Schütze, H. 1998. Automatic Word Sense Discrimination, in *Computational Linguistics*, vol. 24, no. 1.

Schütze, H. 1992a. Word Sense Disambiguation With Sublexical Representations, in *Proc. of the AAAI Workshop on Statistically-Based Natural Language Processing Techniques*.

Schütze, H. 1992b. Context Space, in *Proc. of the AAAI Fall Symposium: Probabilistic Approaches to Natural Language Processing*.

Sinclair, J. (ed.) 1987. Colllins COBUILD English Language Dictionary. Collins (London, England).

Sussna, M. 1993. Word Sense Disambiguation for Free Text Indexing Using a Massive Semantic Network, in *Proc. of the 2nd Int. Conf. on Information and Knowledge Management* (Airlington, USA).

Svartvik, J. (ed.) 1990. The London-Lund Corpus of Spoken English. Lund University Press.

Towell, G. and Voorhees, E.M. 1998. Disambiguating Highly Ambiguous Words, in *Computational Linguistics*, vol. 24, no. 1.

Tsurumaru, H., Hitaka, T., and Yoshida, S. 1986. An attempt to automatic thesaurus construction from an ordinary japanese dictionary, in *Proc. of COLING*.

Tversky, A. 1977. Features of Similarity, in *Psychological Review*, vol. 84, no. 4, pp. 327-354.

Urkia, M. and Sagarna, A. 1990. Terminologia y lexicografia asistidas por ordenador: la experiencia de UZEI, in *Proc. of SEPLN* (Donostia, Basque Country).

Utiyama, M. and Hasida, K. 1997. Bottom-up alingnment of Ontologies, in *Proc. of the IJCAI Workshop on Ontologies and Multilingual Natural Language Processing.*

UZEI lantaldea 1982. Hizkuntzalaritza/1 hiztegia. UZEI (Donostia, Basque Country).

Véronis, J. and Ide N. 1990. Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries, in *Proc. of COLING* (Helsinki, Finland), vol. 2.

Vosse, T. 1992. Detecting and Correcting Morpho-syntactic Errors in Real Texts, in *Proc. of the 3rd Conference on Applied Natural Language Processing* (Trento, Italy), pp. 111-118.

Vosse, T. 1994. The Word Connection: Grammar-based Spelling Error Correction in Dutch, in *Ph.D. thesis*. Unit for Experimental and Theoretical Psychology (Univ. of Leiden, Holland).

Vossen, P. and Serail, I. 1990. Devil: a taxonomy-browser for decompositiona via the lexicon, in *Technical Report*. Faculty of Arts, University of Amsterdam.

Vossen, P., Díez-Orzas, P. and Peters, W. 1997. The Multilingual design of the EuroWordNet Database, in *Proc. of the IJCAI Workshop on Multilingual Ontologies for NLP Applications.*

Vossen, P. 1989. The structure of lexical knowledge as envisaged in the LINKS-project, in J-. Conolly and S. Dik (eds.) *Functional Grammar and the Computer*. Dordrecht: Foris.

Vossen, P. 1990. The end of the chain: Where does decomposition of lexical knowledge lead us eventually?, in *Proc. of the Conference on Functional Grammar*.

Vossen, P. 1996. Right or Wrong: combining Lexical Resources in the EuroWordNet Project, in *Proc. of EURALEX*.

Wilks, Y., Fass, D., Guo, C., McDonald, J.E., Plate, T., and Slator, B.M. 1990. Providing Machine Tractable Dictionary Tools, in *Machine Translation*, no. 5, pp. 99-154.

Wilks, Y., Slator, B.M., and Guthrie, L. 1996. Electric Words: Dictionaries, Computers, and Meanings. The MIT Press (Cambridge, USA).

Microsoft Corporation 1997. Word 97.

Yarowsky, D. 1992. Word sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, in *Proc. of COLING* (Nantes, France), pp. 454-460.

Yarowsky, D. 1993. One Sense per Collocation, in *Proc. of the 5th DARPA Speech and Natural Language Workshop*.

Yarowsky, D. 1994. Decision Lists for Lexical Ambiguity Resolution, in *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Yokoi, T. 1995. The EDR Electronic Dictionary, in *Communications of the ACM*, vol. 38, no. 11.