SNOMED CT sare semantikoa euskaratzeko aplikazioa

Egilea: Olatz Perez de Viñaspre Garralda
Tutorea: Maite Oronoz Anchordoqui

Hizkuntzaren Azterketa eta Prozesamendua

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2013ko otsaila

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

Laburpena

Master bukaerako proiektu honetan, SNOMED CT sare semantikoa euskaratzeko aplikazioaren lehenengo urratsak azaltzen ditugu. Horretarako, SNOMED CTren sakoneko analisia egin dugu, eta bio-zientzien domeinuko euskarazko baliabideak aztertu ditugu. Aplikazioaren diseinu osoa egin dugu, euskarazko ordainak lortzeko algoritmo bat definituz. Algoritmo horren lehenengo urratsa izan da inplementatu duguna, hiztegi espezializatuen parekatzeari dagokiona, alegia. Aplikazioaren hastapen hauetan emaitza itxaropentsuak lortu ditugu.

Abstract

This paper presents the first steps in the development of a system to translate the SNOMED CT ontology into Basque. For this purpose, we analyzed SNOMED CT in depth and we examined the bio-medical domain resources in Basque. We did a full design of the system, defining an algorithm to get the Basque terms. The first step of that algorithm is what we implemented, the part that corresponds to the specialized dictionaries matching. We obtained promising results at the beginnings of the system.

Gaien aurkibidea

Taulen zerrenda

Irudien zerrenda

1 Proiektuaren definizioa

Hizkuntzaren Azterketa eta Prozesamendua Masterreko tesi-lan honen helburua, medikuntzaren domeinuko sare semantiko baten euskaratze erdi-automatikoa definitzea eta garapenaren lehen urratsak ematea da. Aipatutako sare semantikoa SNOMED CT (***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms*) medikuntza-arloko ontologia da, hain zuzen ere.

Ontologia hori medikuntza-arloan orain artean egin den terminologia eleanitz ulergarriena kontsideratzen da, eta bere garapenak elkarreragingarritasun semantikoan aurrerapauso bat suposatu du. Nolabait esateko hizkuntza eta sistema desberdinen arteko dokumentu klinikoen adierazpen eta interpretazio automatikoa eta anbiguotasunik gabea ahalbidetuko duen hiztegi normalizatua da, hiztegi sarreren arteko harremanak zehaztuta daudelarik.

Ontologia horren abantailez gain, SNOMED CT euskaratuz gero, ontologia honi loturik dauden UMLS (***U**nified **M**edical **L**anguage **S**ystem*) moduko beste hainbat baliabide semantiko atzigarri izango ditugu gure hizkuntzan.

UMLS, medikuntza munduan dauden terminologia desberdinak batzeko, AEBko Medikuntzako Liburutegi Nazionalak sortutako proiektu bat da. Bere helburua, biomedikuntza eta osasungintzako terminologia "ulertuko" balute bezala jokatzen duten konputagailu sistemen garapena erraztea da, hizkuntzaren mugetatik at. Berau, SNOMED CT baino askoz ahaltsuagoa bada ere, SNOMED CTren euskarazko bertsioarekin nahikoa litzaiguke proiektu honen baliabideak gure hizkuntzan erabili ahal izateko, SNOMED CT modu independentean erabil daitekeen arren, UMLSko *Metathesaurus* moduluaren parte baita.

SNOMED CT euskaratzeko hainbat ataza garatu ditugu eta euskaratzeaz arduratuko den sistemaren oinarriak ezarri ditugu. Alde batetik SNOMED CT bera aztertu dugu (3.1 atala), bere ezaugarrietan sakonduz, eta eskuragarri dauden ingelesezko eta gaztelaniazko bertsioen arteko aldeak nabarmendu ditugu. Azterketa honen helburua, SNOMED CTren zein bertsio izango genukeen abiapuntu zehaztea izan da. Beste aldetik, SNOMED CTren terminoak aztertu ditugu, hauetan eman daitezkeen patroiak aztertuaz (3.2 atala). Modu honetan Itzultzaile Automatikoek izan ditzaketen gabeziak identifikatu ditugu, baita hauen ekarpenak ere. Analisiaren atalarekin bukatzeko, SNOMED CT eta Gaixotasunen Nazioarteko Sailkapenaren arteko mapaketa sakonki aztertu dugu (3.3 atala).

Aplikazioaren diseinuari dagokionean (4 atala), aplikaziorako definitu dugun algoritmoa azalduko dugu 4.1 atalean. Ostean 4.2 atalean terminologia eskakizunei aurre egiteko moldatu dugun e**X**tended **M**arkup **L**anguage (XML) lengoaiako estandar bat (TBX) definituko dugu; eta azkenik, aplikazioaren klase-diagrama azalduko dugu.

Esan bezala, aplikazioaren diseinuan definitutako lehenengo urratsak inplementatu ditugu master-tesi honetan, eta egindako hurbilpena 5 atalean zehaztuko dugu. Horrela, lehenengo emaitza batzuk jaso ahal izan ditugu eta 6 atalean erakutsiko ditugu. Azkenik, lan honen ondorioak eta etorkizuneko lana zehaztuko ditugu 7 atalean.

1.1 Motibazioa

Euskara normalizazio-fasean dagoen hizkuntza izanik, hainbat alor espezifikoetan honen erabilera mugatua da, eta osasungintzaren arloa muga hauen barnean koka dezakegu. Osasun-langile euskaldunak Osakidetzako osasun-zentroetan barra barra aurkitu badaitezke ere, domeinuan dauden gabeziak medio, euskara erabiltzeko arazoak dituzte. Gabezien artean terminologia glosategi osatu baten falta nabaria da. Honen haritik, SNOMED CTk gabezia hori ase dezake, bere baitan medikuntzako terminologia zabala jasotzen baitu.

Honetaz gain, SNOMED CTk osasun-alorreko informazioa kudeatzeko eta erauzteko abantaila asko ematen dizkigu. Gainera, SNOMED CTren eleaniztasunak informazio erauzketa eleanitza ahalbidetzen du, euskararako dagoen baliabide eskasiari aurre egiten lagundu dezakeena. Euskara moduko hizkuntza gutxitu baterako sare garrantzitsu honen bertsio bat izateak ere motibatzen du lan hau.

2 Aurrekariak

Atal honetan, lehenik eta behin, SNOMED CT ontologia kokatuko dugu gure lanaren oinarria da-eta (2.1 atala); jarraian Gaixotasunen Nazioarteko Sailkapenaz hitz egingo dugu, honen euskarazko banaketa eskura dagoelako (2.2 atala) eta sailkapen honen eta SNOMED CTren arteko mapaketari buruz (2.3 atala) ere arituko gara; azkenik TZOS tresnaren inguruan arituko gara, honen glosategien inguruan, hain zuzen ere (2.4 atala). Eskura izan ditugun euskarazko hiztegi elebidun eta espezializatuak ere aipatuko ditugu atal honetan (2.5 atala).

2.1 SNOMED CT

SNOMED CT (***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms*) medikuntza alorreko ontologia zabala da. Besteak beste, osasun-txostenetan aurki daitezkeen kontzeptuak, deskribapenak eta erlazioak barnebiltzen dituen ontologia da. Medikuntza-arloan orain artean egin den terminologia eleanitz ulergarriena kontsideratzen da, eta bere garapenak elkarreragingarritasun semantikoan aurrerapauso bat egiten du. Nolabait esateko hizkuntza eta sistema desberdinen arteko txosten klinikoen adierazpen eta interpretazio automatikoa eta anbiguotasunik gabea ahalbidetuko duen hiztegi normalizatua da, hiztegi sarreren arteko harremanak zehaztuta daudelarik.

Kontzeptuak deskribapenen bitartez definitzen dira eta termino baliokideak elkartzeko ere balio dute. Kontzeptuak identifikadore batekin adierazten dira eta deskribapenetan definizio zein termino baliokideak agertuko dira. Erlazioen bitartez kontzeptuen arteko harremana zehazten da, egitura hierarkikoa emateaz gain ("is a" harremana) beste harreman motak ere adierazten direlarik (*causative agent* motakoak adibidez edo kausa adierazteko harremana). Horrela, SNOMED CT hierarkiatan banatzen da, 2.1.3 atalean ikusiko dugun bezala. Hiru elementu hauek ikus ditzakegu 1. adibidean.

Kontzeptua: 38907003
Deskribapenak:
Varicella (disorder)
Varicella
Chicken pox
Varicella infection
Erlazioak:
`Varicella | Is a | Viral disease characterized by exanthem`
`Perinatal varicella | Is a | Varicella`
`Varicella | Causative agent | Human herpesvirus 3`
`Varicella | Pathological process | Infectious process`
...

1. adibidea: SNOMED CTren elementu ezberdinak

Bere jatorrizko hizkuntza ingelesa da, eta honentzako 300.000 kontzeptu baino gehiago definituak ditu, baita hauei dagozkien 1.000.000 deskribapen baino gehiago ere. Deskribapenak kontzeptuei dagozkien termino baliokideak dira, alde batetik definizio deskribatzaileak ditugularik (definizioa bera identifikatzeko balio dutenak), eta bestetik, definizioen sinonimoak daudelarik, jarraian ikusiko dugun bezala. SNOMED CTk gaur egunean erabilera zabala du mundu osoan zehar. Hainbat hizkuntzetarako bertsioak daude eskura edota garatze-lanetan, hala nola, holandesa, frantsesa, gaztelania,...

Lan honen motibazio nagusia SNOMED CTren euskaratze erdi-automatikoa egitea da. Izan ere, SNOMED CTren terminoen ezagutzarekin, Itzulpen Automatikoko hainbat teknika erabili nahi dira, ahalik eta zatirik handiena automatikoki itzuli ahal izateko. SNOMED CT beste hizkuntzetara itzulia izan da, zenbaitetan eskuz (danieraren kasu, Petersen (2011)), beste batzuetan itzulpen automatikoa eskuzko itzulpenarekin konbinatuz (txinera adibidez, Zhu et al. (2012)), edota itzulpenean laguntzeko metodo automatiko bat erabiliz (frantsesaren kasuan, Abdoune et al. (2011)). Ataza honetan IHTSDOk argitaratutako SNOMED CTren itzulpenerako gidalerroa (Høy (2010)) ondo aztertu eta gomendioak jarraitzea garrantzitsua da.

2.1.1 Deskribapen motak

Bi deskribapen mota bereizten dira SNOMED CTn: *Fully Specified Name* (FSN), eta *Synonym*.

- *Fully Specified Name*: Kontzeptua identifikatu eta bereizteko erabiltzen diren deskribapenak dira. Ez dira osasun-txostenetan aurkitzen diren terminoak. Orokorrean terminoari jarraiki parentesi artean kategoria semantikoa adierazten duen *semantic tag* deiturikoa zehazten zaio. 2. adibidean *"Myocardial infarction (disorder)"* FSNaren kategoria semantikoa *disorder* litzateke.

- *Synonym*: Osasun-txostenetan aurkitu daitezkeen kontzeptuen deskribapenak dira hauek. Kontzeptu bakoitzerako gutxienez sinonimo bat zehazturik dago eta bi etiketaren arabera sailkaturik daude: *Preferred* eta *Acceptable*.

  - *Preferred Term*: Hobetsitako sinonimoa adierazten da etiketa honen bitartez. Kontzeptu bakoitzerako hobetsitako termino bakarra egon daiteke, hizkuntza edo dialekto bakoitzerako bat. 2. adibidean *"Myocardial infarction"* litzateke *Preferred Term*a.
  - *Acceptable Synonym*: Gainerako sinonimo guztiak dira, hau da, hobetsitakoak ez diren sinonimo guztiak. Kontzeptu bakoitzak nahi adina sinonimo izan ditzake, baita bakar bat ere ez.

***Fully Specified Name* (FSN)**:	*Myocardial infarction (disorder)*
Hierarkia:	*Clinical Finding/disorder*
Semantic tag:	*disorder*
***Preferred Term* (PT)**:	*Myocardial infarction*
Acceptable Synonym:	*Heart attack*
	Infarction of heart
	MI - Myocardial infarction
	Myocardial infarct
	Cardiac infarction

2. adibidea: Kontzeptu baten deskribapen motak

Hemendik aurrera, hobetsitako sinonimoei *Preferred Term* (PT) deituko diegu eta gainerakoei *Acceptable synonym*. Hauek biak batzen dituen *Synonym* terminoa Sinonimo hitzaren bitartez adieraziko dugu, hasiera letra larriz ipiniz. Kontzeptu, Deskribapen eta Erlazioak ere letra larriz idatziko ditugu, hauek SNOMED CTren kontzeptu, deskribapen eta erlazioak direla adierazteko.

Lehen aipaturiko *semantic tag* edo etiketa semantikoak, Kontzeptuari dagokion hierarkia bakarrik adierazi beharrean, kasu batzuetan informazio gehigarria ematen du. *Clinical Finding/disorder* (aurkikuntza/gaixotasun klinikoa) hierarkiaren kasuan adibidez, bi *semantic tag* bereizten dira, *disorder* (gaixotasun) eta *finding* (aurkikuntza). 1. taulan ingelesezko eta gaztelaniazko bertsioetan aurkitu ditugun *semantic tag* kopurua erakusten dugu, SNOMED CTk dituen hierarkiatan sailkaturik.

	Ingelesezko bertsioa		Gaztelaniazko bertsioa	
Hierarchy	**Semantic Tag (ST)**	**FSNen #**	**Semantic Tag (ST)**	**FSNen#**
Clinical Finding/disorder	disorder	94,242	trastorno	82,725
	finding	45,401	hallazgo	36,625
Procedure/intervention	procedure	75,078	procedimiento	59,411
	regime/therapy	3,573	régimen/terapia	2
			régimen/tratamiento	2,773
Organism	organism	35,870	organismo	35,465
Body structure	body structure	26,960	estructura corporal	26,747
	morphologic abnormality	5,259	anomalía morfológica	5,082
	cell	645	célula	640
	cell structure	513	estructura celular	509
Substance	substance	25,834	sustancia	24,918
Pharmaceutical/biologic product	product	24,379	producto	23,854
Qualifier value	qualifier value	10,134	calificador	9,570
Observable entity	observable entity	9,044	entidad observable	8,602
Event	event	8,959	evento	8,587
Situation with explicit context	situation	8,716	situación	5,785
Social context	occupation	6,460	ocupación	4,650
	person	668	persona	432
	ethnic group	366	grupo étnico	283
	religion/philosophy	227	religión/filosofía	217
	life style	30	estilo de vida	25
	social concept	27	contexto social	26
	racial group	21	grupo racial	19
Physical object	physical object	5,148	objeto físico	4,747
Specimen	specimen	1,455	espécimen	1,386
Environment	environment	1,253	medio ambiente	1,162
geographical location	geographic location	619	localización geográfica	619
	environment/location	1	medio ambiente/localización	1
Linkage concept	attribute	1,157	atributo	1,145
	link assertion	8	relación asertiva	8
	linkage concept	1	concepto de enlace	1
Staging and scales	assessment scale	1,125	escala de evaluación	1081
	tumor staging	261	estadificación tumoral	249
	staging scale	41	escala de estadificación	16
Special concept	navigational concept	732	concepto para navegación	725
	namespace concept	153	espacio de nombres	153
	administrative concept	80	concepto administrativo	31
	special concept	31	concepto especial	1
Record artifact	record artifact	318	elemento de registro	234
Physical force	physical force	178	fuerza física	174
Root Metadata Concept	foundation metadata concept	164	metadato fundacional	134
	core metadata concept	31	metadato del núcleo	32

1. taula: Ingelesezko eta Gaztelaniazko bertsioetako *semantic tag*ak

2.1.2 Argitalpen bertsioak

SNOMED CT sei hilean behin eguneratzen da, baita bere itzulpenak ere (gaztelaniazkoa kasu). Bertsio guztiak AEBetako Medikuntza Liburutegiak banatu egiten ditu doan, eta erraz jaso daitezke bertan erregistratuta egonez gero.

Hurrengo lerroetan ikusiko dugun bezala, SNOMED CTren data konkretu bateko banaketan bertsio ezberdinak aurki daitezke. Hala nola SNOMED CTren ikuspegi zein formatu ezberdinek eragindako bertsioak izango dira, baina aukeraketaren arabera SNOMED CT beraren edukian eragina nabaria izango da.

2.1.2.1 Formatu ezberdinak: RF1 eta RF2

Release Format 1 (RF1), SNOMED CTren lehenengo argitalpenetik erabilitako formatua da, hau da, 2002. urtetik aurrera. 2012 urtean zehar RF1 formatua *Release Format 2*rekin ordezkatua izan zen.

RF2 SNOMED CTren erabiltzaileen eskaerengatik sortua izan da. Izan ere, RF1 formatuan hainbat ahultasun aurkitu izan dira, egituraketari dagokionean, eta honi erantzuteko RF2k sendotasun eta trinkotasuna eman dio. Adierazpide logikoan urratsak emateko ere erabili da formatu aldaketa, ontologiekin lan egiteko aukera zabala emanaz.

RF1ean, hiru fitxategi daude SNOMED CTren muinean: Kontzeptuak, Deskribapenak eta Erlazioak. RF2ra igarotzean aldiz, fitxategi kopurua dezente igotzen da, baina fitxategi bakoitzaren edukia mugatuagoa da. Adibidez, Kontzeptuen fitxategian kodeak eta metadatuen kodeak bakarrik agertuko dira, eta Deskribapenen fitxategira jo beharko dugu Kontzeptu baten FSNa ezagutzeko. Izan ere, termino irakurgarriak fitxategi honetan bakarrik aurkituko ditugu, eta gainontzeko fitxategi guztietan, Erlazioak, Kontzeptuak, *Crossmap*ak eta Dialekto ezberdinen informazioa duten fitxategien kasu, metadatu edota Kontzeptuen kodeak baino ez zaizkigu agertuko.

Aipatu beharrik ere ez dago, gure lanetarako RF2 bertsioa erabiliko dugula, RF1 desagertze-bidean egoteaz gain, ahulagoa baita.

2.1.2.2 Ikuspegi ezberdinak: *Full, Snapshot* eta *Delta*

RF2rekin batera, SNOMED CT hiru ikuspegi ezberdin eskaintzen hasi da: *Full, Snapshot* eta *Delta*. Ikuspegiak SNOMED CTren zein zati erakutsiko den zehazten du.

- *Full*: Inoiz argitaraturiko osagai guztiak barnebiltzen dituen ikuspegia da. Inoiz osagairen bat "aktibo" egoeratik "ez-aktibo" egoerara pasatu bada, osagai horren bi sarrera agertuko dira ikuspegi honetan: "aktibo" zegoenekoa eta "ez-aktibo" bilakatu zenekoa. Osagaien bilakaera aztertzeko baliagarria da.

- *Snapshot*: Osagai bakoitzeko bertsio bakarra erakusten du ikuspegi honek, argitalpenaren unera arte argitaraturiko bertsiorik gaurkotuena.

HAP masterra

- *Delta*: Aurreko bertsiotik uneko bertsiora aldatutako osagaiak bakarrik erakusten dituen ikuspegia da. Eguneraketak egiteko baliagarria da, SNOMED CT osoa berraztertu behar izan gabe.

Hiru ikuspegi hauetatik gure lanerako *Snapshot* ikuspegia erabiliko dugu, osagai bakoitzeko bertsio berrituena erakusten duelako eta bertsio zaharkituak alde batera uzten direlako. Horrela, osagai gaurkotuak itzuliko ditugu, eta hemendik aurrerako bertsioetan, *Delta* ikuspegian agertzen diren osagaiak itzuliz gure Euskarazko SNOMED CT bertsioa eguneratuta mantenduko dugu.

2.1.3 Hierarkiak

RF2 formatua ezarri denetik, SNOMED CTk bi erro Kontzeptu ditu, gerora horien azpian 2 hierarkia garatzen direlarik. Lehenengoa metadatuak egituratzeko erabilitako hierarkia da: *Root Metadata Code*. Azken honetan, SNOMED CT beraren informazioa egituratzen da, *Fully Specified Name*, *Preferred*, *Acceptable*, *Current*... moduko elementuen bitartez.

Bigarren hierarkia, SNOMED CTren edukia antolatzeko erabiltzen da: *Root Concept Code*. Honen barruan 19 goi mailako hierarkia definitu dira (*Top level hierarchies*), 1. taulako ezkerreko zutabean ikus ditzakegunak (*Clinical Finding/disorder, Organism, Body structure...*). Eduki-hierarkia hauetan SNOMED CT osatzen duten Kontzeptu guztiak barnebildurik daude, bakoitza dagokion hierarkian antolatuta. Hierarkia bera egituratzeko SNOMED CT Erlazioak erabiltzen dira, *"is a"* Erlazioaren bitartez, umeak gurasoari erreferentzia egiten diolarik.

Itzulpen-lanetarako hierarkia bakoitza ezagutzeak berebiziko garrantzia dauka, izan ere, hauen ezaugarrien arabera taxonomia erabili beharko da, edota ingelesezko jatorrizko terminoa utzi. Beste batzuetan, hobetsitako terminoa taxonomikoa izango bada ere, sinonimoak euskaratzea beharrezkoa izango da, *Organism* hierarkiaren kasu (3.1.4 atalean sakonduko dugu gai hau). Hierarkia hauen itzulpenerako argipenak Høy (2010) txosten teknikoan ematen dira.

SNOMED CTk dituen hierarkien terminoen adibide bana erakusten dugu 2. taulan. Hierarkia bakoitzak dituen *semantic tag*en arabera eman ditugu adibide horiek.

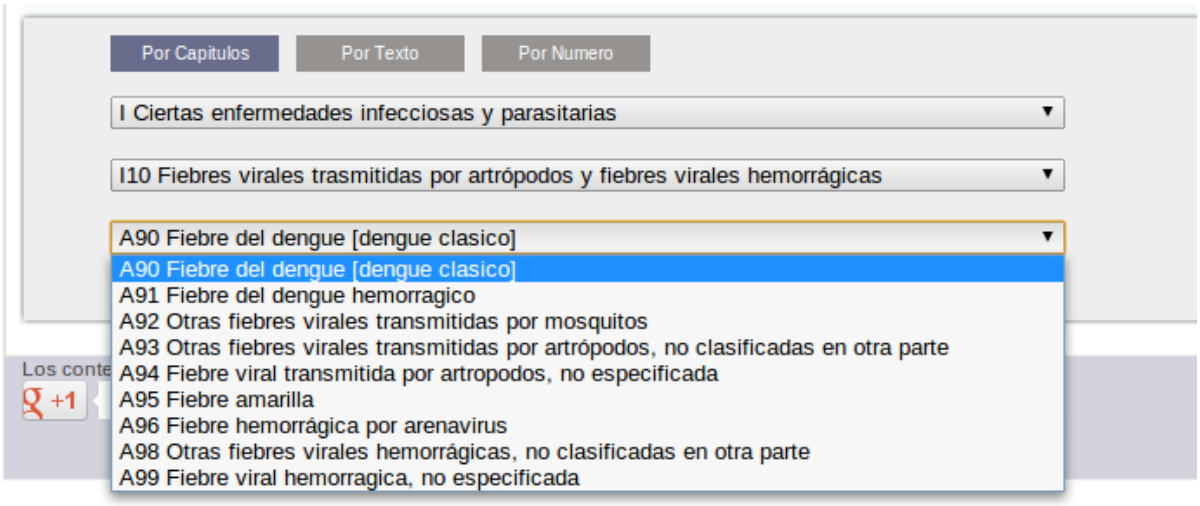
Hierarkia	***Semantic Taga***	**Adibidea**
Clinical Finding/disorder	disorder	Myocardial infarction
	finding	Hyperalphaglobulinaemia
Procedure/intervention	procedure	Eye structure transplantation
	regime/therapy	Pulsed electromagnetic energy to shoulder
Organism	organism	Pelistega europaea
Body structure	body structure	Supratentorial brain structure
	morphologic abnormality	Acute erythremia
	cell	Umbrella cell
	cell structure	Viral envelope
Substance	substance	Bacterial agent
Pharmaceutical/biologic product	product	Naratriptan
Qualifier value	qualifier value	Perinatal period
Observable entity	observable entity	Postvaccination state
Event	event	Flood
Situation with explicit context	situation	Mother smokes
Social context	occupation	Hospital nurse
	person	Homosexual parents (family)
	ethnic group	Irish traveller
	religion/philosophy	Nonconformist religion
	life style	White collar thief
	social concept	Upper class economic status
	racial group	American Indian race
Physical object	physical object	Cardiac compression board
Specimen	specimen	Lumpectomy breast sample
Environment	environment	Psychiatric intensive care unit
geographical location	geographic location	Republic of Serbia
	environment/location	Environment or geographical location
Linkage concept	attribute	Agent relationship
	link assertion	Has problem member
	linkage concept	Linkage concept
Staging and scales	assessment scale	Lequesne index
	tumor staging	pM category
	staging scale	Chest pain rating
Special concept	navigational concept	Enzymes A - L
	namespace concept	Extension Namespace 1000001
	administrative concept	Appointment
	special concept	Special concept
Record artifact	record artifact	Family history section
Physical force	physical force	Vapour pressure
Root Metadata Concept	foundation metadata concept	Referenced component
	core metadata concept	Fully specified name

2. taula: Hierarkien terminoen adibideak

2.2 Gaixotasunen Nazioarteko Sailkapena (GNS-10)

Gaixotasunen Nazioarteko Sailkapena gaixotasunak sailkatzeko irizpideak eskaintzen dituen Munduko Osasun Erakundeak sortutako nazioarteko sailkapena da. Irizpide hauen bitartez, gaixotasunei kode estandar bat edo gehiago esleitzeko aukera ematen zaie, hizkuntza eta herrialde guztietan berdina izango dena.

