WNTERM: zientzia eta teknologiaren espezialitate-ontologia eleaniztuna

Eli Pociello, Antton Gurrutxaga, Mari Susperregi
Elhuyar

Eneko Agirre, German Rigau
Ixa taldea, UPV/EHU

Laburpena

Artikulu honetan, WNTERM lexikoi espezializatu eleaniztuna aurkezten dugu, eta hori egiteko erabili dugun metodologia deskribatzen. WNTERM sortu da EuroWordNeten oinarritutako Multilingual Central Repository (MCR) izeneko baliabidea Elhuyarren Zientzia eta Teknologiaren Hiztegi Entziklopedikoaren *(ZTH) bidez elikatuz. Hala, lan honen emaitza euskarazko eta ingelesezko espezialitate-ontologia eleaniztun bat da, kontzeptuen arteko erlazio taxonomikoak eta bestelako erlazio semantikoak dituena, eta beste wordnetekin ere lotua dagoena.*

0. **Sarrera eta aurkezpen orokorra**

Informazioaren Gizartearen gaur egungo eskakizun eta, aldi berean, aukera nagusietako bat da ezagutzaren kudeaketako sistemak garatzea. Sistema horietako bat, adibidez, web semantikoa da. Horien oinarrian, ontologiak, ezagutza-baseak eta kideko baliabideak daude, hau da, kontzeptuak eta horien arteko erlazioak adierazten dituzten datu-egiturak.

Ixa taldeak[1] arlo horretan euskararentzat ezinbestekoak diren lehen urratsak egin ditu, eta horren emaitza da Euskal WordNet[2] (Pociello et al 2008), 2000. urtean garatzen hasi eta etengabe aberastuz doan ezagutza-base lexikala (EBL). Euskal WordNet, EuroWordNet (Vossen 1998) eta WordNet (Fellbaum, 1998) ereduetan oinarritzen da, eta hitz bakoitzaren adierak egituratutako kontzeptuetara lotzen ditu. Euskal WordNetek izen, adjektibo eta aditzen informazioa du, kontzeptuen inguruan antolatuta. Kontzeptu horiek

[1] http://ixa.si.ehu.es
[2] Euskal WordNet doan kontsultatu eta jaitsi daiteke:[http://ixa2.si.ehu.es/mcr/]

definizioez eta adibideez hornituak daude. Horrez gain, kontzeptuak hierarkia batean antolatuta daude, eta kontzeptuen arteko erlazio lexiko-semantiko ugari topa daitezke (hiponimia, meronimia eta abar).

Euskal WordNet datu-basean, gaztelania, katalana, galiziera, ingelesa eta italiera daude lotuta. Hala, kontzeptuak hizkuntzen artean komunak dira, eta hizkuntza bateko hitz baten adierak beste hizkuntza batean dituen ordainak jakiteko balio du. Wordnet horiek guztiek *Multilingual Central Repository* (MCR) deritzoguna osatzen dute (Atserias et al 2004).

Azken urteotan hizkuntza eta terminologia espezializatuak izan duen gorakadaren eraginez, geroz eta konplexuagoa da informazio terminologikoa jasotzea eta antolatzea. Espezialitate-ontologiak egokiak dira informazio hori guztia antolatzeko, eta ezagutzaren errepresentazioan, kudeaketan eta partekatzean oso lagungarriak direla ere erakutsi dute. Adibidez, badira espazialitate-ontologia aipagarriak e-merkataritzan (UNSPSC,[3] NAICS),[4] medikuntzan (GALEN,[5] UMLS), eta ingeniaritzan —EngMath (Gruber & Olsen 1994), PhysSys (Borst 1997)—. Honezaz gain, azken urteotan, Hizkuntza Prozesamenduan (HP) espezialitate-ontologiak oso erabiliak dira aplikazio eta sistema konputazional ugari garatzeko (Navigli et al 2003; Sagri et al 2003; Stamou et al 2002; Roventini & Marinelli 2003).

Orain dela gutxi arte, ordea, Euskal WordNeten helburua hiztegi orokorra lantzea izan da. Hori dela eta, espezialitate-arloko aplikazioak garatzean, Euskal WordNeten hedadura urria dela ikusi da, arloko hitz, termino eta adierak falta baitzitzaizkion.

Egoera horri aurre egiteko, Euskal WordNet zientzia eta teknologiaren espezializazio-arloetara hedatzea komeni dela ikusi dugu. Horretarako, WNTERM deitu dugun proiektu honetan, Ixa taldeko «Lexikoa eta Semantika» azpitaldea eta Elhuyar Hizkuntza & Teknologia unitateko I+G taldea aritu dira elkarlanean (Pociello et al 2008). Hala, Elhuyar Fundazioak eratutako *Zientzia eta Teknologiaren Hiztegi Entziklopedikoa*n[6] oinarrituta (ZTH), zientzia eta teknologiaren hainbat arlotako ontologiak MCR ereduaren arabera garatu dira, eta WNTERM izeneko ezagutza-base eleaniztunean (euskara eta ingelesa) jasotzen dira. Arlo bakoitzeko kontzeptuen arteko erlazio taxonomiko eta semantikoak ikusgai daude, eta kontzeptu horiek MCR zein ZTH baliabideekin lotuak daude.

