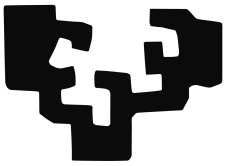
eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Doktorego-tesia

Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea

Olatz Perez de Viñaspre Garralda

2017

(cc)2017 OLATZ PEREZ DE VIÑASPRE GARRALDA (cc by 4.0)

EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea

Olatz Perez de Viñaspre Garraldak Maite Oronoz Anchordoqui eta Jon D. Patricken zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2017ko maiatza.

Amari eta aitari

Eskerrak

Ez dira gauzak apenas aldatu tesia hasi nuenetik! Tesia Donostitik ihesean hasi nuen, bizitokia nire bihotzeko Gasteizen ezarrita, eta orain, nire lagunek dioten bezala "erabat ñoñostituta". Bilakaera honetan tesiak eragin handia izan badu ere, eragin are eta handiagoa izan dute Informatikako txoko honetan aurkitatuko lankide eta lagunek.

Maite. Zuzendari eta laguna. Ezin nezakeen eskerren atalari ekin zurekin hasi gabe. Mila esker urte guzti hauetan lagundu eta babestu nauzun aldi guztiengatik. Zu gabe ez nintzateke gaur tesi-lan hau aurkezten egongo. Eskerrik asko bihotz-bihotzez!

IXAkideei... Aurreko batean, "zer da iXakide izatea?" galderari buruz eztabaida alaiean ibili ginen, eta ezin hitzezko definizio bat adostu. Zer da ba iXakide izatea? Niretzat argi dago lana baino gehio dela, izaera bat da. "Tara" bereziak elkartzen diren talde langile eta alaiaren parte izatea, ta euskaraz.

GuarderiXa edo DuquiXa, GuarderiXa+ edo GuarderiXa^ 2, GuarreriXa edo HipsteriXa, PiratiXa edo MaduriXa. Hamaika aiXialdi ta hamaika izen, ta helburua beti berdina: barre, barre eta barre!

Eta nola ez, bulegokideak. Krisi komitea beti laguntzeko prest, bai arazo linguistiko zein informatiko, ekipo ederra osatzen dugu! Azken aldian desitxuratzen ari bada ere... berriak etorriko dira.

FamiliXa, zuekin bai bizi-pozak bizitakoak! Londonen hasi ta, inbidiXa, bazkariak, afariak... Zuekin bizitako momentuetaz gogoratzean irribarretxoari ezin izaten diot eutsi :-) Berrartu beharko ditugu, ezta?

Gure aditu taldeari, hizkuntzalari zein mediku. Zuen erreferentzia eta ezagutza nirekin partekatzeagatik, eta beti laguntzeko prestutasuna erakusteagatik. Gure lanaren motibazio etengabea izan zarete.

Bilboko lankideei, irakaskuntzan emandako lehenengo urratsak horren errazak eta atseginak egiteagatik. Tesiaren azken txanparekin elkartu izanagatik, emandako babesagatik, tesia garaiz bukatzean eragin zuzena izan baituzue.

Sydneyn egindako egonaldia horren berezia egin zenuten guztiei. Bereziki Jon, Wendy eta Gorkari. Asko ikasi nuen zuei esker, eta munduko beste puntan egon arren, familiartean nengoela sentiarazteagatik.

Kuadrila. Urte hauetan guztietan hurbiletik urrunera, eta hurbilera berriz. Azken txanpa honetan, zuen hurbiltasuna aurrera egiteko indarra bilakatu delako.

Ta "Donostiko kuadrilari", ezin hitzez adierazi zer izan zareten ibilbide honetan guztian niretzat. Etapa berria hasi nuen zuekin, eta etapak luze iraun dezala!

Pixukideak, Donosti nire etxe berria egiteagatik. Egun txarretan etxera heltzean hor egoteagatik, eta baita onetan ere! Pixuan gustura egoteak lanean eragin zuzena izan duelako.

Ta Donostiarrei ere, jakina! Gasteiztarra izaten jarraituko badut ere, Donostia nire bihotzean sartzeagatik. Asko zarete Donostin jarraitzeko arrazoiak egunero erakusten dizkidazuenak, bai ardotxo ederrekin, bai "koplajurik gabe" Kantuari abordatzen, edota beste hamaika ekintzekin.

Etxekoak. Zuena bai babesa. Bizitzako momentu orotan hor egon zarete, gomendioak emateko prest, edo entzuteko bakarrik. Unerik latzenetan ere, indarrak bildu eta nirekin egoteagatik.

Urte hauetan guztietan, lagundu edota entzun nauzuen guztiei. Ez zarete gutxi, eta zerrendatzen hasiz gero, nire zuzendariaren tesia baino luzeagoa aterako da nirea. Eskerrik asko guztiei. Nire zalantzaz beteriko uneetan argia erakutsi didazuenoi, saturazioan saturazioz kaña bat hartzera eraman nauzunoi, eta azken txanpa gogor honetan burutik tesia kendu didazuenoi.

Eta bereziki eskerrak, eskerrak bukatu dela! (Ezeiza, 2002)

Esker instituzionalak

Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Sailari, ikerketalan hau egiteko emandako ikertzaileak prestatzeko bekarengatik (BFI-2011-389).

Laburpena

Testuen prozesamendu automatikoan hizkuntza-baliabideak ezinbestekoak dira, hala nola, datu-base lexikalak edo corpusak. Testu espezializatuetara mugatzen garenean aldiz, baliabide terminologikoek berebiziko garrantzia hartzen dute, eta osasun-zientzien domeinua ez da salbuespena.

Tesi-lan honetan, osasun-zientzien domeinuan euskarak duen lehentasunezko behar bati heldu diogu, eta terminoak automatikoki euskaratzeko sistemak garatu eta ebaluatu ditugu. Horretarako, SNOMED CT, terminologia kliniko zabala barnebiltzen duen ontologia, hartu dugu abiapuntutzat, eta EuSnomed deritzon sistema garatu dugu horren euskaratzea kudeatzeko.

EuSnomedek lau urratseko algoritmoa inplementatzen du terminoen euskarazko ordainak lortzeko:

- Lehenengo urratsak baliabide lexikalak erabiltzen ditu SNOMED CTren terminoei euskarazko ordainak zuzenean esleitzeko. Besteak beste, Euskalterm banku terminologikoa, Zientzia eta Teknologiaren Hiztegi Entziklopedikoa, Giza Anatomiako Atlasa eta Erizaintzako hiztegiak erabili ditugu.

- Bigarren urratserako, ingelesezko termino neoklasikoak euskaratzeko NeoTerm sistema garatu dugu. Sistema horrek, afixu neoklasikoen baliokidetzak eta transliterazio erregelak erabiltzen ditu euskarazko ordainak sortzeko.

- Hirugarrenerako, ingelesezko termino konplexuak euskaratzen dituen

KabiTerm sistema garatu dugu. KabiTermek termino konplexuetan agertzen diren habiaratutako terminoen egiturak erabiltzen ditu euskarazko egiturak sortzeko, eta horrela termino konplexuak osatzeko.

- Azken urratsean, erregeletan oinarritzen den Matxin itzultzaile automatikoa osasun-zientzien domeinura egokitu dugu, MatxinMed sortuz. Horretarako Matxin domeinura egokitzeko prestatu dugu, eta besteak beste, hiztegia zabaldu diogu osasun-zientzietako testuak itzuli ahal izateko.

Garatutako lau urratsak ebaluatuak izan dira metodo ezberdinak erabiliz. Alde batetik, aditu talde txiki batekin egin dugu lehenengo bi urratsen ebaluazioa, eta bestetik, osasun-zientzietako euskal komunitateari esker egin dugun Medbaluatoia kanpainaren baitan azkeneko bi urratsetako sistemen ebaluazioa egin da. Bereziki NeoTermen eta KabiTermen emaitzak izan dira azpimarragarriak, metodo erabat automatikoak erabiliz, doitasun altuko euskarazko ordainak sortzen baitituzte. EuSnomedek, algoritmoaren bitartez sortutako terminoak biltegiratzen eta kudeatzen ditu ere.

Sortutako terminologiaren balioa erakusteko bi aplikazio garatu ditugu: XuxenMed zuzentzaile ortografikoa eta osasun-txostenak euskaraz idazteko laguntza-prototipoa. Xuxen zuzentzaile ortografikoari hiztegia zabaldu diogu osasun-zientzietako terminoak ezagutu ditzan, oker faltsuak ekidinez. Osasun-txostenak euskaraz idazteko laguntza-prototipoari dagokionez, osasun-langileen lana errazte aldera zaintza klinikorako informazioa kudeatzen duen sistema bat sortu dugu, zeinak aurrez definitutako edukia erakusten duen eta idatzitako terminoak ezagutu eta itzultzeko gai den.

"Nazioarteko Doktoregoa" aipamena lortzeko Euskal Herriko Unibertsitatearen eskakizunei jarraituz, tesi-txosten honen ingelesezko bertsio laburtua ondorengo helbide honetan aurki daiteke:

`http://ixa2.si.ehu.eus/~operezdevina001/tesia/thesis_summary.pdf`

Gaien aurkibidea

Laburpena **vii**

Gaien aurkibidea **ix**

1 Tesi-lanaren aurkezpen orokorra **1**
1.1 Sarrera eta motibazioa 1
1.2 Lanaren kokapena 4
1.3 Helburuak 5
1.4 Tesi-txostenaren eskema 7
1.5 Tesi honen garapenetik atera diren argitalpenak 9

2 Kokapena eta aurrekariak **13**
2.1 Osasun-zientzietako terminologia 13
2.2 Terminoen sorkuntza automatikoa 15
2.3 Osasun-zientzietako baliabide terminologikoak 17
2.4 SNOMED CTren itzulpenak 19
2.5 Laburpena eta ondorioak 22

3 SNOMED CTren analisia **25**
3.1 Sarrera 25
3.2 SNOMED CTren inguruko azterketa bibliografikoa 28
3.3 Analisia 30
3.3.1 Egitura hierarkikoa 30

3.3.2 Aberastasun terminologikoa . . . 33
3.3.3 Terminoen deskribagarritasuna . . . 36
3.4 Ingelesezko bertsioa aukeratzeko arrazoiak . . . 38
3.5 Laburpena eta ondorioak . . . 40

4 EuSnomeden diseinua 41
4.1 Deskribapen orokorra . . . 41
4.2 Algoritmoa . . . 42
4.3 Biltegiratzea: TBX . . . 47
4.3.1 SNOMED CTrako TBX formatua . . . 48
4.3.2 Itzulpen-pareen datu-baserako TBX formatua . . . 53
4.4 Klase-diagrama . . . 55
4.5 Laburpena eta ondorioak . . . 58

5 Termino sinpleak 59
5.1 Baliabide lexikalak . . . 59
5.1.1 Aurrekariak . . . 60
5.1.2 Aurre-prozesaketa . . . 63
5.1.3 Hiztegien parekatzea . . . 65
5.2 Termino neoklasikoen sorkuntza . . . 69
5.2.1 Aurrekariak . . . 69
5.2.2 NeoTerm: oinarri-lerroa . . . 73
5.2.3 NeoTerm: transliterazio modulua . . . 80
5.2.4 NeoTerm: identifikazioa fintzeko irizpideak . . . 83
5.3 Ebaluazioaren diseinua . . . 84
5.3.1 Ebaluazio automatikoa . . . 85
5.3.2 Adituen ebaluazioa . . . 85
5.4 Emaitzak . . . 88
5.4.1 Ebaluazio automatikoaren emaitzak . . . 88
5.4.2 Adituen ebaluazioaren emaitzak . . . 94
5.5 Laburpena eta ondorioak . . . 100

6 Termino konplexuak 103
6.1 Termino konplexuen sorkuntza termino habiaratuen bidez . . . 103
6.1.1 AnaMed: Osasun-zientzietarako analizatzailea . . . 104
6.1.2 KabiTerm: termino konplexuen sorkuntza termino habiaratuak baliatuz . . . 109
6.2 Matxinen egokitzapena medikuntzaren domeinura . . . 123

6.2.1 Aurrekariak . . . 123
6.2.2 MatxinMed: sistemaren egokitzapena . . . 130
6.3 Ebaluazioaren diseinua . . . 141
6.4 Emaitzak . . . 146
6.4.1 Medbaluatoiaren emaitzak . . . 147
6.4.2 KabiTermen estaldura SNOMED CTn . . . 153
6.5 Laburpena eta ondorioak . . . 154

7 Zuzentzaile ortografikoa eta osasun-txostenak euskaraz 157
7.1 XuxenMed: osasun-zientzietarako zuzentzaile ortografikoa . . . 157
7.2 Osasun-txostenak euskaraz idazteko laguntza-prototipoa . . . 159
7.3 Laburpena eta ondorioak . . . 164

8 Ondorioak eta etorkizuneko lanak 165
8.1 Ondorio nagusiak . . . 165
8.2 Ekarpenak . . . 170
8.3 Etorkizuneko lanak . . . 174

Bibliografia 179

Eranskinak 195

A TBX formatuaren adibideak 195
A.1 SNOMED CTrentzako TBX formatuaren adibidea . . . 195
A.2 ItzulDBrentzako TBX formatuaren adibidea . . . 197

B Termino neoklasikoen sorkuntzarako erregelak 199
B.1 Euskarazko morfotaktika erregelak . . . 199
B.2 Termino neoklasikoen ingelesa-euskara transliterazio erregelak 201

C Termino habiratuen sorkuntzarako erregelak 203

1 Tesi-lanaren aurkezpen orokorra

1.1 Sarrera eta motibazioa

Euskal Autonomia Erkidegoan euskara eta gaztelania hizkuntzak dira koofizialak baina maiz, osasun-langileekin tratua euskaraz izaten dugun arren, haiek osasun-txostenak gaztelaniaz idazten dituzte. Euskal Herriko gainerako herrialdeetan egoera ez da hobea, eta erderak dira osasun-txostenak idazteko hizkuntza nagusiak. Batzuetan arrazoi desberdinak medio (hezkuntzarako erabilitako hizkuntza, osasunarekin lotutako terminoen ezagutza eza ...), erosoago sentitzen direlako egiten dute hori; besteetan, berriz, euskaraz osasunaren inguruan idazteko gai izan arren, beraien burua erdi behartuta sentitzen dute gaztelaniaz idaztera. Adibidez, Osakidetzak sistema zentralizatua du eta osasun-langile batek idazten duen txostena, gaixoa tratatzen duten gainerako osasun-langileek ere irakur dezakete. Euskaraz idatzi duenaren ondoren datozen irakurleak euskara ulertzeko gai ez badira, gaixoaren segurtasuna kolokan egon daiteke.

Osasun-langileen eta pazienteen arteko harremana gainerako herrialde elebidunetan (edo eleaniztunetan) nolakoa den aztertu nahi izan dugu. Kanadan, elebitasuna ofiziala den zonaldeetan, gaixoaren eta osasun-langilearen arteko harremanerako hizkuntza gaixoak erabakitzen du (Desjardins, 2003) eta osasun-txostenak hizkuntza horretan idazten dira. Belgikan, aldiz, hizkuntza-komunitate bakoitzak bere osasun-zerbitzua kudeatzen duen sistema ez-zentralizatua dauka, eta Bruselaren kasuan, zonalde elebiduna izanik, bi osasun-zerbitzu sistema eskaintzen dira: frantziar komunitatearena eta flan-

diarrena (Gerkens eta Merkur, 2010). Luxenburgon, herrialde txiki bezain anitza izanik, nahiz eta alemana, frantsesa, italiera, ingelesa eta portugesa hizkuntza zabalduak diren, *lingua franca* moduan frantsesa erabiltzen dute (European Observatory on Health Care Systems, 1999).

Gurean aldiz, hizkuntza komunitateak nahastuta daude, eta osasun-zerbitzu bateratua daukagu; *lingua franca* gisa gaztelania erabiltzen dugu Hego Euskal Herrian, baina horrek ez ditu herritar guztien hizkuntza-eskubideak bermatzen. Euskara hizkuntza ofiziala izan arren, osasun-langile guztiek ez dute euskara ezagutzen, eta euskaldunak diren osasun-langileek ahozko komunikazioa euskaraz egiten dute eta berehalako itzulpena egin beharra dute oharrak gaztelaniaz idazteko. Nolanahi ere, gaixoak oraingoz ez du sistematikoki bere arreta euskaraz lortzeko aukerarik eta horretarako bidea ireki nahiko genuke.

Osasun-langileen eta euskararen arteko harremana ezagutzeko asmotan, inkesta bat sortu genuen. Euskaraz aritzeko ohiturak eta arazoak ezagutzeko asmoz, 2014 urteko otsailean osasun alorreko 45 langileri galdeketa bat pasatu genien honakoen inguruan: lan-harremanetan egiten duten euskararen erabilera ezagutu nahi genuen eta, osasun txostenei dagokienez, txostenak edozein hizkuntzatan idazten dituztenean zein zailtasun topatzen dituzten ezagutu nahi genuen; eta txostenak euskaraz idazteko aukera izanez gero, tresnak zein behar ase beharko lituzkeen (ortografia-zuzentzailea, osasun-arloko terminologia euskaraz, txantiloiak etab.) zerrendatu nahi genuen.

Galdeketaren emaitzetatik ateratako ondorioak zerrendatuko ditugu jarraian. Hizkuntzaren erabilerari dagokionez, 41 osasun-langilek gaixoekin komunikazioa euskaraz egiten dute (beste 4ek ez dute erabiltzen) baina 5ek soilik idazten dituzte txostenak euskaraz. Galdetegian, osasun-txosten bat euskaraz idaztera gonbidatu genituen inkestatuak eta honako zailtasunak aurkitu zituzten: euskarazko terminologiaren falta (45etik 32 pertsonak), eredu falta (24 pertsonak) eta ohitura falta (15 pertsonak). Euskaraz idazteko beharrezkoak iruditzen zaizkien baliabideen artean galdetuta, lehentasun handiena eman zieten euskaraz osasun-zientzietako terminologia bateratu eta osatu bati (45etik 39 pertsonak), medikuntzara egokitutako ortografia-zuzentzaile bati (30 pertsonak) eta idazketarako eredu edo txantiloiak izateari (25 pertsonak). Galdetutakoen % 89 prest agertu zen tresna sortuko balitz erabiltzeko; gainerakoek, berriz, bere zalantzak agertu zituzten.

Beraz, inkestaren emaitzetatik ondoriozta dezakegu osasun-zientzietako terminologia bateratu eta osatu bat euskaraz izatea zein garrantzitsua den, gaixo eta osasun-langileen artean euskarazko harremana normalizatzeko, bai-

ta osasunaren munduan egunerokoan euskararen erabilera sustatzeko ere. Hori izan dugu tesi-proiektu honen motibazio nagusia, *osasun-zientzietako terminologia euskaraz izateko urratsak ematea, osasunaren alorrean euskararen normalizazioan aurrera pauso bat egiteko*.

Osasun-zientzietako kontzeptuak deskribatzeko euskarazko terminoak nahi baditugu, dagoeneko beste hizkuntzetarako sortutako sistema terminologiko bat oinarri hartzea iruditu zaigu egokiena. Sistema horien errepaso bat egin ondoren (2. kapitulua), *Systematized Nomenclature Of Medicine – Clinical Terms* edo SNOMED CT euskarara ekartzea erabaki dugu. SNOMED CT da ingelesez, espainieraz eta beste hainbat hizkuntzatan, osasun alorreko terminologia jasotzen duen baliabide zabalduena, eta, era berean, orain arte sortu den terminologia eleaniztun osatuentzat hartzen da[1]. Nolabait esateko, hizkuntza eta sistema desberdinen arteko osasun-txostenen adierazpen eta interpretazio automatikoa eta anbiguotasunik gabea ahalbidetuko duen hiztegi normalizatua da, eta hiztegi-sarreren arteko harremanak zehaztuta ditu. Osasun-txostenetan agertzen diren kontzeptuak, deskribapenak eta erlazioak barnebiltzen ditu. Ingelesa lantzeko sortu zen hasieran, eta 300.000 kontzeptu baino gehiago ditu definituak, baita horiek izendatzeko ingelesezko 1.000.000 termino baino gehiago ere.

Hurrengo kapituluan ikusiko dugun bezala, SNOMED CTren indar nagusia estaldura da. Domeinu klinikoan erabiltzen diren kontzeptuak barnebiltzen ditu, eta espezialitate gehienetako terminologia ere jasotzen du. Horretaz gain, kontzeptuen artean harremanak zehazten ditu: alde batetik, egitura hierarkikoa ematen dioten harremanak ditu, eta, beste aldetik, informazio semantikoa gehitzen duten harremanak, hala nola, agente eragilea (*causative-agent*), aurkikuntzaren tokia (*finding site*), etab. SNOMED CTren barruko harremanez gain, gaur egunean beste baliabide lexikal askoren kontzeptuekin parekatu dituzte SNOMED CTren kontzeptuak. Izan ere, *Unified Medical Language System*en (UMLS) metatesauroaren parte da, eta horri esker UMLSren metatesauro horretan dauden gainerako baliabide terminologikoekin parekatuta dago. Parekatze horri esker, SNOMED CT euskaraz izango bagenu, balibaide terminologiko horiek erabiltzen dituzten teknologiak eskura izango genituzke euskararentzat ere.

Tesi-lan honetan, SNOMED CTren euskarazko bertsioa automatikoki sortzeko algoritmo bat diseinatu eta inplementatu dugu. Algoritmo horrek lau

[1]http://www.snomed.org/snomed-ct/what-is-snomed-ct (2017ko maiatzaren 9an atzitua).

urrats ditu: i) dagoeneko euskararentzat eskuragarri dauden baliabide lexikal elebidunak (edo eleaniztunak) SNOMED CTren terminoekin parekatzen ditugu, ii) hitzen morfosemantika edo lexema-osaera eta transliterazio-erregela batzuk erabiliz, euskarazko termino neoklasikoak ematen ditugu, NeoTerm sistemaren bidez; terminoen luzera handitzen doan heinean, iii) termino batzuetan beste termino batzuk habiaratuak daudela baliatu dugu euskaratzepatroi batzuk definitzeko KabiTerm tresnan, eta azkenik iv) erregeletan oinarritutako Matxin itzultzaile automatikoa egokitu dugu, MatxinMed sortuz. Horretaz gain, egindako lana borobiltzeko, Xuxen zuzentzaile ortografikoa egokitu dugu osasun-zientzietarako (XuxenMed), eta osasun-txosten elebidunak lortzeko prototipo bat garatu dugu. Tesi-lanean egindako lan horiek zein kapitulutan deskribatu ditugun azalduko dugu 1.4 atalean.

1.2 Lanaren kokapena

Tesi-lan hau Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldean sortu da. Taldeak hogeita bederatzi urte inguru daramatza hizkuntzaren tratamendu automatikoa egiten, eta urte horietan batez ere euskara landu dugu. Denbora-tarte horretan, hizkuntzalarien eta informatikarien elkarlanari esker, euskararako sortutako baliabideak eta tresnak[2] ugariak izan dira. Hala nola, *Euskararen Datu-Base Lexikala* (EDBL) (Aldezabal *et al.*, 2001), MORFEUS analizatzaile morfologikoa (Aduriz *et al.*, 1998), hainbat analizatzaile sintaktiko, corpusak, hiztegi elektronikoak, itzultzaile automatikoak, sare semantikoa, eta abar.

Itzulpen automatikoen iker-lerroa IXA taldearen lerro estrategikoetako bat izan da 1988an sortu zenetik. Itzulpen automatikoa ez da ataza erraza, eta horretarako taldeak oinarrizko baliabideak eta tresnak sortzeko estrategia definitu zuen (Sarasola, 2000). Itzultzaile automatikoei dagokienez, esperientzia handia daukan taldea da: Matxin deituriko erregeletan oinarritutako itzultzaile bat garatu du, testuen espainiera-euskara eta ingelesa-euskara itzulpenak egiten dituena (Mayor *et al.*, 2011), baita EuSMT itzultzaile estatistikoa (Labaka, 2010) eta SMatxinT ere, erregeletan oinarritutako metodoak eta metodo estatistikoak erabiltzen dituen itzultzaile hibridoa (Labaka *et al.*, 2014).

[2]Informazio zabalagoa http://ixa.si.ehu.eus web-orrian (2017ko maiatzaren 9an atzitua).

IXA taldean, terminologiaren lanketan, erregistro akademikoan sortzen diren terminoak izan dira ikergai nagusia azken urteetan. EHUko Terminologia Sareak Ehunduz (TSE) programan parte hartuz, komunikazio espezializatuan erabiltzen diren terminoak lantzen ari dira unibertsitateko jarduera akademikoetan erabiltzen diren testuetatik abiatuta, (San Martin, 2013), baita jarduera akademiko horretan sortutako corpusak ere (Zabala *et al.*, 2008).

Ikerketa-lerro horiek guztiak, tesi-lan honetan lagungarriak izan dira, eta gure sistemen garapenean eragin zuzena izan dute.

1.3 Helburuak

Tesi-lan honen helburu nagusia *osasun-zientzien domeinuko testuak automatikoki prozesatzeko euskarazko baliabideak sortzea* da. Motibazioan ikusi dugun bezala, testu espezializatuen prozesamenduan oinarrizko baliabidea terminologia da. Horrenbestez, tesi-lan honen eginkizun nagusia izango da euskarazko terminologia zabal eta bateratua lortzea, teknika automatikoak erabiliz. Hala ere, helburua ez da bakarra, eta jarraian zeharkako helburuak zerrendatuko ditugu.

- **SNOMED CT sakonki ezagutzea**:

  SNOMED CT ontologia erraldoiaren terminologia euskaraz emateko hautua egin dugu, eta, horrenbestez, berau sakonean ezagutzea ezinbestekoa zaigu, eskaintzen dituen abantailak aprobetxatzeko, eta, jakina, euskaratzea diseinatzeko.

- **Dagoeneko euskaraz osasunaren alorrean dauden baliabide lexikalak ezagutzea eta bateratzea**:

  Osasun-zientzien arloan, beste hainbat arlotan bezala, hiztegi espezializatu asko daude eskura, hala nola Euskalterm, Zientzia eta Teknologia hiztegia, etab. Osasun-langileen kezketako bat terminologia bateratuaren falta da, hiztegi ezberdinak izatean erreferentziazko terminologia nahasgarria baita. SNOMED CTren euskaratzeari ekiteko, eskuragarri dauden baliabide horiek guztiak berrerabili nahi ditugu, eta guztien ordainak bateratu nahi ditugu.

- **Osasun-alorreko terminoak automatikoki euskaratzeko sistemak garatzea**:

  Terminoen konplexutasunaren arabera egin nahi dugu euskaratzea, hau da, hitz bakarrekoak euskaratzen hasi, eta hitz anitzekoetan dagoeneko euskaratutakoa erabili. Horrela, hiru sistema nagusi garatuko ditugu:

  - **Termino neoklasikoak euskaratzeko sistema bat garatzea**:

    Termino neoklasikoak, jatorri grekoa edo latindarra duten morfemez osatuta dauden terminoak dira, adibidez, *hipogluzemia* edo *fotodermatitis*. Horien izaera ezagututa (ez dira apenas aldatzen hizkuntza batetik bestera), horiek euskaraz emateko sistema bat garatu nahi dugu.

  - **Termino konplexuen izaera ezagutzea, eta horien euskaratzerako sistema bat garatzea**:

    SNOMED CTn dauden hitz bat baino gehiagoko terminoak (termino konplexuak) euskaratzeko sistema bat garatu nahi dugu, Itzultzaile Automatikoetatik haratago, baliabide lexikalak eta sortutako termino neoklasikoak erabiltzea helburu duena. Euskaratzeko egiturak aztertzeko, analizatzaile bat sortuko dugu, SNOMED CTren terminoak aztertuko dituena.

  - **Itzultzaile automatiko bat osasun-zientzien domeinura egokitzea**:

    Jadanik garatutako itzultzaile automatiko bat osasun-zientzien domeinura egokitu nahi dugu, aurreko sistemek sortutako terminoak eta baliabide lexikaletatik jasotakoekin aberastuta. Horrela, itzultzaile automatikoak terminoak euskaratzean baliagarriak ote diren aztertu nahi dugu.

- **SNOMED CTren euskaratzea kudeatuko duen sistema bat garatzea**:

  SNOMED CTren eduki terminologikoa oso handia da, eta euskaratzeko beharrezko informazioa biltegiratzeaz eta euskaratze-prozesuaz arduratuko den sistema bat garatuko dugu.

- **Osasun-zientzietako euskal komunitatea euskaratzean inplikatzea, ebaluazioan parte hartuaz**:

  Tesi-lan honetan osasun-zientzietako euskal komunitatea inplikatu nahi dugu, eta horretarako gure sistemak ebaluatzeko dinamika bat martxan jarriko dugu, jendearen aktibazioa bilatzeko.

- **Osasun-txostenak euskaraz idazten laguntzeko tresnak prestatzea**:

  Osasungintzako langileei egindako inkestatik ondorioztatu bezala, osasun-txostenak euskaraz idazteko zailtasunak ez dira gutxi. Horregatik, euskarazko idazketa xamurragoa izan dadin, idazketarako laguntzak garatu nahi ditugu aplikazio informatikoen bidez. Besteak beste, osasun-zientzietarako zuzentzaile ortografiko bat egokitu nahi dugu, eta idazketan lagunduko duen prototipo bat diseinatu eta garatu nahi dugu.

1.4 Tesi-txostenaren eskema

Tesi-txostena zortzi kapitulutan banatu dugu. Kapitulu bakoitzean landutako gaiak deskribatuko ditugu labur-labur hurrengo lerroetan.

- 1. kapitulua – Tesi-lanaren aurkezpen orokorra:

  Lehenengo kapituluan, tesi-lan honetan aztertu dugun gaiaren aurkezpen orokorra egin dugu. Horretarako, izandako motibazioa eta izan ditugun helburuak azaldu ditugu. Bukatzeko, tesi-lan honen inguruan argitaratutako artikuluak zerrendatu ditugu.

- 2. kapitulua – Kokapena eta aurrekariak:

  Hasteko, osasun-zientzietako terminologiaren ezaugarriak ikusiko ditugu, eta terminoen sorkuntza automatikoan erabili diren zenbait hurbilpen aurkeztuko ditugu. Horren ostean, gaur egunean gehien erabiltzen diren osasun-zientzietako baliabide terminologikoak aurkeztuko ditugu. Bukatzeko, beste hizkuntzetan egin diren SNOMED CTren itzulpen-prozesuei gainbegiratu bat emango diegu.

- 3. kapitulua – SNOMED CTren analisia:

  Euskaratzea diseinatu ahal izateko, SNOMED CTren sakoneko analisi kuantitatiboa egin dugu. Ingelesezko bertsio internazionala aztertu dugu, eta honen egitura hierarkikoari, aberastasun terminologikoari (sinonimo kopuruari) eta terminoen token kopuruari erreparatu diogu. Bukatzeko, espainierazko bertsioaren gabezien berri emango dugu.

- 4. kapitulua – EuSnomeden diseinua:

  SNOMED CTren eduki terminologikoa ingelesetik abiatuta euskaratzeko garatu dugun EuSnomed aplikazioaren diseinua aurkeztuko dugu. Aplikazio horretarako diseinatu dugun algoritmoa azalduko dugu, algoritmoak dituen lau urrats nagusiekin batera. Horretaz gain, EuSnomedek informazioa biltegiratzeko egokitu dugun formatua aurkeztu dugu.

- 5. kapitulua – Termino sinpleak: baliabide lexikalak eta termino neoklasikoak:

  Termino sinpleak euskaratzeko garatu ditugun euskaratze-algoritmoaren lehenengo bi urratsak aurkeztuko ditugu. Lehenengo urratsak eskuragarri dauden euskararako baliabide lexikal elebidunak erabiltzen ditu. Bigarren urratserako, aldiz, termino neoklasikoak euskaratzeko sistema bat garatu dugu (NeoTerm), zeina termino horien afixuen baliokidetzaz eta transliterazio-erregelez baliatzen den ingelesetik euskarazko terminoak sortzeko.

- 6. kapitulua – Termino konplexuak: termino habiaratuak eta itzultzaile automatikoak:

  Ingelesezko termino konplexuak euskaratzeko, euskaratze-algoritmoaren azkeneko bi urratsei helduko diegu. Alde batetik, termino habiaratuetan[3] oinarritzen den sistema bat garatu dugu KabiTerm deiturikoa. Bestetik, Matxin itzultzaile automatikoa osasun-zientzien domeinura egokitu dugu. Bi sistemak ebaluatzeko, osasungintzako euskal komunitatea inplikatzen duen ebaluazioa egin dugu (Medbaluatoia).

[3]Termino konplexu baten barnean beste termino batzuk agertzen direnean, horiei termino habiaratu deritzegu.

- 7. kapitulua – Zuzentzaile ortografikoa eta osasun-txostenak euskaraz idazteko prototipoa:

  Batetik, osasun-zientzietarako egokitu dugun zuzentzaile ortografikoa aurkeztuko dugu. Bestetik, osasun-txostenak euskaraz idazteko garatu dugun lehen prototipoaren berri emango dugu, honekin osasun-txosten elebidunak lortzeko bidea ireki dugularik.

- 8. kapitulua – Ondorioak eta etorkizuneko lanak:

  Lehenik, aurreko kapituluetan egindako lanetik ateratako ondorioak eta tesi-lan honen ekarpenak zeintzuk diren laburbilduko dugu. Azkenik, eskuartean gelditu zaizkigun etorkizunean egiteko lanak zehaztuko ditugu.

Aurkeztutako kapituluez gain, beste honako eranskin hauek aurki daitezke tesi-txosten honetan:

- A eranskina – TBX formatuaren adibideak.
- B eranskina – Termino neoklasikoen sorkuntzarako erregelak.
- C eranskina – Termino habiaratuen sorkuntzarako erregelak.

1.5 Tesi honen garapenetik atera diren argitalpenak

Tesi-lan honen garapenean, hainbat artikulu argitaratu ditugu. Jarraian, argitalpen horiek zerrendatzen ditugu, kapituluen arabera sailkatuta:

- 4. kapitulua – EuSnomeden diseinua:
  - Perez-de-Viñaspre O., eta Oronoz M. **Translating SNOMED CT Terminology into a Minor Language**. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, 38–45. Association for Computational Linguistics. Gothenburg, Suedia, 2014.

  - Perez-de-Viñaspre O., eta Oronoz M. **An XML Based TBX Framework to Represent Multilingual SNOMED CT for Translation**. *12th Mexican International Conference on Artificial Intelligence, MICAI 2013*. Lecture Notes in Artificial Intelligence, 8265 lib., 419–429. Springer, ISBN 978-3-642-45113-3. Mexiko DF, Mexiko, 2013.

- 5. kapitulua – Termino sinpleak: baliabide lexikalak eta termino neoklasikoak:

  - Perez-de-Viñaspre O., Oronoz M., Agirrezabal M., eta Lersundi M. **A finite state approach to translate SNOMED CT terms into Basque using medical prefixes and suffixes**. *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, 99–103. St Andrews, Eskozia, 2013.
  - Perez-de-Viñaspre O., eta Oronoz M. **SNOMED CT in a language isolate: an algorithm for a semiautomatic translation**. *BMC Medical Informatics and Decision Making*, 15 lib., 2 zenb., S5. BioMed Central. 2015.

- 6. kapitulua – Termino konplexuak: termino habiaratuak eta itzultzaile automatikoak:

  - Perez-de-Viñaspre O., eta Oronoz M. **Osasun-zientzietako terminologiaren euskaratze automatikoaren ebaluazioa, osasungintzako euskal komunitatea inplikatuz**. *II. IkerGazte, Nazioarteko Ikerketa Euskaraz*. Udako Euskal Unibertsitatea. Durango, Euskal Herria, 2017.

- 7. kapitulua – Zuzentzaile ortografikoak eta osasun-txostenak euskaraz idazteko prototipoa:

  - Perez-de-Viñaspre O., Oronoz M., eta Patrick J. **Osasun-txosten elebidunak posible ote?**. *I. IkerGazte, Nazioarteko Ikerketa Euskaraz*, 730–738. Udako Euskal Unibertsitatea, ISBN 978-84-8438-539-4. Durango, Euskal Herria, 2015. IkerGazte Sari Berezia.

Beste honek ere, kapitulu zehatz batekin lotura ez badu ere, tesiarekin zerikusia du:

– Perez-de-Viñaspre O., eta Labaka G. **IXA Biomedical Translation System at WMT16 Biomedical Translation Task**. *Proceedings of the First Conference on Machine Translation (WMT16)*, 477–482. Association for Computational Linguistics. Berlin, Alemania, 2016

2

Kokapena eta aurrekariak

Tesia kokatzea du helburu kapitulu honek, baita bertan egindako lanen aurrekariak azaltzea ere. Osasun-zientzietako terminologiaren, honen sorkuntzarako ikuspegien, sorkuntza automatikoaren eta dagoeneko inplementatuta dauden sistema terminologikoen berrikuspen bat egingo dugu.

Aurrera jarraitu aurretik, tesi-lan honetan "kontzeptu" eta "termino" hitzak maiz azalduko zaizkigunez, komenigarria da hauek definitzea:

- kontzeptua: "ezagutzaren unitate bat, ezaugarrien konbinaketa bakarra" (ISO 1087-1:2000)
- terminoa: "kontzeptu orokor baten hitzezko izendatzea domeinu zehatz batean" (ISO 1087-1:2000)

2.1 Osasun-zientzietako terminologia

Osasun-zientzien domeinuan, eta bereziki medikuntzarenean, kontzeptu berriei izena emateko beharra etengabekoa da. Horretan eragin handia daukate medikuntzako aurrerakuntza azkarrek eta medikuntzan sortzen diren egoera berriek. Egoera honetan, eta kontuan izanik gaur egun ikerketan ingelesa dela hizkuntza nagusia, kontzeptu berri horiei lehen aldiz ingelesez eman ohi zaie izena (ten Hacken eta Panocová, 2015).

Kontzeptuen lehen izendatzerako, hitz-sorkuntza (*word formation*) edo terminologiara ekarrita, termino-sorkuntza deritzon erregela-sistema baliatzen da. Erregela-sistema horren bidez, termino berriak sortzen dira jadanik

existitzen diren elementu lexikalak baliatuz. Termino-sorkuntzarako erabiltzen diren metodo erabilienak adiera-zabalkuntza (*sense extension*) eta mailegatzea (*borrowing*) dira (ten Hacken eta Panocová, 2015).

Adiera-zabalkuntzaren adibide argia aurkitu dezakegu ingelesezko *cell* terminoan. Jatorrian monasterioko edo kartzelako gela txiki bat izendatzeko erabiltzen zen (latinetik *cellulae*, "lekugune txikia") eta metafora erabiliz, hitz horri adiera berri bat esleitu zioten bizidun guztien egitura-unitate mikroskopikoa izendatzeko.

Maileguak ere asko erabili izan dira medikuntzako termino-sorkuntzan. Historikoki, latinera zen nazioarteko komunikazioan erabiltzen zen hizkuntza. Gaur egunean, aldiz, ingelesa bilakatu da komunikazio internazionaletako hizkuntza eta latinetik eta grekeratik hartutako maileguak erabili izan dira ingelesezko kontzeptuak izendatzeko. Termino neoklasiko horiek Europako hizkuntzetarako abantaila izan dira terminoen ulergarritasunari begira, guztientzat baitira maila antzekoan ulergarriak.

Aldakortasunari dagokionez, terminoa, hizkuntzaren beste edozein elementu moduan, aldakortasunarekin lotuta dago (Cabré, 1999). Gainera, terminoek denboran eboluzio bat izan dezakete, deskribatzen duten kontzeptuaren inguruan informazio gehiago lortzen den heinean. Adibidez, kromosoma gehigarri baten agerpenak "Down-en sindromea" sortzen zuela jakin zenean, "trisomia 21" terminoa agertu zen kontzeptuaren kausalitatean arreta ipintzeko (Westman, 2006). Gaur egunean bi terminoak ("Down sindromea" eta "trisomia 21") erabiltzen dira sinonimo gisa gaixotasuna izendatzeko. Bestalde, erregistroarekin zerikusia duen sinonimia ere oso ohikoa da osasun-alorrean. León-Araúz-ek (2015) adierazten duenez, mediku-gaixo elkarrizketa batean, medikuak erabiltzen dituen terminoak ahalik eta ulergarrienak izango dira, eta aldiz, konferentzia batean erabiltzen dituen terminoak estandarizatuak izango dira gehienetan.

Terminoen aldakortasunaz areago, ikerketak aurrera egiten duen heinean, kontzeptu berrien izendatzeko beharra dago, eta horren ondorioz, baliabide terminologikoak etengabean eguneratu behar dira (Bollegala *et al.*, 2015).

Osasunaren alorrean egiten den ikerketan hizkuntza nagusia ingelesa bada ere, termino berriak gainerako hizkuntzetara eramateko presioa handia da. Izan ere, osasungintzan egiten den komunikazioa ez da ikerlarien mailan bakarrik gelditzen, eta paziente zein familiekin hitz egiterako garaian gainerako hizkuntzetan kontzeptuak izendatzeko beharra sortzen da (ten Hacken eta Panocová, 2015).

Euskararen kasuan, normalizazio bidean dagoen hizkuntza bat izanik, alor

askotako terminologia erabat garatu gabe dago. Hori da osasun-zientzien domeinuaren kasua. Gaur egunean, ez dugu terminologia bateratu eta kohesionatu bat, eta hizkuntza minorizatua den heinean, terminologia-plangintza bat behar du, testuinguru soziolinguistikoa kontuan hartuko duena (Cabré, 2003).

Dagoeneko lan handia egin da osasun-zientzien domeinuko terminologia euskaraz lantzeko euskaratzeko eta horren erakusle da Zabala *et al.*-ek (2016) egindako ekarpena. Unibertsitateko osasun-zientzietako ikasleak euskararen normalizazio-prozesuaren agente aktibo bilakatzeko formakuntzaz dihardute lan horretan, hori bideratzeko "Komunikazioa euskaraz: osasun-arloa" ikasgaia baliatuta bidez. Ez da lan isolatu bat, eta osasun-zientzietan euskararen normalizazio prozesuan oso lagungarriak diren baliabide lexikalen sorkuntza etengabekoa da, hala nola, berriki argitaratu den Anatomiako Atlasaren euskarazko argitalpena (Zabala *et al.*, 2012) edo Euskal Herriko Unibertsitatean abian den Terminologia Sareak Ehunduz programa[1], non unibertsitateko jarduera akademikoetan erabiltzen diren testuetatik abiatuta, komunikazio espezializatuan erabiltzen den terminologia eta fraseologia erreala ikusgarri egin nahi diren.

2.2 Terminoen sorkuntza automatikoa

Terminoen sorkuntza automatikoaz ari garenean, jadanik izendapenen bat duen kontzeptu bati termino berriak esleitzeaz ari gara. Hau da, kontzeptu bat izendatzeko aurrekariak izanda (eta horietan oinarrituta), ordainak sortzea. Kontuan izan behar dugu terminoak ez direla itzultzen, baizik eta beste hizkuntzetako ordain egokiak sortzen dira kontzeptu bera izendatzeko, eta kasu honetan, prozesu horretaz ari gara.

Adituek termino-sorkuntza honi bigarren mailako termino-sorkuntza deritzote (Sager, 1997). Izan ere, lehen termino-sorkuntza, kontzeptu bat lehenengo aldiz izendatzen denean egiten da, eta gainerako kasuetan, bigarren mailako termino-sorkuntza egiten da. Bigarren mailako termino-sorkuntzan beraz, kontzeptua izendatzeko eredu bat badago jadanik, eta horrela, proposamenak egin daitezke modu kontrolatuan (Müller, 2015). Euskararen eta osasun-zientzien kasuan, terminologiaren sorkuntza ia osoa bigarren mailakoa izaten da, aurretiaz kontzeptua izendatzeko terminoak egoten baitira

[1]https://www.ehu.eus/ehusfera/tse/ (2017ko maiatzaren 9an atzitua).

euskaran eragina duten ingelera zein espainiera bezalako hizkuntza nagusietan (Zabala *et al.*, 2016).

Terminoak hizkuntza berri batera ekartzea ez da lan erraza izaten. Izan ere, ataza horretarako prest dauden aditu elebidunak aurkitzea ez da lan erraza, eta biomedikuntzaren domeinua horren zabala izatean, are zailagoa da nahikoa aditu aurkitzea azpidomeinu guztien terminologia eskuz sortzeko (Bollegala *et al.*, 2015). Horrenbestez, teknika automatikoak erabiltzeak lana bideragarriagoa egiten du, automatikoki terminoen ordainak lor daitezkeelako, eta adituen lana berrikuspenera mugatzen delako.

Terminologia automatikoki sortzeko garaian, bi metodo multzo bereizten dituzte Langlais *et al.* (2008) lanean.

Alde batetik metodo sortzaileak (ingelesez *generational*) daude, non pertsonen ezagutzan oinarrituta edo corpusean oinarrituta termino berrien edo ezezagunen ordainak lortzen dituzten. Adibidez, Schulz *et al.*-ek (2004) portugesez idatzitako medikuntzako terminoetatik espainierakoak sortzeko erregelak definitu zituzten eta Claveau eta Zweigenbaum (2005) lanean, aldiz, ikasketa automatikoko teknikak erabiliz ingelesetik abiatuta frantsesezko terminoak (eta alderantziz) sortzeko erregelak ikasi zituzten. Langlais *et al.* (2008) lanean, lexikoi elebidun bat erabiltzen dute ikasketa analogikoa[2] (Lepage, 1998; Stroppa eta Yvon, 2005) erabiliz eredu bat sortzeko. Navigli *et al.* (2003) lanean, WordNet-ek (Miller, 1995) eta beraiek garatutako OntoLearn ontologiak emandako kontzeptuen arteko harremanak baliatuz, termino konplexuen (hitz anitzeko terminoen) ordainak lortzen dituzte. OntoLearn ontologia corpus espezializatu elebakarrak erabiliz sortzen dute, eta, Hizkuntzaren Prozesamenduko (HP) teknikak eta teknika estatistikoak erabiliz, terminoen desanbiguazioa eta kontzeptuen arteko harremanak eraikitzen dituzte.

Bestetik, metodo ez-sortzaileak daude, eta horiek corpusetako hitzen itzulpenen identifikazioan oinarritzen dira. Metodo horiek aldiz, mugatuak dira, corpusean agertzen diren termino-ordain pareak baino ezin baitituzte sortu. Mota horretako metodo gehienek corpus paraleloak erabiltzen dituzte terminoen ordainak identifikatzeko. Adibidez, Deléger *et al.*-ek 2006ko lanean corpus paraleloetako hitzen lerrokatzea erabiltzen dute terminologia eleaniztunak zabaltzeko, eta 2009ko lanean medikuntzako terminologia sortzeko (ingelesetik). 2010eko lanean bi geruzetako termino-sorkuntzarako metodoa garatu dute, MedlinePlus-eko (Miller *et al.*, 2000) terminologia frantsesera

[2]Proportzio analogiko bat lau elementuen arteko harremana da [x : y = z : t], non x-ren eta y-ren arteko harremana, z-ren eta t-ren artekoaren berdina den

itzultzeko. Lehenengo hurbilpenean, ezagutzan oinarritzen dira UMLSren metatesauroko informazio kontzeptuala erabiliz (metodo sortzailea). Bigarren hurbilpenean, corpusean oinarritutakoan, aurreko lanean azaldutako metodoa erabiltzen du (Deléger *et al.*, 2009). Bi hurbilpenak konbinatuz, sistemaren estaldura hobetzea lortzen dute (0.51 bien artean), sistemaren doitasuna 0.70 inguru mantenduz.

Azaldu dugun bezala, metodo ez-sortzailearen kalitatea corpusen tamainarekin zuzenki lotuta dago, corpusean aurkitzen ez diren terminoei ezin baitizkiete ordainak eman. Corpus paralelo espezializatu handiak lortzea aldiz, ez da lan erraza izaten. Hori dela eta, corpus konparagarriak erabiltzen dituzten lanak ere argitaratu dira, baina ataza zailtzeaz gain, erroreak sortzeko aukera areagotzen dute. Fung eta Yee-ek (1998) hitzen itzulpenak lortzeko, Informazioaren Berreskurapeneko (*Information Retrival*) teknikak erabiltzen dituzte, itzuli beharreko hitzen inguruko hitzen maiztasunenak baliatuz. Delpech-ek eta lagunek, corpus konparagarriak erabiltzen dituzte lexikoi elebidunak sortzeko (Delpech *et al.*, 2012). Hitzen konposaketaz baliatzen dira, eta morfema mailako baliokidetzak aurkituz, termino-ordain pareak identifikatzen dituzte.

2.3 Osasun-zientzietako baliabide terminologikoak

Azken hamarkadetan baliabide elektronikoak ikaragarri zabaldu dira, eta bereziki testu elektronikoen gorakada da nabarmentzekoa. Gorakada horri esker, ezagutza ordenagailuen eskura jarri da. Ordenagailuek testuan jasota dagoen ezagutza erabili ahal izateko egin beharreko lana, ordea, ez da berehalakoa. Testuan dagoen ezagutza modu automatikoan kudeatu ahal izateko, baliabide terminologikoak oso erabiliak izan dira (Alani *et al.*, 2003; Savova *et al.*, 2010).

MeSH (*Medical Subject Headings*) hiztegi kontrolatu zabal bat da, osasun- eta biologia-zientzietako aldizkarietako artikuluak eta liburuak indexatzeko erabiltzen dena (Lipscomb, 2000). 25.186 gai-sarrera (ingelesez *subject heading*) barnebiltzen ditu, eta hauentzako definizio edo deskribapen bat, baita beste burukoekiko loturak eta sinonimo zerrenda bat ere.

Gaixotasunen Nazioarteko Sailkapena (GNS) edo ingelesez *International Classification of Diseases and Related Health Problems* Munduko Osasun

Erakundeak (*World Health Organization*) sortutako sailkapen bat da, mundu mailan diagnostikoak sailkatzeko erabiltzen dena. Sailkapen-sistema bat da, eta osasun-txostenetan aurkitzen den informazioa kontuan hartuta, diagnostikoa kodetzeko erabiltzen da. Bere helburu nagusia osasun-txostenak sailkatzea da.

Systematized Nomenclature Of Medicine – Clinical Terms edo SNOMED CT (IHTSDO, 2014) osasungintza klinikoko terminologia eleaniztun zabalduena dela esan daiteke. SNOMED CTk, GNSrekin alderatuta, helburu klinikoak kontuan hartuta, osasun-txostenetan dagoen informazioa jasotzeko helburua dauka. Adibidez, SNOMED CTren bitartez, GNS diagnostiko bat automatikoki lor daiteke (Bowman, 2005). SNOMED CTren indar nagusia bere estaldura da (Elkin *et al.*, 2006), izan ere, medikuntzaren esparruko espezialitate askotako terminologia barnebiltzen du, eta horrekin, osasun-txosten elektronikoetan erabiltzen den terminologia zabala. Gainera, munduko herrialde askotan zabaltzen ari den erreferentziazko terminologia da, eta herrialde bakoitzak bere hizkuntza eta beharretara egokitzen du[3]. Dagoeneko SNOMED CT 50 herrialde baino gehiagotan erabiltzen da[4].

Unified Medical Language System (UMLS) aipatutako terminologia guztiak eta askoz gehiago barnebiltzen dituen baliabide bat da (Bodenreider, 2004). Baliabide terminologiko horien guztien arteko mapaketak eskaintzen ditu eta biomedikuntzako kontzeptuen metatesauro eta ontologia bat osatzen du. Horretaz gain, Hizkuntzaren Prozesamendurako (HP) baliabideak ere eskaintzen ditu, helburua biomedikuntzako terminologia "ulertuko" duten konputagailu-sistemen garapena erraztea baitu. Denera ia 200 baliabide barnebiltzen ditu UMLSko metatesauroak[5].

UMLSn integratuta dagoen terminologiaz gain, badaude beste hainbat baliabide, testuen prozesamenduan erabiliak direnak. *Disease Ontology* datu-basea da horietako bat, eta 8.043 gaixotasunen informazio zabala jasotzen du (Schriml *et al.*, 2012; Osborne *et al.*, 2009). Datu-base honetako kontzeptu bakoitzak SNOMED CTrekin, MeSHekin eta GNSekin lotura espezifikoak jasota dauzka, besteak beste. *Medical Dictionary for Regulatory Activities* edo MedDRA aldiz, hamaika hizkuntzetan eskuragarri dagoen medikuntzako terminologia-hiztegi bat da, autoritate arauemaileek erabiltzen dutena farmaziako arautze-prozesuetan, besteak beste (Brown *et al.*, 1999).

[3]http://www.snomed.org/members/ (2017ko maiatzaren 9an atzitua).

[4]https://www.snomedinaction.org/ (2017ko maiatzaren 9an atzitua).

[5]https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html (2017ko maiatzaren 9an atzitua).

Beste ezagutza-base bat *Online Mendelian Inheritance in Man* (OMIM) dugu, zeinak gizakien geneen eta gaixotasun genetikoen ezagutza jasotzen duen (Amberger *et al.*, 2011). MEDIC ere gaixotasunen hiztegi bat da, MeSHen gaixotasunen adar baten moldaketak eta OMIM datu-basearen gaixotasun genetikoak elkartzen dituena (Davis *et al.*, 2012). UMLSren eta SNOMED CTren lizentziak oztopo zirela, *Comparative Toxicogenomics Database* (CTD) proiektuaren baitan, MEDIC sortu zuten eta 2017ko urtarrileko bertsioan 11.000 gaixotasunetik gora ditu.

Ikusi dugun bezala, SNOMED CT erro sendoak dituen biomedikuntzako ontologia da, eta zabalduta dauden beste sailkapen zein ontologiekin lotura estuak ditu. Gainera, azpimarratu dugun moduan, SNOMED CTren estaldura oso zabala da, espezialitate askotako terminologia barnebiltzen duelarik. Gainera, terminologia jasotzeaz gain, kontzeptuen harteko informazioa ere eskaintzen du, ontologia bat baita. Ikusi ditugun baliabide terminologikoen artean SNOMED CT aukeratu dugu oinarri gisa. Horrenbestez, SNOMED CT euskaraz izateak abantaila ugari ekarriko dizkigu, ez bakarrik geroz eta gehiago egonkortzen ari den estandar bat euskaraz ere izango dugulako, baizik eta SNOMED CTrekin lotura duten baliabideak euskaraz erabiltzeko aukera ere izango dugulako.

2.4 SNOMED CTren itzulpenak

Gaur egunean, SNOMED CT AEBetako ingelesez, Erresuma Batuko ingelesez, espainieraz, danieraz eta suedieraz dago eskuragarri, eta frantsesezko, lituanierazko eta beste hainbat hizkuntzetako itzulpenak garatzen ari dira[6]. SNOMED International-en web-orrian zehazten den moduan, SNOMED CTren itzulpen batzuk teknika ezberdinak erabiliz gauzatu dira dagoeneko, eskuzko itzulpenetatik, automatikoetara.

Frantsesaren kasuan, adibidez, frantseserako garatutako baliabide terminologikoak baliatu dute frantsesezko SNOMED CT sortzen laguntzeko (Abdoune *et al.*, 2011). Horretarako, UMLSren metatesauroan integratutako frantsesez dauden lau baliabide terminologiko erabili dituzte (Joubert *et al.*, 2009): SNOMEDen nazioarteko banaketak (SNOMED International), GNS bere 10. bertsioan, MedDRA eta MeSH. Baliabide terminologiko horietatik, termino hobetsiak baino ez dituzte erabili: SNOMED Internationalek 107.900

[6]http://www.snomed.org/snomed-ct/snomed-ct-worldwide/translations-of-snomed-ct (2017ko maiatzaren 9an atzitua).

termino hobetsi ditu frantsesez, GNS10ek 9.306, MesHek 25.186 eta MedDRAk 18.209. SNOMED CTren CORE (*Clinical Observations Recordings and Encoding*) azpimultzoa frantsesera itzultzea zuten helburu. UMLSren CORE azpimultzoak, *National Library of Medicine*-k (NLM) aztertu dituen erakundeen datu-baseetan gehien erabiltzen diren terminoak (14.000) barnebiltzen ditu[7]. Termino horiek UMLS kontzeptuekin parekatu dira (6.800 kontzeptu), horietatik 5.000 baino gehiago SNOMED CTrenak direlarik. Horretaz gain, Merabti *et al.* (2013) lanean, bi estrategia konbinatu dituzte SNOMED CTren terminologia itzultzeko: kontzeptuetan oinarritutakoa eta lexikoan oinarritutakoa. Kontzeptuetan oinarritutakoan, terminoak kontzeptu mailan parekatzen dituzte, hau da, termino batek metatesauroan duen kokapena erabiltzen da, kokapen hori duen ordaina lortzeko. Lexikoan oinarritutakoak baliabide terminologiko elebidunak erabiltzen ditu parekatzea egiteko, aurreko lanetan egin bezala (Joubert *et al.*, 2009; Abdoune *et al.*, 2011).

SNOMED CTren txinerako bertsio bat sortu dute, laguntza automatikoa eta eskuzko lana konbinatuz egin da (Zhu *et al.*, 2012). Eskuzko lanean laguntzeko, Txinan erreferentziazkoa den MedDic deituriko datu-basea erabili dute. Datu-base horrek 2 milioi hiztegi eta balibide terminologiko barnebiltzen ditu, tartean ingelera-txinera medikuntzako hainbat baliabide terminologiko elebidun daudelarik, hala nola GNS9 eta MeSH. Terminoen sorkuntza-prozesuan, jatorrizko terminoen hitzez hitzeko baliokidetza automatikoa lortu dute MedDic erabilita. Hau da, terminoa osatzen duen hitz bakoitzaren ordaina jaso dute MedDic datu-basetik eta hori izan da itzultzaileei emandako proposamen automatikoa. Proposamen hori oinarri hartuta, itzultzaileek zuzenketa-prozesu bat eginda, terminoaren ordain-proposamen ofiziala egiten dute, eta medikuek berrikusi ostean proposatutako ordaina onartzen da.

SNOMED CTren aitzindaria den SNOMEDen espainierara itzultzeko prozedura txinerarako erabilitakoaren antzekoa izan da (Reynoso *et al.*, 2000). Bertan, ingelesezko terminoak aurreitzuliak izan dira hitzez hitzeko itzulpen automatikoa lortuaz, eta medikuek zein medikuntzako ikasleek berrikusi dituzte itzulpen horiek. Prozesua Newmark-en itzulpenaren lau mailak jarraituz egin dute: lehenengo bi mailak medikuek eta ikasleek egin dituzte (testuko eta erreferentziazko mailak), hirugarrena (kohesioa) itzultzaileek osatutako berrikuspenerako talde batek egin du, eta laugarren maila (natu-

[7]https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html (2017ko maiatzaren 9an atzitua).

raltasuna) medikuntzako terminologian aditua den mediku talde batek egin du. SNOMEDen itzulpenerako, ontologiaren kontrolerako eta trinkotasunerako bosgarren maila bat gehitu diote Newmark-en proposamenari Reynoso eta lagunek.

Suediarrek, txinerarako eta espainierarako jarraitutako estrategia bera hartu dute (Klein eta Chen, 2009). Dagoeneko eskura dauden hiztegiak erabiltzen dituzte itzulpenean laguntzeko. Jatorrizko terminoa hiztegietan aurkitzen dutenean, ordain hori erabiltzen dute lehen proposamen bezala (gerora berrikusia dena), eta partzialki aurkitzen dutenean, zati horien ordainak erabiltzen dituzte. Erabilitako hiztegien artean GNS10 emankorrena izan dela azpimarratu dute.

Danieraren kasuan, erabat eskuzko itzulpena egin dute eta kudeaketarako, software baten laguntzaz baliatu dira (Petersen, 2011; Andersen *et al.*, 2007). HealthTerm terminologia kudeatzeko sistemaren itzulpen-modulua erabili dute, 4-5 urrats erabiltzen dituen prozesu bat jarraituz[8]. Høy-k (2006)) itzulpenaren prozesu osoa azaltzen du eta kontuan hartzeko irizpideak ere zerrendatzen ditu. Lan horri esker, SNOMED CTren itzulpenetarako gidaliburua argitaratu zuen (Høy, 2010).

Berrikuspen honekin amaitzeko, SNOMED CTn, hizkuntza ezberdinetan egindako itzulpenen laburpena ikus dezakegu 2.1 taulan. Metodo automatikoak erabili dituztenen artean, parentesi artean adierazi ditugu hiztegien erabilera bakarrik automatizatu dituzten hizkuntzak.

Hizkuntza	**Eskuzkoa**	**Itzulpen Automatikoa**
Frantsesa	✓	✓
Txinera	✓	(✓)
Espainiera	✓	(✓)
Suediera	✓	(✓)
Daniera	✓	

2.1 taula – SNOMED CTren itzulpenen laburpena.

Schulz *et al.* (2013) lanean, SNOMED CTren 500 terminoren alemanierazko ordainak konparatu dituzte. 500 termino horien ordainak teknika ezberdinak erabiliz sortu dituzte: i) medikuntzaren domeinuko itzultzaile

[8]http://www.healthterm.com/translation-of-snomed-ct-into-danish/ (2017ko maiatzaren 9an atzitua).

profesionalek sortatuko ordainak, ii) Google Translate[9] tresnak sortutako ordainak, eta, azkenik, iii) medikuntzako ikasleek sortutako ordainak. Ordain horiek bi adituk (medikuak) ebaluatu zituzten, zuzentasun linguistikoari eta edukiaren zuzentasunari erreparatuz. Konparaketa horren emaitzetan ikus dezakegunez, eskuz sortutako ordainak itzultzaile automatikoak sortutakoak baino hobeak dira, aurreikusten zuten bezala, baina profesionalen eta ikasleen arteko emaitzak berdinak izan dira zuzentasun linguistikoari dagokionez, eta ikasleen ordainak hobeak edukiaren zuzentasunari dagokionez. Lan horretan ondorioztatu zuten moduan, ordainak sortzeko adituak baino alternatiba merkeagoak bideragarriak dira, ikasleak eta itzulpen automatikoko tresnak, hain zuzen ere.

2.5 Laburpena eta ondorioak

Kapitulu honetan tesi-lana kokatu nahi izan dugu, osasun-zientzietako terminologiari eta sorkuntza automatikoari egindako gainbegiratuaren bidez. Gainera, gaur egunean eskuragarri dauden osasun-zientzietako baliabide lexikalak ere aztertu ditugu, eta, bukatzeko, beste herrialde batzuetan SNOMED CT euren hizkuntzetara itzultzeko egindako lanak berrikusi ditugu.

Baliabide lexikalen artean, SNOMED CT ontologia euskaratzeko erabakia hartu dugu. Izan ere, osasun-zientzietako espezialitate askotako terminologia biltzen du, estaldura oso zabala izanik. Gainera, kontzeptuen hartean erlazioak definituta ditu: harreman hierarkikoak, alde batetik (IS-A erlazioak), eta informazio semantikoa gehitzen dutenak, bestetik (*causative-agent* edo *finding site* modukoak). Horrek testuak prozesatzeko garaian abantailak ematen dizkigu, testuan esplizitu agertzen ez den informazioa SNOMED CTtik bertatik jaso daitekeelako, besteak beste. SNOMED CT osasungintza klinikoko terminologia eleaniztun zabalduena dela esan daiteke, eta estandar terminologikotzat dute herrialde askotan. Gaur egunean, 50 herrialde baino gehiagotan erabiltzen da, eta behar ezberdinei erantzuten dioten inplementazioak ezagutzen zaizkio. Hori guztia gutxi balitz, beste sailkapen zein ontologiekin lotura estuak ditu, eta UMLSren metatesauroaren parte da.

Tesi-lan honetan, garatuko ditugun metodoetan, terminoentzako sortuko ditugu ordainak. Termino bat oinarri harturik, horren euskarazko ordain hautagaia proposatuko dugu automatikoki, eta adituen balidazioa egitera-

[9] http://translate.google.com (2017ko maiatzaren 15ean atzitua).

koan, kontzeptu-ikuspegia bermatuko dugu.

Euskarazko osasun-zientzietako corpus paralelo nahikoa ez izatean, terminologia automatikoki sortzeko metodo sortzaileak proposatuko ditugu. Izan ere, euskara normalizazio-prozesuan dagoen hizkuntza izanik, terminologia--falta nabaria dago gaur egunean, eta osasun-zientzien alorra ez da salbuespen bat. Gainera, motibazioa SNOMED CTren terminologia euskaratzea denez, baliteke testuetan zein hiztegietan oraindik jaso gabe dauden terminoak sortu behar izatea. Metodo ez-generazionalek ez dute sistemak ezagutzen ez duen terminoaren ordainik sortzea ahalbidetzen, eta, hortaz, erregeletan oinarritzen diren metodo sortzaileak izango dira garatuko ditugunak.

3

SNOMED CTren analisia

Kapitulu honetan tesi-lan hau horrenbeste baldintzatu duen SNOMED CTren sakoneko analisia aurkeztuko dugu. Aurrena, SNOMED CT kokatuko dugu 3.1 atalean. Bigarrenik, 3.2 atalean SNOMED CTren inguruko hainbat lan azalduko ditugu, eta 3.3 atalean SNOMED CTren bertsio zehatz baten analisia egingo dugu, 2015eko uztaileko ingelesezko bertsioarena hain zuzen ere, hau baita lan honen abiapuntua. Amaitzeko, tesi lan honen abiapuntu izan zen "SNOMED CT sare semantikoa euskaratzeko aplikazioa" Master Tesian (Perez-de-Viñaspre, 2013) egindako espainierazko eta ingelesezko SNOMED CTren bertsioen arteko konparaketaren laburpen bat aurkeztuko dugu 3.4 atalean. Izan ere, azterketa horretatik ateratako ondorioei esker aukeratu dugu ingelesezko SNOMED CT bertsioa abiapuntu euskarazko bertsioa sortzeko.

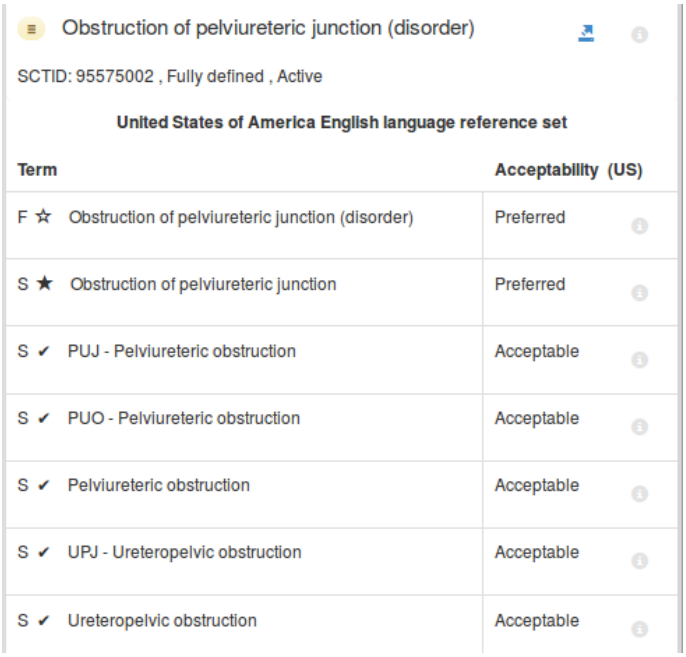
3.1 Sarrera

***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms* edo SNOMED CT (IHTSDO, 2014), osasungintza klinikoko terminologia eleaniztun zabalduena dela esan daiteke. Terminologia klinikoko estandarrak erabiltzeak osasun-arretaren kalitatea hobetu dezake, osasun-txosten elektronikoetako edukia modu trinkoan adierazteko aukera ematen duelako[1]. Nolabait esateko, hizkuntza eta sistema desberdinen arteko txosten klinikoen adierazpen eta in-

[1]http://www.snomed.org/snomed-ct/why-should-i-get-snomed-ct (2017ko maiatzaren 9an atzitua).

terpretazio automatikoa eta anbiguotasunik gabea ahalbidetuko duen hiztegi normalizatua da, hiztegi-sarreren arteko harremanak zehaztuta daudelarik.

SNOMED CTk osasun-txosten elektronikoen muineko terminologia eskaintzen du, eta 296.000 kontzeptu aktibo baino gehiago ditu modu hierarkikoan antolatuta. Humphreys *et al.* (1997) lanean erakusten duten moduan, SNOMED CTk estaldura onargarria eskaintzen du pazienteen baldintzak jasotzeko. Kontzeptuak logika deskribatzailearen axioma bidez deskribatzen dira, eta esanahi berdineko terminoak multzokatzeko erabiltzen dira. Terminoez ari garela, *terminologia* horrela definitzen du Chute-k (2000): kontzeptu bati loturiko hizkuntza-etiketak. Aurretik esan dugun moduan, SNOMED CTren kontzeptu bakoitzak etiketa edo *deskribapen* bat edo gehiago ditu, orokorrean *termino*tzat hartzen direnak.



3.1 irudia – SNOMED CTko *95575002 - Obstruction of pelviureteric junction (disorder)* kontzeptuaren deskribapen motak. Irudia SNOMED CTren nabigatzailetik atera dugu (http://browser.ihtsdotools.org (2017ko maiatzaren 9an atzitua)).

Bi deskribapen mota bereizten dira SNOMED CTn: *Fully Specified Name* (FSN) izendaturikoa eta sinonimo deiturikoa (*Synonym*). FSNak kontzep-

tuak identifikatzeko erabiltzen diren deskribapenak dira eta kontzeptuaren kategoria semantikoa adierazten duen etiketa (*semantic tag*) izaten dute parentesi artean deskribapenaren bukaeran. Kategoria semantiko horri esker, kontzeptuaren hierarkia ere ezagut dezakegu. Sinonimoak berriz, kontzeptua hizkuntza edo dialekto zehatz batean adierazteko erabiltzen dira. Sinonimoen artean ere bi mota bereizten dira: termino hobetsiak eta sinonimo onargarriak. Termino hobetsiak (*Preferred Term* edo PT), hizkuntza edo dialekto zehatz batean sinonimoen artean hobetsiak izateko aukeratuak dira, eta kontzeptu bakoitzeko eta hizkuntza edo dialekto bakoitzeko bakarra egon daiteke.

SNOMED CTren kontzeptu baten eta honen deskribapenen adibide bat ikus dezakegu 3.1 irudian. Bertan, *"95575002 - Obstruction of pelviureteric junction (disorder)"* kontzeptuaren deskribapenak ikus ditzakegu. FSNak lehenengo zutabean "F" etiketa dauka, termino hobetsiak "S★" etiketa lehenengo zutabean, eta *"preferred"* hitza bigarrenean, eta gainerako deskribapenek (sinonimo onargarriek), "S✓" etiketa daukate lehenengo zutabean eta *"acceptable"* hitza bigarrenean.

Aipatutako Chute-ren (2000) lanean, *terminologia kliniko*aren honako definizioa ematen da:

> *"Standardized terms and their synonyms which record patient findings, circumstances, events, and interventions with sufficient detail to support clinical care, decision support, outcomes research, and quality improvement; and can be efficiently mapped to broader classifications for administrative, regulatory, oversight, and fiscal requirements."*

SNOMED CTren lehenengo bertsioa 2002ko urtarrilean kaleratu zenetik[2], 19 herrialdek euren osasun-txosten elektronikoetan erabiltzeko erreferentziazko terminologia izateko aukeratu dute.

[2]http://www.snomed.org/snomed-ct/what-is-snomed-ct/history-of-snomed-ct (2017ko maiatzaren 9an atzitua).

3.2 SNOMED CTren inguruko azterketa bibliografikoa

"Data Analytics with SNOMED CT - Case Studies" txostenean (IHTDSO SNOMED CT, 2015) jasota datorren moduan, SNOMED CT erabilgarria da osasun-txosten elektronikoetatik informazioa jasotzeko eta ondorioz, datuen analisia egin ahal izateko, hala nola ikerketa klinikorako, txosten publikoak garatzeko edo medikuntza aurresalean (*predictive medicine*) erabiltzeko.

Hainbat metodo erabili izan dira medikuntzaren domeinuko ontologiak garatzeko, mantentzeko, parekatzeko eta ebaluatzeko. Yu-k (2006) medikuntzako ontologiak honakoetan erabilgarriak direla azaltzen du: i) terminologiaren kudeaketarako, ii) informazioaren integraziorako, elkarrekintzarako eta partekatzeko, eta iii) ezagutzaren berrerabilpenerako eta erabakiak hartzeko. Gure kasuan, SNOMED CTren euskarazko bertsioa osasun-zientzietako terminologia adosteko erabili nahiko genuke, gure iritziz hori baita lehenengo urratsa datuen analisirako eta erabakiak hartzeko, erremintak garatu aurretik.

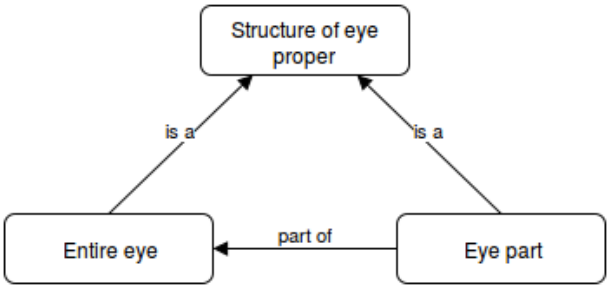
Lan asko argitaratu dira terminologia kudeatzeko sistemak ikuspegi teorikotik aztertu dituztenak. Adibidez, Bakhshi-Raiez *et al.* (2008) lanean, medikuntzako terminologia mantentzeko estandarizazio-egitura bat definitu dute, medikuntzako baliabide terminologiko guztiei aplikagarria dena, baita SNOMED CTri berari ere. Campbell *et al.* (2014) lanean, aldiz, SNOMED CTren terminoen egokitasuna aztertzen dute, patologisten azterketetan ohikoak diren ehunen morfologiak eta ehunen arkitektura nabarienak adierazteko garaian. Horrela, SNOMED CTren adierazgarritasunean hutsuneak identifikatu dituzte.

SNOMED CTren alde teknikoari zein ez-teknikoari (abantailak eta erronkak, adibidez) buruzko deskribapenak ere aurkeztu izan dira. Alde teknikoari dagokienez, Lee *et al.*-ek (2011), SNOMED CTren bertsioak aztertu dituzte. Horretarako 5.182 kontzeptu aztertu dituzte, eta % 41,2an (2.135 kontzeptutan) lau motako aldaketak identifikatu dituzte: i) FSNaren edo termino hobetsiaren forma-aldaketa, ii) kontzeptuaren egoera-aldaketa (aktibo/ez-aktibo), iii) deskribapen-logikaren aldaketa eta iv) hierarkian duen posizioaren aldaketa. SNOMED CTren erabilerari dagokionez, Silva *et al.* (2011) lanean ondorioztatu dute ordenagailu bidezko tomografiak deskribatzeko SNOMED CTren errepresentazio maila egokia dela, eta Maheronnaghsh *et al.* (2011) lanean, lunbagoaren inguruko erabakiak hartzeko laguntza-sistema batean

egindako azterketan, ondorio berera iritsi dira. Elhanan *et al.* (2011) lanean, aldiz, SNOMED CTren alde ez-teknikoari heltzen diote. Horretarako, SNOMED CTren erabiltzaile zuzenei galdeketa bat egin diete, honen estaldurari, kontzeptuen zehaztasunari eta kalitateari buruz. Erabiltzaileen % 42k SNOMED CTren estaldura % 85 baino altuagoa dela deritzo, eta % 60 honen kalitatearekin pozik agertu da. Horretaz gain, erabiltzaileek SNOMED CTren sendotasuna hobetzeko nahia erakutsi dute, baita estalduraren hedapena egiteko nahia ere.

Terminologia-sistemen analisiari esker, SNOMED CTn inkonsistentziak eta akatsak aurkitu izan dituzte hainbat lanetan. Adibidez, Jiang eta Chute (2009) lanean, SNOMED CTren osotasun semantikoa aztertzeko, kontzeptuen analisi formalean oinarritzen den eredu bat erabili dute. Mikroyannidi *et al.* (2012) lanean, aldiz, erregulartasun eta irregulartasun sintaktikoak aztertu dituzte, eta "diseinu-akats" batzuk eta deskribapen osatugabeak daudela ondorioztatu dute.

SNOMED CT sortzeko osasungintzako erreferentziazko bi terminologia-sistema elkartu zituzten: SNOMED RT eta *Clinical Terms*-en 3. bertsioa. Horretarako, 30 editore kliniko elkartu ziren eta, bi terminologia-sistemak aztertuta, adostasuna lortu zuten anatomia-ereduei, sinonimoen erabilerari eta termino hobetsiei dagokienez, besteak beste (Stearns *et al.*, 2001). Adibidez, SNOMED CTk entitate anatomikoak adierazteko "Egitura-Osoa-Zatia" hirukotea deritzon eredu-hurbilpena erabiltzen du, ingelesez *"Structure-Entire-Part"* edo SEP deritzona. Erabil dezagun *eye* (begia) entitatearen adibidea eredu-hurbilpen hori ulertzeko 3.2 irudian. Ikus dezakegunez, *structure of eye proper*, begia adierazteko erabiltzen den kontzeptu orokorrena da, *entire eye*k begi osoari egiten dio erreferentzia eta *eye part*ek, begiaren zati bati.



3.2 irudia – Begiaren "Egitura-Osoa-Zatia" hirukotearen adibidea.

3.3 Analisia

Atal honetan, SNOMED CTren 2015eko uztaileko nazioarteko argitalpenaren analisi kuantitatiboa aurkeztuko dugu. Bertsio hori da 5. eta 6. kapituluetan euskaratzeko abiapuntu gisa erabili duguna.

Gaur egun, SNOMED CT bi formatutan argitaratzen da: *Release Format 1* (RF1) eta *Release Format 2* (RF2). RF1 formatua, SNOMED CTren lehenengo argitalpenetik erabilitako formatua da (2002 urtetik aurrera). RF2 formatua, aldiz, 2012 urtetik aurrera erabiltzen den formatua da, eta SNOMED International-ek esaten duenaren arabera, RF1 erabat ordezkatzera helduko da. RF2 SNOMED CTren erabiltzaileen eskaerengatik sortu da. Izan ere, RF1 formatuan hainbat ahultasun aurkitu izan dira egituraketari dagokionez, eta honi erantzuteko, RF2k sendotasuna eta trinkotasuna eman dio. Adierazpide logikoan urratsak emateko ere erabili da formatu-aldaketa, ontologiekin lan egiteko aukera zabala emanaz.

"SNOMED CT The Release Format 2 Value Proposition" (IHTDSO SNOMED CT, 2013) lanean azaltzen duten moduan, gaur egun RF1 fitxategiak RF2ko datuetatik erauziak dira eraldaketarako baliabideak erabiliz. Zoritxarrez, eraldaketaren ondorioz, RF2 formatuan jasotzen den hainbat informazio galtzea ekidinezina da. Edonola ere, aurretik aipatu moduan, RF1 formatua desagertze-bidean dagoen formatua da.

Hurrengo hiru azpiataletan, SNOMED CTren analisi kuantitatiboan lortutako emaitzak aurkeztuko ditugu, eta honako datuei erreparatuko diegu:

1. *Egitura hierarkikoa*: hierarkia bakoitzean (eta kategoria semantiko bakoitzean) dagoen kontzeptuen populazioa neurtu dugu.

2. *Aberastasun terminologikoa*: atal honetan kontzeptu bakoitzari lotutako sinonimo kopurua neurtu dugu.

3. *Terminoen deskribagarritasuna*: termino edo deskribapen bakoitzaren token kopurua aztertu dugu, horien deskribagarritasunaren azaleko neurri gisa.

3.3.1 Egitura hierarkikoa

SNOMED CT kontzeptu-sistema generiko bat erabiliz antolatuta dago. Hau da, kontzeptuak orokortze-erlazioen bidez lotuta daude (IS-A erlazioak), eta

horrela menpeko kontzeptua gaineko kontzeptutik ezberdintzen duena ezaugarri bereziren bat da. Modu horretan, kontzeptu orokorrenak goiko mailetan kokatzen dira, eta kontzeptu espezifikoenak beheko mailetan. Erlazio horiei esker, SNOMED CTri egitura hierarkikoa ematen zaio. Adibidez, *myocardial infarction* (miokardio-infartu), *myocardial disease* (miokardio-gaixotasun) bat da (is-a erlazioa), *Clinical finding* (aurkikuntza kliniko) hierarkiaren baitan.

Hemeretzi goi-mailako hierarkia daude SNOMED CTn, eta horietan kontzeptu guztiak kokatuta daude (3.1 taula). Hierarkia horiek osasun-zientzien domeinuko hainbat arlo jasotzen dituzte, esaterako, aurkikuntza klinikoak, prozedurak (*Procedure*) eta gorputz-egiturak (*Body structure*). Hala ere, ez dira bakarrik osasun-zientzietako oso espezifikoak diren kontzeptuak bakarrik agertzen. SNOMED CTk gertaerak (*Event*), testuinguru soziala (*Social context*) edo ingurumena edo kokaleku geografikoa (*Environment or geographical location*) moduko hierarkiak ere jasotzen ditu, horietan barnebiltzen diren kontzeptuak osasun-txostenetan esanguratsuak direlako.

Aurretik azaldu dugun moduan, SNOMED CTren kontzeptu oro FSN (*Fully Specified Name*) bakar batekin deskribatuta dago. Deskribapen berezi horiek kontzeptu horren kategoria semantikoa adierazten duen etiketa semantiko (*semantic tag*) batez bukatzen dira (*myocaardial infarction (disorder)*). Etiketa semantiko hau erabiltzen da deskribapen berdina duten baina hierarkia ezberdina duten kontzeptuen artean anbiguotasuna ebazteko. Adibidez, *cyst (disorder)* FSNak pertsona batek kiste bat duenean pairatzen duen diagnostiko klinikoa adierazten du, *cyst (mophologic abnormality)* FSNak, berriz, kistea bera adierazten du.

Hierarkia bakoitzaren eta horien etiketa semantiko bakoitzaren populazioa ere erakusten dugu 3.1 taulan. Erroa den kontzeptua ere taulan sartu dugu *SNOMED CT Concept* hierarkiaren barruan, baina ez dugu analisirako kontuan hartu, kontzeptu bakarra baitago. Hirugarren zutabean ("Kontz." izendatutakoan), etiketa semantiko bakoitzean dagoen kontzeptu kopurua ematen dugu, eta horretarako kontzeptuen FSNak aztertu ditugu. Azkeneko zutabean ("Denera" izenekoan), hierarkia bakoitzak denera duen kontzeptu kopurua jaso dugu, kontzeptuak erlazionatzen dituen "is-a" erlazio motari jarraituz.

Hierarkia guztietan populatuena nahasmendu klinikoena da (*Clinical finding*), kontzeptu guztien ia herena hierarkia horretakoak baitira. Hierarkia horretan, behaketa edo ebaluazio klinikoetako emaitzak barnebiltzen dira eta bi multzotan bereizten dira: nahasmenduak (*disorders*) eta aurkikun-

Hierarkia	**Etiketa semantikoa**	**Kontz.**	**Denera**
Clinical finding	*disorder*	68.839	103.227
	finding	34.388	
Procedure	*procedure*	52.545	55.128
	regime/therapy	2.583	
Organism	*organism*	33.036	33.036
Body structure	*body structure*	25.045	30.871
	morphologic abnormality	4.695	
	cell	627	
	cell structure	504	
Substance	*substance*	25.828	25.828
Pharmaceutical / biologic product	*product*	16.931	16.930
Physical object	*physical object*	14.392	14.393
Qualifier value	*qualifier value*	9.388	9.388
Observable entity	*observable entity*	8.410	8.410
Social context	*occupation*	3.751	4.712
	person	425	
	ethnic group	270	
	religion/philosophy	203	
	social concept	23	
	life style	21	
	racial group	19	
Situation with explicit context	*situation*	4.159	4.159
Event	*event*	3.606	3.606
Environment or geographical location	*environment*	1.197	1.815
	geographic location	617	
	environment / location	1	
Specimen	*specimen*	1.620	1.620
SNOMED CT Model Component	*attribute*	1.136	1.547
	namespace concept	185	
	foundation metadata concept	183	
	core metadata concept	33	
	link assertion	8	
	metadata	1	
	linkage concept	1	
Staging and scales	*assessment scale*	1.110	1.341
	tumor staging	215	
	staging scale	16	
Special concept	*navigational concept*	640	649
	inactive concept	8	
	special concept	1	
Record artifact	*record artifact*	225	225
Physical force	*physical force*	171	171
SNOMED CT Concept	*SNOMED RT+CTV3*	1	1
Denera	-	317.057	317.057

3.1 taula – SNOMED CTren egitura hierarkikoa.

tzak (*findings*). Bi azpihierarkia horiek horren populatuak izanik, hemendik aurrera bi hierarkia berezitu gisa egingo diegu erreferentzia.

Gaur egunean, 19 hierarkietarako 42 etiketa semantiko daude. Etiketa semantiko horiek, batzuetan, hierarkiak baino zehaztasun handiagoa ematen dute. Adibidez, gorputz-egituren hierarkiak lau kategoria semantiko ditu: gorputz-egitura (*body structure*), anormaltasun morfologikoa (*morphologic abnormality*), zelula (*cell*) eta zelula-egitura (*cell structure*). Edonola ere, 11 hierarkietan etiketa semantikoak ez du inolako informazio gehigarririk ematen (bakarra dagoelako).

Taularen azterketari jarraiki, lau kasutan etiketa semantikoak kontzeptu bakar bat duela ikus dezakegu. Kontzeptu horiek, etiketa semantiko ezberdina duten hierarkiaren bi adar elkartzeko erabiltzen dira. Adibidez, "ingurunea / kokalekua" (*environment / location*) etiketa semantikoa, *Environment or geographical location (environment / location)* FSNa duen kontzeptuari dagokio, kasu honetan izen berdina duen hierarkiaren erro-kontzeptua dena. Kontzeptu horretatik, bi adar (ume) ateratzen dira: ingurunea (*environment*) eta kokapen geografikoa (*geographic location*).

Taula ondo aztertuta, inkoherentzia batzuk daudela ohartu gara. Produktu farmazeutiko / biologikoen (*Pharmaceutical / biologic product*) hierarkian, 16.931 kontzeptu daude bere etiketa semantiko bakarrean (*product*), eta 16.930 kontzeptu hierarkiaren populazioan. Kontrakoa gertatzen da objektu fisikoen (*Physical object*) hierarkian, non 14.392 kontzeptu dauden etiketa semantikoan eta 14.393 kontzeptu hierarkian. Salbuespen hori 440245005 identifikadorea duen kontzeptuari dagokio, *dressing medicated with leptospermum honey (product)* FSN duena. Ikusten dugunez, *product* etiketa semantikoa du bere FSNan, baina *"is-a"* erlazioei jarraituz, objektu fisikoen hierarkiakoa da[3].

3.3.2 Aberastasun terminologikoa

Aurretik aipatu bezala, SNOMED CTren helburuetako bat osasun-txostenetan erabiltzen diren kontzeptu kliniko ahalik eta gehien jasotzea da. *Starter Guide* dokumentuan (IHTSDO, 2014) esaten duten bezala, SNOMED CTk ulergarritasuna, erreproduzitzeko gaitasuna eta erabilgarritasuna ziurtatu behar ditu. Horrela, deskribapenek ulergarriak, onargarriak eta pro-

[3]http://browser.ihtsdotools.org/?perspective=full&conceptId1=440245005&edition=en-edition&release=v20170131 (2017ko maiatzaren 9an atzitua).

fesionalentzat esanguratsuak izango dute. Horregatik, SNOMED CTrentzat deskribapenen artean ahalik eta sorta handiena jasotzea garrantzitsua da, eta kontzeptu bat deskribatzeko erabili daitezkeen sinonimo guztiak jaso beharko lituzke.

SNOMED CTren sinonimoen aldakortasunaren ikuspegi orokorra erakusten dugu 3.2 taulan. Lehenengo zutabean, kontzeptuen kopuruak erakusten ditugu, bigarren zutabean, kontzeptu horiei lotutako sinonimo kopurua, hirugarren zutabean, kontzeptuko sinonimo kopuruaren batezbestekoa eta azkeneko zutabean horien mediana erakusten ditugu. Batezbestekoak ematen dizkigun datuak osatzeko ematen dugu mediana, kontzeptuen erdiak baino gehiagok sinonimo bakarra duen ala gehiago duten erakusten baitigu. Gogoan izan behar dugu, FSNak SNOMED CTren kontzeptuak identifikatzeko baino ez direla erabiltzen, eta ez direla testuetan agertzen; hortaz, taulako zenbakietatik at utzi ditugu.

Hierarkia	**Kontzep.**	**Sinonim.**	**Batez.**	**Mediana**
Clinical disorder	68.839	114.830	1,67	1
Clinical finding	34.388	52.857	1,54	1
Procedure	55.128	87.104	1,58	1
Organism	33.036	57.582	1,74	2
Body structure	30.871	59.384	1,92	2
Substance	25.828	43.356	1,68	1
Pharmaceutical / biologic product	16.930	25.179	1,49	1
Physical object	14.393	17.838	1,24	1
Qualifier value	9.388	14.440	1,54	1
Observable entity	8.410	13.253	1,58	1
Social context	4.712	5.893	1,25	1
Situation with explicit context	4.159	6.486	1,56	1
Event	3.606	4.404	1,22	1
Environment or geographical location	1.815	2.305	1,27	1
Specimen	1.620	1.982	1,22	1
SNOMED CT Model Component	1.547	1.956	1,26	1
Staging and scales	1.341	2.421	1,81	2
Special concept	649	894	1,38	1
Record artifact	225	284	1,26	1
Physical force	171	272	1,59	1
SNOMED CT Concept	1	4	4,00	4
Total	317.057	512.724	1,62	1

3.2 taula – Kontzeptuen eta sinonimoen kopuruak SNOMED CTren hierarkietan.

Harritu gaitu SNOMED CTren kontzeptu bakoitzak batezbestean sinonimo bat eta biren artean izateak. Are gehiago, medianaren balioak ikusita, hierarkia gehienetan kontzeptuen erdiak behintzat sinonimo bakarra dauka. Gorputz-egituren hierarkia da batezbesteko altuena duena, bitik oso gertu (1,92) eta organismoen eta, estadifikazioen eta eskalen (*Staging and scales*) hierarkiekin batera 2ko mediana dauka.

Gorputz-egituren hierarkiaren kasuan, SNOMED CTk "Egitura-Osoa-Zatia" hirukote eredu-hurbilpena erabiltzen duela gogora ekarri behar dugu. Alegia, termino hobetsiek hirukotearen egitura jarraitu behar dute, eta ondorioz sinonimo gehiago jasotzen dira. Adibidez, *nose* (sudur) zati anatomikoarentzat, "Egitura" kontzeptuaren termino hobetsia *nasal structure* da, eta sinonimo onargarria *nose*, 3.3 irudian ikus dezakegun bezala. Euskaratzen ari garen SNOMED CTren bertsioan, "Osoa" kontzeptuaren termino hobetsia *entire nose* da, eta sinonimo onargarria *nose* ("egitura" kontzeptuarren berdina) da. Ikusten dugunez, kontzeptu horiek guztiek gutxienez bi sinonimo dituzte.

Term	Acceptability (US)
F ☆ Nasal structure (body structure)	Preferred
S ★ Nasal structure	Preferred
S ✔ Nose	Acceptable

3.3 irudia – SNOMED CTren *45206002 - Nasal structure (body structure)* kontzeptuaren deskribapen guztiak. Irudia SNOMED CTren nabigatzailetik atera dugu (http://browser.ihtsdotools.org (2017ko maiatzaren 9an atzitua)).

Organismoen hierarkiari dagokionez, FSNetan ohikoa da hierarkiaren adierazlea, nazioarteko forma taxonomikoa eta kategoriaren adierazlea (*Genus, Family, Phylom...*) agertzea. Adibidez, *Genus Branchiomyces (organism)* FSNarentzat, *Branchiomyces* termino hobetsia da eta *Genus Branchiomyces* sinonimo onargarria. Horrela, *Linnaean* klase taxonomikoetan ofizialki onartuta dauden organismoek egitura hori jarraitzen dute izendapenei dagokionez, eta ondorioz gutxienez bi sinonimo izaten dituzte.

3.3.3 Terminoen deskribagarritasuna

Baliteke konplexutasuna ezagutzeko, termino bakoitzaren hitz (edo token[4]) kopurua zenbatzea neurri ona ez izatea, baina modu errazean eta sinplean terminoaren izaera erakusten digu. Hau da, terminoen token kopurua neurtuz, SNOMED CTren terminoen konplexutasunaren ideia orokor bat izan dezakegu. Adibidez, ez da berdina *lung cyst* moduko termino motz baten konplexutasuna, edo *ruptured emphysematous bleb of lung* moduko termino luze batena.

Terminoen morfologiaren ikuspegitik, Mayerthaler-ek (1981) azpimarratzen duen moduan, kontzeptu konplexuagoek izen luzeagoak izaten dituzte. Adibidez, *cuvette oximeter* terminoak *oximeter* baino espezifikoagoa den instrumentu bati egiten dio erreferentzia.

Are gehiago, kasu batzuetan terminoen deskribagarritasunarekin lotuta egon daiteke. Gehienetan, terminoa zenbat eta luzeagoa izan, orduan eta espezifikoagoa izango da. Ezaugarri hori SNOMED CTren kontzeptu-sistemaren azalpenean ere aipatzen da. Edonola ere, sinonimoen arteko ezberdintasunak aurkituko ditugu kontzeptu bat deskribatzeko garaian. Adibidez, *apoptosis (morphologic abnormality)* kontzeptuak bi sinonimo ditu: *apoptosis* (grezieratik, narriadura esan nahi du[5]) termino hobetsia eta *gene-directed cell death* sinonimo onargarria. Ikus dezakegun moduan, termino hobetsia zehatzagoa da baina, sinonimo onargarriak informazio lexikal gehiago ematen du.

Token kopuruaren araberako sailkapena egin dugu 3.3 eta 3.4 tauletan, hierarkien arabera. Lehenengo taulak (3.3), 1 eta 6 token arteko kopuruak erakusten ditu, eta 3.4 taulak gainerakoak. Bertan, token kopuru bakoitzeko zenbat termino dauden erakusten dugu alde batetik, eta bestetik, 3.4 taulako azken hiru zutabeetan hierarkia bakoitzean denera dauden sinonimo kopurua, terminoek batezbestean zenbat token dituzten eta horien mediana. Ikus dezakegunez, hierarkia guztiak kontuan izanik, batezbestekoa zein mediana 4 token inguruan daude. Hierarkia populatuenetatik, batezbestekotik ateratzen direnak azpimarratu behar ditugu: organismoak (*Organism*) eta substantziak (*Substance*). Bi hierarkia horien sinonimoek batezbestean 2 token dituzte, horien zehaztasunaren eta laburtasunaren erakusle. Orokorrean, termino gehienak 1 eta 8 token artekoak dira, tarte horretan ia % 95 jaso-

[4]Hitzak zenbatzerako garaian, hitzak eta puntuazio-markak bereizi ditugu, horretarako tokenizatzaile bat erabiliz.

[5]https://en.wiktionary.org/wiki/apoptosis (2017ko maiatzaren 9an atzitua).

Hierarkia	Token kopurua					
	1	**2**	**3**	**4**	**5**	**6**
Clinical disorder	3.863	21.009	25.028	20.732	16.252	10.346
Clinical finding	1.803	8.583	10.953	10.104	8.354	5.187
Procedure	1.996	9.893	15.401	17.049	14.553	10.177
Organism	9.091	32.392	6.346	3.582	1.672	1.453
Body structure	2.593	10.654	12.700	10.689	9.062	5.978
Substance	8.250	13.917	6.900	6.435	3.274	1.681
Pharmaceutical ...	2.616	2.363	4.987	4.537	3.235	2.379
Physical object	946	3.483	4.228	3.680	2.348	1.340
Qualifier value	4.536	4.555	2.760	1.159	717	344
Observable entity	459	2.406	3.394	2.739	1.871	1.059
Social context	904	2.051	1.179	725	466	256
Situation ...	11	401	1.243	1.709	1.272	851
Event	67	173	374	485	522	410
Environment ...	554	752	478	207	93	52
Specimen	9	250	572	339	163	167
... Model Component	240	522	687	156	252	31
Staging and scales	18	119	397	468	411	326
Special concept	19	131	220	140	72	122
Record artifact	2	64	53	34	26	8
Physical force	40	127	48	33	17	6
SNOMED CT Concept	0	0	1	0	0	0
Denera	38.017	113.845	97.949	85.002	64.632	42.173
Portzentajea	% 7,41	% 22,20	% 19,10	% 16,58	% 12,61	% 8,23

3.3 taula – Ingelesezko terminoen token kopuruaren araberako sailkapena (2 tauletan banatu dugu, hau da 1.a).

tzen direlarik. Datu bitxi gisa, bi deskribapen aurkitu ditugu aurkikuntzen hierarkian (*Clinical finding*) 52 token dituztenak.

Lehenengo taularekin jarraituz (3.3 taula), deskribapenen ia larudenak bi tokenez osatuta dagoela ikus dezakegu. Edonola ere, datuak aztertuz, patroi hori hierarkia guztiek jarraitzen ez dutela ikus dezakegu, nahasmenduak (*clinical disorder*), aurkikuntzak (*clinical finding*) eta gorputz-egiturak (*body structure*) kasu. Hierarkia horietan, hiru tokeneko sinonimoak dira gehienak, eta prozeduren hierarkian (*Procedure*) lau tokenetakoak. SNOMED CT osoan, bi tokeneko sinonimoak gehiengoa izatearen arrazoietako bat organismoen hierarkian aurkitzen dugu, hierarkia populatuenetako bat izanik, bertan desoreka oso handia baitago bi tokeneko eta gainontzeko token kopuruetako sinonimo kopuruen artean.

	Token kopurua					
Hierarkia	**7**	**8**	**9+**	**Denera**	**Batez.**	**Mediana**
Clinical disorder	6.748	4.148	6.704	114.830	4,37	4
Clinical finding	2.771	1.665	3.437	52.857	4,67	4
Procedure	6.776	4.198	7.061	87.104	4,90	4
Organism	651	668	1.727	57.582	2,66	2
Body structure	3.981	1.854	1.873	59.384	4,17	4
Substance	1.705	426	768	43.356	3,05	2
Pharmaceutical ...	1.764	1.083	2.215	25.179	4,58	4
Physical object	816	367	630	17.838	3,97	4
Qualifier value	167	71	131	14.440	2,44	2
Observable entity	586	348	391	13.253	4,01	4
Social context	151	68	93	5.893	3,03	2
Situation ...	453	270	276	6.486	4,79	4
Event	389	284	1.700	4.404	9,89	7
Environment ...	17	10	142	2.305	2,91	2
Specimen	170	98	214	1.982	4,81	4
... Model Component	21	13	34	1.956	3,08	3
Staging and scales	275	151	256	2.421	5,48	5
Special concept	41	50	99	894	4,83	4
Record artifact	6	3	88	284	6,67	4
Physical force	1	0	0	272	2,57	2
SNOMED CT Concept	0	0	3	4	16,50	14
Denera	27.489	15.775	27.842	512.724	4,13	4
Portzentajea	% 5,36	% 3,08	% 5,43	% 100		

3.4 taula – Ingelesezko terminoen token kopuruaren araberako sailkapena (2 tauletan banatu dugu, hau da 2.a).

3.4 Ingelesezko bertsioa aukeratzeko arrazoiak

Tesi-lan honekin hasi ginenean, orduko ingelesezko zein espainierazko SNOMED CTren bertsioak aztertu genituen, euskaratzeko abiapuntu onena aukeratzeko. Azterketa horretan, bi bertsio zehatzen konparaketa egin genuen, hierarkietako kontzeptu kopuruei dagokionez, termino hobetsien kopuruei dagokienez eta termino gabeko kontzeptuen kopuruari dagokionez (Perez-de-Viñaspre, 2013). Konparatutako bertsioak, honakoak dira: SNOMED CTren ingelesezko nazioarteko banaketa, 2012ko urtarrilaren 31koa; eta espainierazko banaketa, banaketa internazionalean oinarritua dena, 2012ko apirilaren 30ekoa.

Aurkeratutako espainierazko bertsioan kontzeptu asko aurkitu genituen inolako deskribapenik edo terminorik gabe, 3.5 taulan ikusten dugun mo-

etik. semantikoa	**Kop.**	**Portz.**	**etik. semantikoa**	**Kop.**	**Portz.**
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	etik. semantikorik gabe	1.013	-
			Denera	**40.864**	-

3.5 taula – Gaztelaniazko bertsioan galdutako kontzeptuak, etiketa semantikoen arabera sailkatuta.

duan. Taula horretan, hutsik dauden kontzeptuei dagozkien etiketa semantikoak (*semantic tag*) erakusten ditugu. Denera 34 etiketa ezberdinei dagozkien kontzeptuak dira galdu direnak. Kopuruaren arabera aztertuz gero, *procedure*, *disorder* eta *finding* etiketa semantikoak ditugu gabezia handienarekin (eta aldi berean populatuenak direnak). Ehunekoei begiratuz gero eta kopuru baxuak dituztenak baztertuz, kaltetuenak *situation*, *occupation* eta *regime/therapy* etiketa semantikoak direla esan dezakegu.

Gaineratu beharra dago, une horretan espainierazko bertsioa garapen fasean zegoela oraindik. Gaur egunean egoera asko aldatu bada ere, hartutako erabakia zuzena izan zelakoan gaude. Izan ere, espainierazko bertsioa ingelesezkoaren itzulpena da, eta itzulpen prozesuetan beti egoten dira galerak. Horregatik, ingelesezko bertsioa erreferentziazkoa izatea abiapuntu egokiena iruditzen zaigu, eta espainierazko bertsioa gure euskarazko bertsioa elikatzeko erabiliko dugu 5. kapituluan ikusiko dugun moduan.

3.5 Laburpena eta ondorioak

Kapitulu honetan SNOMED CTren ingelesezko bertsio zehatz baten analisi kuantitatiboa aurkeztu dugu. Analisi horren helburu nagusia SNOMED CTren ezaugarriak aztertzea izan da, honen euskaratzean lagungarria izango zaigun informazioa eskuratzeko.

Lehenik, SNOMED CTren egitura hierarkikoan sakondu dugu, definituta dauden 19 goi-mailako hierarkietako populazioen kopuruak eta horien baitan dauden etiketa semantiko bakoitzaren populazioaren kopuruak ezagutuz. Horrela, populatuenak aurkikuntza klinikoak (*clinical finding*), prozedurak (*procedure*), organismoak (*organism*) eta gorputz-egiturak (*body structure*) barnebiltzen dituzten hierarkiak direla ikusi dugu.

Bigarrenik, SNOMED CTren aberastasun terminologikoa aztertu dugu, kontzeptu bakoitzari lotuta dauden sinonimo kopuruak jasoaz. Zenbakiek erakutsi digutenaren arabera, ingelesezko SNOMED CTk aldakortasun gutxi du, hierarkia gehienetan sinonimo bakarreko kontzeptuak baitira nagusi (1,62 sinonimo batezbestean eta mediana 1).

Jarraian, terminoen deskribagarritasuna kuantifikatu dugu, horien token kopuruazenbatuz. Ikusi dugunaren arabera, bi tokeneko terminoak dira nagusi orokorrean, baina hiru eta lau tokeneko terminoak ere oso ugariak dira (hiruen artean ia % 58 osatzen dute), terminoak trinkoak direla erakutsiz.

Azkenik, tesi-lan honen hastapenetan egindako SNOMED CTren ingelesezko zein espainierazko bertsioen arteko konparaketako datu nabarmenenak ekarri ditugu gogora. Horrela, espainierazko bertsioak zituen gabeziak azaleratu ditugu, garai horretan oraindik garapenean zegoen bertsio bat izanik, hutsune nabariak baitzituen.

SNOMED CTren analisia eginda, SNOMED CT euskaratzeko bi erabaki nagusi hartu ditugu: jatorri-hizkuntza zein izango den eta zein hierarkiarekin hasiko dugun euskaratzea. Esan bezala, SNOMED CTren ingelesezko bertsioa hartuko dugu abiapuntutzat.

Hierarkiei dagokienez, hierarkia populatuenekin hastea erabaki dugu: aurkikuntza klinikoak (eta nahasmenduak, jakina), prozedurak eta gorputz-egiturak. Nahiz eta organismoen hierarkia gorputz-egiturena baino populatuagoa egon, hierarkia berezia da, eta ez dugu aukeratu. Aukeratutako hierarkia horiek erreferentziazko hierarkiak izango dira, eta egingo ditugun esperimentuak hierarkia horiekin ebaluatuko ditugu. Hala ere, garatutako sistemak hierarkia guztiak euskaratzeko erabiliko ditugu.

4

EuSnomeden diseinua

EuSnomed, SNOMED CT sare semantikoa euskaratzeko garatu dugun sistema da. Kapitulu honetan bere diseinuaren zertzelada batzuk aurkeztuko ditugu, sakontasunean sartu gabe. Sistema bera sakonki aurkeztu genuen *Hizkuntzaren Azterketa eta Prozesamendua* masterreko titulua lortzeko master bukaerako proiektuan (Perez-de-Viñaspre, 2013).

Kapitulu honek honako egitura jarraituko du: aurrenik, EuSnomed sistemaren deskribapen orokorra egingo dugu 4.1 atalean eta honen oinarrian dagoen algoritmoa aurkeztuko dugu 4.2 atalean. Jarraian, 4.3 atalean baliabideen biltegiratzea azalduko dugu eta amaitzeko, 4.4 atalean sistemaren klase-diagrama deskribatuko dugu.

4.1 Deskribapen orokorra

EuSnomed, SNOMED CTren eduki terminologikoa euskaratzeko sistema da (beste edozein hizkuntzatan lortzeko, modu errazean egin egokitu daiteke). Sistemaren oinarrian lau urratsetako algoritmo bat dago hurrengo atalean azalduko duguna. Algoritmo horrek jatorri-helburu hizkuntzetarako eskura dauden baliabide lexiko elebidunak berrerabiltzen ditu eta horien hutsuneak betetzeko baliabide berriak sortzen ditu.

Sistemak, exekuzioaren hasieran, alde batetik SNOMED CTren edukia jaso eta hierarkiaka banatu eta biltagiratzen du, eta bestetik, baliabide lexikal eleanitzak biltegiratzen ditu ItzulDB deituriko datu-basean. Biltegiratzearen ezaugarriak zein definitu dugun formatuaren deskribapena, 4.3 atalean

azalduko dugu.

Gogoan izan behar dugu SNOMED CT kontzeptuetan oinarritzen dela, eta guk lortu nahi ditugunak horiek errepresentatzen dituzten terminoen edota deskribapenen ordainak direla. EuSnomed sistemak, deskribapenak jaso eta ordainak lortuko ditu.

Ordainak modu inkrementalean lortuko ditugu, lortutako ordain berriak berrerabiliz. Horrela, aplikazioak deskribapenen token kopuruaren araberako ordena erabiliko du ordainak lortzeko, token bakarrekoekin hasi, eta termino luzeenekin bukatu arte. Token kopuru bakoitzarekin bukatzerakoan, baliabideak birkonpilatuko ditugu, sortutako ordain berriak hurrengo exekuzioetan berrerabil daitezen.

4.2 Algoritmoa

Atal honetan SNOMED CTren deskribapenetarako hizkuntza berri batean ordainak lortzeko helburuarekin definitu dugun algoritmoa deskribatzen dugu.

Algoritmoak kontzeptuen deskribapenak landuko ditu, ez kontzeptuak ezta horien arteko erlazioak ere. Esan bezala, modu inkrementalean egingo du lan, eta horrela, lortutako ordain berriak berrerabiliko dira.

Gure kasu zehatzean, espainierazko zein ingelesezko terminoak euskaratzeko erabiliko badugu ere, algoritmoak edozein hizkuntzatarako balio du, baliabide egokiak izanez gero.

Jatorri-termino bakoitzerako ordain bat edo hainbat sortzen ditu algoritmoak, erabilitako baliabideen arabera. Beti ere, hautagai kopuru minimoa lortzen ahalegintzen da.

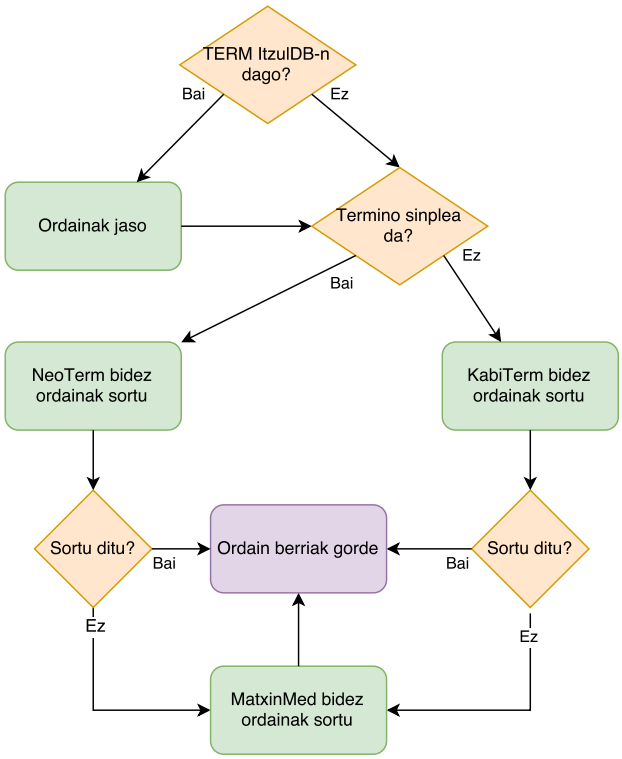
Algoritmoak (ikusi 4.1 algoritmoa) lau urrats nagusi ditu oinarrian (urrats hauek 2-5, 6-11, 13-16 eta 17-21 lerro multzoetan banatuta daude) eta algoritmoaren exekuzio bakoitzean sortutako baliabide berriak ItzulDB deituriko datu-basean gordetzen ditugu.

Jarraian, algoritmoaren sasi-kodea aurkezten dugu 4.1 algoritmoan, eta 4.1 irudian honen errepresentazio grafikoa.

4.1 algoritmoa Terminoetatik ordainak lortzeko algoritmoa.

Sarrera: Terminoa
$TERM \rightarrow$ Jatorri-terminoa (sarrera)
$ItzulDB \rightarrow$ Termino-ordain pareen DBa
$NeoTerm \rightarrow$ Termino neoklasikoen sortzailea
$KabiTerm \rightarrow$ Termino habiaratuen sortzailea
$MatxinMed \rightarrow$ Medikuntzara egokitutako Matxin itzultzaile automatikoa
Irteera: Terminoaren ordaina(k)
$ORDAINAK \rightarrow$ Lortutako ordainen zerrenda

hasiera
baldin $TERM \in ItzulDB$ **orduan**
  $ORDAINAK \leftarrow ItzulDBko\ ordaina(k)$
  itzuli $ORDAINAK$
amaiera baldin
baldin $TERM$ termino sinplea bada **orduan**
  $ORDAINAK \leftarrow NeoTerm(TERM)$
  baldin $ORDAINAK$ ez hutsa **orduan**
    $ORDAINAK\ ItzulDB$n gorde
    itzuli $ORDAINAK$
  amaiera baldin
bestela
  $ORDAINAK \leftarrow KabiTerm(TERM)$
  baldin $ORDAINAK$ ez hutsa **orduan**
    $ORDAINAK\ ItzulDB$n gorde
    itzuli $ORDAINAK$
  bestela
    $ORDAINAK \leftarrow$ MatxinMed($TERM$) ▷ Beti sortuko du ordaina◁
    $ORDAINAK\ ItzulDB$n gorde
    itzuli $ORDAINAK$
  amaiera baldin
amaiera baldin
amaiera



4.1 irudia – Algoritmoaren eskema.

Honakoak dira algoritmoaren lau urratsak:

1. **Baliabide lexikaletatik jasotako terminoak**: osasun-zientzien domeinuko hainbat hiztegi elebidunetatik zein eleanitzetatik erauzitako pareak erabiliko ditugu.

   Gure beharretarako, ingelesa-euskara zein espainiera-euskara hiztegi eleanitzak erabiliko ditugu, hala nola, Euskalterm, ZT Hiztegia, eta abar. Termino-ordain pareak ItzulDB datu-basetik jasoko ditugu. Algoritmoa lehen aldiz exekutatu aurretik, baliabide lexikaletako informazioa ItzulDB datu-basean gordeko dugu.

   Adibidez, 4.2 irudian urrats honen adibide bat ikus dezakegu. *Deoxyribonucleic acid* ingelesezko jatorri-terminoarentzat bi ordain jasotzen ditugu hiztegietatik, “azido desoxirribonukleiko” eta “DNA”. Ikus dezakegunez, ZT Hiztegian eta Euskaltermen agertzen dira bi ordainak. Jatorria jasotzeaz gain, datu gehiago sortu eta gordetzen ditugu pare-

katzean, konfiantza maila, kasua, eta abar. Hurrengo atalean aztertuko dugu informazio hori guztia (4.3 atala).

Jatorri-terminoa: *Deoxyribonucleic acid*
Algoritmoaren urratsak: 1-5 (4.1 algoritmoan)
Ordainak: azido desoxirribonukleiko: (ZT Hiztegia, Euskalterm)
DNA: (ZT Hiztegia, Euskalterm)

4.2 irudia – Baliabide lexikalek emandako ordainen adibidea.

2. **Termino neoklasikoak**: urrats honetan termino neoklasikoen afixu semantikoetaz baliatzen gara ordainak sortzeko. Osasun-zientzien alorreko aurrizki, atzizki zein erroen hiztegi elebidun bat baliatzen dugu sorkuntza egiteko. Afixuen hiztegiaz gain, erreferentzia-hizkuntzaren arau ortografiko zein morfofonologikoak deskribatzen dituzten sorkuntza-erregelak ere erabiltzen ditugu ordainak sortzeko, eta hiztegian afixua ez dagoen kasuetarako, transliterazio-erregelak aplikatzen ditugu.

   Urrats honetarako aparteko programa bat sortuko dugu, NeoTerm, zeinak Egoera Finituko Transduktoreak erabiltzen dituen termino neoklasikoen ordainak sortzeko.

   4.3 irudian urrats honen adibide bat erakusten dugu. *Photodermatitis* ingelesezko jatorri-terminoaren euskaratze-prozesua urratsez urrats erakusten dugu. Aurrena, afixu desberdinak identifikatuko ditugu eta beraien ordainak lortuko ditugu. Bukatzeko afixuak elkartuko ditugu erreferentzia-hizkuntzaren ortografia zein morfofonologia arauak betetzen dituztelarik.

Jatorri-terminoa: *Photodermatitis*
Algoritmoaren urratsak: 6-10 (4.1 algoritmoan)
Sorkuntza-prozesua:
Identifikatutako afixuak: photo+dermat+itis
Itzulitako afixuak: foto+dermat+itis
Sortutako ordaina: Fotodermatitis

4.3 irudia – Termino neoklasikoen sorkuntza-erregelekin lortutako ordainen adibidea.

3. **Termino habiaratuak**: termino konplexuetarako (hitz batez baino gehiagoz osatutako terminoak) termino habiaratuetan oinarritutako sorkuntza-erregelak erabiliko ditugu.

   Termino habiaratuak termino baten barruan dauden beste termino batzuk dira. Adibidez, *fracture of elbow* terminoak, bi termino habiaratu ditu: *fracture* eta *elbow*. Kasu honetan, *fracture* terminoa gaixotasun bat da (gaixotasunen hierarkian dago SNOMED CTn) eta *elbow* gorputz egitura bat da. Informazio horretaz baliatuko gara sorkuntza-erregelak idazteko.

   Aparteko programa bat sortuko dugu urrats honetarako ere, KabiTerm, Egoera Finituko Transduktoreak erabiltzen dituena, besteak beste.

   Aurreko adibidearekin jarraituz, 4.4 irudian termino horren euskaratzea erakusten dugu urratsez urrats.

Jatorri-terminoa: *Fracture of elbow*
Algoritmoaren urratsak: 13-16 (4.1 algoritmoan)
Sorkuntza-prozesua:
Identifikatuta egitura: GAIXOTASUN + of + GORPUTZ_EGITURA
Egitura baliokidea: GORPUTZ_EGITURA+ren + GAIXOTASUN
Sortutako ordaina: Ukondoaren haustura

4.4 irudia – Termino habiaratuen sorkuntza-erregelekin lortutako ordainen adibidea.

4. **Itzultzaile automatikoak sortutako termino konplexuak**: azkenekoz, aurreko urratsetan ordainak lortu ez ditugun terminoetarako itzultzaile automatiko bat egokituko dugu.

   Tesi honen garapenerako, Matxin erabilera orokorreko itzultzaile automatikoa erabiliko dugu. Medikuntzaren domeinura egokituko dugu eta MatxinMed deituko diogu egokitzapen honi. MatxinMed gainerako urratsekin sortuko ditugun termino-ordain pare berriekin elikatuko dugu, osasun-zientzien domeinura egokitzeko.

Horrenbestez, EuSnomed sistemaren oinarria den algoritmoa deskribatu dugu. Jarraian, lexikoaren biltegiratzeaz eta sistemaren klase-diagramaren inguruan arituko gara hurrengo atalean.

Jatorri-terminoa: *Lymphoma of lower esophagus*
Algoritmoaren urratsak: 18-20 (4.1 algoritmoan)
Sortutako ordainak: Beheko esofagoko linfoma

4.5 irudia – MatxinMeden bidez lortutako ordainen adibidea.

4.3 Biltegiratzea: TBX

XML (*e**X**tensible **M**arkup **L**anguage*) datuak modu egituratuan antolatzeko aukera ematen duen etiketa-lengoaia bat da. Datuak egituratzeaz gain, horiei esanahia ere emateko aukera eskaintzen du, elementu bakoitzaren etiketaren bidez. Horrela, elementu bakoitzaren egitura moldatu daiteke elementu horrekin erlazionatutako informaziora egokituz. Azken urteetan, estandar asko definitu dira terminologiaren errepresentaziorako, eta XML lengoaiak berebiziko garrantzia izan du horien artean, hala nola, ***X**ML representation of **L**exicons and **T**erminologies* (XLT) (SALT project, 2000), ***L**exical **M**arkup **F**ramework* (LMF) (ISO-24613:2008) (ISO, 2008), ***D**ictionary **M**arkup **L**anguage* (DML) (Mangeot, 2002) edo ***T**erm**B**ase e**X**change* (TBX) (ISO-30042: 2008) (LISA, 2008 eta Melby, 2012).

Gure lanean, TBX estandarrean oinarritu gara, terminologia kudeatzeko zein elkartrukerako abantailak eskaintzen baititu, eta jadanik beste hainbat baliabidetan integratuta baitago (ZT Hiztegia, adibidez).

TBX, XMLen oinarritutako estandar ireki bat da informazio terminologikoa egituratzeko eta kudeatzeko erabiltzen dena. Datu terminologikoak analizatzeko edo modu deskribatzailean adierazteko aukera ematen du, besteak beste, nahiz eta TBXren eginkizun nagusia datu terminologikoak elkartrukatzea izan. TBXk etiketatze-eredu unibertsala eskaintzen duenez, enpresa eta erakunde ezberdinek TBX euren barne datu terminologikoak kudeatzeko erabiltzeaz gain, elkarren arteko truke eta banaketarako ere erabiltzen dute.

TBXk datu terminologikoak datu-kategorien bitartez egituratzen ditu. 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 multzoa eskaintzen du TBX-k, 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 ***T**erminological **M**arkup **L**anguage*a (TML) definituz.

Gure datu terminologikoak egituratzeko Terminologia Zerbitzurako Online Sistemarako (TZOS) (Arregi *et al.*, 2010) egokitutako TMLaren interpretazio propioa egin dugu gure datu terminologikoak egituratzeko. TZOSek honakoak eskaintzen ditu: i) terminoak jasotzeko ingurunea, ii) jasotako informazioa gordetzeko eta prozesatzeko baliabideak eta iii) kontsultetarako eta elkarrekintzarako interfazea. Hau da, domeinu ezberdinetako terminoak lantzeko eta zabaltzeko zerbitzua eskaintzen du.

Hortaz, jatorrizko formatua ez da TBX hutsa izan, baizik eta TZOS sistemarako egokitutako formatua. Biltegiratu beharreko informazioa desberdina izanik, SNOMED CTren errepresentaziorako formatuaren egokitzapen bat egin dugu eta ItzulDBrako beste bat. Horrela, bi datu-base ditugu XMLz: SNOMED CTrena eta termino-ordain pareena (*ItzulDB*). Jarraian, bi datu-baseen formatuen egokitzapenak azalduko ditugu.

4.3.1 SNOMED CTrako TBX formatua

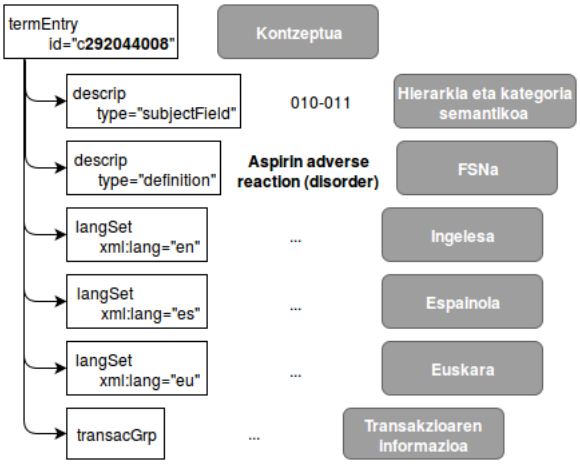
Atal honetan, TBX formatuan bertan adierazi eta gorde beharreko informazioa identifikatu eta kodetzeko hartutako irizpideen berri emango dugu.

Aurretik esan bezala, SNOMED CTren eduki terminologikoa gordetzeko formatua da hemen aurkezten duguna, eta ordain berriak lortu ahala, SNOMED CTren datu-base honetan gordeko ditugu ere. Kasu honetan, ingelesezko zein espainierazko SNOMED CTren bertsioetako eduki terminologikoa gorde dugu, baita lortzen ditugun euskarazko ordainak ere.

Hurrengo lerroetan XML dokumentuaren gorputzaren datu-kategoriak eta beharrezko informazioa kategoria horietan nola kodetzen dugun deskribatuko dugu.

Kontzeptu maila

Kontzeptu mailan, kontzeptuen informazioa eta dagozkien termino zein ordainak sailkatzen ditugu. Kontzeptua goiko elementua da, `termEntry` elementu moduan adierazten duguna eta kontzeptua identifikatzeko SNOMED CTren identifikadorea erabiltzen dugu. 4.6 irudian erakusten dugun adibidean, kontzeptuaren identifikadorea c292044008 da.



4.6 irudia – Kontzeptu baten zuhaitz-egitura.

Jarraian kontzeptu mailan erabiliko ditugun datu-kategoriak azalduko ditugu elementuen arabera sailkatuta. Azalpena ongi jarraitzeko, 4.6 irudiari begiratzea gomendatzen dugu. Datu-kategoriak XMLko elementuen `type` atributu gisa erabiltzen dira.

- `descrip` elementuaren datu-kategoriak `termEntry` barnean:
  - *subjectField*: SNOMED CTren kontzeptuaren hierarkia eta kategoria semantikoa (*semantic tag*) adierazteko erabiliko dugu. Hierarkia horiek adierazteko kode-baliokidetza batzuk definitu ditugu. 4.6 irudiko balioak (010 eta 011) aurkikuntzen eta gaixotasunen (*Clinical Finding/disorder*) hierarkiari eta gaixotasunen (*"disorder"*) kategoria semantikoari dagozkie.
  - *definition*: kontzeptua bera deskribatu eta ulergarri egiten duen adierazpena gordetzeko erabiliko dugu, antzeko kontzeptuetatik bereizteko balioko duena. Horretarako, SNOMED CTren *Fully Specified Name* (FSN) erabiliko dugu, gure kasuan ingeleseko bertsioarena, jatorrizko eta erreferentziazko hizkuntza hori baita. Adibidean, *Aspirin adverse reaction (disorder)*.

- `transacGrp` (transakzioaren informazioa) elementuaren azpi-elementua (hau ez da irudian agertzen):
  - `date`: kontzeptuan azken aldiz aldaketak egindako data gordeko du.

Azkenik, hizkuntza bakoitzaren termino desberdinak gordetzeko elementuak izango ditugu `langSet` elementu moduan definiturik. Elementu honen *lang* atributuaren bitartez terminoen hizkuntza definituko da: euskara, ingelesa edo espainiera. Elementu horien edukia hurrengo atalean aztertuko dugu (4.3.1 atala).

Termino maila: jatorri-terminoak

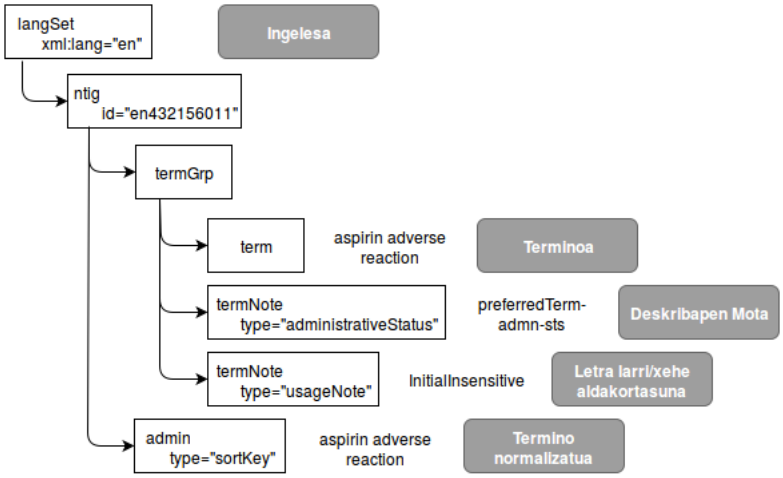
Aurretik azaldu bezala, terminoak hizkuntzaren arabera multzokatuko ditugu, `langSet` elementuaren bitartez. SNOMED CTren itzulpenerako bi motatako terminoak bereiziko ditugu: alde batetik jatorri-terminoak, gure kasuan SNOMED CTtik zuzenean ekarritako ingelesezko eta espainierazko terminoak; eta bestetik, euskarazko ordainak. Termino mota bakoitzerako informazio desberdina beharko dugunez, bereizirik azalduko ditugu hurrengo ataletan.

Esan bezala, jatorri-terminoak SNOMED CTtik zuzenean inportaturiko ingelesezko zein espainierazko terminoak dira. Termino horiek aldaketarik izango ez dutenez, gordeko dugun informazioa SNOMED CTrekin zuzenki lotuta dago.

Termino bakoitza `ntig` (*nesting term information group*) elementuaren bitartez gordeko da eta identifikadoretzat hizkuntzaren ISO 639-1 gakoa ("en" ingeleserako edo "es" espainierarako) eta SNOMED CT deskribapenaren identifikadorea erabiliko ditugu elkarrekin bilduta. 4.7 irudiko adibidean, terminoaren identifikadorea en432156011 da.

Termino bakoitzaren informazioa gordetzeko, hurrengo elementu eta datu-kategoriak erabiliko ditugu (jarraitu azalpena 4.7 irudiarekin):

- `termGrp` elementuan multzokatuta:
  - `term`: terminoa bera gordetzeko elementua izango da. 4.7 irudiko adibidearen kasuan *aspirin adverse reaction* da elementu honen edukia.



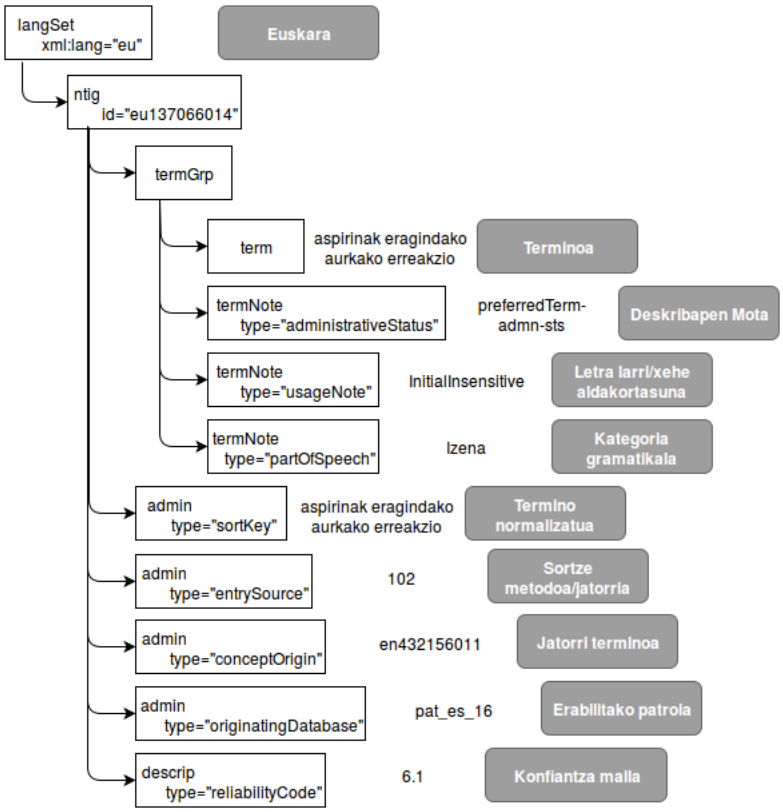
4.7 irudia – Jatorri-termino baten zuhaitz-egitura.

- `termNote` elementuaren datu-kategoria:
  - *administrativeStatus*: terminoak SNOMED CTn daukan onarpen-maila adieraziko da. Hau da, hobetsitako terminoa edo onartutakoa den, edo adituek baztertu duten adieraziko da. Adibidearen kasuan (4.7 irudia), hobetsitako terminoa denez (*Preferred Term*), "preferredTerm-admn-sts" izango da datu-kategoria honen balioa.
  - *usageNote*: terminoaren letra larri/xehe aldakortasuna adieraziko du. Hiru aukera posible jasotzen ditu: *InitialInsensitive* lehen letra aldagarria denean, *Insensitive* letra guztiak aldagarriak direnean eta *Sensitive* letra guztiak aldaezinak direnean.

- `admin` elementuaren datu-kategoriak:
  - *sortKey*: Terminoaren balio normalizatua agertuko da, hau da, terminoaren forma letra xehez eta alfanumerikoak ez diren karaktereak kenduta (hala nola gidoiak eta komak). Espainieraren kasuan diakritikoak ere kentzen zaizkio. 4.7 irudiko adibidean *aspirin adverse reaction* da elementu honen balioa.

Termino maila: ordainak

Atal honetan, SNOMED CTren euskarazko bertsioaren sortze-bidean lortuko ditugun euskal ordainak adierazteko erabiliko dugun informazioaren egituratzea azalduko dugu. Esan bezala, beste edozein hizkuntzetarako ere balio dezake egituratze honek.



4.8 irudia – Euskal ordain baten zuhaitz-egitura.

Jatorrizko terminoetan gordeko den informazioaz gain, beste hainbat datu ere beharko ditugu ordainetan. Dagoeneko azaldutako elementu eta datu-kategoriei honakoak gehitu dizkiegu (erabili bedi 4.8 irudiko egitura azalpenak hobeto ulertzeko):

- `termGrp` elementuan multzokatuta:
  - `termNote` elementuaren *partOfSpeech* datu-kategoria: ordainaren

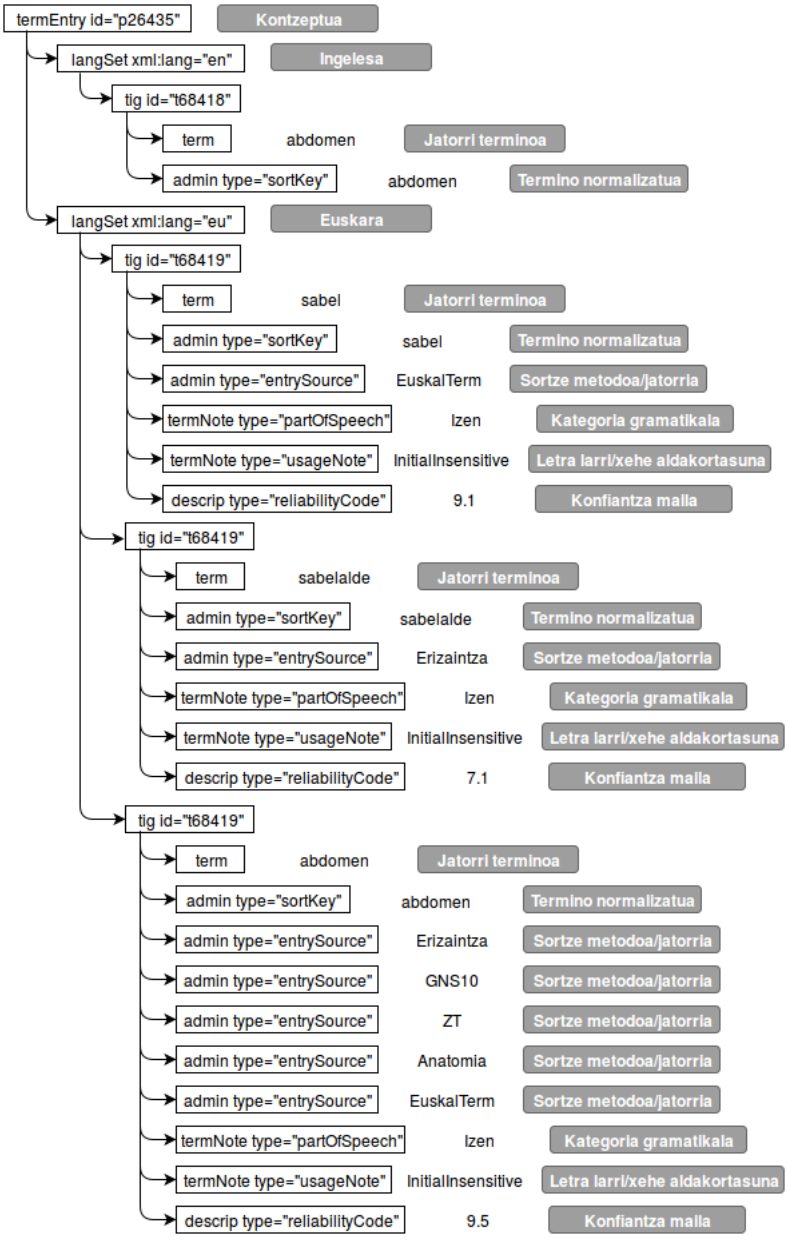
kategoria gramatikala gordeko da hemen, hala nola izena, adjektiboa edo aditza den.

- `admin` elementuaren datu-kategoria gehigarriak:
  - *entrySource*: ordaina lortzeko erabili den modua adieraziko duen kodea da. 4.8 irudian "termino habiaratuen sorkuntza-erregelei" egiten die erreferentzia (102 kodea).
  - *conceptOrigin*: ordaina sortzeko erabili den terminoaren identifikadorea gordeko da hemen. Identifikadore hori jatorri-terminoaren `ntig` elementuaren identifikadorea da. Adibidean, 4.7 irudiko ingelesezko jatorri-terminoari egiten dio erreferentzia (en432156011). Jatorri-termino bat baino gehiago egon den kasuetan, horiek guztiak jasotzen dira.
  - *originatingDatabase*: termino habiaratuen sorkuntza-erregelen bitartez sortzen denean terminoa, datu-kategoria honetan erabilitako erregelaren kodea gordeko da.
- `descrip` elementuaren datu-kategoria:
  - *reliabilityCode*: sortutako ordain bakoitzari konfiantza maila bat esleituko diogu. Zenbaki hori jatorriaren araberakoa izango da, eta zenbat eta baliabide gehiagotatik sortu, konfiantza maila orduan eta altuagoa izango da. Adibidez, Euskaltermi eta ZT Hiztegiari konfiantza maila altuena emango diegu, erreferentziazko baliabide lexikalak baitira horiek, eta automatikoki sortutako terminoei konfiantza maila baxuagoa.

A eranskinean, atal honetan irudien bidez erakutsitako adibideen XML kodea erakusten dugu.

4.3.2 Itzulpen-pareen datu-baserako TBX formatua

Termino-ordain pareen datu-basea (ItzulDB) egituratzeko, deskribatu berri dugun TBX formatuaren bertsio sinpleagoa erabiliko dugu. Parekatzea jatorri-hizkuntzaren termino bakoitzerako egingo dugu, adibidez, ingeleseko termino bat eta berari dagozkion ordainak biltegiratu ditugu ItzulDBn. Hala ere, ez ditugu jatorri-hizkuntzako sinonimoak multzokatu, bilaketek gako bakarra izateko, eta horrela, bilaketa azkarragoa eta sinpleagoa izan dadin.



4.9 irudia – Euskal ordain baten zuhaitz-egitura.

Parekatze bakoitza `termEntry` elementu baten barruan gordeko da. Identifikadorea *p* letraz hasiko da eta jarraian zenbaki bat izango du, parekatze bakoitzerako inkrementatuko dena (p26435, adibidez). 4.9 irudian ItzulDBko

parekatze baten adibidea ikus daiteke.

Terminoak adierazteko aldiz, `langSet` elementuaren barruko `tig` (*term information group*) elementuak erabiliko ditugu. `tig` elementuak `ntig` elementuen antzekoak izango dira, baina sinpleagoak. `tig` elementuak identifikatzeko parekatzeen estrategia berdina erabiliko dugu, baina t letra erabiliz hasieran (t13, adibidez).

Jatorri-terminoen kasuan, terminoa eta bere termino normalizatua gordeko ditugu, `term` elementua eta *sortKey* datu-kategoria erabiliaz. Ordainetan aldiz, informazio gehigarria gordeko dugu, hala nola, *partOfSpeech* datu-kategoriaren bitartez kategoria gramatikala, *entrySource*ren bitartez iturburua(k), *usageNote* erabiliz letra larri/xehe aldakortasuna eta *reliabilityCode*rekin ordain horren konfiantza maila.

4.4 Klase-diagrama

Atal honetan sistema garatzeko diseinatu dugun klase-diagrama azalduko dugu laburki.

EuSnomed sistemak aurreko ataletan azaldutako XML dokumentuetatik jasoko du informazioa eta horietan egiten ditu emaitzen idazketak. Hori horrela izanik, klase-diagrama TBX dokumentuen objektuetan oinarritu dugu. Objektuen bitartez XML dokumentu horietako informazioa jaso eta moldatuko dugu.

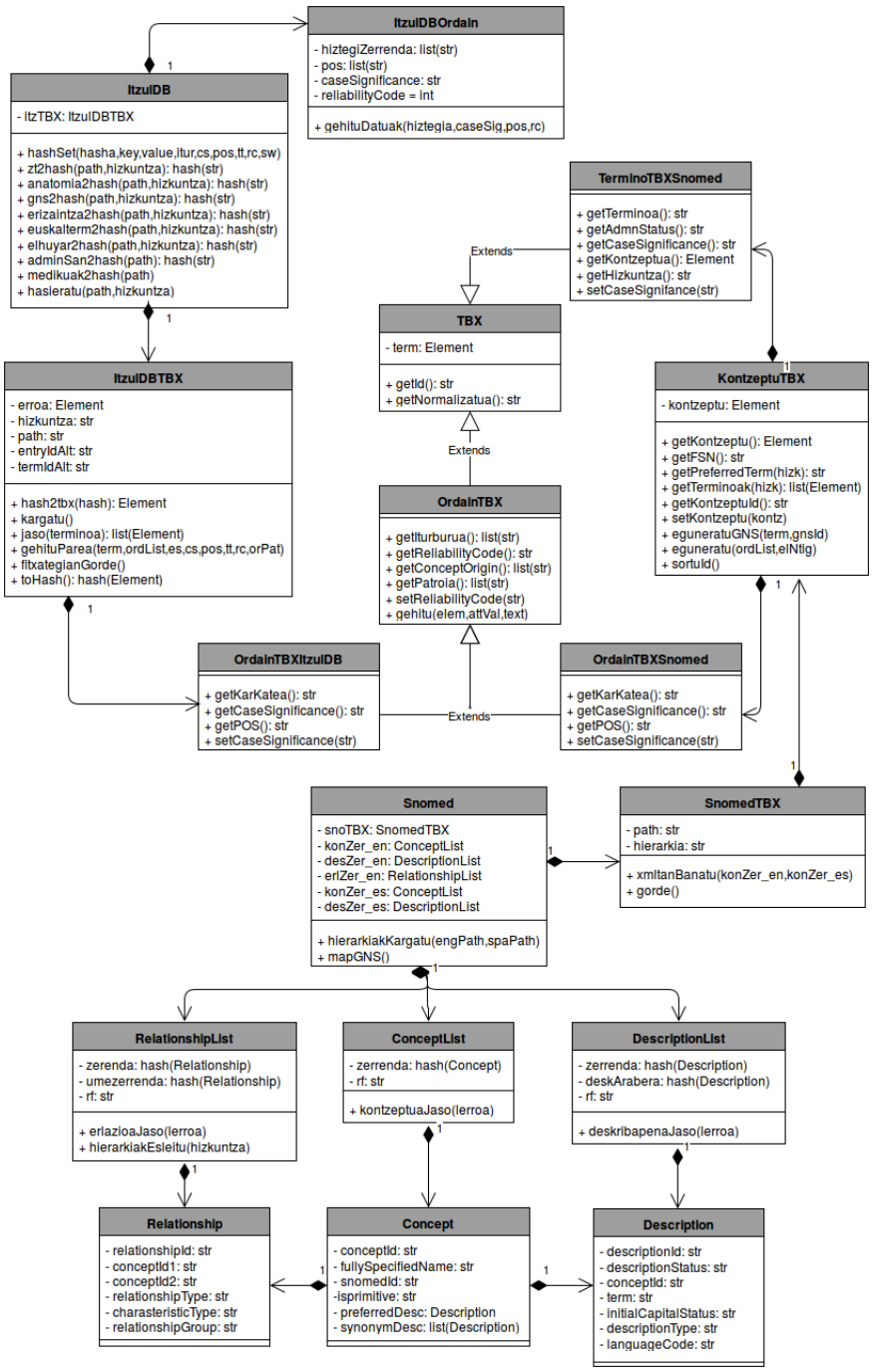
Klase-diagramaren diseinua 4.10 irudian ikus daiteke. Bertan ItzulDB kudeatzeko klaseak eskemaren goiko ezker aldean kokatu ditugu, eta gainontzekoak SNOMED CT kudeatzeko klaseak dira.

Klase bakoitzaren funtzio nagusia azalduko dugu jarraian:

- `ItzulDB`: ItzulDB bera kudeatzeko klasea da. Besteak beste, hiztegi bakoitzaren formatua kontuan hartuta, parekatzeak erauziko ditu ItzulDB beraiekin hasieratzeko.

- `ItzulDBOrdain`: ItzulDBren ordainen informazioa prestatzeko klasea da. TBX formatura pasa aurretik erabiliko da.

- `ItzulDBTBX`: ItzulDBren XML dokumentua kudeatzeko klasea da. Parekatze baten bilaketa, parekatze berria gehitzea edota ItzulDB objektutik jasotako informaziotik XML dokumentua sortzea izango dira klase honen ataza nagusiak.



4.10 irudia – Aplikazioaren diseinurako klase-diagrama definitiboa.

- `TBX`: Klase honek `ntig` edo `tig` elementu bat kapsulatuko du. Bi elementu horiek, identifikadorearen kokapena eta termino normalizatuarena izango dute komunean kasu guztietan (ItzulDB zein SNOMED CT eta termino zein ordain). Hortaz, klase honek bi datu horiek lortzeko metodoak eskainiko ditu.

- `TerminoTBXSnomed`: SNOMED CTren jatorri-terminoen gaineko informazioa eskuratzeaz arduratuko da.

- `OrdainTBX`: TBX formatuan ordainek duten informazio komuna atzituko du, iturburua, konfiantza maila edo jatorri-terminoaren identifikadorea, besteak beste.

- `OrdainTBXSnomed`: SNOMED CTren ordainen gaineko informazioa eskuratzeaz arduratuko da.

- `OrdainTBXItzulDB`: ItzulDBren ordainen gaineko informazioa eskuratzeaz arduratuko da.

- `KontzeptuTBX`: SNOMED CTren kontzeptuei loturiko informazioa eskuratzeaz gain, ordain berriak TBX dokumentuan txertatzeaz arduratuko da klase hau.

- `SnomedTBX`: SNOMED CTren XML dokumentuak kudeatuko ditu. Besteak beste, fitxategiak sortu eta gorde. Kontzeptuak, terminoak eta ordainak kudeatzeko klaseetara lotura egitea ahalbidetuko duen klasea ere bada.

- `Snomed`: SNOMED CTren kudeaketaz arduratuko da, XML dokumentua ikusezina eginez, eta horrela kodetzea erraztuaz. Horretaz gain, bi eginkizun nagusi izango ditu: SNOMED CTren jatorrizko fitxategietatik eduki terminologikoa eta egitura hierarkikoa erauzteko objektuei deitzea eta GNS10 eta SNOMED CTren arteko mapaketa egitea.

- `RelationshipList`: SNOMED CTren erlazioen fitxategitik erlazioak jaso, eta kontzeptuak modu hierarkikoan sailkatuko ditu erlazio horren arabera.

- `Relationship`: erlazio bakoitzaren informazioa kapsulatuko du.

- `ConceptList`: SNOMED CTren kontzeptuen fitxategitik informazioa jasoko du. Kontzeptuen harremanen inguruko informazioa kudeatzeaz ere arduratuko da.
- `Concept`: kontzeptu bakoitzaren informazioa kapsulatzen du eta erlazio zein deskribapenekin loturak kudeatuko ditu.
- `DescriptionList`: SNOMED CTren deskribapenen fitxategitik deskribapenen informazioa jasoko du eta kontzeptuen informazioa osatzeaz arduratuko da. Adibidez, *Fully Specified Name*-ak identifikatu eta dagokion kontzeptuari gehituko dio.
- `Description`: deskribapen bakoitzaren informazioa kapsulatuko du.

4.5 Laburpena eta ondorioak

Kapitulu honetan, garatu dugun EuSnomed sistema aurkeztu dugu. Sistema horrek SNOMED CT medikuntzako sare semantikoa euskaratzea du helburu, nahiz eta modu errazean egokitu daitekeen beste hizkuntza batzuetarako.

SNOMED CTren deskribapenak euskaratzeko, lau urratsetako algoritmoa diseinatu dugu. Lehenengo urratsak baliabide lexikal espezializatu eta elebidun/eleanitzak erabiltzen ditu ordainak lortzeko. Bigarrenak, termino neoklasikoak euskaratzen ditu, afixuen baliokideen eta transliterazioaren bidez. Hirugarren urratsak, termino konplexuen egitura habiaratuan oinarritzen da (termino bat beste baten barruan) ordainak sortzeko patroiak definitzeko. Azkenik, laugarren urratsean, itzultzaile automatiko orokor bat egokitzen da termino konplexuen ordainak lortzeko.

EuSnomed sistemak, SNOMED CTren eduki terminologia erabiltzeaz gain, sortutako baliabideak (termino-ordain pareak) berrerabiliko ditu. Hori horrela izanik, baliabide guztiak biltegiratzeko XMLn oinarritzen den formalismo estandar bat egokitu dugu: TBX. Formalismo horren bitartez, euskaratze prozesuan beharrezkoa izango den informazio guztia modu egituratuan eta ulergarrian biltegiratuko dugu.

Bukatzeko, EuSnomed sistemaren klase-diagrama aurkeztu dugu. Klase-diagrama, biltegiratze-eskeman oinarritu dugu, bertatik jasoko baita euskaratze prozesuaz arduratzen den algoritmoak beharrezko informazioa.

5

Termino sinpleak: Baliabide lexikalak eta termino neoklasikoak

Kapitulu honetan termino sinpleak (hitz bakarreko terminoak) euskaratzeko erabilitako teknikak aurkeztuko ditugu. Alde batetik, SNOMED CTren terminoen ordainak eskura genituen baliabide lexikalak erabiliz (hiztegiak) bilatu ditugu. Beste aldetik, termino neoklasikoak euskaratzeko sistema bat garatu dugu, NeoTerm. Sistema horretarako, termino neoklasikoen afixuen lexikoi elebiduna eta transliterazio erregela multzoa sortu ditugu. Informazio hori transduktoreen teknologia erabiliz konbinatu dugu, eta, horiek baliatuz, termino neoklasikoen euskaratzea bideratu dugu, euskararen arau morfofonologikoak eta ortografikoak errespetatuz.

Jarraian, erabilgarri genituen baliabide lexikalak aurkeztuko ditugu 5.1 atalean; bigarrenik, termino neoklasikoen sorkuntzarako prestatutako sistema azalduko dugu 5.2 atalean; ostean, 5.3 eta 5.4 ataletan, sistemen ebaluazioaren diseinua eta izandako emaitzak aztertuko ditugu; eta bukatzeko, 5.5 atalean, atera ditugun ondorioak eta kapituluaren laburpena aurkeztuko ditugu.

5.1 Baliabide lexikalen eta SNOMED CTren arteko parekatzea

Atal honetan 4. kapituluan definitutako algoritmoaren lehen urratsa azalduko dugu sakonki, dagoeneko sortuta dauden baliabide lexikalen erabilera,

hain zuzen ere. Urrats horretarako erabilitako baliabideen aurkezpena egingo dugu 5.1.1 atalean. Jarraian, 5.1.2 atalean baliabide horien gainean egindako aurre-prozesaketari buruz arituko gara. Bukatzeko, parekatze prozesua azalduko dugu 5.1.3 atalean.

Balibaide lexikalak ustiatzeko motibazio nagusia termino sinpleen euskaratzea izan bada ere, termino konplexu batzuk ere euskaratu ditugu urrats honetan, hiztegietan agertzen direnak, alegia.

5.1.1 Aurrekariak

Terminoak sortzeko algoritmoaren lehen urratserako, osasun-zientziaren domeinuan eskura ditugun euskarazko baliabide lexikal elebidunak bildu ditugu. Hurrengo lerroetan baliabide horiek zerrendatu ditugu:

- **Zientzia eta Teknologiaren Hiztegi Entziklopedikoa (ZT Hiztegia)** (Elhuyar, 2009): izenak dioen bezala, zientzia eta teknologiaren hiztegi bat da hau, Elhuyar Fundazioak garatua. Hiztegiaren izateko arrazoia zientzia eta teknologiari buruzko erreferentzia-informazio fidagarri, landu eta eguneratua eskaintzea da, modu zehatz, argi eta ulergarrian, eta erabiltzaile-multzo zabala gogoan hartuta. Besteak beste, medikuntzako, biokimikako, biologiako, anatomiako eta psikiatriako alorrak aurki daitezke. Euskara, espainiera, ingelesa eta frantsesa barnebiltzen ditu hiztegiak.

- **Euskalterm** (UZEI, 2004): Euskarazko Terminologia Banku Publikoa da, Eusko Jaurlaritzak kudeatzen duena. UZEIk Euskalterm 1986an sortu zuen ordura arte zituen hiztegi terminologikoak bateratuz. Hiztegi berriak integratuz doa orduz geroztik eta Hiztegi Batuko arau berriekin ere eguneratzen du bankua. Euskara, espainiera, ingelesa eta frantsesaz gain, latineko zein alemanierako sarrera batzuk ere baditu.

- **Giza Anatomiako Atlasa** (UPV/EHU Argitalpen Zerbitzua, 2014): giza anatomiako erreferentziazko *"Master" Atlas de Anatomia* testuliburuaren euskarazko itzulpena da. UPV/EHUk Eusko Jaurlaritzaren laguntzaz argitaratu zuen 2014an eta bertan giza gorputza bere osotasunean azaltzen da.

- **Erizaintzako Hiztegia** (EHUko Euskara Zerbitzua eta Donostiako Erizaintza Eskola, 2005): EHUko Euskara Zerbitzuak eta Donostiako Erizaintzako Unibertsitate Eskolak argitaraturiko hiztegi honetan,

erizaintzan erabiltzen diren hainbat termino aurki daitezke. Bertan euskarazko definizioekin osatutako hiztegiaz gain, espainiera-euskara, ingelesa-euskara eta frantsesa-euskara hiztegi elebidunak daude.

- **Gaixotasunen Nazioarteko Sailkapenaren 10. bertsioa (GNS10)** (World Health Organization *et al.*, 1996): Gaixotasunen Nazioarteko Sailkapena gaixotasunak sailkatzeko irizpideak eskaintzen dituen Munduko Osasun Erakundeak sortutako nazioarteko sailkapena da. Irizpide horien bitartez, gaixotasunei kode estandar bat edo gehiago esleitzeko aukera ematen zaie, hizkuntza eta herrialde guztietan berdina izango dena. GNS10ek ***I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems* jatorrizko ingelesezko izena dauka eta 10. bertsioa da (ICD10) gaur egun indarrean dagoena. ***W**orld **H**ealth **O**rganization* (WHO) erakundeak banatu zuen azken berrikuspena 1992 urtean. Euskal Herrian UZEIk euskaratu zuen eta 1996an argitaratu zuen Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusiak.

- **Administrazio Sanitarioko Oinarrizko Hiztegia** (Osakidetza, Euskadiko Osasun Saila eta UZEI, 1999): hiztegi txiki honek administrazio sanitarioan erabiltzen diren terminoak barnebiltzen ditu espainiera-euskara hizkuntza parerako. Eusko Jaurlaritzaren Osasun Sailak, UZEI eta Osakidetzarekin batera argitaratu zuen 1999 urtean.

- **Elhuyar Hiztegia** (Elhuyar, 2007a, b): erreferentziazko hiztegi orokor eleanitza da Elhuyar Hiztegia. Euskara-espainiera, euskara-ingelesa zein euskara-frantsesa hizkuntza pareak eskaintzen ditu. Ez du terminologia espezializatua barnebiltzen.

Atal honetan aurkeztutako baliabide lexikalen laburpena erakusten dugu 5.1 taulan, hiztegi espezializatuak diren eta eskaintzen diren hizkuntza-pareak erakutsiz.

Ikus dezakegunez, euskararen eta ingelesaren, espainieraren eta frantsesaren arteko parekatzeak ditugu hiztegietan eskuragarri. Gaur egunean, SNOMED CT ingelesez zein espainieraz eskuragarri dago beste hizkuntzen artean, baina frantsesezko bertsioa ez dago bukatuta. Hori horrela izanik, ingelesa-euskara eta espainiera-euskara pareetako baliabideak erabiliko ditugu.

Baliabide lexikala	Espe.	en-eu	es-eu	fr-eu
ZT Hiztegia	✗	✗	✗	✗
Euskalterm	✗	✗	✗	✗
Giza Anatomiako Atlasa	✗	✗	✗	✗
Erizaintzako Hiztegia	✗	✗	✗	✗
GNS10	✗	✗	✗	✗
Administrazio Sanitarioko Hiztegia	✗	✗		
Elhuyar		✗	✗	✗

5.1 taula – Baliabide lexikoen laburpena. "Espe." zutabean hiztegi espezializatua den adierazten dugu.

Aipatutako baliabideez gain, SNOMED CTren banaketa ofizialaz arduratzen den SNOMED International erakundeak[1], SNOMED CTren eta GSN10-aren arteko mapaketa banatzen du (IHTSDO, 2012). Mapaketa erdi-automatiko hori *World Health Organization* erakundeak (WHO, GNSren sortzailea) eta SNOMED International erakundeak balioztatua izan da. Mapaketaren helburua SNOMED CTren kontzeptu bakoitzari GNS10en espazio semantikoan tokirik aproposena esleitzea da, GNS10 kode bat esleituz. Horrela, SNOMED CTren kontzeptuaren eta GNS10en kodearen/kodeen arteko estekа sortzen da. GNS10en euskarazko bertsioa izatean, mapaketa horri esker, bertan agertzen diren kontzeptuen euskarazko ordain zuzenak lor ditzakegu.

GNS10 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, osasun-zientzien arloko kontzeptu ezberdinak era hierarkikoan sailkatuz. Garrantzitsua da mapaketaren norantza aipatzea: SNOMED CTren kontzeptuen GNS10 kode baliokideak ematen ditu.

Denera 19.293 SNOMED CT kontzeptu daude parekatuta eta bakoitzerako gehienez 19 GNS10 kode esleitu dira. Hala ere, kasu gehienetan SNOMED CT kontzeptu bakoitzari GNS10 kode bat, bi edo hiru esleitu zaizkio. Mapaketan parte hartzen duten kontzeptuetarako denera 27.167 parekatze definitu dira. Kontzeptu gehienak nahasmenduak dira (23.393 parekatze), baina aurkikuntzak (3.171), gertaerak (184) eta egoerak (413) ere agertzen dira. Azalpena ulergarriago egiteko, hemendik aurrera, mapaketa hitza mapake-

[1]Orain gutxi arte, International Health Terminology Standards Development Organisation (IHTSDO) zeritzon.

ta osoari erreferentzia egiteko erabiliko dugu, eta parekatze hitza kontzeptu bakoitzarena adierazteko.

5.1.2 Aurre-prozesaketa

Hiztegiak, erabili ahal izateko, prestatu edo aurre-prozesatu behar izan ditugu. Adibidez, gure lanetarako testu hutsean (formatu, kolore eta marka berezirik gabe) behar ditugu hiztegiak.

Erizaintzako Hiztegiaren eta Administrazio Sanitarioko Oinarrizko Hiztegiaren kasuan PDF formatuan bakarrik daude eskuragarri, eta, bi zutabetan egituratuta egonik, horien testu formaturako (.txt) itzulpena ez da bat-batekoa izan. Formatu aldaketa horretarako jatorrizko formatua zatikatu egin dugu, eta ostean *Calibre* izeneko aplikazioak eskaintzen duen formatuen aldaketarako tresna erabili dugu. Tresna horrek adierazpen erregularren erabilera ahalbidetzen duenez, hiztegiak testu hutsezko formatura itzultzea posible egin digu.

Dena dela, nahiz eta hiztegi guztiek testu hutsezko formatua izan, hiztegien egitura bata bestetik oso ezberdina da. Hala, hiztegi batzuk termino sarrera bakoitza letra larriz idazten dute, honek berezkoa izan ala ez. Beste batzuk aldiz, letra larriak berezkoak direnean bakarrik erabiltzen dituzte, entitateak edota laburdurak adierazteko, adibidez. Irizpide ezberdin horiek ItzulDB (baliabide lexikal eleanitzak biltegiratzen dituen datu-basea) aberastea zaildu egiten dute, termino berdinaren bi idazketa ezberdin genituelako kasu askotan. Horri aurre egiteko, ItzulDBn terminoak minuskulaz datu-baseratu ditugu. Nahiz eta horrela entitateen berezitasunak galtzen ditugun, erredundantzia ekiditen dugu eta bide batez, baliabideen optimizazioa ere lortzen dugu.

Honetaz gain, euskarazko GNS10aren kasuan terminoak mugagabean agertu beharrean, mugatuan aurkitu ditugu. Kasu horietarako, zaila egin zaigu *a* itsatsia duten terminoak diren erabakitzea. Hortaz, gainerako hiztegiekin alderatu ditugu euskarazko ordainak, eta mugatuaren *a* gabe agertzen badira, forma mugagabea eman diegu; gainerako kasuetan, *a* itsatsia duela suposatu dugu eta bere horretan utzi ditugu.

GNS10aren ingelesezko bertsioari dagokionean, euskarazko bertsioarekin batera eskuragarri dagoenari hainbat termino falta zaizkio. Sarean eskuragarri dauden beste GNS10 fitxategiak aztertu ditugu eta bere terminoen egitura ez dator guk dugun GNS10aren terminoen egiturarekin bat. Terminoen osaeran aurkitu ditugu arazo handienak, terminoak adierazteko beste

formatu edo modu bat erabiltzen baitute. Aurretik aipatu dugun SNOMED CT eta GNS10aren arteko mapaketan, ingelesezko GNS10aren ingelesezko terminoak ere agertzen direnez, eta horien egitura gure ingelesezko GNS10 bertsioaren oso antzekoa denez, berauek erabili ditugu ingelesezko GNS10 elikatzeko. Horrela, ia 500 parekatze gehiago erabilgarri izan ditugu ItzulDB elikatzeko.

Gainerakoan, hainbat ataza txiki zein garrantzitsu egin ditugu hiztegietan, beharrezko informazioa modu egokian jaso ahal izateko:

- Interesatzen zaizkigun alorrak bakarrik erabili ditugu (ZT Hiztegian eta Euskaltermen, adibidez). Besteak beste, medikuntza, biokimika, anatomia, biologia, arnas aparatua, psikologia, eta abar.

- Metadatuak kendu ditugu fitxategi zein termino mailan. Hau da, espainierazko generoari buruzko informazioa, euskarazko deklinabide marka batzuk, eta tankera horretako informazioa kendu dugu. Adibidez, Elhuyar hiztegian *authority* ingelesezko terminoarentzat, euskarazko honako ordaina ematen du, besteak beste: agintari(ak). Kasu horretan, parentesi artean dagoen *ak* hori kendu dugu, "agintari" ordaina utzita.

- Bikoizketak ekidin ditugu eta sinonimoak erauzi ditugu. Hau da, hiztegien eduki guztia bateratu dugu, bikoiztutako sarrerak kenduta, eta hiztegiek emandako ordainak sinonimoetan gehituta. Adibidez, 5.2 taulan, *leprosy* terminoaren hiztegian aurkitu ditugun hiru ordainak ikus ditzakegu, ZT Hiztegian eta Erizaintzako Hiztegian agertzen direnak. Adibide horretan, hiru ordainak jasotzen ditugu *leprosy*ren ordain gisa, baina bikoiztu gabe.

Ordaina	**ZT Hiztegian**	**Erizaintzako Hiztegian**
legen beltz		✓
legen	✓	✓
legenar	✓	✓

5.2 taula – *Leprosy* terminoaren ordainak hiztegietan.

- Zegokion kasuetan kategoria gramatikala jaso dugu, nahiz eta hiztegi gutxi batzuek duten kategoria gramatikala modu sistematikoan adiera-

zita. Adibidez, Elhuyar Hiztegiak "n." erabiltzen du izenak etiketatzeko, "adj." adjektiboetarako (izenondo edo izenlagun bereizi gabe), eta abar.

Aurre-prozesaketaren ondorioz, 5.3 taulan erakusten ditugun ordain kopuruak dira ItzulDBn datu-baseratu ditugunak. Ikus dezakegunez, Euskalterm da terminoen ikuspegitik ordain gehien eskaintzen digun baliabidea. Nahiz eta Elhuyar Hiztegi orokorrak espainiera-euskara parerako dituen sarrerak beste guztien baturaren pare egon, Elhuyar Hiztegiak, orokorra izanik, ez ditu termino espezializatuak barnebiltzen, eta hortaz, osasun-zientzien domeinurako parekatze baliagarriak askoz gutxiago izango direla aurreikus dezakegu.

Baliabide lexikala	**Ingelesa**	**Espainiera**
ZT Hiztegia	9.621	13.568
Euskalterm	37.137	31.535
Giza Anatomiako Atlasa	5.644	6.042
Erizaintzako Hiztegia	5.348	6.183
GNS10	7.057	9.172
Administrazio Sanitarioko Hiztegia	-	1.799
Elhuyar	30.463	90.507
Ordainak denera	**96.138**	**158.806**

5.3 taula – Baliabide lexikoen tamaina.

5.1.3 Hiztegien parekatzea

Algoritmoaren urrats honetan, ingelesezko zein espainierazko baliabideak SNOMED CTren terminoekin parekatu ditugu. Hau da, SNOMED CTren bi bertsioak erabili ditugu kontzeptuen euskarazko ordainak lortzeko. Aurrena, hiztegi espezializatuak erabili ditugu euskarazko ordainak lortzeko, ZT Hiztegia, Euskalterm, eta abar. Ordainak lortu ez diren terminoetarako, Elhuyar Hiztegi orokorra erabili dugu. Izan ere, Elhuyar Hiztegiak sinonimo asko ematen ditu baina orokorregiak kasu askotan. Murriztapenari esker, bakarrik hutsuneak betetzeko erabiltzen dugu Elhuyar, eta gainsorkuntza hein batean kontrolatzen dugu. Adibidez, *scar* terminorako lau ordain jaso ditugu Euskaltermetik: zikatriz, orbain, arrasto eta inpresio; baina Elhuyarreko

ordain gehigarriak ez ditugu jaso: ebakiondo, marka eta aztarna. Ikusten dugunez, Elhuyar zuzenean gehituko bagenu, zazpi ordain izango genituzke *scar* terminorako.

Jarraian, hiztegien parekatze-prozesuan hartutako erabaki nagusiak zerrendatzen ditugu.

- Aurretik aipatu bezala, lortutako zein sortutako parekatze guztiak ItzulDBn gordetzen ditugu, SNOMED CTrekin parekatu ahal izateko informazio bateratu zein egituratuarekin. Horrela, ItzulDBn parekatzeak bilatzeko, SNOMED CTren terminoak ere letra xehez gorde ditugu. Izan ere, aurreko atalean azaldu bezala, hiztegietan letra larri eta xeheen arteko erabilera ez da koherentea, eta zenbaitetan arazoak sortzen dizkigu. Hau da, ItzulDBn termino zein ordainak letra xehez gorde ditugu. Hala ere, SNOMED CTren deskribapenek, letra larri eta xehen inguruko informazioa ematen digute, eta letra xehetara pasatzeko muga adierazita duten kasuetan, ez ditugu aldatu. Adibidez, *Down syndrome* terminoak, letra larri eta xeheak bere horretan mantendu behar direla adierazita duenez, bere horretan gelditu da. Hala ere, ItzulDBn denak letra xehez daudenez, kasu horietan, SNOMED CTko terminoa letra xehetara pasatzen dugu bilaketa egiteko.

- Kasu batzuetan parekatzea ez da bat-batekoa izan. Izan ere, SNOMED CTren hainbat termino pluralean agertzen dira, eta espainierazko kasuan genero femeninoan ere ager daitezke. Kasu horietan, SNOMED CTren terminoak ItzulDBn agertzen ez badira, eta terminoen forma pluralean edota femeninoan badago, bere forma kanonikora pasata (singularra eta maskulinoa) egin dugu parekatzea. Euskaraz normalean generoan bereizketarik egiten ez denez, jatorrizko terminoan genero aldaketek ez dute ordainengan eraginik izango; bai, ordea, numeroaren kasuan. Jatorrian pluralean dauden terminoen forma singularra bilatzean, ordainak singularrean jasotzen ditugu, eta hortaz, ordain horiek pluralera pasatzen ditugu, parekatzea ondo egon dadin. Adibidez, SNOMED CTn *meninges* terminoa agertzen da, pluralean (bai espainieraz, bai ingelesez), baina hiztegietan ditugun sarreretan *meninge* (espainieraz) edo *meninx* (ingelesez) agertzen dira, singularrean. Zuzeneko parekatzea posible ez den arren, SNOMED CTren terminoa singularrera pasata, parekatzea hiztegietan lortzen badugu, euskarazko ordainei forma plurala ematen diegu jatorrizkoarekin bat etor dadin.

- Parekatzearekin bukatzeko, kategoria gramatikala ere jaso behar dugu. Informazio hori, termino habiaratuen sorkuntzarako garrantzitsua da, aurrerago ikusiko dugun bezala. Hala ere, kasu gramatikala oso sarrera gutxietan dator esplizituki jasota. Kasu gramatikalik ez daukaten ordainetarako, terminoen izaeran izenak izatea denez ohikoena, izen kategoria eman diegu hobetsita. Salbuespena, kalifikatzaileen hierarkiarekin (*"Qualifier value"*) egin dugu, eta izenlagun edo izenondo kategoria esleitu diogu ordain bakoitzari, kasuan kasu. Adibidez, *fifth*, *chronic* edo *yellow* SNOMED CTren kalifikatzaileak dira. Ingelesez, zein espainieraz adjektibo kategoria identifikatzearekin nahikoa den arren, euskaraz ez zaigu nahikoa gure interesetarako. Izan ere, izenlagun edo izenondo izan, sintagma barruko hitzen hurrenkeran eragina dauka, eta hain zuzen ere, hurrengo kapituluan ikusiko dugun moduan, hori da termino habiaratuen sorkuntzan behar dugun informazioa. Ordain bat izenlagun edo izenondo den erabakitzeko, atzizkia hartu dugu kontuan. Hobetsitako kategoria izenondoa da, eta "-ko" edo "-(r)en" atzizkia duten kasuetan izenlagun kategoria ematen diegu, Euskaltzaindiaren Hiztegiaren definizioari jarraiki[2].

Aurreko kapituluan aurkeztutako algoritmoak (4.2 atala) sarrera gisa termino bat jasotzen du eta horren ordainak ematen ditu irteeran. Lehen azpimarratu bezala, termino mailan lan egiten du. Hala ere, aurreko atalean aipatu dugun moduan (5.1.1 atala) SNOMED CT eta GNS10aren arteko mapaketa ere eskuragarri dugu, kontzeptu mailako parekatzea egiten duena. Mapaketa hori euskaratze-algoritmotik at dagoen arren, gure erabakia eskura ditugun baliabide guztiak erabiltzea denez, mapaketa hori ere erabili dugu algoritmoa exekutatu aurretik.

Mapaketak jatorritzat SNOMED CT kontzeptua eta helburutzat GNS10 kodeak hartzen ditu, baina horretaz gain, parekatze bakoitzerako hainbat informazio gehigarri eskaintzen du. Besteak beste, mapaketaren nomenklaturari jarraiki, *mapGroup*, *mapPriority* eta *mapRule* atributuak zehazten ditu. *MapGroup* parekatzeak multzokatzeko erabiltzen den atributua da, aurrerago ikusiko dugun bezala, multzo bakoitzean parekatze bat baino gehiago egon daiteke, erabakitzeko testuinguruaren beharra dagoelako. Beste batzuetan aldiz, bi parekatze independentek osatzen dute parekatze

[2]http://www.euskaltzaindia.eus/index.php?sarrera=izenlagun&option=com_hiztegianbilatu&view=frontpage&Itemid=410&lang=eu&bila=bai (2017ko maiatzaren 15ean atzitua).

osoa. Adibidez, 5.4 taulan, SNOMED CTren *Insulin-resistant diabetes mellitus AND acanthosis nigricans (disorder)* kontzeptuaren parekatzea erakusten dugu. Ikus daitekeenez, kontzeptu hori GNS10eko bi kontzepturekin adierazten da *Insulin-dependent diabetes mellitus (with modifiers)* eta *Acanthosis nigricans*. Hala ere, kasu batzuetan SNOMED CTren kontzeptu bat GSN10

mapGroup	**GNS10 kodea**	**GNS10 Terminoa**
1	E10.9	*Insulin-dependent diabetes mellitus (with modifiers)*
2	L83	*Acanthosis nigricans*

5.4 taula – SNOMED CTren *Insulin-resistant diabetes mellitus AND acanthosis nigricans (disorder)* kontzeptuaren GNS10 parekatzea.

kode bakarrarekin parekatzen den arren, testuinguruaren arabera baldintzatua da. Horretarako *mapPriority* eta *mapRule* atributuak erabiltzen dira, besteak beste. *MapPriority* atributua parekatzeen arteko ordena zehazteko erabiltzen da, eta *mapRule* baldintza zehazteko. Adibidez, 5.5 taulan, testuingurua behar duen parekatze multzoa erakusten dugu. `IFA 7200002 | Alcoholism (disorder)` *mapRule*-aren bitartez, 7200002 kontzeptu identifikadorea duen *Alcoholism (disorder)* kontzeptuaren testuinguruan gertatzen bada, F10.2 kodearekin lotu behar duela adierazten da.

mapGroup	***mapPriority***	***mapRule***	**GNS10 kodea**	**GNS10 Terminoa**
1	1	`IFA 7200002 | Alcoholism (disorder)`	F10.2	*Mental and behavioural disorders due to use of alcohol (with modifiers)*
1	2	`IFA 231467000 | Absinthe addiction (disorder)`	F10.2	*Mental and behavioural disorders due to use of alcohol (with modifiers)*
1	3	`OTHERWISE TRUE`	F19.2	*Mental and behavioural disorders due to multiple drug use and use of other psychoactive substances (with modifiers)*

5.5 taula – SNOMED CTren *Psychoactive substance dependence (disorder)* kontzeptuaren GNS10 parekatzea.

Ikusi dugun moduan, parekatzeak ez dira kasu guztietan bat batekoak, baldintza eta testuinguru gabekoak. Hori guztia kontuan izanik, eta gure

atazaren helburuei jarraiki, garatuko dugun sistemarako baliokide gisa har daitezkeen parekatzeak bakarrik izan ditugu kontuan. Hau da, *mapGroup* bakarra dutenak eta baldintzarik ez duten parekatzeak. Irizpide horiei jarraiki, 15.099 parekatze ditugu erabilgarri, "zuzeneko" parekatze gisa kontsideratu daitezkeenak.

SNOMED CT eta baliabide lexikalen arteko parekatzearen emaitzak, baita GNS10ekin mapaketaren emaitzak ere, 5.4 atalean aurkeztuko ditugu.

5.2 Termino neoklasikoen sorkuntza

Atal honetan, EuSnomeden algoritmoaren bigarren urratsa azalduko dugu, termino neoklasikoen sorkuntza automatikoa, hain zuzen ere. Sorkuntza horretako, NeoTerm deituriko sistema garatu dugu, ingelesezko termino neoklasiko bat jasota, euskal ordainak sortzen dituena.

Termino neoklasikoak, jatorri grekoa edo latindarra duten morfemez osatuta dauden terminoak dira, adibidez hipogluzemia edo fotodermatitis.

Lehenik, 5.2.1 atalean ataza honen aurrekariak azalduko ditugu, artearen egoera eta erabilitako teknologiak azalduz hurrenez hurren. Ostean, NeoTerm sistema eta garatu ditugun hiru hurbilpenak aurkeztuko ditugu: oinarri-lerro sistema (5.2.2), transliterazioa barnebiltzen duen sistema (5.2.3) eta doitasunean arreta jartzen duen sistema (5.2.4).

5.2.1 Aurrekariak

Atal honetan termino neoklasikoen sorkuntza automatikoaren artearen egoera azalduko dugu. Horretarako, termino neoklasikoen inguruko azalpenak eta eztabaidak emango ditugu aurrena. Jarraian, beste hizkuntzetarako egindako lanak azalduko ditugu. Bukatzeko, gure sistemaren garapenean erabilitako egoera finituko transduktoreen inguruko azalpenak emango ditugu.

Termino neoklasikoak

Medikuntzaren alorrean idazten diren artikulu zientifiko gehienak ingelesez idazten dira, nahiz eta egileen ama hizkuntza beste bat izan (Giannoni, 2008 eta Gunnarsson, 2009). Panocová (2015) lanean adierazten den bezala, erdiaroan latina zen hizkuntza nagusia medikuntzan, eta XVII. eta XVIII. mendeetan, medikuntzak garapen handia izan zuen garaian, eragin hori oso na-

baria izan zen lan gehienak latinez idatzi baitziren. Ingelesa latinaren tokia hartzen joan den arren *lingua franca* gisa, medikuntzaren terminologia latinezko eta grezierazko elementuen agerpenez josita dago. Are gehiago, Banay-k (1948) dio medikuntzako terminologiaren hiru laurdenak jatorri grekoa duela.

Grezierako eta latineko elementu morfologikoez osatutako terminoei termino neoklasiko konposatuak deritze, eta oso errotuta daude medikuntzaren domeinuan. Termino neoklasikoei izendatze ezberdinak eman izan zaizkie: termino neoklasiko konposatuak, konposatu neoklasikoak, eta abar.

Sorkuntza-elementu neoklasikoak latinetik zein grezieratik datozen elementu morfologikoak dira (Elia *et al.*, 2015). Horiek, hitz tekniko-zientifiko zein hitz arruntak sortzeko erabiltzen dira, beste sorkuntza-elementu batzuekin konbinatuz, edota bestelako hitzekin konbinatuz.

McCray *et al.* (1988) lanean konposatu neoklasikoak bi morfema motetan sailkatzen dituzte: erroak eta amaierakoak. Horiekin batera, aurrizki (*non-*, adibidez *nononcogenic* terminoan) zein atzizki (*-al*, adibidez *nosocomial* terminoan) orokorrak ere erabili dituzte azterketan.

Literaturan ez dago koherentziarik izendatze horien gainean, eta erroei aurrizki ere deitzen diete. Gure lanean, aurrerago ikusiko dugun bezala, erroak eta aurrizkiak elkarrekin landuko ditugu, aurrizki terminoa erabiliz beraiek izendatzeko, zabalduagoa dagoelakoan. Amaierako morfemen kasuan, atzizki terminoaren barruan landuko ditugu.

McCray *et al.* (1988) lan horretan, konposatu neoklasikoen elementuak automatikoki identifikatzeko analisi morfologikorako aplikazioa garatu dute, konposatuak deskonposatu egiten dituena. Sistema horri esker, konposatu neoklasikoen esanahiaren eta elementuen esanahiaren arteko harremana aztertu dute.Wolff *et al.*-ek (1984) baieztatzen duenaren arabera, konposatu neoklasiko baten esanahia, bere zatien esanahien funtzioa da. Hala ere, Bauer-ek (1983) eta Dirckx-ek (1977) horren kontra egiten dute, eta elementuen eta konposatuaren arteko harremana ez dela hain gardena esaten dute.

McCray eta lagunen ondorioen arabera, Bauer-ekin eta Dirckx-ekin bat eginda, konposatu neoklasikoen esanahia ezin da zuzenean bere elementuetatik erauzi. Izan ere, konbinaketak ez du beti zuzenean esanahia deskribatzen, eta modu indeterminatuan funtzionatzen duela dio, horretarako arrazoi oso ezberdinak erakutsiz. Adibidez, *leuko-* (txuria) erroaren kasuan, esanahi ezberdina dauka konposatuaren arabera: *leukotherapy* edo *leukotomy* konposatuek leukozitoekin (odol zelula txuriak) zerikusia duten bitartean, *leukophatia* melaninaren galerarekin dauka zerikusia eta *leukoscope* koloreen

itsutasuna neurtzeko gailua da.

Indeterminazioaren beste adibide bat konposaketaren esanahian ikus dezakegu *-itis* atzikiarekin. *Laryngotracheobronchitis* terminoak "laringearen, trakearen eta bronkioen hantura" esan nahi duen bitartean (hiru erroen juntadura, atzizkiagatik eraldatuak), *phytophotodermatitis* terminoak "dermatitis fototoxikoa landare batzuei eta eguzkiari esposizioak eragindakoa" esan nahi du. Azken kasuan, analisiak ez du atzikiagatik eraldatutako erroen juntadura egitura betetzen.

Leuko-ren kasuan ikusi dugun moduan, *-itis*en kasuan ere, atzizkiaren esanahitik ezin da termino osoaren esanahia zuzenean ondorioztatu. Jatorrian, *-itis* atzizkia, gaixotasunak adjektibo bilakatzen zituen atzizkia bazen ere (grezieraz *nosos* gaixotasuna da), gaur egunean *-itis* atzizkiak hantura adierazten du, *carditis* bihotzaren hantura izanik, edo *hepatitis* gibelaren hantura.

Erabileraren bilakaeraren ondorioz, termino neoklasikoen deskribapena anbiguoa gertatzen da, eta ez dago termino neoklasikoen deskribapen formalik, Ananiadou (1994) lanean berresten den bezala.

Beste hizkuntzetarako sistemak

Termino neoklasikoen sorkuntza automatikoari buruz literaturan aurkitu ditugun lanak corpusetan oinarritutako lanak izan dira. Horretaz gain, corpus elebakar zein elebidunetan oinarrituta, terminoen identifikazioan edota lexikoien aberasketan ere erabili dira teknika antzekoak. Ikus ditzagun sistema horietako batzuk.

Grigonytė *et al.* (2016) lanean afixu neoklasikoen *suedifikazio* (suedierara bihurketa) patroiak aztertu dituzte. Aipatzen dutenaren arabera, suedierazko txosten klinikoetan termino neoklasikoak jatorrizko formarekin erabili izan dira denboran zehar, 1987an ortografia erreforma egin zuten arte. Ortografia erreforma horri esker, termino neoklasikoak *suedifikatu* zituzten, hainbat transliterazio-erregela aplikatuz, suedirako ortografia arauak errespeta ditzaten, hala nola, *ae* edo *oe*, *e* bilakatzen dira, edota *ph*, *f*. Afixu neoklasikoen erabileraren gaineko azterketa honetan, gaur egunean kasu gehienetan termino *suedifikatuak* erabiltzen direla ondorioztatu dute. Azterketa egiteko, txosten klinikoetan termino neoklasikoak identifikatzeko erregela batzuk definitu dituzte, afixuen zerrendetan oinarrituta.

Gooch eta Roudsari (2011) lanean, ingelesezko termino neoklasikoak identifikatzeko modulua garatu dute, erregeletan oinarritutakoa. Morfema neo-

klasikoen zerrenda osatu dute eta aurrizki, erro edota atzizki gisa sailkatu dituzte. Morfema horiek gorputz-egiturei, seinu klinikoei edota posizioari zein deskribapenari loturikoak dira. Termino kliniko bat definitzeko oso esanguratsuak diren atzizkien kasuan (*-itis* edo *-ostomy* bezalakoak), beste multzo bat egin dute. Morfemen multzoak adierazpen erregularren bidez konbinatu dituzte termino neoklasikoak identifikatzeko. Modulu hori, transduktore bat erabiliz inplementatu dute. Tamalez, lan horretan identifikaziora mugatu dira, eta ez dute termino neoklasikoen ordainak sortzeko inolako metodorik garatu.

Hurbilak diren hizkuntzetarako, Schulz *et al.* (2004) lanean kognatuak[3] lotzeko sistema aurkezten dute, portugesetik abiatuta espainierakoak lortzeko. Sistema horrek, erlaziorik ez duten domeinu bereko corpusak erabiltzen ditu eta azpi-hitzen baliokidetzak erabiliz, kognatuak antzematen dituzte. Prozesuan, azpi-hitzak eskuz sortu dituzte, transliterazio arau automatiko batzuen laguntzaz. Azpi-hitzak definitzeko garaian, ez dituzte irizpide linguistikoak kontuan hartu, semantikoei gehiago loturikoak baizik, baina ez dituzte zehaztapenak gehiegi azaltzen.

Hizkuntza hurbilentzako, corpus ez-paraleloetan oinarritzen diren lanak ere aurkitu ditugu, Koehn eta Knight (2002) adibidez. Bestalde, corpus ez-paraleloetan oinarritzen diren hauek, analizatzaile morfologikoen laguntza erabiltzen dute emaitza hobeak lortzeko (Hahn *et al.*, 2001; Namer eta Zweigenbaum, 2004).

Termino neoklasikoen kasuan, Lovis *et al.* (1995) lanean aipatzen den bezala, termino berriak dagoeneko definituta dauden unitate morfologikoak kateatuz sortzen dira, eta medikuntzaren domeinuan teknika hori da termino berriak sortzeko erabiltzen den teknika zabalduenetako bat.

Egoera finituko automatak eta transduktoreak

Hizkuntzaren prozesaketaren alorrean, egoera finituko makinak oso erabiliak izan dira, espezializazio askotan. Horren erakusle da, 12. ediziotik doan *Finite-State Methods and Natural Language Processing* kongresua.

Makina horien erabilera, ikuspegi linguistiko zein konputazionaletik bidezkotu daiteke (Mohri, 1997). Ikuspegi linguistikotik, egoera finituko makinek hizkuntzen azterketa enpirikotik erauzitako fenomeno gehienak modu errazean deskribatzea ahalbidetzen dute. Linguistentzat modu naturalean

[3]Kognatuak jatorri etimologiko berdina duten hitzak dira. Adibidez, italierazko *mangiare* eta frantsezko *manger*.

(Gross, 1987), eta gainera, bisualizazio tresnek gramatika aztertzen eta fintzen lagun dezakete. Ikuspegi konputazionaletik, aldiz, eraginkortasun arrazoiak dira nagusi.

Egoera finituko automatek lengoaia formalak definitzen dituzte, karaktere-kate zehatzak onartu edo baztertzeko gai direnak. Transduktoreak (***F**inite **S**tate **T**ransducers* edo FST-ak), aldiz, bi karaktere-kateen arteko harremana zehazteko gai dira. Hau da, transduktoreak, sarrerako karaktere-kate bat izanda, irteerako karaktere-kate bat sortzeko gai dira.

Morfologiari dagokionean, FSTak maiz erabili izan dira morfologia aberatsa duten hizkuntzen analizatzaileak garatzeko, adibidez, euskararako (Alegria *et al.*, 1996), arabierarako (Beesley, 1996), turkierarako (Oflazer, 1994) edo alemanierarako (Schmid *et al.*, 2004).

Transduktoreen sorkuntzarako hainbat liburutegi daude, Xerox Finite State Transducer edo XFST (Karttunen *et al.*, 1997), Nooj[4] edo AT&T-ren Finite State Machine edo FSM (Mohri *et al.*, 2006) bezalakoak. Guk Foma izenekoa erabili dugu (Hulden, 2009), egoera finituko automatak eta transduktoreak definitzeko balio duen software-tresna askea.

Hurrengo ataletan, termino neoklasikoen euskaratze automatikorako garatutako sistemaren hiru hurbilpenak aurkeztuko ditugu. Aurrenik, oinarri-lerro sistema (*baseline system*) aurkeztuko dugu, eta jarraian, oinarri-lerro sistema horri egin dizkiogun bi hurbilpen: transliteraziorako modulua eta estaldura fintzeko irizpideak.

5.2.2 NeoTerm: oinarri-lerroa

NeoTerm sistema, medikuntzako termino neoklasikoak euskaratzeko diseinatu dugun sistema da. Sistema hori beste hizkuntzetara modu errazean egokitu badaiteke ere, lan honetan ingelesetik abiatuta euskararen kasua bakarrik garatu dugu.

Euskarazko ordainak sortzeko, hiru urrats inplementatu ditugu:

1. Afixuen identifikazioa: jatorri-terminoaren (ingelesezkoak kasu honetan) afixu neoklasikoen identifikazioa, eta horien araberako zatitzea.

2. Afixuen baliokidetzak: jatorri-terminoan identifikatutako afixuen baliokidetzak lortzea helburu hizkuntzan (euskara, gure kasuan).

[4]http://www.nooj4nlp.net/NooJManual.pdf (2017ko maiatzaren 9an atzitua).

3. Afixuen konposaketa: euskararen arau morfofonologikoak errespetatuz afixu baliokideen konposaketa sortzea.

Hurrengo irudian (5.1), *radionecrosis*[5] ingelesezko terminoaren euskaratze prozesua ikus dezakegu urratsez urrats. Lehenik, jatorri-terminoaren afixuak identifikatzen ditugu (*radio+necr+osis* edo *radio+necro+sis*). Ostean, afixu horien baliokideak lortzen ditugu lexikoietatik (radio+nekr+osi edo radio+nekro+si). Eta bukatzeko, afixuak konbinatzen ditugu euskal ordainak lortuz (erradionekrosi).

Jatorri-terminoa: *radionecrosis*
Identifikatutako afixuak: *radio+necr+osis*,*radio+necro+sis*
Afixu baliokideak: radio+nekr+osi, radio+nekro+si
Euskal ordaina: erradionekrosi

5.1 irudia – NeoTermen oinarri-sistemaren sorkuntza prozesua.

Aurreko atalean azaldu bezala, termino neoklasikoak grezieratik zein latinetik eratorritako afixuez osaturiko terminoak dira. Hori horrela izanik, jatorri-terminoaren afixuen identifikaziorako, ingelesezko medikuntzako afixuen bi zerrenda eskuratu ditugu 826 aurrizki eta 143 atzizki lortuz: Stedmanen medikuntza hiztegiko *"Medical Prefixes, Suffixes, and Combining Forms"* (Stedman, 2005) eta Wikipediaren *"List of medical roots, suffixes and prefixes"* (Wikipedia, 2013).

Ordainen sorkuntzarako zerrenda elebakar horietako afixuak banan-banan aztertu ditugu euskarazko baliokidetzak aurkitzeko. Horretarako, afixuen itzulpenak hiztegi espezializatuetako termino-ordain pareetatik ondorioztatu ditugu. Adibide gisa, 5.6 taulan *encephal-*[6] aurrizkiaren euskal baliokidea ondorioztatzeko prozesua erakusten dugu.

Ikus daitekeenez, hiztegiko sarrera guztietan ingelesezko *encephal-* aurrizkia, euskarazko ordainetan entzefal- aurrizki moduan agertzen da. Prozesu hori bera egin dugu zerrendetako 969 afixuekin (ikus http://ixa2.si.ehu.eus/neoterm/ web orria lortu ditugun afixuen baliokidetzak ikusteko).

[5]radionecrosis. (n.d.) Mosby's Medical Dictionary, 8. edizioa. (2009). http://medical-dictionary.thefreedictionary.com/radionecrosis (2017ko maiatzaren 15ean atzitua).

[6]encephal-. (n.d.) Mosby's Medical Dictionary, 8. edizioa. (2009). http://medical-dictionary.thefreedictionary.com/encephal- (2017ko maiatzaren 15ean atzitua).

Ingelesezko terminoak	Euskarazko ordainak
*an**encephal**ia*	an**entzefal**ia
*echo**encephal**ogram*	eko**entzefal**ograma
*electro**encephal**ograph*	elektro**entzefal**ografo
***encephal**itis*	**entzefal**itis
***encephal**omyelitis*	**entzefal**omielitis
***encephal**opathy*	**entzefal**opatia
*leuko**encephal**itis*	leuko**entzefal**itis
...	...

5.6 taula – *encephal-* aurrizkiaren euskal ordainak.

Hiztegiaren sorkuntza prozesu honetan, termino-ordain pareetan aurrizkien eta artizkien arteko bereizketa falta nabaritu dugu, eta sistemaren sinpletasunari begira, artizkiak aurrizki gisa kontsideratzea erabaki dugu. Izan ere, aurkitu ditugun alde guztiak euskararen ortografia arauek eragindako aldaerak izan dira, hitz hasierako *r* letraren agerpenak, adibidez. Horrela, hemendik aurrera aurrizki terminoa erabiltzen dugun aldiro, aurrizki zein artizkiei egingo diegu erreferentzia.

Sortutako hiztegi elebidunetik, bi lexikoi bereizi ditugu, bata aurrizkietarako (826 sarrera) eta bestea atzizkietarako (143 sarrera). Lexikoi horiek, aurrerago ikusiko dugun moduan, egoera finituko transduktoreak definitzeko erabili ditugu.

Lexikoiak definitzeko garaian, hainbat irizpide orokor definitu ditugu, ostean azalduko ditugun afixuen konposaketa-erregelak bete ahal izateko. Irizpide horiek ez dira modu automatikoan landu, aurrizki edo atzizki bakoitzaren izaeraren ezaugarriak adierazteko erabili baititugu.

Definitutako hiru irizpideak azaldu ditugu 5.7 taulan, adibide eta azalpenekin batera.

Aurrizkien lexikoiaren sarrera batzuk ikus ditzakegu 5.2 irudian. Sarreren ezkerreko aldean, ingelesezko aurrizkiak definitzen ditugun bitartean, eskuineko aldean euskarazko ordainak eman ditugu. Kasu honetan, gure lexikoietako sarrerei ez diegu jarraitze-klaserik esleitu, eta horrela, # karakterearekin bukaera elementua dela adierazten dugu.

en	eu	Azalpena
radio- *rhin-* *cineradiography*	rradio- rrin- zinerradiografia	*R* letra ingelesezko afixuen hasieran agertzen denean, *r* gogorra ahoskatu ohi da. Hortaz, euskararako *r* bikoitza erabiliko dugu. Termino hasiera denean, *r* bikoitza soil bilakatuko dugu ortografia arauei jarraituz.
bronch- *isch-* *thorac-* *bronchitis* *ischemia* *thoracic* *thoracotomy*	bronK- isK- torak- bronkitis iskemia toraziko torakotomia	Ingelesezko *ch* fonema, euskarazko *k*-ren baliokide fonologikoa da. Alabaina, *c*-ren ahoskerak aldaerak izaten ditu, jarraian datorkion letraren arabera: *e* edo *i*-rekin agertzen denean, euskaraz *z* erabiliko dugu eta gainerako kasuetan, *k*. Horrela, *K* letra larriz idatzi dugu *k* beti izan behar denean, eta *k* txikia hurrengo karaktereraren arabera aldatu behar dugunean.
bucc- *buccolabial* *buccinator*	buKZ- bukolabial bukzinatzaile	Ingelesezko terminoetan *cc* karaktereak agertzen direnean, hurrengo karakterearen arabera euskarazko *kz* ala *k* izango da, *ch* morfemarekin moduan. Hortaz, *KZ* letra larriz idatziko dugu, morfofonologia arauen arabera aukera dezagun.

5.7 taula – Afixuen baliokidetzen idazketarako irizpideak.

```
1 LEXICON Root
2 abdomin:abdomin #;
3 acanth:akant #;
4 acou:aku #;
5 aden:aden #;
6 adip:adip #;
7 adren:adren #;
```

5.2 irudia – Aurrizkien lexikoiaren sarrera batzuk.

Identifikazioa

Lexikoien "goiko aldearekin" (ingelesezko sarrerekin), 5.3 irudian ikus daitekeen transduktorea definitu dugu. Afixuen lexikoiak jasotzen ditugu aurrena (1-6 lerroak, `.u` aginduarekin, lexikoien ingelesezko sarrerak bakarrik jasotzen ditugu) eta orduan afixuak identifikatzen ditugu 7. lerroko erregelarekin: nahi adina aurrizki ("*" ikurrak 0 edo hainbat adierazten du) eta

derrigorrezko atzizki bakar bat. Ohikoa da -o- bokala, grezierako jatorria duten afixuak konektatzeko erabiltzea eta hortaz, gure erregelak -o- letra afixu gisa ere identifika dezake. Adibidez, 5.1 irudiko adibidean, ingelesezko *radionecrosis* terminoarentzat, 5.3 irudiko transduktorea aplikatu ondoren *radio+necr+osis* edo *radio+nekro+si* zatiak identifikatuak genituzke.

```
1 read lexc aurrizkiak.lex
2 define AURRDEN
3 define AURR AURRDEN.u ;
4 read lexc atzizkiak.lex
5 define ATZDEN
6 define ATZ ATZDEN.u ;
7 regex [[[AURR 0:+] (o 0:+)]* ATZ] ;
```

5.3 irudia – Afixuen identifikaziorako erregelak.

Atzizkia derrigorrezkoa egitea erabaki dugu, atzizkien baliokidetzak aurrizkienak baino konplexuagoak edo ezberdinak direla ikusi dugulako. Hau da, aurrizki gehienek transliterazio arauak jarraitzen dituzten bitartean (ingelesezko *ph*, euskarazko f bezalakoak), atzizkiek aldaera ezberdinak dituzte (ingelesezko *-itis*, euskaraz -itis mantentzen den bitartean, *-sis* atzizkia -si euskaratzen da). Hortaz, atzizkien euskaratzean prozesu konplexuagoek parte hartzen dutenez aurrizki batenean baino, aurrizki batek ezingo du gure transduktorean atzizki baten tokia bete.

Ikusi dugunez, afixuen identifikazioan zenbaitetan anbiguotasuna egoten da, eta hainbat aukera proposatzen ditu NeoTermek. Anbiguotasuna gutxitzeko, afixu gutxien (edo luzeenak) dituzten aukerak erabili ditugu baliokidetzen urratsera joan aurretik.

Ikus dezagun *photodermatitis* terminorako transduktoreak proposatzen dituen afixuen konbinazio ezberdinak:

- *photo+dermat+itis*: 3
- *photo+derm+at+itis*: 4
- *phot+o+dermat+itis*: 4
- *phot+o+derm+at+itis*: 5

Terminoa, 3, 4 edo 5 afixutan banatu daitekeela ikus dezakegu. Zehaztutako irizpideari jarraiki, lehenengo aukera izango da aukeratua afixuen baliokidetzak lortzeko eta euskal ordaina sortzeko.

Radionecrosis bezalako kasuetan anbiguotasuna ekiditeko zaila egin zaigun arren, kasu gehienetan honek ez du ordainen sorkuntzan gainsorkuntzarik eragiten. Izan ere, gehienetan ez da alderik egoten afixu antzekoen baliokideen artean *necr+osis* afixuen konbinaketak, *necro+sis* afixuen baliokideen konbinaketaren berdina baita: nekrosi.

Baliokidetza

Afixuen baliokidetzarako, hau da, ingelesezko afixuak euskarazkoekin ordezkatzeko, 5.4 irudian erakusten dugun transduktorea definitu dugu. Bertan, afixuen lexikoi osoa jasotzen dugu (1-4 lerroak) baita konposaketarako erregelak ere (5-7 lerroetan sinplifikatuta). Baliokidetzarako erregela (8. lerroa) horrela definitu dugu: hitz-hasierako ikurra (^ ikurra) espreski adierazi dugu, jarraian aurrizkia(k) (AURRBAL gisa adierazita) -o- konexio elementuaz konbinatuta agertzen dira (hautazkoa), eta bukatzeko atzizkia (ATZBAL) definitu dugu. Erregela guztiak, azkeneko adierazpen erregularraren bitartez konbinatzen ditugu (9. lerroa).

```
1 read lexc aurrizkiak.lex
2 define AURRBAL
3 read lexc atzizkiak.lex
4 define ATZBAL
5 define HASR r -> e r r || [.#.|^ ] _ ;
6 ...
7 define MORFO HASR .o. ...
8 define BAL (^) [[[AURRBAL +] (o:o +)]* ATZBAL] ;
9 regex BAL .o. MORFO ;
```

5.4 irudia – Afixuen baliokidetzarako erregelak.

Konposaketa

Afixuen konposaketarako, sortutako ordainek euskararen arau ortografikoak eta morfofonologikoak betetzeaz arduratzen diren 28 erregela definitu ditugu (Foman hauek ere). Arau ortografikoen artean v bezalako letren ordezkapenak aurkitzen ditugun bitartean, arau morfofonologikoetan hobikarien eraginez txistukarien sabaikaritzea aurkitzen dugu (euskaraz, entzefalitis esango genuke, eta ez *enzefalitis). Konposaketa erregelak B eranskinean ikus ditzakegu.

Erregela horiek definitzeko, alde batetik Euskaltzaindiaren ortografia arauak izan ditugu kontuan (Euskaltzaindia. Luis Mitxelena, 1968), eta bestetik

enpirikoki aztertutako aldaerak, denera 22 erregela definitu ditugu (B eranskinean laburbildu ditugu).

Aurreko adibideari jarraiki, afixuen konposaketak sortutako aldaeren artean daude gaurko erdaretan *r* letraz hasten diren hitzak. Hitz horiek euskaraz mailegatzean, bokalez hasten dira. Kasu gehienetan *e* letra erabiltzen da horretarako (erreinu, erradio, errezeta,...) nahiz eta kasu bakan batzuetan *a* letra ere erabili izan den (arropa, arrosa,...). Horrela, definitu dugun erregeletako batek, *r* letraz hasten den aurrizki bat aurkitzen duenean *e* letra gehituko dio.

Gure adibidearen erregela ikus dezakegu 5.4 irudiaren 5. lerroan. Bertan, *r* letra *err* letra segidarekin ordezkatzen da, baldin eta *r* letra hori hasieran dagoen (.#. edo ^ ikurren bidez adierazita). Hurrengo taulan (5.8 taula) erregela gehiago ikus ditzakegu.

	en	**eu**	**Adibidea**	**Erregela**
1	^ r	e r r	radionecrosis -> erradionekrosi	r -> err || [.#. | ^] _
2	^ s	e s	spermatocele -> espermatozele	s -> es || [.#.| ^] _ (+) Kon
3	s s	s	dyssomnia -> disomnia	s -> 0 || _ + s
4	n z	n t z	encephalitis -> entzefalitis	n -> n t || _ Txis + Bok
5	r s	r e s	hypersplenism -> hiperesplenismo	r + -> r e || _ Txis Kon
6	r r	r	hyperreflexia -> hiperreflexia	r r -> r || Kon + _
7	m p	n p	symphysis -> sinfisis	m -> n || _ + [b | p | t | f]
8	k	z	thoracic -> toraziko	
9	c c	k	buccolabial -> bukolabial	K Z -> k || _ + [Kon | a | o | u]
...	...	...	...	...

5.8 taula – Erregela morfofonologiko batzuk.

Azkenik, NeoTerm sistemaren hiru urratsak konbinatuta, euskarazko ortografia eta morfofonologia arauak errespetatzen dituzten euskal ordainak sortzen ditugu, 5.1 irudian ikusi dugun sorkuntza prozesua jarraituz.

Emaitzen atalean ikusiko dugun bezala, oinarri-lerro hurbilpen honek oso doitasun altua lortu badu ere, estaldurari dagokionean oso emaitza kaskarrak lortu ditugu. Hori horrela izanik, jarraian azalduko dugun hurbilpena garatu dugu estaldura hobetze aldera.

5.2.3 NeoTerm: transliterazio modulua

Hurbilpen honetan, bi ekarpen nagusi gehitu dizkiogu sistemari estaldura hobetzeko asmoz: hiztegien zabaltzea eta transliteraziorako modulua.

Lexikoiak edo hiztegiak zabaltzeko, McCarthy-ren *Suffix Prefix Dictionary* (McCarthy, 2016) hiztegia eta, garapenean sortutako zalantzazko kasuetan, *Mosby's Medical Dictionary* hiztegia erreferentzia hartuta hainbat aurrizki gehitu ditugu 5.2.2 atalean azaldutako zerrendetan. Modu horretan, aurreko lexikoiekin integratuta, 1.703 aurrizkitako eta 630 atzizkitako lexikoiak osatu ditugu. Bi lexikoiak eskuz sortu ditugu eta aditu batek errepasatu ditu, aurreko atalean azaldutako metodologia berdina jarraituta.

NeoTermen oinarri-lerroaren inplementazioan, medikuntzako termino neoklasiko erabat identifikatuak baino ezin genituen euskaratu. Bestela esanda, adibidez, sistema ez bada *path-* aurrizkia duten terminoak euskaratzeko kapaz, aurrizki hori lexikoian agertzen ez delako, *hypophosphatemia* bezalako terminoak ezin genituen euskaratu, nahiz eta *hypo-*, *phos-* eta *-emia* afixuak lexikoietan agertu. Transliterazioa erabiliz murriztapen hori gainditzea dugu helburu.

Gure lexikoietan agertzen diren aurrizkietatik % 93 transliterazio-erregelen bitartez euskara daitezke, hiztegietako baliokide berdina lortuz. Horrela, afixu bezala identifikatu gabeko zatiak transliteratzeko Fomako 40 erregela definitu ditugu. Erregela horiek sortzeko, oinarri-lerro sisteman definitutako euskararen morfofonologia arauak errespetatzeko erregelak transliteraziorako egokitu ditugu eta berri batzuk definitu ditugu ere. Berriak definitzeko, afixuen konposaketarako erabilitako metodologia berdina jarraitu dugu, eta erregela batzuk enpirikoki ondorioztatu ditugu, euskararen ortografia arauak betetzen dutela bermatuaz.

Adibidez, *v* letra, *b* letrarekin ordezkatzen dugu beti arau ortografikoei jarraituz, edo c letra *z*-rekin *e*, *i* edo *y* letra jarraian izanez gero, eta *k*-rekin gainontzekoetan, arau morfofonologikoek ezartzen duten moduan. Erregela horiek inplizituki hiztegietan askotan aurkitu ditugu, hala nola, *diverticulitis* terminoarekin eta haren euskal ordaina den dibertikulitis terminoarekin (ZT Hiztegia eta Erizaintza Hiztegia), *v* kontsonantea *b* bihurtzen da, eta *c* letra *k*, *u* bokala baitu jarraian. Erregela horietako batzuk erakusten ditugu adibideekin 5.9 taulan.

Hurbilpen honetan, afixuen identifikazioa bi urratsetan egiten dugu. Lehenengo urratsean identifikatzea lortzen ez bada, bigarren urratsean identifikatzen saiatuko da. Bigarren urratsean ere ezin bada identifikatu, terminoa

	en	eu	Adibidea	Erregela
1	^ y	i	yttrium -> itrio	y -> i || [.#.| ^] _ Kon
2	p h	f	pharyngostome -> faringostoma	p h -> f
3	r h	r r	rheostosis -> erreostosis	r h -> r r
4	c k	k	amsinckine -> antsiskina	c k -> K
5	r s	r t s	arsenic -> artseniko	r -> r t || _ Txis Bok
6	x s	s	isoxsuprine -> isosuprina	x -> 0 || _ s
7	c c	k z	occipital -> okzipital	c c -> k z || _ [e | i | y]
...	...	...	...	...

5.9 taula – Transliteraziorako hainbat erregela.

euskaratu gabe gelditzen da.

Aurreneko urratsa, oinarri-lerroaren identifikazio-erregela berdina da (5.5 irudiko 2. lerroa). Bigarren urratsean, identifikazio malguagoa definitu dugu, lexikoian aurkitzen ez diren zatiak ere identifikatuz. Horretako definituko dugun gutxieneko baldintza, terminoaren atzizkia lexikoian agertzea da (5.5 kodeko 3. lerroa). # ikurraren bitartez aurrizkien lexikoian agertzen ez diren zatiak markatuta uzten ditugu, itzulpenean aurrizki baliokidetza egin beharrean transliteraziorako erregelak erabil ditzan baliokideen transduktoreak.

```
1 ...
2 define IDEN1 [[[AURR 0:+] (o 0:+)]* ATZ] ;
3 define IDEN2 [(?+ 0:#+) [AURR 0:+]]* (?+ 0:#+) ATZ;
4 regex IDEN1 .P. IDEN2;
```

5.5 irudia – Afixuen identifikazioa lexikoitik kanpoko zatiekin.

Hurrengo irudian (5.6 irudia) *hypophosphatemia* terminoa NeoTermi esker nola euskaratzen dugun erakusten dugu urratsez urrats.

Jatorri-terminoa: *hypophosphatemia*
Identifikatutako afixuak: *hypo+phos+ph#+at+emia*
Afixu baliokideak: hipo+fos+f+at+emia
Euskal ordaina: hipofosfatemia

5.6 irudia – NeoTermen oinarri-sistemaren sorkuntza prozesua.

Lehenik, afixuak identifikatzen ditugu, identifikatu gabeko karaktere kopurua minimizatuz eta afixu kopuru minimoa aukeratuz (*hypo+phos+ph#+*

at+emia). Ostean, afixuen baliokideak jasotzen ditugu lexikoietatik eta identifikatu gabeko zatiak transliteratzen ditugu (hipo+fos+f+at+emia). Bukatzeko, afixuak konbinatzen ditugu (hipofosfatemia).

Anbiguotasuna mugatzeko irizpide berriak ere definitu behar izan ditugu, bigarren identifikaziorako erregelarako ez baita baliagarria aurretik definitutako irizpidea. Hurrengo adibidearen bidez ikusiko ditugu aurreko irizpidearen arazoak, eta proposatutako irtenbidea.

Diverticulitis terminoaren kasuan, *verticul* zatia ez da afixuen lexikoian agertzen. Ondorioz, bigarren urratsaren bidez identifikatu ditugu afixuak. Transduktoreak afixuen konbinaketa guztiak identifikatzen ditu, eta baita identifikatu gabeko zatien konbinaketak ere. Hurrengo zerrendan ikus daitekeen moduan, transduktoreak lexikoetan agertzen diren afixuak, identifikatu gabeko zatiekin konbinatuta ere proposatzen ditu. Identifikatutako zatien ondoren, zati kopurua eta identikatu gabeko zatiaren karaktere kopurua gehitu ditugu, aurrerago azalduko dugun moduan.

- *diverticul#+itis*: 2 + 10 = 12
- *divertic#+ul+itis*: 3 + 8 = 11
- *di+verticul#+itis*: 3 + 8 = 11
- *di+vertic#+ul+itis*: 4 + 6 = 10

Modu horretan, identifikatutako zati gutxien dituen aukera lehenengoa litzateke (*diverticul#+itis*), baina egokiagoa dirudi afixuen lexikoian agertzen diren zatiak bere aldetik euskaratzea, eta gainerako zatiak transliterazioaren bitartez. Hortaz, kontaketa berri bat gehitu diogu zatien kontaketari: identifikatu gabeko zatiaren karaktere kopurua. Datuen analisiak gidatuta, luzeran motzena den identifikatu gabeko zatia bilatzen dugu, eta aldi berean, lexikoietan agertzen diren afixu gutxien egotea ere bilatzen dugu (*di+vertic#+ul+itis*, kasu honetan).

Emaitzetan ikusiko dugun bezala (ikus 5.4 atala), NeoTermen hurbilpen honetan termino neoklasikoak identifikatzeko hartutako irizpide berriak estaldura asko hobetzen duen arren, positibo faltsu asko ere sortzen ditu. Horien kopurua murrizteko, hurrengo atalean azalduko dugun azken hurbilpena prestatu dugu, identifikaziorako irizpideak mugatuz.

5.2.4 NeoTerm: identifikazioa fintzeko irizpideak

Azken hurbilpenean, adituen gomendioei jarraiki, estaldura fintzeko asmoarekin terminoen identifikaziorako irizpideak mugatu ditugu. Irizpide berri horiekin, *-tion* edo *-able* bezalako atzizkiak dituzten terminoek, baldintza gehiago bete beharko dituzte NeoTermen bitartez euskaratuak izateko.

Horretarako, orain artean erabilitako aurrizki zein atzizkien lexikoien bertsio murriztu berriak sortu ditugu.

Atzizkien kasuan, hitz orokorretan (osasun-zientzien domeinurako espezifikoak ez diren hitzetan) erabiltzen diren atzizkiak baztertu ditugu (*"-tion"* edo *"-able"* moduko atzizkiak). Horretarako, Wiktionary-ren ingelesezko atzizkien hiztegia kontsultatu dugu (Wiktionary, 2014), eta bertan agertzen diren atzizkien esanahia aztertuta, definizio orokorra dutenak baztertu ditugu. Aukeraketa horretan, atzizki neoklasikoak ez direnak ere kendu ditugu (*-hood*, adibidez). Kontziente gara eskuz egindako prozedura horrek, erroreak sor ditzakeela, baina modu horretan, gehien erabiltzen diren atzizkiak detektatu ditugu.

Prozesu honetan, atzizkien bazterketaren inguruko zenbait erabaki hartu behar izan dugu. Adibidez, *-on* atzizkiak, hiru adiera ditu biologiarekin zein kimikarekin erlazioa dutenak, baina atzizkia bera, beste atzizki orokorren bukaera ere izan daiteke, *-tion* edo *-isation* atzizkiena, alegia. Bi atzizki horiek lexikoietatik baztertu ditugunez, eta atzizkien artean *-tion* atzizkia ohikoena denez, *-on* atzizkia ere baztertzea erabaki dugu.

Aurrizkiekin aldiz, ezin izan dugu prozesu bera erabili, eta karaktere kopuruaren arabera baztertu ditugu zenbait aurrizki (3 karaktere edo gutxiagokoak). Izan ere, gure irizpideen arabera aurrizkiak terminoaren edozein posiziotan ager daitezke, eta bi edo hiru karaktereko segida batek ez du zertan aurrizkia izan. Horrela, *an-*, *col-* edo *cyt-* bezalako aurrizkiak baztertu ditugu.

Arrazoi horretaz gain, eraginkortasun arrazoiak ere erabili ahalko genituzke, 1.703 aurrizki errepasatu beharko baikenituzke eskuz, aurretik aztertutako 630 atzizkiez gain.

Bazterketa prozesuaren ondorioz, 241 aurrizki eta 71 atzizki kendu ditugu lexikoietatik, eta horrela 1.462 aurrizkidun eta 559 atzizkidun lexikoi murriztuak osatu ditugu.

Hurrengo zerrendan, urratsez-urrats azalduko ditugu azken hurbilpen honetarako definitu ditugun irizpideak. Lehenengo irizpidearen bitartez terminoaren zatiak ezin baditugu identifikatu, bigarrenarekin saiatuko gara, eta

honekin ere ezin izanez gero, hirugarrenarekin.

1. Terminoaren afixu guztiak lexikoi hedatuetan agertzen diren afixuekin identifikatzen dira (8. lerroa 5.7 irudian). Irizpide hau aurreko bi hurbilpenetan ere erabilitakoa izan da (5.2.2 eta 5.2.3).

2. Gutxienez, terminoaren atzizkia atzizkien lexikoi murriztuan agertzen da (9. lerroa 5.7 irudian).

3. Gutxienez, terminoaren atzizkia atzizkien lexikoi hedatuan agertzen da eta gutxienez aurrizki bat dauka aurrizkien lexikoi murriztuan (10. lerroa 5.7 irudian).

Irizpide horiekin transduktorea definitzeko Foman idatzitako kodea ikus dezakegu 5.7 irudian.

```
1  ...
2  read lexc aurrizkiMugatua.lex
3  define AURRMUGATUA
4  define AURRMUG AURRMUGATUA.u ;
5  read lexc atzizkiMugatua.lex
6  define ATZMUGATUA
7  define ATZMUG ATZMUGATUA.u ;
8  define IDEN1 [[[AURR 0:%+] (o 0:+)]* ATZ] ;
9  define IDEN2 [(?+ 0:#+) [AURR 0:+]]*(?+ 0:#+) ATZMUG;
10 define IDEN3 [(?+ 0:#+) [AURRMUG 0:+]]+(?+ 0:#+) ATZ;
11 regex IDEN1 .P. IDEN2 .P. IDEN3;
```

5.7 irudia – Azken hurbilpenaren afixuen identifikaziorako erregelak.

NeoTermen hiru hurbilpenak azalduta, jarraian baliabide lexikalen parekatzearen emaitzak eta NeoTermen emaitzak aztertzeko diseinatu dugun ebaluazioa azalduko dugu.

5.3 Ebaluazioaren diseinua

Kapitulu honetan azaldutako algoritmoaren lehenengo bi urratsak ebaluatzeko, bi ebaluazio mota egin ditugu. Lehenengoak, ebaluazio automatikoak, tekniken garapenean lagundu digu, lanaren estalduraren eta doitasunaren estimazioa ematen digulako. Bigarrena, adituek egindako ebaluazioa, doitasuna neurtzeko baliagarria izan zaigu.

5.3.1 Ebaluazio automatikoa

Ebaluazio automatikoan SNOMED CTren nahasmenduen (*Disorder*), aurkikuntzen (*Finding*), gorputz-egituren (*Body structure*) eta prozeduren (*Procedure*) hierarkien euskaratzeari buruzko estaldura datuak emango ditugu, modu automatikoan erauzi ditugunak.

Alde batetik, urrats bakoitzaren ekarpena ebaluatuko dugu, lortutako ordain kopurua eta parekatzeak adieraziz. Ordain kopuruarekin euskaraz sortutako terminoak kontatuko ditugu, eta parekatze kopuruarekin, ingelesezko zein espainierazko zenbat terminoren parekatzea lortu dugun adieraziko dugu. Horretaz gain, baliabide lexikalen kasuan, hiztegi bakoitzak egindako ekarpena neurtuko dugu indibidualki, eta jatorri-hizkuntzaren araberako datuak ere jasoko ditugu. Bukatzeko, token kopuruaren araberako emaitzak emango ditugu, baita hierarkia bakoitzerako zenbat kontzeptu euskaratzea lortu dugun ere. Datu horiek guztiak 5.4 atalean aztertuko ditugu sakonki.

NeoTerm sistemaren hurbilpenak ebaluatzeko, urre-patroi gisa hiztegietatik erauzitako ingelesa-euskara pare zuzenak erabili ditugu. Urre-patroia garatzeko eta estaldura zehaztu ahal izateko, pare bakoitzari etiketa bat esleitu diogu terminoa neoklasikoa den ala ez adierazteko. Izan ere, NeoTerm sistema termino neoklasikoak euskaratzeko diseinatua izan da, eta gainerako terminoak ez lituzke euskaratu behar, *dengue* edo *shock* bezalalakoak, adibidez. Eskuzko etiketatze hori, ebaluaziorako urre-patroian bakarrik egin dugu, NeoTerm sistemaren garapenean eskuzko lan handia egin behar izan dugulako, eta baliabideak optimizatu behar izan ditugulako.

Ebaluazioari dagokionean, 2.245 termino erabili dira, horietatik 848 nahasmenduak dira, 375 aurkikuntzak, 774 gorputz-egiturak eta 248 prozedurak.

5.3.2 Adituen ebaluazioa

Adituen ebaluazioan ere, SNOMED CTren lau hierarkia nagusiak erabili ditugu: nahasmenduak, aurkikuntzak, gorputz-egiturak eta prozedurak. Denera 370 kontzeptu eta horiek deskribatzeko erabiltzen diren terminoak aukeratu ditugu. Kopuru hori ez dugu edozein modutan aukeratu, bereziki kalkulatu dugu lagina esanguratsua izan dadin. Horretarako, Franklin eta Agresti (2007) liburuko *How Do We Choose the Sample Size for a Study?* ("Nola aukeratzen dugu azterketa baterako laginaren tamaina?") atalean azaldutako populazioaren proportzioa balioetsiz laginaren tamaina kalkulatzeko formula erabili dugu (5.1 formula).

$$n = \frac{\hat{p}(1-\hat{p})z^2}{m^2} \tag{5.1}$$

Horrela, n, zorizko laginaren tamaina, p populazioaren proportzioaren eta m errore-marjinaren araberakoa izango da. z puntuazioa konfiantza mailan dago oinarrituta, adibidez guk % 95eko konfiantza maila ezarri dugu eta hortaz z = 1,96 izango da eta m = 0,04, banaketa normalari jarraitzen badiogu. Populazioaren proportzioa kalkulatzeko, ebaluazio automatikoan jasotako doitasunaren balioa erabili dugu. Horrela, $\hat{p}$-ren balioa 0,81 da 5.4 atalean ikusiko dugun bezala. Balio horiek kontuan izanda, 5.1 formularen arabera, lagina 370 kontzeptutakoa izatea erabaki dugu.

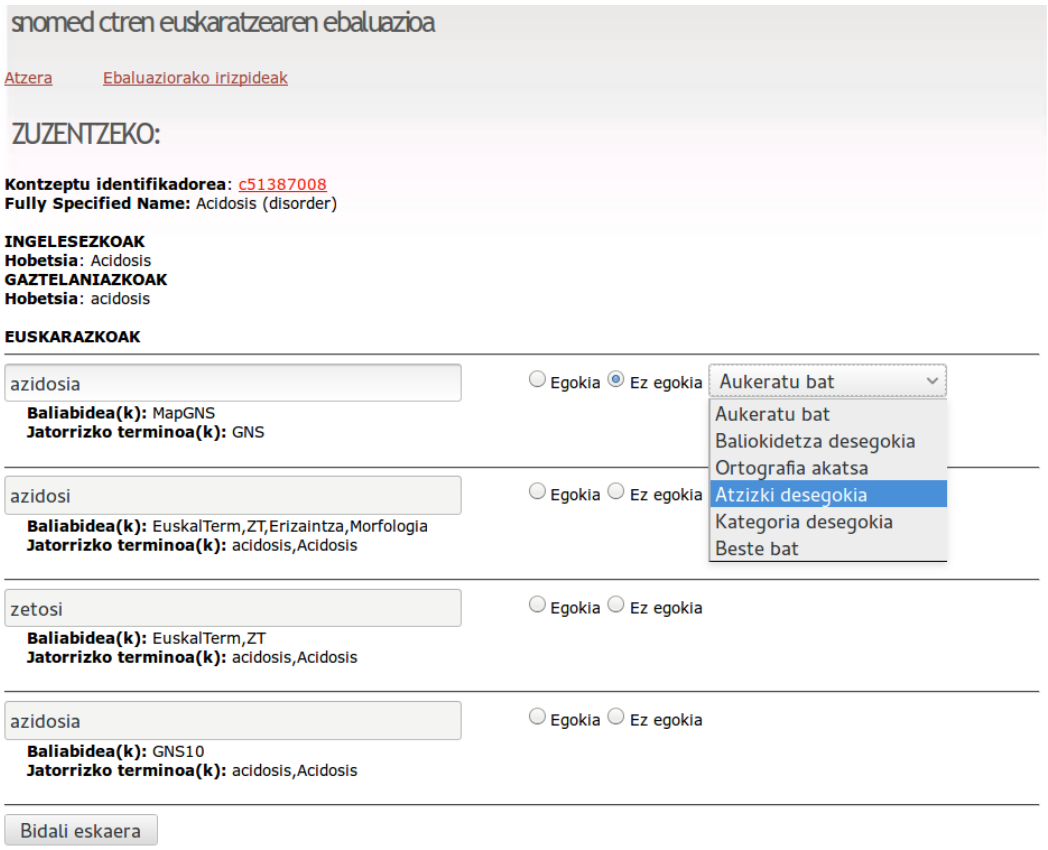
Lagina zoriz aukeratu dugu, euskarazko ordaina jaso duten termino sinple guztien artean (lau hierarkietakoak, beraien arteko proportzio natura-la mantenduta). Hala ere, laginak bete beharreko minimo batzuk zehaztu ditugu, aurretik euskarazko ordainak lortu ditugularik: gutxienez 200 kontzeptuk NeoTerm sistemak emandako ordainen bat izatea, eta gutxienez 100 kontzeptuk baliabide lexikalen ordainen bat izatea. Hau da, 370 kontzeptu landuko ditugu eta horiek deskribatzeko erabilitako terminoak ebaluatuko ditugu, termino horietako ordainak ditugunean.

Ebaluazioan, bi hizkuntzalarik eta bi medikuk parte hartu dute. Lagina, alde batetik hizkuntzalariek ebaluatu dute, eta ostean, lagin berdina medikuek. Horrela, 170 kontzeptu hartu ditugu zoriz ebaluatzaileen arteko adostasuna kalkulatu ahal izateko, eta gainerako 200 kontzeptuak, erdibana banatu ditugu. Horrela, ebaluatzaile bakoitzak euskarazko ordain bat edo hainbat dituzten 270 kontzeptu ebaluatu ditu.

Ebaluatzaileen lana errazteko, interfaze bat prestatu dugu. Interfazean, kontzeptu mailan esanguratsua izan daitekeen informazio guztia erakusten dugu, ebaluatzaileek erabakiak modu erosoan har ditzaten.

Acidosis (disorder) kontzeptuaren ebaluazio-orria 5.8 irudian erakusten dugu. Bertan, kontzeptuaren identifikadorea, SNOMED CTren nabigatzaile ofizialarekin estekatu dugu[7], arakatzaile horretan kontzeptuaren informazio gehigarria eskura daitekeelako (kontzeptuaren egitura hierarkikoa eta hortaz bere guraso eta umeak, adibidez). Identifikadorearekin batera, kontzeptuaren *Fully Specified Name*-a (FSNa) ere agertzen da, kontzeptuaren deskribapen argi eta anbiguotasun gabea izanik, ordainen aukeraketan lagungarria izango

[7]http://browser.ihtsdotools.org/ helbidean topa daiteke nabigatzaile ofiziala (2017ko maiatzaren 9an atzitua).



5.8 irudia – Adituen ebaluaziorako inplementatutako interfazea.

delakoan. Kontzeptu mailako informazioaz gain, ingelesezko zein gaztelaniazko ordain guztiak ere erakusten ditugu, hobetsitako terminoa nabarmenduz.

Bukatzeko, euskarazko ordainak agertzen dira, banan-banan, erabilitako baliabideak adieraziz, baita jatorrizko terminoa zein den (edo zeintzuk diren) adieraziz ere. Ordain bakoitza egokia den edo ez-egokia den erabaki behar dute ebaluatzaileek, eta ez-egokia den kasuetarako, arrazoi bat zehaztu behar dute: “baliokidetza desegokia”, “ortografia akatsa”, “atzizki desegokia”, “kategoria desegokia” edota “beste bat”.

Hurrengo atalean, ebaluazio horien emaitzak aztertuko ditugu.

5.4 Emaitzak

Atal honetan aplikazioaren lehenengo urratsetan lortutako emaitzak aurkeztuko ditugu, baliabide lexikalak erabilita eta NeoTerm erabilita, alegia. Alde batetik, bi urratsen ebaluazio automatikoak erakutsiko ditugu 5.4.1 atalean, eta bestetik, adituen ebaluazioan lortutako emaitzak 5.4.2 atalean.

5.4.1 Ebaluazio automatikoaren emaitzak

Jarraian, automatikoki lortu ditugun emaitzak erakutsiko ditugu. Lehenik eta behin, NeoTerm sistemaren hurbilpen ezberdinen ebaluazioak emango ditugu, sistemaren eboluzioa ikusi ahal izateko. Ostean, SNOMED CTren euskaratzearen estaldurari dagozkion datuak emango ditugu, lau hierarkietarako zenbat kontzeptu eta termino euskaratzeko gai izan garen agerian utziz eta ordainak lortzeko erabilitako baliabide emankorrena zein izan den aztertzeko. Baliabide lexikalen erabilerari buruzko datuak, beraz, estaldurraren datuekin batera emango ditugu, hiztegien doitasuna modu automatikoan neurtzea ezinezkoa zaigulako.

NeoTermen emaitza automatikoak

NeoTerm sistemaren hiru hurbilpenen emaitza automatikoak 5.10 taulan erakusten ditugu. Taulan, egiazko positibo (**E**giazko **P**ositibo), faltsu negatibo (**F**altsu **N**egatibo), faltsu positibo (**F**altsu **P**ositibo) eta egiazko negatibo (**E**giazko **N**egatibo) kopuruak ematen ditugu, doitasuna eta estaldura kalkulatzeko erabili ditugunak. Horretaz gain, F-neurria ere kalkulatu dugu, 5.2 formula erabilita. Kalkulu guztiak termino mailan egin ditugu.

$$F = 2 * \frac{doitasuna * estaldura}{doitasuna + estaldura} \tag{5.2}$$

- Egiazko positiboa (EP): Termino neoklasikoa da, eta sortutako ordaina zuzena da. Zuzena den zehazteko, ordainetako bakarra urre-patroian agertzea nahikoa da.

- Faltsu negatiboa (FN): Termino neoklasikoa da, baina NeoTermek ez du ordainik sortu.

- Faltsu positiboa (FP): Ez da termino neoklasikoa, baina NeoTermek ordaina sortu du, edo terminoa neoklasikoa da, baina ordaina ez da zuzena.
- Egiazko negatiboa (EN): Ez da termino neoklasikoa, eta NeoTermek ez du ordainik sortu.

		EP	FN	FP	EN	Den.	Doi.	Est.	F
Nahasmendu	**Oinarri-lerroa**	289	451	31	77	848	**0,903**	0,391	0,545
	Transliterazioa	615	67	108	58	848	0,851	0,902	**0,875**
	Irizpideak	577	104	102	65	848	0,850	0,847	0,849
Aurkikuntza	**Oinarri-lerroa**	79	171	9	116	375	**0,898**	0,316	0,467
	Transliterazioa	213	29	41	92	375	0,839	0,880	**0,859**
	Irizpideak	178	63	32	102	375	0,848	0,739	0,789
Gorputz-egitura	**Oinarri-lerroa**	121	425	23	205	774	**0,840**	0,222	0,351
	Transliterazioa	322	174	100	178	774	0,763	0,649	**0,702**
	Irizpideak	284	212	91	187	774	0,757	0,573	0,652
Prozedura	**Oinarri-lerroa**	98	77	9	64	248	**0,916**	0,560	0,695
	Transliterazioa	144	16	49	39	248	0,746	0,900	**0,816**
	Irizpideak	140	20	46	42	248	0,753	0,875	0,809
Denera	**Oinarri-lerroa**	587	1.124	72	462	2.245	**0,891**	0,343	0,495
	Transliterazioa	1.295	286	297	367	2.245	0,813	0,819	**0,816**
	Irizpideak	1.179	399	271	396	2.245	0,813	0,747	0,779

5.10 taula – NeoTermen hiru hurbilpenen ebaluazio automatikoa.

Oinarri-lerro sistema da doitasun altuena daukan sistema hierarkia guztietarako, eta nahasmenduen kasuan 0,90 baino altuagoa izatera heltzen da.

Transliterazio moduluaren integrazioarekin, doitasuna 0,08 jaitsi da orokorrean, baina estalduraren kasuan 0,47 irabazi dugu, F-neurrian 0,32ko hobekuntza eraginez. NeoTermen hurbilpen horrekin, oso emaitza onak jaso ditugu, eta tamalez, identifikazioa fintzeko irizpideekin ez dugu gorako joera hori mantendu. Ebaluazio automatikoan erabilitako laginak ez du hobekuntza adierazi, baina baliteke beste lagin batekin edo eskuzko ebaluazioaren ondorioz hobekuntza ikustea, ezarritako irizpideek zentzuzkoak ematen baitute. Hala ere, eskuartean ditugun datuek ez dute hobekuntzarik erakutsi.

Hierarkia batetik bestera, aldeak nabari dira azkeneko hurbilpenari dagokionean: doitasuna pixka bat igotzen da zenbaitetan, eta beste batzuetan jaitsi. Dena dela, orokorrean, aldea arbuiagarria da, eta guztien baturan ez da doitasunean alderik nabaritu. Estaldurari dagokionean, aldiz, emaitzak okertu dira. Estaldura fintzeko irizpideekin, faltsu positiboak gutxitzea zen helburua, eta helburua lortu den arren (26 gutxiago daude), egiazko positiboak nabarmen jaitsi dira (116 gutxiago) eta horrela, estaldura 0,07 jaitsi da.

Emaitza horiek kontuan izanik, identifikazioa fintzeko irizpideek ez diote onura nabarmenik ekarri NeoTerm sistemari. Ondorioz, transliterazio moduluaren hurbilpena da SNOMED CTren ordainak sortzeko erabiliko duguna.

Gogoan hartu behar dugu, ebaluazio honek ez duela kontuan hartu sistemen gainsorkuntza. Beti ere, sortutako ordainen artean zuzena agertzen bada, zuzentzat hartu da sortutakoa. Dena dela, gainsorkuntza oso txikia da hiru hurbilpenetan, eta batazbestean, jatorri termino bakoitzarentzat 1,05 euskal ordain sortzen dira, gainsorkuntza eza izatetik (1etik) oso hurbil.

Ondorengo lerroetan, NeoTerm sistemak egin dituen errore batzuk erakusten ditugu, guri interesgarriak iruditu zaizkigun pare bat fenomeno islatzen dituztenak.

- Zenbakietan nabari ez bada ere, ebaluazio automatikoan erabili dugun urre-patroian hutsuneak aurkitu ditugu. Izan ere, zenbait kasutan ingeles-euskara termino-ordain pareetan, euskarazko ordaina ez dagokio termino neoklasikoari, beste sinonimo bati baizik. Adibidez, *stenocardia* nahasmendurako, hiztegietatik lortzen dugun ordaina "bularreko angina" da, baina horrek ez du esan nahi NeoTermek sortzen duen "estenokardia" okerra denik.

- Horretaz gain, hainbat kasu identifikatu ditugu salbuespen kontsidera daitezkeenak ingeleseko terminoari erreparatzen badiogu. Hau da, Euskal Herriaren hegoaldean, euskararen bizilaguna espainiera da, eta domeinu askotan bezala, medikuntzan ere, termino espezializatuak espainieratik mailegatu dira, eta ez ingelesetik. Horren adibidea da fenilzetonuria terminoaren kasua, ingelesez *phenylketonuria* deritzo, eta hiru afixuk osatzen dute: *phenyl+keton+uria*. Aldaera *keton* afixuan dago. Afixu honen ahoskeran, ingelesez */k/* bada ere, espainieraz, *ceton* afixua erabiltzen da, */z/* fonema duena. Horrela, euskarara ekartzean, afixuaren euskarazko bertsioa "zeton" izan beharko litzateke. Kasu honetan, aurrizki hori ez da gure lexikoian agertzen, eta transliterazio modulua aplikatzean, fenilketonuria euskarazko ordain okerra sortzen dugu.

Azken hurbilpenaren hobekuntza faltari dagokionean, ez dugu ageriko arrazoirik aurkitu ebaluazio testean. Baliteke, adituek egindako ebaluazioan emaitza ezberdinak lortzea, eta hobekuntzarik aurkitzea. Hala ere, ebaluazio honetan, transliterazio modulua bakarrik integratuta duen sistema da emaitza onenak ematen dituena.

SNOMED CTren euskaratze-kopuruak eta estaldura

SNOMED CTren euskaratze-estaldura neurtzeari ekingo diogu jarraian. Baliabide lexikalen ekarpena erakusten dugu 5.11 taulan, jatorri-hizkuntz gisa erabili ditugun ingelesarena eta espainierarena, hain zuzen ere. Zutabeen goiburuei erreparatzen badiegu, "Ord." zutabeetan euskaraz lortutako ordain kopurua adierazten dugu, "Eus." zutabeetan euskaratu ditugun termino kopurua eta "Est" zutabeetan euskaratutako terminoen estaldura.

	Ingelesa			**Espainiera**			**Denera**		
	Ord.	Eus.	Est.	Ord.	Eus.	Est.	Ord.	Eus.	Est.
Nahasmenduak	6.975	5.750	0,050	4.760	3.606	0,040	7.762	9.356	0,045
Aurkikuntzak	2.343	1.635	0,031	2.531	1.340	0,028	3.587	2.975	0,029
Gorputz-egiturak	6.599	4.667	0,054	6.061	3.593	0,064	8.552	8.260	0,058
Prozedurak	1.007	981	0,013	985	590	0,080	1.425	1.571	0,100

5.11 taula – Baliabide lexikalen banakako ekarpena ordain kopuruari dagokionez.

Nahasmenduen hierarkian ingelesaren ekarpena nabarmenagoa da, baita gorputz-egituren hierarkian ere, baina gainerako bi hierarkietan ekarpen antzekoa egiten dute. Kontuan izanda baliabide lexikal berdinak erabili ditugula bi hizkuntzetarako, azpimarratzekoa da ordain kopuruan lortutako ekarpena bi hizkuntzen aldetik. Izan ere, denera lortu ditugun euskarazko ordainak bi hizkuntzen ekarpena gehituaz, hizkuntza bakoitzaren ekarpena baino dezente handiago da.

Hierarkietako ordain kopuruei eta euskaratutako termino kopuruei dagozkion emaitza orokorrak 5.12 taulan ematen ditugu, erabilitako baliabideen arabera sailkatuta. Bertan, baliabide bakoitzarekin lortutako ordainak aurkezteaz gain, ordainak zenbat terminoren bidez lortu ditugun ere erakusten dugu. Horrela, "Ord." zutabeekin ordain kopurua adierazten dugu, eta "Eus." zutabeekin zenbat jatorri-termino euskaratu ditugun.

Ikus dezakegunez, SNOMED CTren eta GNS10ren arteko mapaketa ere gehitu dugu taula horretan, gure sistemaren estalduraren orokortasuna jaso dezagun. Horrela, mapaketa, baliabide lexikalak eta NeoTerm sistemaren estaldura ikus dezagu, eta baita teknika horiekin lortutako estaldura orokorra ere.

NeoTermen ekarpenak mugatua ematen badu ere, kontuan izan termino sinpleak euskaratzeko balio duela bakarrik, eta SNOMED CTn termino

	GNS10 mapaketa		**Baliabide lexikalak**		**NeoTerm**		**Denera**	
	Ord.	Eus.	Ord.	Eus.	Ord.	Eus.	Ord.	Eus.
Nahasmenduak	11.060	-	7.593	9.356	1.951	1.868	20.375	11.290
Aurkikuntzak	2.555	-	3.529	2.975	717	622	6.694	3.619
Gorputz-egiturak	0	-	8.420	8.260	726	676	9.089	8.977
Prozedurak	0	-	1.320	1.337	1.308	1.266	2.561	2.613

5.12 taula – SNOMED CTren euskaratutako termino zein ordainen kopuruak.

sinpleen agerpenekin alderatuz, ekarpena ez da batere mugatua 5.14 taulan ikusiko dugun bezala.

Mapaketa oso emankorra da nahasmenduak euskaratzeko garaian, 11.060 kontzepturen ordainak lortu baitira. Kontuan izan, mapaketa hori kontzeptu mailakoa dela, eta kontzeptu bati euskarazko ordain bakar bat esleitzen zaio.

Baliabide lexikalak izan dira orokorrean ekarpen handiena egin dutenak, bai euskaratutako termino kopuruari dagokionean, bai eta lortutako ordainei dagokionean. Kontuan izan behar da, kasu horretan ingelesezko zein espainierazko terminoen ordainak lortu ditugula.

	Nahasmenduak		**Aurkikuntzak**		**Gorputz--egiturak**		**Prozedurak**	
	Ord.	Bak.	Ord.	Bak.	Ord.	Bak.	Ord.	Bak.
ZT Hiztegia	975	269	401	98	1.754	437	294	106
Euskalterm	5.259	3.945	1.654	1.207	3.932	2.357	729	559
Anatomia	13	12	5	4	3.571	3.120	3	2
Erizaintza	487	103	480	179	1.236	271	235	89
GNS10	2.512	1.730	403	278	425	179	9	9
AdminSan	12	6	12	6	11	6	70	28
Elhuyar	142	136	1.280	1.253	312	296	309	303

5.13 taula – Hiztegien banakako ekarpena ordain kopuruari dagokionez.

Baliabide lexikalen ekarpena zehatzago aztertzeko 5.13 taula ikus dezakegu. Bertan, ordainen jatorriaren banakapena erakusten dugu, hiztegi bakoitzak zenbat ordain eman dituen (“Ord.” zutabeak) eta berak bakarrik horietako zenbat eman dituen erakusten dugu (hizkuntza kontuan izan gabe, “Bak.” zutabeetan). Aipatu nahi dugu, jatorrizko termino baterako ordain

bakarra lortu arren, hori baliabide lexikal ugarietatik errepikatuta lor dezakegula. Adibidez, *eye* terminoarentzat begi ordaina hainbat hiztegietatik lortuko dugu. Horrela, aurreko 5.12 taulako baliabide lexikalen emaitzak ez dira 5.13 taularen batura. Taulak, terminoen "espezializazio maila" edo "ezagutza mailaren" erreferentzia eman ahal digu. Adibidez, begi hitza arrunta izanagatik hainbat iturritatik jasotzen dugu, baina endosteo[8] bezalako termino espezializatua Anatomiako Atlasean baino ez dugu aurkitu.

Adibidez, ZT Hiztegia emankorra dela ematen badu ere, ikus dezakegunez, eskaintzen dituen ordain gehienak gainerako hiztegietan ere agertzen dira. Aldiz, Giza Anatomiako Atlasak, proposatzen dituen ordainak bakarrak dira orokorrean. Honek adierazten digu, Giza Anatomiako Atlasak egiten duen ekarpena oso baliagarria zaigula ordain berriak lortzeko garaian, bereziki, zentzuzkoa den moduan, gorputz-egituren hierarkian.

Estalduraren beste ikuspegi bat ikus dezakegu 5.14 taulan. Bertan, jatorrizko ingelesezko terminoen token kopuruaren arabera ematen ditugu zenbakiak. Horrela, hierarkia bakoitzean, zenbat terminoren ordainak lortu diren ("Eusk." lerroak) erakusten dugu, token kopuruaren arabera. Gainera, denera hierarkia horretan zenbat termino dauden ingelesezko bertsioan ("Den." lerroak) eta euskaratutakoen estaldura ("Estal." lerroak) ematen ditugu. Taulak erakusten digunez, token bakarreko termino gehienak euskaratu baditugu ere (nahasmenduen % 84,5a, aurkikuntzen % 74,7a, gorputz-egituren % 73,6a eta prozeduren %85,5a), oso termino konplexu gutxi euskaratzeko gai izan gara.

Estaldurari dagozkion emaitza orokorrekin bukatzeko, zenbat kontzepturen ordainak lortu ditugun zenbatu dugu. Baliabide lexikalak eta NeoTerm erabilita automatikoki euskaratutako kontzeptuen estaldura erakusten dugu 5.15 taulan. Ikus dezakegunez, bai nahasmenduetan eta bai gorputz-egituretan euskaratutako kontzeptuen portzentajea altua da (% 17tik gorakoa), orain arte garatutako sistemen murriztapena kontuan hartzen badugu. Prozeduren hierarkiari dagokionean, euskaratutako kontzeptuen portzentajea oso baxua da. 5.14 taula gogora ekartzen badugu, gogoratu behar dugu termino sinpleen ia % 90a itzultzeko gai izan garen arren, prozeduren hierarkian termino sinpleen populazioa oso txikia da, eta horrela uler daiteke osotasunean lortutako estaldura baxua.

Hala ere, ezin dugu ahaztu, nahasmenduen hierarkian GNS10 eta SNOMED CTren artean egindako mapaketak egin duen ekarpena (11.000 kon-

[8]https://en.wikipedia.org/wiki/Endosteum (2017ko maiatzaren 9an atzitua).

		token1	2token	3token	4token	≥5token	Denera
Nahas-menduak	**Eusk.**	3.265	2.280	1.252	433	454	7.684
	Den.	3.865	21.003	25.038	20.757	44.167	114.830
	Estal.	0,845	0,109	0,050	0,021	0,010	0,067
Aurki-kuntzak	**Eusk.**	1.449	529	142	60	99	2.279
	Den.	1.940	9.737	11.906	11.317	24.640	59.540
	Estal.	0,747	0,054	0,012	0,005	0,004	0,038
Gorputz--egiturak	**Eusk.**	1.907	2.167	1.024	239	47	5.384
	Den.	2.592	10.863	12.599	10.635	22.695	59.384
	Estal.	0,736	0,200	0,081	0,023	0,002	0,091
Proze-durak	**Eusk.**	1.698	227	48	35	15	2.023
	Den.	1.985	9.892	15.399	17.082	42.746	87.104
	Estal.	0,855	0,023	0,003	0,002	0,001	0,023

5.14 taula – Jatorrizko ingelesezko terminoen token kopuruaren araberako emaitzak terminoak kontuan izanik.

	Nahasmenduak	Aurkikuntzak	Gorputz--egiturak	Prozedurak
Euskaratuak	15.163	4.079	5.432	1.799
Denera	68.815	37.888	30.871	55.128
Estaldura	0,220	0,108	0,176	0,033

5.15 taula – Emaitza orokorrak, GNS10ekin mapaketa, baliabide lexikalak eta NeoTerm erabilita, kontzeptuak kontuan izanik.

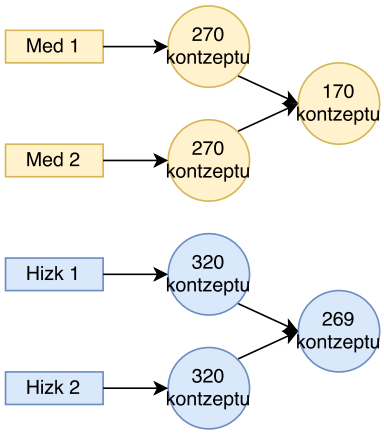
tzeptutik gora), eta Giza Anatomiako Atlasak gorputz-egituretan egin duena (3.000 ordain berri baino gehiago).

Zenbaki horiek ahaztu gabe, hurrengo atalean adituek egindako ebaluazioaren emaitzak aztertuko ditugu, estaldurari dagozkion datuak doitasunarekin osatzeko.

5.4.2 Adituen ebaluazioaren emaitzak

Atal honetan, adituen ebaluazioaren emaitzak erakutsiko ditugu. Ebaluazioaren diseinuan aipatu dugun bezala (5.3 atala), ebaluatzaile bakoitzari 270 kontzeptuz osatutako lagina eman diegu, denera 370 kontzeptuko lagina ebaluatu ahal izateko adostasun neurriak ere kontuan hartuz. Hala ere, hizkuntzalarien arteko ebaluazioan, laginketan arazo bat gertatu da, eta emandako lagina berdina izan da. Horrela, ebaluatzaile bakoitzari 50 kontzeptu

gehigarri eman behar izan dizkiegu laginaren osotasuna ebaluatu ahal izateko. Hortaz, hizkuntzalarien kasuan 269 kontzeptu dituzte komunean eta medikuek 170 6.2 irudian erakusten dugun moduan. Guztien arteko adostasuna neurtzeko lagina ere 170 kontzeptuetakoa da.



5.9 irudia – Adituei banatutako laginaren tamaina.

Ebaluatzaileen arteko adostasun neurriekin hasiko dugu emaitzen atal hau. Adostasun hori neurtzeko neurri asko daude, baina ohikoena kappa (κ) neurria da. Bi ebaluatzaile diren kasuetan Cohenen *kappa* (Cohen, 1960) erabiltzen da, eta bi baino gehiago direnean Fleissena (Gwet, 2014, Artstein eta Poesio, 2008). Izan ere, adostasunen portzentajeak ez du auzakotasuna kontuan hartzen eta hori bera da *kappa*k egiten duena. Horrela, adostasunak kontuan izateaz gain, ebaluatzaile bakoitzak zoriz ebaluatuko balu, ados egoteko proportzioa ere kontuan hartzen du neurri horrek.

Kappa balioaren interpretazioa konplexua den arren, zabaldutako tarteen interpretazio bat dago (Landis eta Koch, 1977), non 0-0,2 balioen artekoa arina den, 0,2-0,4 artekoa dezentekoa, 0,4-0,6 artekoa neurrizkoa, 0,6-0,8 artekoa sendoa eta 0,8-1,0 balioen artekoa ia perfektua.

Kappa neurriak termino mailako adostasunarekin atera ditugu. Kontzeptuen *kappa* ateratzeko arazoak ditugu, kontzeptu bakoitzean hainbat ebaluaziok parte hartzen dutelako, eta batetik bestera kopurua oso aldakorra delako. Kasu batzuetan ordain bakarrak osatzen du kontzeptua, eta hortaz, ebaluazio bakarrean adostasuna behar dugu. Beste kasu batzuetan aldiz, ordain gehiago daude, eta ondorioz, ebaluazio gehiagotan ados egon behar dute ebaluatzaileek. Hori guztia kontuan hartuta, ezin dugu kontzeptu ba-

tean orokorrean, zoriz zenbatetan ados egongo diren kalkulatu, aldakorra baita kopurua.

Aditu talde bakoitzaren barneko *kappa* neurria kalkulatu dugu 5.16 taulan, ebaluatzaile guztien adostasunarekin batera. Bi ebaluatzaileren adostasuna neurtzeko Cohenen *kappa* kalkulatu dugu, eta ebaluatzaile guztiena kalkulatzeko Fleissena. Bertan ikus dezakegunez, ebaluatzaile guztien arteko *kappa* balioa 0,56koa da, hizkuntzalarien artekoa 0,64koa eta medikuen artekoa 0,67; 11 eta 12 puntu altuagoa hurrenez hurren. Horrela, guztien arteko adostasuna neurrizkoa bada ere, aditu talde bakoitzaren adostasuna sendoa da.

	Hizkuntzalariak		**Medikuak**		**Denak**	
	Kontz.	**Termino**	**Kontz.**	**Termino**	**Kontz.**	**Termino**
Denera	269	566	170	346	170	346
Berdin	195	487	124	305	95	237
Ezberdin	74	79	46	41	75	109
Adostasuna	% 72,49	% 86,04	% 72,94	% 88,15	% 55,88	% 68,50
Kappa		0,64		0,67		0,56

5.16 taula – Ebaluatzaileen arteko adostasuna.

Lortutako kappa balio guztiak, Workshop of Machine Translation (WMT) kanpainetan egin izan diren balioen gainetik daude (Bojar *et al.*, 2014). Kappa neurriaren esanahia lausoa da, eta kontuz ibili behar gara honen interpretazioa egiterakoan. Are gehiago, ataza ezberdinetako emaitzak konparatzen ditugunean, sistema ezberdinak, ebaluazio multzo ezberdinak eta ebaluazio-metodoak ezberdinak dituztelako. Hala ere, gure kappa balioak ikerketa honetarako onargarriak dira Landis eta Koch-en (1977) irizpideak jarraituz.

Bi aditu taldeen artean aldea nabaria denez, emaitza orokorrak emateaz gain, aditu taldeen arabera ere emango ditugu. Horrela, 5.17 taulan, ordainen ebaluazioa erakusten dugu, sorkuntza metodoaren arabera sailkatuta. Hau da, SNOMED CTren eta GNS10 arteko mapaketaren bidez lortutako terminoen, baliabide lexikalen bidez lortutako ordainen eta NeoTerm sistemak sortutako ordainen ebaluazioa erakusten dugu.

Emaitzak kalkulatzeko, ebaluazio komunetan bozketa bidez egin dugu kontaketa. Egoki bozkatu dutenak zenbatu ditugu, eta ez-egoki bozkatutakoak, eta bozka gehien jaso dituen aukera da irabazlea. Desadostasun garbia dagoen kasuetan (batek egoki eta besteak ez-egoki) ezin izan dugunez eraba-

kirik, “berdinketa” lerroan zenbatu ditugu eta ez ditugu doitasuna kalkulatzeko erabili. Termino ez-komunen kasuan, ebaluatzaile bakarrak emandako iritzia izan dugu kontuan. Emaitza orokorren kasuan, termino guztiek izan dute ebaluatzaile bat baino gehiago, eta termino bakoitzaren ebaluazioan bi, hiru edo lau ebaluatzailek parte hartu dute. Kasu horretan ere, metodo berdina erabili dugu terminoaren ebaluazioa aukeratzeko. Ezin dugu ahaztu hizkuntzalariek termino gehiago ebaluatu dituztela, eta hortaz, emaitza orokorretan haien eragina nabarmenagoa dela.

	GNS10 mapaketa			**Baliabide lexikalak**			**NeoTerm**		
	Hizk	**Med**	**Den**	**Hizk**	**Med**	**Den**	**Hizk**	**Med**	**Den**
Denera	60	60	60	406	406	406	400	400	400
Egoki	12	35	18	237	299	265	324	319	324
	(% 20)	(% 58)	(% 30)	(% 58)	(% 74)	(% 65)	(% 81)	(% 80)	(% 81)
Ez-egoki	35	19	25	119	86	95	55	66	56
	(% 58)	(% 32)	(% 42)	(% 29)	(% 21)	(% 23)	(% 14)	(% 17)	(% 14)
Berdinketa	13	6	17	50	21	46	21	15	20
	(% 22)	(% 10)	(% 28)	(% 12)	(% 5)	(% 11)	(% 5)	(% 3)	(% 5)
Doitasuna	0,26	0,65	0,42	0,67	0,78	0,74	0,85	0,83	0,85

5.17 taula – Ordainen ebaluazioa, sorkuntza metodoaren arabera sailkatuta.

Taulak erakusten dizkigun emaitzak ikusita, termino neoklasikoetarako garatutako NeoTerm sistemak ematen ditu emaitzarik onenak, aditu talde guztien arabera (0,85eko doitasuna orokorrean eta hizkuntzalarien arabera eta 0,83koa medikuen arabera). SNOMED CTren eta GNS10en arteko mapaketak izan du doitasun baxuena, eta aldi berean desadostasun handiena. Hizkuntzalariek 0,26ko doitasuna eman dioten bitartean, medikuek 0,65ekoa eman diote. Uste dugu, GNS10aren izaerak baldintzatu duela desadostasun nabari hori. Izan ere, aipatu bezala GNS10 sailkapen bat da, eta horrela adierazten dira bertako kontzeptuak ere. Adibidez, “konjuntibitisa, zehaztugabea” deskribapenaren bidez adierazten da konjuntibitis kontzeptua. Gure hipotesiaren arabera, medikuak sailkapen honekin lan egiten ohituta daudenez, arrotza ez egiteaz gain, kontzeptua bera adierazteko deskribapena balkeotzat hartu dute. Gure hipotesiari jarraiki, hizkuntzalariek, alde linguistikoari begiratuta, ez dute termino egokitzat jo. Gainera, badirudi medikuek ez diotela garrantziarik eman terminoen bukaerako “-a” mugatuaren deklinabideari eta egokitzat hartu dituzte mugatuan agertzen diren termi-

noak. Hizkuntzalariek aldiz, terminoak forma mugagabean adierazi behar direla kontuan hartu dute, eta ez-egokitzat hartu dituzte.

Baliabide lexikalen kasuan, ez genuen espero doitasuna NeoTermena baino baxuagoa izatea. Emaitzak hobeto ulertzeko 5.18 taula sortu dugu, baliabide lexikal bakoitzaren emaitzak erakusteko.

		Egoki	**Ez-egoki**	**Berdinketa**	**Denera**	**Doitasuna**
ZT Hiztegia	Hizk	128	4	10	142	0,97
	Med	127	8	7	142	0,94
	Den	135	2	5	142	0,99
Euskalterm	Hizk	96	8	13	117	0,92
	Med	96	16	5	117	0,86
	Den	98	12	7	117	0,89
Anatomia	Hizk	28	17	1	46	0,62
	Med	30	15	1	46	0,67
	Den	29	15	2	46	0,66
Erizaintza	Hizk	92	3	14	109	0,97
	Med	100	7	2	109	0,93
	Den	103	6	0	109	0,94
GNS10	Hizk	18	20	7	45	0,47
	Med	42	1	2	45	0,98
	Den	23	6	16	45	0,79
Elhuyar	Hizk	29	66	13	108	0,31
	Med	57	44	7	108	0,56
	Den	37	53	18	108	0,41

5.18 taula – Baliabide lexikalen emaitzak.

Doitasunari erreparatzen badiogu, ZT Hiztegiak, Euskaltermek eta Erizaintzako Hiztegiak lortutako emaitzak azpimarragarriak dira, 0,9tik gorako doitasuna izan baitute oro har. Giza Anatomiako Atlasak lortutako emaitzak harritu gaitu nabarmen, doitasun altuagoa espero baikenuen. Dena dela, Giza Anatomiako Atlasaren lagina gehienena baino txikiagoa izateak eragina izan dezake. GNS10aren kasua ere nabarmentzekoa da, izan ere, mapaketarekin gertatzen den bezala, hizkuntzalariek oso doitasun baxua eman dioten bitartean, medikuek doitasun altuena eman diote. Kasu horretan ere, lagina Giza Anatomiako Atlasaren antzekoa da tamainan. Elhuyar Hiztegi orokorraren kasuan, aurreikusi genuen bezala, doitasun baxuagoa dauka gainerako hiztegi espezializatuek baino. Dena dela, ezin dugu ahaztu ez dugula sinonimoak lortzeko erabili, eta hortaz, egiten duen ekarpena, sinbolikoa bada ere,

euskaratzean lagungarri izan daiteke. ZT Hiztegiaren kasuan, doitasunaren datuak aztertzen baditugu, denen arteko emaitza aztertzean (0,99), adituek beraien arteko doitasuna (0,97 hizkuntzalariek eta 0,94 medikuek) baina altuagoa dela ikus dezakegu. Kontuan izan, bozketa bidezko emaitzak erakutsi ditugula, eta kasu horietan, bozketaren ondorioz termino gehiago egokitzat hartu ditugu, talde baten barruan berdin berdinduak egon zitezkeen terminoak bestean egokitzat jo direlako.

Jarraian, topatutako fenomeno interesgarrien analisia egingo dugu:

- GNS10aren sortze data dela-eta (1996), gaur egungo Euskaltzaindiaren arauren bat betetzen ez duela ohartu gara, 5.10. adibidean ikus dezakegunez. Horrek arazo bat sortzen digu ordainen sendotasunean. Arazoari konponbidea eman diogu -ejia atzizkidun terminoetan ordezkapen automatikoak eginez.

Ingelesezko terminoa: *Hemiplegia*
GNS10: hemiplejia
ZT Hiztegia eta Erizaintzako Hiztegia: hemiplegia

5.10 irudia – Arau ortografikoak betetzen ez dituen GNS10aren adibide bat.

- GNS10arekin jarraituz, termino bat deskribatzeko ohikoak ez diren egiturak aurkitu ditugu (ikus 5.11. adibidea). Egitura horrek ez ditu termino izateko irizpideak betetzen, tartean "," bat txertatzen duelako, besteak beste. Adibideari jarraiki, horren ordain onargarriagoa "kanpoaldeko goiko ezpain" litzateke.

Ingelesezko terminoa: *External upper lip*
GNS10: goiko ezpaina, kanpoaldea

5.11 irudia – Terminoen egitura ohikoa betetzen ez duen GNS10aren adibide bat.

- Amaitzeko, aurretik aipatu dugun mugatasunaren arazoa identifikatu dugu GNS10aren kasuan bereziki. Bertako termino gehienak forma

mugatuan agertzen dira, eta hori ez da termino baten forma egokia. Aurretik aipatu bezala, arazo horri aurre egiteko, beste hiztegiek sinonimoak eman dituzten kasuetan, bertan terminoa mugatasunaren markarik gabe agertzen bada, GNS10eko ordainean zuzendu egin dugu. Gainerako kasuetan, ezin izan dugu erabaki, medikuntzako atzizki askotan aurkitu baitugu "-a" itsatsia ("-algia", "-kardia", "-zefalia",...).

5.5 Laburpena eta ondorioak

Kapitulu honetan, bereziki termino sinpleen sorkuntzan egindako ekarpena azaldu dugu. Alde batetik, gaur egun eskura ditugun baliabide lexikal espezializatu eta eleanitzak integratu ditugu EuSnomed aplikazioan. Erabilitako baliabideen artean, SNOMED CTren euskaratzean emaitza esanguratsuenak lortu dituztenak ZT Hiztegia (0,99ko doitasuna), Euskalterm terminologia bankua (0,89ko doitasuna) eta Erizaintzako Hiztegia (0,94ko doitasuna) izan dira emaitzen ataleko adituen ebaluazioaren arabera (5.18 taula). Giza Anatomiako Atlasak doitasunari dagokionean ez bada horren nabarmena izan, gorputz-egituren euskaratzean izan duen ekarpena nabarmentzekoa izan da (3.120 ordain berri 5.13 taulan).

Bestetik, termino neoklasikoak euskaratzeko sistema bat sortu dugu, NeoTerm, eta honen hiru hurbilpen garatu ditugu. Lehenengo hurbilpena, oinarri-lerro sistema da, afixu neoklasikoen konposaketan oinarritzen dena. Hurbilpen honek doitasun altua izan arren (0,891), estaldura ez da horren ona (0,343), eta horrela, bigarren hurbilpenean, estaldura hobetzea izan da lehentasuna.

Estaldura hobetzeari begira, bigarren hurbilpenean transliterazio modulu bat integratu dugu, afixuen hiztegiak zabaltzearekin batera. Doitasunean emaitza kaskarragoak lortu baditugu ere (8 puntu gutxiago), estaldura asko igo da (48 puntu), eta horrela, estaldura eta doitasuna orekatzea lortu dugu, 0,81eko F-neurria lortuz.

Azkeneko hurbilpenean, termino neoklasikoen identifikazioa findu nahi izan dugu, termino neoklasikoak ez diren terminoak NeoTermek baztertu ditzan, eta horrela erroreak ez sortzeko. Horretarako, identifikaziorako algoritmoa findu dugu, adituek proposaturiko irizpideak kontuan izanik. Hurbilpen honekin aldiz, ez dugu emaitzak hobetzea lortu, eta bigarren hurbilpenarekin alderatuta, doitasuna bere horretan geratu bada ere, estaldurak 7 puntu behera egin du.

Horrela, EuSnomeden NeoTermen bigarren hurbilpena integratu dugu, transliterazio moduluan oinarritzen dena, alegia.

Emaitzetan ikusi ahal izan dugun bezala (5.14 taulan), termino sinpleen euskaratzean portzentaje altuak lortu baditugu ere (% 75 gora lau hierarkietan), termino konplexuen kasuan estaldurak ematen dizkigun zenbakiak oso baxuak dira.

Hori guztia kontuan izanik, hurrengo kapituluan termino konplexuen euskaratzeari helduko diogu, eta horretarako garatutako sistemak azalduko ditugu, bata termino habiaratuetan oinarrituta, eta bestea itzultzaile automatiko baten domeinurako egokitzapena eginda.

6

Termino konplexuak: termino habiaratuak eta itzultzaile automatiko baten egokitzapena

Kapitulu honetan termino konplexuak euskaratzeko erabilitako teknikak aurkeztuko ditugu. Termino habiaratuen bidezko termino konplexuen sorkuntza azaltzen hasiko gara 6.1 atalean. Jarraian, 6.2 atalean, Matxin Itzultzaile Automatikoaren osasun-zientzien domeinurako egokitzapena azalduko dugu. Hirugarrenik, 6.3 atalean, termino konplexuen sorkuntza automatikoaren ebaluazioaren diseinua aurkeztuko dugu, eta 6.4 atalean ebaluazio horren emaitzak emango ditugu. Bukatzeko, 6.5 atalean, kapituluaren laburpena eta ondorioak aurkeztuko ditugu.

6.1 Termino konplexuen sorkuntza termino habiaratuen bidez

Atal honetan, EuSnomed sistemaren hirugarren urratsa azalduko dugu. Urrats horretarako, KabiTerm deituriko sistema garatu dugu. KabiTerm, termino konplexuen barruan agertzen diren beste terminoetan oinarritzen da termino konplexuak euskaratzeko.

KabiTermen egungo garapenak ingeleseko termino konplexu bat jasotzen du sarreran, eta baliabideak izanez gero, euskarazko ordainak proposatzen ditu. Baliabide horiek, habiaratutako terminoen ordainak eta euskaratze-pa-

troiak dira. Aurrerago azalduko dugun moduan, KabiTermen lana errazteko, AnaMed deituriko analizatzailea prestatu dugu. Analizatzaile hori, KabiTermek beharrezkoa duen informazioa biltzeaz arduratzen da, eta termino habiaratuak identifikatzen eta prestatzen ditu, KabiTerm euskaratzeaz soilik ardura dadin.

Alde batetik, KabiTerm sistemarako diseinatu eta garatu dugun AnaMed analizatzailea azalduko dugu 6.1.1 atalean, eta bestetik, 6.1.2 atalean KabiTerm sistema bera azalduko dugu.

6.1.1 AnaMed: Osasun-zientzietarako analizatzailea

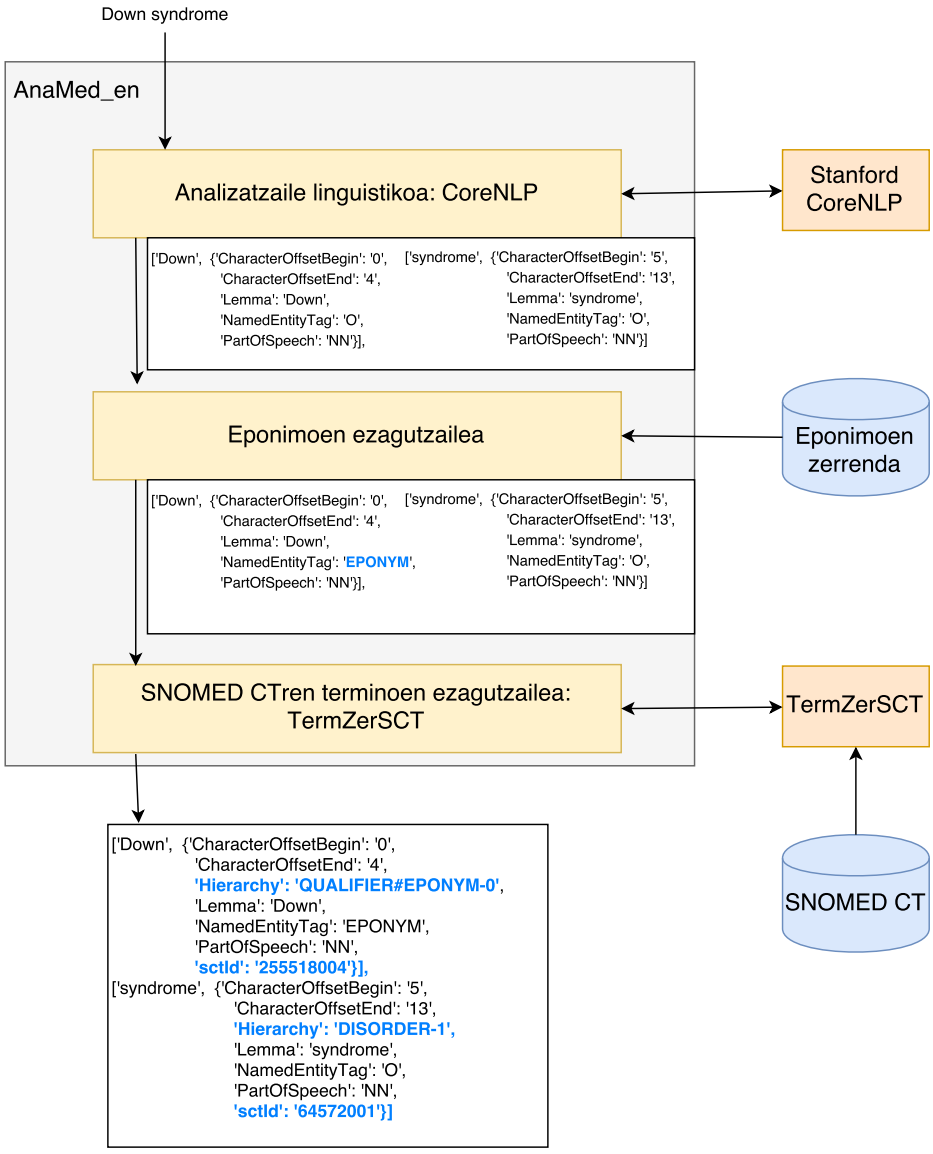
Atal honetan AnaMed osasun-zientzietarako hizkuntza-analizatzailea aurkeztuko dugu. Analizatzaile horrek, informazio linguistikoa analizatzeaz gain, SNOMED CTren terminoak identifikatzen ditu testuan, baita eponimoak ere. Eponimoak kontzeptuen izendatzetan agertzen diren pertsona-izenak dira.

AnaMedek euskaratze prozesurako beharrezkoa iza ndaitekeen informazioa biltzen du. Hasiera batean, KabiTerm sisteman erabiltzeko ingeleserako bakarrik garatu genuen. Hala ere, AnaMed beste hizkuntzetara oso erraz egoki daitekeenez, euskararako ere garatzea erabaki genuen, euskarazko osasun-txostenen idazketan lagungarria izan daitekeelakoan, eta etorkizunean espainierarako ere garatzea aurreikusten dugu.

AnaMedek analizatzaile automatiko bat hartzen du oinarritzat, eta eponimoen eta SNOMED CTren terminoen identifikazioa integratzen ditu. Bere arkitektura 6.1 irudian ikus dezakegu. Hiru modulu ditu: analizatzaile linguistikoa, eponimoen ezagutzailea eta SNOMED CTren termino ezagutzailea.

Ingeleserako analizatzailerako (AnaMed_en) Stanford taldearen CoreNLP tresna (Manning *et al.*, 2014) erabili dugu lehenengo moduluan, horretarako Dustin Smith-en Stanford CoreNLP Pythonerako interfazea erabili dugu[1]. Euskararako (AnaMed_eu), berriz, Eustagger (Ezeiza *et al.*, 1998) erabili dugularik. Analizatzaile linguistikoetatik tokenizatzailea eta etiketatzaile morfologikoak erabili ditugu, tokenen lema eta kategoria gramatikala jasotzeko. Informazio horretaz gain, tokenaren posizioa (ingelesez *offset*) eta AnaMed_en kasuan entitate-ezagutzailearen informazioa ere integratu ditugu (6.1 irudian, lehenengo moduluaren irteera).

[1]https://github.com/dasmith/stanford-corenlp-python (2017ko maiatzaren 9an atzitua).



6.1 irudia – AnaMed analizatzailearen arkitektura.

Esan bezala, bigarren moduluan eponimoen identifikazioa gehitu diogu (ikusi 6.1 irudia). Medikuntzan eponimoak oso ohikoak dira, bereziki gaixotasunak eta sindromeak adierazteko. Adibidez, “Down-en sindromea” edo

"Alzheimer-en gaixotasuna" terminoetakoak[2].

Azkenik, analisia osatzeko asmoz, SNOMED CTren terminoen identifikazioa ere gehitu diogu AnaMedi (6.1 irudiko hirugarren modulua).

Analizatzaile linguistikoei ez zaie inolako egokitzapenik egin tesi-proiektu honetan. Beraz, jarraian, lan honetan sortu ditugun moduluak bakarrik deskribatuko ditugu.

Eponimoen ezagutzailea

Leman aldaketarik gabeko eponimoak aurreko adibidean erakutsitako "Downen sindromea" edo "Alzheimer-en gaixotasuna" bezalakoetan daude, non eponimoa bera esplizituki agertzen den, "Down" eta "Alzheimer" kasu honetan. Horietaz gain, badira hainbat termino eponimoetatik eratorriak izan direnak, daltonismoa kasu. Daltonismo terminoa, John Dalton kimikari britainiarren omenez izendatu zen, bera izan baitzen daltonismoa deskribatzen lehena[3]. Hala ere, inplementatu dugun eponimoen ezagutzaileak ez ditu eponimoetatik eratorritako terminoak identifikatzen, daltonismoa kasu, baizik eta eponimo literalak soilik.

Eponimoen ezagutzailea garatzeko orduan, euskararen gramatika izan dugu buruan. Izan ere, pertsona-izena izan, edo toki-izena izan, euskaraz emango zaion trataera deklinabideari dagokionean ezberdina izango da. Eponimoen definizioan ez dugu adostasunik aurkitu eta batzuetan toki-izenei ere erreferentzia egiten diete[4,5]. Adibidez, toki-izenei dagokionean *Stockholm syndrome* Stockholm hirian gertatutako gertakari baten ostean izendatu zen[6], eta horrela, euskarazko ordaina "Stockholmgo sindrome" da, lekuzko genitiboa erabiliz. Pertsona-izenekin aldiz, edutezko genitiboa (posesiboa) da erabiltzen den deklinabidea, "Weber-en proba" terminoaren kasuan, adibidez.

Eponimoen identifikatzailearekin lanean hasi baino lehen, entitate-ezagutzaileak (Nadeau eta Sekine, 2007; Tjong Kim Sang eta De Meulder, 2003) aztertu ditugu, eta artearen egoerako entitate-ezagutzaileak probatu ditugu SNOMED CTren deskribapenak analizatzeko. Eskuz egindako azterketa horren ondorioz, Stanforden CoreNLP tresnaren entitate-ezagutzaileak

[2]Bi terminoak Euskalterm Terminologia Banku Publikotik erauziak.

[3]https://eu.wikipedia.org/wiki/John_Dalton (2017ko maiatzaren 9an atzitua).

[4]https://en.wikipedia.org/wiki/Eponym (2017ko maiatzaren 9an atzitua).

[5]http://www.dictionary.com/browse/eponym (2017ko maiatzaren 9an atzitua).

[6]https://en.wikipedia.org/wiki/Stockholm_syndrome (2017ko maiatzaren 9an atzitua).

(Finkel *et al.*, 2005) eman ditu emaitza egokienak. Emaitza egokienak izan arren, eponimo gehienak identifikatu gabe gelditzen direnez, Internetetik eponimo ezagunen zerrendak erauzi ditugu, eta zerrenda horietan oinarrituta eponimo-ezagutzaile bat garatu dugu. Horrela, Stanforden CoreNLP tresnaren entitate-ezagutzailearen Pertsona klasekoak eta gure sistemak identifikatutako eponimoak etiketatzen ditugu eponimo gisa.

Eponimoen ezagutzaileak zerrendako eponimoak termino konplexuaren hitzen artean bilatzen ditu. Batzuetan, eponimo konposatuak erabiltzen dira gaixotasunak izendatzeko, *Verner-Morrison syndrome* terminoan, adibidez. Kasu horiek kontuan izanik, eponimoen zerrenda sortzeko garaian, eponimo konposatuetatik eponimo sinpleak erauzi ditugu. Horrela, adibidearen kasuan bi eponimo zerrendaratu ditugu. Ezagutzailearen estaldura hobetzeko, eponimo konposatuak ezagutzeko garaian, eponimoetako bakarra ezagututa eponimo konposatua osotasunean identifikatzen du. Eponimoen zerrenda osatzeko, SNOMED CTren termino guztiak analizatu ditugu, horietan identifikatutako eponimo konposatuetan osagai ezezagunak zerrendara gehituz. Amaieran 3.000 pertsona-izen inguruko zerrenda osatu dugu eponimoak identifikatzeko.

TermZerSCT: SNOMED CTren terminoen ezagutzailea

AnaMeden helburu nagusia termino barruan termino habiaratuak identifikatzea da. Nahiz eta gaur egun termino erauzle asko egon, KabiTermen beharretara egokitzen denik ez dago. Izan ere, gure kasuan ez zaigu terminoak orokorrean identifikatzea interesatzen, baizik eta SNOMED CTren baitan agertzen diren terminoak identifikatzea, eta euren hierarkia erabili nahi dugu.

Hori horrela izanik, TermZerSCT terminologia zerbitzaria prestatu dugu SNOMED CTren terminoen identifikaziorako. SNOMED CTren edukia oso zabala da, 300.000 kontzeptu inguru izanik, eta horren prozesaketa ez da berehalakoa. TermZerSCTri esker, eduki terminologikoaren kudeaketa zerbitzaria abiaratzean egiten dugu, eta zerbitzaria martxan dagoela, SNOMED CTren inguruko informazioa unean jasotzen dugu, itxaron denbora minimoekin.

Esan bezala, zerbitzari horrek SNOMED CTren eduki terminologikoa prestatzen du, bezeroak (AnaMed kasu honetan) behar duen informazioa ahalik eta modu eraginkorrenean jaso ahal izateko. Besteak beste, SNOMED CTren jatorri fitxategietatik abiatuz, aktibo dauden kontzeptuak modu

hierarkikoan sailkatzen ditu, eta horrela, SNOMED CTren kontzeptu identifikadore bat emanda, horren FSN, hobetsitako terminoa edo sinonimoak itzultzeaz gain, dagokion hierarkia ere ezagutu dezakegu. Informazio hori AnaMedetik eta eponimoen ezagutzailetik jasotako informazioari gehitzen zaio, 6.1 irudiko irteeran ikus daitekeen moduan.

Hurrengo taulan (6.1) ikus dezakegun bezala, termino bat izanda (*diabetes mellitus* adibidearen kasuan), bere SNOMED CTren kontzeptu identifikadorea lortu dezakegu, eta behin kode hori dugula, kontzeptuaren informazioa eskuragarri dugu, hala nola FSNa, termino hobetsia (PT), edo sinonimoak.

Azalpena	**Funtzioa**	**Emaitza**
Kodea lortu	`desc2sct`	73211009
Hierarkiak lortu	`sct2hie`	DISORDER
FSN lortu	`sct2fsn`	Diabetes mellitus (disorder)
PT lortu	`sct2term`	Diabetes mellitus
Sinonimoak lortu	`sct2syn`	DM - Diabetes mellitus

6.1 taula – *Diabetes mellitus* terminotik TermZerSCT zerbitzaria erabiliz lor dezakegun informazioa.

Zerbitzaria ingeleserako, espainerarako eta euskararako prestatu dugu, SNOMED CTren eduki terminologikoa hizkuntza horietan baitaukagu. Ingelesezko zein euskarazko bertsioak landuagoak ditugu. Horietarako, lematizatzailearen laguntzaz (Stanford CoreNLP ingeleserako eta Eustagger euskararako), terminoak lematizatuta bilatzeko aukera ere garatu dugu. Espainierakoari etorkizunean lematizatzeko aukera ere gehituko diogu.

AnaMedek, TermZerSCT zerbitzaria erabilita, termino konplexuen barruan dauden termino habiaratuak identifikatzen ditu, eta horiekin termino konplexuaren egitura aztertu dezakegu. Horretaz gain, termino habiaratuak eurak elkartu edo multzokatzen ditu "_" karakterearen bidez. Adibidez, *unstable diabetes mellitus* termino konplexuan, bi termino habiaratu identifikatu ditu: *unstable* kalifikatzailea eta *diabetes mellitus* nahasmendua. Identifikazio horri esker, AnaMedek analisi osoa emateaz gain, KabiTermerako oso erabilgarriak izango zaizkigun termino habiaratuen egitura (QUALIFIER+DISORDER) eta multzokatzea (*unstable diabetes_mellitus*) ere ematen dizkigu.

6.1.2 KabiTerm: termino konplexuen sorkuntza termino habiaratuak baliatuz

Atal honetan KabiTerm sistemaren aurkezpena egingo dugu. KabiTerm, transduktoreak erabiliz terminoen habiaraketan oinarritzen da termino konplexuetarako ordainak sortzeko. Tesi-lan honetan ingelesa-euskara hizkuntza parerako inplementatu ditugun transduktoreak eta aplikazioa aurkeztuko ditugu. Hala ere, sistema modu errazean egokitu daiteke beste edozein hizkuntza paretara.

KabiTermek aurreko atalean azaldutako AnaMed analizatzailetik jasotako informazioa erabiltzen du termino nagusian termino habiaratuak identifikatzeko. AnaMed ezinbestekoa ez izan arren, KabiTermen eraginkortasunean eragin positiboa dauka. Alde batetik, AnaMedek termino habiaratuen multzokatzea prestatzen du, transduktoreen eginkizuna sinplifikatuz. Bestetik, AnaMedek informazio linguistikoa prestatzen du, formen lemak eskainiz, eta horrela, KabiTermi pluralean dauden termino habiaratuen euskaratzeko gaitasuna ematen dio. Hortaz, KabiTermerako AnaMed erabat ezinbestekoa ez bada ere, eraginkortasunean eta emaitzetan eragin positiboa dauka.

Euskaratze-patroiak definitzeko, aurreko kapituluan NeoTerm sistema garatzeko erabilitako tresna bera, Foma (Hulden, 2009), erabili dugu. Foma bidez patroiak egoera finituko transduktoreen bidez inplementatu ditugu. Transduktoreak Fomaz idatzi baditugu ere, transduktoreen konbinaketa eta kudeaketa Python programazio-lengoaian idatzitako aplikazio batekin egin dugu.

Ingelesezko terminoen egituren analisia AnaMeden bidez

Esan dugun moduan, KabiTerm termino habiaratuetan oinarritzen da, hau da, termino konplexu batean agertzen diren beste terminoetan.

AnaMedi esker, SNOMED CTren terminoak termino habiaratuen egituraren arabera sailkatu ditugu. Hurrengo taulan (6.2 taula), egitura horien adibide batzuk ikus ditzakegu. Adibidez, *malignant neoplasm of renal calyx* termino konplexuan bi termino nagusi aurkitu ditugu *malignant neoplasm* nahasmenduen (*disorder*) hierarkiakoa bata eta *renal calyx* gorputz-egituren (*body structure*) hierarkiakoa bestea. Aipatu beharra dugu, habiaratutako termino gehiago ere aurkitzen dituela AnaMedek, hala nola *calyx* gorputz-egitura, *malignant* kalifikatzailea (*qualifier*) edo *neoplasm* nahasmendua.

Habiaratutako termino guztiekin sortu ditugu egiturak (ikusi 6.2 taulan

Terminoa	**Multzokatzea**	**Egitura**
structure of radial tuberosity	*structure of radial_ tuberosity*	structure+of+BODYSTR
Baelz's disease	*Baelz's disease*	EPONYM+'S+DISORDER
malignant neoplasm of renal calyx	*malignant_ neoplasm of renal_ calix*	DISORDER+of+BODYSTR

6.2 taula – AnaMeden bidez lortutako egiturak eta multzokatzeak.

azken zutabea), eta agerpen kopuruaren eta dependentzia kopuruaren arabera sailkatu ditugu. Hau da, egitura zenbat terminotan azaltzen den zenbatu dugu, eta termino hori zenbat terminotan agertzen den habiaratuta kontuan hartu dugu. Egitura horietako gutxi batzuk ikus daitezke 6.3 taulan ("Agerp." izenburuak agerpen kopuruari egiten dio erreferentzia eta "Depen." izenburuak egitura horrek beste terminoekiko duen dependentzia kopuruari). Adibidez, QUALIFIER+DISORDER egiturarako, 4.469 agerpen aurkitu ditugu, eta gainera 74.208 terminoetan agertzen da egitura hau habiaratuta. Lehentasun altuko egitura dela erabaki dugu, agerpen asko izateaz gain, gainerako termino konplexuak euskaratu ahal izateko oso garrantzitsua izango delako. QUALIFIER+*neoplasm* egitura aldiz, bere horretan 4 aldiz agertzen bada ere, beste termino konplexuek lau termino horiekiko duten dependentzia oso altua da (28.642 terminoen barruan agertzen da).

Egitura	**Adibidea**	**Agerp.**	**Depen.**
QUALIFIER+DISORDER	*unstable diabetes mellitus*	4.469	74.208
QUALIFIER+*neoplasm*	*malignant neoplasm*	4	28.642
PROCEDURE+of+BODYSTRUCTURE	*amputation of finger*	5.082	33.181
...	...	...	...

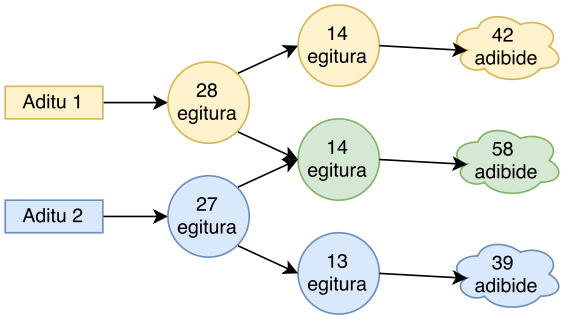
6.3 taula – SNOMED CTren terminoen egituren agerpenak eta dependentziak beste terminoekiko.

Bi adituri egitura horietako adibideak eman dizkiegu, horiei euskarazko ordainak emateko. Izan ere, Zabala *et al.* (2012) lanean erakusten duten bezala, hiztegi zein itzultzaileek jatorrizko terminoaren egituraren kalkoetara jotzeko joera duten bitartean, adituek euskararen berezko egituretara jo ohi dute. Hori horrela izanik, adituek sortutako lagina hobetsi dugu hiztegietatik erauzitako adibideak baino.

Adituek sortutako adibide horiek euskaratze-patroiak definitzeko oinarria izan dira. Domeinuko adituek duten jakintza terminoei ordain egokia eman ahal izateko, ezinbestekoa denez, euskaratzeaz arduratu diren adituak

medikuak izan dira. Adibideak lortzeko prozesua bi fasetan banatu dugu, lehenengo fasean lortutako patroien arabera hurrengo faseko egiturak eta adibideak aukeratu ahal izateko.

Lehenengo faserako 41 egitura aukeratu ditugu eta bakoitzeko gutxienez zoriz erauzitako 3 adibide. Aditu bakoitzari 28 eta 27 egitura eman dizkiogu hurrenez hurren, zeinetatik 14 egitura komunak diren, 6.2 irudian ikus dezakegun moduan. Denera 100 eta 97 adibideren euskarazko ordainak jaso ditugu, euretako 58 termino komunean dituztelarik.



6.2 irudia – Adituei banatutako lagina.

Komunean dituzten adibideen kasuan, adostasun altua neurtu dugu, eta desadostasun kasuekin, beraien arteko adostasuna bilatu dugu, irizpide berdinak erabil ditzaten. Hurrengo taulan (6.4 taula) adibide horietako batzuk ikus ditzakegu, non lehenengo adibidean ordain ezberdinak proposatu dituzten (nahiz eta gerora adostasuna lortu den), bigarrenean adostasuna izan duten, eta hirugarren eta laugarrenetan adibide ezberdinak dituzten. Adibideetan ikus daitekeenez adibide-terminoak ez dira sinpleak eta osasunari buruzko ezagutza behar da euskaratze egokia lortzeko.

Oinarrizko irizpideak argi izanik, bigarren faserako, 25na egitura prestatu ditugu, aditu bakoitzak denera 100 adibide inguru dituelarik. Bi faseak elkarturik, 340 adibide inguru izan ditugu euskaratze-patroiak definitzeko eta horietatik abiatuta 53 euskaratze-patroi definitu ditugu (C eranskinean ikus daitezke).

KabiTerm sistemaren diseinua

Euskaratze-patroiak lortzeko prozesua deskribatu ondoren, jarraian, besteak beste KabiTermek egiten duen patroi horien erabilera aztertuko dugu. Ka-

Ingelesa	**Aditu 1**	**Aditu 2**
cryotherapy to cranial nerve	nerbio kranialaren krioterapia	garezurreko nerbioen krioterapia
calcium regulating agent overdose	kaltzioaren agente erregulatzaileek eragindako gaindosia	kaltzioaren agente erregulatzaileek eragindako gaindosia
open fracture of scaphoid bone of wrist	eskumuturreko eskafoide hezurraren haustura irekia	
adrenergic neurone blocking drug adverse reaction		neurona adrenergikoen blokeatzaileek eragindako kontrako efektua

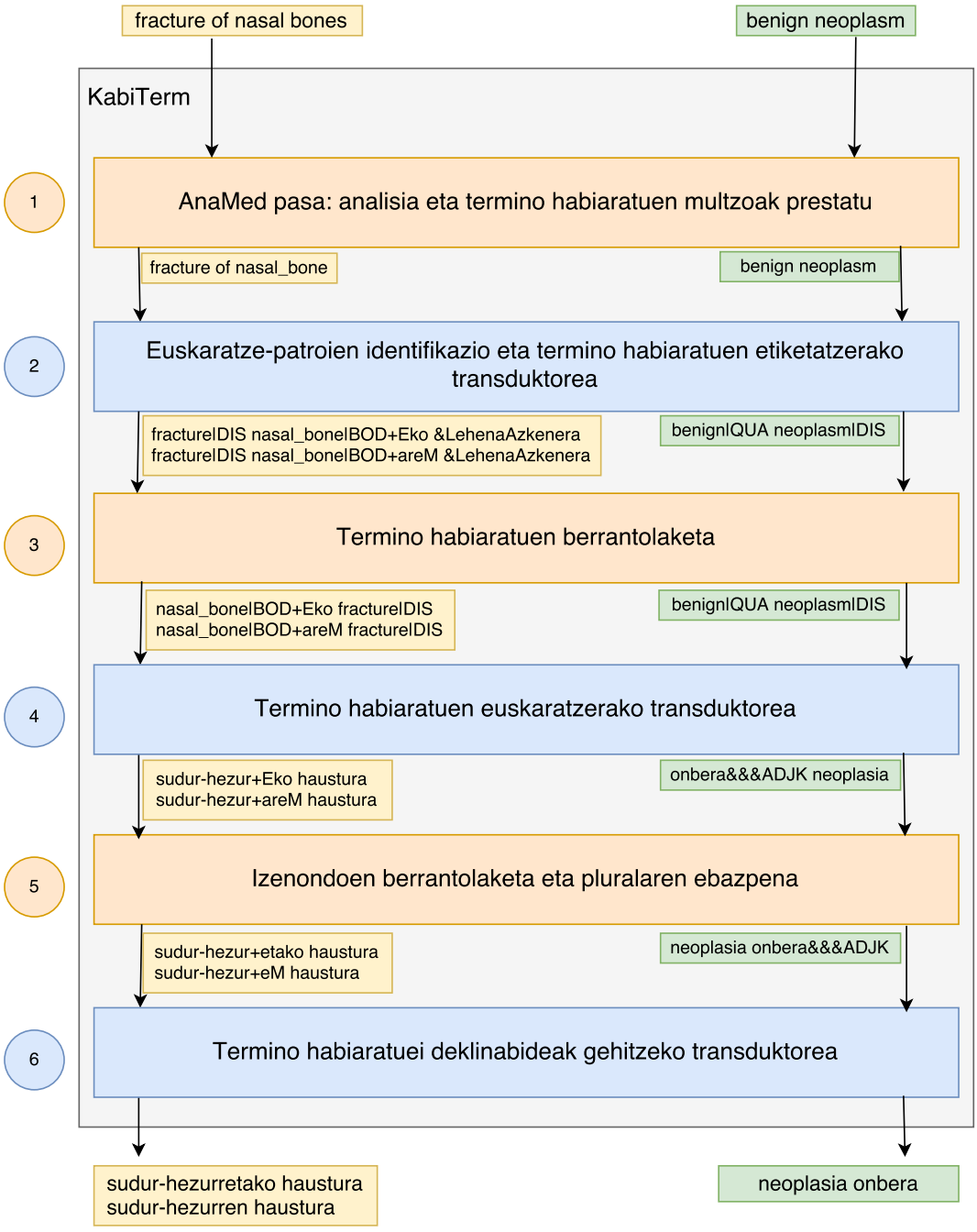
6.4 taula – Adituen euskaratzeen adibide batzuk.

biTermen funtzionamendua 6.3 irudian irudikatu dugu, eta honako pausoak ematen ditu:

1. Aurrena, sarrera terminoa AnaMedek analizatzen du, termino habiaratuak identifikatuz eta multzokatuz: *fracture of nasal bones* terminoaren kasuan, *fracture* nahasmendua da eta *nasal_ bone* gorputz-egitura. Termino habiaratuak multzokatzeaz gain, lematizazioa ere beharrezkoa izan da adibide honetan, *nasal bone* SNOMED CTren terminoa delako, baina ez *nasal bones* forma pluralean.

2. Bigarrenik, euskaratze-patroien identifikazio eta termino habiaratuen etiketatzerako transduktoreari deitzen diogu, eta honek, dagokion euskaratze-patroia aplikatuz, termino habiaratuei euskaratzeko beharrezko etiketak gehitzen dizkie. Kasu honetan, DISORDER+of+BODYSTRUCTURE egitura identifikatu du, eta horri dagokion euskaratze-patroia aplikatu dio: *fracture* terminoari "|DIS" etiketa gehitu dio, nahasmendu bat delako, *nasal_ bone* terminoari aldiz, "|BOD+Eko" etiketa eta "|BOD+areM" etiketak gehitu dizkio, gorputz-egitura izateaz gain, deklinabidea ere gehitu behar izan diolako ("+Eko" eta "+areM" adibidean). Horri guztiari, ordena aldatzeko etiketa ere gehitu dio, lehenengo terminoa amaieran jartzeko ("&LehenaAzkenera").

3. Jarraian, termino habiaratuen berrantolaketa egiten dugu, aurreko urratsak sortutako etiketari jarraituz (&LehenaAzkenera). Horrela, *fracture* terminoa lehenetik azkenera pasatzen da.



6.3 irudia – KabiTermen arkitektura eta funtzionamenduaren adibideak.

4. Laugarrenik, termino habiaratuen euskaratzeko transduktoreari deitzen diogu. Horrela, euskarazko ordainak jasotzen ditugu: "sudur-hezur+Eko haustura" eta "sudur-hezur+areM haustura" lortuz (hierarkien etiketak desagertu dira eta ingelesezko termino bakoitzaren euskarazko ordaina dugu irteeran).

5. Ostean, termino habiaratuetako bat jatorrian pluralean zegoenez, horren deklinabideak eguneratzen ditugu pluralaren forma jaso dezan: "+Eko" "+etako" bihurtzen da, eta "+areM" "+eM". Kasu honetan, "sudur-hezur+etako haustura" eta "sudur-hezur+eM haustura" lortu ditugu. Gogoratu, sarrerako terminoa pluralean dagoela (*fracture of nasal bones*). Adibide honetan gertatzen ez bada ere, izenondoren bat egonez gero, horren berrantolaketa ere gauzatzen da urrats honetan. Izan ere, euskaraz, izenondoak izenaren ondoren kokatzen dira, eta izenlagunak izenaren aurretik. Adibidez, "onbera" izenondoa, izenaren ostean kokatzen da, adibidez, "neoplasia onbera" terminoan ikusten dugun moduan. Izenlagunak, aldiz, izenaren aurrean gelditzen dira ingelesezko adjektiboen antzera, adibidez "jaiotzetiko kiste" terminoan "jaiotzetiko" izenlaguna izenaren aurretik ikus dezakegu.

6. Azkenik, termino habiaratuei deklinabideak gehitzeko transduktoreari deitzen diogu, euskarazko termino konposatuak lortuz: "sudur-hezurretako haustura" eta "sudur-hezurren haustura".

Azaldu berri dugun euskaratze-prozesu honetan hainbat faktore hartu ditugu kontuan. Genitiboaren kasuan, zenbaitetan ez da erraza izaten lekuzko genitiboa ala edutezko genitiboa erabili behar den erabakitzea. Adibidez, *"abdominal aorta"* terminorako, Euskaltermek lekuzko genitiboa, "abdomeneko barrunbe" erabiltzen du eta Anatomiako Atlasak genitiboa erabiltzen du, "abdomenaren barrunbe". Adituekin kontrastatuta, irizpide argia dago: kokapena adierazi nahi denean, lekuzko genitiboa; osoa-zatia erlazioa adierazi nahi denean, edutezko genitiboa (Zabala *et al.*, 2012). Hala ere, irizpide hori automatizatzea ez da argia, eta adituen iritzi jantzia ezinbestekoa da, testuingurua eta horren ulertzea ezinbestekoa delako. Hori horrela izanik, eta tesi honen helburua ez denez erreferentziazko ordainak sortzea, baizik eta ordain hautagaiak sortzea, gainsorkuntzaren bidea aukeratu dugu kasu horietan, eta bi modutara osatu ditugu terminoak.

Bigarren eta laugarren urratsetan, termino habiaratuak identifikatzeko, zein horiek euskaratzeko lexikoi elebidunak erabiltzen ditugu. Lexikoi horiek

SNOMED CTren terminoekin osatu ditugu, jadanik euskaratu ditugun bikoteekin alegia. Hierarkia bakoitzerako (nahasmendu, gorputz-egitura, ...) lexikoi bat sortu dugu, hierarkia bakoitzean ordain ezberdinak izateko aukera dagoelako, eta aldi berean, egituren identifikazioa hierarkien araberakoa delako.

Hurrengo lerroetan, bigarren, laugarren eta seigarren urratsetan sakonduko dugu. Izan ere, hiru urrats horietan euskaratze prozesua gauzatu ahal izateko transduktoreak definitu ditugu, eta gainerako hiru urratsak (lehenengo, hirugarren eta bosgarrena), prozesu prestatzeaz eta kudeatzeaz arduratzen dira.

Euskaratze-patroien identifikazio eta termino habiaratuen etiketatzerako transduktorea

Urrats honetako patroi bakoitza lau erregelen bitartez adierazi dugu. Lehenengoa egitura identifikatzeaz arduratzen da; bigarrenak etiketak gehitzen ditu; hirugarrenak patroiari dagokion identifikatzailea gehitzen dio; eta azkenak aurreko erregelak konbinatzen ditu, kendu beharreko ingelesezko hitzak kentzen ditu eta ordenari dagozkion etiketak prestatzen ditu. Hurrengo irudian (C.1) patroi baten erregela sorta osoa ikus dezakegu.

```
1  define Dis    ?+ @-> ... {|DIS} || Muga _ Muga ;
2  define Bod    ?+ @-> ... {|BOD} || Muga _ Muga ;
3  define SinGEN ?+ @-> ... {+areM} || Muga _ Muga ;
4  define GEL ?+ @-> ... {+Eko} || Muga _ Muga ;
5  define KenOf " " {of} " " -> " " ;
6  define OrdAldatuLehenaAzkenera ?+ @-> ... " " {&LehenaAzkenera} ;
7  ####################
8  define EzDisOfBod HDIS " " {of} " " HBOD;
9  define DisOfBod Dis " " {of} " " (Bod .o. [SinGEN|GEL]);
10 define EtDisOfBod ?+ @-> ... { |pat_or_011};
11 define TrDisOfBod EzDisOfBod .o. DisOfBod .o. KenOf .o.
       OrdAldatuLehenaAzkenera .o. EtDisOfBod;
```

6.4 irudia – KabiTermen identifikazioaren eta etiketatzearen patroi bat.

Alde batetik, etiketak gehitzeko erregela orokorrak ikus ditzakegu 1. lerrotik 6. lerrora: hierarkiaren etiketa gehitzekoa (1. eta 2. lerroak); deklinabide-markak gehitzekoak 3. eta 4. lerroetan; 5. lerroan ingelesezko *of* preposizioa kentzeko erregela; eta azkenik 6. lerroan ordena aldatzeko etiketa gehitzekoa, hasierako elementua amaierara eramatekoa kasu honetan.

Bestetik, DISORDER+of+BODYSTRUCTURE egiturari dagokion euskaratze-patroiaren identifikazio- eta etiketatze-erregelak ikus ditzakegu 8. lerro-

tik 11.era. Aurrena, egitura ezagutzeko erregela definitu dugu `EzDisOfBod` izenarekin: nahasmenduen hiztegiko termino bat (`HDIS` erregelan gordeta ditugu lexikoitik jasotako nahasmenduak), *of* preposizioa eta azkenik gorputz-egituren hierarkiako termino bat (`HBOD`). Bigarrenik, 9. lerroan, termino habiaratuei gehitu beharreko etiketak gehitzen dizkiegu `DisOfBod` erregelaren bitartez. Hamargarren lerroan, patroiaren etiketa orokorra gehitzen diogu (`EtDisOfBod`), garapena kontrolatzeko eta emaitzak emateko erabili duguna. Eta azkenik, aurreko hiru erregelak konbinatzeaz gain, *of* preposizioa kentzeko erregela eta ordena aldatzeko erregelak konbinatzen ditugu `TrDisOfBod` erregelan.

Bigarren urrats honetako euskaratze-patroi batzuren adibideak erakusten ditugu 6.5 taulan. Lehenengo erregela, eponimo batek eta nahasmendu batek osatzen dute. Patroian, termino bakoitzaren hierarkia adierazten dugu: `Epo` erregelak, eponimoen marka gehitzen du, eta `Dis` erregelak nahasmenduen marka. Horretaz gain, euskarazko ordainean ikusten dugunez, eponimoaren eta edutezko genitiboaren deklinabidearen artean marratxo bat gehitu diogu. Hori `MarGEN` erregelaren bidez gehitzen diogu. Kontuan izan, euskaraz eponimoak deklinabide-markarekin eta gabe erabiltzen ditugula. Hau da, "Down-en sindrome" eta "Down sindrome" biak sortuko ditugu. Izan ere, ingelesezko terminoetan eponimoek daudenean, aposizioak[7] sortzeko joera da nagusi (*Down syndrome*). Gaztelaniaz aldiz, *de* preposizioaren bidez osatutako sintagmak sortzen dira (*sindrome de Down*). Euskaraz, bere ezaugarriak kontuan izanik, naturalagoa da aposizioen formatua hartzea ("Down sindrome"), baina espainierarekin egindako kalkoengatik beste forma ere zabaldu da ("Down-en sindrome"). Are gehiago, Euskalterm bezalako erreferentziazko hiztegi terminologikoetan deklinatutako formak agertzen dira nagusiki. Kasu honetan ere, genitiboarekin bezala, gainsorkuntzaren alde egin dugu, eta adituen esku utziko dugu termino hobetsiaren aukeraketa eta sinonimoen baztertzea.

Bigarren adibidean, gorputz-egituraren etiketa gehitzeko `Bod` erregela erabiltzen dugu, eta edutezko genitibo singularraren marka ere gehitzen diogu `SinGEN` erregelaren bitartez. Azkenik, *structure* hitza ez da inongo hierarkiatan agertzen, eta bestelako zerrenda batean gehitu dugu `Bes` etiketa erabiliz. Zerrenda horretan (`Bes` zerrendan) hierarkietan agertzen ez diren hitz edo terminoak agertzeaz gain, patroi zehatzagoak definitzeko erabilitakoak gehi-

[7] Gure kasuan, bi izenez osatutako eraikuntza, non izenetako batek bestea azaltzeko edo zehazteko balio duen.

	Ingelesa	Euskara	Erregela
1	*Down syndrome*	Down-en sindrome Down sindrome	`(Epo .o. (MarGEN)) " " Dis`
2	*head structure*	buruaren egitura	`(Bod .o. SinGEN) " " Bes`
3	*heroin overdose*	heroinak eragindako gaindosi	`(Phar .o. ERGEra) " " Bes`
4	*fracture of hip*	aldakako haustura aldakaren haustura	`Dis" "{of}" "(Bod.o.[GEL|SinGEN])`
5	*benign neoplasm*	neoplasia onbera	`Qua "" Dis`

6.5 taula – KabiTermen identifikazio- eta etiketatze-urratseko erregela batzuk.

tu ditugu. Hori da hirugarren adibidean erakusten dugun kasua. Izan ere, ingelesezko *overdose* terminoa nahasmenduen hierarkiakoa bada ere, erregela berezitu bat sortu dugu termino horretarako, eta beraz, bestelako zerrendan ere gehitu dugu (`Bes`). Kasu honetan, `ERGEra` erregelarekin, ergatiboaren marka gehitzeaz gain, "eragindako" hitza gehitzeko etiketa jartzen diogu.

Laugarren adibidean, bi ordain sortuko ditugu (gutxienez), erregela horren bitartez. Izan ere, DISORDER+of+BODYSTRUCTURE egitura hain da orokorra, gainsorkuntzaren bitartez gorputz-egiturari edutezko genitiboa eta lekuzko genitiboa, biak, gehitzen dizkiogula ordainak lortzeko, lehenago azaldu dugun moduan.

Azkenik, bosgarren adibidean, kalifikatzaile bat eta nahasmendu bat aurkitzen ditu hurrenez hurren, eta horien etiketak gehitzen dizkiegu termino habiaratuei. Kasu honetan ez dira deklinabide-markak beharrezkoak izan, ez eta berrantolaketa etiketarik ere, aurrerago ikusiko dugun bezala, adjektiboen berrantolaketa bakarrik izenondo direnean egingo dugulako. Adibide horretan euskaratze-patroiak hala zehazten duelako egin da termino habiaratuen arteko ordenean aldaketa.

Aurreko taulako (6.5 taula) adibideekin jarraituz, 6.6 taulan lehenengo faseko transduktoreek adibide horietarako sortzen duten irteera erakusten dugu, hurrengo urratsera pasako den informazioa erakustearren.

Termino habiaratuen euskaratzerako transduktorea

Laugarren urrats honetan, lexikoi elebidunak erabiltzen ditugu etiketatutako termino habiaratuak euskaratzeko. Aurreko urratsean bezala, lexikoiak `HDIS` (nahasmenduak) eta `HBOD` (gorputz-egiturak) moduko erregeletan jaso

	Ingelesa	**2. urratsaren irteera**
1	*Down syndrome*	Down|EPO+-+ReM syndrome|DIS Down|EPO syndrome|DIS
2	*head structure*	head|BOD+areM structure|BES
3	*heroin overdose*	heroin|PHAR+ak_eragindako overdose|BES
4	*fracture of hip*	fracture|DIS hip|BOD+ko &LehenaAzkenera fracture|DIS hip|BOD+areM &LehenaAzkenera
5	*benign neoplasm*	benign|QUA neoplasm|DIS

6.6 taula – KabiTermen identifikazio- eta etiketatze-urratseko irteeraren adibide batzuk.

ditugu (hierarkia bakoitzeko lexikoi bat). Terminoak esleituta duen etiketaren arabera dagokion lexikoia aplikatuko zaio 6.5 irudian 1. eta 10. lerroen arteko kodean ikusten dugun moduan. Horiek euskaratu ostean, hierarkia adierazten duten etiketak ezabatzen ditugu 11. lerroan. Hamabigarren lerroan erregelak konbinatzen ditugu, eta azkenik 13. lerroan, konbinatutako erregelak termino habiaratu guztiei aplikatzen dizkiegu. Kontuan izan, termino habiaratu konplexuak (hitz batekoak baino gehiagokoak), azpimarraren bidez multzokatuta egon behar dutela Fomak ondo funtziona dezan. Hortaz, uriunearen bidez bereizitako elementuak, termino ezberdinak kontsideratuko dira.

```
1  define IDIS HDIS "|DIS" ;
2  define IFIN HFIN "|FIN" ;
3  define IEPO HEPO2 "|EPO" ;
4  define IBOD HBOD "|BOD" ;
5  define IPROC HPROC "|PROC" ;
6  define IBEST HBEST "|BES" ;
7  define IPHAR HPHAR "|PHAR" ;
8  define IOBV HOBV "|OBV" ;
9  define IQUA HQUA "|QUA" ;
10 define ITZULE  [ IDIS | IEPO | IBOD | IPROC | IBEST | IPHAR | IFIN |
       IOBV | IQUA] (ETIKETAK);
11 define CLEANUP [ "|BOD" | "|EPO" | "|DIS" | "|FIN" | "|BES" | "|PROC" |
       "|PHAR" | "|OBV" | "|QUA" ] -> 0 ;
12 define ITZUL  ITZULE .o. CLEANUP ;
13 regex ITZUL [" " ITZUL]* ;
```

6.5 irudia – KabiTermen termino habiaratuen euskaratze-transduktorearen patroiak.

Hurrengo taulan (6.7), aurreko transduktoreen irteera (identifikazioa eta etiketatzea) eta urrats honetako irteera erakusten dugu: termino habiara-

tuen euskaratzea. Kontuan izan, hirugarren urratsa ere tartean gauzatu dugula (elementuen berrantolaketa). Taularen laugarren adibidea begiratzen badugu, ordenan aldaketa egon dela ikus dezakegu, hirugarren urratsaren eraginez. Foma erabiliz elementuen arteko ordena aldatzea oso konplexua da, eta hortaz, lehenago erakutsi dugun moduan, transduktoreetatik kanpo kudeatzen dugu termino habiaratuen berrantolaketa.

	2. urratsaren irteera	**4. urratsaren irteera**
1	Down|EPO+-+ReM syndrome|DIS	Down+-+ReM sindrome
	Down|EPO syndrome|DIS	Down sindrome
2	head|BOD+areM structure|BES	buru+areM egitura
3	heroin|PHAR+ak_eragindako overdose|BES	heroina+ak_eragindako gaindosi
4	fracture|DIS hip|BOD+ko &LehenaAzkenera	aldaka+ko haustura
	fracture|DIS hip|BOD+areM &LehenaAzkenera	aldaka+areM haustura
5	benign|QUA neoplasm|DIS	onbera&&&ADJK neoplasia

6.7 taula – KabiTermen termino habiaratuen euskaratze-transduktorearen irteeraren adibide batzuk.

Termino habiaratuei deklinabidea gehitzeko transduktorea

Azkenik, seigarren eta azken urratsean, deklinabideen etiketak baliatzen ditugu euskarazko termino konplexua osatzeko. Transduktore honen deklinabide erregelak Xuxen zuzentzaile ortografikoaren (Agirre *et al.*, 1992) transduktoreetatik jaso ditugu, euskarazko morfologia arauak errespeta daitezen. Adibidez, +Eko marka, lekuzko genitibo mugagabea sortzeko erabiltzen da. Aurreko adibidean bezala, "hezur" hitzarekin konbinatzean ("hezur+Eko"), "r" gogortu egiten da, eta "e" formaren barruan gelditzen da, "hezurreko" forma lortuaz. Aldiz, "birika" hitzarekin konbinatzean ("birika+Eko"), lema "a" letraz bukatzen denez, lekuzko genitiboaren "e" desagertzen da "birikako" forma osatuz.

Kasu honetan ere, ez dugu transduktorearen barruan kudeatu adjektiboen ordenaren afera, elementuen berrantolaketan esan bezala, Foman konplexua delako, eta horregatik bosgarren urratsean izenondoen berrantolaketaz arduratu gara. Gogoan izan, euskaraz, izenondoak izenaren ondoren kokatzen dira, eta izenlagunak izenaren aurretik. Bosgarren adibideari jarraiki (6.8 taula), *benign neoplasm* terminoari dagokion adibideari, 2. faseak

"onbera" termino habiaratua izenondo etiketarekin sortu du (&&&ADJK), eta aplikazio nagusiaren bitartez, "onbera" izenaren ostera pasatzen dugu.

	4. urratsaren irteera	**6. urratsaren irteera**
1	Down+-+ReM sindrome	Down-en sindrome
	Down sindrome	Down sindrome
2	buru+areM egitura	buruaren egitura
3	heroina+ak_eragindako gaindosi	heroinak_eragindako gaindosi
4	aldaka+ko haustura	aldakako haustura
	aldaka+areM haustura	aldakaren haustura
5	onbera&&&ADJK neoplasia	neoplasia onbera

6.8 taula – KabiTermen termino habiaratuei deklinabidea gehitzeko fasearen adibide batzuk.

Euskaratze-patroien zehaztasun batzuk

Patroiak ahalik eta zabalenak izan daitezen saiatu gara, orokortzea posible izan den heinean. Hau da, kasu batzuetan patroia hierarkia oso bateko termino guztiei zabaldu nahi izan badiegu ere, ez da posible izan, hierarkia bereko termino guztiak ez baitira berdin euskaratzen. Hona hemen kontuan izan ditugun fenomeno batzuk:

- Preposizioen euskaratze anitza: ingelesezko preposizio asko kalifikatzaileen hierarkian daude, baina preposizioak ez dira denak berdin euskaratzen. Adibidez, *with* normalean soziatibo deklinabidearekin ordezkatzen da, *on* inesiboarekin, *to* helburuzkoarekin, etab. Gure erregeletan, deklinabide-markak ez dira lexikoi bidez gehitzen, eta erregela zehatzek gehitzen dituzte beharrezkoak direnetan. Deklinabideak lexikoietara eramateak, gure sisteman aldaketa sakonak ekarriko lituzkenez, baztertzea erabaki dugu. Hala ere, egitura zehatz batean agerpen asko dituzten preposizioen kasuan, erregela zehatzak sortu ditugu, hala nola [PROCEDURE]+to+[BODYSTRUCTURE], [PROCEDURE]+on+[BODYSTRUCTURE] edo [DISORDER]+with+[DISORDER].

- Adjektiboen ordena: Kalifikatzaileen hierarkiaren kasuan, hasiera batean orokortzea ezinezkoa zela iruditu zitzaigun, adjektiboek euskaraz

ordenan aldaerak dituztelako. Hala ere, preposizioen kasuan ezinezkoa bada ere, gainerako terminoekin orokortu ahal izan dugu, izenondo eta izenlagun ezberdintasunari esker. Izan ere, hierarkia horretako hitz gehienak adjektiboak dira, eta euskaraz, adjektiboen izenarekiko ordena ez da beti zurruna. Aurretik ere aipatu dugun moduan, izenondoa bada, izenaren ostera mugitzen dugu, eta bestelakoetan dagoen bezala uzten dugu, ingelesezko terminoaren ordena berdinean. [QUALIFIER]+[DISORDER] egirura duten *benign neoplasm* eta *congenital cyst* adibideen kasuan, euskaraz "neoplasia onbera" eta "jaiotzetiko kiste" dira ordainak; "onbera" izenondoa izanik izenaren ostera mugitu dugu, eta "jaiotzetiko" izenlaguna izatean izenaren aurrean utzi dugu. Hori kontuan izanik, adjektiboen kasuan, izenondo ala izenlagun informazioa lexikoietan gorde dugu.

- Genitiboa izenetan: euskaratze-patroi batzuek termino habiaratu bati genitiboaren deklinabidea gehitzea eskatzen dute. Kasu horietan, termino hori izena dela ziurtatu behar izan dugu, gainerako kasuetan ez baitu genitiboaren marka onartzen. Izena ez den kasuetan, terminoa bere horretan utzi dugu, deklinabide markarik gabe. Adibidez, ingelesezko *hypertrophic rhinitis* terminoan, *hypertrophic* gorputz-egituren hierarkiakoa da (anomalia morfologikoak hierarkia honetan kokatzen dira), eta *rhinitis* nahasmendu bat. BODYSTRUCTURE+DISORDER egiturarako euskarazko NAHASMENDU+aren edo +ko GORPUTZ-EGITURA egitura baliokidea sortzen dugu. Adibideari jarraituz, "errinitis hipertrofikoko" edo "errinitis hipertrofikoaren" sortuko genuke, ordain akatsdunak izanik. Izan ere, patroi mota horiek (genitiboa eskatzen dutenak), izenetan betetzen dira, eta ez adjektiboetan.

- Termino habiaratuak pluralean: sistemaren estaldura hobetzeko, terminoen formak lexikoietan ez aurkitzean, horien lemak bilatu ditugu. Izan ere, AnaMedek estrategia berdina erabiltzen du terminoen identifikazioan. Forma pluralean badago, transduktoreei terminoaren forma singularra pasatzen diegu, lexikoietan alferrikako bikoizketak ekiditeko. Horrela, hirugarren faseko transduktoreei deitu aurretik, pluralean zegoen termino habiaratuaren ordainari pluralaren marka gehitzen diogu, deklinabidea ondo gehitu ahal izateko. Adibidez, 6.3 irudiko adibidean, *nasal bones* terminoa ez da SNOMED CTn agertzen, baina bere forma singularra (*nasal bone*) bai, adibidean ikusi dugun moduan, AnaMedi

esker terminoa singularrean mantendu dugu, deklinabide markak gehitu behar izan ditugun arte.

- Euskaratzeak mugatzea: sistemak duen arazo nagusia gainsorkuntzarena da. Gainsorkuntzaren arazoari aurre egiteko, hiztegiak mugatu ditugu. Horretarako, zerrenda beltzak sortu ditugu, garapenean identifikatutako ordain okerrak kenduz, esperientziaren bidez guk geuk ikasitakoa sistemari erakutsita, gure sistema egunetik egunera eraginkorragoa bilakatuz. Adibidez, *head* terminoak, anatomiako hiztegian hamabi ordain ditu, "buru", "humeroaren buru", "falangearen buru", ... dira horietako batzuk. Ikusten dugunez, *head* terminoak adiera horiek guztiak izan ditzakeen arren, ohikoena "buru" ordain orokorragoari dagokion adiera da, eta aurretik izango duen gorputz-egituraren bidez zehaztasuna gehituko zaio. Hori kontuan izanik, "buru" kenduta gainerako adierak zerrenda beltzean gehitu ditugu, gainsorkuntza kontrolatzeko asmoz. Lexikoietan gehitutako 4 ordain baino gehiagoko pareak eskuz berrikusi ditugu. Berrikuspen horri esker, gehienez bi ordain utzi dizkiegu termino horiei. Adibidez, *"head of seventh rib structure"* terminorako, gainsorkutza kontrolatu gabe, KabiTermek 360 ordain sortzen ditu, eta gainsorkuntza kontrolatzeko egindako aldaketekin, 4 ordain baino ez. Agerikoa denez, 360 ordain sortzeak ez du inolako onurarik ekartzen, nahiz eta ordainen artean zuzenak diren hainbat egon.

- Zenbaki ordinalak eta letra bakarreko terminoak: zenbaki ordinalak adierazten dituzten formek (*seventh*, adibidez), hiru adiera izan ditzakete: zenbaki ordinala bera (zazpigarren), zenbakia bera (zazpi) eta zatikia (zazpiren). SNOMED CTren terminoetan, gehienetan zenbaki ordinalari egiten zaio erreferentzia, eta hortaz, zenbaki ordinalaren ordaina baino ez dugu hiztegiratu. Gainera, letra bakarreko terminoei ez diegu ordaina bilatu, eta bere horretan utzi ditugu. Izan ere, SNOMED CTn orokorrean letra bakarrak multzoak edo motak adierazteko erabiltzen dira (*type C thymoma* adibidez), eta ez nota musikal edo bestelako ordain bezala (*C* Do nota adierazteko ere erabiltzen da).

KabiTerm sistema aurkeztu ostean, MatxinMeden aurkezpena egingo dugu jarraian. MatxinMed Matxinen (Mayor *et al.*, 2011) osasun-zientzietarako egokitzapena da, SNOMED CTren terminoekin elikatua dagoena.

6.2 Matxinen egokitzapena medikuntzaren domeinura

Atal honetan algoritmoaren azkeneko urratsa aurkeztuko dugu. Laugarren eta azken urrats honetan, Matxin deituriko Itzultzaile Automatikoa egokitu dugu osasun-zientzien domeinura. Aurrekarien atalean (6.2.1 atala) Itzultzaile Automatikoen inguruan zehaztasun batzuk emango ditugu, eta gaur egunean ingelesa-euskara hizkuntza-parerako eskura dauden Itzultzaile Automatikoak aurkeztuko ditugu. Ostean, 6.2.2 atalean, Matxin sistemari egin diogun egokitzapena azalduko dugu.

6.2.1 Aurrekariak

Itzulpen Automatikoaren helburua ordenagailuen bidez hizkuntza batetik bestera itzulpenak egitea da. Hizkuntzaren Prozesamenduko (HP) tekniken aplikazio izarretakoa izan da hasieretatik, eta gaur egunean ere interes handia sortzen duen gaia da, mundu globalizatu honetan elkar ulertzeko nahia dela-eta.

Itzulpen Automatikoari aurre egiteko hurbilpenak bi multzo nagusitan sailkatuak izan dira: Erregeletan Oinarritutako Itzulpen Automatikoa (EOIA) eta Corpusetan Oinarritutako Itzulpen Automatikoa (COIA). Erregeletan oinarritutakoek, hizkuntzekiko dugun jakintza linguistikoa erabiltzen dute oinarri gisa, eta corpusetan oinarritutakoek aurretik egindako itzulpenak oinarri hartuta, hurbilpen enpirikoak garatzen dituzte.

Atal honen garapenean, unean uneko erreferentziak aipatuko ditugu, baina orokorrean erabili ditugun erreferentziazko lanak Jurafsky eta Martin (2008), Mayor (2007) eta Artetxe (2016) izan dira.

Hurrengo lerroetan Itzulpen Automatikoaren paradigma ezberdinak aurkeztuko ditugu eta, zehazki, ingelesa-euskara hizkuntza parerako dauden sistemak ikusiko ditugu.

Erregeletan Oinarritutako Itzulpen Automatikoa

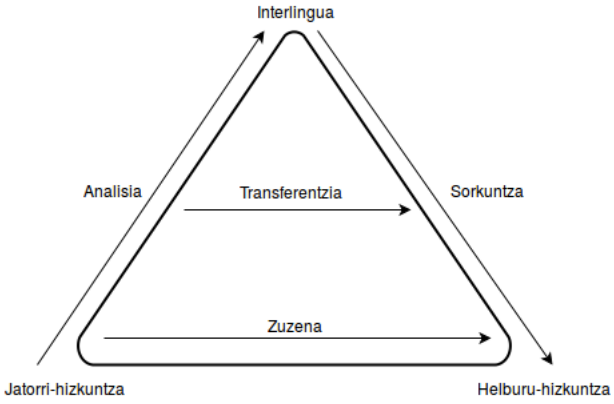
Erregeletan Oinarritutako Itzulpen Automatikoak jatorri-hizkuntzaren eta helburu-hizkuntzaren jakintza linguistikoa erabiltzen du itzulpena egiteko. Itzulpen-prozesua hiru fasetan banatzen da: analisia, transferentzia eta sorkuntza.

Analisian, itzuli beharreko testua ohiko hizkuntzaren prozesamenduko katea erabiliz analizatzen da, normalean analizatzaile morfologikoa, analizatzaile gramatikala (*POS tagger*) eta analizatzaile sintaktikoa barnebiltzen dituena. Analisi horren ondorioz, tarteko errepresentazio bat sortzen da.

Transferentzian, jatorri-hizkuntzaren tarteko errepresentazioa transferitu egiten da helburu-hizkuntzaren tarteko errepresentaziora. Bi motako transferentziak daude: transferentzia sintaktikoa eta transferentzia semantikoa. Transferentzia sintaktikoek lexiko eta egitura mailan lan egiten dute hiztegi elebidunak eta transferentzia-erregelak erabiliz. Transferentzia semantikoek, semantika maila gehitzen diote transferentziari, esanahia adierazteko egitura gehigarriak erabiliz.

Azkenik sorkuntzan, helburu-hizkuntzako tarteko errepresentaziotik itzulpena lortzen da, askotan hiztegi morfologikoak baliatuz.

Hiru motako sistemak bereizten dira erabilitako informazioan eta abstrakzio mailan oinarrituta (Hutchins eta Somers, 1992). Estrategia horiek 6.6 irudian erakusten dugun Vauquois-en triangeluaren bidez adierazi ohi dira: itzulpen zuzenak, transferentzian oinarritutako sistemak eta interlingua bidezko sistemak.



6.6 irudia – Vauquois-en triangelua.

Vauquois-en triangeluaren mutur batean dauden itzulpen zuzenak ez dute analisirik erabiltzen eta beste muturrean dauden interlingua sistemek ez dute transferentziaren beharrik. Triangeluaren gailurrera hurbiltzen garen heinean, tarteko errepresentazioa konplikatuz doa, eta horrela, analisi sakonagoa behar da.

Corpusean oinarritutako itzulpen automatikoa

Corpusean Oinarritutako Itzulpen Automatikoa (COIA) eskuz egindako itzulpenez baliatzen da eta metodo enpirikoak erabiltzen ditu berauek ustiatzeko. Azken urteetan ordenagailuek izan duten garapenari eta Interneten topa daitekeen testu kopuru ikaragarriari esker, corpusean oinarritutako teknikak asko garatu dira. Itzulpen Automatikoari dagokionean, corpus paraleloak[8] dira baliabide nagusia.

Bi motatako COIAk daude, Adibideetan Oinarritutako Itzulpen Automatikoa (AOIA, ingelesez *Example Based Machine Translation* edo EBMT) eta Itzulpen Automatiko Estatistikoa (IAE, ingelesez *Statistical MT* edo SMT). AOIAk itzulpen-prozesuan esaldi-zatiak corpuseko adibideekin parekatzen ditu, horien itzulpen-zatiak identifikatzen ditu eta azkenik zati horiek birkonbinatzen ditu itzulpena sortzeko. IAEek aldiz, corpus paralelo handietatik erauzitako datu estatistikoak erabilita sortzen dituzte itzulpenak, horretarako informazio linguistiko espliziturik erabili gabe.

IAEak izan dira azken urteetan gehien garatu diren sistemak ikerketaren munduan. Hiru motatako sistema estatistikoak bereizten dira: hitzetan oinarritutakoak, hitz-segidetan oinarritutakoak (ingelesezko *phrase-base*) eta hitz-segida hierarkikoetan oinarritutakoak (ingelesez *hierarchical phrase-base*). Hasiera batean hitzetan oinarritutakoak erabiltzen zirenak baziren ere, gaur egunean sistema gehienak hitz-segidetan oinarritutakoak dira[9]. Hala ere, azken urteetan ikasketa sakonaren (*deep learning* ingelesez) agerpen arrakastatsuari esker, tradiziozko IAEak Itzulpen Automatiko Neuronalengatik ordezkatuak izaten ari dira (Sennrich *et al.*, 2016).

Hitz-segidetan oinarritutako IAEetan itzulpen-prozesuan bi eredu sortzen dira: itzulpen-eredua bera, eta hizkuntza-eredua. Itzulpen-eredua sortzeko corpus elebiduna erabiltzen den bitartean, hizkuntza-eredua sortzeko helburu-hizkuntzaren corpus elebakarra erabiltzen da.

Itzulpen-ereduak probabilitatea esleitzen dio jatorri-hizkuntzako esaldiari, helburu-hizkuntzako esaldi batera itzultzen denean. Helburu horrekin, hitz-segidetan oinarritutako ereduek jatorri eta helburu esaldiak parekatutako segmentuetan banatzen dituzte (informazio linguistikoa kontuan hartu gabe) eta parekatze horien probabilitateak kalkulatzen dituzte. Probabilita-

[8]Corpus paraleloetan, testu paraleloak segmentu mailan parekatuak daude. Orokorrean esaldi mailako parekatzea egiten da.

[9]Ingelesez *phrase* hitzak sintagmak adierazten baditu ere, hauek ez dira sintagma linguistikoak, hitz-segidak baizik, eta horregatik guk hitz-segidetan oinarritakoak deritzegu.

teen kalkuluen inguruan informazio gehiago Koehn *et al.* (2003) artikuluan jaso daiteke.

Hizkuntza-ereduak aldiz, helburu-hizkuntzan gertatzen diren hitz-segidei probabilitatea esleitzen die. Eredu asko proposatu badira ere, ohikoenak n-grametan oinarritutakoak dira. Horrela, hitz-segida baten probabilitatea hitz bakoitza aurreko hitzekin agertzeko duen probabilitatearekin lortzen da.

Sistema estatistikoen artean, Moses (Koehn *et al.* 2007) da zabalduenetakoa. Moses Itzulpen Automatiko Estatistikorako kode irekiko tresna da, IAEak entrenatu eta doitzeko (*tuning*) aukera ematen duena. Sistema estatistikoak garatzea asko errazten du Moses sistemak, eta sistema funtzional bat lortzeko beharrezko denbora minimizatzen du.

Corpusean oinarritutako itzulpen automatikoen kalitatea, corpus paraleloen tamainarekin hertsiki lotuta dago. Are nabariagoa izaten da hori hizkuntza parearen ezaugarriak oso ezberdinak direnean (urrutiko hizkuntzak), euskara eta ingelesaren artean, adibidez. Kontuan izan, euskararen eta ingelesaren gramatikak oso ezberdinak direla. Gainera, ingelesa nagusiki hizkuntza analitikoa da, eta euskara hizkuntza eranskaria, morfemak elkartuz hitzak sortzen dituelarik. Alde horiek itzulpengintzarako erronka handia dira, hitzen arteko parekatzea ez delako zuzenekoa eta hitzen hurrenkeran aldaketak egin behar direlako, besteak beste.

Ingelesa-euskara parerako Itzultzaile Automatikoak

Gaur egunean ingelesa-euskara parerako aurkitzen ditugun sistema gehienak erregeletan oinarritutako sistemak dira, hizkuntzen urruntasunak eta corpus-paralelo erraldoien faltak baldintzatuta ziurrenik. Izan ere, ikusi dugun moduan, hizkuntza urrunak direnez, corpus are handiagoa beharrezkoa da itzulpenaren kalitatea ziurtatzeko.

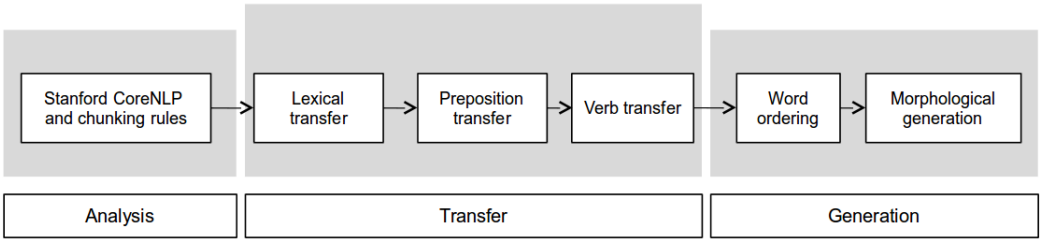
Hurrengo lerroetan, ingelesa-euskara hizkuntza parerako eskuragarri dauden itzultzaile automatikoak deskribatuko ditugu:

- **Matxin** (Mayor *et al.*, 2011) IXA taldean garaturiko sistema librea da. Jatorrian gaztelania-euskara hizkuntza parerako garatu bazen ere, ingelesa-euskara parerako ere egokitua izan da (Aranberri *et al.*, 2015). Erregeletan oinarritutako itzulpen-sistema da, kode irekikoa. Transferentzian oinarritutako sistemen arkitektura klasikoa jarraitzen du, hiru modulu nagusi barnebiltzen dituena: jatorri-hizkuntzaren analisia, jatorritik helbururako transferentzia eta helburu-hizkuntzaren sorkuntza (6.7 irudia).

Analisiaren moduluan, Stanforden CoreNLP tresna erabiltzen da ingelesezko analisia egiteko. Analisitik honako informazioa jasotzen da: hitzen informazioa (kategoria gramatikala eta flexio morfologikoa), zatiak (zatien arteko dependentzia-harremanak, ingelesez *chunk* deritzo) eta esaldi mota.

Transferentziaren moduluan, bi motatako informazioa kudeatzen da: ezagutza lexikala eta egiturazkoa. Transferentzia lexikala hiztegien bitartez lemen ordain egokiak eskuratzeaz arduratzen da, eta egiturazko transferentzia, ezaugarri morfosintaktikoez arduratzen da, zati eta hitz egokiei esleituz.

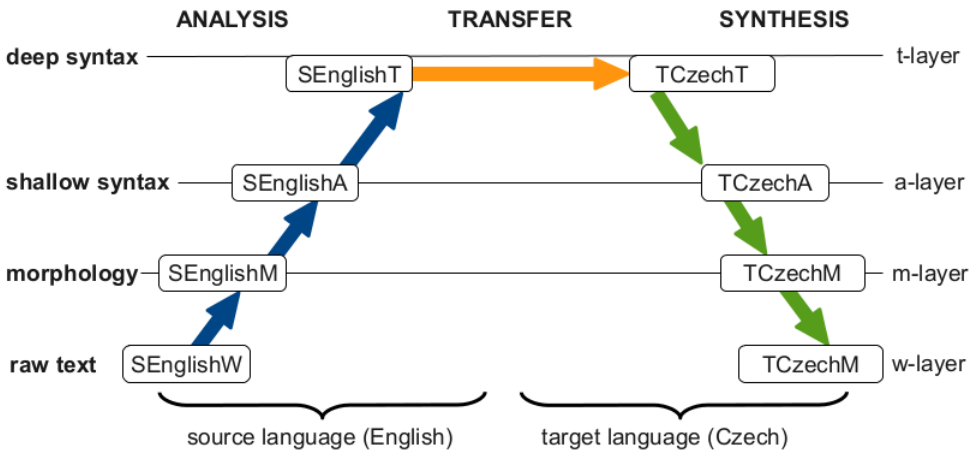
Sorkuntza modulua aldiz, bi urrats nagusitan banatuta dago. Lehenengoa zatien barruko hitzen ordenaz arduratzen da, baita zatien arteko ordenaz ere. Horretaz gain, zati mailan jasotako informazioa, flexioa jaso behar duen hitzari esleitzen zaio (euskararen kasuan zatiaren azken elementuari). Bigarren urratsean, sorkuntza morfologikoa egiten da, hiztegi morfologiko baten laguntzaz (kasu honetan Euskararen Datu-Base Lexikala, EDBL (Aldezabal *et al.*, 2001) erabiltzen da), eta jatorrizko lemetatik forma egokia sortzen da.



6.7 irudia – Matxinen arkitektura orokorra. Irudia, Aranberri (2016) artikulutik jaso dugu.

- **TectoMT** erregeletan oinarritutako itzulpen-sistema da, modulartasun handia duena (Popel eta Žabokrtskỳ, 2010). Sistema, sintaxian oinarritzen da transferentzia egiteko garaian, eta Matxinek baino analisi sakonagoa egiten du, abstrakzio maila altuagoa lortuz. Horretarako, tektogramatikaz baliatzen da (Hajicová, 2000), dependentzia-zuhaitz sintaktiko sakonen bidez hizkuntza errepresentatzen duena. Erregeletan oinarritutako sistema bada ere, teknika estatistikoak erabiltzen ditu

itzulpen-prozesuaren hainbat modulutan. Jatorrian ingelesetik txekierara itzultzeko sistema bada ere, berriki sistema ingelesa-euskara pareraako egokitu da QTleap[10] proiektuaren baitan (Aranberri *et al.*, 2016b).



6.8 irudia – TektoMTren arkitektura orokorra. Irudia Popel eta Žabokrtskỳ (2010) artikulutik jaso dugu.

- **Google** enpresaren Google Translate dohaineko itzulpen-zerbitzua aski ezaguna da mundu osoan zehar. 2001 urtean argitaratu zutenetik 2005-2006 urte ingurura arte, Google Translate Systran deituriko erregeletan oinarritutako sisteman oinarritzen zen. Hasiera horietan, ingelesaren eta beste zortzi hizkuntzen arteko itzulpenak egiten zituen. 2005 urtetik aurrera, sistema estatistikoak erabiltzen hasi ziren, gerora hizkuntza guztietara zabaltzeko eta urte bereko *NIST DARPA TIDES Machine Translation Evaluation* izeneko txapelketan lehenengo postua eskuratu zuten arabiera-ingelesa eta txinera-ingelesa sistema estatistikoekin[11]. Estatistikan oinarritutako itzulpen-sistemek Europar Batasunaren eta Nazio Batuen dokumentazio paraleloa erabiltzen dute eta baita webetik erauzitako datu paraleloak ere.

  2010etik aurrera, euskararako *alpha* bertsioa argitaratu zuten, eta gaur egunean 90 hizkuntzetarako eskaintzen dituzte itzulpen-sistemak. Goo-

[10] http://qtleap.eu (2017ko maiatzaren 9an atzitua).

[11] NISTetik jasota http://www.itl.nist.gov/iad/mig//tests/mt/2005/doc/mt05eval _official_results_release_20050801_v3.html (2017ko maiatzaren 9an atzitua).

gle Translate tresnak, corpus paralelo txikia duten hizkuntza pareetarako, ingelesa *pibote* hizkuntza gisa erabiltzen du itzulpenak egiteko. Hau da, jatorri-hizkuntzatik, aurrena ingelesera itzultzen du, eta ostean ingelesetik helburu-hizkuntzara.

Oso zehaztasun gutxi ezagutzen dira Google Translate sistemaren inguruan, eta garatzaileek azkeneko berriak eta egitura orokorrari buruzko informazio minimoa baino ez dute argitaratzen. Horrela, sistema dohainekoa izan arren, kodea ez dago eskura egokitzapenak egin ahal izateko.

- **Lucy** sistema Eusko Jaurlaritzaren proiektu bati esker gaztelania-euskara eta beranduago ingelesa-euskara pareetarako garatu zen. Oso zaila da itzultzaile komertzial honi buruzko informazioa lortzea. Komertziala den heinean, ez dute sistemaren zehaztasunik eman inongo ikerketa kongresutan edo aldizkaritan. Hala baina, sarean aurkitutako aurkezpen baten diapositibetan (Gieselmann, 2008) eta Helduen Alfabetatze eta Berreuskalduntzerako Erakundearen (HABE) web-orrian lortutako informazioari[12] esker, sistemaren datu orokor batzuk jakin ditzakegu: i) erregeletan oinarritzen den sistema hau, analisi morfologiko zein sintaktikoan oinarritzen da; ii) transferentzia lexikala eta egiturazkoa egiteaz gain, testuinguruaren transferentzia ere egiten du; eta iii) sorkuntza morfologikoaz gain, helburu-hizkuntzaren araberako prozesuak ere egiten ditu.

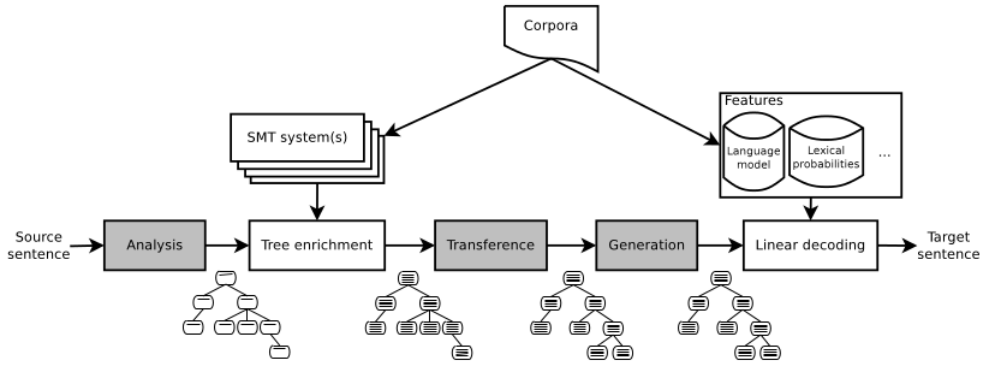
- **EuSMT** IXA taldean garatutako sistema estatistikoa da. Mosesen oinarrituta, bi hurbilpen garatu dira ingelesa-euskara hizkuntza parerako. Lehenengo hurbilpena, oinarri-lerroa deiturikoa (**EuSMTo**), Mosesen bidez sortutako hitz-segidetan oinarritutako IAE sistema estandarra da. Entrenamenduan erabilitako corpus-paraleloaren % 85a Elhuyarren itzulpen-memorietatik erauzia izan da eta gainerako % 15a webetik automatikoki erauzia izan da PaCo2 (San Vicente eta Manterola, 2012) tresna erabiliz.

  Bigarren hurbilpenean, segmentazioa gehitu zaio sistemari (**EuSMTs**). IAE sistemek hizkuntza antzekoen artean funtzionatzen dute hobeto, hau da, ezaugarri gramatikal antzekoak dituztenen artean. Kasu honetan, ingelesa hizkuntza nagusiki analitikoa da, morfema bakoitzerako

[12] http://www.habe.euskadi.eus/s23-4728/es/contenidos/noticia/tzulpenautomatikoa_mintegia16/es_def/index.shtml (2017ko maiatzaren 9an atzitua).

hitz bat duelarik eta euskara aldiz, eranskaria da, morfemak elkartuz hitzak sortzen dituelarik. Bi hizkuntzen arteko loturan laguntzeko segmentazioa erabiltzen da (Al-Haj eta Lavie, 2012; Naradowsky eta Toutanova, 2011). Segmentazioaren bidez, hizkuntza eranskarien hitzak morfemetan banatzen dira, eta horrela hizkuntza analitikoarekin parekatzea errazten da.

- **SMatxinT** sistema EuSMTren bi hurbilpenak eta Matxin hibridatuz IXA taldean sortutako sistema da, España Bonet *et al.* (2011) eta Labaka *et al.* (2014) jarraituz. Egileek azaltzen duten bezala, orokorrean EOIA sistemak berrantolaketa sintaktikoan hobeak dira eta COIA sistemek lexikoaren aukeraketa hobea egiten dute. Oinarri horrekin sortu zuten SMatxinT sistema hibridoa (6.9 irudia).



6.9 irudia – SMatxinTren arkitektura orokorra non EOIA moduluak grises nabarmenduak agertzen diren. Irudia España Bonet *et al.* (2011) artikulutik jaso dugu.

Amaitzeko, ingelesa-euskara hizkuntza parerako azaldu ditugun sistemen laburpena erakusten dugu 6.9 taulan.

6.2.2 MatxinMed: sistemaren egokitzapena

Azken urteetan itzultzaile automatikoak domeinu zehatzera egokitzeko joera zabaldu da, frogatu baita domeinura mugatzean, domeinu horretako itzulpenen kalitatea hobetzen dela. Erregeletan oinarritutako sistemak domeinu zehatzetara egokitzeko, hiztegiak eta gramatikak egokitu beharko liratekeen

	EOIA	COIA
Matxin	✓	
TectoMT	✓	(✓)
Google		✓
Lucy	✓	
EuSMT		✓
SMatxinT	✓	✓

6.9 taula – Ingelesa-euskara hizkuntza parerako sistemen laburpena, non EOIAk Erregeletan Oinarritzen diren Itzultzaile Automatikoak diren eta COIAk Corpusetan Oinarritzen diren Itzultzaile Automatikoak diren.

arren, horretarako eskuzko lan handia egin behar denez, ohikoa da egokitzapena hiztegiak zabaltzera mugatzea (Weijnitz *et al.*, 2004). IAEak aldiz, domeinu zehatzetara egokitzeko, domeinu horretako corpus-paraleloak behar dira.

Matxin aukeratu dugu osasun-zientzien domeinura egokitzeko sistema moduan, ingelesa-euskara parerako ditugun baliabideak kontuan izanik, erregeletan oinarritutako sistema bat baita gure ustez bideragarriena. Euskararako osasun-arloko corpus-paraleloen urritasunak Itzultzaile Automatiko Estatistiko bat egokitzea oraingoz oso zaila egiten du. Gainera, Matxin IXA taldean bertan garatua izanik, egokitzapenean lagundu ahal gaitun ikerlari-taldea hurbil dugu. Horretaz gain, hiztegia modu errazean zabal daiteke, eta inbertsio txikiarekin, sistema domeinura egokitu dezakegu.

Atal honetan beraz, Matxin itzultzaile automatikoaren osasun-zientzien domeinurako egokitzapena, MatxinMed, azalduko dugu.

Gogora ekarri behar dugu, EuSnomeden laugarren urratsa dugula hau, eta aurreko urratsetan garatu ditugun teknikek kale egiten dutenean erabiliko dugula MatxinMed. Kontuan izanik Itzultzaile Automatikoak esaldiak itzultzeko diseinatuta daudela, terminoen euskaratzean ez ditugu kalitate altuko emaitzak espero, baina jatorrizko terminoak euskaratu gabe uztea baino irtenbide hobea iruditzen zaigu. Horrela, SNOMED CTren euskarazko bertsioaz arduratuko diren adituei erraztasunak eman nahi dizkiegu, euskarazko proposamenetatik abiatuko baitira terminoak aukeratzerako garaian, eta ez hutsetik.

Hurrengo lerroetan, Matxini egindako egokitzapenak azalduko ditugu, MatxinMed sortzeko egindako aldaketak, hain zuzen ere. Egokitzapen ho-

riek, Matxinen transferentzia lexilakaren moduluan egin ditugu. Lehenik eta behin, Matxinen domeinua aukeratzeko funtzionalitatea eta hiztegiaren zabaltzea azalduko dugu. Bigarrenik, NeoTerm sistemaren integrazioaren berri emango dugu. Jarraian, ordainen arteko ordena zehazteko erabilitako hizkuntza-eredua aurkeztuko dugu. Azkenik, termino konplexuak ezagutzeko erregelen berri emango dugu.

Domeinuen aukeraketarako funtzioa eta hiztegiaren zabaltzea

Aurreko atalean azaldu bezala, Matxin Erregeletan Oinarritutako Itzultzaile Automatikoa da. Lexikoa aukeratzeko baliabide nagusia hiztegia da, Matxinek formatu zehatz batean gordeta duena. Hiztegi horrek, transferentzia lexikoa egiteko beharrezko informazio guztia jasotzen du, hitzen itzulpenak eta hitzaren kategoria gramatikala, besteak beste.

Matxinek hiztegia XMLz gordetzen du, eta hiztegi horren sarrera batzuen informazioa erauzi dugu 6.10 taulan. Taulako bigarren lerroan, toki-izen bat dago, *croatia*, eta horrekin batera euskarazko ordaina (Kroazia) zenbatgarren adiera den eta kategoria gramatikala ("NP_IZE_LIB", "IZE" etiketak izena adierazten du, eta "LIB" etiketak leku-izen berezia dela). Hirugarren lerroan *croatian* adjektiboa eta bere informazioa agertzen da ("NP_ADJ_IZO" etiketarekin, "ADJ" etiketak adjektiboa adierazten du, eta "IZO" etiketak izenondoa dela) eta seigarrenean "cross" izen arrunta ("NP_IZE_ARR", "ARR" etiketak izen arrunta dela adierazten du).

Ingelesa	**Euskara**	**Adiera**	**Kategoria**	**Domeinua**
colon	bi_puntu	1	NN_IZE_ARR	
croatia	Kroazia	1	NN_IZE_LIB	
croatian	kroaziar	1	NN_ADJ_IZO	
cross	gurutze	1	NN_IZE_ARR	
crown	koroa	1	NN_IZE_ARR	Med

6.10 taula – Matxinen hiztegiaren sarrera batzuen informazioa.

Hiztegi hori aberasteko orduan eta osasun-zientzietako terminoak sartzerakoan, modu errazean bereiz daitezen, hiztegi-sarrerei ezaugarri berri bat gehitu diegu domeinua adierazten duena, XMLean `dom` etiketaren bitartez. Ezaugarri horretan "Med" etiketa erabili dugu MatxinMed sortzeko. Adibidean ikusten dugunez (6.10 taula), zazpigarren lerroan *crown* izen arrunta,

domeinu orokorrekoa izateaz gain, osasun-zientzien domeinukoa ere denez (hagineko koroa, adibidez), domeinuko etiketa gehitu diogu.

Matxinen hiztegiko sarrerak ataletan edo *section*-etan multzokatuta daude. Guk SNOMED CTren euskaratze-algoritmoari esker euskaratu ditugun ordainak gehitu ditugu hiztegira. Horretarako, automatikoki dagoeneko hiztegiratutako sarrerak errepasatu ditugu, eta sarrera aurkitzekotan domeinuari dagokion etiketa gehitu diogu. Bestelakoetan, hiztegian atal (`section`) berri bat gehitu diogu hiztegiari, kasu honetan *medicine* deiturikoa, eta bertan parekatze berriak gehitu ditugu. Hurrengo adibidean (6.11 taula), *medicine* ataleko sarrera batzuk ikus ditzakegu, denak osasun-zientzien domeinukoak.

Ingelesa	**Euskara**	**Adiera**	**Kategoria**	**Domeinua**
colon	kolon	2	NN_IZE_ARR	Med
noradrenaline	noradrenalina	1	NN_IZE_ARR	Med
noradrenaline	norepinefrina	2	NN_IZE_ARR	Med
steatopygia	esteatopigia	1	NN_IZE_ARR	Med
sacculotomy	sakulotomia	1	NN_IZE_ARR	Med
cholangiohepatitis	kolangiohepatitis	1	NN_IZE_ARR	Med

6.11 taula – Matxinen hiztegi espezializatuaren sarrera batzuen informazioa.

Aipatu dugun moduan, hiztegia elikatuz Matxini domeinu zehatz batera egokitzeko aukera gehitu diogu. Hori horrela izanik, `sense` eta `dom` ezaugarriek berebiziko garrantzia hartzen dute diseinu berrian, bien edukiari begiratuta egiten baita hitzen edo terminoen aukeraketa.

Matxinek, hitz baten ordaina aukeratzeko, hiztegi-sarreraren `sense` ezaugarriari begiratzen dio eta balio baxuena duen parekatzea aukeratzen du ordaina lortzeko. Aukeraketa konplexuagoak ere baimentzen ditu Matxinek, testuinguruaren arabera ordain baten alde egiteko. Arau zehatzak definitu behar dira, ordea, horretarako.

Hori horrela izanik, funtzionalitate berria definitu dugu, domeinua zehaztuta, Matxinek domeinu horretako ordainak hobetsi ditzan. Aukera eman diogu nahi adina domeinu zehazteko, eta haien artean hierarkia bat sortzeko. Horretarako erabiltzaileak hierarkien arteko ordena konfiguratu beharko du. Hitzaren ordaina bilatzean, lehen domeinuan ordainik ez balego, bigarren domeinukoak bilatzen ditu, gero hirugarrenekoak, etab. Domeinuaren barruan ordain bat baino gehiago egonez gero, horien artean `sense` balio txikiena duena aukeratzen du.

Adibidez, demagun pediatriaren inguruko testu bat euskaratu nahi dugula. Kasu horretan, pediatria izango da lehenengo domeinua edo domeinu nagusia, eta medikuntza bigarren mailakoa. Horrela, aurrena pediatriako parekatzeak bilatzen ditu MatxinMedek, ez aurkitzekotan medikuntzakoak, eta azken unean parekatze guztiak izango lituzke kontuan. Nahi beste domeinu zehaztu ahalko ditugu, eta MatxinMedek lehenengotik hasita, denak aztertuko ditu ordaina aurkitu arte.

Hiztegiari erreparatzen badiogu (6.10 eta 6.11 taulak), *colon* terminoaren bi adierak ikus ditzakegu. Hiztegiaren atal orokorrean (6.10 taula), "bi puntu" gisa euskaratzen den bitartean, medikuntzaren domeinuan (6.11 taula) kolon da. Transferentzia lexikalean egindako egokitzapen honi esker, 6.12 taulako adibidean ikus dezakegun bezala, MatxinMed *"He has colon cancer"* bezalako esaldiak ondo itzultzeko gai da, transferentzia lexikala egiterakoan parekatze hobea aukeratzeko baliabideak dituelako.

Jatorrizko esaldia	*He has colon cancer.*
Matxinen itzulpena	Hark **bi puntu** minbizia dauka.
MatxinMeden itzulpena	Hark **kolon** minbizia dauka.

6.12 taula – Matxin eta MatxinMeden itzulpenen adibidea.

NeoTerm integratzeko modulua

Transferentzia lexikalari, domeinua aukeratzeko funtzionalitateaz gain, medikuntzaren domeinuan erabiltzeko transliterazioan oinarritutako modulua gehitu diogu. Modulu horren bitartez, aurreko kapituluan (5.2.1 atalean) azaldutako NeoTerm sistema Matxinen integratu dugu. Izan ere, MatxinMeden hiztegian SNOMED CTren eduki terminologikoa da gehitu duguna, baina egon litezke termino batzuk SNOMED CTn agertzen ez direnak. Horietarako, NeoTerm bidez euskarazko ordaina lortzen dugu. Azken finean, sortutako baliabideak berrerabiltzeko hautua egin dugu modulu hori integratzean.

Matxinek, hitz baten ordaina hiztegian aurkitzen ez duenean, jatorrizko hitza bere horretan uzten du, eta kasuan kasu, deklinabidea gehitzen dio. Hurrengo adibidean (6.13 taulan) ikus dezakegu Matxinek *cholangiohepatitis* terminoa ez duenez hiztegian aurkitu, bere horretan uzten duela, eta singular mugatuaren deklinabidea gehitzen diola (cholangiohepatitisa). Aldiz, hizte-

gian aurkitzen duen kasuetan, adibidean MatxinMeden hiztegian aurkitzen denez, terminoaren ordaina erabiltzen du (kolangiohepatitisa).

Jatorrizko esaldia	*He has cholangiohepatitis.*
Matxinen itzulpena	Hark **cholangiohepatitis**a dauka.
MatxinMeden itzulpena	Hark **kolangiohepatitis**a dauka.
Jatorrizko esaldia	*He has cholangio**hypo**hepatitis.*
MatxinMeden itzulpena	Hark **cholangiohypohepatitis**a dauka.
MatxinMed NeoTermekin	Hark **kolangiohipohepatitis**a dauka.

6.13 taula – NeoTermen integrazioaren adibidea.

Adibide horretan, ez bada NeoTermen beharrik ikusi ere, Matxinen funtzionamendua erakusteko balio izan digu hitz ezezagunen aurrean. Hori kontuan izanik, NeoTerm integratzearekin, ezagutzen ez ditugun termino neoklasikoak ere euskaratzen ditugu.

Adibidez, aurreko terminoari, *hypo* afixua gehitzen badiogu, *cholangiohypohepatitis* terminoa sortuz (existitu ez arren eta zentzurik ez izan arren, adibide gisa erakutsi nahi dugu), MatxinMedek NeoTerm gabe, Matxinen itzulpen berdina emango liguke, eta NeoTermen integrazioari esker, euskarazko itxura duen itzulpena sortzen du, 6.13 taularen bigarren zatian ikus dezakegun bezala.

Hizkuntza-Eredua

Ikusi dugun moduan, hainbat ingeleseko terminok euskarazko ordain bat baino gehiago dituzte (4. eta 5. lerroak 6.11 taulan, non adibidez *noradrenaline* bi modutan adierazten den euskaraz, noradrenalina eta norepinefrina). Matxinen oraingo inplementazioak, ordainen gaineko desanbiguaziorako tresnak nahikoa garatu gabe ditu, eta hortaz definitzen dugun ordenak berebiziko garrantzia dauka, lehenengo ordaina izango baita uneko bertsioak erabiliko duena. Hala ere, lehenengo ordaina bakarrik erabiliko bada ere, interesgarria iruditu zaigu ordain guztiak hiztegian gehitzea, etorkizunean desanbiguatzeko aukera garatuz gero, hiztegia osatuta izango dugulako.

SNOMED CTren ordainen arteko ordena zehazteko (`sense` etiketaren bidez adierazten dena), hizkuntza-eredu bat osatu dugu. Aurrekarien atalean ikusi dugun moduan (6.2.1 atala), hizkuntza-ereduak Estatistiketan Oinarritutako Itzultzaileetan erabiltzen diren eredu probabilistikoak dira, helburu-

-hizkuntzarako sortzen direnak. Eredu horiek testu-kate bat helburu-hizkuntzako baliozko esaldia izateko probabilitatea adierazten dute, eta helburu nagusia itzulpena helburu-hizkuntzaren ahalik eta antzekoena dela bermatzea da.

Hizkuntza-eredu bat sortzeko, helburu-hizkuntzaren corpus bat beharrezkoa da. Gure kasuan, euskarazko corpusa behar dugu eta gainera, osasun-zientzien domeinukoa. Domeinuko corpusa osatzeko, honako iturriak erabili ditugu:

- Udako Euskal Unibertsitatearen (UEU) osasun-zientzietako liburuak: UEUk euskarazko liburugintza akademikoan lan handia egin du azken hamarkadetan. Osasun-zientzien domeinuan 15 liburu argitaratu dituzte eta bertatik 300.000 token inguru erauzi ditugu.

- Euskal Herriko Unibertsitatearen medikuntzako ikasleen apunteak: medikuntza fakultateko euskal adarreko ikasleek euren apunteak biltegi batean bildu dituzte euskarazko materiala elkartuz. Bertako apunteak jaso ditugu, eta testua automatikoki erauzi dugu. 1.200.000 token inguru bildu ditugu horrela.

- Elhuyar Fundazioaren itzulpen-memoriak: Elhuyarrek urteetako esperientzia dauka liburuen eta oro har testuen euskaratzean. Prozesu horretan sortutako itzulpen-memoria horietatik azpicorpus paralelo bat erauzi dugu osasun-zientzien domeinukoa ia 1.000.000 token dituena.

- Osasungoa Euskalduntzeko Erakundearen (OEE) osasun biltzarren txostenak: OEE osasun-arloan euskara sustatu eta bultzatu nahi duen elkartea da. Elkarte honek urtero osasun biltzarra antolatzen du osasun-arloko gai ezberdinak jorratuz eta biltzar horiei esker idatzizko material asko sortu dute euskaraz. Biltzar horietako 1996-2014 urte tarterako txostenak jaso ditugu. Hala ere, urte batzuetako materiala erauzteko arazoak izan ditugu formatu kontuak direla medio. Denera 15 urtetako jardunaldietako testua gehitu diogu corpusari, 400.000 tokenetik gora lortuz.

- Osakidetzaren zabalkunderako txostenak: Euskal Autonomia Erkidegoaren osasun-zerbitzuak zabalkunderako prestatutako 41 txostenetatik 600.000 tokenetatik gora gehitu dizkiogu corpusari. Txosten horiek

> alor ezberdinetakoak dira, erizaintza, etika sanitarioa, kudeaketa sanitarioa, lan-osasuna, lehen mailako arreta edo osasun mentalari buruzkoak.

Horrela 3.500.000 token baino gehiagoko corpusa osatu dugu. Kontuan izan, corpusa automatikoki sortu dugula eta zenbait kasutan .pdf zein .docx formatua duten fitxategietatik erauzi behar izan dugula testua. Hori horrela izanik, baliteke formatu akatsen bat aurkitzea corpusean. Horretaz gain, medikuntzako ikasleen apunteen kasuan bereziki, beste hizkuntzetako testu zati gutxi batzuk aurkitu ditugu, ingelesez zein espainieraz, eta baita euskarazko ortografia akats batzuk ere. Izan ere, ikasleen apunteak dira, ez dira argitaratutako testu zainduak, eta izaera horretako testuen ezaugarriak dituzte. Corpusaren tamaina handia izanik, aurkitu ditzakegun akatsen eragina txikitzen da eta horrela, gure interesekoa den hizkuntza-eredua sortzeko baliagarria zaigu.

Kontuan hartu behar dugu erabili dugun corpusaren izaera. Osasun-zientzien domeinukoa izan arren, dokumentu akademikoak dira gehienak (liburuak eta ikasgaietako apunteak). SNOMED CTren izaera aldiz, ezberdina da, terminologia klinikoa baita nagusi, baina alor zehatz horretan dugun hutsunea betetzeko baliagarria izango zaigu, alor klinikoko corpusa sortzen den bitartean.

Hizkuntza-eredua sortzeko, corpusa aurreprozesatu behar izan dugu. Alde batetik lerro hutsak kendu ditugu eta letra xehetara pasa dugu. Bestetik, corpusa tokenizatu dugu. Ereduaren entrenamendurako Moses sistemaren eredu sortzailea erabili dugu, Mosesek hizkuntza-ereduak prestatzeko modulu bat baitauka, guk erabili duguna.

Bi hizkuntza-eredu entrenatu ditugu, hirugrametan eta bosgrametan oinarritutakoak. Bi ereduen emaitzak azaletik aztertu ditugu eta emaitzek ez dute nabarmentzeko aldaketarik erakutsi. Bosgramak erabiltzea erabaki dugu azkenean, emaitzetan ageriko eraginik izango ez badu ere. Hurrengo taulan (6.14 taula) hirugrametan zein bosgrametan oinarritutako hizkuntza-ereduen emaitzen arteko konparazioa ikus dezakegu. Kontuan izan behar dugu, gure kasuan ordainen arteko aukeraketa egiteko erabiliko dugula, eta ondorioz, hizkuntza-ereduak esleitzen dien probabilitateen artean (eskala logaritmikoan) altuena daukan ordaina aukeratuko dugula (kasu honetan, "ezkerreko besoaren haustura" eta "ezkerreko besoaren haustura irekiaren infekzio" terminoen alde egingo genuke, zenbakiak negatiboak baitira). Ikus dezakegunez, bi ereduek oso zenbaki antzekoak ematen dituzte, eta gure ataza

zehatzari dagokionez, konparatzen ari garen ordainen arteko probabilitateen aldeak arbuiagarriak dira (1,19767 eta 1,181446). Horrela, esan bezala, bosgramak erabiltzeko hautua egin dugu egitura konplexuagoetan lagungarriak izan daitezkeelakoan.

	Hirugramak	**Bosgramak**
ezkerreko besoaren haustura	-12,771531	-12,795929
ezkerreko besoko haustura	-13,969201	-13,977375
Aldea	1,19767	1,181446
ezkerreko besoaren haustura irekiaren infekzio	-23,259136	-23,283657
ezkerreko besoko haustura irekiaren infekzio	-24,456806	-24,465101
Aldea	1,19767	1,181446

6.14 taula – Hirugramak eta bosgramak konparatzen.

Domeinuko corpusaz gain, euskarazko corpus orokorra ere erabili dugu. Azken hori Elhuyar Fundazioaren itzulpen-memorietako euskarazko esaldiekin osatuta dago nagusiki, eta publikoki eskuragarri dauden Eusko Jaurlaritzaren[13] eta Gipuzkoako Foru Aldundiaren[14] itzulpen-memoriekin osatuta dago. Itzulpen-memoria horiei, Elhuyar Fundazioak webetik erauzitako corpus paraleloaren euskarazko esaldiak (San Vicente eta Manterola, 2012) gehitu zaizkio, baita Euskal Herriko Unibertsitateak euskaratutako hainbat liburu ere, besteak beste. Horrekin guztiarekin, corpus orokorrak 100.000.000 token baino gehiago ditu. Corpus honi esker, osasun-zientzien corpuseko gabeziak estali nahi ditugu, 100 milioi tokenetan euskararen ezaugarri gehienak agertuko direlakoan.

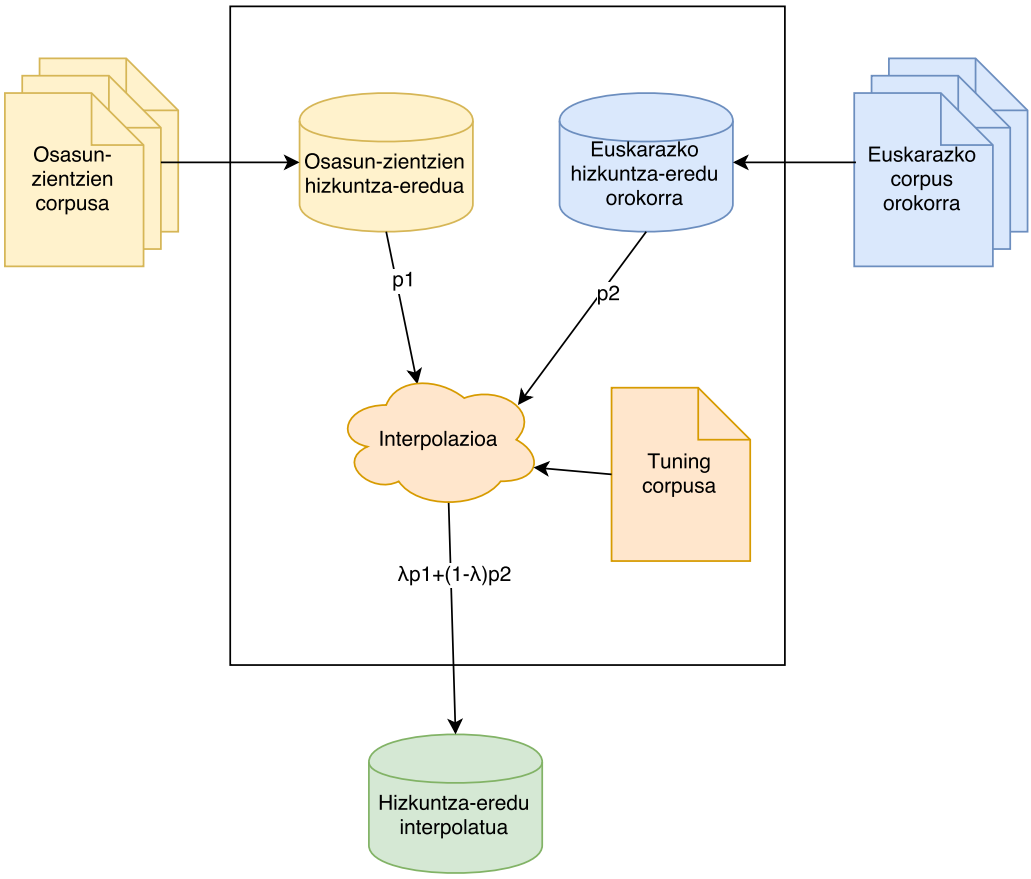
Hizkuntza-ereduak bi corpusekin entrenatu ditugu (bosgrametan oinarrituta) eta sortutako ereduak interpolatu egin ditugu. Interpolazioak corpusak biltzeko aukera emateaz gain, *tuning*-corpusa deritzon erreferentziazko corpus batekin optimizatzeko aukera ematen du. Gure kasuan, osasun-zientzien corpusa eta euskarazko corpus orokorra elkartu ditugu, eta SNOMED CTren ordain batzuekin osatutako corpusarekin optimizatu dugu.

Izan ere, *tuning* egiteko, erreferentziazko corpus txiki bat erabili ohi da. Esan dugun bezala, teknika horren helburua, entrenatutako eredua optimizatzea da, sortutako probabilitateei pisuak esleituz, itzulpenak *tuning*-corpuseko esaldiei ahalik eta gehien hurbil daitezen. Gure kasuan, SNOMED

[13] http://www.ivap.euskadi.eus/ivapeko-itzulpen-zerbitzu-ofizialeko-itzulpen-memoriak/r61-vedorok/eu/ (2017ko maiatzaren 9an atzitua).

[14] http://www.gipuzkoa.eus/imemoriak/ (2017ko maiatzaren 9an atzitua).

CTren terminoak sortu nahi ditugunez, baliabide lexikalekin euskaratutako terminoetatik, ordain bakarra dutenak aukeratu ditugu *tuning*-corpus moduan. Izan ere, termino horiek ordain bakarra izatean, ez dago anbiguotasunik, eta baliabide lexikaletatik erauzi ditugunez, proposatutako ordainak linguistikoki zuzenak izateko bermeak dituztela aurreikusi dugu. Corpus horrek 23.852 euskarazko termino eta 34.583 token ditu.



6.10 irudia – Hizkuntza eredua SNOMED CTren terminologiari doitzeko interpolazioa.

Interpolazioa irudikatu dugu 6.10 irudian. Ikus dezakegunez, osasun-zientzien corpuserako eta corpus orokorrerako hizkuntza-ereduak sortzen ditugu, eta horien probabilitateak (p1 eta p2) jasotzen ditu interpolazioa egiten duen programak. Interpolazioan, ereduen probabilitateei pisuak ematen zaizkie (λ), *tuning* corpuseko testuari modu optimoan egokitzeko, hain zuzen ere.

Termino konplexuen integrazioa

Matxinek Hitz Anitzeko Unitate Lexikalak (HAULak) ezagutzeko modulu berezi bat dauka. Bertan, HAULen identifikaziorako beharrezko informazioa jasotzen da erregelen bitartez. Osasun-alorreko HAULen identifikaziorako definitutako erregelen adibideak erakusten ditugu 6.11 irudian.

Erregela horien osaketa modu automatikoan egin dugu, horretarako Matxinek daukan aplikazio baten bitartez. Aplikazio horrek HAULen zerrenda jasotzen du, eta horien analisi linguistikoa Stanforden CoreNLP tresnaren bidez eginez, erregelak osatzen ditu. Analisi linguistikotik kategoria gramatikalak erabiltzen ditu, eta horrela HAULaren azkeneko hitza deklinatuta ager daitekeen edo ez ondorioztatzen du.

Adibidez, *no known allergies* terminoaren kasuan, azken hitza (*allergies*) pluralean dagoen izen bat da. Orokorrean, termino baten izendatzean barruko hitzak espreski pluralean agertzen badira, forma horretan agertzea beharrezkoa izan ohi da. Hori horrela izanik, HAUL hori ezagutu ahal izateko, forma guztiak bere horretan agertu behar dira. *Alfentanil allergy* terminoarekin aldiz, azken hitza izen singularra izanik, pluralean zein singularrean ager daitekeela iritzi dio Matxinek, eta horrela, azken hitzaren lema bilatuko du forma beharrean. Adibide horietan oinarrituz automatikoki sortu dizkiegun erregelak 6.11 irudian ikus ditzakegu.

```
1 {pattern: (/alfentanil/ [ lemma:"allergy" ]),
      result: Concat("alfentanil_allergy", "|", $0[1].tag)}
2 {pattern: (/no/ /known/ /allergies/),
      result: Concat("no_known_allergies", "|", $0[1].tag)}
```

6.11 irudia – Matxinen HAULen ezagutzarako automatikoki sortutako erregelak.

Sortutako erregela horiekin, Matxin HAUL berriak identifikatzeko gai da, gure kasuan osasun-alorreko termino konplexuak identifikatzeko, eta 6.15 taulan erakusten dugun adibidean bezala, HAULen identifikazio zuzenari esker, termino konplexuak dituzten esaldiak hobeto itzultzeko aukera dugu.

Jatorrizko esaldia	*He has Osler syndrome.*
Matxinen itzulpena	Hark **sindrome Osler** dauka.
MatxinMeden itzulpena	Hark **Osler-en sindromea** dauka.

6.15 taula – Matxin eta MatxinMeden itzulpenen adibidea.

Tesi-lan honetarako egokitu dugun Matxini MatxinMed deitzen badiogu ere, egokitzapena orokorra izan da, eta horrela Matxin edozein domeinutara egokitzeko aukera eman diogu. Hau da, Matxin berak hainbat domeinu izan ditzake integratuta, eta parametro bidez zein domeinu erabili nahi dugun erabaki ahalko dugu, sistema berdina erabiliz. Hala ere, lan honen ulermenean laguntzeko, osasun-zientzietarako egokitutako Matxini MatxinMed deituko diogu.

Hurrengo atalean, MatxinMeden eta KabiTermen ebaluazioaren diseinua azalduko dugu, osasun-zientzietako euskal komunitateari esker egin dugun Medbaluatoia kanpaina, hain zuzen ere.

6.3 Ebaluazioaren diseinua

Azkeneko bi urratsak ebaluatzeko, metodologia ezberdina erabili dugu. Aurreko kapituluan termino sinpleen adituen ebaluaziorako, 4 ebaluatzailek hartu zuten parte: bi hizkuntzalarik eta bi medikuk. Laginari dagokionean, 370 kontzeptu eta 766 termino ebaluatu zituzten, adituei eskuzko lan handia eskatuz.

Kapitulu honetako sistemak ebaluatzeko aldiz, osasun-zientzien euskal komunitatea inplikatzen duen ebaluazioari heldu diogu: Ebaluatoiaren (Aranberri *et al.*, 2016a) moldaera bat egin dugu, eta Medbaluatoia deritzon kanpaina diseinatu eta martxan jarri dugu. Ebaluatoia, 2015ean IXA taldeak Itzultzaile Automatikoak sailkatzeko erabili zuen euskal hiztunen komunitatea inplikatuz, eta oso emaitza interesgarriak jaso zituen.

Itzultzaile Automatikoen ebaluazioan ohikoena neurri automatikoak ateratzea bada ere, BLEU (Papineni *et al.*, 2002) izanik zabalduena, pertsonen ebaluazioa ekidinezina da sortzen dugun tresna pertsonek erabiltzea nahi baldin badugu. Hori horrela izanik, ikerketa gehienetan ebaluatzaile talde txiki batek parte hartzen du, nagusiki ikerketa taldeko bertako ikerlariek, oso lagin txikia ebaluatuz. Kontuan izan beahr dugu, BLEU neurtzeko urre-patroi lagina beharrezkoa da ebaluatu ahal izateko, eta gure kasuan ez dugu erreferentziazko termino-ordain pareen lagin nahikorik. Gainera, gure kasuan terminoentzat automatikoki sortutako ordainak direla ebaluatu nahi ditugunak, eta BLEU bezalako metrikek esaldiak ebaluatzeko diseinatu dira.

Ebaluatoiak aldiz, komunitatean oinarritutako ebaluazioa proposatzen du. Izan ere, euskaldunok komunitate kontzientziatu eta dinamikoa osatzen dugu, eta inbertsio handiegirik egin gabe, metodo azkarra da konfiantza

maila nahikoa duten ebaluazioak jasotzeko. Jakina den bezala, hizkuntza gutxituen komunitateek, haien hizkuntzek biziraun dezaten lagungarriak izan daitezkeen ekitaldietan zintzoki parte-hartzeko prestutasun osoa erakusten dute.

Esan bezala, Ebaluatoiak Itzultzaile Automatikoak bere horretan ebaluatzeaz haratago, horien sailkapena egitea du helburu. Bost sistema ebaluatu zituzten 2015eko kanpainan, guztiak ingelesa-euskara hizkuntza parerako (Aranberri, 2016): EuSMT oinarri-lerroa, EuSMT segmentazioarekin, Matxin ENEUS, SMatxinT sistema hibridoa eta Google Translate. Esaldiak izan ziren ebaluatzen ziren unitateak.

Alegria *et al.* (2013)-en ikasketei esker, Ebaluatoian diseinatutako ataza ahalik eta sinpleena da. Horretarako, bikotekako konparaketaren metodoa erabili zuten esaldi hautagaiak ebaluatzeko. Parte-hartzaileei jatorrizko esaldia eta bi itzulpen automatiko erakusten zaizkie, eta beraien eginbehar bakarra bi itzulpenen arteko onena aukeratzea da. Metodo horrek, esfortzu kognitibo txikiagoa eskatzen du beste metodo batzuk baino, ebaluatzaileen arteko adostasun handiagoa lortuz. Agerian geratu zen hori, ebaluatzaileen arteko adostasuna, 0,49 eta 0,53 arteko *kappa* neurriak lortu baitzituen Ebaluatoiak. Aurretik WMT (Workshop on Statistical Machine Translation) edizioetan egindako ebaluazio antzekoetan 0,075 eta 0,324 arteko *kappa* neurriak lortu ziren (Bojar *et al.*, 2014).

Guk Ebaluatoian garatutako metodologia erabili dugu terminoen ordain hautagaiak sailkatzeko. Horrela, esaldiak beharrean, parte-hartzaileek termino konplexuak (2 eta 8 token artekoak) ebaluatuko dituzte. Ebaluatoian ebaluatu beharreko esaldiak, esaldi konplexuak ziren gehienetan, eta horrela, gure ataza errazagoa izango dela aurreikusten dugu. Gainera, parte-hartzaileak alorrean adituak direla kontuan hartzen badugu, zailtasuna jaitsiko da Ebaluatoiarekin alderatuz.

Aurreko guztia kontuan hartuta, atazari eginkizun berri bat gehitzea erabaki dugu Medbaluatoian: ordain hautagaien artean zuzena(k) aukeratzeko aukera ere ematea. Horrela, automatikoki sortutako ordainen zuzentasuna ere neurtu ahalko dugu.

Horretaz gain, parte-hartzaileen profila osatzeko eremuak ere egokitu ditugu. Izan ere, Ebaluatoiak edozein adin, ikasketa-maila edo jakintza-alorreko jendearen parte-hartzea bilatzen zuen bitartean, gure kasuan osasun-zientzien alorrak baino ez ditugu aurreikusten, eta ikasketa mailari dagokionean, unibertsitate-graduko ikasle, espezialitateko ikasle zein ikasketak bukatutakoak izango dira parte-hartzaileak.

Medbaluatoia KabiTerm eta MatxinMed ebaluatzeko erabili dugu, hortaz bi ebaluaziorako multzo lortu ditugu. Alde batetik, KabiTerm ebaluatzeko, hiru sistema konparatu ditugu: Google Translate oinarri-lerroko sistema gisa, KabiTerm sistema eta MatxinMed-1 (KabiTermekin sortutako terminorik gabeko MatxinMeden bertsioa). Multzo honi KabiTermen multzoa deituko diogu.

Bestetik, MatxinMed bera ebaluatuko dugu. Horretarako, oinarri-lerroko sistema berdina erabiliko dugu, Google Translate, eta MatxinMeden bi aldaera erabiliko ditugu. Lehenengoa, MatxinMed-1, KabiTermen ebaluazioan erabiliko dugun berdina, eta bigarrengoari, MatxinMed-2, KabiTerm sistemak sortutako ordainak ere integratu dizkiogu. Multzo horri MatxinMeden multzoa deritzogu. Kontuan izan, MatxinMed-1 eta MatxinMed-2ren artean, MatxinMed-2k hiztegi zabalagoa duela termino konplexuei dagokionean. Horrela, termino motzetan aldea aurreikusten ez badugu ere, termino luzeetan aldea egon daiteke. Adibidez, MatxinMed-1ek *entire deep pectoral muscle* terminoa "sakoneko pektorala osoa" moduan euskaratzen duen bitartean, MatxinMed-2k "bularreko muskulu sakon osoa" euskaratzen du, *deep pectoral muscle* dagoeneko hiztegiratuta duelako.

Hurrengo taulan (6.16 taula), bi ebaluazioetan parte hartuko duten sistemak erakusten ditugu.

	KabiTerm	**Google**	**MatxinMed-1**	**MatxinMed-2**
KabiTermen multzoa	✓	✓	✓	
MatxinMeden multzoa		✓	✓	✓

6.16 taula – Medbaluatoiaren bidez ebaluatutako sistemak.

Kapitulu honetako ebaluazioetan, aurreko ebaluazioetan bezala, nahasmenduen, aurkikuntzen, gorputz-egituren eta prozeduren hierarkiak izan ditugu aztergai. Token kopuruei dagokienean, 2 eta 8 token artean dituzten terminoekin ibili gara, lehenago ikusi dugun bezala (3 kapituluan) 8 token arteko terminoak lagin osoaren % 92a baino gehiago baitira, eta hortaz, laginketa esanguratsua lortu dugu. Izan ere, geroz eta token kopurua handiagoa, orduan eta termino konplexuagoak dira, eta horien azterketa, sorkuntza automatikoa zein ebaluazioa asko zailtzen dira.

Ebaluatoian erabilitako laginaren tamaina bera erabiliko dugu Medbaluatoiaren kanpainarako: 500 jatorri-termino. Kontuan izanda gure kasuan bi ebaluazio egingo ditugula, ebaluazio bakoitzerako 500 termino aukera-

tu ditugu SNOMED CTren termino konplexuetatik, eta multzo horiek token kopuruaren eta hierarkiaren arabera estratifikatu ditugu (Ripley, 2009). Hau da, 6.17 taulan ikusten dugun bezala, SNOMED CTren lagin guztiaren token kopuruen eta hierarkien proportzioak kontuan hartuta jaso ditugu ingelesezko bertsiotik terminoak. Horretarako lau hierarkien arteko proportzioak kalkulatu ditugu, eta ostean hierarkia horien barnean token kopuruen proportzioak.

	Nahasmenduak		**Aurkikuntzak**		**Gorputz-egiturak**		**Prozedurak**	
	Prop.	**Kop.**	**Prop.**	**Kop.**	**Prop.**	**Kop.**	**Prop.**	**Kop.**
2 token	0,20	42	0,18	18	0,20	12	0,13	11
3 token	0,24	43	0,23	16	0,23	31	0,20	19
4 token	0,20	38	0,21	24	0,19	25	0,22	27
5 token	0,16	35	0,18	11	0,16	11	0,19	24
6 token	0,10	23	0,11	7	0,11	13	0,13	12
7 token	0,07	16	0,06	8	0,07	8	0,09	11
8 token	0,04	4	0,03	6	0,03	0	0,05	7
Denera	0,37	201	0,17	90	0,19	98	0,28	111

6.17 taula – Ebaluazio multzoetako kopuruak eta proportzioak.

Ebaluazioen kopuruak erabakita, SNOMED CTtik ingelesezko jatorri--terminoak erauzi ditugu. KabiTermen ebaluazio multzorako, zoriz aukeratu ditugu 500 termino konplexu, beti ere, KabiTerm jatorri-terminoen ordainak sortzeko gai bada. MatxinMeden multzoa sortzeko aldiz, Kabitermek itzuli ezin dituen jatorri terminoetatik 500 jaso ditugu. Izan ere, MatxinMed soilik KabiTermek euskaratzen ez dituen terminoak euskaratzeko erabiliko dugu, eta horrela egin dugu ebaluaziorako lagina lortzeko ere.

KabiTermen multzorako, terminoen gainsorkuntza kontrolatzeko beharra izan dugu. Izan ere, gainerako sistemek ordain bakarra ematen duten bitartean, KabiTermek hainbat ordain emateko joera dauka, eta sistemak ez leudeke baldintza berdinetan lehiatzen. Lehia orekatzeko, gure osasun-zientzien euskarazko hizkuntza-eredua erabili dugu. KabiTermek proposaturiko ordain guztien artean, hizkuntza-ereduan probabilitate altuenarekin agertzen den ordaina aukeratu dugu. Tamalez, horrek ez du ziurtatzen ordain onena izango denik, baina metodo automatikoen artean ezagutzen dugun metodo fidagarriena da.

Denera 1.000 termino ebaluatu ditugu, eta termino bakoitzarekin hiru sistema-bikote (hiru sistema). Horrela, 3.000 ebaluazio behar izan ditugu. Subjektibotasunari aurre egiteko, ebaluazio bakoitza 5 erabiltzailek egitea

erabaki dugu. Horrela, denera 15.000 ebaluazio behar izan ditugu Medbaluatoia kanpaina bukatzeko. Gainera, erabiltzailearen arreta neurtzeko, kontrolerako ebaluazioak gehitu ditugu. Kontrolerako ebaluazio horiek, eskuz emandako ordain zuzenez, eta oso nabariak diren ordain okerrez osatuta daude. Kontrolerako ebaluazioetatik herena gaizki eginez gero, parte-hartzailea kanporatu egiten da.

Kontrolerako ebaluazioak beharrezkoak dira, erantzunen fidagarritasuna ziurtatu ahal izateko, bereziki komunitate zabalekin lan egiten badugu. Ahal den neurrian, erantzun gaiztoak edo hizkuntza gaitasun nahikoa ez duten parte-hartzaileak identifikatu beharko genituzke, beraien erantzunak alboratzeko. Ebaluatoian bezala, ebaluazioa egiten duten bitartean kontrolerako ebaluazioak gaizki egiten dituzten parte-hartzaileak kanporatzeko erabakia hartu dugu: bost ebaluazioetatik bat, kontrol-ebaluazioa izango da, eta horietatik herena gaizki eginez gero, parte-hartzailea kanporatua izango da.

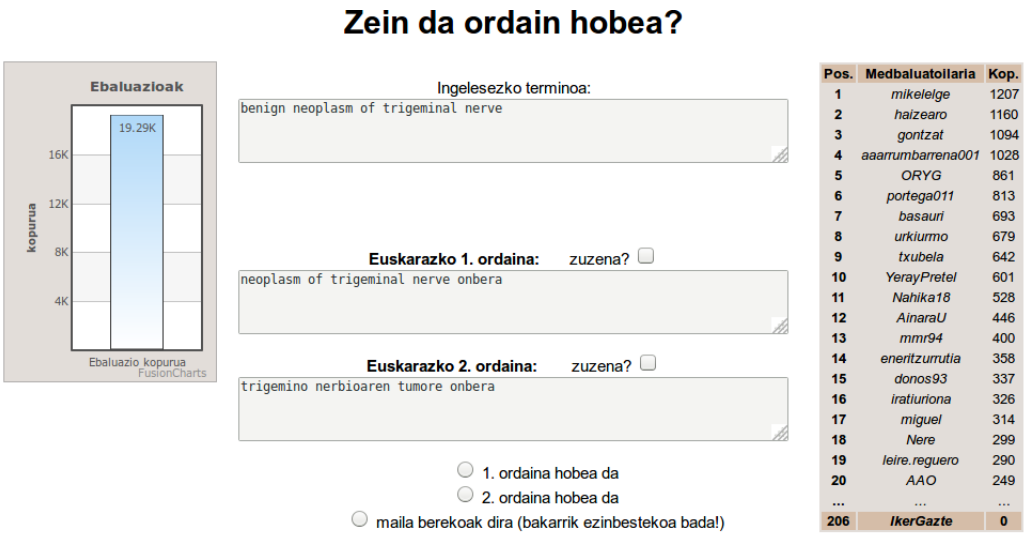
Kontuan izanik parte-hartzaile batek gehienez 1.000 ebaluazio egingo dituela, eta bost ebaluaziotik bat kontrolekoa izango dela esan dugunez, 250 kontrolerako termino-ordain beharko ditugu. Kontrolerako terminoak medikuek sortutako laginetik hartu ditugu. Hau da, KabiTerm sistema garatzeko adituei eskatutako termino eta euren ordainen zerrendatik. Hautagaiak sortzeko, batetik medikuek proposaturiko ordain zuzenak erabili ditugu, eta bestetik Matxin itzultzaile orokorrak emandako ordainak eskuz okertuta, antonimoak, ezeztapenak eta testuingurutik kanpoko hitzak erabili ditugu (ikus adibideak 6.18 taulan).

Jatorri-terminoa	*burn of vagina and uterus*
Ordain hobea	bagina eta umetokiko erredura
Ordain okerragoa	erre ezazu utero baginako eta
Jatorri-terminoa	*excision of fimbrial cyst*
Ordain hobea	finbriako kistearen erauzketa
Ordain okerragoa	fimbrial cysteko hanka

6.18 taula – Kontrolerako termino-ordainen bi adibide.

Medbaluatoiaren zabalkunderako Medikuntzako Fakultateko euskara taldeen gela guztietik pasa ginen kanpainaren hasierako egunean. Horretaz gain, mezu elektroniko bat zabaldu genuen, profesionalen artean parte-hartzea sustatzeko. Zabalkundean gakoak izan dira Osasungoa Euskalduntzeko Erakundea eta EHUko Euskara Errektoreordetza.

Interfazea hurrengo irudian (6.12) ikus dezakegu. Bertan, ingelesezko terminoa eta euskarazko bi ordainak agertzeaz gain, ordainen artean onena aukerako formularioa agertzen da. Gainera ordain bakoitzaren ondoan, ordaina bera zuzen gisa aukeratzeko laukitxoa ere eskaintzen da. Ebaluaziotik kanpo, denera egindako ebaluazio kopurua ere erakusten zaie parte-hartzaileei, eta lehia sustatzeko parte-hartzaile aktiboenen *ranking*a ere, 6.12 irudiko bi aldeetan ikus daitekeen bezala.



6.12 irudia – Medbaluatoiaren ebaluazio baten interfazea.

Hurrengo atalean, Medbaluatoia kanpainaren emaitzak erakutsiko ditugu, ebaluatzaileen arteko adostasun neurriak eta sistemen arteko konparaketak azalduz.

6.4 Emaitzak

Atal honetan Medbaluatoiaren kanpainan lortutako emaitzak aurkeztuko ditugu eta bukaeran, SNOMED CTren estaldurari lotutako emaitzak ere emango ditugu.

6.4.1 Medbaluatoiaren emaitzak

Medbaluatoiaren kanpaina oso arrakastatsua izan dela azpimarratu beharrean gaude. 2016ko urriaren 10an abiatu genuen, eta hiru astetarako aurreikusitako lana aste bakarrean bukatu zuten parte-hartzaileek. Denera 217 parte-hartzaile izan ditugu kanpainan, eta euretako 13 baztertuak izan dira kontrol-ebaluazioetan kale eginagatik.

Parte-hartzaileen profila

Hurrengo taulan, 6.19 taula, parte-hartzearen inguruko laburpena erakusten dugu. Ikus daitekeenez, kanporatutako 13 horiei, ebaluaziorik egin ez duten beste 13 gehitu behar dizkiogu parte-hartzaile ez baliagarriak zenbatzeko garaian. Horrela, denera 191 baliozko parte-hartzaile izan ditugu. Horien mediana 24 ebaluaziokoa izan da, eta batezbestean 100,25 ebaluazio egin dituzte.

	Parte-hartzaile kopurua	**%**
Parte-hartzaileak denera	217	
Kanporatuak	13	5,99
Ebaluaziorik gabe	13	5,99
Parte-hartzaile baliozkoak	191	88,02
	Ebaluazio kopurua	
Parte-hartzaile baliozkoen mediana	24	
Parte-hartzaile baliozkoen batezbestekoa	100,25	

6.19 taula – Medbaluatoiko parte-hartzearen laburpena.

Adin-tarteari dagokionean (6.20 taula), parte-hartzailearen profila nagusiki gaztea dela esan dezakegu. 19-25 adin-tartea da jendetsuena, parte-hartzaileen % 40 baino gehiago adin-tarte horretakoa izanik. Ez da harritzekoa, eta ikasketa mailari dagokion taularekin bat dator (6.21 taula), bertan ikasketa maila ezberdinetan dauden parte-hartzaileak % 55a baino altuagoa baita (1. eta 6. maila arteko portzentajeak gehituta). Adin-tarte horretaz gain, 26-45 adin-tartean portzentajea handia bildu da, % 40 inguru, eta horiek ikasketak bukatuta dutenekin bat egiten dute (% 41,94).

Medbaluatoiaren oinarria teknologia izateak, parte-hartzaileen profila asko mugatu duela esan dezakegu, adin-tarte zehatz bateko jendearentzat erakargarriagoa eginez.

Adin-tartea	Parte-hartzaile kopurua	%
<18	13	5,99
19-25	94	43,32
26-35	35	16,13
36-45	52	23,96
46-55	19	8,76
56-65	4	1,84
>65	0	0,00

6.20 taula – Medbaluatoiko parte-hartzaileen adin-tarteak.

Ikasketa maila	Parte-hartzaile kopurua	%
1. maila	33	15,21
2. maila	11	5,07
3. maila	6	2,77
4. maila	18	8,30
5. maila	19	8,76
6. maila	10	4,61
Erresidentzia	15	6,91
Ikasketak bukatuta	91	41,94
Bestelakoak	14	6,45

6.21 taula – Medbaluatoiko parte-hartzaileen ikasketa mailak.

Esan bezala, Medbaluatoian automatikoki sortutako osasun-zeintzietako terminologia ebaluatu nahi izan dugu. Hori horrela izanik, parte-hartzaileak zein osasun-zientzietako alorretan lan egiten edo ikasten duten jakin nahi izan dugu. Emaitzak 6.22 taulan ikus ditzakegu. Nagusiki medikuak eta medikuntzako ikasleak izan dira parte-hartzaileak (% 61,75), baina erizainen eta erizaintzako ikasleen parte-hartzea ere nabarmentzekoa izan da % 21,20a izanik.

Hizkuntzen ezagutzaren inguruan, Medbaluatoiaren atazan euskara zein ingelesa jakitea beharrezkoa zen parte-hartzaileentzat. Hala ere, beraien lana berrikuspena izanik, ez dugun gutxieneko hizkuntza gaitasunik eskatu. Hurrengo tauletan ikus dezakegunez (6.23 eta 6.24 taulak), euskarazko maila

Alorra	Parte-hartzaile kopurua	%
Erizaintza	46	21,20
Medikuntza	134	61,75
Farmazia	6	2,77
Fisioterapia	17	7,83
Bestelakoak	14	6,45

6.22 taula – Medbaluatoiko parte-hartzaileen alorrak.

altuena nagusi izan den bitartean (parte-hartzaileen % 93,09 C1-C2 mailarekin), ingelesezko maila erdi-mailakoa izan da nagusi (% 48,85a B1-B2 mailarekin), nahiz eta goi-mailako (% 30,87) eta behe-mailako (% 20,28) asko ere egon.

Euskara maila	Parte-hartzaile kopurua	%
A1-A2	2	0,92
B1-B2	13	5,99
C1-C2	202	93,09

6.23 taula – Medbaluatoiko parte-hartzaileen euskara mailak.

Ingelesa maila	Parte-hartzaile kopurua	%
A1-A2	44	20,28
B1-B2	106	48,85
C1-C2	67	30,87

6.24 taula – Medbaluatoiko parte-hartzaileen ingeles mailak.

Parte-hartzaileen arteko adostasuna

Jarraian, parte-hartzaileen arteko adostasuna kalkulatu dugu. Horretarako, ohikoena *kappa* (κ) neurria kalkulatzea da. Aurreko kapituluan azpimarratu genuen moduan bi ebaluatzaileen arteko adostasuna neurtzeko Cohenen κ

(Cohen, 1960) erabiltzen da, eta bi baino gehiago direnetan Fleissena (Gwet, 2014, Artstein eta Poesio, 2008). Kontuan izan, bai Cohenen zein Fleissen *kappa* neurriak kalkulatzeko garaian, ebaluatzaile berdinek lagin berdina ebaluatzen dutela ulertzen dela.

Gure kasuan, 200 ebaluatzailetik gora ditugu (baliozkoak 191), eta bakoitzak ebaluatzen duen lagina ezberdina da, bai kopuruz zein ebaluazioz. Hori kontuan izanik, eta nahiz eta *kappa* neurriek mugak dituzten, beste ebaluazio batzuetan neurtutakoarekin konparatzeko helburuarekin atera dugu Cohenen *kappa*. Aurreko itzulpen automatikoaren WMT kanpainetan Cohenen *kappa* (Bojar *et al.*, 2014) neurriak atera dituzte, baita Ebaluatoiaren kanpainan ere (Aranberri *et al.*, 2016a).

Sistema-parea	***Kappa***
Google - KabiTerm	0,36
Google - MatxinMed-1	0,37
KabiTerm - MatxinMed-1	0,37
Google - MatxinMed-2	0,30
Google - MatxinMed-1	0,30
MatxinMed2 - MatxinMed-1	0,48

6.25 taula – Sistema pare bakoitzak lortutako adostasuna (*kappa*).

Kappa neurrien taulan ikus daitekeenez, 6.25 taulan, lortutako *kappa* balioak 0,30 eta 0,48 artean kokatuta daude. Aurreko kapituluan aipatzen genuen interpretazioaren harira (Landis eta Koch, 1977), non 0-0,2 artekoa adostasun arina den, 0,2-0,4 dezentekoa, 0,4-0,6 moderatua, 0,6-0,8 sendoa eta 0,8-1,0 ia perfektua, gure balioak dezentekoak direla ondorioztatu dezakegu.

Lortutako *kappa* balio guztiak, itzulpen automatikoaren WMT kanpainetan lortu izan diren balioen barruan daude (Bojar *et al.*, 2014), langa gorenetik gertu. Emaitzak ez dira aurrikusitakoak bezain onak izan (Ebaluatoiaren 0,49 eta 0,53 arteko *kappa* kontuan izanik), eta baliteke gehitu dugun eskakizun berriak (zuzenak diren aukeratzeko funtzionalitatea) horretan eragin zuzena izatea. Hala ere, lortutako adostasuna balio onargarrien artean dagoela uste dugu.

Ebaluazio kopuruak

Medbaluatoian zehar, parte-hartzaileei jatorrizko terminoa, eta bi euskarazko ordain erakutsi dizkiegu, eta bien arteko onena aukeratu behar izan dute, horrela, ebaluazio bat eginez. Ez dugu onena aukeratzeko irizpiderik eman, bakoitzaren irizpideen arabera onena iruditzen zaiena jaso nahi izan dugulako.

Jatorrizko termino bakoitzerako eta sistema pare bakoitzeko bost ebaluazio jaso nahi izan ditugu. Hala ere, web-aplikazioaren konfigurazioa dela-eta, batzuetarako 7 ebaluazio ere lortu ditugu. Ebaluazio gehigarri horiek ez dira akastunak eta erabiltzea erabaki dugu. Hurrengo taulan (6.26 taula), sistema bakoitzerako lortutako ebaluazio kopuruak erakusten ditugu[15]. Gogoratu behar da 2.500 ebaluazio behar genituela sistema-pare bakoitza ebaluatu ahal izateko (500 termino, eta 5na ebaluatzaile). Taulan ikus dezakegunez hortik gorako kopuruak lortu ditugu, baina aldakortasun txikiarekin (2.523 eta 2.540 artean). Hortaz, lortutako emaitzak konparagarriak dira.

	Google - KabiTerm	**Google - MatxinMed-1**	**KabiTerm - MatxinMed-1**
KabiTermen multzoa	2.529	2.523	2.527
	Google - MatxinMed-2	**Google - MatxinMed-1**	**MatxinMed-2 - MatxinMed-1**
MatxinMeden multzoa	2.540	2.535	2.535

6.26 taula – Sistema-pare bakoitzak denera lortutako ebaluazioak.

Emaitzak

Ebaluazioetatik sistema onena aukeratzeko honako estrategia jarraitu dugu: bi sistemen arteko bozka diferentzia bi baino handiagoa bada, zalantzarik gabeko irabazlea da (taulan "X.sistema++" moduan kodetu dugu, non X-k sistema adierazten duen). Sistemen arteko diferentzia 1 edo 2 bada, sistema irabazlea dela ere deritzogu (taulan "X.sistema+" bezala kodetu dugu). Bi sistemek bozka kopuru berdina lortzen badute, orduan berdintza dagoela deritzogu. Adibidez, termino baten ebaluazioan, Google-ek bozka bat jaso badu eta KabiTermek 4, bien arteko ezberdintasuna 3 denez KabiTermen alde, "2.sistema++"-n bozka bat gehituko da.

[15]Kontaketetan ez ditugu kontrolerako ebaluazioak kontuan izan, noski.

	Google - KabiTerm	**Google - MatxinMed-1**	**KabiTerm - MatxinMed-1**
1.sistema++	6,8 (34)	13,2 (66)	**46,4(232)**
1.sistema+	3,2 (16)	9,2 (46)	14,4 (72)
berdin	3,2 (16)	7,8 (39)	5,6 (28)
2.sistema+	13,4 (67)	15,2 (76)	12,4 (62)
2.sistema++	**73,2(366)**	**54,4(272)**	21,0(105)

6.27 taula – Medbaluatoia kanpainaren sistemen ebaluazioaren emaitzak (KabiTermen multzoa).

	Google - MatxinMed-2	**Google - MatxinMed-1**	**MatxinMed-2 - MatxinMed-1**
1.sistema++	19,4 (97)	21,6(108)	6,6 (33)
1.sistema+	12,0 (60)	16,4 (82)	13,0 (65)
berdin	7,0 (35)	8,8 (44)	**62,4(312)**
2.sistema+	22,0(110)	19,4 (97)	14,6 (73)
2.sistema++	**39,8(199)**	**34,0(170)**	3,6 (18)

6.28 taula – Medbaluatoia kanpainaren sistemen ebaluazioaren emaitzak (MatxinMeden multzoa).

Emaitzetan ikus dezakegunez, KabiTermen multzoan (6.27 taula), KabiTerm sistema izan da emaitza onenak lortu dituena alde handiarekin. Bereziki handia izan da Google-en itzultzailearekin alderatuz, % 86,6 kasutan izan baita hobea (% 73,2 gehi % 13,4). MatxinMed-1ekin egindako konparazioan aldea txikiagoa bada ere, nabarmenki jaso ditu emaitza hobeak (% 46,4 kasutan nabarmen irabazi du, eta MatxinMed-1ek aldiz % 21,0 kasutan).

MatxinMeden multzoari dagokionean (6.28 taula), emaitzak horren nabarmenak ez badira ere, MatxinMeden bi bertsioek Google Translatek baino emaitza hobeak lortu dituzte. Izan ere, irabazten duten portzentajeak gehituta, % 61,8 kasutan MatxinMed-2k irabazi du (% 22,0 gehi % 39,8), eta Google-ek aldiz % 31,4 kasutan (% 19,4 gehi % 12,0). Dena dela, MatxinMeden bi bertsioak konparatzen ditugunean, emaitzek ez dute sistema baten edo beste baten alde egiten. Izan ere, kasuen % 62,4tan sistemek kalitate berdina erakutsi dute, % 19,6tan MatxinMed-2k emaitza hobeak (portzentajeak gehituta) eta % 18,2tan MatxinMed-1ek hobeak. Ezin dugu ahaztu, bi sistemen aldea KabiTermek sortutako terminoetan datzala. Horrela, MatxinMed-2k

KabiTermek euskaratutako termino habiaratuak aurkitzen dituenean bakarrik sortuko ditu ordain ezberdinak, eta esku artean izan dugun laginean kasu gutxitan gertatu da hori. Hala ere, ezin dugu kasu horietarako zein sistema den hobea ondorioztatu, emaitza oso antzekoak lortu baitituzte bi sistemek. Azpimarratzekoa da, KabiTermen multzoan agertzen diren terminoen egitura mugatua dela, eta egitura horretan MatxinMedek emaitza hobeak lortzen dituela Google-ekin alderatuz.

Termino zuzena aukeratzeko funtzionalitate berriak ez du arrakastarik izan. Uste dugu, ordain "ona" eta ordain "hobea" kontzeptuen arteko bereizketa ez dela garbi geratu, parte-hartzaile batzuekin kontrastatu ostean "hobetzat" hartzen zuten hori "zuzena" markatzen baitzuten, nahiz eta ez izan erabat zuzena. Hortaz, emaitza esanguratsuak ez direnez, ez ditugu erakutsiko.

6.4.2 KabiTermen estaldura SNOMED CTn

Jarraian, SNOMED CTren euskaratzen lortu ditugun estalduraren inguruko datuak emango ditugu. Hau da, ingelesezko nahasmenduen, aurkikuntzen, gorputz-egituren eta prozeduren hierarkietako terminoen zein proportzio euskaratu dugun KabiTermi esker. Kontuan izan, MatxinMedi esker, SNOMED CT bere osotasunean euskaratzen dugula, eta beraz, MatxinMeden inguruko estaldura datuak ez dira esanguratsuak; KabiTermen emaitzak bai, aldiz. Horregatik, 6.29 taulan KabiTermen estalduraren inguruko datuak ematen ditugu, SNOMED CTren zenbat termino eskaratzeko gai izan den eta zenbat ordain sortu dituen adieraziz.

	Nahasmendu	**Aurkikuntza**	**Gorputz--egitura**	**Prozedura**
Terminoak denera	114.830	52.857	59.384	87.104
Euskaratuak	26.136	4.054	12.497	10.651
Terminoen %	% 22,76	% 7,67	% 21,04	% 12,23
Ordainak	102.724	15.868	43.913	34.232

6.29 taula – KabiTermen SNOMED CTren euskaratze-estaldura.

Ikusten dugunez, bereziki azpimagarria da nahasmenduen hierarkian egindako ekarpena, 26.136 termino euskaratzeko gai izan baita (114.830 terminotatik). Horrek termino guztien % 22,76 termino euskaratu dituela esan nahi

du, oso emaitza ona gure iritziz. Gorputz-egituren hierarkiako emaitzak ere portzentajeei dagokionez antzekoak izan dira (% 21,04). Aurkikuntzen eta prozeduren hierarkietan ez ditugu horren emaitza onak lortu portzentajeei dagokionean. Azpimarratzekoa da, landu ditugun egitura gehienak nahasmenduak eta gorputz-egiturak deskribatzeko egiturak izan direla, nahasmenduen kasuan agerpen kopuruaren eraginez eta gorputz-egituren kasuan termino habiaratuen egiturak oso errepikakorra dutelako.

Azpimarratzekoa da ere ordainen sorkuntzan KabiTermek duen gainsorkuntza. Asko mugatzea lortu dugu eta batezbeste 3-4 ordain sortzen ditu termino bakoitzeko (adibidez, 26.136 termino euskaratu ditu 102.724 ordain sortuaz, hau da, 3,9 ordain terminoko).

6.5 Laburpena eta ondorioak

Kapitulu honetan, termino konplexuen euskaratzerako egindako ekarpena aurkeztu dugu. Bi sistema garatu eta ebaluatu ditugu: KabiTerm eta MatxinMed. KabiTermek termino habiaratuen egitura baliatzen du ingelesezko termino konplexuak euskaratzeko. Ideia nagusia zera da: termino konplexuen barruan maiz beste termino batzuk agertu ohi dira habiaratuta. Habiaratutako terminoa, barrukoa, euskaratuta badago, euskaratze-patroiak definitu ahal dira termino konplexu osoaren ordainak lortzeko. Helburu horrekin termino habiaratu horien SNOMED CTren hierarkiak erabili ditugu terminoen egiturak aztertzeko eta euskaratze-patroiak definitzeko.

Bestalde, MatxinMed, erregeletan oinarritutako Matxin itzultzaile automatikoaren osasun-zientzien domeinura egindako egokitzapena da. Matxini hainbat ekarpen egin dizkiogu, domeinu zehatzetara egokitu ahal izateko. Horretarako, hiztegiari ezaugarri berri bat gehitu diogu, eta honi esker, lexikoaren transferentzia egiteko momentuan, Matxinek ordainak domeinuaren arabera aukeratzen ditu. Osasun-zientzien kasurako, hiztegia zabaldu dugu jadanik euskaratu ditugun SNOMED CTren termino-ordain pareekin. Gainera, anbiguotasun kasuetarako, ordainen arteko ordena zehazteko hizkuntza-eredua ere sortu dugu. Horretaz gain, modulu gehigarri bat ere gehitu diogu, aurreko kapituluan (5 kapitulua) aurkeztu dugun NeoTerm sistema integratzeko eta termino ezezagunak euskaratu ahal izateko. Azkenik, termino konplexuak identifikatzeko erregelak gehitu ditugu, horien hiztegiko ordainak erabili ahal izateko.

Bi teknikak Medbaluatoia deritzon kanpainan ebaluatu ditugu, eta artea-

ren egoerako Google Translate tresnarekin konparatu ditugu. Emaitza oso positiboak izan dira, bai parte-hartzeari dagokionean, bai eta gure sistemen kalitatea, Google-ekin alderatzeari dagokionez. Bereziki azpimarragarriak dira KabiTermek lortutako emaitzak, gainerako sistemei nabarmenki irabazi baitie.

Ebaluazioak oso erantzun positiboa izan du komunitatearen aldetik, eta lanean jarraitzeko irrika zabaldu digute. Medbaluatoiak euskal osasun-zientzien komunitatea inplikatzeko tresna oso erabilgarria dela frogatu du, eta etorkizunean antzeko ekintzak abiatzea aurreikusten dugu, SNOMED CTren euskarazko bertsioaren baliozkotzea eta zuzentzea, adibidez.

7

Zuzentzaile ortografikoa eta osasun-txostenak euskaraz idazteko prototipoa

Sarrerako kapituluan aipatu dugun moduan (1 kapitulua), osasun-arloko terminologia euskaraz izatea ezinbestekoa da etorkizunean osasunean euskaraz lan egiteko aplikazioak garatu nahi baditugu. Osasun-zientzietako hiztegia garatzen joan garen heinean, jarraian aurkeztuko ditugun tresnak garatu ditugu. Batetik, 7.1 atalean deskribatu dugun XuxenMed zuzentzaile ortografiko egokitua dugu, eta bestetik, 7.2 atalean aurkeztuko dugun osasun-txostenak euskaraz idazteko prototipoa (Perez-de-Viñaspre *et al.*, 2015) garatu dugu.

7.1 XuxenMed: osasun-zientzietarako zuzentzaile ortografikoa

Ortografia-zuzentzaileek maiz eraginkortasuna galdu egiten dute hizkuntza oso teknikoa erabiltzerakoan. Izan ere, terminologia espezializatua ez da hiztegi orokorretan agertzen, eta zuzentzaileak hiztegi horietatik elikatzen dira. Hortaz, testu oso espezializatua idazten badugu, testuaren zati handiak gorriz azpimarratuta ager daitezke, akatsak balira bezala, nahiz eta termino horiek domeinuan zuzenak izan. Horrelakoetan erabiltzaileak zuzentzailea erabiltzeari utz diezaioke, lagundu ordez, lana ematen diolako benetako

akatsak bereizteak. Arazo horri aurre egiteko, Xuxen (Agirre *et al.*, 1992) zuzentzailea oinarri hartu dugu eta berau osasunaren domeinuko hiztegi espezializatuetako hitz-sarrerekin elikatu dugu. Xuxenen gehitu dugun hiztegi espezializatua osatzeko, 5. kapituluan lortutako euskarazko terminoak erabili ditugu, egokitzapena egiteko unean eskura genituenak dira-eta. Prozesu honetan, 32.728 lema gehitu ditugu Xuxenen hiztegian eta XuxenMed izena eman diogu egokitzapen honetatik sortutako zuzentzaileari.

Xuxen zuzentzaile orokorrak alta-txosten batean egiten dituen zuzenketak ikus ditzakegu 7.1 irudian, eta 7.2 irudian aldiz medikuntzara egokitutakoak egiten dituen zuzenketak. Ikus daitekeenez, Xuxen orokorrak (7.1 irudia) medikuntzari loturiko zenbait hitz okertzat hartzen ditu ("sinusal", "gastroskopioa", "behazunbidea" eta "Adenokartzinoma"), eta XuxenMedek (7.2 irudia), aldiz, zuzentzat ditu.

ELEKTROKARDIOGRAMA (EKG).	Erritmo sinusal normala.
GASTROSKOPIA.	Anestesia topikoa jarri ostean, gastroskopioa ahotik barrena sartuta: hestegorriak kalibre eta mukosa normala du; urdailaren zabalgarritasuna eta uzkurtasuna normalak dira; ez da odol-aztarnarik ikusten, ez eta lesiorik ere; piloro normala. Duodeno-
ABDOMENEKO OTA.	Gibela, barea, pankrea, behazuna, behazunbidea, giltzurrunen-gaineko guruina eta giltzurrunak normalak. Itsuaren paretaren gizentze bat ikusten da, neoplasia itxurakoa. Ez da adenopatiarik ikusten.
ANATOMIA PATOLOGIKOA.	Adenokartzinoma.

7.1 irudia – Xuxen ortografia-zuzentzaileak zuzendutako testua.

ELEKTROKARDIOGRAMA (EKG).	Erritmo sinusal normala.
GASTROSKOPIA.	Anestesia topikoa jarri ostean, gastroskopioa ahotik barrena sartuta: hestegorriak kalibre eta mukosa normala du; urdailaren zabalgarritasuna eta uzkurtasuna normalak dira; ez da odol-aztarnarik ikusten, ez eta lesiorik ere; piloro normala. Duodeno-
ABDOMENEKO OTA.	Gibela, barea, pankrea, behazuna, behazunbidea, giltzurrunen-gaineko guruina eta giltzurrunak normalak. Itsuaren paretaren gizentze bat ikusten da, neoplasia itxurakoa. Ez da adenopatiarik ikusten.
ANATOMIA PATOLOGIKOA.	Adenokartzinoma.

7.2 irudia – XuxenMed medikuntzako ortografia-zuzentzaileak zuzendutako testua.

XuxenMed Firefox nabigatzailerako gehigarri moduan prestatu dugu, eta hobetze prozesuan egonik, oraindik publikoki zabaldu ez badugu ere, gurekin harremanetan ipinita banatzeko prestutasun osoa dugu. Etorkizunean, Firefox nabigatzaileaz haratago, ohiko testu-editoreetan ere gehitzeko moduluak garatu nahi ditugu.

Ortografia-zuzentzailea egokitu ondoren, osasun-txostenak idazteko lehen prototipoa sortzeari ekin genion, eta hori da hurrengo atalean azalduko duguna.

7.2 Osasun-txostenak euskaraz idazteko laguntza-prototipoa

Atal honetan osasun-txosten elebidunak lortzeko lehen prototipoaren garapena aurkezten dugu. Prototipoa lortzeko zaintza klinikorako informazioa kudeatzeko sistema bat sortu dugu horretarako prestatuta dagoen *Innovative Clinical Information Management Systems*ek (iCIMS) garatutako *Clinical Care Information System*[1] softwareari esker. Sistema honen oinarrian dauden terminologia-zerbitzariak euskararako eta espainierarako sortu ditugu (*CliniTermServer* tresnaren moldapenak, jatorrizkoa *Health Language Analytics*-ek garatu duena). Prototipoa eraikitzeko, Donostia Unibertsitate Ospitaleko alta-txostenen bildumako (Joanes Etxeberri Saria V. Edizioa, 2014) "Digestio-aparatua" espezialitateko eredua erabili dugu. Horrela, euskaraz alta-txostenak idazteko eta modu errazean haren terminologiaren espainierako baliokideak lortzeko prototipoa eraiki dugu.

Aipaturiko alta-txostenen bilduman, Donostia Unibertsitate Ospitaleko espezialitate ezberdinetako alta-txostenen ereduak agertzen dira, euskaraz idatzitako alta-txostenen adibideak, hain zuzen ere. Hogeita hamar espezialitate eta berrogeita hiru alta-txosten barnebildurik daude bilduma horretan eta benetan erabiltzen diren egiturak jasotzen dira bertan. Gainera, esfortzu berezia egin dute egileek bertan ageri den euskarazko terminologia zuzena izan dadin.

Clinical Care Information System edo CCIS, zaintza klinikorako informazioa kudeatzeko sistemari deritzo iCIMS enpresak. Sistema honek gaixoaren informazioa kudeatzeko aukera zabaltzeaz gain, osasun-arloko langileek euren azterketa eta prozeduren ondorioz idatziz jaso nahi duten informazioa modu antolatu eta egituratuan idazteko aukera ematen du. Gainera, gaixoaren historia klinikoa modu bateratuan gordetzeko aukera ematen du eta horrela kudeaketa errazten du. iCIMSek CCISak diseinatzeko softwarea eskaintzen du *formBuilder* txantiloi-sortzaileari esker, eta ondorioz, egokitutako CCIS berriak sor daitezke modu errazean. Software hori oso malgua da aldaketak

[1] http://www.icims.com.au/solutions/products/ (2017ko maiatzaren 9an atzitua).

eta eguneraketak egiteko garaian.

Garatu dugun prototipoa CCIS sistema bat da, eta *formBuilder* txantiloi--sortzailea erabili dugu honen hezurdura diseinatzeko. Diseinu honetan, lehen txantiloia osatzeko Donostia Unibertsitateko Ospitaleko "digestio-aparatua" espezialitaterako alta-txosten eredua erabili dugu, dagoeneko esan dugun moduan.

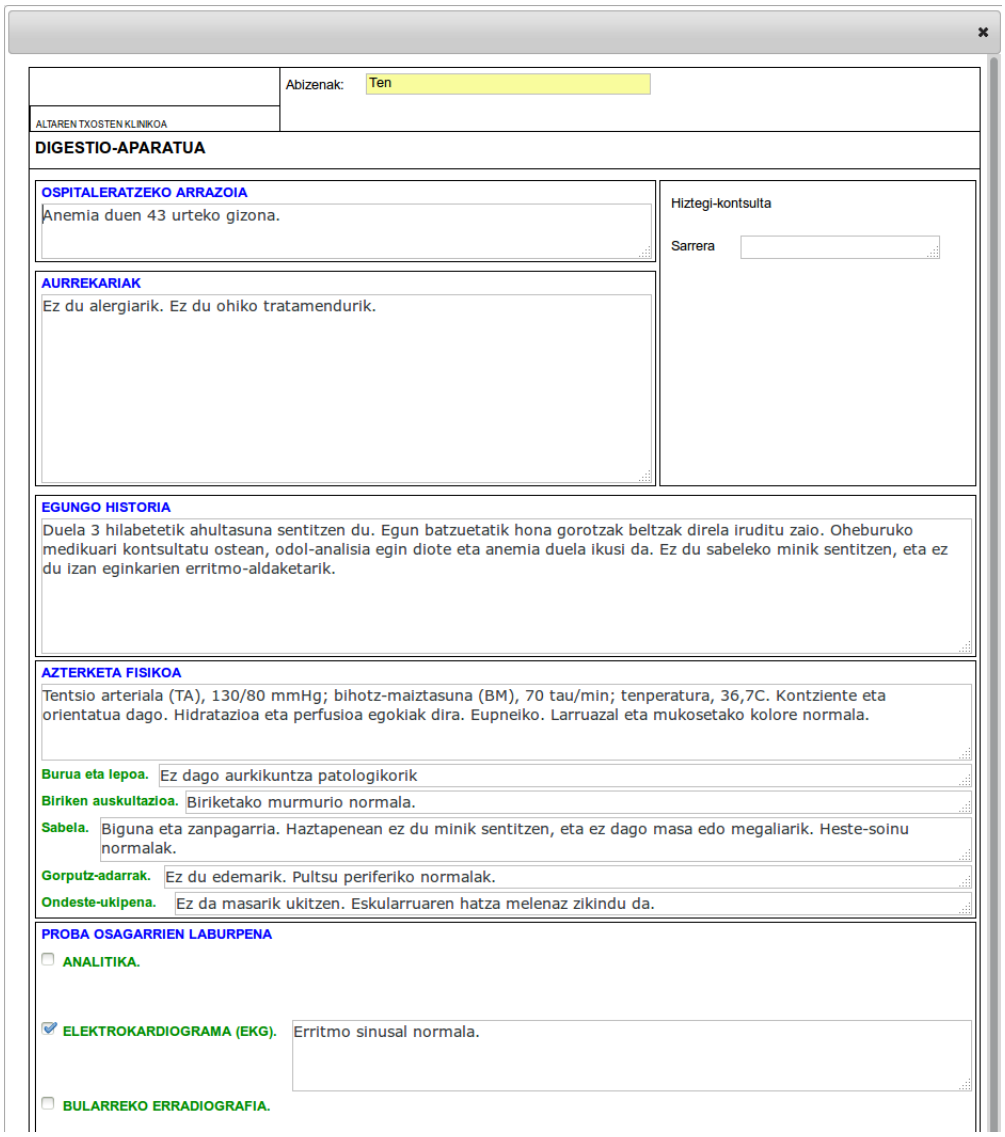
Alta-txostena idazteko sistemak duen txantiloiaren zati bat ikus dezakegu 7.3 irudian. Bertan ikus daitekeen bezala, atal eta azpiatal ezberdinak definituta daude ("Ospitaleratzeko arrazoia", "Aurrekariak" e.a.) horiek guztiak alta-txostenen bildumatik jasotako ereduei jarraituz. Atal batzuk hautazkoak izanik (ikusi "Proba osagarrien laburpena" atala), horiek klik erraz batean agertu edo desagertuko dira, alta-txostena betetzen ari den osasun-langileak beharretara egokitutako alta-txostena idatz dezan.

Gainera, osasun-langileen gomendioei jarraituz, atal batzuei eduki lehenetsia esleitu diegu. Eduki lehenetsi honek gaixoaren egoera "normala" adierazten du. Adibidez, "Aurrekariak" atalean, egoera normala "Ez du alergiarik. Ez du ohiko tratamendurik." izango litzateke, hau da, alteraziorik gabeko egoera. Modu horretan, osasun-langileak egoera normaletik at dagoen osasun-egoera topatzen duenean bakarrik aldatu beharko du eremuaren edukia, idazketa-lana sinplifikatuz. Eskakizun hori, langileen denbora-kudeaketari erantzuten dion behar bat da; izan ere, gaixoari eskaini ahal dioten denbora laburra da, eta egoeraren alteraziorik ez dagoenean osasun-txostenen atal guztiak bete behar izatea denbora-galeratzat hartzen dute.

Hiztegi elebidunean termino baten bilaketa egiteko eremua definitu dugu (ikus 7.3 irudian "Hiztegi-kontsulta" aukera). Dena dela, funtzionalitate hori ez da oraindik CCIS tresnan eskaintzen.

Prototipoaren azalpenekin jarraitu ahal izateko, aurrenik *CliniTermServer* terminologia-zerbitzariaren gainean egindako ekarpenak azalduko ditugu. Ikusi dugun bezala, CCIS sistemako txantiloian testu-kutxak daude. Testu--kutxa horietan automatikoki ezagutzen dira osasun-terminoak *CliniTermServer* terminologia-zerbitzariari esker.

*CliniTermServer*ek osasun-arloko terminologia eta beste hainbat pseudotermino eta zerrenda barnebiltzen ditu. Besteak beste, osasun-txostenetan maiz erabiltzen diren esamoldeak eta izen-sintagmak jasotzen dira, baita salbuespeneko jokabidea dutenak ere. Zerbitzari honek terminologia jasotzeaz gain, berau prozesatzeko softwarea ere eskaintzen du, esaterako, gordeta duen terminologia testuan identifikatzeko gai da, eta terminoarekin loturiko informazioa eskaintzen du.



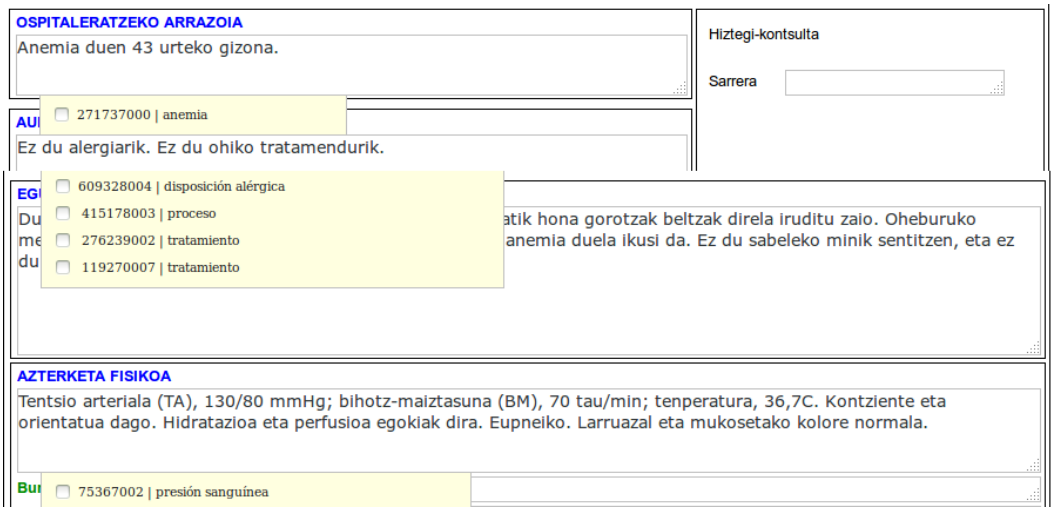
7.3 irudia – "Digestio-aparatua" espezialitaterako garaturiko CCIS sistema.

CliniTermServer terminologia-zerbitzaria ingeleserako bakarrik eskaintzen du *Health Language Analitics* enpresak, eta guk berau euskararako eta espainierarako egokitu dugu. Terminologia zerbitzari hori beste hizkuntzetara egokitzeko, hizkuntzaren menpekoak diren prozesamendurako tresnak

behar izan ditugu, tokenizatzailea eta lematizatzailea, hain zuzen ere.

CliniTermServer egokitzeko honako tresnak erabili ditugu: espainieraren kasuan hizkuntzaren prozesamendurako Freeling kode irekiko liburutegiaren 3.1 (Padró eta Stanilovsky, 2012) bertsioa izan da eta euskararako IXA taldeak garaturiko Eustagger etiketatzaile/lematizatzailea (Ezeiza *et al.*, 1998).

Hizkuntza prozesatzeko tresnak erabiltzeaz gain, eduki terminologikoz elikatu behar izan ditugu *CliniTermServer*eko espainierazko eta euskarazko zerbitzariak. Espainierarako sortutako zerbitzarirako, SNOMED CTren nazioarteko banaketaren espainierazko bertsioa jaso dugu, 2014ko apirilaren 30eko banaketa[2], hain zuzen ere. Euskarazko zerbitzarirako aldiz, tesi-lan honetan garatutako SNOMED CTren 2014ko apirilean genuen bertsioa erabili dugu, Perez-de Viñaspre eta Oronoz (2015) lanean argitaratutakoa, hain zuzen ere.



7.4 irudia – "Digestio-aparatua" espezialitaterako garaturiko CCIS sistema, termino klinikoak identifikatuta dituelarik.

Prestatu ditugun zerbitzari terminologikoek, bi funtzionalitate nagusi eskaintzen dizkigute: i) testuan agertzen diren SNOMED CTren termino edo deskribapenak identifikatzea eta ii) SNOMED CTren kontzeptu baten identifikadorea jaso eta honen adierak ematea. Aipatu moduan, hiru zerbitzari ezberdin prestatu ditugu: ingeleserako (HLA enpresak eskainia), euskararako

[2]http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html (2017ko maiatzaren 9an atzitua).

eta espainierarako (guk egokituak). Lan horretarako, hizkuntza ezberdineta-ko zerbitzariak konbinatu ditugu, eta euren artean komunikatzeko SNOMED CTren kontzeptu-identifikadorea erabili dugu, kontzeptu bat identifikatu, eta beste hizkuntza batean ordainak emateko. Horren adibideak 7.4 irudian ikus ditzakegu, euskarazko zerbitzariak testuan dauden terminoak identifikatu, eta espainierazko zerbitzariak horien espainierazko baliokideak ematen di-tuenak: "alergiarik" deskribapenerako *"disposición alérgica"* edota "tentsio arteriala" terminorako *"presión sanguínea"*.

Itzulpen prozesua nola egiten den ikusteko, azter dezagun "Ez du aler-giarik" esaldia 7.4 irudian. Euskarazko esaldia tokenizatu ostean, esaldia-ren tokenak euren lemarekin ("Ez izan alergia") ordezkatu dira. Ondoren, lema guztiak SNOMED CTren kontra parekatzen ditu zerbitzariak. Kasu honetan, "alergia" terminoak bakarrik jaso du SNOMED CTren kontzeptu--identifikadorea ("609328004"). Kontzeptu-identifikadore hori espainierazko zerbitzariari pasatakoan, *"disposición alérgica"* terminoa itzuli digu, hau da, espainierazko SNOMED CTn kontzeptu horri loturik dagoen hobetsitako ter-minoa.

Terminologia-zerbitzarien konbinazio honi esker, osasun-langileek esan-guratsuak diren terminoak aukeratu ahal izango dituzte, eta kontzeptu bati dagozkion ordain guztien artean egokia zein den aukeratzeko. Hau da, de-sanbiguatzeko aukera dute, itzulpen automatikoan akats gehien sortzen duen ataza sinplifikatuz.

Prototipo hau web-aplikazio bat izanik, 7.1 atalean aurkeztutako Xuxen-Med zuzentzaile ortografikoa erabili daiteke, idazketan lagungarria izango dena.

Laburbilduz, sortu dugun prototipoa iCIMS enpresaren *formBuilder* tres-narekin egin dugu, eta, ikusi dugun moduan, prototipoa sortzeko abantaila asko eskaintzen ditu, eramangarritasuna, aldakortasuna eta berehalakota-suna besteak beste. Hala ere, ez da guk garatu nahi dugun proiekturako diseinatua izan eta hainbat muga jartzen dizkio lortu nahi genukeen produk-tuari. Horregatik, iCIMSeko softwaretik ideiak jaso eta kode irekiko beste prototipo bat garatzen hasi gara. Garatzen ari garen softwareak oso modu errazean txantiloi berriak sortzea ahalbidetzen digu, eta dagoeneko AnaMed analizatzailea integratu diogu idazten den testuan SNOMED CTren termi-noak identifikatu ahal izateko.

7.3 Laburpena eta ondorioak

Tesi-lanean sortutako euskarazko osasun-arloko terminoak bi aplikaziotan erabili ditugu. XuxenMed zuzentzaile ortografikoan eta osasun-txostenak euskaraz idazteko prototipoan.

Bietan une zehatz batean, 2014 urtean, eskura genituen baliabideak erabili ditugu, baina modu errazean eguneratu daitezke aplikazio horien hiztegiak gaur egunean dugun lexikoarekin.

XuxenMed tresna xumea izanagatik ere, oso erabilgarria dela uste dugu. Hala ere, XuxenMeden terminoak integratu baditugu ere, ez dugu ebaluazio berezirik egin bere funtzionamendu egokia egiaztatzeko. Gure ustez, etorkizunari begira, Xuxenen egin zenari jarraiki, lexikoaren eguneraketetan akatsak gertatu diren egiaztatzeko zerrendak sortu beharko lirateke, aurrez ondo egiten zena ez galtzeko eta maiztasun handiko hitz edota terminoen prozesaketa egokia bermatzeko, dela erabilera orokorreko zein teknikoko hiztegia erabiltzaileak darabilkina.

Osasun-txostenak idazteko sistemak ordea bide luzeagoa du. Esan dugun moduan, kode irekiko bideari ekin diogu eta etorkizunean idazketa prozesua errazteko baliabide gehiago gehitu nahi dizkiogu, adibidez, terminoak idatzi ahala automatikoki osatzeko proposamenak luzatzea edo SNOMED CTren eduki eleanitzean termino baten ordainak bilatzeko aukera izatea. Eta jakina, osasun-txostenen itzulpen automatikoaren ikerketa-lerroari ekin nahi diogu.

8

Ondorioak eta etorkizuneko lanak

Tesi-lan honetan hizkuntza minorizatu batentzako osasun-alorreko terminologia automatikoki sortzeko algoritmo bat diseinatu eta garatu dugu. Horrela, baliabide lexikalak berrerabiltzeaz haratago, corpus paralelorik gabe, terminoak sistematikoki sortzeko bi sistema garatu ditugu (NeoTerm eta KabiTerm), eta Matxin itzultzaile automatikoari domeinuetara egokitzeko funtzionalitatea gehitu diogu. Horretaz gain, osasun-txostenak euskaraz idazteko laguntzak ere inplementatu ditugu.

Terminologia euskaratzeko, SNOMED CTren eduki terminologikoa hartu dugu erreferentzia moduan. Osasun-zientzietako domeinuan, SNOMED CT da baliabide terminologiko kliniko eleaniztun osatuena, eta zehaztasun eta estaldura handia dauka. Mundu osoan da erabilia gaur egun, eta informazio klinikoa kodetzeko, eraузteko edota aztertzeko erabiltzen da.

Jarraian, tesi-lanean zehar egindako lanen ondorio nagusiak (8.1 atala) eta tesi-lanaren garapenetik egindako ekarpen nagusiak (8.2 atala) zerrendatuko ditugu. Bukatzeko, etorkizunean tesi-lan honek irekita utzi dizkigun bideak eta ikerketa-lerroak azalduko ditugu 8.3 atalean.

8.1 Ondorio nagusiak

Gure helburu nagusia osasun-zientzien domeinuko testuak automatikoki prozesatzeko euskararako baliabideak sortzea izan da. Horretarako, terminologia euskaratzea urrats ezinbestekoa dela deritzogu, eta hori izan da tesi-lan honen eginkizun nagusia. Literaturan argitaratutako lanak aztertu ditugu, eta

tesi-lanarekin hasi aurretik corpusetan oinarritutako lanak aurkitu ditugu, nagusiki. Euskara bezalako hizkuntza minorizatu baterako teknika horiek ez dira aplikagarriak, osasun-zientzietako corpus paralelo zein konparagarririk ez dagoelako. Horrenbestez, corpusak erabiltzen ez dituzten teknikak proposatu eta garatu ditugu tesi-lan honetan. Jarraian lanaren ezaugarri nagusiak laburbiltzen ditugu:

Hurrengo lerroetan ondorio nagusiak zerrendatuko ditugu gaiaren arabera sailkatuta.

- **Iturria:** Terminologia euskaratzeko teknikak edo sistemak garatu aurretik, SNOMED CT bera aztertu dugu, analisi kuantitatiboa eginez. SNOMED CT eleaniztuna izanik, iturri moduan ingelesa edo espainiera hartuko dugun erabakitzen lagundu digu analisi horrek. Kasu honetan, ingelesezko bertsioa hartu dugu abiapuntutzat, SNOMED CTren jatorrizko bertsioa izateaz gain, tesi-lana hasi genuenean espainierazko bertsioa ez zelako egonkorra, eta ondorioz gabeziak zituelako. Erreferentziazko hierarkiak ere aukeratu ditugu, eta hierarkia populatuenekin hastea erabaki dugu: aurkikuntza klinikoak (eta nahasmenduak, jakina), prozedurak eta gorputz-egiturak. Hierarkia horietarako eman ditugu emaitzak baina hierarkia guztietako terminoak euskaratu ditugu. SNOMED CTk osasun-txostenetan esanguratsua den edozein informazio ere jasotzen badu ere, hierarkia populatuenak eta aldi berean terminologia klinikoarekin lotura estuena dutenak aukeratu ditugu.

- **SNOMED CTren euskaratzea kudeatzen duen sistema**: SNOMED CT euskaratzeko EuSnomed sistema diseinatu eta garatu dugu. Sistema horren baitan, terminologia euskaratzeko lau urratsetako algoritmoa diseinatu dugu. Lehenengo urratsak baliabide lexikal espezializatu eta elebidun/eleaniztunak erabiltzen ditu ordainak lortzeko. Bigarrenak, termino neoklasikoak euskaratzen ditu, afixuen baliokideen eta transliterazioaren bidez. Hirugarren urratsa, termino konplexuen egitura habiaratuan oinarritzen da ordainak sortzeko patroiak definitzeko. Azkenik, laugarren urratsak, itzultzaile automatiko orokor bat egokitzen du termino konplexuen ordainak lortzeko. Algoritmoa inplementatzeaz gain, sistema honek informazioaren biltegiratzeaz ere arduratzen da, baita sortutako ordain berrien berrerabiltzeaz ere.

- **Termino sinpleen euskaratzea**: Lehenengo bi urratsak bereziki termino sinpleak euskaratzeko diseinatu baditugu ere, baliabide lexikalen

kasuan edozein luzeratako terminoak euskaratzeko gai gara. Erabilitako baliabide lexikalen artean, SNOMED CTren euskaratzean emaitza hoberenak lortu dituztenak ZT Hiztegia (0,99ko doitasuna), Euskalterm terminologia bankua (0,89ko doitasuna) eta Erizaintzako Hiztegia (0,94ko doitasuna) izan dira adituen ebaluazioaren arabera. Giza Anatomiako Atlasa doitasunari dagokionez ez bada ere horren nabarmena izan, gorputz-egituren euskaratzean ekarpena nabarmena da. Termino neoklasikoak euskaratzeko, NeoTerm deituriko sistema sortu dugu, eta honen hiru hurbilpen garatu ditugu. Lehenengo hurbilpena oinarrilerro sistema da, afixu neoklasikoen konposaketan oinarritzen dena. Hurbilpen honek doitasun altua izan arren (0,89), estaldura txikia du (0,34). Bigarren hurbilpenean, estaldura hobetzea izan dugu lehentasuna. Estaldura hobetzeari begira, bigarren hurbilpenean transliterazio modulua integratu dugu, afixuen hiztegiak zabaltzearekin batera. Doitasunean emaitza kaxkarragoak lortu baditugu ere (8 puntu gutxiago), estaldura asko igo da (48 puntu), estaldura eta doitasunaren arteko oreka lortuaz (0,81eko F-neurria). Azkeneko hurbilpenean, termino neoklasikoen identifikazioa findu nahi izan dugu, termino neoklasikoak ez diren terminoak NeoTermek bazter ditzan, erroreak ekiditeko. Horretarako, identifikaziorako algoritmoa findu dugu, adituek proposaturiko irizpideak kontuan izanik. Hurbilpen honekin, aldiz, ez dugu emaitzak hobetzea lortu, eta bigarren hurbilpenarekin alderatuta, doitasuna bere horretan geratu bada ere, estaldurak 7 puntu behera egin du. Emaitzak aztertuta, NeoTermen bigarren hurbilpena da EuSnomeden integratu duguna, transliterazio-moduluan oinarritzen dena, alegia. Euskaratze-algoritmoaren bi urrats horiekin SNOMED CTren termino sinpleen % 75etik gora euskaratzea lortu dugu erreferentziazko lau hierarkietan (nahasmenduak, aurkikuntzak, gorputz-egiturak eta prozedurak).

- **Termino konplexuen euskaratzea**: Termino konplexuak euskaratzeko, aldiz (algoritmoaren azken bi urratsak), termino habiaratuetan oinarritzen den KabiTerm sistema garatu dugu, eta Matxin Itzultzaile Automatikoa osasun-zientzien domeinura egokitu dugu MatxinMed deitu dugun bertsioan. KabiTermek termino konplexuen barruan agertzen diren beste terminoek osatzen duten egitura baliatzen du euskaratze-patroi batzuen bitartez euskarazko ordaina sortzeko. Termino habiaratu horien SNOMED CTren hierarkiak erabili ditugu terminoen egiturak aztertzeko eta euskaratze-patroiak definitzeko. KabiTerm eta

MatxinMed Medbaluatoia deritzon kanpainan ebaluatu ditugu, eta artearen egoerako Google Translate tresnarekin konparatu ditugu. Emaitza oso positiboak izan dira, bai parte-hartzeari dagokionez, baita gure sistemen kalitateari dagokionez Google-ekin alderatzean. Bereziki azpimarragarriak dira KabiTermek lortutako emaitzak, gainerako sistemei nabarmenki irabazi baitie. Medbaluatoiak euskal osasun-zientzien komunitatea inplikatu du sistemen ebaluazioa egiteko. Horrela, aditu talde txiki batek ebaluatu beharrean, komunitate zabalagoak ebaluatu ditu KabiTerm eta MatxinMed. Ebaluazioak oso erantzun positiboa izan zuen komunitatearen aldetik (217 parte-hartzaile eta 15.000 ebaluaziotik gora), eta lanean jarraitzeko irrika zabaldu digute. Medbaluatoiak euskal osasun-zientzien komunitatea inplikatzeko tresna oso erabilgarria dela frogatu du, eta etorkizunean antzeko ekintzak abiatzea aurreikusten dugu, SNOMED CTren euskarazko bertsioaren baliozkotzea eta zuzentzea helburu, adibidez.

- **SNOMED CTren euskaratzearen estaldura-emaitzak**: EuSnomed sistemaren estaldura-emaitza orokorrak 8.1 taulan ikus ditzakegu (doitasunari dagozkionak kapituluetan zehar eman ditugu). Bertan ikus dezakegunez, MatxinMedek aurreko urratsetan euskaratu gabe gelditu diren termino guztiak euskaratzen ditu (hitz bakarreko gehienak bere horretan uzten baditu ere), eta horrela SNOMED CTren euskarazko *alpha* bertsio osatua lortu dugu. MatxinMed alde batera utzita, lortutako emaitzak token kopuruen arabera ikus ditzakegu 8.2 taulan. Ikusten dugunez, token bakarreko termino gehienak termino sinpleak euskaratzeko teknikek euskaratzen dituzte, eta aurkikuntzen hierarkia kenduta, bi tokeneko terminoen estaldura altua lortzen dute. Azpimarratzekoa da, KabiTerm garatzeko landu ditugun egiturak gehienak nahasmenduak eta gorputz-egiturak deskribatzeko terminoak izan direla, eta horregatik bi hierarki horietarako lortzen ditugu estaldura daturik altuenak (% 30 inguru). Izan ere, nahasmenduen termino askoz gehiago ditugu beste hierarkietakoak baino, eta gorputz-egituren kasuan, termino konplexuen egitura oso errepikakorra da.

	Nahasmenduak			Aurkikuntzak			Gorputz-egiturak			Prozedurak		
	Eus.	**Estal.**	**Ord.**	**Eus.**	**Estal.**	**Ord.**	**Eus.**	**Estal.**	**Ord.**	**Eus.**	**Estal.**	**Ord.**
GNS10	11.060	-	-	2.555			-	-	-	-	-	-
Bal. lex.	5.750	0,050	6.975	1.635	0,030	2.343	4.667	0,079	6.599	747	0,086	981
NeoTerm	1.868	0,016	1.951	622	0,001	717	676	0,011	726	1.266	0,015	1.308
KabiTerm	26.136	0,228	102.724	4.054	0,077	15.868	12.497	0,210	43.913	10.651	0,122	34.232
MatxinMed	81.076	0,706	81.076	46.539	0,881	46.539	41.535	0,699	41.535	74.436	0,855	74.436
Denera	114.830	1,000	180.952	52.857	1,000	95.089	59.384	1,000	92.029	87.104	1,000	109.200

8.1 taula – EuSnomeden SNOMED CTren euskaratze-estaldura.

		token1	2token	3token	4token	≥5token	Denera
Nahasmen.	**Eusk.**	3.265	8.335	9.966	5.614	6.574	33.754
	Den.	3.865	21.003	25.038	20.757	44.167	114.830
	Estal.	0,845	0,397	0,398	0,271	0,149	0,294
Aurkikun.	**Eusk.**	1.449	2.442	1.413	568	439	6.311
	Den.	1.940	9.737	11.906	11.317	24.640	59.540
	Estal.	0,747	0,251	0,119	0,050	0,018	0,106
Gorputz--egiturak	**Eusk.**	1.907	4.308	4.631	3.444	3.550	17.840
	Den.	2.592	10.863	12.599	10.635	22.695	59.384
	Estal.	0,736	0,397	0,368	0,324	0,156	0,300
Prozedur.	**Eusk.**	1.698	3.295	2.744	2.456	2.471	12.664
	Den.	1.985	9.892	15.399	17.082	42.746	87.104
	Estal.	0,855	0,333	0,178	0,144	0,058	0,145

8.2 taula – Jatorrizko ingelesezko terminoen token kopuruaren araberako emaitza orokorrak estaldurari dagokionez.

- **Sortutako gainerako aplikazioak**: EuSnomed sistemaren baitan terminologia euskaratzeko garatutako sistemez gain, euskarazko SNOMED CTren erabilgarritasuna frogatu nahi izan dugu osasun-txostenen idazketarako laguntzak garatuz. Xuxen zuzentzailearen medikuntzarako bertsioa (XuxenMed) sortu dugu, euskarazko ordainak hiztegiratu ditugularik. Gainera, euskaraz osasun-txostenak idazten laguntzeko prototipo bat sortu dugu, eta bertan idatzitako SNOMED CTren terminoak identifikatzen ditugu.

Osasungintza euskalduntzeko lan handia egiteke badago ere, hemen aurkeztu dugun lana argi-izpi bat izan daiteke. Oraindik orain osasun-txosten elebidunak sortzeko proiektua hastapenetan badago ere, oinarri itxaropentsua ezarri dugu etorkizunean osasun-txosten elebidunak gure osasun-sistemaren egunerokoan erabili ahal izateko. Modu honetan, osasun-langile eta gaixo euskaldunon hizkuntza-eskubide batzuk bermatzeko bidea erraz daiteke.

8.2 Ekarpenak

Tesi-lan honetan egin dugun ekarpen nagusia terminologiaren sorkuntza automatikorako teknika berriak garatzea izan da. Literaturan corpus elebidun zein elebakarretan oinarritutako teknikak erabiltzen dituzten lan asko aurki

badaitezke ere, gutxi dira hurbilpen horretatik at dauden lanak. Tesi-lan honekin, euskara bezalako hizkuntza minorizatuetarako interesgarriak izan daitezkeen erregeletan oinarritzen diren sistemak garatu ditugu. Gainera, garatu ditugun metodo automatikoak erabilgarriak direla frogatu dugu, itzultzaile eta adituen lana errazten duelako.

Ekarpen nagusi horretaz gain, ekarpen gehiago ere egin ditugu tesi-lan honetan zehar, eta jarraian zerrendatuko ditugu:

- **SNOMED CTren euskaratzea kudeatzen duen EuSnomed sistema garatu dugu.** (4. kapitulua)

  Euskaratze-algoritmoa inplementatzen duen EuSnomed sistema garatu dugu. Algoritmoa inplementatzeaz gain, euskaratze-prozesu osoa kudeatzen du, baliabide lexikalen integraziotik, emaitzen kalkuluak egitera. Sistema, exekuzioan lortzen dituen euskarazko ordain berriak berrerabiltzeko diseinatuta dago, eta, horrela, algoritmoaren urrats bakoitzak aurreko urratsak sortutako ordainak ere berrerabil ditzake. Kode guztia GitHuben eskuragarri dago[1].

- **Termino neoklasikoak euskaratzeko NeoTerm sistema garatu dugu.** (5. kapitulua)

  NeoTerm ingelesezko termino neoklasikoak euskaratzeko erregeletan oinarritzen den sistema da. Horretarako, osasun-zientzietako termino neoklasiko horien afixuen hiztegi elebiduna sortu dugu eta transliterazio erregelak definitu ditugu, modu errazean. Kode hau ere GitHuben eskuragarri dago[2]. Afixuen ingeles-euskara pareak hainbat adibideekin batera web-orri batean jarri ditugu eskuragarri eta NeoTermen demoa ere erabil daiteke[3].

- **Termino habiaratuetan oinarrituz, termino konplexuak euskaratzen dituen KabiTerm sistema garatu dugu.** (6. kapitulua)

  Termino konplexuak euskaratzeko KabiTerm sistema sortu dugu. Horrek, ingelesezko termino konplexuak euskaratzen ditu, termino konplexuen barruan agertzen diren termino habiaratuen egitura baliatuz.

[1] https://github.com/olatz87/euSnomed
[2] https://github.com/olatz87/NeoTerm
[3] http://ixa2.si.ehu.eus/neoterm/ (2017ko maiatzaren 9an atzitua).

Hau da, termino habiaratuek osatzen dituzten egituretarako euskaratze-patroi batzuk definitu ditugu, euskarazko termino konplexuen egitura sortzeko. Kode guztia GitHuben eskuragarri dago[4].

- **Matxini domeinu zehatzera egokitzeko funtzionalitatea gehitu diogu eta MatxinMed sortu dugu.** (6. kapitulua)

  Matxin erregeletan oinarritzen den Itzultzaile Automatikoa da, testua espainieratik euskarara eta ingelesetik euskarara itzultzen duena. Guk Matxini domeinuetara egokitzeko funtzionalitatea gehitu diogu. Domeinu bat baino gehiago aukera daiteke, horien arteko lehentasunak zehaztuz. Funtzionalitatea gehitzeaz gain, osasun-zientzietarako bertsioa garatu dugu, MatxinMed, horretarako SNOMED CTren eduki terminologikoa integratu diogularik. Kodea laster izango da eskuragarri Matxinen GitHub orrian[5].

- **Osasun-zientzien euskal komunitatea Medbaluatoia kanpainan inplikatu dugu.** (6. kapitulua)

  Termino konplexuen euskaratzea ebaluatzeko, osasun-zientzien euskal komunitatea inplikatzen duen ebaluazioari heldu diogu. Horretarako Ebaluatoiaren moldaera egin dugu, eta Medbaluatoia deritzon kanpaina diseinatu eta martxan jarri dugu. Kanpainak oso erantzun positiboa izan zuen komunitatearen aldetik eta horrelako iniziatiba gehiagoren beharra agerian utzi dugu.

- **Osasun-zientzietarako analizatzailea (AnaMed) eta SNOMED CTn oinarritzen den terminologia-zerbitzaria (TermZerSCT) garatu ditugu.** (6. kapitulua)

  AnaMed osasun-zientzietarako hizkuntza-analizatzailea da, informazio linguistikoa analizatzeaz gain, SNOMED CTren terminoak testuan identifikatzen dituena, baita eponimoak ere. TermZerSCT, aldiz, terminologia-zerbitzaria da, zeinak SNOMED CTren inguruko informazioa kudeatzea ahalbidetzen duen, prozesamendu-denbora minimoarekin. Kode guztia GitHuben dago eskuragarri[6,7].

[4] https://github.com/olatz87/KabiTerm
[5] https://github.com/matxin/matxin
[6] https://github.com/olatz87/anaMed-en
[7] https://github.com/olatz87/TermZerSCT

- **SNOMED CTren euskarazko *alpha* bertsioa sortu dugu.**

  Euskaratze-algoritmoari esker, SNOMED CTren euskarazko *alpha* bertsioa sortu dugu, metodo erabat automatikoekin. SNOMED CTren euskarazko lehen bertsioa izateak interes handia piztu du, bereziki Osakidetzaren Euskara Zerbitzuan, eta euskarazko SNOMED CT gainbegiratua sortzeko hitzarmena sinatu dugu Osakidetzak eta IXA taldeak lankidetzan.

- **Euskarazko osasun-zientzien corpusa osatu dugu.** (6. kapitulua)

  Tesiarekin hasi ginen unean, ez zegoen euskarazko osasun-zientzien domeinuko inolako corpusik. MatxinMeden garapenerako, osasun-zientzien euskarazko hizkuntza-eredu baten beharra izan genuen, sinonimoen artean "hoberena" aukeratu ahal izateko. Horretarako, iturri ezberdinetatik izaera ezberdinetako testuak bildu ditugu, testu-liburuetatik hasita ikasleen apunteetara. Tamalez, ezin dugu corpusa argitaratu, jabetza- eta lizentzia-arazoak direla medio, baina sortu dugun hizkuntza-eredua argitaratu dezakegu, eta eskuragarri jarriko dugu laster.

- **Zuzentzaile ortografikoa osasun-zientzietara egokitu dugu.** (7. kapitulua)

  XuxenMed deitu dugun zuzentzaile ortografikoa Xuxen zuzentzailearen egokitzapena da. Momentuz, Firefox nabigatzailerako plugina prestatu dugu bere *alpha* bertsioan eta eskaerapean banatzen dugu. Azken datuekin eguneraketa egiten dugunean, plugina publikoki eskuragarri jarriko dugu.

- **Euskarazko osasun-txostenen idazketarako laguntzaile baten prototipoa sortu dugu.** (7. kapitulua)

  Osasun-langile euskaldun askok adierazi dutenez, osasun-txostenak euskaraz idazteko zailtasunak dituzte. Hori dela eta, Donostia Ospitaleko Alta Txostenen bilduma oinarri harturik, idazketarako bi prototipo garatu ditugu. Lehenengorako, iCIMS enpresak garatutako softwarea erabili dugu. Kode jabeduna izanik, ezin dugu kode hori zabaldu. Bigarrena, aldiz, Django-n oinarritzen den web-aplikazioa da, eta kode guztia GitHuben eskuragarri dago[8].

[8] https://github.com/olatz87/OsatEus

- **Osasun-txosten elebidunen ikerlerroa ireki dugu eta gaia mahai gainean jarri dugu.** (7. kapitulua)

  Osasun-zientzietan, eta bereziki alor klinikoan, euskararen normalizazioa lortu nahi badugu, euskarak osasun-txostenetan presente egon behar du. Euskal Herriko osasun-sistema guztietan langile elebakar asko dagoenez, euskara hutseko osasun-txostenek gaixoaren segurtasuna kolokan jar dezakete. Horretarako, osasun-txostenen denbora errealeko itzulpen erdi-automatikoa egitea proposatu dugu, SNOMED CT terminologia eleaniztunaren iturria delarik. Hizkuntza kontrolatua definitzeko beharra ikusten dugu itzulpenaren kalitatea bermatu ahal izateko. Ikerlerro hau oraindik hastapenetan egon arren, Osakidetzak interes handia erakutsi du honek aurrera egin dezan.

- **Euskara oinarri hartuta, osasun-arloko langileak eta informatikariak elkarlanean ipini ditugu.**

  Osasun-arloko langileekin elkarlanerako zubiak eraiki ditugu, euskara normalizatzeko helburuarekin. Elkarlan oso emankorra izan da, eta etorkizunean garatuko ditugun proiektu eta ideia asko definitzeko baliagarria izan zaigu. Guri, informatikarien ikuspegitik beraien ezagutza ezinbestekoa izan zaigu, eta beraiei, ezagutzen ez zituzten ideia eta aukera berriak ezagutzea ekarri die.

8.3 Etorkizuneko lanak

Batetik, tesi-lan honetan guztiz amaitu gabe gelditu diren lanak, eta, bestetik, lan honi jarraipena emateko dauden ikerlerroetako batzuk zerrendatuko ditugu jarraian:

- **SNOMED CTren euskarazko bertsio egonkorra lortzea, adituek banan-banan kontzeptuak gainbegiratuta**.

  Jakina da automatikoki proposaturiko terminoak ez direla zuzenak izango ehuneko ehunean. Guk tesi-lan honetan sortu dugun SNOMED CTren bertsio automatikoa adituek errepasatu beharko dute, eta kasuan kasu terminoak zuzendu, berriak proposatu edota automatikoki proposaturikoa berretsi beharko dute. Gainera, SNOMED CTren bertsio ofiziala osatzeko, kontzeptu bakoitzaren sinonimoen artean bat aukeratu beharko dute adituek hobetsi gisa.

- **Automatikoki proposaturiko terminoen eta adituek aukeraturiko terminoen arteko aldakortasuna neurtzea**.

  Adituek gainbegiratutako SNOMED CTren euskarazko bertsio egonkorra lortzen dugunean, guk automatikoki proposaturiko termino edo deskribapenen eta adituek aukeratutakoen arteko aldakortasuna neurtu ahal izango dugu. Honek oso ikerketa-gai interesgarria irekitzen du, eta tesi-lan honetan garatutako sistemen sakoneko ebaluazioa berma lezake.

- **Akronimoen itzulpen automatikoa egitea**.

  Akronimoen itzulpena landu gabe geratu zaigu tesi-lan honetan. Esku artean dugun domeinuan berebiziko garrantzia dute, ordea, akronimoek.

- **AnaMed analizatzailea espainierara egokitzea**.

  Dagoeneko AnaMed ingeleserako eta euskararako garatu dugu eta oso interesgarria litzateke espainierara ere egokituko bagenu. Izan ere, Hego Euskal Herrian idazten diren osasun-txostenetan gaztelania da nagusi, eta berauek prozesatzen lagungarria izan daiteke AnaMed. IXA taldeak garatutako FreelingMed tresna izango dugu abiapuntu, dagoeneko SNOMED CTko terminoak identifikatzen baititu.

- **KabiTermen, MatxinMeden eta AnaMeden demoak sortzea.**

  Sortu ditugun tresnak erabiltzaileen esku utzi nahi ditugu, eta horretarako demoak prestatu nahi ditugu. Kodea eskuragarri egon arren, osasun-arloko langileentzat ez da baliagarria, eta demo batekin gure sistemen indarguneak ikusi eta baliatu ahalko dituzte.

- **EuSMT sistema domeinura egokitzea**.

  Hiztegi elebidunak eta corpus elebakarrak erabilita, EuSMT itzultzaile estatistikoa, osasun-zientzien domeinura egokitu nahi dugu, eta MatxinMeden emaitzekin alderatu. Corpus paraleloa lortu bitartean zaila izango da estatistikan oinarritutako itzultzaile automatikoekin lan egitea, baina lan oso interesgarriak aurkitu ditugu corpus elebakarretan eta hiztegietan oinarritutakoak. Kontuan izan behar dugu Ebaluatoia kanpainan sistema honek jaso zuela emaitzarik onena, eta emaitza on horiek domeinura ekarri nahi ditugu.

- **Sistema beste baliabide terminologiko batzuk euskaratzeko egokitzea**.

  EuSnomedentzat diseinatu dugun algoritmoa beste baliabide terminologiko batzuk euskaratzeko erabili nahi dugu. Izan ere, Osakidetzako Euskara Zerbitzuak interes berezia dauka GNS10 euskaratzeko, eta proiektu bat abiatu dugu gure tresna egokitzeko. Horretarako, KabiTermerako patroi berriak definitu beharko ditugu GNS10ek dituen terminoen egitura oso berezia delako (sailkapenera bideratuta dago). Adibidez *"Salmonella infection, unspecified"* moduko deskribapenak oso maiz ikus ditzakegu, eta SNOMED CTn ez.

- **XuxenMed ohiko testu-editoretan erabiltzeko prestatzea**.

  Momentuz, XuxenMed Firefox nabigatzailean erabiltzeko baino ez dugu prestatu. Etorkizunean, XuxenMeden hiztegia eguneratzeaz gain, testu-editoreetan erabiltzeko moduluak garatu nahi ditugu.

- **Euskarazko SNOMED CT komunitatean baliozkotzeko sistema prestatzea**.

  Medbaluatoia SNOMED CTren terminologia balidatzeko egokitu nahi dugu. Izan ere, kanpainan zehar hainbat parte-hartzailek zuzentzeko aukera eskatu ziguten, eta gehiagorako gogoarekin geratu ziren. Jendearen prestutasuna kontuan izanik, euskarazko SNOMED CT osoaren zuzenketa egin nahiko genuke horrelako kanpainaren bidez. Horrela, jende gehiagoren oniritziarekin esleitu ahalko dizkiogu SNOMED CTri sinonimoak. Termino hobetsia aukeratzeko ataza aditu talde txiki baten esku utziko genuke, terminologoak eta osasun-langileak barnebilduz.

- **Osasun-txostenen itzulpen automatikoa egitea**.

  Alde batetik, garatu dugun prototipoaren kode irekiko bertsio malguagoa sortu nahi dugu. Izan ere, iCIMSeko *formBuilder*ek prototipoa sortzeko abantaila asko eskaintzen dizkigu, eramangarritasuna, aldakortasuna eta berehalakotasuna besteak beste. Hala ere, ez da guk garatu nahi dugun proiekturako diseinatua izan eta hainbat muga jartzen dizkio lortu nahi genukeen produktuari. Horregatik, iCIMSeko softwaretik ideiak jaso eta kode irekiko beste prototipo bat garatzen hasi gara, itzulpen automatikoa ahalbidetzeko, hiztegi dinamikoetan

aurka kontsultak bermatzeko, terminologia berria identifikatu ahal izateko, etab.

Euskara normalizazio-prozesuan dagoen hizkuntza izanik, termino berriak etengabe sortzen dira. Gure nahia da prototipoak automatikoki identifikatu ez dituen terminoak terminologia-zerbitzarira eskuz gehitzeko aukera ematea. Horrela, termino edota adiera berriak eskuz identifikatu ahal izango ditugu eta euskarazko SNOMED CT aberastu.

Dagoeneko lehen urratsak eman ditugu aplikazio berri honetan, eta, TermZerSCT eta AnaMed sistemak erabiliz, euskarazko testuetan terminoak identifikatzeko eta horien ordainak emateko gai gara.

- **Eta bukatzeko, guztion etorkizuneko lana da tesi-lan honetan sortutako euskarazko SNOMED CTren bertsioa hobetu eta erabiltzea, gaixoen egunerokoan euskarazko terminoak naturaltasunez txertatzeko.**

Bibliografia

Abdoune H., Merabti T., Darmoni S.J., eta Joubert M. Assisting the Translation of the CORE Subset of SNOMED CT Into French. In Moen A., Andersen S.K., Aarts J., eta Hurlen P., editors, *Studies in Health Technology and Informatics*, 169 lib., 819–823, 2011.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., eta Urizar R. A framework for the automatic processing of Basque. *Proceedings of the Workshop on Lexical Resources for Minority Languages*, Granada, Spain, 1998.

Agirre E., Alegria I., Arregi X., Artola X., de Ilarraza A.D., Maritxalar M., Sarasola K., eta Urkia M. Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. *Proceedings of NAACL-ANLP'92, 119-125. Povo Trento. 1992.*, 1992.

Al-Haj H. eta Lavie A. The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. *Machine translation*, 26(1-2):3–24, 2012.

Alani H., Kim S., Millard D.E., Weal M.J., Hall W., Lewis P.H., eta Shadbolt N.R. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: a General Lexical Basis for the Automatic Processing of Basque. *IRCS Workshop on linguistic databases. Philadelphia (USA).*, 2001.

Alegria I., Artola X., Sarasola K., eta Urkia M. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203, 1996.

Alegria I., Cabezon U., de Betono U.F., Labaka G., Mayor A., Sarasola K., eta Zubiaga A. Reciprocal enrichment between basque Wikipedia and machine translation. *The People's Web Meets NLP*, 101–118. Springer, 2013.

Amberger J., Bocchini C., eta Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human mutation*, 32(5):564–567, 2011.

Ananiadou S. A methodology for automatic term recognition. *Proceedings of the 15th conference on Computational linguistics-Volume 2*, 1034–1038. Association for Computational Linguistics, 1994.

Andersen U., Lerche J., Petersen P.G., Bernstein K., *et al.*. Adapting SNOMED CT for use in Denmark-the tools and the process of concept based translation. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 2613. IOS Press, 2007.

Aranberri N. Ebaluatoia: crowd evaluation of English-Basque machine translation. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2016.

Aranberri N., Labaka G., de Ilarraza A.D., eta Sarasola K. Exploiting portability to build an RBMT prototype for a new source language. *Proceedings of EAMT*, 2015.

Aranberri N., Labaka G., de Ilarraza A.D., eta Sarasola K. Ebaluatoia: crowd evaluation for English–Basque machine translation. *Language Resources and Evaluation*, 1–32, 2016a.

Aranberri N., Labaka Intxauspe G., Jauregi O., Díaz de Ilarraza Sánchez A., Alegría Loinaz I., eta Agirre Bengoa E. Tectogrammar-based machine translation for English-Spanish and English-Basque. *Procesamiento del Lenguaje Natural*, 73–80, 2016b.

Arregi X., Arruarte A., Artola X., Lersundi M., Santander G., eta Umbelina J. TZOS: Terminologia Zerbitzurako On-line sistema. In UPV/EHU E.H.U., editor, *Ugarteburu Terminologia Jardunaldiak 2010*, 136–153, 2010.

Artetxe M. Distributional semantics and machine learning for statistical machine translation. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2016.

Artstein R. eta Poesio M. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Bakhshi-Raiez F., Cornet R., eta F. de Keizer N. Development and Application of a Framework for Maintenance of Medical Terminological Systems. *Journal of the American Medical Informatics Association, JAMIA*, 15(5): 687–700, 2008.

Banay G. An introduction to medical terminology, Greek and Latin derivations. *Bulletin of the Medical Library Association*, 36(1):1–27, Jan 1948.

Bauer L. *English word-formation*. Cambridge university press, 1983.

Beesley K.R. Arabic finite-state morphological analysis and generation. *Proceedings of the 16th conference on Computational linguistics-Volume 1*, 89–94. Association for Computational Linguistics, 1996.

Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

Bojar O., Buck C., Federmann C., Haddow B., Koehn P., Leveling J., Monz C., Pecina P., Post M., Saint-Amand H., *et al.*. Findings of the 2014 workshop on statistical machine translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 12–58. Association for Computational Linguistics Baltimore, MD, USA, 2014.

Bollegala D., Kontonatsios G., eta Ananiadou S. A Cross-Lingual Similarity Measure for Detecting Biomedical Term Translations. *PloS one*, 10(6): e0126196, 2015.

Bowman S.E. Coordination of SNOMED-CT and ICD-10: getting the most out of electronic health record systems. *Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems/AHIMA, American Health Information Management Association*, 2005.

Brown E.G., Wood L., eta Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug safety*, 20(2):109–117, 1999.

Cabré M.T. Elementos para una teoría de la terminología: hacia un paradigma alternativo. *La terminología. Representación y comunicación*, 69–92, 1999.

Cabré M.T. Terminología y normalización lingüística. *Espezialitate hizkerak eta terminologia jardunaldiak*, 11–25. UPV/EHU Argitalpen Zerbitzua, 2003.

Campbell W., Campbell J., West W., McClay J., eta Hinrichs S. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. *Journal of the American Medical Informatics Association*, 21(5):885–892, 2014.

Chute C.G. Clinical Classification and Terminology: Some History and Current Observations. *Journal of the American Medical Informatics Association*, 7(3):298–303, 2000.

Claveau V. eta Zweigenbaum P. Translating biomedical terms by inferring transducers. *Conference on Artificial Intelligence in Medicine in Europe*, 236–240. Springer, 2005.

Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1), 1960.

Davis A.P., Wiegers T.C., Rosenstein M.C., eta Mattingly C.J. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, 2012:bar065, 2012.

Deléger L., Merabti T., Lecrocq T., Joubert M., Zweigenbaum P., eta Darmoni S. A twofold strategy for translating a medical terminology into French. *AMIA Annual Symposium Proceedings*, 2010 lib., page 152. American Medical Informatics Association, 2010.

Deléger L., Merkel M., eta Zweigenbaum P. Using word alignment to extend multilingual medical terminologies. *the Proceedings of Language Resources and Evaluation 2006, Workshop on Acquiring and representing multilingual, specialized lexicons: the case of biomedicine*, 9–14, 2006.

Deléger L., Merkel M., eta Zweigenbaum P. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701, 2009.

Delpech E., Daille B., Morin E., eta Lemaire C. Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. *arXiv preprint arXiv:1210.5751*, 2012.

Desjardins L. Le santé des francophones du Nouveau-Brunswick. Petit-Rocher, Société des Acadiens et des Acadiennes du Nouveau-Brunswick, 2003.

Dirckx J.H. *Dx+ Rx: A physician's guide to medical writing*. Macmillan Reference USA, 1977.

EHUko Euskara Zerbitzua eta Donostiako Erizaintza Eskola. *Erizaintzako Hiztegia*. UPV/EHU Argitalpen Zerbitzua, 2005.

Elhanan G., Perl Y., eta Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. *Journal of the American Medical Informatics Association*, 18(1), 2011.

Elhuyar. *Elhuyar Hiztegia Euskara/Gaztelania Castellano/Vasco*. Elhuyar, 2007a.

Elhuyar. *Elhuyar Hiztegia Euskara/Ingelesa English/Basque*. Elhuyar, 2007b.

Elhuyar. *Elhuyar Zientzia eta Teknologiaren Hiztegi Entziklopedikoa*. Elhuyar Edizioak & Euskal Herriko Unibertsitatea, 2009.

Elia A., Maisto A., eta Pelosi S. Morphological Analysis and Generation of Monolingual and Bilingual Medical Lexicons. *International Workshop on Systems and Frameworks for Computational Morphology*, 148–165. Springer, 2015.

Elkin P.L., Brown S.H., Husser C.S., Bauer B.A., Wahner-Roedler D., Rosenbloom S.T., eta Speroff T. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clinic Proceedings*, 81 lib., 741–748. Elsevier, 2006.

España Bonet C., Màrquez Villodre L., Labaka G., Díaz de Ilarraza Sánchez A., eta Sarasola Gabiola K. Hybrid machine translation guided by a rule-based system. *Machine translation summit XIII: proceedings of the 13th machine translation summit, September 19-23, 2011, Xiamen, China*, 554–561, 2011.

European Observatory on Health Care Systems. Luxembourg: Health system review. *Health Systems in Transition*, 1999.

Euskaltzaindia. Luis Mitxelena. 0. araua. ortografia, 1968.

Ezeiza N., Alegria I., Arriola J.M., Urizar R., eta Aduriz I. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 380–384. Association for Computational Linguistics, 1998.

Finkel J.R., Grenager T., eta Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370, 2005. URL `http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf`.

Franklin C. eta Agresti A. *Statistics: The art and science of learning from data*, chapter Statistical Inference: Confidence Intervals, 385–389. Upper Saddle River, New Jersey: Pearson Prentice Hall, 2007.

Fung P. eta Yee L.Y. An IR approach for translating new words from nonparallel, comparable texts. *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 414–420. Association for Computational Linguistics, 1998.

Gerkens S. eta Merkur S. Belgium: Health system review. *Health Systems in Transition*, 12(5):1–266, 2010.

Giannoni D.S. Medical writing at the periphery: The case of Italian journal editorials. *Journal of English for Academic Purposes*, 7(2):97–107, 2008.

Gieselmann P. Architecture of the Lucy translation system. *Second machine translation marathon, Wandlitz, Berlin*, 28, 2008.

Gooch P. eta Roudsari A.V. Automated recognition and post-coordination of complex clinical terms. *ITCH*, 8–12, 2011.

Grigonytė G., Kvist M., Wirén M., Velupillai S., eta Henriksson A. Swedification patterns of Latin and Greek affixes in clinical text. *Nordic Journal of Linguistics*, 39(01):5–37, 2016.

Gross M. The use of finite automata in the lexical representation of natural language. *LITP Spring School on Theoretical Computer Science*, 34–50. Springer, 1987.

Gunnarsson B.L. *Professional discourse*. Bloomsbury Publishing, 2009.

Gwet K.L. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

Hahn U., Honeck M., Piotrowski M., eta Schulz S. Subword segmentation–leveling out morphological variations for medical document retrieval. *Proceedings of the AMIA Symposium*, page 229. American Medical Informatics Association, 2001.

Hajicová E. Dependency-based underlying-structure tagging of a very large Czech corpus. *TAL. Traitement automatique des langues*, 41(1):57–78, 2000.

Høy A. Coming to terms with SNOMED CT® terms: linguistic and terminological issues related to the translation into Danish. *Budin G, Laurén C, Picht H et al., Terminology Science and Research*, 8, 2006.

Høy A. Guidelines for Translation of SNOMED CT. Barne-txostena version 2.0, International Health Terminology Standards Development Organization IHTSDO, 2010.

Hulden M. Foma: a Finite-State Compiler and Library. *Proceedings of EACL 2009*, 29–32, Stroudsburg, PA, USA, 2009. URL `http://dl.acm.org/citation.cfm?id=1609049.1609057`.

Humphreys B.L., McCray A.T., eta Cheh M.L. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *Journal of the American Medical Informatics Association*, 4(6):484–500, 1997.

Hutchins W.J. eta Somers H.L. *An introduction to machine translation*, 362 lib. Academic Press London, 1992.

IHTDSO SNOMED CT. SNOMED CT The Release Format 2 Value Proposition. Barne-txostena, IHTDSO, April 2013.

IHTDSO SNOMED CT. Data Analytics with SNOMED CT âĂŞ Case Studies. Barne-txostena, IHTDSO, May 2015.

IHTSDO I.H.T.S.D.O. Mapping SNOMED CT to ICD-10 Technical Specifications. Barne-txostena, International Health Terminology Standards Development Organisation, 2012.

IHTSDO I.H.T.S.D.O. SNOMED CT Starter Guide. February 2014. Barne-txostena, International Health Terminology Standards Development Organisation, 2014.

ISO. Language Resource management – Lexical Markup Framework (LMF). Barne-txostena, International Organization for Standarization, Geneva, Switzerland, June 2008.

Jiang G. eta Chute C. Auditing the Semantic Completeness of SNOMED CT Using Formal Concept Analysis. *Journal of the American Medical Informatics Association*, 16(1), 2009.

Joanes Etxeberri Saria V. Edizioa, editor. *Donostia Unibertsitate Ospitaleko alta-txostenak*. Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea, 2014.

Joubert M., Abdoune H., eta Fieschi M. Assisting the Translation of SNOMED CT into French using UMLS and four Representative French-language Terminologies. *Substance*, 22:7–3, 2009.

Jurafsky D. eta Martin J.H. *Speech and language processing*. Prentice Hall, 2008.

Karttunen L., Gaál T., eta Kempe A. Xerox finite-state tool. *Rapport technique, Centre de recherche Xerox de Grenoble*, 1997.

Klein G.O. eta Chen R. Translation and Localization of SNOMED CT–Strategies and description of a pilot project. *Nursing Informatics*, 2009.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., *et al.*. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics, 2007.

Koehn P. eta Knight K. Learning a Translation Lexicon from Monolingual Corpora. *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9*, ULA '02, 9–16, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1118627.1118629`.

Koehn P., Och F.J., eta Marcu D. Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48–54. Association for Computational Linguistics, 2003.

Labaka G. *EUSMT: incorporating linguistic information into SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.* Doktoretza-tesia, Euskal Herriko Unibertsitateko Donostiako Informatika Fakultatea, 2010.

Labaka G., España-Bonet C., Màrquez L., eta Sarasola K. A hybrid machine translation architecture guided by syntax. *Machine translation*, 28(2):91–125, 2014.

Landis J.R. eta Koch G.G. The measurement of observer agreement for categorical data. *biometrics*, 159–174, 1977.

Langlais P., Yvon F., eta Zweigenbaum P. Analogical translation of medical words in different languages. *Advances in Natural Language Processing*, 284–295. Springer, 2008.

Lee D., Cornet R., eta Lau F. Implications of SNOMED CT versioning. *International Journal of Medical Informatics*, 80:442–453, 2011.

León-Araúz P. Chapter two Term Variation in the Psychiatric Domain: Transparency and Multidimensionality. *Word Formation and Transparency in Medical English*, 33–54, 2015.

Lepage Y. Solving analogies on words: an algorithm. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, 728–734. Association for Computational Linguistics, 1998.

Lipscomb C.E. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

LISA. Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX). Barne-txostena, Localization Industry Standards Association, 2008.

Lovis C., Michel P., Baud R., eta Scherrer J. Word Segmentation Processing: A Way To Exponentially Extend Medical Dictionaries. *MEDINFO*, 8:28–32, 1995.

Maheronnaghsh R., Nezareh S., Sayyah M.K., eta Rahimi-Movaghar V. Developing SNOMED-CT for Decision Making and Data Gathering: A Software Prototype for Low Back Pain. *Acta Medica Iranica*, 51(8):548–53, September 9 2011.

Mangeot M. An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language. *LREC Workshop on International Standards of Terminology and Language Resources Management*, 37–44, Las Palmas, Spain, May 2002.

Manning C.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J., eta McClosky D. The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Mayerthaler W. *Morphologische natuerlichkeit*, 28 lib. Akademische Verlagsgesellschaft Athenaion, 1981.

Mayor A. *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Doktoretza-tesia, Euskal Herriko Unibertsitateko Donostiako Informatika Fakultatea, 2007.

Mayor A., Alegria I., Diaz de Ilarraza A., Labaka G., Lersundi M., eta Sarasola K. 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.

McCarthy E.M. http://macroevolution.net, 2016.

McCray A.T., Browne A.C., eta Moore D.L. The semantic structure of neoclassical compounds. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 165. American Medical Informatics Association, 1988.

Melby A.K. Terminology in the Age of Multilingual Corpora. *The Journal of Specialised Translation*, 18:7–29, July 2012.

Merabti T., Soualmia L., Grosjean J., Letord C., eta Darmoni S. Assisting the Translation of SNOMED CT into French. *Studies in health technology and informatics*, 192:47–51, 2013.

Mikroyannidi E., Stevens R., Iannone L., eta Rector A. Analysing Syntactic Regularities and Irregularities in SNOMED-CT. *Journal of Biomedical Semantics*, 3(8), 2012.

Miller G.A. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Miller N., Lacroix E.M., eta Backus J.E. MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. *Bulletin of the Medical Library Association*, 88(1):11, 2000.

Mohri M., Pereira F., Riley M., eta Allauzen C. AT & T FSM Library Finite-State Machine Library. Barne-txostena, AT&T Labs-Research, NJ, USA, 2006.

Mohri M. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311, 1997.

Müller P.O. *Word-formation: An International Handbook of the Languages of Europe*, 40 lib., chapter Word-formation and technical languages, 2251–2266. Walter de Gruyter GmbH & Co KG, 2015.

Nadeau D. eta Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Namer F. eta Zweigenbaum P. Acquiring meaning for French medical terminology: contribution of morphosemantics. *Actes 10 th World Congress on Medical Informatics*, 535–539, 2004.

Naradowsky J. eta Toutanova K. Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 895–904. Association for Computational Linguistics, 2011.

Navigli R., Velardi P., eta Gangemi A. Ontology learning and its application to automated terminology translation. *IEEE Intelligent systems*, 18(1): 22–31, 2003.

Oflazer K. Two-level description of Turkish morphology. *Literary and linguistic computing*, 9(2):137–148, 1994.

Osakidetza, Euskadiko Osasun Saila eta UZEI. *Administrazio Sanitarioko Hiztegia*. UZEI, 1999.

Osborne J.D., Flatow J., Holko M., Lin S.M., Kibbe W.A., Zhu L.J., Danila M.I., Feng G., eta Chisholm R.L. Annotating the human genome with Disease Ontology. *BMC genomics*, 10(1):S6, 2009.

Padró L. eta Stanilovsky E. FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

Panocová R. Chapter four Transparency and Use of Neoclassical Word Formation in Medical English. *Word Formation and Transparency in Medical English*, 73–97, 2015.

Papineni K., Roukos S., Ward T., eta Zhu W.J. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics, 2002.

Perez-de-Viñaspre O. SNOMED CT sare semantikoa euskaratzeko aplikazioa. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2013.

Perez-de-Viñaspre O., Oronoz M., eta Patrick J. Osasun-txosten elebidunak posible ote? *I. IkerGazte, Nazioarteko Ikerketa Euskaraz*, 730–738, 2015. ISBN 978-84-8438-539-4.

Perez-de Viñaspre O. eta Oronoz M. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC Medical Informatics and decision making*, 15(2):S5, 2015.

Petersen P.G. How to Manage the Translation of a Terminology. Presentation at the IHTSDO October 2011 Conference and Showcase, October 2011.

Popel M. eta Žabokrtskỳ Z. TectoMT: modular NLP framework. *International Conference on Natural Language Processing*, 293–304. Springer, 2010.

Reynoso G.A., March A.D., Berra C.M., Strobietto R.P., Barani M., Iubatti M., Chiaradio M.P., Serebrisky D., Kahn A., Vaccarezza O.A., *et al.*. Development of the Spanish version of the Systematized Nomenclature of Medicine: methodology and main issues. *Proceedings of the AMIA Symposium*, page 694. American Medical Informatics Association, 2000.

Ripley B.D. *Stochastic simulation*, 316 lib. John Wiley & Sons, 2009.

Sager J.C. Term Formation. *Handbook of Terminology Management*, 1:25–41, 1997.

SALT project. SALT project – XML representations of Lexicons and Terminologies (XLT) – Default XLT Format (DXLT). Barne-txostena, SALT project, 2000. Reference name of working document: DXLT specification draft 1b.

San Martin I. Terminologia Sareak Ehunduz: Unibertsitateko ikasgeletan erabiltzen den terminologia erreala ikusgai egitea helburu duen programa. *A: ALBERDI, Xabier*, 2013.

San Vicente I. eta Manterola I. PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Sarasola K. Strategic priorities for the development of language technology in minority languages. *Workshop on Developing language resources for minority languages: re-useability and strategic priorities. Second International Conference on Language Resources and Evaluation*, 2000.

Savova G.K., Masanz J.J., Ogren P.V., Zheng J., Sohn S., Kipper-Schuler K.C., eta Chute C.G. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5): 507–513, 2010.

Schmid H., Fitschen A., eta Heid U. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. *LREC*, 2004.

Schriml L.M., Arze C., Nadendla S., Chang Y.W.W., Mazaitis M., Felix V., Feng G., eta Kibbe W.A. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

Schulz S., Bernhardt-Melischnig J., Kreuzthaler M., Daumke P., eta Boeker M. Machine vs. Human Translation of SNOMED CT Terms. In et al. C.L., editor, *MEDINFO 2013*, 581–584, 2013.

Schulz S., Markó K., Sbrissia E., Nohama P., eta Hahn U. Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. *Proceedings of the 20th international conference on Computational Linguistics*, page 813. Association for Computational Linguistics, 2004.

Sennrich R., Haddow B., eta Birch A. Edinburgh neural machine translation systems for WMT 16. *arXiv preprint arXiv:1606.02891*, 2016.

Silva T.S.D., MacDonald D., Paterson G., Sikdar K.C., eta Cochrane B. Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) to represent computed tomography procedures. *Computer Methods and Programs in Biomedicine*, 101(3):324 – 329, 2011. ISSN 0169-2607. URL `http://www.sciencedirect.com/science/article/pii/S0169260711000125`.

Stearns M., Price C., Spackman K., eta Wang A. SNOMED clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium*, 662–666, 2001.

Stedman T.L. *Stedman's Medical Dictionary*, chapter Medical Prefixes, Suffixes, and Combining Forms. Lippincott Williams & Wilkins, twenty-eighth edition edition, 2005.

Stroppa N. eta Yvon F. An analogical learner for morphological analysis. *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 120–127. Association for Computational Linguistics, 2005.

ten Hacken P. eta Panocová R. Introduction: Medical Language, Word Formation and Transparency. *Word Formation and Transparency in Medical English*, 13–32, 2015.

Tjong Kim Sang E.F. eta De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 142–147. Association for Computational Linguistics, 2003.

UPV/EHU Argitalpen Zerbitzua, editor. *Giza anatomiako atlasa*. UPV/EHU Argitalpen Zerbitzua, 2014.

UZEI. Euskalterm Terminologia Banku Publikoa. `http://www.euskadi.net/euskalterm`, 2004.

Weijnitz P., Forsbom E., Gustavii E., Pettersson E., eta Tiedemann J. MT Goes Farming: Comparing Two Machine Translation Approaches on a New Domain. *LREC*, 2004.

Westman J.A. *Medical genetics for the modern clinician*. Lippincott Williams & Wilkins, 2006.

Wikipedia. List of medical roots, suffixes and prefixes – Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=List_of_medical_roots,_suffixes_and_prefixeso, 2013.

Wiktionary. Category:English suffixes – Wiktionary, a wiki-based Open Content dictionary. http://en.wiktionary.org/wiki/Category:English_suffixes, 2014.

Wolff S. *et al.*. The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, 23(4):195–203, 1984.

World Health Organization, Osasun Saila, eta UZEI. *GNS-10 (Gaixotasunen eta horiekin lotutako osasun-arazoen nazioarteko sailkapen estatistikoa - 10. berrikuspena)*. Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia, 1996.

Yu A.C. Methods in biomedical ontology. *Journal of Biomedical Informatics*, 39(3):252 – 266, 2006. ISSN 1532-0464. URL http://www.sciencedirect.com/science/article/pii/S1532046405001310.

Zabala I., Aierbe A., Aldezabal I., Aranzabe M., Arregi X., Arriola J.M., Elordui A., Elosegi A., Elosegi K., Ezeiza J., *et al.*. GARATERM: diskurtso akademiko-profesionalaren didaktika eta garapena uztartzeko tresna informatikoen diseinua eta integrazioa helburu duen proiektua. *Ugarteburu I. Eta P. Salaburu (arg.) Espezialitate hizkerak eta terminologia Jardunaldiak III. Bilbo: Euskara Institutua–EHU*, 2008.

Zabala I., Martin I.S., eta Lersundi M. Learning terminology in order to become an active agent in the development of Basque biomedical registers. *Language Learning in Higher Education*, 6(1):145–165, 2016.

Zabala I., San Martin I., Lersundi M., Azkue J.J., eta Mendizabal J.L. The Elaboration of Human Anatomy Terminology for the Basque Language: the Contribution of Translators, Linguists and Experts. *Terminàlia*, 15–25, 2012.

Zhu Y., Pan H., Zhou L., Zhao W., Chen A., Andersen U., Pan S., Tian L., eta Lei J. Translation and Localization of SNOMED CT in China: A pilot study. *Artificial Intelligence in Medicine*, 54(2):147–149, 2012.

A

TBX formatuaren adibideak

Eranskin honetan, lexikoaren biltegiratzeko erabilitako TermBase eXchange (TBX) formatuaren egokitzapenaren adibideen XMLak erakusten ditugu.

A.1 SNOMED CTrentzako TBX formatuaren adibidea

Atal honetan SNOMED CTren eduki terminologikoa egituratuta gordetzeko definitu dugun TBX formatuaren adibideak erakusten ditugu. Alde batetik A.1 irudian, kontzeptu mailan gordetzen den informazioaren adibidea erakusten dugu. Bestetik, A.2 irudian, jatorri terminoei buruz gordetzen denaren adibidea ikus dezakegu, eta azkenik A.3 irudian, euskarazko ordainen inguruan gordetako informazioaren adibidea. Hiru irudien artean kontzeptu baten informazio guztia eskuratuko genuke.

```
1  <termEntry id="c292044008">
2    <descrip type="subjectField">010-011</descrip>
3    <descrip type="definition">Aspirin adverse reaction (disorder)</
         descrip>
4    <langSet xml:lang="en">...</langSet>
5    <langSet xml:lang="es">...</langSet>
6    <langSet xml:lang="eu">...</langSet>
7    <transacGrp>
8      <date>2013-01-04T18:46:12.954+01:00</date>
9    </transacGrp>
10 </termEntry>
```

A.1 irudia – Kontzeptu baten adibidea TBXn.

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

A.2 irudia – Jatorri-terminoen adibidea TBXn.

```
<langSet xml:lang="eu">
  <ntig id="eu100014">
    <termGrp>
      <term>aspirinak eragindako aurkako erreakzio</term>
      <termNote type="administrativeStatus">preferredTerm-admn-sts</
          termNote>
      <termNote type="partOfSpeech">izena</termNote>
      <termNote type="usageNote">InitialInsensitive</termNote>
    </termGrp>
    <admin type="sortKey">aspirinak eragindako aurkako erreakzioa</admin
        >
    <admin type="entrySource">102</admin>
    <admin type="conceptOrigin">en432156011</admin>
    <admin type="originatingDatabase">pat_es_16</admin>
    <descrip type="reliabilityCode">6.1</descrip>
  </tig>
</langSet>
```

A.3 irudia – Euskal ordainen adibidea TBXn.

A.2 ItzulDBrentzako TBX formatuaren adibidea

Atal honetan, SNOMED CTren euskaratze prozesurako lortutako terminoordain pareak gordetzeko sortutako TBX fitxategiaren sarrera baten adibidea erakusten dugu (A.4 irudia).

```
<termEntry id="p26435">
  <langSet xml:lang="en">
    <tig id="t68418">
      <term>abdomen</term>
      <admin type="sortKey">abdomen</admin>
    </tig>
  </langSet>
  <langSet xml:lang="eu">
    <tig id="t68419">
      <term>sabel</term>
      <admin type="sortKey">sabel</admin>
      <admin type="entrySource">EuskalTerm</admin>
      <termNote type="partOfSpeech">Izen</termNote>
      <descrip type="reliabilityCode">9.1</descrip>
      <termNote type="usageNote">InitialInsensitive</termNote>
    </tig>
    <tig id="t68420">
      <term>sabelalde</term>
      <admin type="sortKey">sabelalde</admin>
      <admin type="entrySource">Erizaintza</admin>
      <termNote type="partOfSpeech">Izen</termNote>
      <descrip type="reliabilityCode">7.1</descrip>
      <termNote type="usageNote">InitialInsensitive</termNote>
    </tig>
    <tig id="t68421">
      <term>abdomen</term>
      <admin type="sortKey">abdomen</admin>
      <admin type="entrySource">Erizaintza</admin>
      <admin type="entrySource">GNS10</admin>
      <admin type="entrySource">ZT</admin>
      <admin type="entrySource">Anatomia</admin>
      <admin type="entrySource">EuskalTerm</admin>
      <termNote type="partOfSpeech">Izen</termNote>
      <descrip type="reliabilityCode">9.5</descrip>
      <termNote type="usageNote">InitialInsensitive</termNote>
    </tig>
  </langSet>
  <transacGrp>
    <date>2016-06-13</date>
  </transacGrp>
</termEntry>
```

A.4 irudia – KabiTermen identifikazioaren eta etiketatzearen patroi bat.

B

Termino neoklasikoen sorkuntzarako erregelak

B.1 Euskarazko morfotaktika erregelak

```
define HASR r -> e r r || [.#.|%^] _ ;
define HASS s -> e s || [.#.|%^] _ (%+) Kon ;
define CLEANUP [%+|%^] -> 0;
define KERR f -> 0 || _ %+ f ,,
         s -> 0 || _ %+ s ,,
         p -> 0 || _ %+ p ,,
         l -> 0 || _ %+ l ,,
         n -> 0 || _ %+ n ,,
         m -> 0 || _ %+ m ,,
         t -> 0 || _ %+ t ,,
         g -> 0 || _ %+ g ,,
         k -> 0 || _ %+ k,,
         K -> 0 || _ %+ k,,
         k -> 0 || _ %+ K,,
         K -> 0 || _ %+ K,,
         d -> 0 || _ %+ d;
define TXISAB %+ -> %+ t || [ l | n | m | r ] _ Txis ("+") Bok,,
              %+ -> %+ t || [ l | n | m ] _ x ("+") Bok,,
           n -> n t || _ Txis "+" Bok,,
           l -> l t || _ Txis "+" Bok,,
           r -> r t || _ Txis "+" Bok,,
               0 -> t || [ l | n | m ] _ x "+" Bok;
define RTXIS r %+ -> r e || _ Txis Kon;
define RKON r r -> r || Kon %+ _;
define RERREP r -> 0 || _ r r ;
define MN m -> n || _ %+ [ b | p | t | f ] ;
define XS x %+ -> 0 || _ s,,
```

```
28          x %+ -> s || _ ezSH Kon;
29
30  define KZ k -> z || [ezRNL|%+|%^] _ %+ [e|i|y],,
31          k -> t z || [r|n|l] _ %+ [e|i|y];
32
33  define HKON t %+ h -> t ,,
34          p %+ h -> f,,
35       k %+ h -> K ,,
36       w %+ h -> w ,,
37       K %+ h -> K;
38
39  define CC K Z -> k z || _ %+ [e|i],,
40            K Z -> k || _ %+ [Kon|a|o|u];
41  define SPN %+ p n -> n || s _;
42
43  define KONPT p -> t || Kon _ z ,,
44               p -> 0 || Kon _ KonEzSNLR ,,
45            p -> 0 || s _ [n|s] ,,
46               k -> 0 || Kon _ t ;
47  define OUS %+ o s o -> o || z e _,,
48             %+ o s o -> o || n e _;
49  define TM t -> t a || _ %+ m e n [d|t] u .#. ;
50  define AEMIA a e m i a -> e m i a || _ .#. ;
51  define NTIA n t i a -> n t z i a || _ .#. ;
52  define ZIO z i o -> t i o || s %+ _ .#. ;
53  define SIO i o [i|n] -> i o || [s|x] _ ;
54  define TION t i %+ o [i|n] -> z i o || _ .#.;
55  define OSPITAL h -> 0 || _ o s p i t a l ;
56
57
58
59  define KCLEAN K -> k ,,
60                Z -> z ;
61
62  define DESR s -> 0 || d e _ %+ [r|n] ;
```

B.1 irudia – Euskarazko morfotaktika erregelak.

B.2 Termino neoklasikoen ingelesa-euskara transliterazio erregelak

```
define PH p h -> f;
define TH t h -> t;
define CH c h -> K;
define RH r h -> r r;
define SH s h -> s;
define WH w h -> w;
define LH l h -> l;
define H PH .o. TH .o. CH .o. RH .o. WH .o. SH .o. LH;
define YHAS y -> j || %^ _ Bok,,
            y -> i || %^ _ Kon;
define Y y -> i ;
define Q q -> K ;
define C c -> k || [ezC] _ [a|o|u|ezHC|%#],,
          c -> z || [ezC] _ [e|i|y];
define V v -> b;
define CK c k -> K ;
define NM m -> n || _ [b|p];
define NT n -> n t || _ Txis Bok,,
           l -> l t || _ Txis Bok,,
       r -> r t || _ Txis Bok,,
       m -> n t || _ Txis Bok;

define KA k -> K || [ezC] _ ;
define X x -> 0 || _ s,,
          x -> s || _ Kon,,
     x -> t s || [ n | r ] _ Bok;
define CHA c h a -> t x a || %^ _ ;

define Ald YHAS .o. CHA .o. NM .o. H .o. X .o. Q .o. CK .o. KA .o. C .o.
     NT .o. Y .o. V;

define CC  c c -> k z || _ [e|i|y],,
       c c -> k || _ [a|o|u|Kon],,
       c c -> K Z || _ %# ;

define BIK m m -> m || [ezI] _ ,,
           n n -> n ,,
           l l -> l ,,
       f f -> f,,
       s s -> s,,
       t t -> t,,
       g g -> g,,
       k k -> K,,
       K K -> K,,
       d d -> d,,
       p p -> p;

define Trans Ald .o. CC .o. BIK ;
```

B.2 irudia – Termino neoklasikoen ingelesa-euskara transliterazio erregelak.

C

Termino habiratuen sorkuntzarako erregelak

```
read lexc scriptak/foma/lex/disorder.lex
define HDISDENA;
define HDIS HDISDENA.u;

read lexc scriptak/foma/lex/finding.lex
define HFINDENA;
define HFIN HFINDENA.u;

read lexc scriptak/foma/eponimoak.lex
define HEPODENA;
define HEPO HEPODENA.u;

read lexc scriptak/foma/lex/bodystructure.lex
define HBODDENA;
define HBOD HBODDENA.u;

read lexc scriptak/foma/lex/procedure.lex
define HPROCDENA;
define HPROC HPROCDENA.u;

read lexc scriptak/foma/procedureSimple.lex
define HPROCSIMPLDENA;
define HPROCSIMPL HPROCSIMPLDENA.u ;

read lexc scriptak/foma/lex/qualifier.lex
define HQUALDENA;
define HQUA HQUALDENA.u;

read lexc scriptak/foma/lex/pharmproduct.lex
read lexc scriptak/foma/lex/substance.lex
union net
```

```
define HPHARDENA;
define HPHAR HPHARDENA.u;

read lexc scriptak/foma/lex/observable.lex
define HOBVDENA;
define HOBV HOBVDENA.u;

read lexc scriptak/foma/besteak.lex
define HBESTDENA;
define HBES HBESTDENA.u;


define Muga [ " " | .#. ];

define CLEAN2 " " " " -> " " ;

define HEPO2 HEPO ("-" HEPO) ("-" HEPO);

##################ETIKETAK GEHITZEKO
##################Deklinabideak###############
define MarGEN  ?+ @-> ... {+-+ReM} || Muga _ Muga ;
define SinABS ?+ @-> ... {+a} || Muga _ Muga ;
define SinGEN ?+ @-> ... {+areM} || Muga _ Muga ;
define MugaGEN ?+ @-> ... {+ReM} || Muga _ Muga ;
define SinSOZ ?+ @-> ... {+arekiM} || Muga _ Muga ;
define GEL ?+ @-> ... {+Eko} || Muga _ Muga ;
define SinGEL ?+ @-> ... {+ako} || Muga _ Muga ;
define MugaDAT ?+ @-> ... {+ri} || Muga _ Muga ;
define SinDAT ?+ @-> ... {+ari} || Muga _ Muga ;
define INE ?+ @-> ... {+Ean} || Muga _ Muga ;
define SinKAU ?+ @-> ... {+agatikako} || Muga _ Muga ;
define ERGEra ?+ @ -> ... {+ak_eragindako} || Muga _ Muga ;
define HEL ?+ @ -> ... {+Erako} || Muga _ Muga ;

#################Hierarkiak###################
define Epo ?+ @-> ... {|EPO} || Muga _ Muga ;
define Dis    ?+ @-> ... {|DIS} || Muga _ Muga ;
define Fin    ?+ @-> ... {|FIN} || Muga _ Muga ;
define Bes    ?+ @-> ... {|BES} || Muga _ Muga ;
define Bod    ?+ @-> ... {|BOD} || Muga _ Muga ;
define Proc    ?+ @-> ... {|PROC} || Muga _ Muga ;
define Phar  ?+ @-> ... {|PHAR} || Muga _ Muga ;
define Obv  ?+ @-> ... {|OBV} || Muga _ Muga ;
define Qua ?+ @-> ... {|QUA} || Muga _ Muga;
################

define OrdAldatuLehenaAzkenera ?+ @-> ... " " {&LehenaAzkenera} ;
define OrdAldatuLehenaEtaAzkena ?+ @-> ... " " {&LehenaEtaAzkena} ;
define OrdAldatuAzkenaLehenera ?+ @-> ... " " {&AzkenaLehenera} ;

define OrdZurruna ?+ @-> ... " " {&OrdZurruna};

define Txur ?+ " " ?+ ;
define Elkar " " -> "_" || ?+ _ ?+;
define Elkartu Txur .o. Elkar ;

###############POS kendu
```

```
define KenAp {'s} -> 0 ;
define KenOf " " {of} " " -> " " ,,
             " " {with} " " -> " ",,
         " " {on} " " -> " " ,,
         " " {to} " " -> " " ;

#1#########[EPONYM]+[DISORDER]###+
define EpDis (Epo .o. MarGEN) " " Dis;
define EzEpDis HEPO2 " " HDIS;
define EtEpDis ?+ @-> ... { |pat_es_01};
define TrEpDis EzEpDis .o. EpDis  .o. EtEpDis;

#2########[EPONYM]+'s[POS]+[DISORDER]###
define EpPOSDis (Epo .o. MarGEN) " " {'s} " " Dis;
define EzEpPOSDis HEPO2 " " {'s} " " HDIS;
define EtEpPOSDis ?+ @-> ... { |pat_es_02};
define TrEpPOSDis EzEpPOSDis .o. EpPOSDis .o. KenAp .o. EtEpPOSDis;

#3#######[BODYSTRUCTURE]+structure[NN]###
define BodStr (Bod .o. SinGEN) " " Bes;
define EzBodStr HBOD " " {structure};
define EtBodStr ?+ @-> ... { |pat_es_03};
define TrBodStr EzBodStr .o. BodStr  .o. EtBodStr ;

#4#####structure[NN]+of[IN]+[BODYSTRUCTURE]###
define StrOfBod  Bes " " {of} " " (Bod .o. SinGEN);
define EzStrOfBod  {structure} " " {of} " " HBOD;
define EtStrOfBod ?+ @-> ... { |pat_es_04};
define TrStrOfBod EzStrOfBod  .o. StrOfBod .o. KenOf .o.
    OrdAldatuLehenaAzkenera .o. EtStrOfBod ;

#5#####deficiency[NN]+of[IN]+[PHARMPRODUCT|SUBSTANCE]###
define DefOfPhar  Bes " " {of} " " (Phar .o. MugaGEN);
define EzDefOfPhar  {deficiency} " " {of} " " HPHAR;
define EtDefOfPhar ?+ @-> ... { |pat_es_05};
define TrDefOfPhar EzDefOfPhar  .o. DefOfPhar .o. KenOf .o.
    OrdAldatuLehenaAzkenera .o. EtDefOfPhar ;

#6######neoplasm[NN]+of[IN]+[CLINICAL_FIN]+of[IN]+[BODYSTRUCTURE]###
define NeoOfFinOfBod Bes " " {of} " " (Fin .o. GEL) " " {of} " " (Bod .o
    . [GEL|SinGEN]);
define EzNeoOfFinOfBod {neoplasm} " " {of} " " HFIN " " {of} " " HBOD;
define EtNeoOfFinOfBod ?+ @-> ... { |pat_es_06};
define TrNeoOfFinOfBod EzNeoOfFinOfBod  .o. NeoOfFinOfBod .o. KenOf .o.
    OrdAldatuLehenaAzkenera .o. EtNeoOfFinOfBod ;

#7#####[PHARMPRODUCT|SUBSTANCE]+allergy[NN]###
define PharAll (Phar .o. SinDAT) " " Bes;
define EzPharAll HPHAR " " {allergy} ;
define EtPharAll ?+ @-> ... { |pat_es_07};
define TrPharAll EzPharAll .o. PharAll .o. EtPharAll ;

#8#####[PHARMPRODUCT|SUBSTANCE]+measurement[NN]###
define PharMea (Phar .o. SinGEN) " " Bes;
define EzPharMea HPHAR " " {measurement};
define EtPharMea ?+ @-> ... { |pat_es_08};
define TrPharMea EzPharMea .o. PharMea  .o. EtPharMea ;
```

```
142
143 #9#####[PHARMPRODUCT|SUBSTANCE]+antibody[NN]+measurement[NN]###
144 define PharAntMea Phar " " (Bes .o. SinGEN) " " Bes;
145 define EzPharAntMea HPHAR " " {antibody} " " {measurement};
146 define EtPharAntMea ?+ @-> ... { |pat_es_09};
147 define TrPharAntMea EzPharAntMea .o. PharAntMea .o. EtPharAntMea;
148
149 #10#####[PHARMPRODUCT|SUBSTANCE]+overdose[NN]###
150 define PharOve (Phar .o. [ERGEra]) " " Bes;
151 define EzPharOve HPHAR " " {overdose};
152 define EtPharOve ?+ @-> ... { |pat_es_10};
153 define TrPharOve EzPharOve .o. PharOve  .o. EtPharOve;
154
155 #11#####[PHARMPRODUCT|SUBSTANCE]+poisoning[NN]###
156 define PharPoi (Phar .o. [ERGEra]) " " Bes;
157 define EzPharPoi HPHAR " " {poisoning};
158 define EtPharPoi ?+ @-> ... { |pat_es_11};
159 define TrPharPoi EzPharPoi .o. PharPoi .o. EtPharPoi;
160
161 #12#####benign[JJ]+[CLINICAL_DIS]#######open[JJ]+[CLINICAL_DIS]...
162 define BenDis Bes " " Dis;
163 define EzBenDis [{open}|{closed}|{benign}|{malignant}|{accidental}|{
        intentional}|{superficial}|{chronic}] " " HDIS ;
164 define EtBenDis ?+ @-> ... { |pat_es_12};
165 define TrBenDis EzBenDis .o. BenDis  .o. EtBenDis;
166
167 #13######lesion[NN]+of[IN]+[BODYSTRUCTURE]######fracture[NN]+of[IN]+[
        BODYSTRUCTURE]
168 define LesOfBod Bes " " {of} " " (Bod .o. [GEL|SinGEN]);
169 define EzLesOfBod [{lesion}|{fracture}] " " {of} " " HBOD;
170 define EtLesOfBod ?+ @-> ... { |pat_es_13};
171 define TrLesOfBod EzLesOfBod .o. LesOfBod .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtLesOfBod ;
172
173 #14#####finding[NN]+of[IN]+[OBSERVABLE]###
174 define FinOfObv Bes " " {of} " " (Obv .o. SinGEN);
175 define EzFinOfObv {finding} " " {of} " " HOBV ;
176 define EtFinOfObv ?+ @-> ... { |pat_es_14};
177 define TrFinOfObv EzFinOfObv .o. FinOfObv .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtFinOfObv ;
178
179 #15#####entire[JJ]+[BODYSTRUCTURE]###
180 define EntBod Bes " " Bod;
181 define EzEntBod {entire} " " HBOD ;
182 define EtEntBod ?+ @-> ... { |pat_es_15};
183 define TrEntBod EzEntBod .o. EntBod .o. OrdAldatuLehenaAzkenera .o.
        EtEntBod ;
184
185 #16#####[PHARMPRODUCT|SUBSTANCE]+adverse[JJ]+reaction[NN]##
186 define PharAdvReac (Phar .o. [ERGEra]) " " [ Bes];#SinKau Kendu dut
187 define EzPharAdvReac HPHAR " " {adverse_reaction};
188 define EtPharAdvReac ?+ @-> ... { |pat_es_16};
189 define TrPharAdvReac EzPharAdvReac .o. PharAdvReac  .o. EtPharAdvReac;
190
191 #17#####[CLINICAL_DIS]+of[IN]+undetermined[JJ]+intent[NN]###
192 define DisOfUndInt Dis " " {of} " " [Elkartu .o. (Bes .o. GEL)];
193 define EzDisOfUndInt HDIS " " {of} " " {undetermined} " " {intent};
```

```
194 define EtDisOfUndInt ?+ @-> ... { |pat_es_17};
195 define TrDisOfUndInt EzDisOfUndInt .o. DisOfUndInt .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtDisOfUndInt;
196
197 #18#####[PHARMPRODUCT|SUBSTANCE]++[CC]+[PHARMPRODUCT|SUBSTANCE]#
198 define PharPlusPhar Phar " " Bes " " Phar;
199 define EzPharPlusPhar HPHAR " " [{+}|{%}] " " HPHAR ;
200 define EtPharPlusPhar ?+ @-> ... { |pat_es_18};
201 define TrPharPlusPhar EzPharPlusPhar  .o. PharPlusPhar .o.
        EtPharPlusPhar;
202
203 #19#####serum[NN]+[PROCEDURE] eta plasma[NN]+[PROCEDURE] eta urine[NN]+[
        PROCEDURE]##
204 # define SerProc Bes " " (Proc .o. INE);
205 # define EzSerProc [{serum}|{plasma}|{urine}] " " HPROC ;
206 # define EtSerProc ?+ @-> ... { |pat_es_19};
207 # define TrSerProc EzSerProc .o. SerProc  .o. EtSerProc;
208
209 #20#####[PROC]+on[]+[BODY] (ON, QUAL MOTA BAT DA)##
210 define ProcOnBody Proc " " {on} " " (Bod .o. INE);
211 define EzProcOnBody HPROC " " {on} " " HBOD;
212 define EtProcOnBody ?+ @-> ... { |pat_es_20};
213 define TrProcOnBody EzProcOnBody .o. ProcOnBody .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtProcOnBody;
214
215 #21####[PROC]+to[]+[BODY]##
216 define ProcToBody Proc " " {to} " " (Bod .o. HEL);
217 define EzProcToBody HPROC " " {to} " " HBOD;
218 define EtProcToBody ?+ @-> ... { |pat_es_21};
219 define TrProcToBody EzProcToBody .o. ProcToBody .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. OrdZurruna .o. EtProcToBody;
220
221 ###OROKORRAK
222
223 #1#####[PROCEDURE]+of[IN]+[BODYSTRUCTURE]######+
224 define ProcOfBod [Proc] " " {of} " " (Bod .o. [SinGEN|GEL]);
225 define EzProcOfBod [HPROC] " " {of} " " HBOD;
226 define EtProcOfBod ?+ @-> ... { |pat_or_010};
227 define TrProcOfBod EzProcOfBod .o. ProcOfBod .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtProcOfBod;
228
229 define DisOfBod [Dis] " " {of} " " (Bod .o. [SinGEN|GEL]);
230 define EzDisOfBod [HDIS] " " {of} " " HBOD;
231 define EtDisOfBod ?+ @-> ... { |pat_or_011};
232 define TrDisOfBod EzDisOfBod .o. DisOfBod .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtDisOfBod;
233
234 define BodOfBod [Bod] " " {of} " " (Bod .o. [SinGEN|GEL]);
235 define EzBodOfBod [HBOD] " " {of} " " HBOD;
236 define EtBodOfBod ?+ @-> ... { |pat_or_012};
237 define TrBodOfBod EzBodOfBod .o. BodOfBod .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtBodOfBod;
238
239 #2####[PHARMPRODUCT|SUBSTANCE]+[CLINICAL_DIS]###+
240 define PhDis (Phar .o. [ERGEra]) " " Dis;
241 define EzPhDis HPHAR " " HDIS ;
242 define EtPhDis ?+ @-> ... { |pat_or_02};
```

```
243 define TrPhDis EzPhDis .o. PhDis .o. EtPhDis ;
244
245 #3#######[BODYSTRUCTURE]+[PROCEDURE]###+#####[BODY]+[DIS]##
246 define BodProc (Bod .o. [SinGEN]) " " Proc ;
247 define EzBodProc HBOD " " HPROC;
248 define EtBodProc ?+ @-> ... { |pat_or_030};
249 define TrBodProc EzBodProc .o. BodProc .o. EtBodProc ;
250
251 define BodDis (Bod .o. [SinGEN]) " " [Dis];
252 define EzBodDis HBOD " " [HDIS];
253 define EtBodDis ?+ @-> ... { |pat_or_031};
254 define TrBodDis EzBodDis .o. BodDis .o. EtBodDis ;
255
256 #4#####[CLINICAL_DIS]+with[IN]+[CLINICAL_DIS]#
257 define DisWithDis (Dis .o. SinABS) " " {with} " " (Dis .o. SinSOZ);
258 define EzDisWithDis HDIS " " {with} " " HDIS;
259 define EtDisWithDis ?+ @-> ... { |pat_or_04};
260 define TrDisWithDis EzDisWithDis  .o. DisWithDis .o. KenOf .o.
        EtDisWithDis ;
261
262 #5#####[PROCEDURE]+of[IN]+[BODYSTRUCTURE]+of[IN]+[BODYSTRUCTURE]#
263 define ProcOfBodOfBod Proc " " {of} " " (Bod .o. SinGEN) " " {of} " " (
        Bod .o. GEL);
264 define EzProcOfBodOfBod HPROCSIMPL " " {of} " " HBOD " " {of} " " HBOD;
265 define EtProcOfBodOfBod ?+ @-> ... { |pat_or_05};
266 define TrProcOfBodOfBod EzProcOfBodOfBod .o. ProcOfBodOfBod .o. KenOf .o
        . OrdAldatuLehenaEtaAzkena .o. EtProcOfBodOfBod;
267
268 #6#####[CLINICAL_FIN]+of[IN]+[PHARMPRODUCT|SUBSTANCE]###
269 define FinOfPhar Fin " " {of} " " (Phar .o. SinGEN) ;
270 define EzFinOfPhar HFIN " " {of} " " HPHAR ;
271 define EtFinOfPhar  ?+ @-> ... { |pat_or_06};
272 define TrFinOfPhar EzFinOfPhar .o. FinOfPhar .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtFinOfPhar;
273
274 #7#####[PROCEDURE]+[PROCEDURE]###
275 define ProcProc (Proc .o. SinGEN) " " Proc ;
276 define EzProcProc HPROC " " HPROC;
277 define EtProcProc ?+ @-> ... { |pat_or_07};
278 define TrProcProc EzProcProc .o. ProcProc .o. EtProcProc;
279
280 #8#####[PROCEDURE]+of[IN]+[CLINICAL_DIS]###
281 define ProcOfDis Proc " " {of} " " (Dis .o. SinGEN);
282 define EzProcOfDis HPROC " " {of} " " HDIS;
283 define EtProcOfDis ?+ @-> ... { |pat_or_08};
284 define TrProcOfDis EzProcOfDis .o. ProcOfDis .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtProcOfDis;
285
286 #9#####[BODY]+[FIN]##
287 define BodFin (Bod .o. GEL) " " Fin ;
288 define EzBodFin HBOD " " HFIN ;
289 define EtBodFin ?+ @-> ... { |pat_or_09};
290 define TrBodFin EzBodFin .o. BodFin .o. EtBodFin;
291
292 #10####[QUAL]+[DIS]+of[IN]+[BODY] ##
293 define QualDisOfBody Qua " " Dis " " {of} " " (Bod .o. [GEL|SinGEN]);
294 define EzQualDisOfBody HQUA " " HDIS " " {of} " " HBOD;
```

```
295 define EtQualDisOfBody ?+ @-> ... { |pat_or_10} ;
296 define TrQualDisOfBody EzQualDisOfBody .o. QualDisOfBody .o. KenOf .o.
        OrdAldatuAzkenaLehenera .o. EtQualDisOfBody ;
297
298 #11####[QUAL]+[DIS] ## ### [QUAL]+[PROC]...
299 define QualDis Qua " " [Dis] ;
300 define EzQualDis HQUA " " [HDIS] ;
301 define EtQualDis ?+ @-> ... { |pat_or_110} ;
302 define TrQualDis EzQualDis .o. QualDis .o. EtQualDis ;
303
304 define QualProc Qua " " [Proc] ;
305 define EzQualProc HQUA " " [HPROC] ;
306 define EtQualProc ?+ @-> ... { |pat_or_111} ;
307 define TrQualProc EzQualProc .o. QualProc .o. EtQualProc ;
308
309 define QualObv Qua " " [Obv] ;
310 define EzQualObv HQUA " " [HOBV] ;
311 define EtQualObv ?+ @-> ... { |pat_or_112} ;
312 define TrQualObv EzQualObv .o. QualObv .o. EtQualObv ;
313
314 define QualFin Qua " " [Fin] ;
315 define EzQualFin HQUA " " [HFIN] ;
316 define EtQualFin ?+ @-> ... { |pat_or_113} ;
317 define TrQualFin EzQualFin .o. QualFin .o. EtQualFin ;
318
319 define QualBod Qua " " [Bod] ;
320 define EzQualBod HQUA " " [HBOD] ;
321 define EtQualBod ?+ @-> ... { |pat_or_114} ;
322 define TrQualBod EzQualBod .o. QualBod .o. EtQualBod ;
323
324 #12###[QUAL]+[QUAL]+[BODY] ###[QUAL]+[QUAL]+[DIS] ##
325 define QualQualBody Qua " " Qua " " [Bod] ;
326 define EzQualQualBody HQUA " " HQUA " " [HBOD] ;
327 define EtQualQualBody ?+ @-> ... { |pat_or_120} ;
328 define TrQualQualBody EzQualQualBody .o. QualQualBody .o. EtQualQualBody
        ;
329
330 define QualQualDis Qua " " Qua " " [Dis] ;
331 define EzQualQualDis HQUA " " HQUA " " [HDIS] ;
332 define EtQualQualDis ?+ @-> ... { |pat_or_121} ;
333 define TrQualQualDis EzQualQualDis .o. QualQualDis .o. EtQualQualDis ;
334
335 #13###[QUAL]+[BODY]+of+[BODY] ##
336 define QualBodyOfBody Qua " " Bod " " {of} " " (Bod .o. [GEL|SinGEN]) ;
337 define EzQualBodyOfBody HQUA " " HBOD " " {of} " " HBOD;
338 define EtQualBodyOfBody ?+ @-> ... { |pat_or_13} ;
339 define TrQualBodyOfBody EzQualBodyOfBody .o. QualBodyOfBody .o. KenOf .o
        . OrdAldatuAzkenaLehenera .o. EtQualBodyOfBody;
340
341 #14###[SUBS]+[QUAL]+[SUBS] ##
342 define SubsQualSubs Phar " " Qua " " Phar ;
343 define EzSubsQualSubs HPHAR " " HQUA " " HPHAR ;
344 define EtSubsQualSubs ?+ @-> ... { |pat_or_14} ;
345 define TrSubsQualSubs EzSubsQualSubs .o. SubsQualSubs .o. EtSubsQualSubs
        ;
346
347 #15###[PHAR|SUBS]+[QUAL]+[QUAL] ##
```

```
348 define PharQualQual Phar " " Qua " " Qua;
349 define EzPharQualQual HPHAR " " HQUA " " HQUA;
350 define EtPharQualQual ?+ @-> ... { |pat_or_15} ;
351 define TrPharQualQual EzPharQualQual .o. PharQualQual .o. EtPharQualQual
        ;
352
353 #16###[QUAL]+[BODY]+[BODY] ##
354 define QualBodyBody Qua " " (Bod .o. SinGEN) " " Bod ;
355 define EzQualBodyBody HQUA " " HBOD " " HBOD;
356 define EtQualBodyBody ?+ @-> ... { |pat_or_16} ;
357 define TrQualBodyBody EzQualBodyBody .o. QualBodyBody .o. EtQualBodyBody
        ;
358
359 #17###[PROC]+of+[QUAL]+[BODY] ##
360 define ProcOfQualBod Proc " " {of} " " Qua " " (Bod .o. SinGEN) ;
361 define EzProcOfQualBod HPROC " " {of} " " HQUA " " HBOD ;
362 define EtProcOfQualBod ?+ @-> ... { |pat_or_17} ;
363 define TrProcOfQualBod EzProcOfQualBod .o. ProcOfQualBod .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtProcOfQualBod;
364
365 #18###[FIN]+of+[BODY] ##
366 define FinOfBod Fin " " {of} " " (Bod .o. [SinGEN|GEL]);
367 define EzFinOfBod HFIN " " {of} " " HBOD ;
368 define EtFinOfBod ?+ @-> ... { |pat_or_18} ;
369 define TrFinOfBod EzFinOfBod .o. FinOfBod .o. KenOf .o.
        OrdAldatuLehenaAzkenera .o. EtFinOfBod;
370
371 #19###[QUAL]+[PHARM]+[DIS] ##
372 define QualPharDis Qua " " (Phar .o. SinGEN) " " Dis ;
373 define EzQualPharDis HQUA " " HPHAR " " HDIS ;
374 define EtQualPharDis ?+ @-> ... { |pat_or_19} ;
375 define TrQualPharDis EzQualPharDis .o. QualPharDis .o. EtQualPharDis;
376
377 #20###[BODY]+[QUAL] ##
378 #define BodyQual (Bod .o. SinGEN) " " Qua;
379 define BodyQual Bod " " Qua;
380 define EzBodyQual HBOD " " HQUA;
381 define EtBodyQual ?+ @-> ... { |pat_or_20} ;
382 define TrBodyQual EzBodyQual .o. BodyQual .o. EtBodyQual;
383
384 #21###[FIN]+[QUAL] eta [PROC]+[QUAL] #
385 define FinQual [Fin] " " Qua;
386 define EzFinQual [HFIN] " " HQUA;
387 define EtFinQual ?+ @-> ... { |pat_or_210} ;
388 define TrFinQual EzFinQual .o. FinQual .o. EtFinQual ;
389
390 define ProcQual [Proc] " " Qua;
391 define EzProcQual [HPROC] " " HQUA;
392 define EtProcQual ?+ @-> ... { |pat_or_211} ;
393 define TrProcQual EzProcQual .o. ProcQual .o. EtProcQual ;
394
395 #22###[QUAL]+[DIS]+[BODY]
396 define QualDisBody Qua " " Dis " " (Bod .o. SinGEN) ;
397 define EzQualDisBody HQUA " " HDIS " " HBOD;
398 define EtQualDisBody ?+ @-> ... { |pat_or_22} ;
399 define TrQualDisBody EzQualDisBody .o. QualDisBody .o.
        OrdAldatuAzkenaLehenera .o. EtQualDisBody ;
```

```
400
401 #espezifikoak
402 regex TrEpDis .o. CLEAN2 ;
403 regex TrEpPOSDis .o. CLEAN2 ;
404 regex TrBodStr .o. CLEAN2 ;
405 regex TrStrOfBod .o. CLEAN2 ;
406 regex TrDefOfPhar .o. CLEAN2 ;
407 regex TrNeoOfFinOfBod .o. CLEAN2 ;
408 regex TrPharAll .o. CLEAN2 ;
409 regex TrPharMea .o. CLEAN2 ;
410 regex TrPharAntMea .o. CLEAN2 ;
411 regex TrPharPoi .o. CLEAN2 ;
412 regex TrPharOve .o. CLEAN2 ;
413 regex TrLesOfBod .o. CLEAN2 ;
414 regex TrBenDis .o. CLEAN2 ;
415 regex TrFinOfObv .o. CLEAN2 ;
416 regex TrDisOfUndInt .o. CLEAN2 ;
417 regex TrPharAdvReac .o. CLEAN2 ;
418 regex TrEntBod .o. CLEAN2 ;
419 regex TrProcOnBody .o. CLEAN2 ;
420 regex TrProcToBody .o. CLEAN2 ;
421
422 #orokorrak
423 regex TrProcOfBod  .o. CLEAN2 ;
424 regex TrDisOfBod  .o. CLEAN2 ;
425 regex TrBodOfBod  .o. CLEAN2 ;
426 regex TrPhDis .o. CLEAN2 ;
427 regex TrBodProc .o. CLEAN2 ;
428 regex TrBodDis .o. CLEAN2 ;
429 regex TrDisWithDis .o. CLEAN2 ;
430 regex TrProcOfBodOfBod .o. CLEAN2 ;
431 regex TrPharPlusPhar .o. CLEAN2 ;
432 regex TrFinOfPhar .o. CLEAN2 ;
433 regex TrProcOfDis .o. CLEAN2 ;
434 regex TrBodFin .o. CLEAN2 ;
435 regex TrQualDisOfBody .o. CLEAN2 ;
436 regex TrQualDis .o. CLEAN2 ;
437 regex TrQualProc .o. CLEAN2 ;
438 regex TrQualObv .o. CLEAN2 ;
439 regex TrQualFin .o. CLEAN2 ;
440 regex TrQualBod .o. CLEAN2 ;
441 regex TrQualQualBody .o. CLEAN2 ;
442 regex TrQualQualDis .o. CLEAN2 ;
443 regex TrQualBodyOfBody .o. CLEAN2 ;
444 regex TrSubsQualSubs .o. CLEAN2 ;
445 regex TrQualDisBody .o. CLEAN2 ;
446 regex TrQualBodyBody .o. CLEAN2 ;
447 regex TrProcOfQualBod .o. CLEAN2 ;
448 regex TrFinOfBod .o. CLEAN2 ;
449 regex TrQualPharDis .o. CLEAN2 ;
450 regex TrBodyQual .o. CLEAN2 ;
451 regex TrFinQual .o. CLEAN2 ;
452 regex TrProcQual .o. CLEAN2 ;
453 regex TrSubsQualSubs .o. CLEAN2 ;
```

C.1 irudia – Termino habiratuen sorkuntzarako erregelak.