***I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems* jatorrizko ingelesezko izena dauka eta 10. bertsioa da (ICD-10) gaur egun indarrean dagoena. ***W**orld **H**ealth **O**rganization* (WHO) erakundeak banatu duen azken berrikustea 1992. urtean. Euskal Herrian UZEIk euskaratu zuen eta 1996an argitaratu zuen Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusiak. Gaur egun Espainiako Osasun Ministeritzak GNS-9 bertsioa erabiltzen du informe klinikoen diagnostikoak sailkatzeko eta GNS-10 heriotza-kausa sailkatzeko. 1. irudian adibide bat erakusten dugu. Bertan "antropodoek sortutako sukar birikoak" ikus ditzakegu eta hauen arabera "dengearen sukar" gaixotasunari "A90" kodea esleituko genioke.



1. irudia: GNS-10aren adibide bat

WHO erakundea GNS-11 bertsioa lantzen hasia da jada, eta berau garatzeko oinarri ontologikoa SNOMED CT da. Hortaz, esan dezakegu GNSren 11. bertsioak SNOMED CT izango duela oinarri.

2.3 SNOMED CT eta GNS-10en arteko mapaketa

Hurrengo lerroetan 2012ko abuztuan IHTSDO erakundeak ofizialki zabaldutako SNOMED CT eta GNS-10ren arteko mapaketa dugu mintzagai (Organisation (2012)). Mapaketa honek ontologiaren eta sailkapenaren arteko mapaketa erdi-automatikoa eskaintzen du, zeina WHOk (GNSren sortzailea) eta IHTSDOk (SNOMED CTren kudeatzailea) balioztatua

den. Mapaketa honen helburua SNOMED CT Kontzeptu baten esanahia identifikatuz, Kontzeptuaren GNS-10 espazio semantikoaren tokirik aproposena aurkituz eta GNS-10 sailkapen kodea(k) esleituz, SNOMED CT Kontzeptuaren eta GNS-10 kodearen/kodeen arteko esteka sortzea da. Horrek SNOMED CTren itzulpena eskainiko digu GNS-10 dela-medio.

GNS sailkapenak eta SNOMED CTk oso egitura ezberdina daukate. Azken finean, bata gaixotasunen sailkapena izatera mugatzen den bitartean, besteak ontologia batek eman dezakeen sakontasuna barnebiltzen du, medikuntzako arloko kontzeptu ezberdinak era hierarkikoan sailkatuaz. Garrantzitsua da mapaketaren norantza aipatzea: SNOMED CT Kontzeptuen GNS10 kode baliokideak ematen ditu. Mapaketa hau egiteko, SNOMED CTren testuinguru falta bere gain hartzen du parekatzeak.

Denera 19.293 SNOMED CT Kontzeptu parekatu dira, bakoitzarentzat 0tik 19ra GNS-10 kode esleitu direlarik. 0 kasua berezia da, garapen kontuetarako erabiltzen dutena, eta orokorrean esan daiteke bat eta hiru artean egoten dela parekatze kopurua. Azalpena ulergarriago egiteko, hemendik aurrera, mapaketa hitza mapaketa osoari erreferentzia egiteko erabiliko dugu, eta parekatzea hitza Kontzeptu bakoitzaren mapaketa zehatza adierazteko.

Mapaketan parte hartzen duten Kontzeptuetarako denera 27.167 parekatze definitu dira. Kontzeptu hauek hiru hierarkia ezberdinetakoak dira: *Clinical findings* (*disorders* 23.393 parekatze eta *findings* 3.171), *Events* (184 parekatze) eta *Situations with explicit context* (413 parekatze).

Mapaketa honen sakoneko analisia 3.3 atalean egiten dugu, mapaketaren egitura aztertuz, honen zailtasun eta arazoak ulertzeko.

2.4 Terminologia Zerbitzurako On-line Sistema (TZOS)

Terminologia Zerbitzurako On-line Sistemak (Arregi et al. (2010)) terminoak jasotzeko ingurunea, jasotako informazioa gordetzeko eta prozesatzeko baliabideak eta kontsultetarako eta elkarrekintzarako interfazea eskaintzen du. Hau da, domeinu ezberdinetako terminoak lantzeko eta zabaltzeko zerbitzua eskaintzen du.

Unibertsitateko adituak (irakasleak kasu honetan) euren jarduera aurrera eramateko beharrezkoa duten euskarazko terminologia sortzen eta eguneratzen dabiltza etengabe. Orain artean, terminologia mailan egindako esfortzu guztia ikasgaiaren baitan geratzen zen, zabalkunde berezirik izan gabe. Terminologia lan horri zabalkundea emateko intentzioz sortu zen TZOS besteak beste.

TZOSen sortzaileek terminologia jarduera kolektibo gisa ulertzen dute, eta bide horretan TZOSek bitartez terminologia-sortzaile potentzialen komunitateari rol aktiboagoa eskaini nahi diote. Elkarlaneko terminologia (*collaborative terminology*) deitu izan zaion molde horretara hurbiltzeko saioa da, nolabait.

TZOSen baitan Euskal Herriko Unibertsitateko hainbat fakultatetako irakasle elkarlanean ari dira euren jakintza arloko glosategiak osatzeko. Glosategi hauen erreferentzia hizkuntza euskara bada ere, beste hizkuntzekin termino-ordain pareak osatu dituzte. Horrela, hiztegi funtzioa ere barnebiltzen du.

HAP masterra

Master-tesi honi dagokionean, bio-zientzien inguruko glosategiak interesatzen zaizkigu, hala nola, anatomiako glosategia, farmaziakoa, eta abar. Gaur egun glosategiak bukatu gabe daude, eta oraindik garatze fasean dagoen anatomiako glosategiko bertsio ez-bukatua erabili dugu. Horrela, glosategiaren zati bat erabili ahal izan dugu aplikazioaren garapen-fase honetan. Anatomiako glosategiaren 7.017 sarreretatik, 2.576 sarrera bakarrik izan ditugu erabilgarri.

2.5 Hiztegi espezializatuak

SNOMED CTren euskaratze honetarako, hainbat hiztegiren laguntza izan dugu. Bio-zientzien domeinuko hiztegiak bilatu ditugu, eta baita arloka sailkatutako hiztegi espezializatuak ere. Bi multzotan sailkatu ditzakegu batu ditugun hiztegiak: gaztelania-euskara hizkuntzen artekoa bakarrik, eta ingeleseko ordaina ere duten hiztegiak:

- Ingelera-gaztelania-euskara:
  - **Zientzia eta Teknologiaren Hiztegi Entziklopedikoa**: izenak dioen bezala, zientzia eta teknologiaren hiztegi bat da hau, Elhuyarrek garatua. Hiztegiaren izateko arrazoia da zientzia eta teknologiari buruzko erreferentzia-informazio fidagarri, landu eta eguneratua eskaintzea, modu zehatz, argi eta ulergarrian, eta erabiltzaile-multzo zabala gogoan hartuta. Besteak beste, medikuntzako, biokimikako, biologiako, anatomiako eta psikiatriako alorrak aurki daitezke.
  - **Erizaintzako Hiztegia**: EHUko Euskara Zerbitzuak eta Donostiako Erizaintzako Unibertsitate Eskolak argitaraturiko hiztegi honetan, erizaintzan erabiltzen diren hainbat termino aurki daitezke. Bertan euskarazko hiztegiaz gain, Gaztelania-Euskara (4.155 sarrera), Ingelesa-Euskara (4.671 sarrera) eta Frantsesa-Euskara hiztegi elebidunak daude.
  - **GNS-10**: 2.2 atalean ikusi dugun bezala Gaixotasunen Nazioarteko Sailkapenaren euskarazko, gaztelaniazko eta ingelesezko bertsioak eskuragarri daude.
- Gaztelania-euskara
  - **Drogei buruzko hiztegia**: Drogamendekotasunei buruzko Dokumentazio Zentroak 2002. urtean kaleratutako hiztegi honetan, gaztelaniako terminoen euskal ordainak izateaz gain, hauen definizio elebiduna ere eskaintzen du. Drogari eta drogaren munduari nolabait dagozkion adierazpenak biltzen ditu.
  - **Administrazio Sanitarioko Oinarrizko Hiztegia**: Osakidetzak eta UZEIk kaleratutako hiztegia da, zeinetan izenak dioen bezala osasun alorreko administraziorako oinarrizko terminoak datozen.

3 Analisia

Atal honetan hiru azterketa mota egin ditugu. Alde batetik SNOMED CTren ezaugarriak eta honen gaztelaniazko eta ingelesezko bertsioen azterketa alderatzailea egin dugu, SNOMED CT euskaratzeko jatorri bertsioa zehazteko asmoz (3.1 atala). Bestetik, SNOMED CTren terminoak aztertu ditugu eta hauen gainean itzulpen automatikoak izan ditzakeen ekarpenak eta gabeziak identifikatu ditugu (3.2 atala). Azkenik, SNOMED CT eta GNS--10ren arteko mapaketaren analisia egingo dugu 3.3 atalean.

3.1 SNOMED CT: Gaztelaniazko eta Ingelesezko bertsioen azterketa

Gure helburu nagusia SNOMED CTren sakoneko azterketa izan da, bere indarguneak eta ahulguneak identifikatuz. Azterketa honek SNOMED CTren zein zati edo hierarkia itzuli behar ditugun lehenik erabakitzen lagunduko digu, baita zein izan behar den jatorri hizkuntza ere (ingelesa edo gaztelania). Gainera, SNOMED CTren ikuspegi errealista izaten ere lagunduko digu.

3.1.1 Abiapuntua

SNOMED CTk bertsio ezberdinak ditu eskura, 2.1 atalean azaldu dugun moduan. Azterketarekin hasi aurretik, erabiliko dugun SNOMED CT bertsioen ezaugarrien errepaso azkarra egingo dugu:

- Formatua: *RF2*.
- Ikuspegia: *Snapshot*.
- Bertsioa: SNOMED CTren ingelesezko nazioarteko banaketa 2012ko urtarrilaren 31koa; gaztelaniazko banaketa, banaketa internazionalean oinarritua dena, 2012ko apirilaren 30koa.
- Definizioak: Soilik aktibo dauden Deskribapenak hartu ditugu kontuan, euren Kontzeptuaren egoera alde batera utzita. SNOMED CTren inplementazio-gidan aktibo ez dauden Kontzeptuak eginkizun historikoetarako baliagarriak dira, aurretik noizbait aktibo egon baitira.

Hurrengo ataletan honako atazak dira aztertuko ditugunak:

1. SNOMED CTren zenbaki orokorrak, hala nola, *Fully Specified Name* kopuruak, Sinonimoen batezbestekoak, eta abar (3.1.2 atala).
2. Dialektoen eragina Sinonimoetan (ingelesezko bertsioan) (3.1.3 atala).
3. *Preferred Term*en (PT) gabezia gaztelaniazko bertsioan (3.1.4 atala).

HAP masterra

4. FSNen terminoen eta Sinonimoen arteko antzekotasuna (3.1.5 atala).

5. Sinonimoen luzera hitz kopurua kontuan izanik (3.1.6 atala).

6. Gaztelaniazko bertsioan agertzen ez diren Kontzeptuen kopurua eta hauen *semantic tag*a (3.1.7 atala).

7. *Semantic tag* bakoitzaren kopuruak hizkuntza bakoitzean (3.1.8 atala).

Atazak aztertu ostean, azterketa honetatik ateratako ondorio nagusiak ere azalduko ditugu.

3.1.2 Zenbaki orokorrak

Kontzeptuen, FSNen, Sinonimoen, *Preferred Term*en eta *Acceptable Synonym*en kopuruak (# tauletan) eta batezbestekoak erakutsiko ditugu atal honetan. Aurretik azaldu bezala Sinonimoak *Preferred Term*ek eta *Acceptable Synonyms*ek osatzen dituzte (azken bien baturak Sinonimoen kopurua eman beharko luke).

Jarraian dugun 3. taulan datu interesgarri asko atera daitezke. Alde batetik, gaztelaniazko bertsioan Kontzeptu kopurua ingelesezko bertsioarekin alderatuz baxuagoa dela ikus daiteke, hau da, ez dira Kontzeptu guztiak gaztelaniara itzuli. Galdutako 40.864 Kontzeptu hauek 3.1.7 atalean aztertuko ditugu.

Horretaz gain, ingelesezko bertsioan Kontzeptu bakoitzerako 1,07 *Preferred Term* daudela ikus daiteke, hau da, Kontzeptu bakoitzerako PT bat baino gehiago. Gaztelaniazko bertsioan aldiz, kontrakoa gertatzen da, Kontzeptuak baino PT gutxiago daude. *Preferred Term*en inguruko datuei buruz 3.1.3 eta 3.1.4 ataletan hausnartuko dugu.

Azkenik, ikus daitekeen moduan, gaztelaniazko bertsioan Kontzeptuetako batek ez dauka FSN aktiborik (354.648 Kontzeptuetatik, 354.647 FSN daude). FSN gabeko Kontzeptuaren *Preferred Term*a *"lesión de médula espinal - sin fractura vertebral"* litzateke. *Preferred Term*a aktibo baldin badago ere, ez dago hau identifikatzen duen FSNrik, eta horrek gaztelaniazko bertsioari sendotasuna kentzen dio.

	Ingelesezko bertsioa		**Gaztelaniazko bertsioa**	
	#	**Batezbestekoa**	#	**Batezbestekoa**
Kontzeptuak	395.512	-	354.648	-
FSN	395.512	1,00	354.647	1,00
Sinonimoak	619.742	1,57	446.768	1,26
Preferred Terms	422.841	1,07	328.069	0,93
Acceptable Synonyms	196.901	0,50	118.699	0,34

3. taula: Zenbaki orokorrak eta batezbestekoak

3.1.3 *Preferred Term*ak eta dialektoak

Aurreko taulan (3. taula), ingelesezko bertsioan Kontzeptuak baino *Preferred Term* gehiago daudela ikus daiteke (395.512 Kontzeptu eta 422.841 PT). Hau ulertu ahal izateko ingelesezko bertsioak bi dialekto barnebiltzen dituela kontuan izan behar dugu: Britainia Handiko (BH) ingelesa eta AEBko ingelesa. Horrela, 4. taulak Kontzeptu bakoitzeko PT bakarra dagoela erakusten du (3. taulako Kontzeptu kopurua), bat hizkuntzaren dialekto bakoitzeko.

Hurrengo taulako (4. taula) zenbakietan ikus daitekeen moduan, dialekto bakoitzerako definitutako 395.512 *Preferred Term*etatik, kasuen %93an termino berdina partekatzen dute. *Fully Specified Name* guztiak berdinak dira bi dialektoetarako. *Acceptable Synonyms*ei dagokienean, Britainia Handiko ingeleserako gutxi batzuk gehiago definitu dira.

	BH ingelesa	AEB ingelesa	Partekatutako terminoak
Deskribapenak	977.704	977.872	940.322
Fully Specified Names	395.512	395.512	395.512
Sinonimoak	582.360	582.192	543.691
Preferred Terms	395.512	395.512	368.183
Acceptable Synonyms	186.848	186.680	175.508

4. taula: BH eta AEB ingelesa

Horretaz gain, 4. taulan ikus dezakegunez, 368.183 *Preferrend Term* dira bi dialektoek partekatzen dituztenak, gainerako 27.329 PTak dialekto bakoitzaren berezitasunei egokitzen zaizkien terminoak izanik.

SNOMED CTren euskaratze lanetarako, ingelesezko bertsioari etekin handiena ateratzeko asmoz, bi dialektoak izango ditugu kontuan, eta biak itzultzen saiatuko gara. Hau horrela izanik, hurrengo ataletan ingelesezko bertsioa osotasunean hartuko dugu kontuan, eta ez bi bertsio bailiran.

3.1.4 Gaztelaniazko bertsioaren *Preferred Term*en gabezia

Jadanik aipatu dugun moduan (gogoratu 2.1 atala), Kontzeptu bakoitzak Sinonimo bat izan behar du *Preferred Term* gisa markatuta. Gaztelaniazko bertsioan 26.579 Kontzeptuen *Preferred Term*a falta dela ohartu gara, hau da, Kontzeptu guztien %7,5a falta da (3. taula).

Gabezia duten Kontzeptuak euren *semantic tag*aren arabera sailkatu ditugu 5. taulan. Kontzeptuen %86,67a *"organismo" semantic tag*ari dagokio. *Semantic tag* hori, *Organism* goi mailako hierarkiari dagokio, zeinetan pertsonen zein animalien medikuntzan esanguratsuak diren organismoak barnebiltzen diren (IHTSDO (2012)).

Hierarki horrek 35.319 *Fully Specified Name* ditu *"organismo" semantic tag*arekin etiketatuta. Hortaz, hierarkia honetako %65tik gora Kontzeptuk ez dauka *Preferred Term*ik gaztelaniazko bertsioan. Itzulpen gidalerroaren (Høy (2010)) aholkuei jarraitzen badiegu, bertsio honek nazioarteko izen taxonomikoak izan beharko lituzke PT gisa:

HAP masterra

> *"Unless it is clearly in conflict with national language policy, names of organisms should be retained as universal (international) scientific terms and should adhere to the accepted orthography, especially with respect to the application of upper and lower case letter conventions in the individual words."*

Organism hierarkiari dagokion Kontzeptu bat erakusten dugu 3. adibidean. Notazio zientifikoan dagoen *Fully Specified Name*a dauka eta organismoa deskribatzen duen *Acceptable Synonym* bat ere bai. Kasu gehienetan egitura hau errepikatzen da, gaztelaniazko bertsioak *Preferred Term*a galtzen duelarik.

Kontzeptu-identifikadorea: 64922006
FSN: *Pennisetum purpureum (organismo)*
Sinonimoa: *pasto de elefante*

3. adibidea: *Organism* hierarkiako PT gabeko termino bat

Gaztelaniazko *Semantic tag*a	**Agerpen kopurua**
organismo (organism)	23.037
producto (product)	1.787
calificador (qualifier)	821
sustancia (substance)	321
localización geográfica (geographic location)	271
estructura corporal (body structure)	265
hallazgo (finding)	22
estadificación tumoral (tumor staging)	13
trastorno (disorder)	12
objeto físico (physical object)	8
religión/filosofía (religion/philosophy)	6
procedimiento (procedure)	4
metadato fundacional (foundational core metadata)	4
grupo étnico (ethnic group)	4
entidad observable (observable entity)	1
concepto para navegación (navigational concept)	1
atributo (attribute)	1
escala de evaluación (assessment scale)	1
Denera	**26.579**

5. taula: Galdutako *Preferred Term*en *semantic tag*ak

*Preferred Term*az gain, 18.647 Kontzeptuk ez dute ez *Preferred Term*ik ez eta *Acceptable Synonym*ik. *Fully Specified Name*a daukate soilik. IHTSDO (2012) erakundeak dioen

bezala, *"typically the Fully Specified Name will not be a Term that would be used in a clinical record"*. Hau da, FSNak ez dira txosten klinikoetan agertzen diren terminoak izango, hortaz, *Preferred Term* zein *Acceptable Synonym*ak esleitzearen garrantzia areagotu egiten da.

3.1.5 *Fully Specified Name*n eta Sinonimoen arteko konparaketa

FSNak txosten klinikoetan ez direla agertzen argitu dugu aurreko 2.1 atalean. Orokorrean FSNak *Preferred Term*ari *semantic tag*a gehituaz osatzen dira. Gure helburu nagusia SNOMED CT euskaratzea denez, garrantzitsua deritzogu fenomeno honen maiztasuna neurtzea. Horrela, bi terminoak (FSN eta PT) itzultzea ekidin dezakegu. PTak bakarrik itzultzen baditugu, FSNa dagokion *semantic tag*a gehituz sor dezakegu.

*Semantic tag*ik gabeko FSNen agerpenak Sinonimoen artean neurtu ditugu 6. taulan.

FSN aurikitua non:	**Ingelesa**	**Gaztelania**
Preferred Term	372.874	317.706
Acceptable Synonym	20.746	6.981
Ez da agertzen	1.892	29.960
Denera	**395.512**	**354.647**

6. taula: FSN vs Sinonimoak

Ikus dezakegunez, ingelesezko bertsioan Sinonimoen artean agertzen ez diren FSNen kopurua arbuiagarria da, soilik guztien %0,5a. Hurrengo adibidean (4. adibidea) FSNaren bitartez ordezkatutako Kontzeptu bat ikus dezakegu, zeina ez den bere Sinonimoen artean agertzen.

Kontzeptuaren identifikadorea: 29605007
FSN: Speech fluency, function (observable entity)
Sinonimoa: Speech fluency

4. adibidea: FSN eta Sinonimo ezberdina duen termino bat

Gaztelaniazko bertsioak ez duela ingelesezko bertsioak adina sendotasunik egiazta daiteke, bere FSNen %8,4 baino gehiago ez baitira Sinonimoen artean agertzen. Dena dela, datu hauek ez dira harrigarriak, 3.1.4 atalean aztertu dugun bezala, 26.579 Kontzeptu baitaude *Preferred Term*ik ez dutenak.

3.1.6 Sinonimoen luzera hitz kopuruaren arabera

Atal honetan terminoen luzera aztertuko dugu, euren hitz kopurua kontuan izanda. Aurreko atalean ikusi dugun bezala (3.1.5 atala), Sinonimoen artean agertzen ez diren FSNen kopurua oso baxua da. Hau horrela izanik, atal honetan aztertuko ditugun terminoak Sinonimoei dagozkien terminoak soilik izango dira, hau da, ez ditugu *Fully Specified Name*ak kontuan hartuko.

HAP masterra

Sinonimoen deskribapen motari dagozkion terminoen luzeraren araberako kopuruak eta batezbestekoak erakusten ditu 7. taulak. Honakoak dira kontaketa egiteko erabili ditugun neurriak: hitz bakarreko terminoak, bi hitzetakoak, hirukoak, laukoak eta lau hitz baino gehiagokoak. Egitura sinpletzat hitz bakarrekoak, bikoak eta hirukoak kontsideratu ditugu, eta lau hitzekoak edota gehiagokoak, termino konplexuak.

Ataza hau garatzeko termino errepikatuak baztertu ditugu bi aldiz zenbatuak izan ez daitezen. Dena dela, baztertze hori egin gabe ere zenbatzea erabaki dugu, errepikaturik termino hauek anbiguoak izan baitaitezke. Ondorengo 5. adibidean anbiguotasun horren kasu bat ikus dezakegu.

Sinonimoa: *Eye*
1. **Deskribapenaren identifikadorea:** 135599015
   FSN: *Structure of eye proper (body structure)*
2. **Deskribapenaren identifikadorea:** 365613018
   FSN: *Entire eye (body structure)*

5. adibidea: Anbiguoa den Sinonimo bat

	Ingelesezko bertsioa				**Gaztelaniazko bertsioa**			
			Bakarra				**Bakarra**	
	#	%	#	%	#	%	#	%
Hitz bakarra	44.868	%7,24	35.376	%6,69	26.475	%5,93	22.118	%5,24
Bi hitz	146.735	%23,68	120.930	%22,86	58.598	%13,12	54.126	%12,81
Hiru hitz	126.203	%20,36	105.078	%19,87	61.567	%13,78	57.751	%13,67
Lau hitz	105.588	%17,04	89.613	%16,94	67.962	%15,21	65.305	%15,46
Hitz anitz	196.348	%31,68	177.928	%33,64	232.166	%51,96	223.161	%52,82
Denera	**619.742**	**%100**	**528.925**	**%100**	**446.768**	**%100**	**422.461**	**%100**

7. taula: Terminoen luzera

Ingelesezko bertsioari arreta jartzen badiogu, egitura oso sinplea duten terminoek (hitz bakarrekoak eta bikoak) termino guztien ia heren bat osatzen dute (%7,24 + %23,68). Gainera, hiru hitzeko terminoak gehitzen baditugu ingelesezko bertsioaren Sinonimoen erdia lortzen dugu.

Gaztelaniazko bertsioak ez ditu horren zenbaki itxaropentsuak eskaintzen, terminoen erdia baino gehiago hitz anitzekoen taldean sailkatzen baita, eta egitura sinplea duten terminoak (hitz bakarrekoak, bikoak eta hirukoak) osoaren herena baino ez dute osatzen.

Kontuan izan behar dugu SNOMED CTn karaktere berezien erabilera zabala dela. 93.449 termino aurkitu ditugu ingelesezko bertsioan hurrengo zerrendako karaktereren bat dutenak: “,”, “#”, “%”, “:”, “/”, “(”, “)”, “¿”, “¡” edo “^”. Termino hauek zenbatu ditugu itzulpenerako arazoak sor ditzaketela uste dugulako.

Gaztelaniazko bertsioan mota horretako termino gutxiago aurkitu dugu, 78.798 hain zuzen ere. Dena dela, Sinonimo gutxiago daudela kontuan hartzen badugu, ehunekotan portzentaia handiagoa da: %17a, ingelesezko bertsioaren %14aren aurrean.

HAP masterra

Terminoen luzeraren kopuruak izanik, 8. taulan hitz bakarrekoen agerpenak zenbatu ditugu gainerako luzeradun terminoen barruan. Itzulitako terminoen erabilerak termino luzeagoen itzulpenean nola lagun dezakeen aztertzeko egin dugu hori. Adibidez, *"Aspirin adverse reaction"* SNOMED CTren terminoaren barruan beste SNOMED CT termino baten agerpena dago: *"Aspirin"*.

	Ingelesezko bertsioa		**Gaztelaniazko bertsioa**	
		Bakarra		**Bakarra**
Bi hitzetako terminoak	164.764	95.608	83.758	57.469
Hiru hitzetako terminoak	172.387	92.399	156.301	107.329
Lau hitzetako terminoak	158.298	86.028	219.251	152.793
Hitz anitzeko terminoak	313.660	184.926	1.270.037	910.625
Inoiz ez	19.202	16.545	10.186	9.575
Denera	**809.109**	**458.961**	**1.729.347**	**1.228.216**

8. taula: Agerpen kopurua

Aurreko taulan agertzen diren kopuruen artean (8. taula), kontuan izan behar dugu SNOMED CTk termino gisa zenbaki zein karaktere bereziak definitu dituela, "¡=" edo "0.5" adibidez. Hauek emandako kopuruak puztu ahal izan ditu, baina hauek ere termino "normal" gisa kontsideratu behar ditugu, ziurrenik hauetan notazio aldaketak egin beharko baititugu. Adibidez, zenbaki errealen kasuan ingelesez puntua erabiltzen den bitartean ("0.5") euskaraz komarekin idazten dira ("0,5");

Aurreko taulako datuak aztertuz (8. taulakoak), hitz bakarreko terminoen agerpen totala gainerako terminoetan, gaztelaniazko bertsioan ingelesezko bertsioan baino askoz altuagoa da: ia bikoitza. Dena dela, gaztelaniazko bertsioaren datuak sakonki aztertuz gero, hitz bakarrekoen eragina batez ere hitz anitzeko terminoetan aurkitzen dela ikus daiteke, eta eragina oso baxua dela egitura sinplea duten terminoetan.

Ingelesezko bertsiotik jasotako ondorioak aldiz, aurretik aipatutakoaren ezberdina da. Jada aztertu dugun bezala, egitura sinplea duten terminoek osoaren %50a osatzen dute eta hitz bakarrekoen eragina termino hauetan esanguratsua da.

3.1.7 Gaztelaniazko bertsioaren Kontzeptuen gabezia

Atal honetan, 9. taularen bitartez gaztelaniazko bertsioan galdutako Kontzeptuei dagozkien *semantic tag*ak aztertu ditugu. 34 *semantic tag* ezberdin aurkitzen ditugu galdutako Kontzeptu hauei dagozkienak. Kopuruaren arabera aztertuz gero, *"procedure"*, *"disorder"* eta *"finding" semantic tag*ak ditugu gabezia handienarekin. Dena dela, hurrengo atalean ikusiko dugun bezala (3.1.8 atala), hiru *semantic tag* horiek dira Kontzeptu gehien dituztenak.

Ehunekoei begiratuz gero eta kopuru baxuak dituztenak baztertuz, *"situation"*, *"occupation"* eta *"regime/therapy" semantic tag*ak direla kaltetuenak esan dezakegu.

semantic tag	#	%	**semantic tag**	#	%
procedure	11.398	%16,17	ethnic group	83	%22,68
disorder	10.537	%11,31	specimen	69	%4,75
finding	8.534	%18,91	morphologic abnormality	54	%1,06
situation	2.931	%33,63	administrative concept	49	%61,25
occupation	1.792	%27,82	assessment scale	44	%3,99
regime/therapy	785	%22,05	special concept	29	%96,67
substance	652	%2,55	staging scale	25	%60,98
product	484	%1,99	tumor staging	13	%4,96
qualifier value	483	%4,81	attribute	10	%0,87
observable entity	408	%4,54	religion/philosophy	10	%4,41
physical object	381	%6,91	navigational concept	5	%0,69
event	377	%4,21	cell	3	%0,47
person	230	%34,74	physical force	3	%1,69
body structure	157	%0,58	racial group	2	%9,52
organism	127	%0,36	cell structure	1	%0,19
environment	90	%7,19	social concept	1	%3,70
record artifact	84	%26,42	*semantic tag* gabe	1.013	-
			Total	**40.864**	-

9. taula: Gaztelaniazko bertsioan galdutako Kontzeptuak

3.1.8 *Semantic tag*en populazioa

Azkeneko atalean *semantic tag* bakoitzaren termino kopurua jasotzen duen taula erakusten dugu (10. taula). Ikus daitekeen moduan, populazio altuena duten *semantic tag*ak Clinical Finding/disorder eta Procedure hierarkiei dagozkienak dira. Horrek FSN guztien %55a suposatzen du ingelesezko bertsioan, eta %51a gaztelaniazkoan.

	Ingelesezko bertsioa		Gaztelaniazko bertsioa	
Hierarkia	**Semantic Tag (ST)**	**FSN #**	**Semantic Tag (ST)**	**FSN #**
Clinical Finding/disorder	disorder	94.147	trastorno	82.630
	finding	45.362	hallazgo	36.586
Procedure/intervention	procedure	74.748	procedimiento	59.081
	regime/therapy	3.573	régimen/terapia	2
			régimen/tratamiento	2.773
Organism	organism	35.722	organismo	35.319
Body structure	body structure	26.942	estructura corporal	26.729
	morphologic abnormality	5.198	anomalía morfológica	5.021
	cell	644	célula	639
	cell structure	513	estructura celular	509
Substance	substance	25.824	sustancia	24.908
Pharmaceutical/biologic product	product	24.377	producto	23.852
Qualifier value	qualifier value	10.133	calificador	9.569
Observable entity	observable entity	9.027	entidad observable	8.585
Event	event	8.959	evento	8.587
Situation with explicit context	situation	8.715	situación	5.784
Social context	occupation	6.459	ocupación	4.649
	person	668	persona	432
	ethnic group	366	grupo étnico	283
	religion/philosophy	227	religión/filosofía	217
	life style	30	estilo de vida	25
	social concept	27	contexto social	26
	racial group	21	grupo racial	19
Physical object	physical object	5.134	objeto físico	4.733
Specimen	specimen	1.454	espécimen	1.385
Environment	environment	1.253	medio ambiente	1.162
geographical location	geographic location	619	localización geográfica	619
	environment/location	1	medio ambiente/localización	1
Linkage concept	attribute	1.157	atributo	1.145
	link assertion	8	relación asertiva	8
	linkage concept	1	concepto de enlace	1
Staging and scales	assessment scale	1.102	escala de evaluación	1.058
	tumor staging	261	estadificación tumoral	249
	staging scale	41	escala de estadificación	16
Special concept	navigational concept	731	concepto para navegación	724
	namespace concept	153	espacio de nombres	153
	administrative concept	80	concepto administrativo	31
	special concept	31	concepto especial	1
Record artifact	record artifact	318	elemento de registro	234
Physical force	physical force	178	fuerza física	174
Root Metadata Concept	foundation metadata concept	134	metadato fundacional	134
	core metadata concept	31	metadato del núcleo	31

10. taula: Ingelesezko eta gaztelaniazko bertsioen *semantic tag*ak.

3.1.9 Ondorioak

Aurreko ataletan aztertutako elementuak kontuan izanik eta egindako irakurketak kontuan izanik, atera dezakegun lehenengo ondorioa itzulpenerako jatorri bertsioa zein izango den: ingelesezko bertsioa.

3.1.6 atalean ikusi dugun bezala, ingelesezko bertsioak konplexutasun gutxiago erakusten du terminoen konplexutasunari dagokionean. Horrela egitura sinpleagoekin ontologiaren portzentaia altuagoa itzul dezakegu. Gainera, 3.1.1 atalean ikusi dugun bezala, gaztelaniazkoa ingelesezko bertsiotik sortua da, eta 3.1.7 eta 3.1.4 ataletan ikusi dugun moduan, horrek terminoen gaineko gabezia dakar. SNOMED CTren hierarkia aukeraketa egiterako garaian, 3.1.8 atalean ikusi dugun bezala *Clinical Finding/disorder* eta *Procedure* hierarkiak populatuenak izanik, hauek itzultzen hasi beharko ginateke. Irizpide honi jarraiki, *Organism* eta *Body structure* hierarkiak izango lirateke zerrendan hurrengoak. Dena dela, 3.1.4 atalean *Organism* hierarkiari buruz esan dugunaren harira, *Preferred Term*a ez itzultzearen irizpidea zehazten du IHTSDOk, baizik eta izen taxonomikoa erabiltzearena.

Itzulpenerako bertsio egokiena ingelesezkoa dela erabaki ondoren, itzulpen automatikorako baliabideak aztertzeari ekingo diogu, hauen euskarazko ordainak sortzeko asmoarekin.

Aurrera jo aurretik, azalpenak argiago egiteko ingelesezko zein gaztelaniazko terminoei jatorri-termino edo termino deituko diegun bitartean, euskarazkoei ordain deituko diegu.

3.2 Euskarazko ordainen sorkuntzarako Itzulpen Automatikoaren ekarpen eta gabeziak

Azterketa honetan medikuntzako domeinuan izen sintagmak itzultzeko Itzultzaile Automatikoak (IA) dituen gabeziak aztertu ditugu. Azterketa hau hurrengo atalean (4.1 atala) azaltzen dugun algoritmoa diseinatzeko oinarria izan da, gure helburua terminoen ordainak lortzea baita, IAk eskaintzen dizkigun abantailak baliatuz.

3.2.1 Atazaren aurkezpena

Gure hasierako planteamendua hierarkia hauen egitura aztertzea izan da, honen bitartez terminoen itzulpen automatikoa egiteko patroiak definitu asmoz. Hasieratik lan zama handia izango zela bagenekien ere, azterketarekin aurrera joan ahala, hilabeteak hartu zitzakeen ataza zela ohartu ginen.

Horren ondorioz, *"Finding"* hierarkiako 45.125 termino multzokatu ditugu, eta itzulpena era automatikoan egiteko erraztasun gehien dituzten multzoak aukeratu ditugu. Multzokatze hori eskuz egindako lana izan da eta egitura antzekoa duten terminoak elkartuz egin dugu.

Ataza honen zailtasunaz jabetzeko adibide bat aurkeztuko dugu jarraian (6. adibidea).

Termino taldea: *Adverse reaction to <X>*
<X>ren esanahaia: sendagai baten izena
Termino bat: *Adverse reaction to aspirin*
Terminoaren hierarkia: *Finding* (Aurkikuntza)
Gaztelaniazko ordaina: *Reacción adversa a la aspirina*

6. adibidea: Termino multzo baten adibidea

Termino honen gaztelaniazko baliokidea berehalakoa izan badaiteke ere (*"Reacción adversa a la aspirina"*), euskarazko baliokidea sortzeko hainbat aukera edota zalantza sortzen zaizkigu. Termino multzo hau *"Finding"* hierarkiari dagokio, eta hortaz terminoa sintoma bat da.

Aurreko adibidearekin jarraituz, 7. adibidean, informatikari, hizkuntzalari zein medikuen ekarpenetatik lortutako euskal ordain hautagaiak erakusten ditugu.

Terminoaren ordain hautagaiak: (*Adverse reaction to aspirin*)
1. Aspirinaren aurkako erreakzioa
2. Aspirinarekiko aurkako erreakzioa
3. Kontrako erreakzioa aspirinari
4. Aspirinaren albo ondorio kaltegarria
5. Aspirinaren erreakzio kaltegarria

7. adibidea: *Adverse reaction to aspirin* terminoaren ordain hautagaiak

Adibideko ordain hautagai bakoitzaren inguruan sortu zaizkigun kezka edo zalantzak aipatuko ditugu jarraian:

1. **Aspirinaren aurkako erreakzioa:** Aspirinaren ezaugarria dela dirudi, eta ez gaixoaren sintoma bat.
2. **Aspirinarekiko aurkako erreakzioa:** Determinatzaileekin arazo linguistikoak daude: -kiko -ko parea ez da egokia.
3. **Kontrako erreakzioa aspirinari:** Nominalizazioetan ez da datiboa erabiltzen, soilik genitibo eta lokatiboa erabiltzen dira.
4. **Aspirinaren albo ondorio kaltegarria:** "Albo ondorio"ren zuzentasun eta zehaztasuna zalantzagarria da.
5. **Aspirinaren erreakzio kaltegarria:** Jakintzat ematen da erreakzioa kaltegarria dela, baina terminoak berak ez du hori zehazten.

Adibide zehatz horretan, lehenengo aurkeztu dugun izen sintagma litzateke medikuak proposaturikoa (aspirinaren aurkako erreakzioa), eta linguistikoki zuzena dela ziurtatu digute hizkuntzalariek. Hortaz, gure proposamena berau litzateke *Preferred Term* gisa azaltzeko. Dena dela, aurretik aipaturiko SNOMED CTren egitura gure alde daukagu horrelako egoeretan, linguistikoki balizkoak badira nahi ditugun guztiak barnebiltzeko aukera ematen baitigu SNOMED CTk, Sinonimoen bitartez.

Hortaz, ikusi dugun bezala terminoen euskarazko baliokideen sorkuntza ez da berehalakoa eta aditu ezberdinen iritzia jaso behar dugu hauek ondo sortzeko. Domeinu honetarako adituak medikuak eta hizkuntzalariak izango dira.

Egitura antzekoa duten terminoen multzokatzea eskuz definitutako erregela batzuk proposatzeko egin genuen, gero hauek automatikoki aplikatzean lortutako emaitzen gaineko hausnarketa aurkezteko asmoz. Baina erregelak automatikoki aplikatzeko modua ez da berehalako lana, hainbat tresnen erabilera eskatzen baitigu: hiztegi elebidunak, sortzaile linguistikoa,... Hauek guztiak *Matxin* itzultzaile automatikoan barnebilduta daude Mayor et al. (2011). *Matxin* erregeletan oinarrituriko itzultzaile automatikoa da.

Hortaz, guk hutsetik sortutako tresnak definitzen hasi aurretik jadanik eskura ditugun tresnen emaitzak aztertu ditugu, eta gure esku egon daitezkeen hobekuntzak txertatzeko hastapenak prestatu ditugu.

Hortaz, egindako lanari probetxua ateratzeko, ingelesez dauden Kontzeptuetatik abiatu gara, hauek baitira multzokatuta ditugunak.

Alde batetik, IXA taldeak garatutako *Matxin* erregeletan oinarritutako itzultzaile automatikoa erabili dugu; eta bestetik *Google*ren *Translate* tresna erabili dugu, zeina estatistikan oinarritutako itzultzaile automatikoa den.

3.2.2 Estatistikan oinarritutako IA

*Google*ren *Translate* tresna interneten dauden testu paraleloetan oinarritzen da estatistika hutsaren bitartez itzulpen automatikoa egiteko. Hainbat hizkuntzen arteko itzulpenak

eskaintzen ditu eta euren artean ingelesa eta euskararen arteko itzulpenak eskaintzen ditu. *Google*ek ez du erabiltzen dituen teknikei buruz berri ematen, hortaz ez dakigu era garbian itzultzaile hauen funtzionamendua zein den.

Aipatzekoa da *Google*ek euskara bezalako hizkuntz gutxituekin darabilen politika: hizkuntza bakarrarentzat sortzen du itzultzailea, ingelesa gure kasuan, eta beste hizkuntzaren baten itzulpena egiteko, adibidez gaztelaniatik euskarara, lehenik ingelesera itzultzen du eta ostean ingelesezko itzulpena euskaratu egiten du. Hau horrela izanik, ingelesa ez den beste edozein hizkuntzarekin egindako itzulpenen kalitatea asko jaisten da, itzulpen erroreak metatu egiten baitira. Harrigarria da erro itzulpena ingelesa-euskara izatea, izan ere gure ezagutzak dio testu paralelo gehiago egongo direla gaztelania-euskara parerako ingelesa-euskararako baino.

Dena dela, euskararen itzultzailea oraindik *Alpha* bertsioan dago eta horregatik jarraian erakutsiko dugun bezala emaitzak ez dira oso erabilgarriak.

Aurkitu dugun tresna honen arazo nagusia deklinazioa da. Adibidez, *"Bleeding from hip"* terminoa itzultzeko, ez da "from" itzultzeko gai izan eta "Aldakako from hemorragia" izan da proposaturiko itzulpena. Ikus daitekeenez, tresnak ez du "from" preposizioaren itzulpena aurkitu euskararentzat, ez baita konturatu euskaraz hau deklinazioaren bitartez adierazitako hitza dela. Antzeko egoera aurkitu dugu "to" eta "at" preposizioekin.

Honetaz gain, itzultzaile honek beste eragozpen bat sortzen digu: guk izen sintagmak itzuli nahi ditugu, baina itzultzaileak esaldiak dira itzultzen dituenak. *"Born in Andorra"* izen sintagmaren kasuan "Andorran jaio zen" itzultzen du. Ikus daitekeenez esanahiean eta linguistikoki zuzena bada ere, izen sintagma izatetik esaldi bilakatu du.

Lehenago aipatu dugun estatistika hutsaren erabilerak beste motako ondorioak ere ekarri dizkigu: ordena. Nahiz eta euskara ordena libreko hizkuntza izan, askatasun hau sintagmen artean ematen den ezaugarria da, eta ez sintagma barruan. *Google Translate* tresnak ingelesezko ordena mantentzen du askotan sintagmaren barruan, euskararen ordena beste bat izanik. Adibidez, *"Blood group AB"* sintagmaren itzulpentzat "Odol-talde AB" ematen du, "AB odol-talde" izan beharrean.

Alabaina, dena ez da negatiboa. Sintagma batzuen itzulpenak esperotakoak baina hobeak izan dira 8. adibidean ikusi daitekeenez bezala.

Lack of <X> egitura duten terminoen itzulpenak:
Terminoa: *Lack of emotional response*
Ordaina: Erantzun emozionala ez izatea

Terminoa: *Lack of energy*
Ordaina: Energia eza

Terminoa: *Lack of exercise*
Ordaina: Ariketa eza

8. adibidea: *Lack of <X>* egitura duten terminoen itzulpenak estatistikan oinarritutako IAren bidez

Ikus daitekeenez, adibideko sintagmak dezente sinpleak badira ere, interesgarria da nola jatorri egitura berdinari, itzulpen egitura ezberdinak esleitzen dizkien ("ez izatea" eta "eza"). Adibideak ugariak dira, bai aldekoak eta bai kontrakoak, baita itzultzen jakin ez dituenak ere, eta ingelesez utzi dituenak ere.

3.2.3 Erregeletan oinarritutako IA

Matxin gaztelaniatik euskararako lehenengo Itzultzaile Automatikoa da, IXA taldeak garatua. Erregeletan oinarritutako Itzultzaile Automatikoa da, hau da, teoria linguistikoetan zein hiztunen erabileran oinarriturik definitu diren erregela finko batzuetan oinarrituta dago. Erregela hauek jatorrizko testua jasotzen dute, eta hobekien egokitzen zaion erregela multzoa aplikatuz, testuaren euskarazko bertsioa osatzen du *Matxin*ek.

Orain arte gaztelaniatik euskarako itzulpenak bakarrik egiten bazituen ere, berriki ingelesa-euskara itzulpen erregelak lantzen hasiak dira IXA taldean. Itzultzaile Automatikoa hastapen prozesuan badago ere oso tresna interesgarria da gure lanerako, erregela berriak definitzea ahalbidetzen digulako, beharrezkoak diren baliabide linguistikoak barnebilduta izanik.

Aurreko Itzultzaile Automatikoaren kasuan bezala (3.2.2 atala), hainbat egitura antzekoa duten termino multzoak tresna hau erabiliz itzuli ditugu, eta hauen emaitzak aztertu. Erregeletan oinarritutako IA izanik, egitura linguistiko berdina duten termino guztientzat, erregela berdinak aplikatuko dizkio. Hau abantaila handia da termino berriak sortzeko, talde osoaren egitura baita eztabaidatu beharko dena, eta ez banan banakako terminoak. Hau da, behin egitura definiturik, multzo osoarentzat baliagarria izango da.

Tamalez, sistema hastapenetan dagoenez, ezin diogu oraindik benetako zukua atera. Dena dela, zenbait egituratan emaitza interesgarriak eman dizkigu hurrengo adibideetan ikus daitekeen bezala (9. adibidea).

Terminoa: *Lack of exercise* **Ordaina:** Ariketaren falta **Terminoa:** *Loss of appetite* **Ordaina:** Jateko gogoaren galera **Terminoa:** *Pain in upper limb* **Ordaina:** Goiko gorputz-adarrean oinazea

9. adibidea: Egitura ezberdina duten terminoen itzulpenak erregeletan oinarritutako IAren bidez

Adibide hauetan ikus dezakegunaz gain, *Matxin*ek beste abantaila bat eskaintzen digu: hitza ezagutu ala ez, deklinatu egiten du. Medikuntza bezalako domeinu batean, hitz tekniko asko ez dira itzultzen, baina hizkuntzaren ezaugarri linguistikoetara moldatu behar dira. Hortaz, zenbaitetan hiztegian ez agertzearen gabeziari etekina atera ahal izango diogu.

HAP masterra

Aipatu dugun bezala, Itzulpen Automatikorako metodologia honek, gure beharretara egokituriko erregelak definitzea ahalbidetzen digu. 10 adibidean agertzen diren erregelen antzekoak definitu ahalko genituzke, formatu egoki bat emanaz (hauek definizio posible bat baino ez dira):

```
"Blister of <X> without infection" -> "Infekziorik gabeko baba <X>[INE]"
                                      edo "Baba <X>[INE], infekziorik gabe"
                                      edo "..."
```

10. adibidea: Erregelen definizioak

Erregela hauetan ere, aukera bat baino gehiago definitu ditzakegu, sinonimo bezala erabili ahal izateko SNOMED-CT egiturari jarraiki.

3.2.4 Ondorioak

Itzultzaile Automatiko hauen gaineko proba lan honetan eragin zuzena izan du jatorri hizkuntzak. Ingelesezko Kontzeptuak izanik itzulitakoak, ez dugu Kontzeptu hauen baliokideak diren gaztelaniazko terminoekin probarik egin. Azken horiek eskura ditugu eta interesgarria litzateke hauekin ere probaren bat egitea. Alabaina, SNOMED CTren gaztelaniazko bertsioaren terminoen egitura ez da ingelesezkoa bezain sendoa 3.1 atalean ikusi dugun moduan.

Erregeletan oinarritutako Itzultzaile Automatikoak (*Matxin*), domeinuari egokitzea ahalbidetzen digu, (Alegria et al., 2011) artikuluan erakusten den bezala. Izan ere domeinu orokorreko hainbat balizko hitz eta izen sintagma medikuntzaren domeinura aplikatuz gero, zuzentasuna galdu egiten da. Horregatik, *Matxin* domeinuari egokitzeko urratsak eman beharko ditugu, besteak beste, *Elhuyar*ren Zientzia eta Teknologia Hiztegia integratuz.

Google Translate tresnak aldiz, termino proposamenak sortzeko ataza honetan ez dirudi gehiegi lagun diezagukeenik. Izen sintagma barruko hitz solteak itzultzeko hainbat hastarna ematen badizkigu ere, izen sintagma bere osotasunean itzultzeko arazoak ditu. Hortaz, adituari zenbait kasu konkretutan laguntzeko erabili ahal izango dugu tresna hau, inoiz ere ez automatikoki itzulpenak proposatzeko.

Ondorio orokor gisa esan dezakegu, SNOMED CTren terminoen egitura dela medio, itzultzaile automatikoak erabili aurretik termino sinpleak beste metodo batzuk erabiliz itzultzeko aukera aztertu beharko dugula, eta azken aukera gisa Itzultzaile Automatikoen itzulpenak erabili. Izan ere *Matxin*ekin lan egiten hasi aurretik honen ingeles-euskara bertsioa findu beharko da, oraindik orain, hastapenetan baitago. Gainera, hiztegi espezializatuak eskura ditugu terminoen euskaratzearekin hasteko.

Itzultzaile Automatikoen azterketa egin ostean, SNOMED CT eta GNS-10ren arteko mapaketaren analisia egingo dugu hurrengo atalean.

3.3 SNOMED CT eta GNS-10en arteko mapaketaren analisia

SNOMED CTren eta GNS-10en arteko mapaketa aurkeztu dugu dagoeneko 2.2 atalean. Oraingoan, mapaketan sakondu egingo dugu, eskaintzen duen informazioa aztertuz eta gure lanerako baliagarria zaiguna aukeratuz.

3.3.1 Mapaketaren ezaugarriak

Mapaketak jatorritzat SNOMED CT Kontzeptua eta helburutzat GNS-10 kodeak hartzen ditu, baina horretaz gain, parekatze bakoitzerako hainbat informazio gehigarri eskaintzen du. Besteak beste, mapaketaren nomenklaturari jarraiki: *mapGroup*, *mapPriority*, *mapCategory* eta *mapRule* atributuak zehazten ditu.

- *mapGroup*: jatorri Kontzeptua (SNOMED CT) zehazteko beharrezkoak diren parekatzeak multzokatzeko erabiltzen den zenbakia da. 1 eta 4 artean egoten da zenbakia. 11. adibidean bi *mapGroup* dituen Kontzeptu baten informazioa ikus dezakegu. Kasu honetan, SNOMED CT Kontzeptuak era agerian (AND baten bitartez) bi kontzeptu ezberdin elkartzen ditu bakar batean. GNS-10en aldiz kontzeptu hauek bi kode ezberdinetan errepresentatuta daude. Mota horretako Kontzeptuek ez dituzte beti bi kontzeptu ezberdin horren agerian adierazten, baina argiago ikusten da modu honetan *mapGroup*en eginkizuna.

SNOMED CT kodea: 9859006
SNOMED CT Kontzeptua: *Insulin-resistant diabetes mellitus AND acanthosis nigricans (disorder)*
***mapGroup:* 1**
GNS-10 kodea: `E10.9`
GSN-10 terminoa: *Insulin-dependent diabetes mellitus (with modifiers)*
***mapGroup:* 2**
GNS-10 kodea: `L83`
GSN-10 terminoa: *Acanthosis nigricans*

11. adibidea: SNOMED CT eta GNS-10en arteko mapaketan, bi *mapGroup* dituen Kontzeptua.

Mapaketa osotasunean aztertuz eta *mapGroup*en kantitateak kontuan hartuz, 8 kasutan SNOMED CT Kontzeptuak 4 *mapGroup* ezberdin ditu, 111 kasutan 3 *mapGroup*, 2.468 kasutan 2 eta 21.984 kasutan *mapGroup* bakarra.

- *mapPriority*: zenbaki honek parekatzeak exekuzio-garaian *mapGroup* bakoitzaren barnean zein ordenatan egikarituko diren adierazten du. Parekatzearen baldintzak betetzen badira, lehenengo parekatze hori bakarrik hartuko da kontuan. Kasu honetan balioak 1etik 18raino doaz. Orokorrean baldintzak dituzten parekatzeak joaten

dira aurrena, gerora kasu orokorra agertzen delarik (12. adibidea), baina kasu gehienetan parekatze bakarreko *mapGroup*ak aurkitzen dira. Kasu honetan balioak 1etik 18raino doaz.

SNOMED CT kodea: 317006
SNOMED CT Kontzeptua: *Reactive hypoglycemia (disorder)*
***mapPriority:* 1**
GNS-10 kodea: `K91.1`
GSN-10 terminoa: *Postgastric surgery syndromes*
Oharra: `IF LATE DUMPING SYNDROME CHOOSE K91.1`
***mapPriority:* 2**
GNS-10 kodea: `E16.1`
GSN-10 terminoa: *Other hypoglycaemia*
Oharra: `ALWAYS E16.1`

12. adibidea: SNOMED CT eta GNS-10en arteko mapaketan, *mapGroup* bakarra bi parekatzeekin.

Bi parekatze dituen *mapGroup* bat duen Kontzeptu bat ikus dezakegu 12. adibidean. Kasu horretan, lehentasun altuena duen parekatzea *"Postgastric surgery syndromes"* GNS-10 terminoari dagokiona da. Mapaketak parekatze hau betetzen den ala ez aztertuko du aurrerago azalduko dugun *mapRule*ren bitartez. Parekatzearen baldintza betetzekotan `K93.1` kodea esleituko dio Kontzeptuari, eta ezezko kasuan, hurrengo parekatzea aztertuko du. Kasu honetan baldintza beti egiazkoa izango denez `E16.1` kodea esleituko litzateke.

*mapPriority*k ematen dituen zenbakietatik, *mapGroup* ezberdinen tamaina ondorioztatu daiteke, 11. taulan ikus dezakegun moduan. Oso *mapGroup* handiak egon arren, ez dira oso ugariak eta orokorrean SNOMED CT Kontzeptu batek parekatze bat ala bi izaten dituela ondoriozta daiteke.

Parekatze kopurua	Agerpen kopurua	Parekatze kopurua	Agerpen kopurua
1	18.113	7	69
2	1.544	8	36
3	660	9	27
4	327	10	13
5	180	11	3
6	103	18	1

11. taula: *mapGroup* ezberdinen tamainen kopuruak

- *mapCategory*: atal honek parekatzearen egoera adierazten du, edizioko egoera ere barnebiltzen duelarik. Dokumentazioan egoera gehiago aurkitzen badira ere, honakoak dira argitaraturiko azken bertsioan agertu diren egoerak, eta berauek ulertzeko adibideak:

– *Map of source concept is context dependant.* Parekatzea testuinguruaren menpe egotea: 5.961 kasutan. Ikus dezagun 13. adibidea. Kontzeptu honi GNS-10 kode bat esleitu ahal izateko informazio gehigarria behar dugu. Kasu honetan GNS-10k *Rheumatic mitral stenosis with regurgitation (disorder)* gaixotasunari lotuta bi kode esleituta ditu. Lehenengoak beste gaixotasun batekin batera ematen den kasuetan esleitzen du `I08.0` kodea, eta bigarrenak gaixotasun hau bakarrik agertzen den kasuetan. Oharretan ikusten da adibidea `MAP OF SOURCE CONCEPT IS CONTEXT DEPENDENT` kategoriakoa dela.

SNOMED CT kodea: 787001
SNOMED CT Kontzeptua: *Rheumatic mitral stenosis with regurgitation (disorder)*
***mapPriority:* 1**
GNS-10 kodea: `I08.0`
GSN-10 terminoa: *Disorders of both mitral and aortic valves*
Oharra: `IF RHEUMATIC MITRAL VALVE STENOSIS AND AORTIC VALVE INSUFFICIENCY CHOOSE I08.0 | MAP OF SOURCE CONCEPT IS CONTEXT DEPENDENT`
***mapPriority:* 2**
GNS-10 kodea: `I05.2`
GNS-10 terminoa: *Mitral stenosis with insufficiency*

13. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, testuinguruaren menpe

– *Map source concept cannot be classified with available data.* Eskuragarri dagoen informazioa parekatzea egiteko nahikoa ez izatea: 1.031 kasutan. 14. adibidean SNOMED CT kodeak ez dauka informazio nahikorik GNS-10ren kode bat esleitzeko. Kontzeptu honen umeak ez daude erabat parekatuta, eta hortaz informazio falta dago parekatzea burutzeko.

SNOMED CT kodea: 219006
SNOMED CT Kontzeptua: *Current drinker of alcohol (finding)*
Oharra: `DESCENDANTS NOT EXHAUSTIVELY MAPPED | MAP SOURCE CONCEPT CANNOT BE CLASSIFIED WITH AVAILABLE DATA`

14. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, informazio faltaren adibidea.

– *Mapping guidance from WHO is ambiguous.* WHOk zehaztutako irizpideak anbiguoak izatea: 3 kasutan bakarrik eman da egoera hau mapaketan. Kasu horietako bat da 15. adibidean erakusten duguna. Ikus daitekeenez, umeak erabat parekatuta ez daudela esateaz gain, ez du inolako informazio gehigarririk eskaintzen.
– *Source SNOMED concept is ambiguous.* SNOMED CT Kontzeptua anbiguoa izatea: 283 kasutan. Adibidean ikus dezakegunez (16. adibidea), SNOMED CT Kontzeptuak ez du inolako zehaztapenik ematen (*"Reduced mobility" (finding)*) eta Kontzeptua anbiguotzat jo da.

HAP masterra

SNOMED CT kodea: 31574009
SNOMED CT Kontzeptua: *Systolic murmur (finding)*
Oharra: `DESCENDANTS NOT EXHAUSTIVELY MAPPED | MAPPING GUIDANCE FROM WHO IS AMBIGUOUS`

15. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, irizpide anbiguoen adibidea

SNOMED CT kodea: 8510008
SNOMED CT Kontzeptua: *Reduced mobility (finding)*
Oharra: `SOURCE SNOMED CONCEPT IS AMBIGUOUS`

16. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, Kontzeptu anbiguo baten adibidea.

- *Map source concept is properly classified*. Egokiro sailkatua egotea: 19.888 kasutan. Maizen gertatzen den egoera da hau eta parekatzea erabat sailkatutzat ematen da. 17. adibidean parekatze zuzen bat erakusten dugu, eta terminoa berdinaren bitartez ordezkatzen dituzte SNOMED CT Kontzeptu eta GNS-10 kodea.

SNOMED CT kodea: 36348003
SNOMED CT Kontzeptua: *Primary hyperparathyroidism (disorder)*
GNS-10 kodea: `E21.0`
GSN-10 terminoa: *Primary hyperparathyroidism*
Oharra: `MAP SOURCE CONCEPT IS PROPERLY CLASSIFIED`

17. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, egokiro sailkatutako adibidea.

- *mapRule*: adierazpen honek parekatzea gauzatzeko baldintza zehazten du. Adierazpenak 'egiazko' ala 'faltsu' balioa jasoko du baldintza exekuzio-garaian ebaluatu ostean. Soilik 'egiazko' balioa hartzen duen kasuetan egingo da parekatzea.

  Hurrengo hiru forma ezberdinetako bat jaso dezake *mapRule* adierazpenak:

  - `TRUE`: parekatzerako baldintza zehatzik behar ez denetan, *mapRule* adierazpenak jasotzen duen forma da. Kasu honetan parekatzea beti gauzatuko da (18. adibidea).
  - `IFA SCTID | FULLY SPECIFIED NAME | [=VALUE]`: honakoa parekatzea gauzatzeko bete beharreko baldintza zehazteko forma da. Orokorrean testuinguruan agertzen diren beste SNOMED CT Kontzeptuen agerpenaz galdetzen du, eta berau irakurgarria egiten duen FSNa ere agertzen da. Balio posibleen agerpena beharrezkoa balitz, azken atalean hauen agerpenerako tartea dago.
  - `OTHERWISE TRUE`: azken *mapPriority* zenbakia duen parekatzearen *mapRule* adierazpena izan ohi da, aurreko parekatzeetako baldintzak betetzen ez diren kasuetan betetzeko.

HAP masterra

SNOMED CT kodea: 297009
SNOMED CT Kontzeptua: *Acute myringitis (disorder)*
GNS-10 kodea: `H73.0`
GSN-10 terminoa: *Acute myringitis*
***mapRule*:** `TRUE`

18. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, egokiro sailkatutako adibidea.

Aurreko bi formak erakusten dituen adibidea erakusten dugu jarraian (19. adibidea). Hiru parekatze dituen SNOMED CT Kontzeptu bat da. Lehenengo parekatzearen baldintza alkoholismoa izatea da. Alkoholismoa balu, `F10.2` GNS-10 kodea esleituko litzateke, eta ez balu, hurrengo parekatzea aztertuko litzateke. Bigarren parekatzearen baldintza absenta edariari menpekotasuna izatea da. Kasu honetan, baiezkoa balitz lehenengo baldintzaren kode berdina jasoko luke (`F10.2`) eta ezezkoan, azken parekatzera joko genuke. Horretan, `OTHERWISE TRUE` adierazpenaren bitartez egiazkotzat hartuko genuke, eta `F19.2` kodearekin egingo genuke parekatzea.

SNOMED CT kodea: 2403008
SNOMED CT Kontzeptua: *Psychoactive substance dependence (disorder)*
***mapPriority:* 1**
GNS-10 kodea: `F10.2`
GSN-10 terminoa: *Mental and behavioural disorders due to use of alcohol (with modifiers)*
***mapRule*:** `IFA 7200002 | Alcoholism (disorder) |`
***mapPriority:* 2**
GNS-10 kodea: `F10.2`
GSN-10 terminoa: *Mental and behavioural disorders due to use of alcohol (with modifiers)*
***mapRule*:** `IFA 231467000 | Absinthe addiction (disorder) |`
***mapPriority:* 3**
GNS-10 kodea: `F19.2`
GSN-10 terminoa: *Mental and behavioural disorders due to multiple drug use and use of other psychoactive substances (with modifiers)*
***mapRule*:** `OTHERWISE TRUE`

19. adibidea: SNOMED CT eta GNS-10ren arteko mapaketan, egokiro sailkatutako adibidea.

Mapaketa hau eskuzko lan sakon baten ondorioz gauzatu bada ere, garatzaileek honen errore tasa %3,3 eta %6,4 artean kokatzen dute.

HAP masterra

3.3.2 Ondorioak

Ikusi dugun moduan, parekatzeak ez dira kasu guztietan bat batekoak. Hau guztia kontuan izanik, eta gure atazaren helburuei jarraiki, garatuko dugun sistemarako baliokide gisa har daitezkeen parekatzeak bakarrik izan ditugu kontuan. Horretarako honako irizpide guztiak betetzen dituzten parekatzeak erabiliko ditugu:

- ***mapGroup*:** *mapGroup* bakarra dutenak.
- ***mapPriority*:** lehentasun altuena dutenak.
- ***mapCategory*:** egokiro sailkatuta daudenak, *Map source concept is properly classified* balioa dutenak.
- ***mapRule*:** baldintzarik ez dutenak, `TRUE` balioa dutenak.

Irizpide hauei jarraiki, 15.099 parekatze ditugu erabilgarri, "zuzeneko" parekatze gisa kontsideratu daitezkeenak.

Parekatze hauek, eta 2. atalean aipatutako hiztegi espezializatuak, SNOMED CTren ingelesezko bertsioa euskaratzen hasteko erabiliko ditugu, Itzultzaile Automatikoek oraindik dituzten galerak ikusi ostean. Euskaratzerako, hau da, euskal ordainak lortzeko, diseinatutako aplikazioaren atal nagusiak azalduko ditugu jarraian (4. atala).

4 Diseinua

Aplikazioaren diseinuarekin zuzenki lotuta dauden hiru atal aztertuko ditugu hurrengo lerroetan. Lehenik aplikaziorako diseinatu dugun algoritmoa azalduko dugu. Jarraian, *TermBase eXchange* (TBX) formatuaren egokitzapena deskribatuko dugu. Eta azkenik, aplikazioaren eskakizunetara egokitzen den klase-diagrama aurkeztuko dugu.

4.1 Aplikazioaren algoritmoa

Atal honetan SNOMED CTren terminoak euskaratzeko helburuarekin definitu dugun algoritmoa deskribatzen da.

Gogoan izan behar dugu, SNOMED CT Kontzeptuetan oinarritzen dela, eta hauek errepresentatzen dituzten terminoak direla guk euskaratu nahi ditugunak.

Azken finean, algoritmo honen bitartez, euskal ordainen sorkuntza edota itzulpena deskribatuko dugu, gerora SNOMED CTren euskaratzeko aplikazioan integratuko dena.

4.1.1 Algoritmoaren testuingurua

Euskal ordainen sorkuntzarako algoritmo honek ez du terminoaren jatorri-hizkuntza kontuan hartzen, ez eta ordainaren egokitasuna ere. Bi ataza horietaz aplikazioaren beste atal bat arduratu beharko da.

Aplikazioak ingelesezko nahiz gaztelaniazko baliabideak jasota, terminoaren jatorri-hizkuntza jakinik baliabide egokienak igorri eta abiatu beharko ditu algoritmoaren funtzionamendua zuzena izan dadin. Hortaz, jarraian erakusten dugun algoritmoan ez zaio jatorri-hizkuntzari inolako erreferentziarik egingo.

Hizkuntzaz gain, kontuan izan behar dugu sorturiko euskarazko ordainak, ordain-hautagaiak izango direla. Hauen zuzentasuna adituen esku gelditzuko bada ere (medikuak zein hizkuntzalariak), aplikazioak gure esku dagoen guztia ahalbidetu beharko du. Tartean, ordain-hautagaiaren erabilpena corpus espezializatuan aztertu beharko du aplikazioak, eta honen arabera konfiantza maila ezberdina emango zaio ordain-hautagaiari. Honetaz gain, adituei ordain-hautagaia balioztatzea eskatuko zaienean, honen agerpenen adibideak corpusetik erauzi beharko ditu.

Gainera, SNOMED CTren Kontzeptu bakoitzaren zein termino itzuliko diren erabaki beharko du aplikazioak, baita hauen ordainen errepikapenak egiaztatu ere. Izan ere, sinonimo asko dituen Kontzeptu batek euskal ordain errepikatuak sor ditzake.

Erabilgarri ditugun baliabideak kontuan izanda, hainbat terminoetarako euren definizioa ere, eskura izango dugu. Definizio hauekin zer egingo dugun ere erabaki beharko da. Izan ere, nahiz eta SNOMED CTn horretarako tokirik ez izan, oso interesgarria izan daiteke definizio hauek SNOMED CTrekin lotzea.

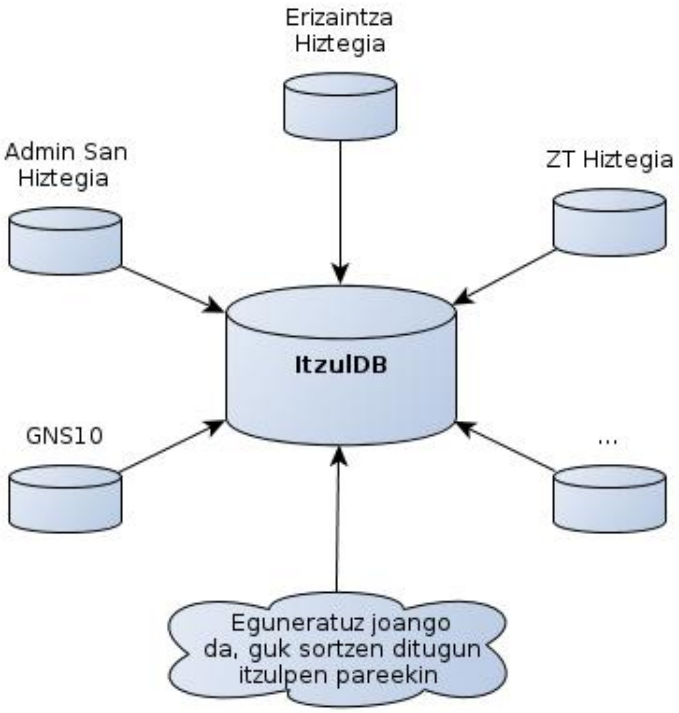
Algoritmoa deskribatzen hasi aurretik, kontuan hartu behar dugu algoritmoak sarrera gisa termino bat jasoko duela eta horren ordaina lortzen saiatuko dela. SNOMED CT eta GNS-10ren arteko mapaketak aldiz Kontzeptu mailan egiten du lan. Hau horrela

izanik, algoritmoa martxan jarri aurretik SNOMED CT eta GNS-10ren arteko mapaketa eta ordainen esleipena eginda egongo da.

4.1.2 Algoritmoaren deskribapena

Hurrengo lerroetan definitzen den algoritmoak, ingelesezko edota gaztelaniazko jatorrizko terminoak euskaratzea du helburu. Zenbaitetan terminoaren euskal ordain bakarra lortuko bada ere, algoritmoa ordain ezberdinen zerrenda itzultzeko diseinatuta dago, baliabide ezberdinetatik elikatuz euskal ordain egokiena sortzeko.

Algoritmoan erabiliko ditugun baliabideak azalduko ditugu aurrena. *ItzulDB* oinarrian TBX XML estandarra duen datu-basea da eta 2. irudian ikusten dugun moduan eratuta egongo da. Bertan osasun arloko hainbat hiztegi eta Gaixotasunen Nazioarteko Sailkapenaren 10. bertsioaren (GNS10) euskarazko itzulpena egongo dira integraturik. Dena dela, egitura honek *ItzulDB*ren hasierako egoera deskribatzen du bakarrik, algoritmoa aurrera egin ahala osatuz joango baita, sortzen diren itzulpen-pare berriak txertatuz joango baikara. Itzulpen pare berri hauek atalase bat gainditzen duten ordain-hautagaiak izango dira jatorri-terminoarekin parea osatzen dutelarik. Adibidez, *"abortus"* jatorri-terminoaren ordain-hautagaiak "abortu", "abortatze", "haur-galtze" eta "hilaurtze" izango lirateke, guztien artean parekatze bakarra osatzen dutelarik.



2. irudia: *ItzulDB*ren eskema

Itzulpen pareez gain, hainbat gramatika erabili beharko ditugu. *MorDB*n maila morfo-

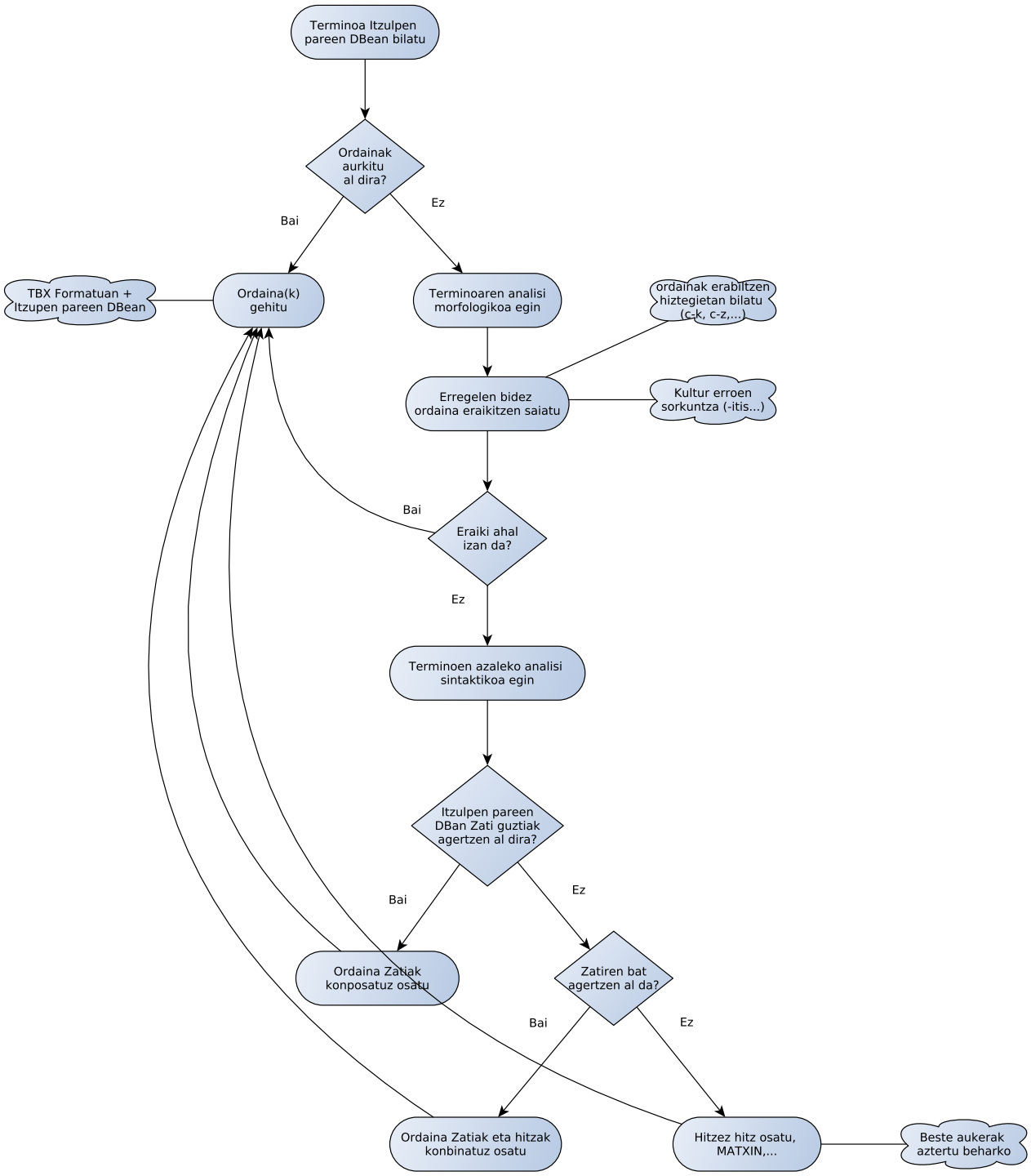
logikoan definituko ditugun sorkuntza-patroien datu-basea izango da. Bertan kultur erro sorkuntza zein idazketa aldaketa-patroiak egongo dira definiturik. *ZatDB* datu-basean aldiz, sintaxi mailako patroiak egongo dira definiturik.

Jarraian algoritmoaren sasi-kodea ikus dezakegu, eta 3. irudian honen errepresentazio grafikoa.

Sarrera: Terminoa
TERM → Itzuli beharreko terminoa (sarrera)
MorDB → Maila morfologikoko sorkuntza patroien DBa
ZatDB → Zati mailako orkuntza patroien DBa
ItzulDB → Itzulpen pareen DBa
Irteera: Terminoaren euskal ordaina(k)
EusOr → Sarrera terminoaren loturik dauden euskal ordainen zerrenda

```
 1: hasiera
 2: baldin TERM ∈ ItzulDB orduan
 3:    EusOr ← ItzulDB ordaina(k)
       ▷ItzulDBn sarrera bakoitzak konfiantza pisu bat edukiko du◁
       ▷Hiztegiek konfiantza pisu altuena izango dute◁
 4:    bueltatu EusOr
 5: amaiera baldin
 6: baldin EusOr hutsa orduan
 7:    TERMen analisi morfologikoa egin
 8:    baldin MorDBko patroiren bat betetzen bada orduan
 9:       EusOr ← patroien bidezko sorkuntza
          ▷c-k, c-z... trukaketa, kultur erroen sorkuntza (-itis...)...◁
10:       baldin EusOr sortu ahal izan bada orduan
11:          bueltatu EusOr
12:       amaiera baldin
13:    amaiera baldin
14:    baldin PatDBko patroirik ez edo EusOr ezin izan bada sortu orduan
15:       TERMen azaleko analisi sintaktikoa egin
16:       baldin Zati guztiak ∈ ItzulDB orduan
17:          EusOr ← ZatDB erregelak
18:          bueltatu EusOr
19:       bestela
20:          EusOr ← MATXINMed ordaina jaso
             ▷MATXIN Medikuntzarako moldatu beharko da, hiztegia, erregela bereziak...◁
21:          bueltatu EusOr
22:       amaiera baldin
23:    amaiera baldin
24: amaiera baldin
25: amaiera
```

1. algoritmoa: SNOMED CT terminoak euskaratzeko algoritmoa.



3. irudia: Algoritmoaren eskema

4.1.3 Adibideak

Deskribatutako algoritmoa hobeto ulertzen laguntzeko hainbat adibide aurkeztuko ditugu jarraian, adibide bakoitzean algoritmoaren kasu bat irudikatzen delarik.

1. ***ItzulDB*n agertzen den ordaina**

   Adibide hau itzulpen algoritmoaren kasurik sinpleena da. Terminoa hiztegietan agertzen denez, *ItzulDB*n gordeta dauden ordain ezberdinak jasoaz, euskal ordaina lortuko genuke.

Sarrera terminoa: *Deoxyribonuclic acid* **Algoritmoko pausuak:** 1-11 **Irteerako ordainak:** Azido desoxirribonukleiko, ADN, DNA

20. adibidea: *ItzulDB*n agertzen den ordaina

2. **Sorkuntza-erregelekin sortzen den ordaina**

   Termino hau ez da *ItzulDB* aurkitzen, hortaz terminoaren analisi morfologikoa eginik, kultur erro baliokidetza, eta aldaketa ortografikoak aplikatuko lirateke, irteerako euskal terminoa osatuaz. Kontuan izan beharko dugu, hemendik sortzen diren euskal ordainek euskarazko osasun-arloko corpusean duten agerpena aztertu beharko dela, honi konfiantza maila egokia esleitu ahal izateko.

Sarrera terminoa: *Photodermatitis* **Algoritmoko pausuak:** 1-4, 12-17 Aplikaturiko erregelak: *Ortografikoa:* ph ← f *Kultur erro baliokidetza:* -itis ← -itis(a) **Irteerako ordainak:** Fotodermatitis

21. adibidea: Sorkuntza-erregelekin sortzen den ordaina

3. **Zati mailako sorkuntza-erregelez sortzen den ordaina**

   Adibide honetan, terminoa hiztegietan ez agertzeaz gain, maila morfologikoan definitutako sorkuntza patroirik ez du betetzen. Dena dela, terminoaren zati ezberdinak *ItzulDB*n gordeta dauden itzulpen-pareen artean aurkitzen dira eta zati mailako sorkuntza-erregelak aplikatu daitezke. Hortaz, zatiak elkartzeko erregelak aplikaturik ordain ezberdinak sortuko lirateke.

   Algoritmoaren kasu honetan hainbat egoera ezberdin eman daitezke *ItzulDB*n gordeta dauden zatien arabera. Izan ere, algoritmoaren estrategia zatirik luzeenaren itzulpena jasotzea izango da, eta hau posible izango ez balitz, zati txikiagoak hartuz joango litzateke, terminoaren elementu guztiak itzuli arte.

HAP masterra

Sarrera terminoa: *Deoxyribonucleic acid sample*
Algoritmoko pausuak: 1-4, 12-13, 20-24
*ItzulDB*n agertzen diren zatiak:
1. zatia: Deoxyribonucleic acid
2. zatia: sample
Irteerako ordainak: Azido desoxirribonukleiko(aren)? lagin, ADN lagin, DNA lagin

22. adibidea: Zati mailako sorkuntza-erregelez sortzen den ordaina

Hurrengo adibidean (23. adibidea) erregela hauen beste kasu bat erakusten da, zeinetan zati osoaren itzulpen parerik ez dagoen, baina zati txikiago batzuenak bai.

Sarrera terminoa: Photodermatitis due to sun
Algoritmoko pausuak: 1-4, 12-13, 20-21, 25-27
Erregelak aplikatzeko elementuak:
1. zatia: Photodermatitis
2. zatia: due to <X>
hitza: sun
Irteerako ordainak: Eguzkiaren ondorioz sorturiko fotodermatitis

23. adibidea: Zati mailako sorkuntza-erregeleze sortzen den beste ordain bat

4. ***MatxinMed*en bidez sortzen den ordaina**

Gainerako urratsen bidez terminoa sortzea lortzen ez badugu, *Matxin* erabil dezakegu. Honetarako prestatu egin beharko dugu, erregela batzuk gehitzeaz gain, osasun-alorreko hiztegiak ere sartuaz. Hurrengo adibidean gaur egun *Alpha* bertsioan dagoen *Matxin* erabiliz lorturiko irteera erakusten da, inolako aldaketarik oraindik egin gabe.

Sarrera terminoa: *Partial excision of oesophagus and interposition of colon*
Algoritmoko pausuak: 1-4, 12-13, 20-21, 28-30
Irteerako ordainak: Esofagoaren zati baten excisiona eta interpositiona bi puntua

24. adibidea: *MatxinMed*en bidez sortzen den ordaina

Lexikoa, jatorri-termino zein ordainak, XMLz biltegiratuko ditugu, eta horretaz arituko gara hurrengo atalean.

4.2 Aplikaziorako egokitutako TBX formatua

***T**erm**B**ase e**X**change* (TBX) *e**X**tensible **M**arkup **L**anguage*a (XML) oinarritutako estandar ireki bat da informazio terminologikoa egituratzeko erabiltzen dena. Datu terminologikoa darabilten prozesu ezberdinak jasotzeko diseinatuta dago, hala nola analisiak edo errepresentazio deskribatzailea. TBXren eginkizun nagusia datu terminologikoa trukatzea dela esan daiteke. Etiketatze eredu unibertsala eskainiaz, enpresa eta erakunde ezberdinek ez dute TBX euren barne datu terminologikoa kudeatzeko bakarrik erabiltzen, elkarren arteko truke eta banaketarako ere erabiltzen dute.

TBXk datu-kategorien bitartez datu terminologikoa egituratzen du. Datu-base terminologikoetan datu-kategoriak kudeatzeko TBXk bi modulu ezberdin eskaintzen ditu, biak XMLz zehaztuta. Lehenengoak erabiltzaileari oinarrizko egitura zehazten ahalbidetzen dion bitartean, bigarrenak datu-kategorien gaineko murrizketak zehazteko eta identifikatzeko formalismoa eskaintzen du. Lehenetsitako datu-kategoria multzo bat eskaintzen du TBX, era honetan erabiltzaileak banatzaileei kontsultatu gabe datuak interpretatzeko aukera dauka. Dena dela, TBX formatua malgua da eta erabiltzaile talde bakoitzak, bere eskakizunak kontuan hartuta, datu-kategoria propioak defini ditzake, haien beharretara egokitutako TML (***T**erminological **M**arkup **L**anguage*) bat definituz.

Gure kasuan 2.4 atalean azaldutako TZOS sistemarentzat egokitutako TMLaren interpretazio bat egin dugu gure datu terminologikoak egituratzeko.

Gure jatorrizko formatua ez da TBX hutsa izan, baizik eta TZOS sistemarentzat egokitutako formatua (2.4 atala). Alde batetik SNOMED CTren errepresentaziorako definitutako TBX formatua azalduko dugu (4.2.1 atala). Bestetik, 4.2.2 atalean *ItzulDB*ren egitura sinpleagora egokitzen den formatuaren berri emango dugu.

Izan ere, SNOMED CT euskaratzeko datu-base batean gordeko dugu. Hortaz bi datu-base izango ditugu: SNOMED CTrena eta itzulpen-pareena (*ItzulDB*). SNOMED CTren datu-basean euskaratzerako beharrezko den Kontzeptuetako informazioa egongo da gordeta, tartean ingelesezko zein gaztelaniazko Sinonimoak. Gainera, SNOMED CTko terminoen euskarazko ordainak lortu ahala datu-base horretan txertatuko ditugu. *ItzulDB*n aldiz, baliabide lexikoetatik erauzitako termino-ordain pareak egongo dira bakarrik, SNOMED CTren informaziorik agertzen ez delarik.

4.2.1 SNOMED CTrentzako TBX formatua

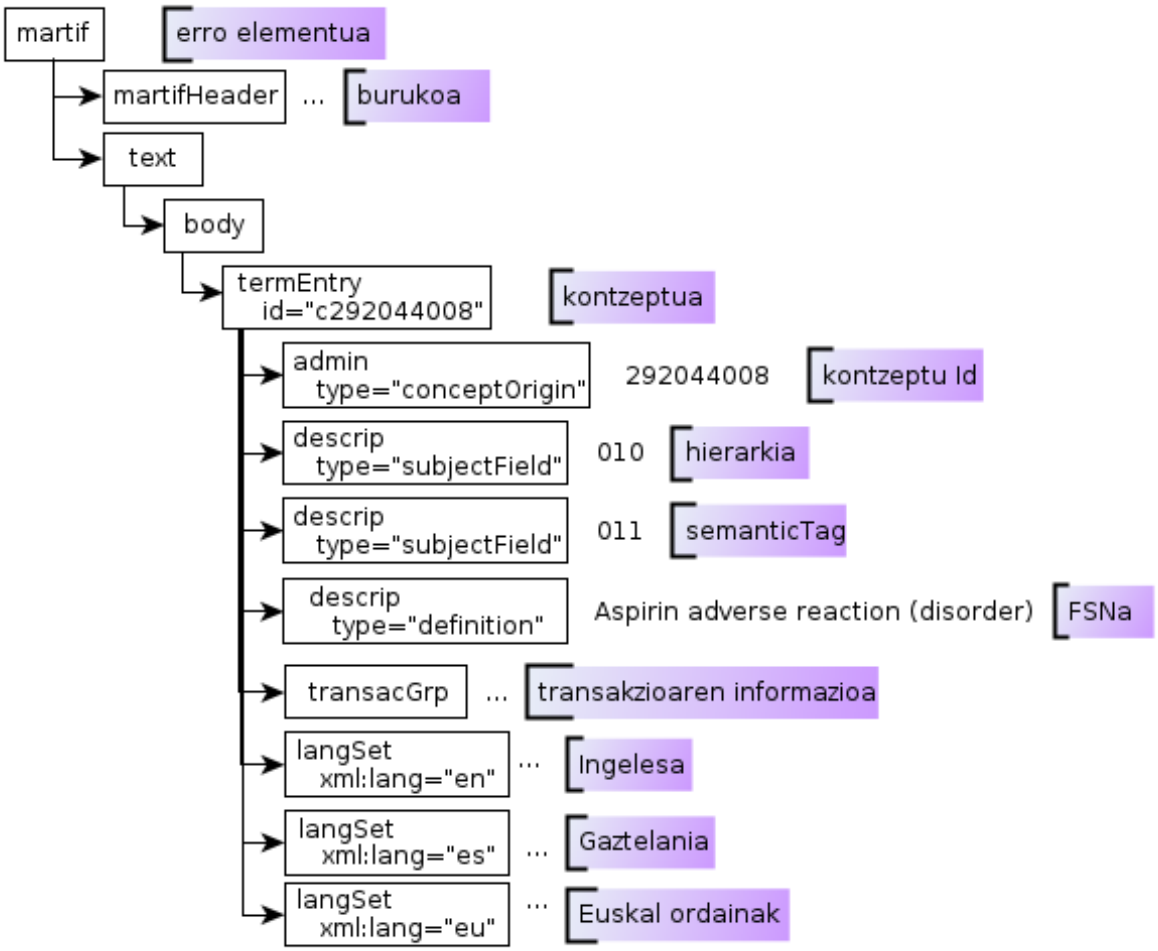
Atal honetan, TBX formatuan bertan adierazi eta gorde beharreko informazioa identifikatu eta kodetzeko hartutako irizpideen berri emango dugu.

Aurretik esan bezala, SNOMED CT gordetzeko formatua izango da hemen aurkeztuko duguna, baina euskal ordainak lortu ahala, SNOMED CTn datu-base honetan gordeko ditugu, euskarazko SNOMED CT ingelesezko zein gaztelaniazko bertsioekin batera gordeaz.

Horretarako, XML dokumentuaren goiburukoa alde batera utziko dugu eta gorputzean jarriko dugu arreta. Izan ere, goiburukoan dokumentuari buruzko azalpenak agertzen dira (izenburua, adibidez), eta gorputzean datuak gordetzen dira. Hurrengo azpiataletan elementu bakoitzari emango zaion interpretazioa azalduko dugu, hau da, erabiliko diren

datu-kategoriak azalduko ditugu. Datu-kategoriak elementuen *type* atributu gisa erabiltzen dira.

4.2.1.1 Kontzeptu maila



4. irudia: Kontzeptu baten zuhaitz egitura

Kontzeptua goiko elementua izango da, `termEntry` elementu bezala adierazia. Kontzeptua identifikatzeko SNOMED CTren Kontzeptu-identifikadore zenbakia erabiliko dugu, TBX eskakizunak betetzeko aurretik "c" letra ipiniko diogularik. "c292044008" izango da `termEntry`ren identifikadorea 4. irudiaren kasuan.

Jarraian Kontzeptu mailan erabiliko ditugun datu-kategoriak azalduko ditugu elementuen arabera sailkatuta. Azaplena ongi jarraitzeko, 4. irudiari begiratzea gomendatzen da.

- `admin` elementuaren datu-kategoriak `termEntry` barnean:

  - *conceptOrigin*: SNOMED CTren Kontzeptu-identifikadorea gordetzeko erabiliko dugu. Aurreko adibideari jarraiki (4. irudia), 292044008 balioa du, hori baita *"aspirin adverse reaction"* terminoarekin lotutako Kontzeptuaren identifikadorea.

- `descrip` elementuaren datu-kategoriak `termEntry` barnean:

  - *subjectField*: SNOMED CTren Kontzeptuaren hierarkia eta *semantic tag*a adierazteko erabiliko dugu. Hierarkia hauek adierazteko A.1 eranskineko 24. taulan kode-baliokidetzak erakusten ditugu. 4. irudiko balioak (010 eta 011) *Clinical Finding/disorder* hierarkiari eta *"disorder" semantic tag*ari dagozkie.
  - *definition*: Kontzeptua bera deskribatu eta ulergarri egiten duen adierazpena gordetzeko erabiliko dugu, antzeko Kontzeptuetatik bereizteko balioko diguna. Hemen helburu berdina duen SNOMED CTen *Fully Specified Name* (FSN) erabiliko dugu, ingeleseko bertsioarena hain zuzen ere. Adibidean, *"Aspirin adverse reaction (disorder)"*.
  - *explanation*: euskal hiztegietan agertzen diren Kontzeptu honen definizioak agertuko dira. Irudiko adibideko Kontzeptua ez denez hiztegietako sarrera bat, ez dugu horren definiziorik gorde.

  Jarraian azaltzen den datu-kategoriak `transGrp` elementuaren barnean multzokatuko dira, bakoitza elementu ezberdin bati lotuko zaiolarik. Berezitasun honetaz gain, transakzioak adierazteko datu-kategoria eta elementu hauek termino mailan ere erabiliko ditugu, hortaz, jarraian azalpen orokorra emango dugu, berezitasunak argi utziaz.

- `transGrp` elementuaren elementu eta datu-kategoriak (hauek ez ditugu irudian topatuko):

  - `transac` *transactionType*: transakzio beraren mota adierazten du, aldaketa, sorkuntza, inportazioa edo onarpena izan den. Jaso ditzakeen balioak eta euren azalpena A.2 eranskineko 26. taulan aurkitzen da. Kontzeptuen kasuan beti inportazioak izango dira, kontzeptuak SNOMED CTtik jasoko baititugu, eta berdina jatorri-terminoekin. Euskal ordainak izango dira balio ezberdinak jasoko dituztenak.
  - `transacNote` *responsability*: datu-kategoria honek transakzioa burutu duenaren ardura adierazten du. A.2 eranskineko 27. taulan definiturik dauden ardura ezberdinak agertzen dira. Sorkuntza edota inportazio motako transakzioetan arduraduna beti kudeatzailea izango da (*admin*), aplikazioaren esku egongo baitira transakzio mota hauek. Dena den, gure eginkizunerako transakzioa egin duen pertsonaren izena ere gordetzea interesatzen zaigu, eta hortaz, datu-kategoria honen bigarren agerpenak arduradunaren izen-abizenak agertuko dira.

Azkenik, hizkuntza bakoitzaren termino ezberdinak gordetzeko elementuak izango ditugu `langSet` bezala definiturik. Elementu honen *lang* atributuaren bitartez terminoen hizkuntza definituko da: euskara, ingelesa edo gaztelania. Elementu hauen edukia hurrengo atalean aztertuko dugu (4.2.1.2 atala).

Hurrengo adibidean (25. adibidea), Kontzeptu baten egitura erakusten dugu aurretik azalduriko elementuen bitartez adierazita. 4. irudiaren adibide berdinari jarraitzen dio.

```
<termEntry id="c292044008">
  <admin type="conceptOrigin">292044008</admin>
  <descrip type="subjectField">010</descrip>
  <descrip type="subjectField">011</descrip>
  <descrip type="definition">Aspirin adverse reaction (disorder)</descrip>
  <!--<descrip type="explanation">Hiztegietako definizioa, balego</descrip>-->
  <transacGrp>
    <transac type="transactionType">importation</transac>
    <date>2013-01-04T18:46:12.954+01:00</date>
    <transacNote type="responsibility">admin</transacNote>
    <transacNote type="responsibility">Olatz Perez de Vinaspre</transacNote>
  </transacGrp>
  <langSet xml:lang="en">...</langSet>
  <langSet xml:lang="es">...</langSet>
  <langSet xml:lang="eu">...</langSet>
</termEntry>
```

25. adibidea: Kontzeptu baten adibidea TBXn

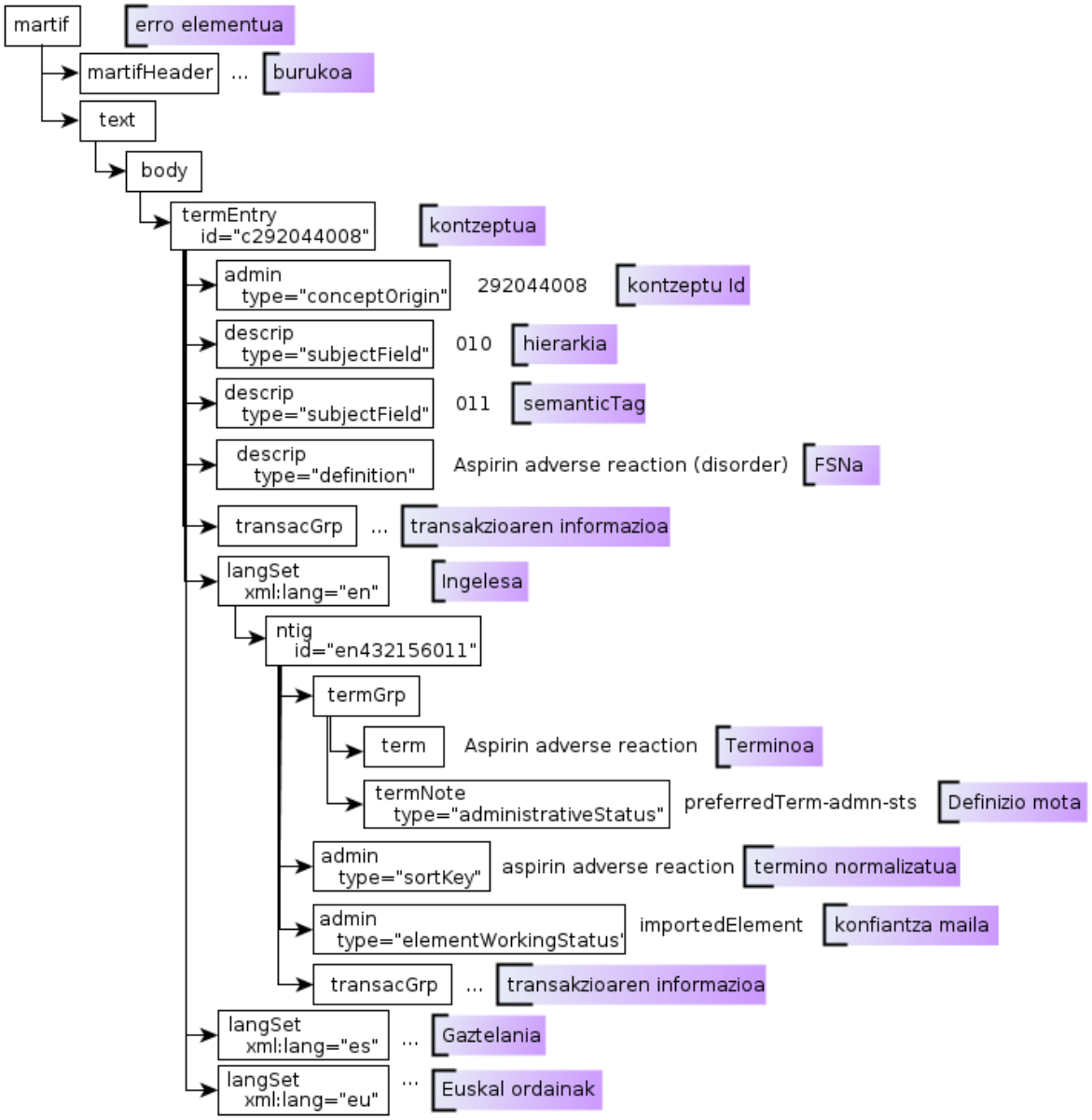
4.2.1.2 Termino maila

Aurretik azaldu bezala, terminoak hizkuntzaren arabera multzokatuta egongo dira, `langSet` elementuaren bitartez. SNOMED CTren euskaratze-lan honetarako bi motatako terminoak bereiziko ditugu: alde batetik jatorri-terminoak izango direnak, hau da, SNOMED CTtik zuzenean ekarri ditugun ingelesezko eta gaztelaniazko terminoak; eta bestetik, euskal ordainak. Termino mota bakoitzerako informazio ezberdina jasoko dugunez, bereizirik azalduko ditugu hurrengo ataletan.

4.2.1.2.1 Jatorri-terminoak

Jatorri-terminoak SNOMED CTtik zuzenan inportaturiko ingelesezko zein gaztelaniazko terminoak dira. Termino hauek gainean aldaketarik izan ez dutenez, gordeko dugun informazioa SNOMED CTkin zuzenki lotuta egongo da.

Termino bakoitza `ntig` elementuaren bitartez gordeko da eta identifikadore bezala hizkuntzaren gakoa ("en" edo "es") eta SNOMED CT Deskribapenaren identifikadorea erabiliko ditugu elkarrekin bilduta. 5. irudian daukagun adibidean, "en432156011" izango da terminoaren identifikadorea.



5. irudia: Jatorri-termino baten zuhaitz egitura

Termino bakoitzaren informazioa gordetzeko, hurrengo elementu eta datu-kategoriak erabiliko ditugu (jarraitu azalpena 5. irudiarekin):

- `termGrp` elementuan multzokatuta:
  - `term`: terminoa bera gordetzeko elementua da. 5. irudiko adibidearen kasuan *"Aspirin adverse reaction"* da elementu honen edukia.
  - `termNote` elementuaren datu-kategoria:

    * *administrativeStatus*: Terminoak SNOMED CTn daukan onarpen-maila adieraziko da. Besteak beste, termino hobetsia ala onartua den, edota adituek baztertu duten adieraziko da. A.2 eranskineko 28. taulan datu-kategoria honen balioen kodeak eta esanahaiak azaltzen dira. Adibidearen kasuan (5. irudia) terminoa *Preferred Term*a denez, "preferredTerm-admn-sts" izango da datu-kategoria honen balioa.

- `admin` elementuaren datu-kategoriak:
  - *sortKey*: Terminoaren balio normalizatua agertuko da, hau da, terminoaren forma minuskulaz eta alfazenbakizkoak ez diren karaktereak kenduta. Gaztelaniaren kasuan diakritikoak ere kenduko zaizkio. 5. irudiko adibidean "aspirin adverse reaction" da elementu honen balioa.
  - *elementWorkingStatus*: Termino batek daukan konfiantza maila adierazteko balio du. A.2 eranskineko 25. taulan datu-kategoria honen balioen kode ezberdinak eta esanahaiak azaltzen dira. Jatorri-terminoen kasuan, *importedElement* balioa izango dugu kasu guztietan.

Aurretik aipatu dugun bezala (4.2.1.1 atalean), transakzioak adierazteko `transacGrp` elementua erabiliko dugu, jatorrizko terminoen kasuan Kontzeptuen elementu zein datu berdinak dituelarik.

Hurrengo adibidean (26. adibidea), jatorri-terminoak TBX formatuan nola gordeko diren erakusten dugu. 25. adibidearen jarraipena da eta 5. irudiko adibidearen edukiari gaztelaniazko terminoak gehitzen dizkio.

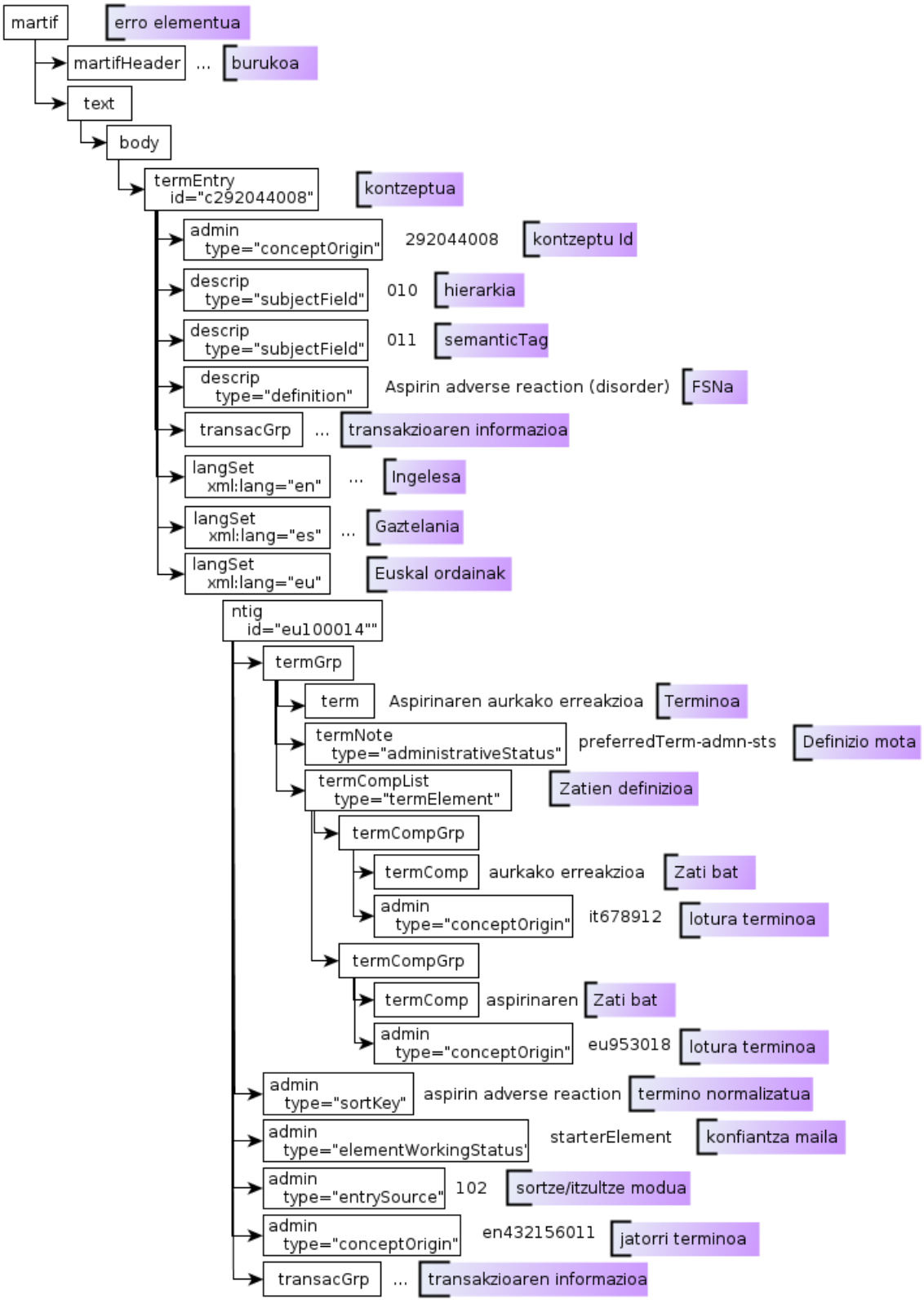
```
<langSet xml:lang="en">
  <ntig id="en432156011">
    <termGrp>
      <term>Aspirin adverse reaction</term>
      <termNote type="administrativeStatus">preferredTerm-admn-sts</termNote>
    </termGrp>
    <admin type="sortKey">aspirin adverse reaction</admin>
    <admin type="elementWorkingStatus">importedElement</admin>
    <transacGrp>...</transacGrp>
  </ntig>
</langSet>
<langSet xml:lang="es">
  <ntig id="es1299655018">
    <termGrp>
      <term>reacción adversa al ácido acetilsalicílico</term>
      <termNote type="administrativeStatus">preferredTerm-admn-sts</termNote>
    </termGrp>
    <admin type="sortKey">reaccion adversa al acido acetilsalicilico</admin>
    <admin type="elementWorkingStatus">importedElement</admin>
    <transacGrp>...</transacGrp>
  </ntig>
  <ntig id="es1328574019">
    <termGrp>
      <term>reacción adversa a la aspirina</term>
      <termNote type="administrativeStatus">admittedTerm-admn-sts</termNote>
    </termGrp>
    <admin type="sortKey">reaccion adversa a la aspirina</admin>
    <admin type="elementWorkingStatus">importedElement</admin>
    <transacGrp>...</transacGrp>
  </ntig>
</langSet>
```

26. adibidea: Jatorri-terminoen adibidea TBXn

4.2.1.2.2 Ordainak

Atal honetan, SNOMED CTren euskarazko bertsioaren sortze-bidean lortuko ditugun euskal ordainak adierazteko erabiliko dugun informazioaren egituratzea azalduko dugu.

Jatorrizko terminoetan gordetzen den informazioaz gain, beste hainbat datu ere gorde behar dugu. Dagoeneko 4.2.1.2.1 atalean azaldutako elementu eta datu-kategoriei honakoak gehituko dizkiegu (erabili bedi 6. irudiko egitura azalpenak hobeto ulertzeko):



6. irudia: Euskal ordain baten zuhaitz egitura

- `termCompList` Elementu honetan informazio linguistikoa bilduko dugu. Jarraian elementu eta datu-kategoria ezberdinak azalduko ditugu:
  - Bi motako `termCompList` erabiliko ditugu:
    * *termElement*: terminoa osatzen duten hitzak edo hitz multzoak gordeko direla adierazteko erabiltzen da.
    * *morphologicalElement*: terminoa osatzen duten morfema esanguratsuak gordeko direla adierazteko erabiltzen da.

    6. irudiaren adibidean *termElement* motako `termCompLista` definitu dugu.
  - `termCompGrp` elementuaren bitartez terminoa osatzen duten zati ezberdinak identifikatzen dira, zatia bera zehaztuz eta honen termino jatorria identifikatuz:
    * `termComp`: terminoaren zatia zehazten du. Hitz osoak izan daitezke, edota hitz-erroak (sorkuntza morfologikoan erabili direnak). 6. irudiari dagokionean, bi zati ezberdindu ditugu: "aurkako erreakzioa" eta "aspirinaren".
    * `admin` *conceptOrigin*: zein den jatorrizko terminoa identifikatzen du. Honetarako SNOMED CTko jadanik sortutako beste Deskribapen baten identifikadorea erabiliko da, eta hau ez balego, horren *ItzulDB*ko sarreraren identifikadorea. Beti ere euskarazko termino bati egingo dio erreferentzia. Adibidearekin jarraituz, "aurkako erreakzioa" *ItzulDB*n gordeta dagoen termino bati lotuta dago ("it678912") eta "aspirinaren" zatiak aldiz SNOMED CTko beste termino bati ("eu953018").
- `admin` elementuaren datu-kategoria gehigarriak:
  - *entrySource*: ordaina lortzeko erabili den bidea adierazten duen kodea da. A.2 eranskineko 29. taulan kode hauen esanahia adierazten da. 6. irudian "sintaxi mailako erregelei" egiten die erreferentzia (102 kodea).
  - *conceptOrigin*: ordain hau sortzeko erabili d(ir)en terminoaren identifikadorea. Kontzeptu mailako itzulpenak egingo direnez, identifikadore hau jatorri-terminoaren `tig` elementuaren identifikadorea izango da. Adibidean, 5. irudiko ingelesezko jatorri-terminoari egiten dio erreferentzia ("en432156011").
- `descrip` elementuaren datu-kategoria:
  - *context*: Medikuntzako adibidetegitik edota Corpusetik jasotako adibideak. Hasiera batean medikuntzako adibidetegitik bakarrik erauziko ditugu. Terminoaren testuingurua eta erabilera aztertzeko baliagarria zaigu. 6. irudiko ordainaren erabilera kasurik ez dugu gorde.

Terminoaren zatiak bereizteko beharrak, SNOMED CTren itzulpenean eguneraketak egiteko nahiari erantzuten dio. Termino batzuen itzulpenak egiteko aurretik itzulitako SNOMED CTren termino-ordainak erabiliko ditugunez, zati mailako harremanak gorde

nahi ditugu. Izan ere, eta aurretik aipatu bezala, adituek euskal ordainak zuzenduko dituzte, eta zuzenketa hauek burutzean zuzendutako ordainak beste ordain batzuekin loturarik balu, aldaketak barreiatzeko erabiliko genituzke lotura horiek. Barreiatutako aldaketa hauek ez dute aurrez sortutako ordaina ordezkatuko, baizik eta hautagai bat gehituko liokete.

Ordain baten errepresentazioa ikus dezakegu 27. adibidean. Irizpideei erantzunez, aurreko adibideetatik (26. eta 25. adibideak) itzuliko genukeen terminoaren errepresentazioa da, baita 6. irudiarena ere.

```
<langSet xml:lang="eu">
  <ntig id="eu100014">
    <termGrp>
      <term>Aspirinaren aurkako erreakzioa</term>
      <termNote type="administrativeStatus">preferredTerm-admn-sts</termNote>
      <termCompList type="termElement">
        <termCompGrp>
          <termComp>aurkako erreakzioa</termComp>
          <admin type="conceptOrigin">it678</admin>
        </termCompGrp>
        <termCompGrp>
          <termComp>aspirinaren</termComp>
          <admin type="conceptOrigin">eu953018</admin>
        </termCompGrp>
      </termCompList>
    </termGrp>
    <admin type="sortKey">aspirinaren aurkako erreakzioa</admin>
    <admin type="elementWorkingStatus">starterElement</admin>
    <admin type="entrySource">102</admin>
    <admin type="conceptOrigin">en432156011</admin>
    <descrip type="context">Medikuntzako adibidetegitik edota Corpusetik
      jasotako adibideak.</descrip>
    <transacGrp>
      <transac type="transactionType">origination</transac>
      <date>2011-07-24T18:18:54</date>
      <transacNote type="responsibility">admin</transacNote>
      <transacNote type="responsibility">Olatz Perez de Vinaspre</transacNote>
    </transacGrp>
  </tig>
</langSet>
```

27. adibidea: Euskal ordainen adibidea TBXn

4.2.2 Itzulpen-pareen datu-baserako TBX formatua

Diseinatu berri dugun TBX formatua itzulpen-pareen datu-basea (*ItzulDB*) egituratzeko ere erabiliko dugu. Dena dela, aurreko atalean zehaztutako elementu eta datu-kategoria gutxi batzuk erabiliko ditugu termino-ordain pareak gordetzeko. Parekatzea jatorri-hizkuntzaren termino bakoitzeko egingo da, adibidez, ingeleseko termino bat eta berari dagozkion ordainak. Ez ditugu jatorri-hizkuntzako sinonimoak kontuan hartu, bilaketa gako bakarra izateko, eta horrela, bilaketa azkarragoa eta sinpleagoa izan dadin.

Parekatze bakoitza `termEntry` elementu baten barruan gordeko da, eta identifikadorea "p" letraz hasi eta jarraian zenbatzaile bat joango da. 28. adibidean *ItzulDB*ren parekatzen baten adibidea ikus daiteke. Parekatzearen barruan *elementWorkingStatus* datu-kategoria esleituko zaio, berau hiztegietatik erauzi den ala beste nolabait sortua izan den adierazteko.

Terminoak adierazteko aldiz, `langSet` elementuaren barruko `tig` elementuak erabiliko ditugu, terminoaren hizkuntza adierazita utziz. `tig` elementuak `ntig` elementuen antzekoak dira, baina sinpleagoak. Hauek identifikatzeko parekatzeen estrategia berdina erabiliko dugu, baina "t" letra erabiliz hasieran ("t13", adibidez).

Jatorri-terminoen kasuan, terminoa eta honen termino normalizatua erabiliko dugu, `term` elementua eta *sortKey* datu-kategoria erabiliaz. Ordainentzat aldiz, informazio gehigarria gordeko dugu, hala nola, *elementWorkingStatus* datu-kategoriaren bitartez, ordain horren konfiantza maila neurtuko dugu, eta *entrySource*ren bitartez iturburua zein izan den gordeko dugu.

Hiztegietan dauden parekatzeen kasuan, terminoen idazketan hiztegi ezberdinek ez dituztenez irizpide berdinak jarraitu, hauen termino normalizatua gordeko dugu bakarrik, baina iturburua gordetzearekin batera, honen jatorrizko parekatzea gordeko dugu. 12. taulan ikus dezakegunez, ZT Hiztegiak eta Anatomiako glosategiak termino guztien lehenengo letra minuskulaz idazten duten bitartean, Erizaintzako Hiztegiak terminoaren eta lehen ordainaren hasierako letra larriz idazten du.

Hiztegia	**Jatorri-terminoa**	**Euskal ordainak**
ZT Hiztegia	abdomen	abdomen
Anatomiako glosategia	abdomen	abdomen
Erizaintzako Hiztegia	Abdomen	Abdomen, sabelalde

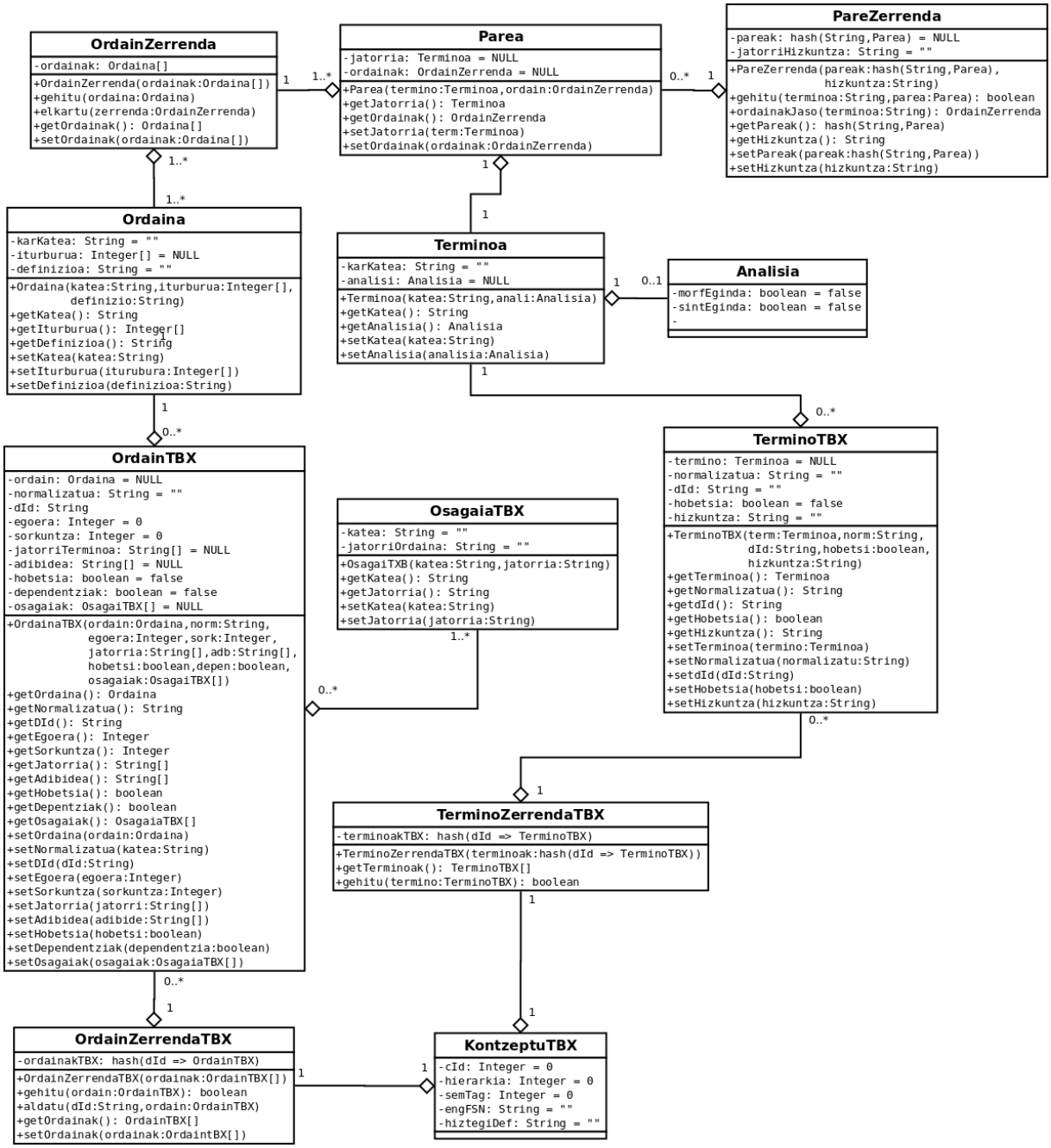
12. taula: Hiztegi ezberdinen sarrerak "abdomen" terminorako

```
<termEntry id="p6">
  <admin type="elementWorkingStatus">importedElement</admin>
  <langSet xml:lang="en">
    <tig id="t13">
      <term>abdomen</term>
      <admin type="sortKey">abdomen</admin>
    </tig>
  </langSet>
  <langSet xml:lang="eu">
    <tig id="t14">
      <term>abdomena</term>
      <admin type="sortKey">abdomena</admin>
      <admin type="elementWorkingStatus">starterElement</admin>
      <admin type="entrySource">GNS10##Abdomen ##Abdomena</admin>
    </tig>
    <tig id="t15">
      <term>abdomen</term>
      <admin type="sortKey">abdomen</admin>
      <admin type="elementWorkingStatus">starterElement</admin>
      <admin type="entrySource">ZT##abdomen##abdomen</admin>
      <admin type="entrySource">Anatomia##abdomen##abdomen</admin>
      <admin type="entrySource">Erizaintza##Abdomen##Abdomen</admin>
    </tig>
    <tig id="t16">
      <term>sabelalde</term>
      <admin type="sortKey">sabelalde</admin>
      <admin type="elementWorkingStatus">starterElement</admin>
      <admin type="entrySource">Erizaintza##Abdomen##sabelalde</admin>
    </tig>
  </langSet>
</termEntry>
```

28. adibidea: *ItzulDB*ren parekatze bat

4.3 Klase-diagrama

Atal honetan aplikazioa garatzeko diseinatutako klase-diagramak azalduko ditugu. Hasiera batean 7. irudiko diagrama diseinatu genuen.



7. irudia: Aplikazioaren diseinurako hasierako klase-diagrama

Bertan XML dokumentuetako ezaugarri eta elementu guztiak objektuetara eramaten genituen, baina horrela, ez genituen XML dokumentuak kudeatzeko abantailak aprobetxatzen: sinpletasuna eta azkartasuna. Era honetan klase-diagrama berri bat egin genuen, oraingoan erabiliko genituen teknologiekin bat egiten zuelarik.

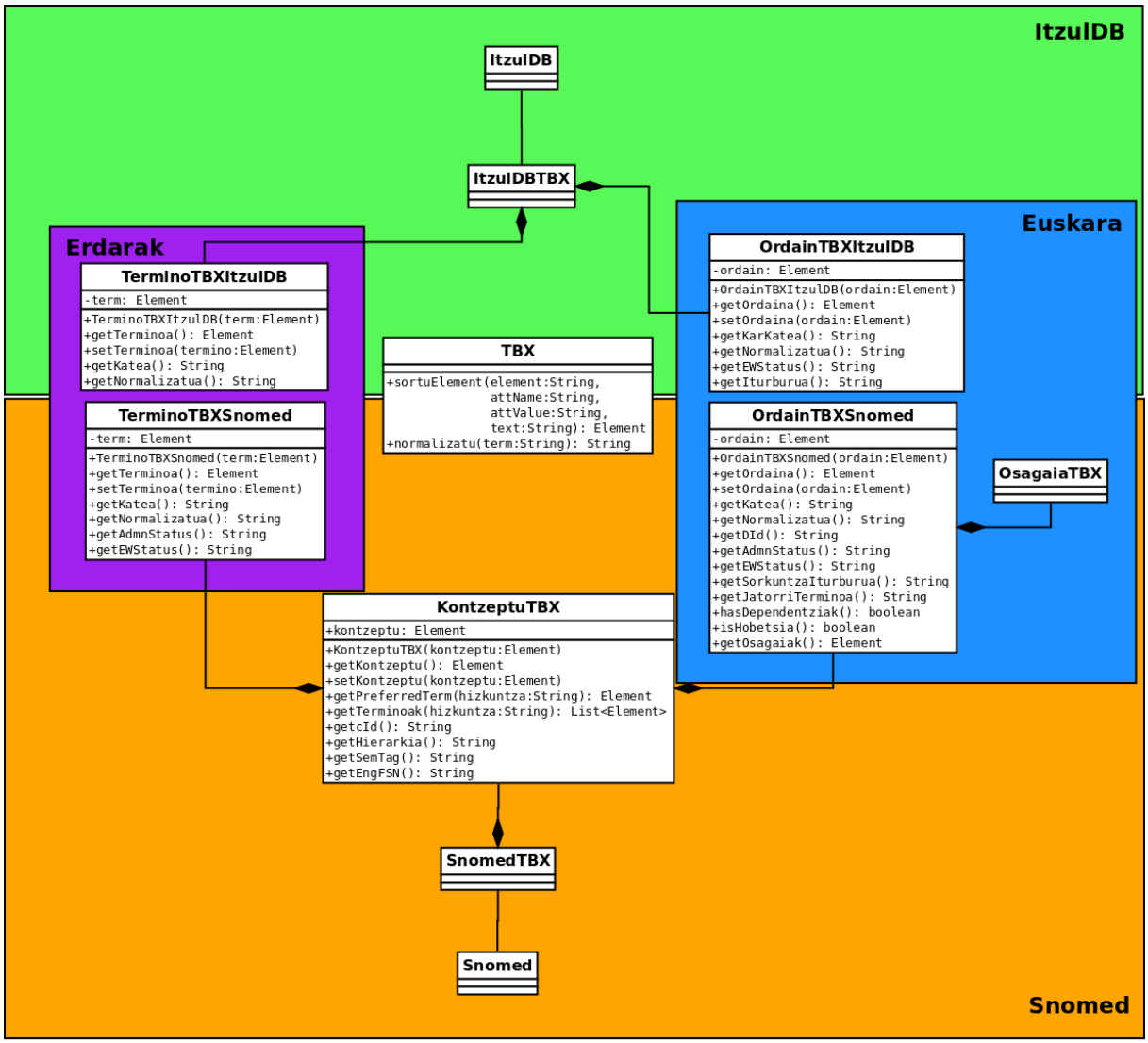
Aplikazioan XML dokumentuak etengabe irakurri eta aldatzen ari garenez, zentzuzkoagoa dirudi XML dokumentuak prozesatzeko dauden teknologiak erabiltzea. Horrela, objektuak sortzen eta ezabatzen ibiltzea aurrezten dugu.

Hau horrela izanik, XML dokumentuak Java lengoaian kudeatzeko JDOM teknologiaren abantailez baliatzea erabaki genuen, eta horrela, diseinu berri bat egin genuen. Aplikazioan inplementatutako klase-diagrama 8. irudian ikus daiteke. Bertan *ItzulDB* kudeatzeko klaseak eskemaren goiko aldean kokatu ditugu, eta SNOMED CT kudeatzekoak beheko aldean. Hizkuntzei dagokienean, ezkerreko aldean erdarak kudeatzeko objektuak definitu ditugu, hau da, jatorri-terminoak kudeatzeko objektuak, eta eskuineko aldean aldiz, euskal ordainak kudeatzekoak.

Kasu honetan datu guztiak XML dokumentuetan gelditu ko dira, eta objektuen bitartez XML dokumentu horietako informazioa jasoko dugu. Horrela klase-diagrama asko sinplifikatu zaigu, objektuen arteko harremanak gutxituz. Klase-diagrama bi zatitan bana dezakegu, alde batetik *ItzulDB* kudeatzeko zatia izango genuke, eta bestetik SNOMED CT kudeatzeko zatia.

Klase bakoitzaren funtzio nagusia azalduko dugu jarraian:

- `ItzulDB`: *ItzulDB* bera kudeatzeko objektua izango da. Besteak beste, *ItzulDB* hasieratzeko beharrezko informazioa erauziko du hiztegietatik.
- `ItzulDBTBX`: *ItzulDB*ren XML dokumentua kudeatzeko objektua izango da. Parekatze baten bilaketa edota `ItzulDB` klaseak emandako informaziotik XML dokumentua sortzea izango dira klase honen ataza nagusiak.
- `TerminoTBXItzulDB`: *ItzulDB*ren jatorri-terminoen gaineko informazioa eskuratzeaz arduratuko da.
- `OrdainTBXItzulDB`: *ItzulDB*ren ordainen gaineko informazioa eskuratzeaz arduratuko da.
- `Snomed`: SNOMED CTren kudeaketaz arduratuko da. SNOMED CTren bertsio berri bat kargatzeko aplikazioei deitzeko eginkizuna izango du, baita GNS-10 eta SNOMED CTren arteko mapaketa egiteko ardura ere bai. Honetaz gain, SNOMED CTn termino berri bat gorde nahi denean, klase honek SNOMED CTren identifikadore balizkoa sortu beharko du (IHTSDO (2012)).
- `SnomedTBX`: SNOMED CTren XML dokumentuak kudeatuko ditu. Besteak beste, fitxategiak sortu edota ordenatu. Kontzeptuak, terminoak eta ordainak kudeatzeko klaseetara lotura egitea ahalbidetuko duen klasea ere bada.
- `KontzeptuTBX`: Kontzeptuari loturiko informazioa eskuratzeaz gain, ordainak txertatzeaz arduratuko da klase hau.



8. irudia: Aplikazioaren diseinurako klase-diagrama definitiboa

- `TerminoTBXSnomed`: SNOMED CTko jatorri-terminoen gaineko informazioa eskuratzeaz arduratuko da.
- `OrdainTBXSnomed`: SNOMED CTko ordainen gaineko informazioa eskuratzeaz arduratuko da.
- `OsagaiaTBX`: Ordain barruko zati edo morfemak kudeatzeaz arduratuko da.
- `TBX`: Klase estatiko honek TBX formatuarekin zerikusia duten eragiketak egiteko tresnak eskaintzen ditu, XML dokumentuak kudeatuko dituzten beste klaseek erabil ditzaten. Besteak beste elementu berri bat TBX formatu egokian sortzea izango da honen eginkizun nagusia.

5 Inplementazioa

Orain arte deskribatu dugun aplikazioaren hastapenak inplementatu ditugu. *ItzulDB* eta SNOMED CTren XML dokumentuak prest izanik, euskal ordainak lortzeko lehen urratsak eman ditugu.

Inplementazioari dagokionean, Java lengoaiaz idatzi dugu aplikazioa. Lengoaia horri lotuta, XML dokumentuak errepresentatzeko JDOM (Hunter eta McLaughlin (2012)) liburutegia erabili dugu. Liburutegi horrek, dokumentuari zuhaitz egitura ematen dio eta horrela elementuz-elementu arakatzea era erraz eta eraginkor batean egitea ahalbidetzen du, XML dokumentuetan irakurketak, aldaketak eta idazketak sinplifikatuz.

Diseinaturiko XML dokumentuen konplexutasuna (4.2 atala) oztopo izan da JDOMek eskaintzen dituen bilaketa sinpleak egiteko garaian. Diseinatutako XML dokumentuen atributuen erabilpen zabalak, bilaketak egiteko garaian deserosotasunaz gain errekurtsoen erabilpen zabala dakar. Honi aurre egiteko, JDOMek berak erabat integratuta dakarren XPATH erabili dugu.

XPATH (***X****ML* ***Path*** *Language*) XML dokumentua arakatzeko eta prozesatzeko adierazpenak eraikitzea ahalbidetzen duen hizkuntza da. Atributu gabeko testuan bilaketak eta hautapenak egiteko adierazpen erregularren ideia antzekoari jarraitzen dio. Izan ere, XPATHek XML dokumentuaren egitura hierarkikoa jarraituz bilaketak eta hautapenak egitea ahalbidetzen du, honen elementuen atributuak kontuan hartuaz.

Lehenik, SNOMED CT eta GNS-10ren arteko mapaketa inplementatu dugu, Kontzeptu mailako mapaketa izanik, algoritmoaren aipatutako urratsekin hasi aurretik inplementatzea erabaki dugu. Izan ere, algoritmoa terminoak itzultzeko diseinatuta dago, eta ez Kontzeptuak. Algoritmoari dagokionez lehenengo pausoa inplementatu dugu orain arte.

Ordainak lortzeko lehen urrats hauen inplementazioa ez da konplexua izan, aurretik egindako diseinuari esker, bilaketa eta parekatzeak bat batekoak izan baitira. Dena dela, aipamen berezia merezi du inplementazioan baliabide ezberdinen prestaketak ekarri dituen arazo eta erabaki hartzeak. Hurrengo ataletan bi baliabide nagusiak (*ItzulDB* eta SNOMED CTren XML dokumentua) sortze bidean sortutako gorabeherak azalduko ditugu.

5.1 Hiztegiak eta GNS-10 itzulpen-pareen aberasketan

Itzulpen-pareen datu-basea (*ItzulDB*) 2.5 eta 2.4 ataletan aztertu ditugun hiztegi zein glosategietatik osatu dugu. Hasiera batean ingelesezko SNOMED CT bertsioa bakarrik izango dugunez itzulpenak egiteko iturburu, ingelesezko *ItzulDB* sortu dugu, eta gaztelaniazkoa inplementaziorako prest gelditu bada ere, ez dugu gaztelaniazko datu-basea sortu.

Era honetan, ingelesezko *ItzulDB*a osatzeko, Erizaintzako Hiztegia, ZT Hiztegi Entziklopedikoa, Anatomiako glosategiaren lanerako bertsio bat eta euskarazko zein ingelesezko GNS-10ak erabili ditugu.

Hiztegiak erabili ahal izateko hauen aurre-prozesaketa egin behar izan dugu. Erizaintzako Hiztegiaren kasuan PDF formatuan eskuragarri dago bakarrik, eta bi zutabetan egituratuta egonik honen TXT formatura itzulpena ez da bat-batekoa izan. Formatu al-

daketa horretarako jatorrizko formatua zatikatu egin dugu, eta ostean zati horien formatu aldaketarako *Calibre* aplikazioak eskaintzen duen formatuen aldaketarako tresna erabili dugu. Tresna horrek adierazpen erregularren erabilera ahalbidetzen duenez, hiztegia TXT formatura itzuli ahal izan dugu.

Dena dela, nahiz eta hiztegi guztiek TXT formatua izan, hiztegi bakoitzaren egitura oso ezberdina da bata bestearengandik. Hala nola, hiztegi batzuk termino sarrera bakoitza letra larriz idazten dute, honek berezkoa izan ala ez. Beste batzuk aldiz, berezko letra larriak bakarrik erabiltzen dituzte, entitateak edota laburdurak adierazteko, adibidez. Irizpide ezberdin hauek *ItzulDB* aberastea zaildu egiten dute, termino berdinaren bi idazketa ezberdin genituelako kasu askotan. Honi aurre egiteko *ItzulDB*n termino normalizatuak gordetzea erabaki dugu. Nahiz eta horrela entitateen berezitasunak galtzen ditugun, erredundantzia ekiditen dugu eta baliabideen optimizazioa lortzen dugu ere.

Erizaintzako Hiztegiak eta ZT Hiztegiak euskal terminoen definizioak ematen dituzte. Dena dela, definizio horiek ez datoz termino-ordain pareekin batera, eta hortaz definizioen aberasketa aplikazioaren azken faserako uztea erabaki dugu, algoritmoaren atal guztiak bukatu ostean gauzatzeko.

GNS-10ari dagokionean, eskura daukagun ingelesezko GNS-10 bertsioa ez dago erabat osatua, hainbat terminoren falta dago. Sarean eskuragarri dauden beste GNS-10 fitxategiak aztertu ditugu eta hauen egitura ez dator guk dugun GNS-10ren egiturarekin bat. Terminoen osaeran aurkitu ditugu arazo handienak, terminoak adierazteko beste formatu bat erabiltzen baitute.

Azkenean, SNOMED CT eta GNS-10en arteko mapaketa erabili dugu ingelesezko GNS--10ko hutsuneak betetzeko. Mapaketan GNS-10en ingelesezko terminoak ere agertzen dira eta hauen idazketa gure ingelesezko GNS-10 bertsioaren oso antzekoa denez, berauek erabili ditugu ingelesezko GNS-10 elikatzeko. Horrela, ia 500 parekatze gehiago erabilgarri izan ditugu *ItzulDB* elikatzeko.

Kontraesan bakarra aurkitu dugu elikatze hau burutzeko garaian: gure GNS-10 AEBko ingelesez idatzita dago, eta mapaketarena Britainia Handiko ingelesez. Kontraesan honek aldiz, berehalako konponketa dauka, 3.1 atalean ikusi dugun bezala bi dialektoen Sinonimoak jaso baititugu itzulpenerako. Hortaz, mapaketaren kasuan kode bidez egingo dugunez nabarituko ez bada ere, hiztegi gisa erabiltzean ere ez digu arazorik sortuko, termino baten parekatzea aurkitzen ez badu, bere sinonimoa den beste terminoa aurkituko duelako.

Dialektoen aldaketaren kasu baten adibidea erakusten du 29. adibideak.

SNOMED CT terminoak:
US ingelesa: *Gastric hemorrhage*
BH ingelesa: *Gastric haemorrhage*
GNS-10ko terminoak:
Ingelesezko GNS-10: *Gastrointestinal hemorrhage, unspecified*
Mapaketa: *Gastrointestinal haemorrhage, unspecified*

29. adibidea: AEB eta BHko ingelesen erabilera GNS-10n eta SNOMED CTn

Dena dela, euskarazko GNS-10aren kasuan terminoak mugatu forman aurkitu ditugu. 28. adibidea gogora ekarriz, GNS-10ren hizkuntzen arteko parekatzeak *abdomen* ingelesezko terminoaren ordaina "abdomena" dela ikus dezakegu. Momentuz ez diogu aparteko tratamendurik eman, eta horrela gorde ditugu ordainak. Etorkizunean termino hauen gaineko aurre-prozesaketa beharrezkoa izango da, terminoen forma mugagabea lortzeko.

```
<termEntry id="p6">
  <admin type="elementWorkingStatus">importedElement</admin>
  <langSet xml:lang="en">
    <tig id="t13">
      <term>abdomen</term>
      <admin type="sortKey">abdomen</admin>
    </tig>
  </langSet>
  <langSet xml:lang="eu">
    <tig id="t14">
      <term>abdomena</term>
      <admin type="sortKey">abdomena</admin>
      <admin type="elementWorkingStatus">starterElement</admin>
      <admin type="entrySource">GNS10##Abdomen ##Abdomena</admin>
    </tig>
    <tig id="t15">
      <term>abdomen</term>
      <admin type="sortKey">abdomen</admin>
      <admin type="elementWorkingStatus">starterElement</admin>
      <admin type="entrySource">ZT##abdomen##abdomen</admin>
      <admin type="entrySource">Anatomia##abdomen##abdomen</admin>
      <admin type="entrySource">Erizaintza##Abdomen##Abdomen</admin>
    </tig>
    <tig id="t16">
      <term>sabelalde</term>
      <admin type="sortKey">sabelalde</admin>
      <admin type="elementWorkingStatus">starterElement</admin>
      <admin type="entrySource">Erizaintza##Abdomen##sabelalde</admin>
    </tig>
  </langSet>
</termEntry>
```

28. adibidea: *ItzulDB*ren parekatze bat

Baliabide hauek guztiak aplikazioaren lehenengo exekuzioan (edota baliabide berriren bat gehitzean) gehituko zaizkio *ItzulDB*ri, eta gainerako exekuzioetan XML dokumentua zuzenean kargatu egingo da.

HAP masterra

5.2 SNOMED CTren egokitzapena

SNOMED CTren analisia egiteko erabili ditugun irizpide antzekoak erabili ditugu SNOMED CTtik terminoak erauzteko (3.1 atala). Alde batetik RF2ren *Snapshot* banaketa erabiliko dugu, eta bestetik, aplikazioa inplementatzeko garaian eskuragarri zeuden SNOMED CTren azken bertsioak erabili ditugu: ingelesezko nazioarteko banaketa 2012ko uztailekoa da, eta gaztelaniazko banaketa 2012ko urrikoa. Hau da, analisikoa baino bertsio berriagoak. Deskribapenen kasuan ere aktibo daudenak hartu ditugu oinarri gisa.

SNOMED CTren egitura kontuan hartzen badugu eta 3.1 ataleko ondorioei jarraituz, berau hierarkiaka banatu dugu. Banaketa honen bitartez SNOMED CTren tamaina txikiagotzea lortzen dugu eta XML dokumentuen kudeaketa arintzea.

Aplikazioa lehen aldiz exekutatzen denean (edo SNOMED CTren bertsioa eguneratzen dugunean), lehenik eta behin SNOMED CT hierarkiaka banatu egingo du, ostean 4.2.1 atalean ikusi dugun bezala, SNOMED CT XML dokumentuan gordetzeko. Horrela hierarkia bakoitzeko XML dokumentu bat izango dugu, nahiz eta dokumentu ezberdineko identifikadoreen artean bateragarritasuna mantendu den (ez dira identifikadorerik errepikatu).

Clinical Finding/disorder hierarkiaren kasuan banaketa *semantic tag*aren arabera egin dugu, hierarkia hau populatuena izanik, bere tamaina handiegia zelako ordenagailuaren memorian kargatzeko. Hau horrela izanik, *"disorder"* ("alterazio") *semantic tag*a duten Kontzeptuak alde batetik gorde ditugu, eta *"finding"*ak ("aurkikuntzak") bestetik.

Dena dela, hierarkiaka banatzeak arazo berri baten sorrera ekarri digu: kontzeptu guztiak ez daude SNOMED CTren Erlazioen fitxategian. Izan ere, hierarkiak zuzenki banatu ahal izateko SNOMED CTren erro kontzeptutik abiatuz, ontologia osoa zeharkatu dugu hierarkia bakoitzaren Kontzeptuak sailkatuz. SNOMED CTren Erlazioak aztertu ditugunean *semantic tag* biren kasua eskuz aldatu behar izan dugu, *namespace concept* eta *record artifact*ena hain zuzen ere. Bi hauek metadatuen erro Kontzeptuari loturik agertzen zitzaizkigunez hasiera bateko zenbakietan ez ziren agertzen.

Horrela, 13. taulan hierarkien banaketari esker lortu ditugun kopuru errealak erakusten ditugu, banaketa egin gabeko zenbakiekin alderatuz. Ikus daitekeenez Kontzeptu asko galdu egiten dira, ia 100.000 Kontzeptu hain zuzen ere. Galera honen arrazoia Kontzeptuen aktibo egoeran datza. Gogora ekartzen badugu 3.1 atalean esaten zena, Deskribapena aktibo egon arren, posible da horri dagokion Kontzeptua jadanik aktibo ez egotea. Hori da 100.000 Kontzeptu horien kasua.

Hortaz, zeharka bada ere, ez ditugu aktibo dauden Deskribapen guztiak kontuan hartu, baizik eta Kontzeptu aktiboetatik abiatu gara, egitura sendoa duen SNOMED CT euskaratu ahal izateko, ahulguneak alde batera utziaz.

Gaztelaniazko terminoen daturik ez dugu lortu, gaztelaniazko Sinonimoak bakarrik gehitu baitizkiogu SNOMED CTri. Diseinuari jarraiki (4.2 atala) ingelesezko SNOMED CT da oinarria, eta berau gaztelaniazko SNOMED CTrekin osatu dugu.

		Banatu gabe	**Hierarkia banatuta**
Hierarchy	**Semantic Tag (ST)**	**FSNen #**	**FSNen#**
Clinical Finding/disorder	disorder	94.242	65.386
	finding	45.401	33.204
Procedure/intervention	procedure	75.078	50.587
	regime/therapy	3.573	2.466
Organism	organism	35.870	32.465
Body structure	body structure	26.960	25.519
	morphologic abnormality	5.259	4.450
	cell	645	627
	cell structure	513	509
Substance	substance	25.834	23.777
Pharmaceutical/biologic product	product	24.379	17.143
Qualifier value	qualifier value	10.134	8.982
Observable entity	observable entity	9.044	8.254
Event	event	8.959	3.661
Situation with explicit context	situation	8.716	3.238
Social context	occupation	6.460	3.851
	person	668	423
	ethnic group	366	262
	religion/philosophy	227	203
	life style	30	21
	social concept	27	26
	racial group	21	19
Physical object	physical object	5.148	4.508
Specimen	specimen	1.455	1.331
Environment geographical location	environment	1.253	1.094
	geographic location	619	617
	environment/location	1	1
Linkage concept	attribute	1.157	1.122
	link assertion	8	8
	linkage concept	1	1
Staging and scales	assessment scale	1.125	1.055
	tumor staging	261	214
	staging scale	41	16
Special concept	navigational concept	732	640
	namespace concept	153	153
	administrative concept	80	-
	special concept	31	1
Record artifact	record artifact	318	223
Physical force	physical force	178	171
Root Metadata Concept	foundation metadata concept	164	164
	core metadata concept	31	30

13. taula: Hierarkiaka sailkatzearen ondorioak populazioari dagokionean. HAP masterra

SNOMED CT XML dokumentuetan gordetzeko prestatzeaz gain, berau ordenatu egin dugu, *Preferred Term*en hitz luzeraren arabera, lehenik hitz bakarreko terminoak itzuli ahal izateko datu-basean bilaketak egiten ibili behar izan gabe.

Azkenik SNOMED CT identifikadoreak modu egokian sortu ditugu euskarazko ordain berrietarako, IHTSDO (2012) txosten teknikoan emandako irizpideei jarraiki.

6 Emaitzak

Atal honetan aplikazioaren lehenengo hastapenetan lortutako emaitzak aurkeztuko ditugu. Alde batetik emaitzen alde kuantitatiboa erakutsiko dugu, itzulpen kopuruen informazioa emanaz (6.1). Bestetik, aplikazioak sortu dituen ordainen adibideak emango ditugu, zenbakiak alde batera utziaz (6.2).

6.1 Emaitzak zenbakitan

Jarraian, lortu ditugun emaitzen taulak erakutsiko ditugu. Atal honen baitan *"Clinical Finding/disorder"*, *Procedure* eta *"Body Structure"* hierarkie dagozkien emaitzak erakutsiko ditugu, bukaeran gainontzeko hierarkien emaitzen batura azalduko delarik.

Bi taula motatan antolatu ditugu emaitzen datuak. Alde batetik, emaitza orokorrak ematen dizkigun taula daukagu (14. taula, adibidez). Bertan hiztegiekin lortutako parekatze kopurua jasotzen dugu, hau da, *ItzulDB*ko zenbat parekatze egon diren jasotzen dugu. Honetaz gain SNOMED CT eta GNS-10en arteko mapaketaren bitartez egindako parekatze kopurua ere jasotzen dugu (ikus "GNS-10 mapaketan" lerroa mota honetako tauletan).

Hiztegien kasuan, alde batetik, orokorrean zenbat terminoren parekatzea lortu den erakusten dugu, eta bestetik, zehazki hiztegi bakoitzetik zenbat ordain lortu diren. Azal dezagun hori 14. taulan erreparatuta, *"disorder" semantic tag*ari dagokion taula horretan, 2.953 gaixotasunen termino parekatu dira hiztegiekin, baina lortutako ordain kopurua handiagoa da, 3.718 hain zuzen ere ("Ordain kopurua" zutabearen batura). Honen arrazoia, parekatzeen bitartez termino bakoitzarentzat ordain bat baino gehiago egon daitekeela da. 30. adibidean ikusi daitekeen moduan, *"Microcephaly"* ingelesezko terminoaren ordaina hiztegi ezberdinek ematen dute, nahiz eta ordaina berdina izan. Horrela, hiru hiztegiren ordaina izan arren, parekatze bakarra gertatu da.

Ingelesezko terminoa: *Microcephaly*
Ordaina: mikrozefalia
Iturburu hiztegiak:
ZT Hiztegia
Erizaintzako Hiztegia
GNS-10

30. adibidea: Parekatze bakarrean hainbat hiztegi

Gainera, parekatze bakar gisa ere kontatu ditugu, jatorri-termino berdina duten hiztegietako sinonimoak, hau da, 31. adibidearen kasuan Erizaintzako Hiztegiak hiru ordain sortuko ditu "Leprosy" terminoarentzako eta ZT Hiztegiak bi, parekatze bakarra izan arren. Guztira hiru ordain lortu dira iturri desberdinetatik, baina SNOMED CTko termino bakarra itzuli da.

Ingelesezko terminoa: *Leprosy*
1. ordaina: legen beltz
Iturburu hiztegiak:
Erizaintzako Hiztegia
2. ordaina: legen
Iturburu hiztegiak:
ZT Hiztegia
Erizaintzako Hiztegia
3. ordaina: legenar
Iturburu hiztegiak:
ZT Hiztegia
Erizaintzako Hiztegia

31. adibidea: Parekatze bakarrean hainbat ordain

Emaitza orokorrekin bukatzeko, denera lortutako ordainak zenbatu ditugu eta baita zenbat Kontzepturen ordainak lortu diren ere. Azken honen portzentaia ere jasotzen da 14. taulan eta honen antzekoetan.

Alterazioen ("disorder") taulari begiratuta ikus dezakegu gaixotasunen %19,32a itzuli dela, Kontzeptu berarekin loturiko termino batek baino gehiagok lortu duelarik ordaina. Oso portzentaia altua lortu dugula esan daiteke, kontuan izaten badugu aplikazioaren hastapenetan gaudela.

Aurkitutakoak	**Ordain kopurua**	**Parekatze kopurua**
Hiztegietan	-	2.953
ZT Hiztegia	1.055	-
Erizaintzako Hiztegia	435	-
Anatomiako Glosategia	3	-
GNS-10	2.225	-
GNS-10 mapaketan	-	11.228
	Kopurua	**Portzentaia**
Ordainak denera	14.568	-
Kontzeptuak denera	12.630	%19,32 (65.386)

14. taula: *Clinical Finding/disorder* hierarkiaren *"disorder" semantic tag*aren emaitzak.

*"Disorder" semantic tag*erako sortutako bigarren taulan (15. taula), itzuli diren jatorri-terminoen hitz kopuruei erreparatzen diogu. Horrela, bat, bi, hiru, lau edo lau hitz baino gehiagoko terminoen itzulpen kopurua erakusten dugu, hauen portzentaiak ere agertzen direlarik. Taula horretan, denera itzuli diren terminoen portzentaia altua ez bada ere (%2,75), hitz bakarreko terminoei erreparatzen badiegu emaitza itxaropentsuak jaso ditugu, %22,23a itzuli dugu, hain zuzen ere.

Aurkitutakoak	Itzulitako terminoak	Portzentaia
Hitz bakarrekoak	884	% 22,23 (3.977)
Bi hitzekoak	897	% 4,41 (20.319)
Hiru hitzekoak	501	% 1,98 (25.213)
Lau hitzekoak	264	% 1,29 (20.398)
Hitz gehiagokoak	407	% 1,08 (37.539)
Denera	2.953	%2,75 (107.446)

15. taula: *Clinical Finding/disorder* hierarkiaren *"disorder" semantic tag*aren emaitzak jatorri-terminoen hitz kopuruka.

"Finding" (aurkikuntza) *semantic tag*aren emaitzei erreparatzen badiegu (16. eta 17. taulak), *"disorder"*n jaso ditugun emaitzak baino baxuagoak lortu ditugula nabaria da. Hitz bakarreko terminoen portzentaia mantendu bada ere, gainerako terminoetan beherakada nabaria izan da, eta horrela Kontzeptuen %6,54a baino ez dugu euskaratzea lortu. Horren arrazoia, terminoen izaeran datza, oso termino espezializatuak dira, domeinura egokitutako hiztegi batean ere agertzen ez diren terminoak.

Aurkitutakoak	Ordain kopurua	Parekatze kopurua
Hiztegietan	-	560
ZT Hiztegia	332	-
Erizaintzako Hiztegia	339	-
Anatomiako Glosategia	10	-
GNS-10	193	-
GNS-10 mapaketan	-	1.878
	Kopurua	**Portzentaia**
Ordainak denera	2.575	-
Kontzeptuak denera	2.172	%6,54 (33.204)

16. taula: *Clinical Finding/disorder* hierarkiaren *"finding" semantic tag*aren emaitzak.

Aurkitutakoak	Itzulitako terminoak	Portzentaia
Hitz bakarrekoak	382	% 22,21 (1.720)
Bi hitzekoak	79	% 0,90 (8.821)
Hiru hitzekoak	28	% 0,23 (12.034)
Lau hitzekoak	19	% 0,18 (10.552)
Hitz gehiagokoak	52	% 0,30 (17.440)
Denera	560	%1,11 (50.567)

17. taula: *Clinical Finding/disorder* hierarkiaren *"finding" semantic tag*aren emaitzak hitz kopuruka.

Procedure hierarkiari dagokionean SNOMED CT eta GNS-10en arteko mapaketaren falta nabaria da. 3.3 atalean aztertu dugun moduan, mapaketa honen izaerak ez ditu prozedurak parekatzen, gaixotasunen sailkapen bat baita, eta ez prozedurena. Horrela, hiztegien ekarpen urriarekin, Kontzeptuen %0,51a bakarrik itzultzea lortu dugu.

Aurkitutakoak	**Ordain kopurua**	**Parekatze kopurua**
Hiztegietan	-	297
ZT Hiztegia	278	-
Erizaintzako Hiztegia	197	-
Anatomiako Glosategia	4	-
GNS-10	4	-
GNS-10 mapaketan	-	0
	Kopurua	**Portzentaia**
Ordainak denera	362	-
Kontzeptuak denera	271	%0,51 (53.049)

18. taula: *Procedure* hierarkiaren emaitzak.

Aurkitutakoak	**Itzulitako terminoak**	**Portzentaia**
Hitz bakarrekoak	259	% 13,43 (1.929)
Bi hitzekoak	35	% 0,37 (9.425)
Hiru hitzekoak	3	% 0,02 (15.908)
Lau hitzekoak	0	% 0,0 (16.914)
Hitz gehiagokoak	0	% 0,0 (37.893)
Denera	297	%0,36(82.069)

19. taula: *Procedure* hierarkiaren emaitzak hitz kopuruka.

Jarraian 20. eta 21. tauletan *Body structure* hierarkiaren emaitzak erakusten ditugu. Kasu honetan Anatomiako Glosategiaren ekarpena bereziki aipagarria da, 1.981 ordain ematen baititu. Horrela, hitz bakarreko terminoen portzentai oso altua itzultzea lortu dugu, %37,30 hain zuzen ere.

Aurkitutakoak	**Ordain kopurua**	**Parekatze kopurua**
Hiztegietan	-	2.391
ZT Hiztegia	1.616	-
Erizaintzako Hiztegia	978	-
Anatomiako Glosategia	1.981	-
GNS-10	384	-
GNS-10 mapaketan	-	0
	Kopurua	**Portzentaia**
Ordainak denera	3.848	-
Kontzeptuak denera	2.201	%7,08 (31.105)

20. taula: *Body structure* hierarkiaren emaitzak.

Aurkitutakoak	**Itzulitako terminoak**	**Portzentaia**
Hitz bakarrekoak	1.004	% 37,30 (2.692)
Bi hitzekoak	1.050	% 9,12 (11.519)
Hiru hitzekoak	266	% 2,12 (12.575)
Lau hitzekoak	61	% 0,56 (10.903)
Hitz gehiagokoak	10	% 0,05 (21.631)
Denera	2.391	%4,03 (59.320)

21. taula: *Body structure* hierarkiaren emaitzak hitz kopuruka.

Azkenik, gainerako hierarkien emaitzak taula bakarrean elkarturik emango ditugu. *Clinical Finding/disorder*, *Procedure* eta *Body structure* hierarkiak itzuliz hasiko ginela esan dugu 3.1 atalean. Dena dela, aplikazioaren-garapen fase honetan ditugun baliabideak gainerako hierarkietara zabaltzeak ez digu lanik suposatzen, eta emaitza interesgarriak lor ditzakegu. Hortaz, jarraian erakusten ditugun 22. eta 23. tauletan hierarkia horietako emaitzak batuta agertzen dira.

Aurkitutakoak	**Ordain kopurua**	**Parekatze kopurua**
Hiztegietan	-	3.014
ZT Hiztegia	3.186	-
Erizaintzako Hiztegia	1.426	-
Anatomiako Glosategia	144	-
GNS-10	60	-
GNS-10 mapaketan	-	436
	Kopurua	**Portzentaia**
Ordainak denera	4.311	-
Kontzeptuak denera	3.178	%7,08 (31.105)

22. taula: Gainerako hierarkien emaitzak.

Aurkitutakoak	**Itzulitako terminoak**	**Portzentaia**
Hitz bakarrekoak	2.428	% 9,22 (26.327)
Bi hitzekoak	526	% 0,82 (64.477)
Hiru hitzekoak	41	% 0,13 (32.873)
Lau hitzekoak	6	% 0,03 (23.233)
Hitz gehiagokoak	13	% 0,04 (29.986)
Denera	3.014	%1,70 (176.886)

23. taula: Gainerako hierarkien emaitzak hitz kopuruka.

6.2 Emaitzak adibidetan

Atal honetan, emaitzen adibide batzuk emango ditugu. Ikus dezakegunez, 30. adibidean termino bakarrarentzat hainbat hiztegik ordain berdina eman dute. Kasu hau oso maiz gertatu da, 32. adibidean ere ikusi daitekeen bezala. Ez dira beti hiztegi berdinak ordain bera ematen dutenak, hauen artean aldaketak egoten dira. Normala den moduan, *"Microcephaly"* alterazioa, gaixotasunen sailkapenak itzuli du, eta, *"Metatarsus"* gorputz zatia, Anatomiako Hiztegiak.

Ingelesezko terminoa: *Microcephaly*
Ordaina: mikrozefalia
Iturburu hiztegiak:
ZT Hiztegia
Erizaintzako Hiztegia
GNS-10

30. adibidea: Parekatze bakarrean hainbat hiztegi

Ingelesezko terminoa: *Metatarsus*
Ordaina: metatartso
Iturburu Hiztegiak:
ZT Hiztegia
Anatomiako Hiztegia
Erizaintzako Hiztegia

32. adibidea: Ordainean egitura aldaketak.

Hiztegiek askotan ordainak ematerako garaian bat egiten badute ere, beste askotan ez dira bat etortzen, hiztegi bakoitzaren ekarpena beharrezkoa eginez.

Gainera, 31. adibidean ikusi dugun bezala, termino bakarrarentzat ordain ezberdinak jaso izan ditugu, zenbaki orokorretan agerian ikusi dugun moduan (parekatzeak baino ordain gehiago baitaude).

Ingelesezko terminoa: *Leprosy*
1. ordaina: legen beltz
Iturburu hiztegiak:
Erizaintzako Hiztegia
2. ordaina: legen
Iturburu hiztegiak:
ZT Hiztegia
Erizaintzako Hiztegia
3. ordaina: legenar
Iturburu hiztegiak:
ZT Hiztegia
Erizaintzako Hiztegia

31. adibidea: Parekatze bakarrean hainbat ordain

GNS-10en sortze data dela eta (1992), gaur egungo Euskaltzaindiaren hainbat arau betetzen ez direla ohartu gara, 33. adibidean ikus dezakegunez. Horrek arazo bat sortzen digu ordainen sendotasunean eta ongi aztertu beharko dugu gertaera hauen kasuistika, konponbidea aurkitu ahal izateko.

Ingelesezko terminoa: *Hemiplegia*
GNS-10: hemiplejia
ZT Hiztegia eta Erizaintzako Hiztegia: hemiplegia

33. adibidea: GNS-10en eguneratze falta.

Terminoen egitura aldaketak ere aurkitu ditugu, GNS-10ari lotuta (ikus 34. adibidea). Egitura honek ez ditu termino izateko irizpideak betetzen, tartean "," bat txertatzen duelako. Adibideari jarraiki, horren ordain zuzenetako bat "kanpoaldeko goiko ezpaina" litzateke.

Ingelesezko terminoa: *External upper lip*
Ordaina: goiko ezpaina, kanpoaldea
Iturburu Hiztegia: GNS-10

34. adibidea: Ordainean egitura aldaketak.

Azkeneko adibidea, 35. adibidea, termino luze baten ordainari dagokio. Aplikazioaren hastapenetan ez genuen termino luzeak itzuliko genituenik espero, baina emaitza interesgarriak jaso ditugu alde honetatik.

Ingelesezko terminoa: *Anterior external vertebral venous plexus*
Ordaina: aurre-alboko zain-sare bertebrala
Iturburu Hiztegia: Anatomiako glosategia

35. adibidea: Termino luze baten itzulpena

7 Ondorioak eta etorkizuneko lana

Atal honetan master-lan honen ondorio nagusiak aipatuko ditugu (7.1 atala), eta etorkizunerako gelditu zaizkigun atazen azalpena emango dugu (7.2 atala).

7.1 Ondorioak

Lehenik eta behin, master-tesi honetarako ezarri genituen helburuak bete ditugula aipatu beharra dago. Batetik, SNOMED CT euskaratzeko aplikazioaren diseinua bere osotasunean egin dugu, eta bestetik, aplikazioaren hastapenak inplementatu ditugu.

Aplikazioaren diseinuari dagokionean, TBX formatuaren egokitasuna frogatu dugu aplikazioaren inplementazioaren lehen urrats honetan.

Emaitzen atalean ikusi dugun bezala (6 atala), aplikazioaren hastapenak izateko, emaitza itxaropentsuak jaso ditugu. Honek aurrera jarraitzeko bidea irekitzen digu, algoritmoan zehaztutako urratsak inplementatzeko, hain zuzen ere.

Aipatu beharra dago, emaitzak esperotakoak baina dezente hobeak izan direla, hiztegien ekarpena eta GNS-10 eta SNOMED CT arteko mapaketaren ekarpena balio handikoa dela erakutsiaz. Mapaketari dagokioenean, GNS-10aren baitan gaixotasunak aurkitzeaz gain, sintomak, ez-ohiko aurkikuntzak eta kanpo-arrazoia duten zauriak ere barnebiltzen dituenez, SNOMED CTren gaineko ekarpena handia izan da.

Dena dela, txosten honetan agerian utzi dugun bezala, ez da ataza erraza hemen aurkeztu duguna. Algoritmoan zehazten den urrats bakoitzak aldez aurretik egin beharreko lan handia dakar, linguistikoa zein informatikoa, eta bidelagunen beharra azaleratu zaigu. Izan ere, oraindik algoritmoaren urrats asko inplementatzeko gelditzen zaizkigu.

Medikuntzaren domeinurako euskarazko baliabide lexiko osatu eta bateratu baten hastapenak egin ditugu, gerora osasungintzako langileek erabili ahal izan dezaten, eta horrela, euskararen erabilera osasungintzan sustatu dadin. Oraindik lan asko egiteke badago ere, hastapen honek motibazio gehigarria ematen du baliabide lexiko honen osatzeari begira.

Azkenik, aplikazioak sortu dituen (eta sortuko dituen) euskal ordainek adituen, hizkuntzalarien eta medikuen, berrikuspena beharko dute, gure ezagutza ez baita nahikoa horien zuzentasuna bermatzeko.

7.2 Etorkizuneko lana

Etorkizunerako lan asko gelditu zaizkigu mahai gainean, nahiz eta jadanik mami askoko lana burututa egon, eta garrantzitsuena dena, aplikazioaren diseinu osoa eginda egon.

TBX formatudun XML dokumentuari dagokionean, hiztegietako definizioak txertatu gabe gelditu dira. Hainbat definizio eskuratu baditugu ere, hauek ez dira XML dokumentuan txertatu. Ezta adibidetegitik jaso behar genituen erabilera-kasuak ere. Bi eginkizun hauek aplikazioaren azken faserako uztea erabaki dugu.

Lortu ditugun ordainen gainean irizpideak zehaztu behar ditugu, eta beharrezkoa denean ordain hauek aldatu beharko ditugu, hala nola, GNS-10ek ematen dizkigun ordainen lexemak lortzea eta berauek gordetzea.

SNOMED CTren terminoen artean medikuntza-alorreko terminoez gain, hiztegi orokorreko sarrerak ere aurki ditzakegu, adibidez, *"Speaks"* edota *"Speaks fluently"* (biak *Clinical Finding/disorder* hierarkiakoak). Termino hauek ez ditugu hiztegi espezializatuetan aurkituko, bai ordea hiztegi orokorretan. Hau horrela izanik, hiztegi orokor bat txertatzea aztertu behar dugu, honen onurak eta arazoak zehaztuz. Horretarako Elhuyar Hiztegia daukagu integraziorako prest. Hiztegiekin jarraituz, EuskalTerm hiztegi espezializatua eskuratu nahi dugu, hau ere aplikazioan txertatzeko. Hiztegi honetan ere, bio-zientzien domeinuko alorrei dagozkien terminoak bakarrik erabiliko genituzke.

Aplikazioari 4.1 atalean azaldutako algoritmoaren garatu gabeko funtzionalitateak gehitu behar dizkiogu, hala nola, sorkuntza morfologikoa, sintaxi mailako erregelak edota itzultzaile automatikoen egokitzapena.

Gainera, ordainen erabilera aztertzeko, domeinuko Corpus baten beharra daukagu, ordain hauen erabilerarekin batera, hauen gaineko konfiantza maila esleitu ahal izateko. Eta sortutako ordainen gaineko zuzenketa eta balioztatzea egiteko interfaze bat garatu beharko dugu, adituek interfaze horren gainean lan egin dezaten. Interfaze hori garatze-bidean dago.

Honetaz gain, 4.2 atalean diseinatu dugun bezala, elkarren arteko dependentziak dituzten terminoen arteko eguneraketak egiteko atala inplementatu beharko dugu; adibidez, adituek "aspirina" terminoa zuzenduko balute, demagun, "azido azetilsaliziliko" terminoarengatik, zuzenketa honek "aspirinaren aurkako erreakzio" terminoan eragina izan dezan. Eguneratze honetan ez dugu terminoen ordezkatzea proposatzen, baizik eta termino berri baten gehikuntza egitea, adituek balioztatu arte.

Bukatzeko, ordainen sorkuntzari dagokionean (sorkuntza morfologikoa zein sintaxi-mailakoa), aholkularitza eskainiko diguten adituak ere beharko ditugu.

Erreferentziak

Hocine Abdoune, Tayeb Merabti, Stéfan J. Darmoni, eta Michel Joubert. Assisting the Translation of the CORE Subset of SNOMED CT Into French. In Anne Moen, Stig Kjær Andersen, Jos Aarts, eta Petter Hurlen, editors, *Studies in Health Technology and Informatics*, volume 169, pages 819–823, 2011.

Iñaki Alegria, Unai Cabezon, Gorka Labaka, Aingeru Mayor, eta Kepa Sarasola. Matxin-Informatika: versión del traductor Matxin adaptada al dominio de la informática. In *XXVII CONGRESO DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL. Huelva*, 2011.

Xabier Arregi, Ana Arruarte, Xabier Artola, Mikel Lersundi, Gotzon Santander, eta Joseba Umbelina. TZOS: Terminologia Zerbitzurako On-line sistema. In Euskal Herriko Unibertsitatea UPV/EHU, editor, *Ugarteburu Terminologia Jardunaldiak 2010*, pages 136–153, 2010.

Jason Hunter eta Brett McLaughlin. JDOM, 2012. URL `http://www.jdom.com/`.

Asta Høy. Guidelines for Translation of SNOMED CT. Technical Report version 2.0, International Health Terminology Standards Development Organization IHTSDO, 2010.

International Health Terminology Standards Development Organisation IHTSDO. SNOMED CT Technical Implementation Guide. January 2012 International Release. Technical report, International Health Terminology Standards Development Organisation IHTSDO, 2012.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, eta Kepa Sarasola. Matxin, an Open-source Rule-based Machine Translation System for Basque. *Machine Translation*, 25:53–82, 2011. ISSN 0922-6567. URL `http://dx.doi.org/10.1007/s10590-011-9092-y`. 10.1007/s10590-011-9092-y.

International Health Terminology Standards Development Organisation. Mapping SNOMED CT to ICD-10 Technical Specifications. Technical report, International Health Terminology Standards Development Organisation, 2012.

Palle G. Petersen. How to Manage the Translation of a Terminology. Presentation, October 2011.

Yanhui Zhu, Huiting Pan, Lei Zhou, Wei Zhao, Ana Chen, Ulrich Andersen, Shuxiang Pan, Lixin Tian, eta Jianbo Lei. Translation and Localization of SNOMED CT in China: A Pilot Study. *Artificial Intelligence in Medicine*, 54(2):147–149, February 2012.

A Eranskina: TBX formatua

A.1 Hierarkien kode-baliokidetzak

Hierarkia	**Kodea**	**Semantic tag (ST)**	**Kodea**
Clinical Finding/disorder	010	disorder	011
		finding	012
Procedure/intervention	020	procedure	021
		regime/therapy	022
Organism	030	organism	031
Body structure	040	body structure	041
		morphologic abnormality	042
		cell	043
		cell structure	044
Substance	050	substance	051
Pharmaceutical/biologic product	060	product	061
Qualifier value	070	qualifier value	071
Observable entity	080	observable entity	081
Event	090	event	091
Situation with explicit context	100	situation	101
Social context	110	occupation	111
		person	112
		ethnic group	113
		religion/philosophy	114
		life style	115
		social concept	116
		racial group	117
Physical object	120	physical object	121
Specimen	130	specimen	131
Environment or geographical location	140	environment	141
		geographic location	142
		environment/location	143
Linkage concept	150	attribute	151
		link assertion	152
		linkage concept	153
Staging and scales	160	assessment scale	161
		tumor staging	162
		staging scale	163
Special concept	170	navigational concept	171
		namespace concept	172
		administrative concept	173
		special concept	174
Record artifact	180	record artifact	181
Physical force	190	physical force	191
Root Metadata Concept	200	foundation metadata concept	201
		core metadata concept	202

24. taula: TBXrentzako hierarkien eta *semantic tag*en kodeak

A.2 Kudeaketarako datu-kategorien balio posibleak

A.2.1 elementWorkingStatus

Kodea	en	eu	Azalpena
importedElement	Imported	Inportatua	SNOMED CTetik inportaturiko terminoak izango dira.
starterElement	Starter	Hasierakoa	Fidagarritasun maila bajuko terminoak izango dira. Adibidez erregelen bidez sortutako ordainak.
workingElement	Working	Lantzen	Fidagarritasun maila altuko terminoak izango dira. Adibidez hiztegien bidez lortutako ordainak.
consolidatedElement	Consolidated	Kontsolidatua	Adituek balioztatutako terminoak izango dira.
archiveElement	Archived	Artxibatua	Adituek baztertutako terminoak izango dira.

25. taula: *elementWorkingStatus* kode posibleen esanahia

A.2.2 transactionType

Kodea	en	eu	Azalpena
origination	Origination	Sorkuntza	Ordainak algoritmoaren bitartez sortu denean erabiliko da.
modification	Modification	Aldaketa	Adituek terminoaren gainean aldaketak egiten dituztenean erabiliko da.
importation	Importation	Inportazioa	SNOMED CTtik termino ala Kontzeptu bat inportatzen denean erabiliko da.
approval	Approval	Onarpena	Aditu zein hizkuntzalariek terminoa onartzen dutenean erabiliko da.

26. taula: *transactionType* kode posibleen esanahia

A.2.3 responsability

Kodea	en	eu	Azalpena
admin	Administrator	Kudeatzailea	Diseinuaz eta garapenaz arduratu den pertsona. Bakarra egon daiteke.
doctor	Doctor	Medikua	Medikuak izango dira eta aditu papera jokatuko dute. Bat baino gehiago egon daitezke.
linguist	Linguist	Hizkuntzalaria	Hizkuntzalariak izango dira, aukeran terminologian adituak. Terminoen zuzentasuna bermatuko dute. Bat baino gehiago egon daitezke.
other	Other	Besterik	Beste motako adituak izango dira. Bat baino gehiago egon daitezke.

27. taula: *responsability* kode posibleen esanahia

A.2.4 administrativeStatus

Kodea	en	eu	Azalpena
preferredTerm-admn-sts	Preferred Term	Hobetsitako terminoa	SNOMED CTren *Preferred Term*en baliokidea.
admittedTerm-admns-sts	Admitted Term	Onartutako terminoa	SNOMED CTren *Accepted Term*en baliokidea
deprecatedTerm-admns-sts	Deprecated Term	Baztertutako terminoa	Adituek baztertutako terminoak

28. taula: *administrativeStatus* kode posibleen esanahia

A.2.5 entrySource

Baliabide orokorra	**Baliabidea**	**Kodea**
Mapaketa	GNS-10 - SNOMED CT	000
Hiztegiak	Elhuyar Hiztegia	001
	ZT Hiztegia	002
	Erizaintza Hiztegia	003
	Administrazio Sanitarioko Hiztegia	004
	Anatomiako Hiztegia	005
	Gaixotasunen Nazioarteko Sailkapena	006
Sorkuntza-erregelak	Maila morfologikoko erregelak	101
	Sintaxi mailako erregelak	102
Besteak	Matxin-Med itzultzaile automatikoa	200

29. taula: *entrySource* kode posibleen esanahia