Artikulu honen helburua da WNTERM aurkeztea, eta WNTERM lortzeko jarraitutako prozedura deskribatzea. Horrenbestez, hau da artikuluaren egitura: 1. eta 2. ataletan, proiektuan erabili diren baliabideak labur aurkeztuko dira: MCR, Euskal WordNet eta ZTH. Ondoren, 3. atalean, espeziali-

---

[3] http://www.unspsc.org
[4] http://www.naics.com
[5] http://opengalen.org
[6] http://www.zientzia.net/hiztegia

tate-ontologia eratzeko jarraitu dugun metodologia deskribatuko da. Azkenik, 4. atalean, ondorioak eta hemendik aurrerako lanak aipatuko dira.

1. **The Multilingual Central Repository**

WNTERM hizkuntza batzuetako wordnetetara lotzeko, Multilingual Central Repository (Atserias et al 2004) egiturari jarraitzea erabaki dugu. MCR ereduak ingelesezko, gaztelaniazko eta euskarazko wordneten arteko loturak ezartzen ditu, eta gainera, Euskal WordNet eta WNTERM aldi berean lantzeko eta garatzeko aukera eskaintzen du.

MCR interfaze eleaniztuna da, non Europa Batzordeko «MEANING: Developing Multilingual Web-Scale Language Technologies» (IST-2001-34460) proiektuan (Rigau et al 2003) aztertu den informazio guztia integratzen den. Ezagutza-base honek EuroWordNeten[7] eredua (Vossen 1998) jarraitzen du: hizkuntza bakoitzeko izen, aditz, adjektibo eta adberbioak *synonym set* edo *synset*etan (sinonimo-multzotan) antolatzen ditu, eta synset horien arteko erlazio semantikoak jasotzen (hiperonimia, hiponimia, meronimia, holonimia), hierarkia bat osatuz. Horietako synset bakoitza kontzeptu lexikal bati dagokio, eta gehienek glosa bat dute, kontzeptuak adierazten duenaren azalpen edo adibide banarekin.

MCRk sei hizkuntzatako wordnetekin egiten du lan: euskara, katalana, galiziera, ingelesa, italiera eta gaztelania. Sei hizkuntza horietako izen, aditz, adjektibo eta adberbioen adieren inbentarioa da, eta, EuroWordNeten eredua jarraitzen duenez, hizkuntza guztiak lotuta daude. Beraz, MCR neurri handiko baliabide linguistiko eleaniztuna da.

Gaur egun, MCRk synset arteko 934.771 erlazio semantiko jasotzen ditu, eta Princetoneko WordNet baino lau aldiz handiagoa da (WordNet 3.0 bertsioak 235.402 erlazio semantiko ditu).

Hurrengo ataletan (1.1 eta 1.2), WordNet espezialitate-ontologiaren eta Euskal WordNeten berri emango dugu, WNTERM proiektuaren garapenerako funtsezkoak izan baitira.

1.1. *WordNet «domeinu»- edo arlo-ontologia*

Magnini & Cavagliàren (2000) lanean oinarrituta, *WordNet Domains* izeneko *domain ontology* edo arlo-ontologia eratu da,[8] non synsetak erdi-automatikoki etiketatu diren arlo-marka batekin edo gehiagorekin. Arlo-marka horiek hierarkikoki antolatuta daude, eta synsetak arloaren arabera antolatu

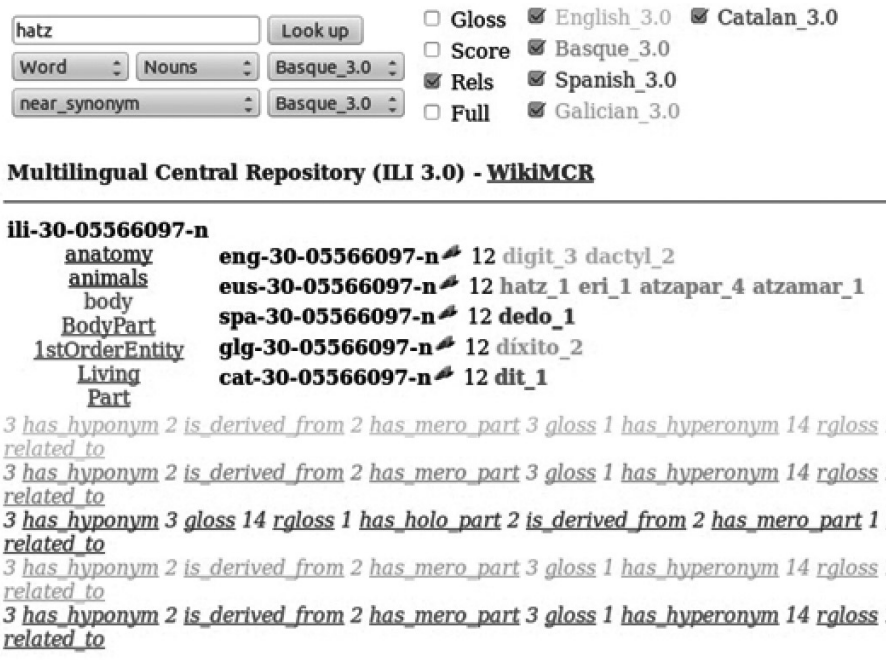
[7] http://www.illc.uva.nl/EuroWordNet
[8] http://wndomains.fbk.eu/

dira: *Transport*, *Sports*, *Medicine*, *Gastronomy*[9] eta antzekoetan. Esate baterako, *jokatu* aditzak kirol-arloko adiera duenean (*futbolean jokatu* diogunean, adibidez), synsetak *free time* arloaren marka darama; *zuzen jokatu* konbinazioko adiera duenean, berriz, synsetak *psychology* marka du. WNTERM proiekturako erabili dugun bertsioa 171 arlo-marka dituen hierarkia da.

1.2. *Euskal WordNet*

Euskal WordNet (Pociello et al 2008) Ixa taldeak garatutako ezagutza-base lexiko-semantikoa da, WordNeten (Fellbaum 1998)[10] eta horren ildotik sortutakoetan oinarritua (EuroWordNet eta MCR).

Euskal WordNet WordNeten gainean garatu da, Vossenen (1998) *expand approach*a jarraituz; hots, euskarako ordainak, WordNeteko hierarkiari jarraituz, bertako synsetei zuzenean esleitu zaizkie. Bestalde, euskarako kontzeptuak edo synset berriak ere gehitu dira: *ikastola, txakolina, trikitia*…



1. IRUDIA

Hatz *hitzaren kontsulta Euskal WordNeten*

[9] Domeinuen izendapena ingelesez dago MCRn, wordnet guztietan domeinu-izen horiek erabili ahal izateko. Guk ere halaxe erabiliko ditugu artikuluan zehar.

[10] http://wordnet.princeton.edu

1. irudian, Euskal WordNeten kontsulta-interfazearen irudia dugu. Bertan ikus daitekeenez, euskarazko wordneta, MCRren egiturari esker, beste hizkuntzetako wordnetekin (katalana, galiziera, ingelesa, italiera eta gaztelania) lotua dago, eta beste edozein hizkuntzatako wordnet batera lotu daiteke automatikoki, % 100eko fidagarritasunarekin.

Horrenbestez, Euskal WordNetek erabilera ugari izan ditu HP arloan eta aplikazio konkretuetan. Arlo zientifikoan, Ixa taldeak eta Elhuyarrek frogatu dute WordNet erabilgarria dela analisi semantikoan, hala nola adiera-desanbiguazioan (Martínez 2005), rol semantikoen detekzioan (Zapirain et al 2007), ikasketarako ariketen sorkuntza automatikoan (Aldabe 2011), edo sentimendu-markak dituzten lexikoak sortzeko (San Vicente et al 2013).

Gaur egun, Euskal WordNetek 33.302 synset ditu, 50.841 adiera eta 26.999 lema.

2. *Zientzia eta Teknologiaren Hiztegi Entziklopedikoa* (ZTH)

Elhuyarrek, bere hiztegigintza espezializatuko jardueran, hainbat baliabide terminologiko eratu ditu, horietatik nabarienetakoa *Zientzia eta Teknologiaren Hiztegi Entziklopedikoa*.[11] Egun, hiztegi honetan 50 jakintza-arlotako 23.000 kontzeptu baino gehiago jasotzen dira. Hiztegi honen helburua da zientzia eta teknologiari buruzko erreferentzia-informazio fidagarri, landu eta eguneratua eskaintzea, modu zehatz, argi eta ulergarrian, eta erabiltzaile-multzo zabala gogoan izanik. Zientzia eta teknologiako arlo guztietako informazioa biltzen duen lehen euskarazko hiztegi entziklopedikoa da, eta 200 adituk baino gehiagok parte hartu dute hiztegiko informazioa eta artikuluak osatzen.

Hiztegi honen datu-basean, kontzeptu bakoitza fitxa bat da, non kontzeptu horren informazio guztia jasotzen den: terminoak (euskaraz, gaztelaniaz, frantsesez eta gaztelaniaz), definizioa eta jakintza-arloa. Kontzeptu bat arlo batekin baino gehiagorekin egon daiteke lotua. 1. taulan eman ditugu ZTHn bereizitako arloak, eta bakoitzean den kontzeptu-kopurua. Hala ere, kontzeptuen artean ez dute erlazio semantikoen bidezko loturarik. Horregatik, ZTHren hurrengo urratsetako bat WNTERM ezagutza-base lexiko-semantikoaren bidetik egingo da, datu-base terminologikoan kontzeptuen arteko erlazio hierarkiakoak adierazita egoteak aukera emango baitu hiztegiaren edukiak egituratzeko, erlazionatzeko, eta, azken buruan, erabiltzaileari informazio aberatsagoa eskaintzeko.

[11] http://zthiztegia.elhuyar.org/

1. TAULA

ZTHko kontzeptu-kopurua arloka

Abeltzaintza	114	Ingurumena	119
Aeronautika	314	Itsasoa	286
Albaitaritza	28	Kimika	1.726
Anatomia	698	Logika	54
Antropologia	61	Marrazkigintza	4
Argazkigintza	196	Matematika	949
Arkitektura	157	Materialak	75
Armagintza	98	Meatzaritza	17
Arrantza	28	Medikuntza	2.409
Astronomia	532	Metalurgia	352
Astronautika	152	Meteorologia	417
Automobilgintza	223	Mikologia	133
Biokimika	614	Mikrobiologia	414
Biologia	816	Mineralogia	434
Botanika	1.391	Nekazaritza	307
Ekologia	354	Orokorra	193
Elektronika	541	Ozeanografia	136
Elikagaigintza	78	Paleontologia	127
Eraikuntza	342	Psikiatria	78
Estatistika	168	Teknologia	1.025
Fisika	1.612	Teknologia elektrikoa	561
Fisiologia	214	Teknologia mekanikoa	315
Genetika	171	Telekomunikazioak	537
Geografia	128	Trenbidea	43
Geologia	1.306	Zoologia	1.889
		GUZTIRA	**23.656**

3. **WNTERM garatzeko metodologia**

Atal honetan, WNTERM garatzeko jarraitu ditugun pausoak deskribatuko ditugu. Lehenik, bi baliabideak (MCR eta ZTH) alderatu ditugu, bi baliabideetako terminoak eta arlo-markak kontuan hartuz. Gero, WNTERM ezagutza-basean sartu beharreko kontzeptuak eta terminoak nola aukeratu ditugun azalduko dugu.

