Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

Informatika Fakultatea

# Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz

**Izaskun Etxeberria Uztarrozek**

Informatikan Doktore titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

Donostia, 2016ko ekaina

Lengoaia eta Sistema Informatikoak Saila

Informatika Fakultatea

# Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz

Izaskun Etxeberria Uztarrozek Iñaki Alegriaren eta Montse Maritxalarren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua

Donostia, 2016ko ekaina

*Jesusi, Iñakiri eta Maddiri*

*Amari eta aitari*

# Eskerrak

Askotan pentsatu izan dut zer nolako esker-hitzak idatziko nituzkeen atal honetan, inoiz tesia amaitzera iritsiko banintz. Horiek idazteko unea iritsi da, niri sinesgaitza iruditu arren.

Askotan pentsatu izanak, ordea, ez dit askorik laguntzen oraintxe bertan, eta ez dut hitz egokirik aurkitzen nire esker on guztia adierazteko hainbeste laguni.

- Iñaki eta Montse zuzendariak, lan guztian zehar "xuxen-xuxen" gidatu nauzuelako. Ezingo nituzke zuek baino zuzendari hobeak aurkitu poliki-poliki aurrera egiteko. Iñaki, zure lana, eta bereziki, zure adorea funtsezkoak izan dira hasieratik bukaeraraino.

- Larraitz eta Mans, lan hau ez litzatekeelako honaino iritsiko zuek gabe. Tesi-lan hau talde-lana da, eta zuena ere bada hemen jasotakoa.

- Xabier Artola, eskatu dizudan guztietan nirekin egoteko tartea egin duzulako zure agenda konplikatu horretan.

- Alicia, txostenaren azken bertsiora iristeko eman didazun laguntza eta animo guztiagatik.

- Arantza D., oso lanpetuta ibili arren, beti izan zarelako niri laguntzeko prest zure atea jo dudanean.

- Itziar Irigoien, estatistikako kontuekin lagundu didazulako zuregana joan naizen bakoitzean.

- IXA taldeko lagunak, askotariko zalantzak argitzeagatik bidean zehar: programekin, artikuluekin, burokraziako paperekin... Asko zarete eta guztiak zaituztet gogoan nahiz eta zerrenda ez eman.

- Fakultateko hainbat lagun, sail guztietakoak, hainbeste urtetan elkarrekin egindako lanarengatik, eta azken urte hauetan hainbeste animo emateagatik. Guztia behar izan dut honaino iristeko. Asko zarete eta guztiei ematen dizkizuet eskerrak, baina ezin dut bukatu izen batzuk aipatu gabe: Agus, Olatz Arbelaitz, Olatz Arregi, Txelo eta Amaya.

- Josef Novak ikerlaria, Phonetisaurus tresna maisuki garatu duzulako. Funtsezkoa izan da zure tresna lan honetan, eta galderaren bat egin dizudanean berehala erantzun didazu beti, hor nonbaitetik.

- J. Porta, Y. Scherrer eta T. Erjavec ikerlariak, gurekin konpartitu dituzuelako zuen ikerkuntzako datuak eta metodoak. Zuen lankidetzarik gabe tesi-lan hau hanka-motz geratuko zen.

- Etxekoak, nire bihotzekoak: Maddi, Iñaki eta Jesus. Lan hau niretzat garrantzitsua zela ulertu duzuelako. Zuek pairatu duzue zenbait kezkak eta estresak niregan eragindako umore txarra, eta zuei lapurtu dizkizuet ordu asko. Zorionez, bukatzen ari naiz.

- Ama eta aita, nire eredua zaretelako eta beti hor zaudetelako.

- Anai-arrebak –Eukene, Agus, Alex, Nekane, Josu– eta familiako gainerakoak, zuek ere hor zaudetelako beti.

Guztiak zaituztet gogoan. Mila esker!

# Gaien aurkibidea


**I Tesi-lanaren nondik norakoak** **1**

I.1 Sarrera . . . . . 1
I.2 Lanaren kokapena . . . . . 3
I.3 Helburuak . . . . . 6
I.4 Txostenaren egitura . . . . . 7
I.5 Argitalpenak . . . . . 9

**II Aldaera-corpusak** **11**

II.1 Sarrera . . . . . 11
II.2 Testu historikoen erabilpena liburutegi digitaletan . . . . 12
II.3 Euskarazko corpus historikoa . . . . . 16
II.3.1 Euskal literaturako klasiko digitalizatuak . . . . . 16
II.3.2 Hainbat obraren aukeraketa . . . . . 19
II.3.3 Lehenengo azterketa . . . . . 20
II.3.4 Corpusaren prestaketa: *Gero* . . . . . 25
II.3.5 Corpusaren prestaketa: *Peru Abarka* . . . . . 31
II.4 Euskarazko corpus dialektala . . . . . 34
II.5 Beste hizkuntza batzuetako corpusak . . . . . 35
II.5.1 Gaztelaniazko corpusa . . . . . 36
II.5.2 Eslovenierazko corpusa . . . . . 41

**III Metodoen aurkezpena eta hautaketa** **47**

III.1 Sarrera . . . . . 47
III.2 Azterketa bibliografikoa . . . . . 48
III.2.1 Erregeletan oinarritutako metodoak . . . . . 49
III.2.2 Aldaera fonologikoak ikasteko metodoak . . . . . 50
III.3 Fonologiaren inferentzia . . . . . 54
III.3.1 Erregela fonologikoak . . . . . 58
III.3.2 Lehenengo metodoa: lexdiff . . . . . 59
III.3.3 Bigarren metodoa: ILP motako algoritmoa . . . . . 62

III.3.4 Hirugarren metodoa: WFST teknologia . . . . . 64
III.4 Esperimentuak . . . . . 67
III.4.1 Esperimentuen diseinua . . . . . 67
III.4.2 Oinarri-lerroak . . . . . 69
III.4.3 Lehenengo metodoaren emaitzak (lexdiff) . . . . . 71
III.4.4 Bigarren metodoaren emaitzak (ILP) . . . . . 74
III.4.5 Hirugarren metodoaren emaitzak (WFST) . . . . . 76
III.4.6 Informazio morfologikoaren erabilera . . . . . 78
III.5 Esperimentuen ondorioak eta erabakiak . . . . . 82

**IV Testu historikoen normalizazioa WFST erabiliz 85**
IV.1 Sarrera . . . . . 85
IV.2 WFST teknologia . . . . . 86
IV.2.1 Teknologiaren oinarriak . . . . . 86
IV.2.2 Aplikazioak . . . . . 89
IV.3 Phonetisaurus tresnaren deskribapen zehatza . . . . . 90
IV.3.1 Eredu estatistikoak . . . . . 92
IV.3.2 Azken urratsa: deskodeketa . . . . . 96
IV.3.3 Phonetisaurusen erabilera beste lan batzuetan . . . 97
IV.3.4 Phonetisaurusen erabilera aldaeren normalizazioan 97
IV.3.4.1 Datu-prestaketa . . . . . 98
IV.3.4.2 Datuen lerrokatzea . . . . . 98
IV.3.4.3 *Joint n-gram* ereduaren entrenatzea . . . . . 99
IV.3.4.4 Deskodeketa . . . . . 99
IV.4 *Gero* corpusa: esperimentuak eta emaitzak . . . . . 100
IV.4.1 Oinarrizko doikuntza (1. esperimentua) . . . . . 103
IV.4.2 Maiztasunaren eragina (2. esperimentua) . . . . . 108
IV.4.3 Hitz estandarren eragina (3. esperimentua) . . . . 113
IV.4.4 Azken ebaluazioa: test-corpusa . . . . . 115
IV.4.5 Ondorioak . . . . . 118
IV.5 *Peru Abarka* corpusa: esperimentuak eta emaitzak . . . . 119
IV.5.1 Oinarrizko doikuntza (1. esperimentua) . . . . . 120
IV.5.2 Hitz estandarren eragina (3. esperimentua) . . . . 122
IV.5.3 Azken ebaluazioa: test-corpusa . . . . . 124
IV.6 Esperimentuak gaztelaniarekin eta eslovenierarekin . . . . 127
IV.6.1 Gaztelania zaharra: esperimentuak . . . . . 127
IV.6.1.1 Esperimentuak FL-EM datu-multzoarekin . . . 127
IV.6.1.2 Esperimentuak IMPACT datu-multzoarekin . . 132
IV.6.1.3 Esperimentu berria: datu-multzoak gurutzatu . 133
IV.6.2 Esloveniera historikoa: esperimentuak . . . . . 134

**V Morfologiaren ekarpena 139**
V.1 Sarrera . . . 139
V.2 Segmentazio morfologiko ez-gainbegiratua . . . 140
V.2.1 Bibliografia . . . 140
V.2.2 Hitzen segmentazioa Morfessor bitartez . . . 142
V.2.2.1 Corpusa . . . 143
V.2.2.2 Esperimentuak . . . 144
V.3 Normalizazioa morfologiaren bitartez . . . 148
V.3.1 Ikasteko informazio berria: *analisia-analisia* . . . 148
V.3.2 Segmentazio-eredua hobetzeko saioak . . . 150
V.4 Morfologiaren ikasketa modu gainbegiratuan . . . 152
V.4.1 Morfemen lerrokatzea . . . 152
V.4.2 Estandarraren hedapena eta ebaluazioa . . . 154
V.5 WFST sistemak: emaitzen analisia . . . 156
V.6 Ondorioak . . . 161

**VI Ondorioak, ekarpenak eta etorkizuneko lanak 163**
VI.1 Sarrera . . . 163
VI.2 Ondorio nagusiak . . . 164
VI.3 Ekarpenak . . . 167
VI.4 Etorkizuneko lanak . . . 170

**Bibliografia 173**

**ERANSKINAK 1**

**A Brat anotazio-tresna erabiltzeko gida laburra 1**
A.1 Sarrera . . . 1
A.2 Anotazio-tresna abiatzeko urratsak . . . 1
A.3 Etiketatzea . . . 2
A.3.1 Testuan ageri diren “markak” . . . 3
A.3.2 Etiketak . . . 5
A.4 Etiketatzearen inguruko oharrak . . . 10
A.5 Anotazio-prozesua azkartzeko aukera . . . 13

**B Phonetisaurus tresnaren erabilera 15**
B.1 Sarrera . . . 15
B.2 Tresnaren urratsak eta komandoak . . . 15
B.2.1 Datu-prestaketa . . . 15

B.2.2 Datuen lerrokatzea . . . . . 16
B.2.3 *Joint n-gram* ereduaren entrenatzea . . . . . 17
B.2.4 Deskodeketa . . . . . 18

**C Izen propioak *Gero* corpusean 19**

# Irudien zerrenda


II.1 *Gero* obrako paragrafo baten irudia. . . . . 27
II.2 Testu zati bat Brat aplikazioarekin anotatzeko prest. . . . . 30
II.3 Testu zati bat Brat aplikazioarekin anotatu eta gero. . . . . 30
II.4 *Peru Abarka* obraren testu zati bat. . . . . 32
II.5 Testu zati bat anotatzeko prest: AUTO etiketa automatikoak. 33

III.1 Proposatutako hautagaien lehen iragazkia . . . . . 57
III.2 Doitasunaren eta estalduraren arteko konpentsazioa. . . . . 74

IV.1 Egoera finituko transduktore baten irudia. . . . . 87
IV.2 FL-EM multzoa. Esperimentuetarako banaketa . . . . . 128
IV.3 FL-EM multzoa. Ikaketa-kurbak . . . . . 130

V.1 Morfessor. Tokenak vs formak oinarrizko ereduarekin . . . . . 146
V.2 Morfessor. Gamma banaketaren eragina segmentazio-ereduan . 148
V.3 Morfessor. Segmentazio-ereduaren adibide batzuk . . . . . 151

A.1 Brat anotazio-tresna. Ongietorriko leihoa . . . . . 2
A.2 Anotatzeko informazioaren antolaketa fitxategietan . . . . . 3
A.3 Anotatu beharreko testu baten hasierako irudia . . . . . 4
A.4 Aplikazioan sartzeko leihoa . . . . . 6
A.5 Erabiltzailea eta pasahitza . . . . . 6
A.6 Anotaziorako aurredefinitu diren etiketak . . . . . 7
A.7 *Aldaera* etiketari lotutako informazioa . . . . . 8
A.8 Testu baten itxura anotazioa egin ondoren . . . . . 10
A.9 Nola elkartu bi hitz anotazio berean (I) . . . . . 11
A.10 Nola elkartu bi hitz anotazio berean (II) . . . . . 12
A.11 Elkartutako bi hitzen itxura anotatu ondoren . . . . . 13
A.12 Brat anotazio-tresnaren aukerak aldatzeko leihoa . . . . . 14

B.1 Phonetisaurus. Ikasteko informazioaren adibideak . . . . . 16
B.2 Phonetisaurus. Ikasteko datuak lerrokatzearen ondoren . . . . 17

# Taulen zerrenda


II.1 Analizatutako obrak: baztertu diren token kopuruak. . . . . 21
II.2 Analizatutako obrak: token eta forma kopuruak. . . . . 22
II.3 Analizatutako obrak: transliterazio-erregelen efektua. . . . . 24
II.4 *Gero*: ikasteko eta testeko zatiak. . . . . 28
II.5 *Peru Abarka*: ikasteko eta testeko zatiak. . . . . 32
II.6 Lapurtera/Estandarra corpuseko adibideak. . . . . 35
II.7 Lapurtera/Estandarra corpusa: ikasteko eta testeko zatiak. . . 35
II.8 FL-EM multzoa: kategoria bakoitzaren sarrera kopurua. . . . . 36
II.9 FL-EM multzoko adibide batzuk: egungo lema eta kategoria. . 37
II.10 FL-EM multzoko adibide batzuk: egungo lema eta analisi morfosintaktikoa. . . . . 37
II.11 FL-EM multzoko adibide batzuk: egungo forma . . . . . 38
II.12 FL-EM multzoko adibide batzuk: egungo forma bat baino gehiago . . . . . 38
II.13 FL-EM multzoa: esperimentuetarako bikote kopurua. . . . . 39
II.14 IMPACT multzoko adibide batzuk. . . . . 41
II.15 IMPACT multzoa: esperimentuetarako bikote kopurua. . . . . 41
II.16 Eslovenierazko corpusak: *goo* eta *foo*. . . . . 43
II.17 Eslovenierazko lexikoien ezaugarriak. . . . . 44
II.18 Eslovenierazko lexikoietako adibide batzuk. . . . . 45
II.19 Eslovenierazko lexikoien hainbat kopuru. . . . . 45

III.1 Aldaera fonologikoen ikasketa. Laburpena . . . . . 55
III.2 Lapurtera/Estandarra corpus paraleloaren hainbat kopuru. . . 68
III.3 Oinarri-lerroko hiru sistemen emaitzak . . . . . 70
III.4 Lexdiff metodoaren emaitza batzuk . . . . . 72
III.5 Lexdiff metodoa: hiru emaitza onenak . . . . . 73
III.6 ILP metodoaren emaitza batzuk . . . . . 75
III.7 WFST metodoa: doikuntzako esperimentuen emaitzak . . . . . 77
III.8 Metodo guztien emaitzen laburpena . . . . . 78

III.9 WFST metodoa: doikuntzako esperimentuak bi transduktore berriekin . . . 81
III.10 WFST metodoaren emaitzak hiru transduktoreekin . . . 82
III.11 Laburpena. Hiru metodoekin lortutako emaitza onenak . . . 83

IV.1 Phonetisaurus: sarrerako lexikoia . . . 98
IV.2 Phonetisaurus: lexikoiko datuen lerrokatzea . . . 99
IV.3 Phonetisaurus. Deskodeketa-urratsaren adibideak . . . 100
IV.4 *Gero* corpusa: ikasteko eta testeko zatiak. . . . 101
IV.5 *Gero*. 1. esperimenturako kopuruak . . . 103
IV.6 *Gero*. Oinarri-lerroaren emaitzak 1. esperimentuan . . . 104
IV.7 *Gero*. Oinarrizko esperimentua WFST metodoarekin. . . . 106
IV.8 *Gero*. WFST metodoaren emaitzak 1. esperimentuan (I) . . . 106
IV.9 *Gero*. WFST metodoaren emaitzak 1. esperimentuan (II) . . . 107
IV.10 *Gero*. 2. esperimenturako kopuruak . . . 108
IV.11 *Gero*. Oinarri-lerroaren emaitzak 2. esperimentuan (I) . . . 109
IV.12 *Gero*. Oinarri-lerroaren emaitzak 2. esperimentuan (II) . . . 109
IV.13 *Gero*. WFST metodoaren emaitzak 2. esperimentuan (I) . . . 110
IV.14 *Gero*. WFST metodoaren emaitzak 2. esperimentuan (II) . . . 111
IV.15 *Gero*. WFST metodoaren emaitzak 2. esperimentuan (III) . . . 111
IV.16 *Gero*. WFST metodoaren emaitzak 2. esperimentuan (IV) . . . 112
IV.17 *Gero*. 3. esperimenturako kopuruak . . . 113
IV.18 *Gero*. WFST metodoaren emaitzak 3. esperimentuan (I) . . . 114
IV.19 *Gero*. WFST metodoaren emaitzak 3. esperimentuan (II) . . . 115
IV.20 *Gero*. Azken ebaluaziorako kopuruak . . . 117
IV.21 *Gero*. Azken ebaluazioko emaitzak . . . 117
IV.22 *Peru Abarka* corpusa: ikasteko eta testeko zatiak . . . 119
IV.23 *Peru Abarka*. 1. esperimenturako kopuruak . . . 120
IV.24 *Peru Abarka* corpusa. Oinarri-lerroaren emaitzak 1. esperimentuan . . . 121
IV.25 *Peru Abarka*. WFST metodoaren emaitzak 1. esperimentuan . 121
IV.26 *Peru Abarka*. 3. esperimenturako kopuruak . . . 123
IV.27 *Peru Abarka*. WFST metodoaren emaitzak 3. esperimentuan . 123
IV.28 *Peru Abarka*. Azken ebaluaziorako kopuruak . . . 124
IV.29 *Peru Abarka*. Azken ebaluazioko emaitzak . . . 125
IV.30 FL-EM multzoa. Esperimentuen emaitzak . . . 130
IV.31 IMPACT multzoa. Esperimentuen emaitzak . . . 133
IV.32 FL-EM eta IMPACT multzoak batera: esperimentuen emaitzak . . . 134
IV.33 Eslovenierazko lexikoien ezaugarriak. . . . 135

IV.34 Esloveniera. Esperimentuen emaitzak . . . . . 136
IV.35 Esloveniera. CSMT eta WFST sistemen emaitzak . . . . . 137

V.1 Morfessor. Esperimentuetarako multzoen tamainak . . . . . 143
V.2 Gero corpusa. WFST sistema berria: *analisia-analisia* . . . . . 150
V.3 Analizatzaile hedatuekin lortutako emaitzak . . . . . 155
V.4 WFST sistemaren eta analizatzaile hedatuen konbinazioa . . . 156
V.5 Hiru WFST sistemen konbinazioa: bozketa . . . . . 158
V.6 WFST sistemen osagarritasuna: orakulua . . . . . 158
V.7 Balizko orakulu onenaren muga . . . . . 160

C.1 Emaitza berriak izen propioen tratamendua aldatuta . . . . . 20

# I. KAPITULUA

## Tesi-lanaren nondik norakoak

### I.1 Sarrera

Hizkuntzaren azterketa eta prozesamenduaren barnean, tesi-lan hau testu ez-estandarren ikertze-arloan kokatzen da, zehazki, euskarazko testu ez-estandarren arloan. Oro har, testu estandarrekin alderatuta, testu ez-estandarrek ezaugarri bereziak uzten dituzte agerian maila lexikoan, morfologikoan edota fonologikoan, eta hala, haien prozesaketa erronka bat da.

Euskararen kasuan, oso jatorri eta garai ezberdineko testuak sartzen dira testu ez-estandarren multzoan. Alde batetik, ez-estandarrak dira euskararen estandarizazio-prozesua baino lehen idatzitako testu guztiak, hau da, lau mendetan zehar sortutako testuak (XVI. mendetik XX. mende erdira artekoak). Beste aldetik, estandarizazio-prozesuaren ondoren idatzitako guztia ez da estandarra izan, noski, dialekto aberats eta ugariak baititu euskarak eta horiek ere erabili dira eta erabiltzen dira egun. Halaber, testu ez-estandar ugari aurkitzen dugu gaur egun sare sozialetan, askotan sailkatzeko zailak —ez dago argi dialektalak diren edo beste fenomeno batzuen adierazleak diren— baina, dena dela, ez-estandartzat jo behar dira horiek ere.

Jatorria alde batera utzita, testu ez-estandarren prozesatzeak, oro har, arazo berberarekin egiten du topo beti: hizkuntza prozesatzeko tresna gehienak (NLP, *Natural Language Processing* tresnak) hizkuntza estandarretan idatzitako testuak prozesatzeko garatu izan dira, eta testu ez-estandarrekin erabiltzen direnean, asko jaisten da haien errendimendua.

Baina, beharrezkoa al da testu ez-estandarrak prozesatzea? Non sortzen da halako beharra? Alde batetik, liburutegi digitalen hedapena dago. Gaur egun liburutegi digital ugari daude Internet bitartez atzigarri, eta aukera

berri bat eskaintzen diote halako liburutegiek publiko zabalari: orain arte adituek soilik zuten hainbat dokumentu preziatu eta urri kontsultatzeko aukera. Baina dokumentu horietako asko aspaldikoak dira, ez daude egungo hizkuntzaren arabera idatziak eta bertan kontsultak egitea ez da egungo dokumentuetan bezain erraza: testu ez-estandarrak dira.

Liburutegi digitalez gain, ikertze-arlo berri bati buruz hitz egiten da gaur egun gero eta gehiago: Humanitate Digitalen arloa (*Digital Humanities*). Informazio digitala prozesatzeko baliabideak eta indarrak humanitateen ikerkuntzaren eskura jartzea da arlo berri horren jomuga (Piotrowski, 2012), eta horrek zera eskatzen du, humanitateetako tradiziozko metodo kualitatiboak eta egungo metodo kuantitatiboak zein aplikazioak biltzea. Egungo aplikazioak dira informazio-berreskuratzea (IR, *Information Retrieval*), informazio-erauzketa (IE, *Information Extraction*), testu-analisia, datu-meatzaritza (*Data mining*) eta abar. Azken urteetan egin dira ikerketa batzuk, non bi arloetako metodoak eta teknikak konbinatu diren. Horien artean koka daiteke, esaterako, Pettersson eta Nivre ikerlarien lana (2011), non suediera historikoan idatzitako testuetatik aditzak automatikoki erauzten dituzten, historialariek ikertu ahal dezaten horietan oinarrituta suediar gizartearen bizibideak 1550–1800 garaian.

Nazioarteko hainbat ekimen dago humanitate digitalen arloarentzat azpiegitura egokiak sortzeko (Piotrowski, 2012). Horien artean daude, esaterako, Europak babestutako CLARIN[1] (*Common Language Resources and Technology Infraestructure*) eta DARIAH[2] (*Digital Research Infrastructure for the Arts and Humanities*) proiektuak. Humanitate digitalen arlo berri horretan, beraz, giza-zientzientzat interesekoak diren testu historikoak prozesatu behar dira beste hainbat testuren artean, hau da, testu ez-estandar ugari prozesatu behar dira.

Bestalde, ikertze-arlo berriak sortzen ari dira orain arte banatuak ikusten ziren arloen ikerkuntzak gerturatzen ari diren heinean, eta hala gertatzen ari da, esaterako, soziolinguistika eta linguistika konputazionalaren artean (Nguyen *et al.*, 2015). Hizkuntza giza-fenomeno bat da, uneoro aldatzen ari den fenomenoa, eta hizkuntzalaritza konputazionaleko ikerlariek gero eta interes handiagoa dute hizkuntzaren giza-ikuspegi horretan. Nguyen *et al.* egileen lanean (2015) "soziolinguistika konputazionala" arlo berria aipatzen da, non ikerketa konputazionala soziolinguistikako hainbat gairi aplikatzen zaion. Gai horien artean daude hizkuntza eta identitate soziala, hizkuntzaren erabilera giza-erlazioetan edota komunikazioa hizkuntza anitzetan. Arlo

[1] `http://www.clarin.eu/` (2016-03-15ean atzitua)
[2] `http://www.dariah.eu/`(2016-03-15ean atzitua)

horietako ikerketa asko sare sozialen inguruan egiten dira, hau da, hedapen azkarra eta erabilpen zabala duten sareetan, Twitter, LinkedIn, Facebook eta antzekoetan. Milioika lagunek erabiltzen dituzten sareak dira, eta bertan sortzen diren edota bertatik hedatzen diren testuen gaineko ikerketak interes handia piztu du gizarte-zientzietan. Testu horietako asko bat-batekoak dira eta ez dituzte arau estandarrak jarraitzen, eta hala, testu historikoen antzera, haien prozesaketa ere erronka bat da.

Sare sozialen hedapen azkarrarekin batera, iritziak emateko joera izugarri zabaldu da sarean. Edozein gairi edo produkturi buruz iritzia emateko guneak zein *blog*ak gero eta ugariagoak dira Interneten, eta jendeak besteen iritziak bilatzen eta jarraitzen ditu (Pang eta Lee, 2008). Interes horrek bultzatu du, hein batean behintzat, iritziak kontuan hartzen dituzten sistemak eraiki nahi izatea, eta hala, iritzien eta sentimenduen analisiaren ikerketa oso popularra bilakatu da azken urteetan. Iritziak, sentimenduak eta subjektibitatea analizatu nahi dira testuetan, eta, lehen aipatu den moduan, testu horietako asko ez-estandarrak dira.

Aurreko aipamenak kontuan izanik, agerikoa da testu ez-estandarren prozesatzea intereseko gai bat dela gaur egun, eta IXA ikerketa-taldeak[3] interesa du gai horretan sakontzeko, batez ere, euskararen ikuspuntutik. Hizkuntzaren prozesamenduan sortzen diren erronka berriei aurre egitea izan da beti taldearen helburua, eta testu ez-estandarrak prozesatzea erronka horien artean kokatzen den gaia da. Tesi-lan honen helburua da bide hori jorratzeko lehen urratsa ematea.

Ikusi dugunez, jatorri ezberdineko testu ez-estandar asko dago, baina tesi-lan honetan testu historikoetan zentratuko gara batez ere. Lehenengo esperimentuak testu dialektalekin egingo baditugu ere, lanean zehar ikusiko dugu ikerketa nagusia testu historikoen gainean egin dugula. Halere, esan beharra dago aukera izan dugula gure ikerketa aplikatzeko egungo testu ez-estandarrak normalizatzeko saioetan, eta hala, 2013ko SEPLN kongresuan antolatutako *Tweet-Norm* izeneko tailerrean hartu genuen parte, non txioak normalizatzea zen planteatutako ataza (Alegria *et al.*, 2013).

## I.2 Lanaren kokapena

Sarreran aipatu denez, NLP tresna estandarrak ezin dira bere horretan erabili testu ez-estandarrak prozesatzeko, eta hala eginez gero lortutako errendimendua txikia da. Bi aukera plantea daitezke arazo horri irtenbide bat emateko: (1) NLP tresna berriak eraiki edota garatutakoak moldatu estan-

[3] `http://ixa.si.ehu.eus/Ixa`

darrak ez diren testuak prozesatzeko, edo (2) testu ez-estandarrak normalizatu[4] lehendabizi, gero tresna estandarrekin prozesatu ahal izateko.

Lehenengo irtenbideak, tresna berriak sortzeak edo dauden tresnak egoki moldatzeak, denbora asko eskatzen du, baliabide berriak sortzea edo dauden baliabideak egokitzea ez baita berehalako lana, eta, guk dakigunik, ez da hizkuntza zabalduenen artean jorratu den bidea. Euskararen esparrura etorrita, dialektoak ugariak direla kontuan izanik, ez da bideragarria dialekto bakoitzeko tresna berriak gauzatzen hastea, nahiz eta bizkaierarekin egin zen saio bat (Alegria *et al.*, 2010). Euskalkien kasuan, gainera, kontuan hartu behar edozein hizkuntzaren baliabide zein tresna konputazionalak garatu baino lehen, hizkuntza bera deskribatu behar dela zehatz, "arautu" egin behar dela hizkuntza, eta horrez gain, erabiltzaileen artean hedatu behar da arau hori, gerora prozesatu ahal izango diren testuak sortuko badituzte erabiltzaile horiek. Hori guztia lan zaila eta korapilatsua bilakatzen da euskalkien testuinguruan.

Bigarren irtenbideak, berriz, testuak normalizatzearen bideak, aukera ematen du aurretik garatutako tresnak moldatutako testu "berri" edo normalizatu horietan erabiltzeko, eta hori da hainbat hizkuntzatan jarraitu den bidea.

Normalizazioaren prozesua aurrera eramateko gaur egun erabiltzen diren teknikak hiru multzotan bana daitezke, oro har:

- Erregeletan oinarritutako metodoak dira hasieran batez ere erabili izan diren metodoak. Metodo horietan, testuetako aldaeretan gertatzen diren fenomenoak aztertzen dira, eta eskuz idazten dira erregela fonologikoz osatutako gramatikak. Erregela horiek aldaerei aplikatuz gero, dagozkien forma estandarrak aurkitzeko aukera dago.

- Teknika ez-gainbegiratuak deitutakoak (*unsupervised*), erabat automatikoak dira eta heuristikoetan edo etiketatu gabeko corpusetan oinarritzen dira. Teknika horien artean edizio-distantzia (batez ere, *Levenshtein* distantzia) edota distantzia fonetikoa (adibidez, *Soundex* algoritmoarekin lortua) dira erabilienak, aldaeretatik gertuen dauden forma estandarrak lortzeko. Halako metodoak oinarri-lerro gisa erabiltzen dira maiz, horiekin konparatzeko proposatzen diren sistema berriak.

- Ikasketa automatikoa oinarri duten teknikak ere aplika daitezke ataza ebazteko. Teknika horien bitartez, sistemari hainbat adibide ematen

[4]NLP arloan, *normalizazio* edo *kanonikalizazio* terminoak erabiltzen dira forma ez-estandarra eta estandarraren arteko mapaketa egiten duen prozesua izendatzeko.

zaizkio ikas dezan, eta gero eskatzen zaio ikasitakoa aplikatzea adibide berrietan. Ikasketarako adibideak lortzeko eskuzko lana behar denez, teknika gainbegiratuak direla esan ohi da. Ikusiko dugunez, ikasitakoa erregelak izan daitezke ala ez.

Normalizazioaren inguruko lanen azterketa bibliografikoan ikusiko dugunez (III. kapituluan), ataza ebazteko hainbat sistema eskuz idatzitako erregeletan oinarritu dira. Izan ere, ez-estandarra eta estandarraren arteko aldaketak nahiko erregularrak dira maiz, fonologiaren ildotik datozen aldaketak direlako. Baina badira beste arrazoi batzuk —aldaketa lexiko-morfologikoak, ortografia-konbentzioak— aldaketa ez hain erregularrak sortzeko, eta horregatik erregelen bitartez lortzen diren emaitzak ez dira beti nahi bezain onak. Euskararen morfologia aberatsa dela eta, aukera izango dugu tesi-lan honetan mota guztietako fenomenoak aztertzeko.

Eskuz idatzitako erregela-sistemen bideragarritasunaren ildotik, hainbat baliabide behar dira erregelak idazteko. Hasteko, bi eratako adituak behar dira: testu historikoen azterketan adituak, alde batetik, testuak aztertzeko eta bertan gertatzen diren fenomenoak identifikatzeko; eta erregelen formalizazioan adituak, beste aldetik, identifikatutako fenomenoak erregelen bitartez adierazteko, ongi zehaztuz erregela bakoitzak behar duen testuingurua. Testuen azterketak zein erregelen formalizazioak denbora luzea eraman dezake testu kopurua handia bada. Gainera, erregelak idatzi ondoren aplikatu egin behar dira normalizazioa egiteko, eta ongi analizatu behar dira lortzen diren emaitzak. Askotan, erregelak aplikatzeko algoritmoa ez da sinplea: Zer egin erregela bat baino gehiago aplika badaiteke une jakin batean? Zenbat erregela aplikatuko dira? Zer ordenatan? Zein erregelaren bitartez lortuko da emaitza onena? Kontuan izan behar da une bakoitzean aplikatzen den erregela erabakigarria dela atzetik aplika daitezkeenak aukeratzeko.

Beraz, eskuz idatzitako erregelak erabiltzen dituzten metodoak asko erabili badira ere, dena ezin da ebatzi erregelen bitartez, eta gaur egun beste zenbait aukera esperimentatzen ari dira zer nolako emaitzak lor daitezkeen aztertzeko.

Horixe izan da, hain zuzen ere, tesi-lan honetan hartu dugun bidea. Gure lanaren jomuga euskal testu ez-estandarren normalizazioaren inguruan kokatzen da, eta ez dugu planteatzen eskuz idatzitako erregeletan oinarritutako sistema bat garatzea, esan bezala, denbora luzea eta azterketa sakona eskatuko lukeen bidea baita hori. Lan honetan ikasketa automatikoan oinarritzen diren metodoak aplikatu nahi ditugu euskarazko testu ez-estandarretan normalizazioaren ataza ebazteko, eta horrez gain, gure metodoek lortzen dituzten emaitzak konparatu nahi ditugu beste metodo batzuek lor-

tzen dituztenekin. Konparaketa hori egin ahal izateko, beharrezkoa izango da, euskaraz gain, beste metodo horiek lantzen dituzten hizkuntzekin ere lan egitea.

## I.3 Helburuak

Tesi-lan honetan galdera nagusi honi erantzuten saiatuko gara:

Morfofonologia konputazionaleko tresnak erabiliz, ikas daiteke metodo bat euskarazko aldaerei, diakronikoei zein dialektalei, dagozkien forma estandarrak automatikoki esleitzeko?

Galdera horrek beste galdera batzuk dakartza: Zer doitasun lor daiteke ataza horretan? Zenbatekoa izan behar du ikasteko corpusak?

Galdera horiei erantzun ahal izateko, helburu hauek definitu ditugu tesi-lan honetan:

1. Aztertzea euskarazko testu ez-estandarren testuingurua jakiteko zer iturri dauden halako testuak eskuratzeko, horrela erabaki ahal izateko zein testurekin egin behar den lan.

2. Fonologian oinarritutako hainbat metodo lantzea normalizazioaren ataza ebazteko, eta horien emaitzak konparatzea literaturan aurkitutako metodoek lortutakoekin. Konparatu nahi direnak, batez ere, erregeletan oinarritzen diren metodoak dira, horien emaitzak gainditzen diren ala ez jakiteko.

3. Ohiko metodoak eta erregela-sistemak, ia denak, fonologian oinarrituak dira, eta ikertu nahi dugu ea morfologiaren gainerako osagaiek (lexikoa, morfotaktika) eragina izan dezaketen normalizazioaren atazan.

4. Landuko dugun sistemaren emaitzak beste sistema batzuek lortzen dituztenekin konparatu ahal izateko, beharrezkoa izango da sistema hori datu berrietan aplikatzea. Datu horiek eskuratzeko, beraz, beste hainbat ikerlariren kolaborazioa lortu behar dugu.

5. Ikasketa automatikoan oinarritzen diren metodoak erabili behar ditugunez, neurtu beharra dago ikasteko zenbat informazio behar den emaitzek kalitate minimoa izan dezaten.

6. Liburutegi digitalen munduan zabaltzen diren aukerak ikusirik, urrats bat aurrera eman nahi dugu, bide horretan aztertzeko ea metodologia

zehatz bat proposa dezakegun, lagunduko duena euskal testu historikoen prozesaketa gauzatzen egungo tresnen bitartez.

Tesi-lanaren helburuak finkatu ditugunez honez gero, une egokia da planteatzen dugun lanaren testuingurua ongi zehazteko. Lehenik eta behin argitu nahi dugu lan honetan ez dugula planteatzen ikasketa-metodo berriak asmatzea, baizik eta daudenak erabiltzea, ebaluatzea eta horien artean egokiena aukeratzea. Horrekin batera, metodoen arteko osagarritasuna ere aztertu nahi dugu.

Kontuan izan behar den beste ezaugarri garrantzitsua da zer nolako informazioa erabili behar dugun normalizazioaren ataza ebazteko. Izan ere, hitz-mailako informazioa erabiltzea da hasieratik planteatzen duguna, hau da, morfologia mailakoa soilik, testuen sintaxian sartu gabe.

Azkenik, argitu beharra dago ez ditugula testuetako aldaera guztiak landuko normalizazioan. Izan ere, lan honen abiapuntua hitz ez-estandarren detekzio automatikoan datza, eta argi izan behar dugu detekzio horretatik at geratuko direla zenbait aldaera, hitz estandar baten forma dutelako. Erroreen analisiaren arloan *real word errors* esaten zaie kasu horiei, hau da, testuinguruaren arabera erroreak dira baina hitz zuzenen forma dute. Aldaeren arlora etorrita, badira aldaeren artean egungo hitz estandarren forma dutenak, baina, esan bezala, horien detekzioa zein tratamendua tesi-lan honen esparrutik kanpo geratzen da.

## I.4 Txostenaren egitura

Tesi-txosten hau sei kapitulutan eta hainbat eranskinetan egituratuta dago:

1. Esku artean dugun I. kapituluaren xedea da lanaren motibazioa zein kokapena egitea IXA taldearen barruko estrategian, eta lanaren helburuak zehaztea. Halaber, txostenean topatuko dugun informazioaren berri ematen da kapitulu honetan, eta tesi-lanarekin loturik egin diren argitalpenak zerrendatzen dira.

2. II. kapituluan lanean zehar erabili diren corpusen berri ematen da. Dagoeneko aipatu da euskarazko testuekin aritzea dela tesi-lan honen helburu nagusia, baina beharrezkoa dela beste hizkuntza batzuetako testuak edota datuak ere lantzea, alderaketak egin ahal izateko. Hala, lanean erabili diren corpus guztien xehetasunak ematea da bigarren kapituluaren helburua.

3. III. kapituluaren xedea bikoitza da. Kapituluaren hasieran azterketa bibliografikoa egiten da jakiteko zein diren normalizaziorekin lotutako lanetan planteatu diren metodoak eta lortutako emaitzak. Horren ondoren, tesi-lan honetan probatu ditugun metodoak aurkezten dira eta metodo horien ebaluazioa egiten da corpus jakin baten gainean. Ebaluazio horren helburua metodo onena aukeratzea da, hurrengo kapituluan metodo hori aplikatzeko gainontzeko corpusetan.

4. Esperimentuei dagokien kapitulu nagusia IV. kapitulua da. Aukeratutako metodoan sakondu ondoren haren oinarriak eta funtzionamendua argi uzteko, metodoaren doikuntza egiten da hainbat esperimenturen bitartez, eta azkenean, metodoaren portaera ebaluatzen da euskarazko bi corpusetan, bietan antzeko emaitzak lortzen ote diren aztertzeko. Kapitulua bukatzeko, gure metodoa beste bi corpusetan aplikatzen da: gaztelaniazko zein eslovenierazko corpusetan. Esperimentu berrien helburua bikoitza da: alde batetik, agerian uztea metodoa hizkuntzarekiko independentea dela, eta, beste aldetik, metodoak lortzen dituen emaitzak konparatzea beste metodo batzuek lortutakoekin.

5. V. kapituluan urrats bat aurrera ematen saiatu gara, eta bide berri bat jorratu nahi izan dugu aldaeren normalizazioa ebazteko, informazio morfologiko sakonagoa baliatuta. Kapituluan zehar informazio berri hori lortzeko jarraitu diren bideak azaltzen dira, eta baita informazio berri horrekin lortutako emaitzak ere, azterketa hori euskarazko corpus historiko bakar batean eginez.

6. Tesi-txostena amaitzeko, VI. kapituluan egindako lanetik ateratako zenbait ondorio ematen dira, eta etorkizunerako zabalik geratzen diren bideak azaltzen dira. Horrekin batera, tesi-lan honek egin dituen ekarpenak zerrendatzen dira.

7. Aurreko sei kapituluez gain, hiru eranskin gehitu zaizkio tesi-txosten honi, non, batez ere, erabilitako tresnen inguruko xehetasunak ageri diren. Hala, A eranskinean *Brat* anotazio-tresnari buruzko argibideak ematen dira, eta anotazio-prozesuan erabili diren etiketak deskribatzen dira. B eranskina *Phonetisaurus* tresna erabiltzeko gida laburra da, tresna hori berria izan baita guretzat, eta proposatzen dugun normalizazio-metodoaren oinarria da. Azkenik, C eranskinean, izen propioek sortutako arazoaren inguruko hausnarketa zein analisia egiten da.

## I.5 Argitalpenak

Tesi-lan honekin zuzenean lotutako argitalpenak dira:

- Hulden M., Alegria I., Etxeberria I., eta Maritxalar M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties,* 39–48 or. Edinburgo. Association for Computational Linguistics. (Hulden *et al.*, 2011)

- Etxeberria I., Alegria I., Hulden M., eta Uria L. (2014). Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural*, 52: 13–20 or. (Etxeberria *et al.*, 2014)

- Etxeberria I., Alegria I., Uria L. (2015). Induction of Phonology and Morphology for the Normalization of Historical Texts. *22nd International Conference on Historical Linguistics,* 107-109 or. Napoli.

- Etxeberria I., Alegria I., Uria L. eta Hulden M. (2016). Evaluating the Noisy Channel Model for the Normalization of Historical Texts: Basque, Spanish and Slovene. *International Conference on Language Resources and Evaluation, LREC 2016.* 1094–169 or. (Etxeberria *et al.*, 2016)

Tesi-lan honekin zeharka lotutako argitalpenak dira:

- Uria L., Hulden M., Etxeberria I. eta Alegria I. (2011). Recursos y métodos de sustitución léxica en las variantes dialectales en euskera. *Proceedings of the Workshop on Iberian Cross-Language NLP tasks, (ICL 2011)* 70–76 or. Huelva. (Uria eta Etxepare, 2011)

- Alegria I., Etxeberria I., eta Labaka G. (2013). Una cascada de transductores simples para normalizar tweets. *Proceedings of the Tweet Normalization Workshop co-located with 29th Conference of the Spanish Society for Natural Language Processing (SEPLN 2013),* 15–19 or. Madrid.

Bestelako argitalpenak:

- Etxeberria I., Alegria I. eta Leturia I. (2012). Ortografia-erroreak eta konpetentzia-erroreak Webeko euskarazko testuetan. *EKAIA Euskal*

*Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria*, 24: 219–236 or.

- Alegria I., Etxeberria I., Hulden M., eta Maritxalar M. (2009). Porting basque morphological grammars to foma, an open-source tool. *Finite-State Methods and Natural Language Processing,* 105–113. Springer.

# II. KAPITULUA

## Aldaera-corpusak

### II.1 Sarrera

Tesi-lan honen ikertze-lerro nagusia euskarazko testu ez-estandarrak normalizatzeko metodoen azterketan, ebaluazioan eta konbinazioan datza, eta horretarako ezinbestekoa da lehendabizi testu horiek hautatzea, garbitzea eta corpus aproposak prestatzea. Horrez gain, lan honetan aztertu nahi dugu ea euskarazko testuak normalizatzeko aukeratzen ditugun metodoak eta horien emaitzak, konparagarriak ote diren beste hizkuntza batzuetan erabilitako metodoek lortutakoekin; eta horretarako ere, ezinbestekoa da hizkuntza horietan erabili diren corpusak eskura izatea.

Euskarazko testu ez-estandarrei dagokienez, jatorri ezberdinetako testuak dituzten bi corpus erabili ditugu: alde batetik, zenbait testu dialektal jasotzen dituen corpus bat, eta, beste aldetik, zenbait testu historiko jasotzen dituen beste corpus bat. Beste hizkuntza batzuetako testuei dagokienez, berriz, hizkuntza horietako zenbait ikerlarirekin izan dugun kolaborazioa baliatu dugu haiek sortutako edota erabilitako datuak eskuratzeko eta gure esperimentuetan erabiltzeko.

Tesi-lan hau burutzeko erabili diren corpus horien guztien xehetasunak azaltzea da kapitulu honen helburua. Aurrera egin baino lehen, ordea, testu historikoen inguruko hausnarketa egin nahi dugu, izan ere, testu ez-estandarrekin egin diren lanetan behin eta berriro ageri da ingelesezko *historical texts* erreferentzia, hots, *testu historikoak*. Beharrezkoa jotzen dugu zehaztea zein motatako testuak sartzen diren multzo horren barruan, eta zehaztapen horren bila Piotrowski-ren liburura (2012) jo dugu.

Piotrowskik dioenez, testu historikoak dira hizkuntza historikoetan ida-

tzirik daudenak, eta jarraian argitzen du ez dagoela ikuspegi bakarra *hizkuntza historikoa* zer den definitzeko unean. Alde batetik, historikoa omen da "iraganari dagokiona eta egun erabiltzen ez dena", eta definizio horren arabera, latina eta antzinako grekoa hizkuntza historikoak lirateke. Baita antzinako ingelesa, antzinako frantsesa edo antzinako alemana ere, egungo hiztunek ez baitituzte ulertzen hizkuntza horietan idatzitako testuak. Baina, zer gertatzen da denboran gertuago dauden testuekin? Orain dela berrehun edo hirurehun urte idatzitako testuekin? Egungo hiztun asko gai dira testu horiek ulertzeko, baina askotan glosategi bat behar izaten dute hainbat termino ongi ulertzeko, horren arrazoia izanik jadanik erabiltzen ez diren terminoak direla (arkaikoak), edota esanahia aldatu zaien terminoak direla. Ikuspuntu horretatik, testu horiek ere *historikoak* kontsidera daitezke Piotrowskiren arabera, eta hori izan da, hain zuzen ere, kontuan hartu dugun ikuspegia euskal testu historikoak aukeratzeko unean.

Kapituluaren II.2 atalean liburutegi digitalei eta testu historikoei buruz arituko gara, eta arreta berezia eskainiko diogu euskararen egoerari arlo horretan. Gero, II.3 atalean lan honetan zehar sortu dugun euskarazko corpus historikoaren oinarriak eta ezaugarriak azalduko ditugu, eta corpus hori nola sortu dugun argituko dugu. II.4 atalean, lehendabiziko esperimentuak egiteko erabili dugun corpus dialektalaren xehetasunak emango ditugu eta bukatzeko, II.5 atalean, erabili ditugun beste bi hizkuntzatako corpusen hainbat zehaztasun emango ditugu.

## II.2 Testu historikoen erabilpena liburutegi digitaletan

Liburutegiek, artxiboek eta museoek dokumentu historiko asko jasotzen dituzte, hala nola liburuak, egunkariak, gutunak, aktak, laborategi-oharrak eta abar, eta ugariak dira herrialde askotan dauden proiektuak dokumentu horiek digitalizatzeko (Piotrowski, 2012). Digitalizazio-prozedurak dokumentu horien babesa eta iraupena bermatzen du eta, aldi berean, dokumentuen atzipena izugarri errazten denez, haien erabilpena hedatzen du, adituen eremutik publiko zabalaren eremura zabalduz.

Hizkuntza prozesatzeko tresnak aplikatu nahi badira testuetan, lehenik eta behin testu digitalizatua behar da, hau da, testua formatu elektronikoan behar da. Hori ez da inolako arazoa egungo testuak aztertu nahi badira, ugariak baitira testuak formatu elektronikoan eskuratzeko aukerak (Internetetik hasita). Testu historikoen testuingurua, aldiz, ezberdina da. Dokumentu historiko asko jadanik digitalizatuak izan dira hainbat museo, liburutegi eta antzeko instituziok bultzatutako proiektuei esker. Dokumentuak aukeratze-

ko irizpideak eta digitalizazioa aurrera eramateko metodoak, ordea, ez dira berberak izan kasu guztietan. Halako hainbat proiektu aipatzen ditu Piotrowskik bere liburuan (Piotrowski, 2012): Europeana[1], Wikisource[2], Google Books[3], HathiTrust[4], Text Creation Partnership[5] edota Project Gutenberg[6]. Proiektu horien helburuak ez dira berdinak, eta testuen ezaugarriak ere ez, beraz, ez dago prozedura orokor bat testuak digitalizatzeko, eta bietara egokitu beharra dago prozedura: testuen ezaugarrietara eta proiektuaren beharretara.

Dokumentu historikoak digitalizatzeko lehenengo urratsa, oro har, dokumentuen irudi digitalak lortzea da, hots, faksimileak, horretarako eskaner bat edo kamera digital bat erabilita. Irudiei metadatuak erantsiz gero (egilea, data, dokumentu mota, deskribapena eta abar) eta *online* atzitzeko aukera emanez gero, dokumentu horien atzipena inoiz baino azkarragoa eta errazagoa bilakatzen da, eta edozeinek, edozein tokitatik eta edozein unetan atzi ditzake dokumentu horiek.

Bigarren urratsa testua bera eskuratzea da, hau da, testua formatu digitalean lortzea. Horretarako, testua transkribatu behar da, eta bi aukera nagusitzen dira transkripzio-lan hori egiteko: OCR (*Optical Character Recognition*) teknikak erabiltzea edota eskuzko transkripzioa egitea. OCR teknikak inprimatutako obrei aplikatzen zaizkie baldin eta obra horien ezaugarriek hori ahalbidetzen badute: irudiaren kalitatea, letra mota, hizkuntza bera eta abar. Testuak eskuz idatzitakoak badira, berriz, eskuzko transkripzioa izaten da aukera bakarra, eta horretarako maiz erabiltzen den teknika ingelesez *double-keying* deitutakoa da: bi pertsonek (gutxienez) editatzen dute testua, eta gero transkripzioak konparatzen dira erroreak detektatzeko.

Hirugarren urratsa, erroreak detektatzeko eta zuzentzeko eman ohi den urratsa da, eta batzuetan beste zenbait eguneraketa egiteko erabiltzen da. Kontuan izan behar da testu historikoen edizioak kritikoak zein paleografikoak izan daitezkeela. Edizio kritikoetan, askotan, zenbait aldaketa egiten dira jatorrizko testuaren gainean, esaterako grafian. Edizio paleografikoetan, berriz, ahalik eta aldaketa gutxien egiten dira jatorrizko testuarekiko (nahiz eta baten bat egin). Edizio digitala inprimatutako edizio kritiko edo paleografiko batetik abia daiteke, baina urrats honetan egiten den prozesatzeak ere eragina izan dezake edizio digitala mota batekoa edo bestekoa izate-

[1] `www.europeana.eu` (2016-02-21ean atzitua)
[2] `wikisource.org` (2016-02-21ean atzitua)
[3] `books.google.com` (2016-02-21ean atzitua)
[4] `www.hathitrust.org` (2016-02-21ean atzitua)
[5] `www.textcreationpartnership.org` (2016-02-21ean atzitua)
[6] `www.gutenberg.org` (2016-02-21ean atzitua)

ko. Hala, bukaeran lortzen den edizio digitalaren fideltasun-maila jatorrizko bertsioarekiko, proiektuaren helburuen araberakoa da, eta ongi zehaztu beharreko kontua da. OCR teknikak eta OCR osteko prozesatzea ikergai oso interesgarriak dira (Reynaert, 2008) baina tesi-lan honetatik kanpo geratzen dira.

Azkenik, testua digitalizatua izanik, hizkuntza prozesatzeko egungo tresnak eta teknikak aplika daitezke testu horien gainean: testu-bilaketa, lematizazioa, analisi morfologikoa zein sintaktikoa, itzulpen automatikoa eta abar. Baina teoriak hori esaten badu ere, ongi dakigu testu historiko horien prozesaketa ez dela berehalakoa: egungo NLP tresnek, testu historikoetan adituak ez diren erabiltzaileek bezalaxe, arazoak dituzte testu horiek prozesatzeko. Arazo nagusienetarikoa ortografiaren ildotik dator, testu historikoek ez baitute egungo ortografia estandarra jarraitzen, eta NLP tresnek, berriz, ortografia estandar bat izan ohi dute oinarrian. Ortografiaren ildotik, hiru ezaugarri aipatzen dira Piotrowskiren liburuan (2012) testu historikoak lantzen direnean:

1. Hizkuntzen ortografia ez da estatikoa eta, noizbehinka bada ere, arau ortografikoak edo konbentzioak aldatu egiten dira: adibidez, alemanaren azken erreforma 1996koa da, eta nederlanderarena 1995ekoa. Logikoa da, beraz, aldaera diakronikoak izatea ortografia dela eta.

2. Ortografia estandar nagusiaren kontzeptua hizkuntzetan, nahiko berria da, XX. mende ingurukoa edo: alemanaren ortografia, esaterako, ez zen formalki erregulatu 1901era arte, eta portugesarena 1911ra arte. Hala, formalizazio-prozesuak garatu baino lehen hizkuntzek ez zuten ortografia "zuzenik", eta ohikoa zen idazteko joera ezberdinak izatea hizkuntza baten barruan (*written dialects* terminoa erabiltzen du Piotrowskik). Horrez gain, hitz beraren idazkera ezberdin asko erabil zitezkeen tokian tokiko ahoskera ezberdinak islatu nahian. Beraz, aldaera sinkronikoak ere ohikoak dira hainbat testu historikotan.

3. Testu historikoak ez ziren formatu digitalean sortu, noski, eta NLP tresnak aplikatzeko guztiz beharrezkoa da testuen transkripzioa. Beraz, tresnak ez zaizkie aplikatzen testu "originalei", baizik eta horietatik egin diren interpretazioei, eta horietan, akatsak egon daitezke. Testu historiko digitalizatuen ezaugarri honi *ziurgabetasuna* deritzo Piotrowskik (*uncertainty*).

Testuen normalizazioa modu automatikoan lortzerik balego, aukera legoke aipatutako aplikazioak testu historikoetan aplikatzeko. Adibide argienetari-

koa informazioa berreskuratzeko aplikazioetan (IR) aurkitzen dugu: onuragarria izango litzateke galdera egungo hitzak erabiliz egitea eta dokumentu historikoetan aurkitzea hitz horiei dagozkien hitz historikoak, halakoak baleude; horrela ez litzateke beharrezkoa izango testu historikoetan aditua izatea informazioa eskuratu ahal izateko.

Euskararen alorrera etorrita, euskaraz, beste hainbat hizkuntzatan bezala, gero eta gehiago dira Internet bitartez atzi daitezkeen euskal testu zaharren faksimileak, bai eskuz idatzitako testuenak bai inprimatutako testuenak. Liburutegi digitalek eta nazioarteko beste hainbat proiektuk asko erraztu dute dokumentu historikoen atzipena. Europeana proiektuaren bitartez, esaterako, erraz eta azkar atzi daiteke Bernard Etxepareren *Linguae Vasconum Primitiae* (1545) obraren lehen edizioko ale bakarra; orain dela gutxi arte, kontsulta hori egiteko Parisera joan beharra zegoen, Frantziako Liburutegi Nazionalera (Bilbao Telletxea eta Gómez López, 2014).

Proiektu internazionalez gain, bertako hainbat liburutegi digital ditugu, batzuk euskal testuetan espezializatuak, euskal testu zaharren faksimileak atzitzeko aukera eskaintzen dutenak. Horien artean daude, esaterako: Liburuklik[7], Euskal Memoria Digitala[8], Gipuzkoako Foru Aldundiko Repositorio Digitala[9], Nafarroako liburutegi digitala[10], Bizkaiko Foru Aldundiko liburutegi digitala[11] edota Euskaltzaindiako Azkue liburutegia[12].

Euskarazko testuen normalizazioaren atazan aritu gara ikertze-lan honetan, eta testu ez-estandarren iturri nagusienetarikoa testu historikoak dira. Baina normalizazioaren inguruan lanean hasi ginenean, 2011n gutxi gorabehera, ez genuen euskal corpus historiko digitalizatu apropos bat lanean hasteko, eta hala, hori prestatu bitartean, corpus paralelo dialektal batekin ekin genion gure proiektuari. Hain zuzen ere, testu dialektalek eta historikoek antzeko ezberdintasunak azaleratzen dituzte testu estandarrekin konparatuz gero, eta aukera izan genuen erraz eskuratzeko dialektala/estandarra corpus paralelo bat, Baionako IKER ikerketa-taldearen kolaborazioari esker. Beraz, corpus dialektal horrekin egin genituen gure lehenengo esperimentuak, eta bitartean, corpus historikoa nola osatu eta prestatu aztertzen hasi ginen, esperimentuetarako beharrezko informazioa zein zen

[7] `www.liburuklik.euskadi.net` (2016-02-21ean atzitua)
[8] `www.memoriadigitalvasca.es` (2016-02-21ean atzitua)
[9] `meta.gipuzkoakultura.net` (2016-02-21ean atzitua)
[10] `administracionelectronica.navarra.es/binadi/busqueda.aspx?idioma=eu-ES` (2016-02-21ean atzitua)
[11] `www.bizkaia.eus/home2/Temas/DetalleTema.asp?Tem_codigo=2542` (2016-02-21ean atzitua)
[12] `www.euskatzaindia.eus` (2016-02-21ean atzitua)

ongi zehaztu eta gero.

Kapitulu honen hurrengo bi ataletan euskarazko bi corpus horien ezaugarriak azalduko ditugu: lehendabizi corpus historikoaren xehetasunak emango ditugu, eta gero corpus dialektalekoak.

## II.3 Euskarazko corpus historikoa

Tesi-lan honetan euskal testu historikoen corpus bat osatu dugu esperimentuak aurrera eraman ahal izateko, eta osatze-lan horretan egin beharreko lehenengo urratsa testuen aukeraketa eta prestaketa izan da.

Euskal testu historiko gisa literaturako klasikoak hartu ditugu gogoan, eta corpusa prestatzeko testuak aukeratu baino lehen, bibliografian eta Interneten aritu gara bila eskura zer dagoen jakiteko. Gure asmoa da jadanik digitalizatuak dauden euskal literaturako klasiko batzuk aukeratzea, tesi-lan honetatik at geratzen baita faksimileei OCR teknikak aplikatzen hastea testua eskuratzeko. Prozesu hori luzea izan daiteke eta, gainera, hizkuntzalari adituen lana eskatzen du.

Hainbat obratan oinarrizko analisia egin eta gero (II.3.3 atalean egiten da) ideia da euskal literaturako obra bat aukeratzea, horrekin corpus paralelo txiki bat prestatzeko esperimentuetarako. Esperimentu horien xedea da euskarazko testu historikoen normalizazioaren inguruko emaitzak zein ondorioak ateratzea. Gero, metodoa eta ondorioak berresteko, bigarren obra bat aukeratu nahi da, horretan normalizazioko esperimentuak errepikatzeko, horrela aztertu ahal izateko emaitzak parekoak ote diren.

### II.3.1 Euskal literaturako klasiko digitalizatuak

Euskal literaturako klasikoak lau mendetan zehar banatzen badira ere, Bilbao Telletxea eta Gómez López (2014) lanean aipatzen denez, XVIII. mendearen erdira arte argitaratutako obra kopurua txikia da benetan. Berrehun urtetan, 1545–1749 bitartean, 48 orri baino gehiagoko 40 obra berri publikatu ziren; 10 besterik ez epe horretako lehen ehun urtetan. Gero, XVIII. mendetik aurrera, progresioa askoz nabarmenagoa da. Aipatutako lanean euskal testu zaharren biltegien on-gaitzak aztertzen dira eta bost proiekturen berri ematen da. Bost proiektu horietatik lauk webgunea daukate eta bertan dauden testuak aztertu ditugu gure lanerako behar ditugunak aukeratzeko.

1. **Euskal Testuen Gordailua**[13]. Euskal testu zaharrak eskaintzen dituen ataria da eta UPV/EHUko webgunean kokatuta dago. Atariaren arduraduna Patxi Salaberri Muñoa da, UPV/EHUko irakaslea, eta atariaren helburua euskal testuen bertsio elektronikoak biltzea da. Atarian bertan ageri denez, 2000ko abenduaren 29az geroztik ez da edukirik gaurkotu.

   Gordailu horren "Idazleak eta idazlanak" atala izan da gehienbat aztertu duguna. Bertan 53 obra daude eta horietako bakoitzaren testua eskuratzeko aukera dago. Ez dago digitalizazioari buruzko oharrik, eta eransten den informazio bakarra, obra bakoitzaren bukaeran, bertsio digitalizatuaren egileari buruzkoa da gehienetan (kasuren batean, egilearen ordez digitalizaziorako erabili den edizio inprimatua zehazten da). Aipatzekoa da bertan ageri diren testu klasiko askoren bertsio digitalizatuaren egilea Josu Lavin dela, euskarazko testuen digitalizazioaren aitzindaria, hain zuzen ere.

2. **Klasikoen Gordailua**[14]. Gordailu hau *Armiarma*[15] euskal literaturari dagokion atariaren barruan kokatzen da eta bertan ageri diren obra guztiak publikoak dira. Atariaren aurkezpenean ageri diren bi paragrafo ekarri ditugu hona ("KG-ri buruz" atalean daude):

   *Euskal testuak dira gordailuan jaso ditugunak, ohikoan «klasiko» izendapenarekin ezagutzen ditugunak. Literaturaren alorreko testuak dira gehienak, baina ez bakarrak. Euskal literaturako lehen agerpenetatik 1936ko gerrak ezartzen duen mugara arte ekoitzi diren euskarazko literatur testuak biltzeko asmoa du guneak.* [...]

   *Klasikoen Gordailuan bi ahaleginen batuketa dago. Euskal Testuen Gordailua, batetik, eta Susa literatur argitaletxea, bestetik. Patxi Salaberri Muñoa idazle eta irakasleak egin eta zuzendu duen eta azken urteetan geldirik izan den Euskal Testuen Gordailua dago honen oinarrian. Susa literatur argitaletxeak Armiarma literatur atariaren oinarrizko zutabe ikusten zuen euskal literaturako testu klasikoak sarean jartzea, eta asmo bien batzetik sortu da KLASIKOEN GORDAILUA. Bien elkarlanetik sortu da, eta hala jarraituko.*

   Beraz, webgune horretan aipatutako *Euskal Testuen Gordailua* atariko obra guztiak eta askoz gehiago daude. Gordailuak dituen atalen artean, "Idazleak eta idazlanak" atala da gehien interesatu zaiguna.

[13] `www.vc.ehu.es/gordailua` (2016-02-11n atzitua)
[14] `klasikoak.armiarma.eus` (2016-02-12an atzitua)
[15] `armiarma.eus`

Bertan obra asko daude, 496 obra Bilbao Telletxea eta Gómez Lópezen artikuluan (2014) diotenez, eta aukera dago testuak zuzenean atarian kontsultatzeko edota norberaren diskoan gordetzeko. Aipatutako artikuluan diote, ordea, webgune horretan eskaintzen diren testuen grafia modernizatu egin dela, eta zenbait kasutan testuek errore asko dituztela:

*(1) El texto en sí, que, a diferencia de Euskal Testuen Gordailua, se presenta con la grafía modernizada; el problema es que los textos contienen muchos errores, ya que en ocasiones la modernización parece haberse realizado de forma (semi)automática.*

Halere, atarian ez dugu informaziorik aurkitu modernizazio-prozesu horri buruz.

3. **Andres de Poza ataria**[16]. Atari hau Carmen Isasik zuzendutako *Seminario Alfonso Irigoien* ikerketa taldearen proiektua da, eta atariaren aurkezpenean diotenez *"bertsio ugari dituzten testuen argitalpenean espezializatutako webgune bat izan nahi du"*. Hala, testuen edizio ezberdinak edota itzulpen ezberdinak lerrokatuta ikusteko aukera eskaintzen du atariak.

   Ez da, beraz, euskal testuak bereziki biltzen dituen ataria, baina badira bertan euskarara itzulitako lau testu zahar:

   - *Platica lelengoá*, Rafael Mikoleta (1653).
   - *San Frances Sales Genevaco Ipizpicuaren Philotea eta Chapeletaren Andre Dana Mariaren ohoretan devocionearequin erraiteco antcea*, Silvain Pouvreau (1664).
   - *Gudu espirituala*, Silvain Pouvreau (1665).
   - *Philotea edo devocioneraco bide erakusçailea*, Joannes Haraneder (1749).

   Euskaraz dauden testu horiek espresuki editatuak izan dira taldearen proiekturako, eta Josu Bijuesca eta Ana Toledo filologoak izan dira edizioaren arduradunak. Interneten argitaratu dituzte edizioak "Creative Commons BY-NC-ND" lizentziapean eta obra bakoitzaren dokumentazioan jarraitutako edizio-irizpideak azaltzen dituzte. Irizpideetan argi uzten dute zein izan diren jatorrizko testuarekiko egin diren aldaketak edota moldaketak, eta nola adierazi diren zalantzak

[16] `andresdepoza.com` (2016-02-12an atzitua)

eta antzeko kontuak testuan zehar (editoreen oharrak, hondatutako testuak eta abar).

4. **Lazarragaren eskuizkribua: edizioa eta azterketa**[17]. Webgune horretan *Monumenta Linguae Vasconum* ikerketa taldearen proiektuaren berri aurkitzen dugu, egile bakar baten obra bat editatzeko eta aztertzeko: *Lazarragaren eskuizkribua* obra. Eskuizkribua 2004an aurkitu zen eta haren gaineko ikerketa-lana oraindik martxan dago.

   Hainbat testuz osatutako bilduma da eta datazioa ziurra ez bada ere, badirudi Lazarragak gutxi gorabehera 1567tik 1602ra bitartean idatzitako lanak kopiatu zituela eskuizkribu horretan. Guztira 51 orrik osatzen dute eskuizkribua, gehienak euskaraz idatziak (% 88 inguru), eta gainerakoak gaztelaniaz idatziak. Eskuizkribua ez da osorik aurkitu eta, gutxienez, beste hamasei orri galdu dira ezagutzen den zatian. Edukiari dagokionez, testu nagusian bi atal bereizten dira: lehenik, artzain-liburu bat (testu nagusiaren % 32) gehienbat prosan idatzia, tartekatuak dituen arren kantu moduko poema batzuk; bigarrenik, askotariko poema eta kantuen bilduma.

   Eskuizkribuaren testua oraindik lantzen ari dira ikerlariak, eta hala argitzen dute webgunean bertan:

   *Dena den, zatikako beste azterketa batzuk ere argitaratu diren arren, oraindik finkatu gabe dago testua bera, eta edizio fidagarri, arduratsu eta zintzoa egin behar da, pasarte zail ugariak argitzen ahaleginduko dena, zailtasun asko baititu testuak hala paleografiari dagokionean nola hizkuntzari dagokionean; ahalegin horretan kokatzen da Monumenta Linguae Vasconum ikerketa-taldearen lana.*

### II.3.2 Hainbat obraren aukeraketa

Aurreko atalean aipatutako atarietatik zenbait obra eskuratu eta aztertu ditugu lehenik, horien artean aukeratzeko eta prestatzeko gure esperimentuetan erabiliko ditugun corpusak. Azkenean, *Klasikoen Gordailua* webgunetik ez dugu obrarik hartu, Bilbao Telletxea eta Gómez López (2014) artikuluan esaten baita zenbait testuk errore asko dituztela (agian grafia-eguneraketagatik). Bestalde, Lazarragaren eskuizkribuari dagokion testua ere ez dugu jaso: prosan idatzitako atala nahiko motza da, eta, gainera, testua bera oraindik lantzen ari direla diote webgunean. Beraz, beste bi webguneetatik hartu ditugu zenbait obra.

[17] `lazarraga.com` (2016-02-12an atzitua)

**Andres de Poza** atarian dauden lau obretatik, hiru eskuratu ditugu: Silvain Pouvreauren bi lanak, *San Frances Sales Genevaco Ipizpicuaren Philotea eta Chapeletaren Andre Dana Mariaren ohoretan devocionearequin erraiteco antcea* (1664) eta *Gudu espirituala* (1665), eta Joannes Haranederren *Philotea edo devocioneraco bide erakusçailea* (1749). Obrak aukeratzeko irizpideetako bat tamaina izan da, eta horregatik ez dugu hartu atari horretako laugarren obra, Rafael Mikoletarena, besteak baino dezente motzagoa baita.

**Euskal Testuen Gordailua** atarian obra asko daude eta bakar batzuk aukeratu ditugu, kontuan izanik haien tamaina, garrantzia, gaia, garaia, euskalkia eta bertsio digitalizatuaren egilea. Azken ezaugarri horri garrantzi handia eman diogu. Digitalizazioaren ezaugarriak eta irizpideak ez dira webgunean ageri, eta ez dakigu bertan ageri diren bertsioak eskuzko transkripzioak diren edo beste teknika batzuen bitartez lortu diren, beraz, iruditu zaigu egile bakar baten digitalizazioak aukeratuz gero, irizpide horiek antzekoak izan daitezkeela obra guztietan. Hala, atari honetatik eskuratutako obra guztiak Josu Lavinek digitalizatuak dira: *Linguae Vasconum Primitiae* (Bernard Etxepare, 1545); *Gero* (Pedro Agerre *Axular*, 1643); *Laborantzako liburua* (Jean Pierre Duvoisin, 1858); *Bi saindu Hescualdunen Bizia* (Franzisko Laphitz, 1867); *Peru Abarka* (J. A. Mogel, 1881) eta *Kresala* (Txomin Agirre, 1908).

### II.3.3 Lehenengo azterketa

Bederatzi obra eskuratu ditugu Internetetik eta horietan egin dugun lehenengo analisia izan da kontatzea zenbat hitz ez diren estandarrak, hau da, zenbat hitz ez datozen bat egungo hitz estandarrekin. Halako hitzei OOV (*Out of Vocabulary*) deitu ohi zaie.

Lortutako kopuruak ahalik eta zehatzenak izatearren, azterketa egin baino lehen beharrezkoa izan da testuak pixka bat *garbitzea*. Hainbat obraren hasieran edota bukaeran ageri dira obratik at kontsidera ditzakegun atalak: argitalpenari buruzko informazioa, hitzaurrea, indizea eta abar. Horiek markatzeko eskuz sartu ditugu XML etiketak : `<front>` eta `<back>`, gero erraz kentzeko obraren atal horiek programa sinple baten bitartez.

Horren ondoren, obrako testua "tokenizatu" dugu, hau da, hitzetan banatu dugu testua eta hitz horiei hasierako zein bukaerako puntuazio-ikurrak kendu dizkiegu. Tokenizazioa egin ondoren, hainbat token alboratu dira analisian zarata gutxitzeko: letrarik ez duten tokenak (zifrak), letra larriz soilik osatutako tokenak eta oraindik karaktere arraroren bat duten tokenak (*ver-*

| **Egilea** | **Mendea** | **Euskalkia** | **Tokenak** | **Alboratuak** |
|---|---|---|---|---|
| Etxepare | XVI | Behe Nafarrera | 6.613 | 9 |
| Axular | XVII | Lapurtera | 98.708 | 1.399 |
| Pouvreau (1) | XVII | Lapurtera | 72.630 | 1.159 |
| Pouvreau (2) | XVII | Lapurtera | 35.761 | 252 |
| Haraneder | XVIII | Lapurtera | 74.673 | 1.100 |
| Duvoisin | XIX | Lapurtera | 26.297 | 531 |
| Laphitz | XIX | Nafar Lapurtera | 32.469 | 33 |
| Mogel | XIX | Bizkaiera | 23.987 | 907 |
| Agirre | XX | Bizkaiera | 34.004 | 121 |

II.1 Taula: Analizatutako obren token kopuruak eta baztertutakoen kopuruak. Baztertu dira analisirako: letrak ez dituzten tokenak, letra larriz soilik osatuak daudenak eta zerbait arraroa dutenak (ez dira letraz soilik osatutakoak). Obrak identifikatzeko egileen izenak erabili dira taulan. Pouvreau idazlearen kasuan bi obra aztertu direnez, zenbakiz markatu dira: (1) zenbakia *Philotea* lanari dagokio eta (2) zenbakia, berriz, *Gudu Espirituala* lanari.

*tutean?Gaineracoan*, *gurutcefic[_atua]*...). Datu horien guztien laburpena II.1 taulan bildu dugu.

Hurrengo urratsa obra bakoitzean ageri diren OOV motako token kopurua (hitz bakoitza ageri den bezain beste aldiz) zein OOV forma kopurua (hitz bakoitza, soilik behin) aztertzea izan da[18]. Azterketa hori egiteko, Xuxen zuzentzaile ortografikoaren egokitzapen bat erabili dugu (Alegria *et al.*, 2009a), eta II.2 taulan jaso ditugu kopuruak. Arreta jartzen bada portzentajeetan, ikusten da obra zaharrenak (XVI., XVII., XVIII. mendekoak) direla, oro har, OOV portzentaje handienak dituztenak: % 60 edo % 80 inguru, tokenak edo formak kontatzearen arabera. XIX. eta XX. mendeko lanetan nabarmen txikiagoak dira portzentaje horiek, 20 puntu inguru txikiagoak, hain zuzen ere. Dena den, badira bietan salbuespenak. Alde batetik, Axularren *Gero* obran, XVII. mendekoa, portzentajeak askoz txikiagoak dira denboran

[18]Corpusen tamainari buruzko datuak ematen direnean, ohikoa izaten da ingelesezko *token* eta *type* terminoak erabiltzea. *Token* terminoak adierazten du corpusaren hitzak ageri diren moduan hartzen direla kontuan, berdin dio zenbat aldiz errepikatzen diren. *Type* terminoak, berriz, adierazten du hitz bakoitza behin bakarrik hartzen dela kontuan eta ez ageri den bezainbeste aldiz. Zilegi bekigu txosten honetan zehar *token* terminoa erabiltzea oraintxe azaldu dugun adierarekin, eta *type* terminoaren adierarako *forma* terminoa erabiltzea.

| **Egilea** | **Mendea** | **Tokenak** | **OOV tokenak** | **Formak** | **OOV formak** |
|---|---|---|---|---|---|
| Etxepare | XVI | 6.604 | 4.412 | 2.675 | 2.185 |
| | | | % 66 | | % 81 |
| Axular | XVII | 97.309 | 27.567 | 16.955 | 8.832 |
| | | | % 28 | | % 52 |
| Pouvreau (1) | XVII | 71.471 | 41.256 | 13.954 | 11.308 |
| | | | % 57 | | % 81 |
| Pouvreau (2) | XVII | 35.509 | 20.901 | 7.528 | 5.933 |
| | | | % 58 | | % 78 |
| Haraneder | XVIII | 73.573 | 45.866 | 20.222 | 17.514 |
| | | | % 62 | | % 86 |
| Duvoisin | XIX | 25.766 | 6.420 | 6.461 | 2.423 |
| | | | % 24 | | % 37 |
| Laphitz | XIX | 32.436 | 12.502 | 7.154 | 4.246 |
| | | | % 38 | | % 59 |
| Mogel | XIX | 23.080 | 15.989 | 7.428 | 6.123 |
| | | | % 69 | | % 82 |
| Agirre | XX | 33.883 | 13.371 | 8.988 | 4.335 |
| | | | % 39 | | % 48 |

II.2 Taula: Analizatutako obren OOV kopuruak, bai token gisa kontatuta, bai forma gisa kontatuta.

gertuen dauden beste obrekin konparatuta; eta beste aldetik, Mogelen *Peru Abarka* obran portzentajeak askoz handiagoak dira berarengandik gertu daudenekin konparatuta.

Hein batean, salbuespen horiek jatorrizko bertsio digitalizatuaren edizio motarekin lotzen dira (kritikoa vs. paleografikoa), baina seguru asko, hori ez da arrazoi bakarra. Obran erabiltzen den euskalkia ere arrazoi garrantzitsua izan daiteke OOV kopurua handiagoa edo txikiagoa izateko.

*Andres de Poza* ataritik eskuratu ditugun hiru obretan, OOV tokenen portzentajeak altuak dira eta obra horietako testua irakurtzen hasita, berehela igartzen da grafia ez dagoela eguneratuta. Webgunean bertan honako esaldi hau irakur daiteke *Philotea* lanaren edizio-irizpideetan:

*Joera nagusia jatorrizko testua bere horretan uztea izan da, bai grafiari dagokionez, bai eta puntuazioaz den bezain batean ere. Puntuazioan ohartu gabeko zeinu gutxi batzuk (punturen bat eta zenbait koma) txertatu dira*

*zentzuak halakorik eskatzen zuela argi ikusi denean.*

EHUko gordailutik eskuratutako obretan, berriz, denetarik aurkitzen dugu. Guztiak dira Josu Lavinek digitalizatuak baina grafia aldetik dauden ezberdintasunak oso nabariak dira. Axularren obran, *Gero*n, OOV tokenen portzentajea oso baxua da, baina kasu horretan argitu ahal izan dugu Josu Lavinek Villasantek egindako argitalpena erabili zuela iturri gisa. Argitalpen horri buruz, honako informazio hau aurkitu dugu EHUko Euskara Institutuaren atarian[19]:

*Hurrengo argitalpena, bosgarrena, 1964koa da, Donostian Izarra irarkolan inprimatua eta Bartzelonako Juan Flors argitaratzailearen etxeak plazaratua, Villasantek paraturik; lehen argitalpenaren araberakoa bada ere, "ortografia gaurkotuaz eta hatxe letra errespetatuaz" eginiko argitalpena dela adierazten du Villasantek (1972: 127), hain zuzen ere, euskara batua Arantzazuko 1968ko biltzarrean abian jarri aurretik ortografia estandarra izan zitekeenaren alde eginiko saiakera garrantzitsuetako bat.*[20]

Beraz, Axularren obran badakigu nondik datorren grafiaren eguneraketa. Beste obretan ez dugu zehazterik izan zein izan den abiapuntuko edizioa digitalizazioa egiteko, baina gure ustez, horrek izan behar du arrazoietako bat portzentajeak handiagoak edo txikiagoak izateko.

Analisiarekin bukatzeko, ikusi dugunez OOV portzentaje altuak dituzten obretako grafia asko aldatzen dela egun estandarra kontsideratzen den grafiarekiko, azkeneko proba bat egitea erabaki dugu obra horien artean bat aukeratu baino lehen. Obra horietan *c*, *ç*, *q* eta *v* letrak erruz ageri direnez, transliterazio-erregela batzuk aplikatu dizkiegu obra guztiei, aurreko portzentajeak zenbateraino aldatzen diren ikusteko. Badakigu erregela horiek automatikoki aplikatzean ez dela beti asmatuko egiten diren aldaketetan, baina esperimentu honen helburua da jakitea zenbateko garrantzia duen grafiak OOV portzentaje handi horietan.

Honako hauek izan dira aplikatutako erregelak (letra larriekin zein xeheekin) eta aplikatutako ordenan ekarri ditugu hona:

- