Testuliburuetatik domeinu-modulu eleaniztunak eraikitzen

(Building multilingual domain modules from textbooks)

Angel Conde[1], *Ana Arruarte*[2], *Mikel Larrañaga*[2]*, *Jon A. Elorriaga*[2], *Ruben Urizar*[3]

[1] Softwarea produzitzeko teknologiak, IKERLAN, J.M. Arizmendiarrieta pasealekua, 2; 20500 Arrasate

[2] Lengoaia eta Sistema Informatikoak Saila, Informatika Fakultatea, UPV/EHU, 649 Postakutxa, 20080 Donostia

[3] Hizkuntzaren eta Literaturaren Didaktika Saila, Hezkuntza, Filosofia eta Antropologia Fakultatea, UPV/EHU, Oñati plaza 3, 20018 Donostia

* mikel.larranaga@ehu.eus

DOI: 10.1387/ekaia.16284

Jasoa: 2016-04-28 Onartua: 2016-06-29

Laburpena: Lan honetan LiDom Builder tresnaren analisia, diseinua eta ebaluazioa aurkezten dira. Domeinu Modulu Eleaniztunak testuliburu elektronikoetatik era automatikoan erauztea ahalbidetzen du LiDom Builderek. Ezagutza eskuratzeko, Hizkuntzaren Prozesamendurako eta Ikaste Automatikorako teknikekin batera, zenbait baliabide eleaniztun erabiltzen ditu, besteak beste, Wikipedia eta WordNet.

Hitz gakoak: Teknologian oinarritutako hezkuntzarako tresnak, domeinu-modulu eleaniztuna, sorkuntza automatikoa.

Abstract: This paper presents the analysis, design and validation of LiDom Builder, a framework for the automatic generation of Multilingual Domain Modules from electronic documents. LiDom Builder uses Natural Language Processing and Machine Learning techniques, along with multilingual resources such as Wikipedia and WordNet to fulfil its work.

Keywords: Technology Supported Learning Systems, Multilingual Domain Modules, Automatic Acquisition.

1. **SARRERA**

Hizkuntza da jakintza transmititzeko darabilgun komunikazio-metodo ohikoena eta testuinguru elebidunak eta eleaniztunak errealitatea dira Guztiontzako Hezkuntza ardatz duen gizartean, non berdintasuna, gizarte-kohesioa eta herritartasun aktiboa sustatzen baitira. Errealitate konplexu horretara egokitzea eta kalitatezko hezkuntza eskaintzea da, orduan, hezkuntza-sistemaren erronka garrantzitsuenetakoa [1].

Teknologikoki garatutako gizarteetan, hezkuntza elebidunak eta eleaniztunak eragin zuzena izan du, oro har, Informazio eta Komunikazio Teknologietan (IKT) eta, bereziki, Teknologian Oinarritutako Hezkuntzarako Tresnetan. Teknologian Oinarritutako Hezkuntzarako Tresnak —esaterako, Tutore Adimendunak, Moodle[1] edo Blackboard[2] moduko ikasketa kudeatzeko sistemak— eta Coursera[3] edo edX[4] gisako lineako ikastaro ireki masiboak lantzeko plataformak ezinbesteko bihurtu dira hainbat hezkuntza-erakundetan [2].

Teknologian Oinarritutako Hezkuntzarako Tresnek domeinu-modulua —hots, ikasi beharreko domeinuaren adierazpen pedagogikoa— behar dute. Domeinu-modulua da Teknologian Oinarritutako Hezkuntzarako edozein tresnaren muina, hark adierazten baitu ikasleek ikasi beharreko ezagutza guztia [3].

Domeinu-modulua sortzea ez da lan arina, ordea. Ikasi beharreko gaiak adierazteaz gain, identifikatu behar dira horien arteko erlazio pedagogikoak, ikasketa-saioak nola planifikatu zehazten dutenak, eta ikasteko erabiliko diren hezkuntzarako baliabideak (definizioak, adibideak, ariketak eta abar).

Azkeneko urte hauetan, berrerabilpena bultzatzeko saiakerak egin dira hezkuntzarako informatikan. Batetik, hezkuntza-baliabideak deskribatzeko estandarrak sortu dira [4]. Bestetik, hezkuntza-baliabide berrerabilgarriak —Ikaste Objektuak— garatu dira. Gero eta ohikoagoak dira ARIADNE [5, 6], Merlot [7] edo GLOBE[5] gisako biltegi-sareak, ikastaro berriak sortzeko behar diren baliabideak eskaintzen dituztenak.

Domeinu-modulua sortzea zaila bada, are zailagoa da domeinu-modulua era automatikoan edo erdiautomatikoan erauztea, dokumentu elektronikoetatik edo corpusetatik abiatuta. Egiteko horren inguruan argitaratu diren lanak, gainera, domeinu-modulu elebakarretara mugatuta daude [8, 9, 10, 11, 12, 13]. Euskararen kasuan, *DOM-Sortze* inguruneak ahalbidetzen du

[1] http://moodle.org
[2] http://www.blackboard.com
[3] https://es.coursera.org/
[4] https://www.edx.org/
[5] http://www.globe-info.org/

domeinu-moduluak sortzea. Euskaraz idatzitako testuliburu elektronikoetatik domeinu-modulua era erdiautomatikoan erauzteko tresna da *DOM-Sortze* [14, 15, 16], eta domeinuarekiko independentea da. Domeinu-modulu elebakarretik domeinu-modulu eleaniztunerako bidean, *LiDom Builder* tresna *DOM-Sortze* ingurunearen bilakaera dela esan genezake.

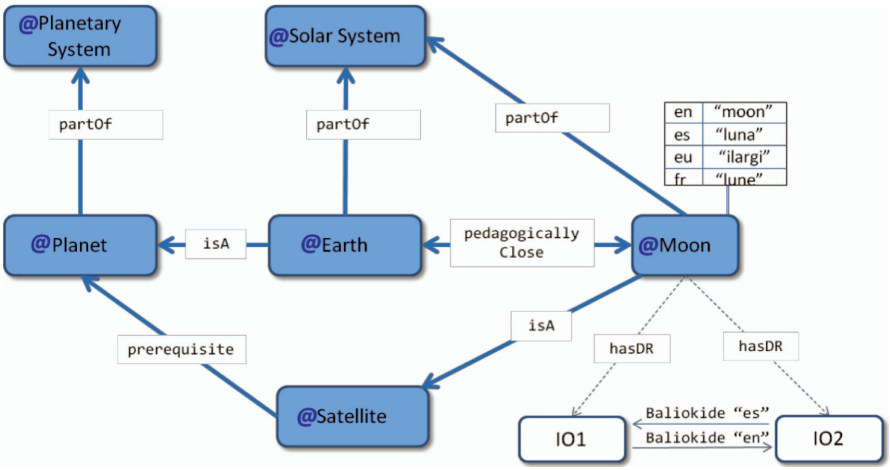
Lan honetan, *LiDom Builder* aurkezten da, liburu elektronikoetatik domeinu-modulu eleaniztunak modu automatikoan sortzeko tresna. 2. atalean, domeinua ikuspegi eleaniztun batetik adieraztea ahalbidetzen duen mekanismoa aurkezten da. 3. atalean, berriz, *LiDom Builder*en ezaugarri nagusiak azaltzen dira. Testuliburuetatik domeinu-modulu eleaniztunak eraikitzeko, *LiDom Builder* hiru modulu nagusitan oinarritzen da: *LiTeWi*n*, LiReWi*n eta *LiLoWi*n. Modulu bakoitzaren ezaugarriak eta balidatzeko egin diren esperimentuak aipatzen dira, hurrenez hurren, 4., 5. eta 6. ataletan. Bukatzeko, ondorioak eta etorkizuneko lerroak aurkezten dira 7. atalean.

2. **DOMEINU MODULU ELEANIZTUNAK ADIERAZTEA**

*LiDom Builder*en testuinguruan, domeinu-modulu eleaniztunak bi mailako ezagutza jasotzen du: batetik, Ikaste Domeinuaren Ontologia (IDO), non hizkuntza ezberdinetan etiketatutako gaiak eta horien arteko erlazio pedagogikoak jasotzen baitira eta, bestetik, Ikaste Objektuak (IO), hau da, hezkuntzarako erabiliko diren edukiak, metadatuekin etiketatuak (definizioak, adibideak, ariketak eta abar). *LiDom Builder*ek aukera ematen du onartutako hizkuntza guztietan domeinuaren gaiak adierazteko. Gai bakoitza lotuta dago hizkuntza bakoitzean dagokion etiketa baliokidearekin. Gainera, IOak deskribatzeko, metadatu aberastuak erabiltzen dira, hizkuntza desberdinetan parekideak diren baliabide didaktikoak lotzeko.

Domeinu-modulu eleaniztun zati baten adibidea erakusten da 1. Irudian. Bertan adierazten dira domeinuko gai nagusiak —*Planetary System, Solar System, Planet, Earth, Moon, Satellite*— eta horien itzulpenak —adibidean, *Moon*en itzulpenak: euskaraz, *Ilargi;* gaztelaniaz, *Luna,* eta frantsesez, *Lune*). Horrez gain, gai nagusien arteko erlazio pedagogikoak ere zehazten dira. Lau erlazio pedagogikorekin lan egin dugu: *isA*rekin eta *partOf*ekin, egitura adierazteko; *prerequisite*ekin, ordena adierazteko, eta *pedagogicallyClose*rekin, gertutasun pedagogikoa adierazteko. *Earth isA Planet* erlazioak adierazten du *Earth* gaia *Planet*en mota zehatz bat dela. *Planet partOf Planetary System* erlazioak, ordea, *Planet* gaia *Planetary System*aren zati bat dela erakusten du, hots, *Planetary System* landutzat emateko, *Planet* ikasi behar diren gaietako bat dela. Halaber, *Satellite prerequisite Planet* erlazioak adierazten du, *Satellite* ikasten hasi aurretik, ikasleak *Planet* dagoeneko landuta izan behar duela, eta *Earth pedagogicallyClose Moon* erlazioak, berriz, bi gaiak oso gertu daudela adierazten du

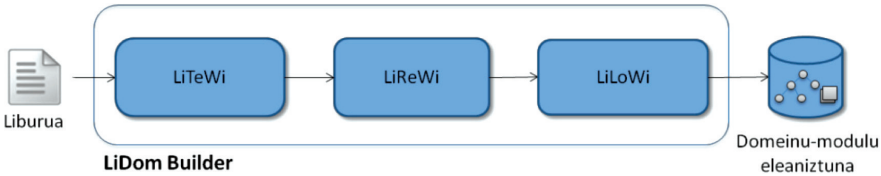
eta litekeena dela, adibidez, biak batera lantzea. Gai eta erlazio pedagogikoez gain, IO eleaniztunak ere jaso behar dira domeinu-modulu eleaniztunean. IO bakoitzak lotuta egon behar du beste hizkuntzetan baliokideak diren IOei. Helburu hori betetzeko, IO bakoitzaren metadatuak aberastu dira, baliokideen arteko loturak deskribatzeko.



1. irudia. Domeinu-modulu eleaniztun zati baten adibidea.

3. LIDOM BUILDER: DOMEINU MODULU ELEANIZTUNAK ERAIKITZEKO TRESNA

Testuliburuetatik domeinu-modulu eleaniztunak eraikitzeko, *LiDom Builder* hiru modulu nagusitan oinarritzen da (2. irudia). LiTeWi eta LiReWi moduluak IDO eleaniztuna eraikitzeaz arduratuko dira, eta LiLoWi, aldiz, IO eleaniztunak sortzeaz.



2. irudia. Domeinu-modulu eleaniztunaren sorkuntza-prozesua.

LiDom Builder tresnan, hasiera batean, hizkuntza jakin batean idatzitako dokumentu batetik erauziko da domeinu-modulua, eta baliabide eleaniztunak erabiliko dira, gerora, bai gaiak bai IOak beste hizkuntzetan ere

lortzeko. Baliabide eleaniztunei dagokienez, Wikipediatik eta WordNetetik eratorritako zenbait ezagutza-base erabiliko dira.

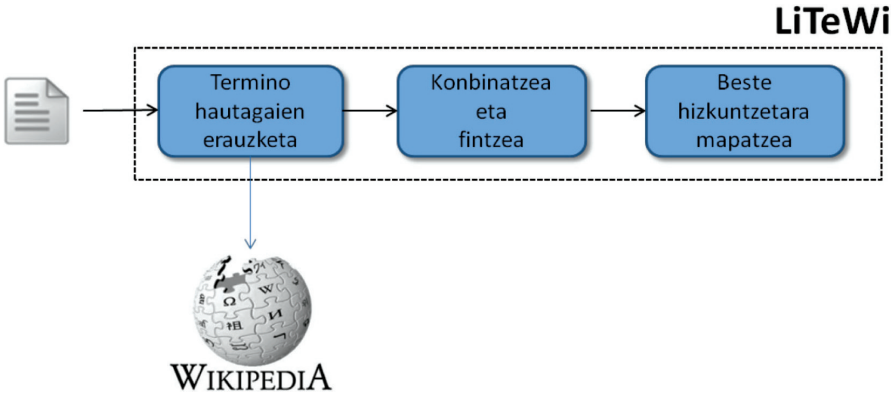
LiDom Builder balidatzeari dagokionez, modulu bakoitza bere aldetik testatu eta ebaluatu da, bai urre-patroiaren eredua bai aditu-ebaluazioa baliatuz. Urre-patroiari dagokionez, hiru neurri erabili dira: *estaldura*, hauteman beharreko elementu guztietatik, aurkitu direnen ehunekoa; *doitasuna*, aurkitutako elementuen artean, zuzenak direnen ehunekoa, eta *F1 neurria*, estaldura- eta doitasun-neurrien batezbesteko harmonikoa. Gainera, Wikipedia eta WordNet ezagutza-baseen integrazioak IOen erauzketari ekarri dion hobekuntza ere ebaluatu da. *LiDom Builder*en eraginkortasuna bermatzeko, beharrezkoa izan da doitze-prozesu bat egitea, sistemaren atalase eta faktore egokienak ezartzeko.

Lan honetan, ingelesez idatzitako liburuek osatu dute informazio-iturri nagusia bai doitze-prozesuetan bai ebaluazio-prozesuetan. Zehazki, honako testuliburu hauek erabili dira, testu lauzko formatuan: *Principles of Object Oriented Programming* [17] —30.000 hitzez osatua—, *Introduction to Astronomy* [18] —110.000 hitzez osatua— eta *Introduction to Molecular Biology* [19] —70.000 hitzekoa—.

Jarraian, aipatutako modulu bakoitza xehetasun handiagoz azalduko da.

4. LITEWI: GAI ERAUZKETA ELEANIZTUNA IKASTE DOMEINUAREN ONTOLOGIETARAKO

LiTeWi [20] moduluak, edozein ikaste-domeinutako testuliburu batetik abiatuta, identifikatzen ditu Ikaste Domeinuaren Ontologia (IDO) bati dagozkion hainbat gai, zenbait hizkuntzatan. Prozesua hiru urratsetan egiten da (ikus 3. irudia):



3. irudia. LiTeWi-ren prozesua.

1. **Gai hautagaiak erauztea**: *LiTeWi*k IDO bati dagozkion gai hautagaien zerrenda identifikatzen du, horretarako, gainbegiratu gabeko zenbait datu-erauzle paraleloan exekutatuz. Datu-erauzketarako tekniken artean daude, besteak beste, TF-IDF [21], Cvalue [22], KP-Miner [23] eta Shallow Parsing Grammar [24].
2. **Konbinatzea eta fintzea**: Urrats honetan, lortutako gaiak konbinatu eta fintzen dira eta azkeneko termino-zerrenda osatzen da. Horretarako, lehenik, (1) gai-zerrenda normalizatu eta iragazi egiten da, errefus-hitzen zerrenda bat (*stopwords*) aplikatuz; ondoren, (2) zerrendako gaiak Wikipediako artikuluekin mapatzen dira, *Wikimine*ren [25] konfigurazio jakin bat erabiliz; gero, (3) gai bat artikulu batekin baino gehiagorekin mapatu bada, Milnek eta Wittenek proposatutako Global Disambiguation [26] prozesua erabiltzen da, esanahia desanbiguatzeko eta artikulu bakar batekin mapatzeko, eta, bukatzeko, (4) Wikiminer Comparing Service [18] erabiliz, gai bakoitzak domeinuarekin daukan gertutasuna kalkulatzen da eta zerrendatik ezabatzen dira aldez aurretik zehaztutako atalase bat gainditzen ez duten horiek.
3. **Mapatzea**: Wikipedia baliatuz, azken zerrendako gaiak mapatzen dira beste hizkuntzetako gai baliokideekin. Horretarako, Wikipedian hizkuntza desberdinetako artikuluek beren artean duten baliokidetasuna erabiltzen du *LiTeWi*k.

Aipatu bezala, *LiTeWi* testatu eta ebaluatu egin da, bai urre-patroiaren eredua bai aditu-ebaluazioa baliatuz, bi domeinuko liburutan: astronomia eta biologia. Liburu bakoitzaren glosategia urre-patroi gisa erabili da. Astronomiako liburuaren glosategiak 378 sarrera ditu, eta horietatik 322 gaik daukate sarrera ingeleseko Wikipedian. Biologiako liburuaren kasuan, glosategiak 274 sarrera ditu, eta 220 gaik daukate sarrera Wikipedian. Horrez gain, aditu-talde batek aztertu ditu automatikoki erauzitako gaiak, domeinuarekin erlaziorik ote duten ala ez zehazteko. Programazio-domeinuko liburua, aldiz, doitze-prozesuetarako baino ez da erabili.

Gai-erauzketari dagokionez, astronomiako domeinuan, 1.545 gaik osatu dute zerrenda. Horietatik 275 glosategian jasotakoak dira, eta 1.217 gaik lotura dute domeinuarekin. Biologiaren domeinuan, aldiz, zerrenda 635 gaik osatu dute, glosategian agertzen diren 165ek eta domeinuari dagozkion 455ek. 1. taulak jasotzen ditu liburu bakoitzean lortutako emaitzak, konbinatze- eta fintze-urratsen ostean.

Eleaniztasunaren aldetik, neurtu da domeinuari lotutako gaietatik zenbat mapatu diren, Wikipediaren medioz, beste hizkuntzetara, hau da, domeinuari lotuta dauden gaietatik zenbatek daukaten itzulpenena beste hizkuntzetan. 2. taulak jasotzen ditu urrats honen emaitzak. Astronomia-liburuaren kasuan, erauzitako 1.545 gaietatik, 1.236 gairentzat (% 80) lortu da

1. taula. Konbinatze- eta fintze-urratsen osteko emaitzak.

	Urre-patroia			Aditu-ebaluazioa
	Estaldura (%)	Doitasuna (%)	F1 neurria (%)	Zuzentasuna (%)
Astronomia	17,96	72,55	28,79	78,77
Biologia	27,09	57,29	36,77	71,65

gaztelaniazko itzulpena; 1.297 gairentzat (% 84), frantsesezko itzulpena, eta 602 gairentzat (% 32), euskarazkoa. Biologiaren kasuan, 635 gaietatik, 476 (% 75) eskuratu dira gaztelaniaz; 469 (% 74), frantsesez, eta 203 (% 32), euskaraz.

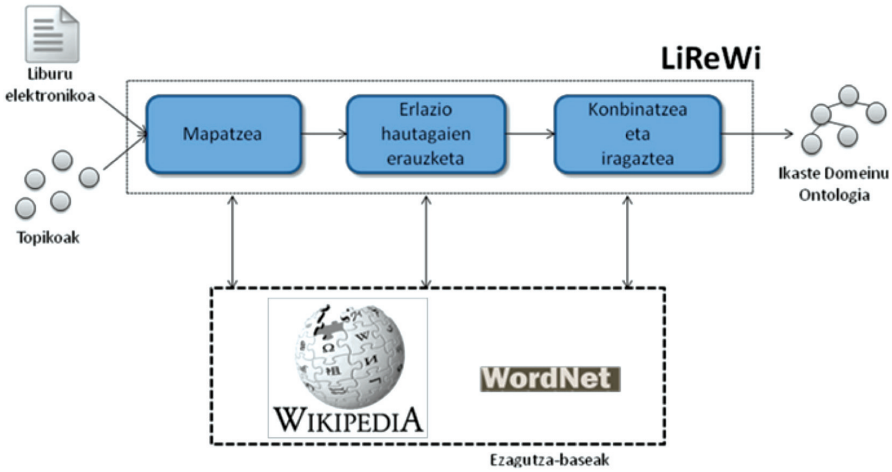
2. taula. Hizkuntzen arteko gai-mapatzea.

	Hizkuntzen arteko mapatzea			
	Ingelesa	Gaztelania	Frantsesa	Euskara
Astronomia	1.545	1.236	1.297	602
Biologia	635	476	469	203

5. LIREWI: IKASTE DOMEINUAREN ONTOLOGIETARAKO ERLAZIOEN ERAUZKETA

LiReWi moduluak erlazio pedagogikoez aberastuko du IDO, beti ere, testuliburua abiapuntu gisa erabilita. *LiReWi*k, zenbait teknika eta ezagutza-base konbinatuz, lau motako erlazio pedagogikoak erauziko ditu: *isA, partOf, prerequisite* eta *pedagogicallyClose*. *LiReWi*k ere hiru urratsetan lortuko ditu erlazioak (ikus 4. irudia):

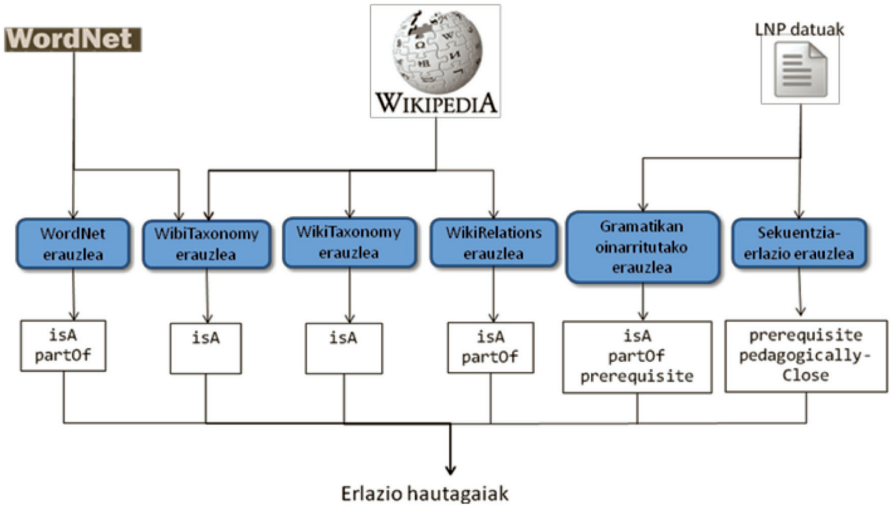
1. **Mapatzea**: Hasteko, *LiReWi*k ontologiako gaiak mapatuko ditu erlazioak erauzteko erabiliko diren ezagutza-baseetara. Ezagutza-baseen artean daude Wikipedia, WordNet [27], WikiTaxonomy [28], WibiTaxonomy [29] eta WikiRelations [30]. Aurreko urratsean, *LiTeWi*k mapatu ditu gaiak Wikipediara, beraz, ez da beharrezkoa mapatzea berriro egitea ez ezagutza-base horretara ez bertatik eratorritako ezagutza-baseetara. Nahikoa izango da, beraz, *WordNet*eko sarrerekin mapatzea. Horretarako, *LiTeWi*k Navigli eta Ponzettoren [31], eta Fernandoren [32] lanetan proposatutako mapatzeak baliatzen ditu. Bi hurbilpen horiekin emaitza desberdina lortzen denean, anbiguotasun-arazo bat dago, eta hori ebazteko, Page-Rank desanbiguazio-prozesua egiten da, UKB [33] erabiliz.



4. irudia. LiReWi-ren prozesua.

2. **Erlazio hautagaien erauzketa**: Gero, erlazio hautagaiak erauzteko, konkurrenteki exekutatuko ditu erlazio-erauzle batzuk, bakoitza teknika desberdin batean oinarritzen dena (ikus 5. irudia). Tekniken artean daude, besteak beste, taxonomiatan oinarritutakoak, gramatikatan oinarritutakoak eta agerralditan oinarritutakoak. Erabiltzen diren taxonomiak eta tresnak, zehazki, honako hauek dira: (i) WordNet, (ii) WibiTaxonomy, (iii) WikiTaxonomy, (iv) WikiRelations, (v) Larrañagak eta bestek [15] euskararako garatutako gramatikaren parekoa ingelesarentzat eta, bukatzeko, (vi) garatu berria den sekuentzia-erlazioen erauzle bat [24].
3. **Konbinatzea eta iragaztea**: Azken urratsean, lortutako emaitza guztiak konbinatu eta iragaziko ditu *LiReWi*k, eta erlazio pedagogikoen azken multzoa lortuko du. Horretarako, lehenik, erauzle batek baino gehiagok erauzitako erlazioen konfiantza konbinatu eta egokitzen du; zenbat eta erauzle gehiagok erlazio bat erauzi, orduan eta konfiantza-maila handiagoa izango du erlazio horrek. Gatazkak ebazteari ekiten dio orduan, sendotasun eza eta kontraesanak saihesteko asmoz [24]. Bukatzeko, aldez aurretik ezarritako gutxieneko ziurtasun-maila gainditzen ez duten erlazioak ezabatuko dira.

*LiReWi*ren kasuan, eta ebaluazioari dagokionez, berriro ere bi teknika erabili dira: urre-patroiaren eredua eta aditu-ebaluazioa, eta bi liburu: programazio-domeinukoa, doitze-prozesuetarako, eta astronomia-domeinukoa, aldiz, ebaluaziorako.



5. irudia. LiReWi-n erabiltzen diren erlazio-erauzleak.

Sarrera gisa, gai-multzo bat eskatzen du *LiReWi*k. Gai horiek *LiTeWi*k erauzitako gaien azpimultzo batek osatu ditu. Zehazki, glosategian jasotako 199 gai erabili dira, CVALUE neurriari erreparatuz gero, domeinuarekin erlazio handiena zutenak.

Bestalde, aditu-talde batek ezarri ditu urre-patroi gisa erabiliko diren gaien arteko erlazioak (*isA, partOf, prerequisite* eta *pedagogicallyClose*). Guztira, 174 erlazio identifikatu ditu: 80 *isA*, 69 *partOf*, 10 *prerequisite* eta 15 *pedagogicallyClose*.

*LiReWi*k 266 erlazio erauzi ditu, guztira. Urre-patroiari erreparatuz gero, estaldura % 36,21 izan da, eta doitasuna, aldiz, %50,57.

Bestalde, aditu-ebaluazio bat egin da, automatikoki erauzitako erlazioetatik zuzenak zenbat diren neurtzeko. Automatikoki erauzitako 266 erlazioetatik, 117 (% 43,98) izan dira zuzenak. Adituen arteko Fleiss's kappa [34] koefizienteak erakutsi du, 0,974 pisuarekin, adituen arteko adostasun-maila ia erabatekoa izan dela, erlazioak ebaluatzeko orduan [35]. Esperimentu honetan, aditu guztiek aho batez adostutako erlazioak baino ez dira zuzentzat hartu.

3. taulak jasotzen ditu astronomiako domeinuan lortutako emaitza orokorrak, konbinatze- eta fintze-urratsen ostean, eta *4*. taulak erakusten ditu erlazio-mota bakoitzari dagozkion emaitzak.

3. taula. LiReWi-ren emaitza orokorra astronomia-domeinuan.

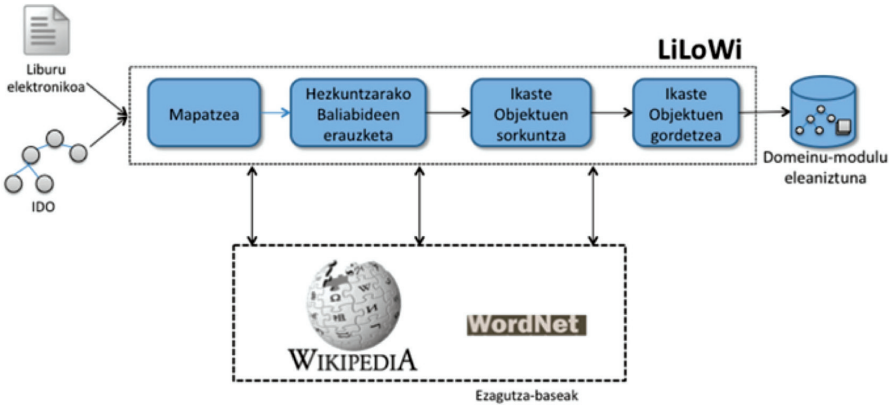
	Urre-patroia			Aditu-ebaluazioa
	Estaldura (%)	Doitasuna (%)	F1 neurria (%)	Zuzentasuna (%)
LiReWi	36,21	50,57	42,2	43,98

4. taula. LiReWi-ren erlaziokako emaitza.

		Urre-patroia			Aditu-ebaluazioa
	Erlazioak	Estaldura (%)	Doitasuna (%)	F1 neurria (%)	Zuzentasuna (%)
isA	213	30,38	76,25	43,44	40,84
partOf	37	51,34	27,54	35,85	51,36
prerequisite	10	30,00	30,00	30,00	80,00
pedagogicallyClose	6	50,00	33,00	39,75	50,00

6. LILOWI: IKASTE OBJEKTU ELEANIZTUNEN ERAUZKETA

LiLoWi moduluak IOak —batzuk, eleaniztunak— lau urrats nagusitan erauziko ditu, abiapuntuko testuliburutik ez ezik Wikipedia edo WordNet moduko ezagutza-baseetatik ere (ikus 6. irudia).



6. irudia. LiLoWi-ren prozesua.

1. **Mapatzea:** Hasteko, *LiLoWi*k ontologiako gaiak mapatzen ditu Wikipedia eta WordNet ezagutza-baseetara. *LiDom Builder*en aurreko

urratsak egikaritu badira, ontologiako gaiak dagoeneko ezagutza-base horietara mapatuta egongo dira.

2. **Hezkuntza Baliabideen erauzketa:** Gero, Hezkuntza Baliabideak erauzten ditu konkurrenteki, exekuta daitezkeen bi prozesu egikarituz. (1) Lehenik, gramatikan oinarritutako prozesuak [36] erregelak erabiltzen ditu, hizkuntza jakin batean zenbait baliabide-mota desberdin erauzteko, adibidez, definizioak, ariketak eta abar. 5. taulak patroi bat erakusten du ingelesezko adibideak aurkitzeko, eta haren patroi baliokidea, euskararako. (2) Bigarren, baliabide berriak erauzten dira ezagutza-base desberdinetatik. Wikipediatik, erreferentzia bibliografikoak eta definizio eleaniztunak eskuratzen dira; WordNet-etik, aldiz, definizio berriak.
3. **Ikaste Objektuen sorkuntza:** Urrats honetan, IOak sortzen dira, Hezkuntza Baliabideetatik abiatuta. Batetik, metadatuak sortzen dira modu automatikoan, besteak beste, SamgI [37] etiketatzailea erabiliz. Bestetik, lotura ezartzen da hizkuntza desberdinetan parekideak diren IOen artean, gerora, domeinu eleaniztuna kudeatu ahal izateko.
4. **Ikaste Objektuak gordetzea:** Bukatzeko, sortutako IO guztiak Ikaste Objektuen Biltegian gordetzen dira. Erabilitako biltegia ARIADNEren teknologian [6] oinarritzen da.

5. taula. Adibideak aurkitzeko patroi bat.

	Euskara	Ingelesa
Patroia	**Adibidez**, @*topic*	**for instance**, @*topic*
Adibidea	Uretan, **adibidez**, *hidrogeno-* eta *oxigeno-atomoak* daude.	**For instance**, there are *hydrogen* and *oxygen atoms* in water.

*LiLoWi*ren ebaluazioari dagokionez, bai urre-patroiaren eredua bai aditu-ebaluazioa erabili dira, bi alderdi nagusi ebaluatzeko. Batetik, moduluak ingeleseko IOak identifikatzeko duen ahalmena neurtzeko. Bestetik, Wikipedia eta WordNet gisako ezagutza-baseen integrazioak IO berriak erauzteko —besteak beste, IO eleaniztunak— dakarren onura.

Gramatikaren ebaluazioa programazioko liburuaren gainean egin zen, eta aurretik argitaratutako lan batean deskribatzen da xehetasunez [36]. Ebaluazioa egiteko, aditu-talde batek liburua aztertu eta bertan zeuden hezkuntza-baliabideak identifikatu zituen eta urre-patroia osatu zuen. Gerora, automatikoki erauzitako IOak ebaluatu ziren, bai urre-patroiari erreparatuz bai egokitasuna aztertuz (aditu-ebaluazioa).

Urre-patroiari erreparatuz, % 75,93 estaldura lortu zen, guztira (% 62,86, definizioentzat eta % 100, ariketentzat). Doitasuna neurtzeko,

adituek erauzitako IO bakoitza banan-banan aztertu zen, egokia zen ala ez erabakitzeko. Doitasuna, guztira, % 86,79 izan zen (% 86,42, definizioentzat eta % 100, ariketentzat).

Bestalde, Wikipediaren eta WordNeten integrazioak IOak erauztean dakarren onura ebaluatzeko asmoz, bigarren esperimentu bat egin zen programazioko liburuarekin, aurrekoan erabili zen ontologia berbera berrerabiliz. Ontologia 82 gaik osatzen zuten. Oraingo honetan, bi alderdi ebaluatu nahi ziren: batetik, ezagutza-base horien integrazioak nola hobetzen zuen IO-gaien bikoteen estaldura —hau da, zenbat gairentzat aurkitu den, gutxienez, IO bat— eta, bestetik, zer eragin zuen integrazioak IO eleaniztunen erauzketan.

IO-gaien estaldurari erreparatuz gero, aurreneko esperimentuan 21 gairentzat lortu ziren IOak. Definizioen kasuan, 19 gaik baino ez zeukaten IO mota hori esleiturik (% 19,51). Ezagutza-baseen integrazioaren ostean, 46ra igo zen estalitako gaien kopurua (% 56,10). Integrazioak, gainera, *erreferentziak* IO mota berria erauztea ere ahalbidetu zuen 12 gairentzat. IO eleaniztunei dagokienez, definizioak aurkitu ziren hiru hizkuntzatan: gaztelaniaz, 36 gairentzat (% 43,90); euskaraz, 9 gairentzat (% 10,97), eta frantsesez, 36rentzat (% 43,90).

6. taulak laburtzen ditu integrazioaren ondoren lorturiko emaitzak.

6. taula. IO-gai estaldura.

	Definizioak				Erreferentziak
	Ingelesa	Gaztelania	Euskara	Frantsesa	
Kopurua	46	36	9	36	12
Estaldura (%)	56,10	43,90	10,97	43,90	14,63

Argi utzi nahi da ez dela, inondik inora, lan honen helburua Wikipedia gisako ezagutza-baseen edukien egokitasun pedagogikoa aztertzea. IO erauzketari dagokionez, lan honetan, Wikipedia ez da baliabide eleaniztun bat besterik, IO eleaniztunak erauzteko, ahalmen izugarria eskaintzen duena.

7. ONDORIOAK ETA ETORKIZUNERAKO LANA

Artikulu honetan, *LiDom Builder* aurkeztu da, testuliburu elektronikoetatik era automatikoan domeinu-modulu eleaniztunak erauzten dituen

tresna. Ezagutza eskuratzeko, *LiDom Builder*ek, Hizkuntzaren Prozesamendurako eta Ikaste Automatikorako teknikekin batera, zenbait bialiabide eleaniztun erabiltzen ditu, besteak beste, Wikipedia eta WordNet. Laburpen gisa, esan genezake lau direla *LiDomBuilder*ek domeinu-modulu eleaniztunaren arloari egin dizkion ekarpen nagusiak:

— Mekanismo egokia definitzea domeinu-modulu eleaniztunak adierazteko.
— *LiTeWi*ren garapena. Modulu honek ahalbidetzen du testuliburuetatik terminologia eleaniztuna erauztea, hezkuntzarako ontologiak sortzeko balio duena.
— *LiReWi*ren garapena. Modulu honek testuliburuetatik erlazio pedagogikoak erauzten ditu, hezkuntzarako ontologiak sortzeko balio dutenak.
— *LiLoWi*ren garapena. Wikipedia eta WordNet ezagutza-baseak testuliburuarekin batera erabilita, modulu honek IO eleaniztunak erauzten ditu.

Probetarako erabili diren baliabideek, ingelesezko hainbat dokumentuk eta zenbait baliabide eleaniztunek —besteak beste, Wikipediak eta WordNetek— ontologia eleaniztunak eta IO eleaniztunak identifikatzeko balio izan dute, hots, Ikaste Domeinu Eleaniztunak sortzeko. Bibliografian aztertutako lanen artean antzeko asmoarekin garatutako sistemarik ez aurkitzeak galarazi du lortutako emaitzen alderaketa zuzena egitea. *DOM-Sortze* [15,16] da lan honetatik gertuen dagoen lana, baina, aipatu bezala, domeinu-modulu elebakarretara mugatzen da. Edozein kasutan, esan behar da *DOM-Sortze*rekin alderaketa egin ahal izan denean, egin dugula, eta, kasu guztietan, emaitzak hobetu egin direla *LiDom Builder*ekin [24]. Horren arrazoi nagusia argia da: *DOM-Sortze* garatzeko orduan erabilitako teknikaz gain, teknika eta metodo berri ugari integratzen ditu *LiDom Builder*ek, bai gai-erauzketari erreparatuta, bai erlazio pedagogikoei eta IO erauzketari dagokienez ere. Ezin ahaztu, gainera, hasierako dokumentu edo testuliburuez gain, *LiDom Builder*ek baliabide eleaniztun berriak ere integratu dituela bere baitan, emaitzak aberastu eta hobetzearren.

Etorkizunean, *LiDom Builder* aberasteko asmoa dago hizkuntza berriak txertatzen laguntzeko tresna batekin.

8. **ESKER ONA**

EHUko UFI11/45 Prestakuntza eta ikerkuntza Unitatearen eta GIU16/20 ikerkuntza-taldeen babesa jaso du lan honek.

9. **BIBLIOGRAFIA**

[1] UNESCO, 2003, *Education in a multilingual world*.

[2] PARSAD, B., & LEWIS, L., 2008, *Distance Education at Degree-Granting Postsecondary Institutions: 2006--07*.

[3] ANDERSON, J.R., 1988, «*The Expert Module*», Foundations of Intelligent Tutoring Systems, Lawrence Erlbaum Associates, Inc., 21-54.

[4] LTSC, 2001, *1484.12.1 IEEE LTSC Draft Standard for Learning Object Metadata*.

[5] DUVAL, E., FORTE, E., CARDINAELS, K., VERHOEVEN, B., DURM, R.V., HENDRIKX, K., FORTE, M. W., EBEL, N., MACOWICZ, M., WARKENTYNE, K. & HAENNI, F., 2001, «*The ARIADNE Knowledge Pool System*», Communications of the ACM, **44**(5), 72-78.

[6] TERNIER, S., VERBERT, K., PARRA, G., VANDEPUTTE, B., KLERKX, J., DUVAL, E., ORDONEZ, V. & OCHOA, X., 2009, «*The Ariadne Infrastructure for Managing and Storing Metadata*», IEEE Internet Computing, **13**(4), 18-25.

[7] CAFOLLA, R., 2006, «*Project Merlot: Bringing Peer Review to Web-based Educational Resources*», Journal of Technology and Teacher Education, **14**(2), 313-323.

[8] LU, R., CAO, C., CHEN, Y. & HAN, Z., 1995, «*On Automatic Generation of Intelligent Tutoring Systems*», Proceedings of the 7th International Conference on Artificial Intelligence in Education, AIED 1995, AACE, 67-74.

[9] LENTINI, M., NARDI, D. & SIMONETTA, A., 2000, «*Self-instructive spreadsheets: an environment for automatic knowledge acquisition and tutor generation*», International Journal on Human-Computer Studies, **52**(5), 775-803.

[10] DE HOOG, R., BARNARD, Y. & WIELINGA, B. J., 1999, «*IMAT: Re-using Multi-media Electronic Technical Documentation for Training*», Business and Work in the Information Society: New Technologies and Applications, IOS Press, 415-421.

[11] VERBERT, K., OCHOA, X. & DUVAL, E., 2008, «*The ALOCOM Framework: Towards Scalable Content Reuse*», Journal of Digital Information, **9**(1).

[12] ZOUAQ, A., & NKAMBOU, R., 2009, «*Enhancing Learning Objects with an Ontology-Based Memory*», IEEE Transactions on Knowledge and Data Engineering, **21**(6), 881-893.

[13] ALDABE, I., & MARITXALAR, M., 2014, «*Semantic Similarity Measures for the Generation of Science Tests in Basque*», IEEE Transactions on Learning Technologies, **7**(4), 375-387.

[14] LARRAÑAGA, M., 2012, *Semi-Automatic Generation of Learning Domain Modules for Technology Supported Learning Systems using Natural Language Processing Techniques and Ontologies*, Tesia, Euskal Herriko Unibertsitatea UPV/EHU.

[15] LARRAÑAGA, M., CONDE, A., CALVO, I., ELORRIAGA, J.A. & ARRUARTE, A., 2014, «*Automatic Generation of the Domain Module from Electronic Text-*

books. Method & Validation», IEEE Transactions on Knowledge and Data Engineering, **26**(1), 69-82.

[16] Larrañaga, M., Conde, A., Calvo, I., Arruarte, A. & Elorriaga, J.A., 2014, «*Ikaste-domeinuaren sorkuntza erdiautomatikoa*», EKAIA Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria, **0**(26).

[17] Wong, S., & Nguyen, D., 2010, *Principles of Object-Oriented Programming*.

[18] Morison, I., 2008, *Introduction to Astronomy and Cosmology*, Wiley.

[19] Raineri, D., 2001, *Introduction to molecular biology*, Blackwell Science, Malden, MA.

[20] Conde, A., Larrañaga, M., Arruarte, A., Elorriaga, J. A. & Roth, D., 2016, «*LiTeWi: A Combined Term Extraction and Entity Linking Method for Eliciting Educational Ontologies From Textbooks*», Journal of the Association for Information Science and Technology, **67**(2), 380-399.

[21] Salton, G., & Buckley, C., 1988, «*Term-weighting approaches in automatic text retrieval*», INFORMATION PROCESSING AND MANAGEMENT, 513-523.

[22] Frantzi, K., Ananiadou, S. & Mima, H., 2000, «*Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method*», International Journal on Digital Libraries, **3**(2), 115-130.

[23] El-Beltagy, S. R., & Rafea, A., 2009, «*KP-Miner: A keyphrase extraction system for English and Arabic documents*», Information Systems, **34**(1), 132-144.

[24] Conde, Á., 2016, *LiDom builder: Automatising the construction of multilingual domain modules*, Tesia, Euskal Herriko Unibertsitatea UPV/EHU.

[25] Milne, D., & Witten, I. H., 2013, «*An open-source toolkit for mining Wikipedia*», Artificial Intelligence, **194**, 222-239.

[26] Milne, D., & Witten, I. H., 2008, «*Learning to link with wikipedia*», ACM Press, 509.

[27] Fellbaum, C., 1998, *WordNet: An Electronic Lexical Database*, MIT Press.

[28] Ponzetto, S. P., & Strube, M., 2007, *Deriving a large scale taxonomy from Wikipedia*, Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, 1440-1445.

[29] Flati, T., Vannella, D., Pasini, T. & Navigli, R., 2014, *Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project*, Association for Computational Linguistics, Baltimore, Maryland.

[30] Nastase, V., & Strube, M., 2008, «*Decoding Wikipedia Categories for Knowledge Acquisition*», Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI Press, Chicago, Illinois, 1219-1224.

[31] Navigli, R., & Ponzetto, S. P., 2012, «*BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*», Artificial Intelligence, **193**, 217-250.

[32] Samuel Fernando, 2013, *Enriching Lexical Knowledge Bases with Encyclopedic Relations*, University of Sheffield.

[33] AGIRRE, E., & SOROA, A., 2009, «*Personalizing PageRank for Word Sense Disambiguation*», Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2009, The Association for Computer Linguistics, 33-41.

[34] FLEISS, J.L., 1971, «*Measuring nominal scale agreement among many raters.*», Psychological bulletin, **76**(5), 378.

[35] LANDIS, J.R., & KOCH, G.G., 1977, «*The measurement of observer agreement for categorical data*», biometrics, 159-174.

[36] CONDE, A., LARRANAGA, M., CALVO, I., ARRUARTE, A. & ELORRIAGA, J.A., 2012, «*Automating the Authoring of Learning Material in Computer Engineering Education*», Proceedings of 2012 Frontier in Education Conference, Seattle, USA, 1376-138.

[37] MEIRE, M., OCHOA, X. & DUVAL, E., 2007, «*SAmgI: Automatic Metadata Generation v2.0*», Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, ED-MEDIA 2007, AACE, 1195-1204.