3.1. *Baliabideen arteko alderaketa*

Aipatu dugun bezala, lehenengo egitekoa bi baliabideak (MCR eta ZTH) erkatzea izan da. Lan hau automatikoki egin da, eta emaitzak kualitatiboki aztertu dira. Terminoen grafia ez ezik, arloaren bat-etortzea ere hartu dugu kontuan. Horretarako, lehenik bi baliabideen arlo-marken mapaketa bat egin behar izan da.

3.1.1. Arlo-marken eskuzko mapaketa

MCRko eta ZTHko kontzeptuak sailkatzeko sistemak eta arlo-markak desberdinak dira. MCRn 171 arlo-marka daude hierarkikoki antolatuta (adibidez, *cinema*, *radio, post, tv, telegraphy* eta *telephony, telecommunications* arloaren azpiarloak dira). ZTHk 34 jakintza-arlo daude, eta ez dago azpiarlo-markarik.

Bi baliabideen arteko bat-etortzea neurtzeko arloen informazioa nahitaezkoa denez, lehenik eskuz mapatu ditugu MCRko 171 arlo-markak ZTHko 34 markekin. ZTHtik 18 arlo lotu dira MCRko arlo bakarrarekin, eta 16 MCRko arlo batekin baino gehiagorekin (esaterako, ZTHko *Telekomunikazioak* MCRko *cinema*, *radio*, *post*, *tv*, *telegraphy*, *telephony*rekin). MCRko arloei dagokienez, 64k ZTHko arlo bana dute, baina 20 ZTHko arlo bati baino gehiagori lotu zaizkio (hala nola, MCRko *town planning,* ZTHko *Arkitektura* eta *Eraikuntza* arloekin erlazionatu dena). MCRko 87 arlok ez dute baliokiderik izan mapatzean, espezifikotasun handiko arloak direlako (*fencing*, *paranormal*, *numismatics*), edo zientzia eta teknologiaren arlokoak ez direlako (*religion*, *sport*, *gastronomy*).

3.1.2. Terminoen arteko erkaketa automatikoa eta kasuistika

Ingelesezko terminoetan oinarritutako erkaketa automatiko bat egin dugu lehenik. Euskal WordNetekin alderatuta, ingeleseko wordnetak terminolo-

2. taula

Erkaketa automatikoan erabilitako kodeak

Kodea	Deskribapena
1 - 0	Ingelesezko terminoa MCRn dago, eta ZTHn ez.
1 - 1	Ingelesezko terminoa MCRn eta ZTHn dago, arlo berean.
1 - 2	Ingelesezko terminoa MCRn eta ZTHn dago, baina ez arlo berean.
1 - 3	Ingelesezko terminoa MCRn dago, eta ZTHn ez. Baina MCRko termino honen sinonimo bat badago ZTHn, eta, gainera, arlo berekoa da.
1 - 4	Ingelesezko terminoa MCRn dago, eta ZTHn ez. Hala ere, MCRko termino honen sinonimo bat badago ZTHn, baina arlo ezberdinekoak dira.
0 - 1	Ingelesezko terminoa ZTHn dago, eta MCRn ez.
1 - 1	Ingelesezko terminoa ZTHn dago eta MCRn dago, arlo berean.
2 - 1	Ingelesezko terminoa WNen eta MCRn dago, baina ez arlo berean.
3 - 1	Ingelesezko terminoa ZTHn dago, eta MCRn ez. Baina ZTHko termino honen sinonimo bat badago MCRn, eta, gainera, arlo berekoa da.
4 - 1	Ingelesezko terminoa ZTHn dago, eta MCRn ez. Hala ere, ZTHko termino honen sinonimo bat badago MCRn, baina arlo ezberdinekoak dira.

gia zientifiko-teknologiko handiagoa du. Euskal WordNet garatzean, hiztegi orokorra (eta ez zientifikoa) lantzea izan zen helburu nagusia. Hori dela eta, ingelesetik abiatzea erabaki genuen, bat-etortze handiagoa lortuko genuelakoan.

Sistemak baliabide bateko ingelesezko termino bat bestean aurkitzen duenean, egiaztatzen du bietan arlo-marka bera duen (arlo-mapaketan oinarrituta betiere); iturburu-baliabideko terminoa helburu-baliabidean ez badago, egiaztatzen du haren sinonimoa, iturburu-baliabidean egonez gero, helburuan dagoen. Horiek kontuan izanda, bi baliabideetako termino guztiak 2. taulan ikusten ditugun kodeekin etiketatu dira.

3.1.3. Erkaketaren ondorio batzuk

3. taulak erakusten ditu erkaketa automatikoaren emaitzak. 40.665 kasutan, bi baliabideen arteko bat-etortzeren bat gertatu da (erkatze-kodeetan 0 balioa ez dutenen batura).

3. TAULA

MCR eta ZTH baliabideen bat-etortzea termino-kopurutan

Kodea	Terminoak
1-0	112.307
0-1	10.436
1-1	8.678
1-2	9.039
1-3	5.911
1-4	6.660
2-1	7.187
3-1	1.636
4-1	1.554

Bistakoa da MCRk ZTHn ez dauden termino ugari dituela (112.307 ingelesez), eta ZTHk dituen baina MCRn ez dauden termino-kopurua txikiagoa dela (10.436 ingelesez). Datu hori ez da harritzekoa, ZTHn zientzia eta teknologiako arloko kontzeptuak bakarrik jasotzen direla kontuan izanda.

Kontuan izan beharrekoa da terminoen arteko mapaketa egiterakoan irizpide zorrotz bat erabili dugula («bat-etortze hertsia» izendatu duguna); hau da, terminoak berdin idatzia eta arlo berean sailkatua behar du bi baliabideetan. Haatik, maiz ingelesezko terminoen aldaerak topatu ditugu. Adibidez, ZTHn *videotape*, *vanilla plant*, *goldfish* eta *desertization* terminoak daude, eta MCRk, berriz, *video tape*, *vanilla* and *gold-fish*, edo *desertification*. Beraz, 3. taulako emaitza MCRren eta ZTHren arteko benetako bat-etortzearen lehen estimazio bat baino ez da.

3. taulak ere erakusten digu bat etorritako termino gehienek arlo-marka desberdinak dituztela (1-2, 1-4, 2-1 eta 4-1 kodedunak); guztira, 24.440 dira. Termino horiek guztiak eskuz landu behar dira, gerta baitaiteke kontzeptu bera ez adieraztea. Esaterako, *horn* terminoa ingelesez automatikoki 2-1 gisa etiketatu da, hau da, terminoa MCRen eta ZTHn dago, baina arlo-marka desberdina da bietan. MCRen *horn* terminoa *anatomy* arloari dagokio, eta ZTHn *zoology* arloari, eta bi terminoek kontzeptu bera adierazten dute ('zenbait animaliaren adar itxurako luzakina'). Bestalde, beste *horn* bat dago MCRen, *transport* arloa duena ('autoak soinua egiteko duen gailua'), eta termino honek ez du zerikusirik ZTHko *zoology* arloko terminoarekin.

Hurrengo atalean, erkaketa automatikoaren 0-1 terminoen eskuzko orrazketari buruz arituko gara.

3.2. *Erkaketa automatikoren eskuzko orrazketa*

WNTERM eratzeko, lehenik, 0-1 kasuak eskuz orraztu ditugu (MCRn ez dauden ZTHko terminoak). Orrazketaren helburua da kontzeptu berritzat ditugun 0-1 terminoei MCRko hiperonimo bana egokitzea, arlo-ontologiako kontzeptu-hierarkian kokatzeko. Hori egitean ikusi dugu erkaketa automatikoak 0-1 gisa markatu dizkigun termino batzuk (% 5) ez direla benetan berriak, dagoeneko MCRko kontzepturen batean badaudelako. Kasu horietan, MCRko terminoa:

— ZTHko terminoaren grafia-aldaera da (IDAZK).
— ZTHko terminoaren termino sinonimoa da (SIN).

Beraz, 0-1 terminoen lagin bat orraztu ondoren, 0-1 kasuen eskuzko orrazketaren helburua findu egin dugu: 0-1 terminoei hiperonimoa esleitu aurretik, ziurtatu behar izan dugu adierazten duten kontzeptua ez dagoela lehendik MCRn.

Ondorengo ataletan, erkaketa automatikoaren errore-kasuistika landuko dugu, eta arazo horiei nola egin diegun aurre azalduko.

3.2.1. ZTHko terminoa MCRko terminoaren grafia-aldaera denean

Termino beraren aldaerak aurkitu ditugu bi baliabideetan, 4. taulan ditugu adibide batzuk.

Marratxoek (*marvel of peru / marvel-of-peru*), pluralaren erabilerak (*enteroviruses / enterovirus*), hitzen arteko hutsuneek (*wild cat / wildcat*), letra larriaren erabilerak (*Bing Bang / bing bang*) eta antzeko alde tipografikoek eragin dute erkaketa automatikoak MCRko eta ZTHko terminoak desberdintzat jotzea.

Horrelakoei IDAZK marka gehitu diegu datu-basean, eta ZTHko terminoari MCRko zein synset dagokion zehaztu ere, WNTERMen sinonimo gisa ager daitezen. Guztira, horrelako 163 adibide landu ditugu.

4. TAULA

MCR eta ZTH baliabidetan sinonimoak diren baina era desberdinean idatzita dauden terminoen adibideak

ZTH	MCR
anesthesic	anesthetic
wild cat	wildcat
Bing Bang	bing bang
chabacite	chabazite
enteroviruses	enterovirus
acrocephalia	acrocephaly
marvel of peru	marvel-of-peru
St. John's wort	St John's wort

3.2.2. TERMINOAK MCRN SINONIMO BAT DUENEAN

Beste zenbaitetan, elkarren aldaera izan gabe ere sinonimoak diren terminoak ageri dira bi baliabideetan. Erkaketa automatikoak MCRn ZTHko termino berbera aurkitu ez duenez, MCR aberasteko kontzeptu berri gisa proposatzen du termino hori. Kasu horiek automatikoki ezin atzemanezkoak dira, eta eskuz aurkitu eta ebaluatu behar izan ditugu.

5. TAULA

MCR eta ZTH baliabideetan kontzeptu bera adierazten duten terminoen adibideak

ZTH	MCR
gas pipeline	gas line
stomach pain	stomachache
riparian wood	riparian forest
spermatic duct	ejaculatory duct
muscle gluteus	gluteus
radioactive fallout	radioactive dust
military aircraft	military plane
IOP	interoperability
rhinocerotid	rhinoceros
worm gearing	worm gear
biofuel	biomass
achalasia	cardiospasm
antitetanic	tetanus antitoxin
paracetamol	acetaminophen
neomenia	new moon
antibacterial agent	bacteriostat
phoenicopterid	flamingo

Horrek ez du adierazi nahi erkaketa automatikoak sinonimia tratatzen ez duenik. 3-1 eta 1-3 kodedun bat-etortze mota sinonimoen detekzioan oinarritua dago, baina, horretarako, ZTHn dauden sinonimoetako batek 1-1 kodea izan behar du aurretik MCRrekiko konparazioan. Atal honetan aztergai ditugun kasuak, berriz, sistema horren bidez atzeman ez diren sinonimia-kasuak dira (0-1 kodea dutenak); guztira, 357 dira eta SIN etiketaz markatu ditugu. 5. taulan, horietako kasuak ditugu.

3.2.3. KATEGORIA

Erkaketa automatikoan erabilitako irizpideak ez du kategoria kontuan hartzen; hori dela eta, zenbaitetan kategorien arteko interferentziak gertatu dira.

6. TAULA

Erkaketa automatikoan gertatutako kategoria-arazoen adibideak

ZTH	MCR proposamena	MCR termino egokia
cetacean	cetacean (adjektiboa)	cetacean (izena)
scarf	scarf (izena)	scarf (aditza)

Batzuetan, sistemak MCRko eta ZTHko terminoak parekatu arren (ikus 6. taulako *scarf*), bi baliabideetako terminoek ez dute adiera bera, bata aditza zelako (ZTHko aditzak ‘akoplatu’ adiera du) eta bestea izena (MCRkoa, ‘bufanda’ adierarekin). Erkaketa automatikoa egitean sistemak kategoriari buruzko informazioa jaso balu, horrelako akatsak eragotziko ziren, baina ZTHn, beste hainbat baliabide terminologikotan bezala, kategoria ez dago zehaztuta.

Bestalde, errore gehienek izen-adjektibo nahastearekin zerikusia dute, adjektibo ugari izen gisa ere erabil baitaitezke (*cetacean, geosynclinal*, *creeper, triacid*). ZTHren definizioan oinarritu gara kontzeptu horien kategoria erabakitzeko. Hala ere, komeni da gai hau sakonago aztertzea, baita MCRn adjektiboak nola antolatzen diren ere; ez dugu baztertzen horrelako terminoak izen eta adjektibo gisa sortzea WNTERMen, eta MCRko erlazio semantiko batekin lotzea (*XPOS_EQ_near_synonymy*).

3.2.1, 3.2.2 eta 3.2.3an azaldu ditugun erroreek erkaketa automatikoan erabilitako «bat-etortze hertsia» irizpidearekin harreman zuzena dutela kontuan izanik, 0-1 motakoak ez diren beste kodeak ere eskuz orraztea erabaki dugu (3-1 kasuak lehendabizi, eta dagoeneko hasiak gara 2-1 eta 4-1 kodedunekin), aurrerago egin ditzakegun erkaketa automatiko berriak hobetzeari begira.

Hurrengo ataletan, eskuzko orrazketaren ondoren WNTERM ezagutza-basea bera sortzeko erabili dugun metodologia deskribatuko dugu.

7. TAULA

MCR aberasten. MCRn ez dauden eta ZTHtik hartu ditugun terminoen kopurua, WNTERMeko arloen arabera antolatuta

Arloa	Ingelesezko terminoak	Euskarazko terminoak	Arloa	Ingelesezko terminoak	Euskarazko terminoak
Abel.	30	56	Ingur.	114	92
Aeron.	184	162	Itsas.	104	113
Albait.	15	13	Kim.	754	732
Anat.	96	97	Log.	17	18
Antr.	34	31	Mar.	3	1
Arg.	13	10	Mat.	479	503
Arkit.	32	27	Mater.	32	41
Arm.	22	20	Meatz.	7	4
Arr.	9	9	Med.	511	477
Astron.	219	203	Metal.	168	143
Astronaut.	157	129	Meteorol.	228	222
Autom.	118	100	Mikol.	36	23
Biokim.	283	244	Mikrob.	251	246
Biol.	243	239	Miner.	223	224
Bot.	283	269	Nekaz.	159	116
Ekol.	213	216	Orok.	61	70
Elektron.	465	431	Ozean.	98	92
Elik.	35	39	Paleont.	77	80
Eraik.	110	115	Psikiatr.	23	19
Estat.	177	188	Teknol.	524	443
Fis.	829	870	Teknol.Elektr.	413	385
Fisiol.	48	54	Teknol. Mek.	192	167
Genet.	60	62	Telekom.	594	563
Geogr.	22	21	Tren.	10	11
Geol.	755	789	Zool.	507	648
Inform.	399	396	**GUZTIRA**	**10.436**	**10.223**

3.3. *Kontzeptu-hautaketa*

Urrats honetan, WNTERM osatuko duten kontzeptuak hautatu ditugu. Ingelesa hartu dugu abiapuntu gisa (Euskal WordNet ingelesezko wordnetaren azpimultzo bat delako, eta ZTHren euskarazko eta ingelesezko edukiak baliokideak direlako). Oinarri gisa MCRn dagoeneko dauden kontzeptuak hartu ditugu (synsetak), eta horiek guztiak zuzenenean WNTERMen kopiatu. ZTHko kontzeptu bat MCRen ez dagoenean, WNTERMen txertatu da, eta eskuz esleitu zaio MCRko hiperonimo bat.

Hala, eskuz mapatutako arlo-marka bat duten MCRko kontzeptu guztiak, ZTHn egon ala ez, WNTERMera kopiatu dira. Bestela esanda, 1-0, 1-1, 1-3 eta 3-1 gisa etiketatu diren MCRko termino guztiak zuzenean txertatu ditugu

WNTERMen. 1-2 eta 1-4 gisa etiketatu diren terminoak oraingoz alde batera utzi dira, 3.1.3 atalean aipatu dugun bezala, termino horiek, MCRn eta ZTHn arlo-marka desberdinak dituztenez, aurretik eskuz orraztea komeni delako, kontzeptu bera adierazten dutela ziurtatzeko.

Lehenik, 0-1 gisa markatutako kontzeptuak orraztu ditugu, hau da, MCRn ez dauden baina ZTHn daudenak. Orrazketa honen helburuetako bat izan da ziurtatzea kontzeptu hori ez dagoela MCRn (ikus 3.2 ataleko adibideak). 4. taulako terminoekin gertatu den bezala, egiaztatzen badugu termino horri dagokion kontzeptua MCRn badagoela (aldaera edo sinonimo batez adierazia), kontzeptua kopiatu egiten dugu WNTERMen. Aldiz, kontzeptua berria bada, WNTERMen kopiatu eta MCRko hiperonimo bat egokitzen zaio.

8. TAULA

WNTERMek gaur egun duen kontzeptu-kopurua mapaketa-kodeen arabera

Kodeak	WNTERM kontzeptuak	Izenak	Adjektioak	Aditzak
1 - 0	30.984	26.653	3.054	1.009
1 - 1	6.197	5.810	248	70
1 - 3	3.453	3.219	126	110
3 - 1	901	878	22	1
0 - 1	8.688	8.286	363	13
GUZTIRA	50.223	44.846	3.813	1.203

8. taulak WNTERMen gehitu den kontzeptu-kopurua erakusten du. Kontzeptu gehienak nominalak dira (44.846). Adjektiboak (3.813) eta aditzak (1.203) ere gehitu dira.

3.4. *Termino-hautaketa*

Hirugarren urratsa terminoei dagokie. Ingelesezko eta euskarazko terminoak automatikoki gehitu dira MCRtik eta ZTHtik WNTERMen dagokien kontzeptuetara. Aurreko urratsean WNTERMen egongo diren kontzeptuak hautatu ditugunez, kontzeptu horiek bi baliabideetan dituzten terminoak bilatu eta WNTERNen kopiatu ditugu. 9. taulan eman ditugu MCRtik eta ZTHtik WNTERMen gehitu ditugun terminoen kopuruak.

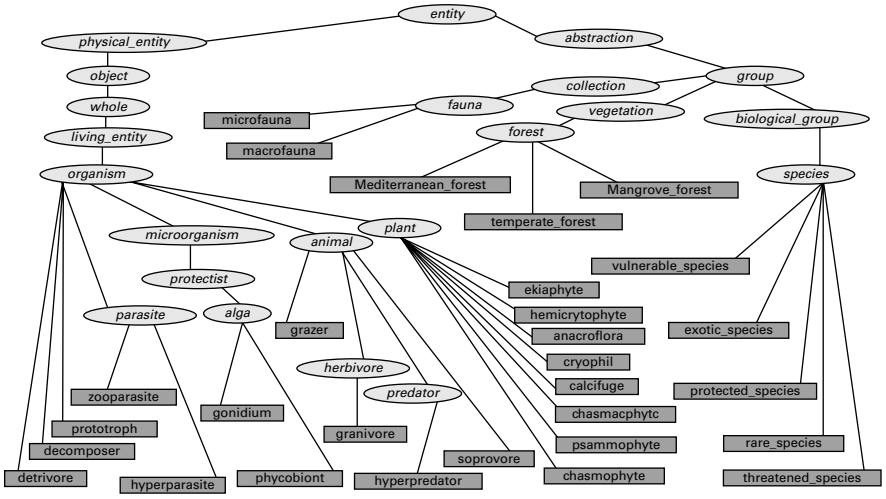
9. TAULA

WNTERMek gaur egun dituen terminoen kopurua mapaketa-kodeen arabera antolatua

Kodea	Ingelesezko terminoak	Euskarazko terminoak
1 - 0	112.307	22.702
0 - 1	11.676	11.142
1 - 1	6.871	8.618
1 - 3	5.911	4.430
3 - 1	1.134	1.620
GUZTIRA	137.899	48.512

3.5. *Espezialitate-ontologiaren egitura hierarkikoa*

Azkenik, hiperonimoen loturak erabili ditugu espezialitate-ontologiaren kontzeptuen hierarkia egituratzeko, hots, WordNeteko hiperonimo-egituran oinarritu gara espezialitate-ontologiako terminoak antolatzeko. MCRtik datozen kontzeptuen hiperonimoak kopiatu ditugu WNTERMen. ZTHtik hartu diren kontzeptu berriei, aldiz, eskuz esleitu diegu MCRko hiperonimoa, eta



2. IRUDIA

WNTERMeko azpihierarkia, Ekologia arloari dagokiona. ZTHtik txertatu diren kontzeptuak laukizuzen baten barruan, eta MCRko hiperonimoak borobil baten barruan

horiek ere WNTERMera eraman ditugu, aurreko atalean aipatu dugun bezala. Informazio gutxi eskaintzen duten hiperonimoak egon daitezke hierarkian; esate baterako, 2. irudiko *macrofauna* kontzeptuaren hiperonimoen artean, *collection* eta *group* hiperonimoak. Horrelako hiperonimo hutsalak automatikoki detektatu eta ezabatu daitezke Vossenek (2001) garatutako sistemarekin.

Gerora, espezialitate-adituek beharrezkoa iritziko baliote, hierarkia orrazteko eta hobetzeko aukera izango genuke.

Beraz, emaitza wordnet berri bat litzateke (zehazkiago, zientzia eta teknologiako arloen ontologia), beste wordnetetara lotzeko aukera eskaintzen duena.

4. **Ondorioak**

WNTERM zientzia eta teknologiaren ontologia da, ingelesezko eta euskarazko wordnetetatik eta ZTHtik abiatuta garatu dena. WordNet eleaniztunak errepresentatzeko MCRren arkitekturan oinarrituta dago. Artikulu honetan, WNTERM garatzeko metodologia eta WNTERM bera azaldu ditugu. WNTERMen arlo bakoitzeko kontzeptuen arteko erlazio semantikoak eta taxonomikoak ikusgai daude, eta kontzeptu horiek WordNet eta ZTHrekin lotuak daude. Hala, WNTERMek:

— espezialitate-terminologia jasotzen du
— termino guztien artean erlazio semantikoak definitzen ditu
— euskarazko eta ingelesezko wordnetetara lotua dago
— beste edozein wordnetera lotzeko aukera eskaintzen du
— ZTHra lotua dago

Bestalde, bi baliabide hauek kontuan hartuta, proiektu honetan beste helburu batzuk ere bete dira:

— Euskal WordNet terminologiarekin aberastu, eta aldi berean, espezializazio-arlo batzuetarantz hedatu dugu hiztegi terminologiko eta corpus berezituak erabiliz. Horretarako metodologia finkatu eta gauzatu dugu.
— ZTH hierarkikoki antolatu dugu MCRko arkitekturaren arabera.
— Proiektuan erabiltzen diren baliabide guztiak (MCR, ZTH eta WNTERM) lotu ditugu.

Gerora begira, WNTERMek MCRetik jaso duen hierarkia berrantolatu nahi genuke, eta hierarkia terminologikoa diseinatu arloko adituen irizpideen arabera.

Aipamenak

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini & P. Vossen. 2004. «The MEANING Multilingual Central Repository». *Proceedings of the 2nd Global WordNet Conference.* Brno, Txekiar Errepublika. 23-30

Borst, W.N. 1997. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. doktorego-tesia. Centre for Telematics and Information Technology, University of Tweety. Enschede, Herbehereak.

Fellbaum, C. 1998. *WordNet: An electronic Lexical Database*. Cambridge, Massachusetts: The MIT Press.

Gruber, T.R. & F. Olsen. 1994. «An ontology for Engineering Mathematics». In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning*. Bonn, Alemania.

Magnini, B. & G. Cavaglià. 2000. «Integrating subject field codes into WordNet». *Proceedings of LREC*. 1. Atenas, Grezia. 413-1.418

Navigli, R., P. Velardi. & A. Gangemi. 2003. «Ontology Learning and Its Application to Automated Terminology Translation». *IEEE Intelligent Systems*. 18-1, 22-31.

Pociello, E., A. Gurrutxaga, M. Susperregi, E. Agirre & G. Rigau. 2008. «WNTERM: Combining the Basque WordNet and a Terminological Dictionary». *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC)*. Marrakech, Maroko.

Rigau, G., E. Agirre & J. Atserias. 2003. «The MEANING project». *Proceedings of the XIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*. Alcalá de Henares, Madril. 307-308.

Roventini, A. & R. Marinelli. 2004. «Extending the Italian WordNet with the Specialized Language of the Maritime Domain». *Proceedings of the 2nd Global WordNet Conference.* Brno, Txekiar Errepublika. 92-99.

Sagri, M.T., D. Tiscornia & F. Bertagna. 2004. «Jur-WordNet». In P. Sojka et al (arg.), *Second International WordNet Conference*. Brno, Txekiar Errepublika. 305-310.

Stamou, S., A. Ntoulas, J. Hoppenbrouwers, M. Saiz-Noeda & D. Christodoulakis. 2002. «EUROTERM: Extending EWN using the expand and merge model». *Proceedings of the 1st Global WordNet Conference.* Mysore, India.

Vossen, P. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

Vossen, P. 2001. «Extending, Trimming and Fusing WordNet for Technical Documents». In *Proceedings of the NAACL Workshop on Extending Wordnet.* Pittsburgh, Pennsylvania.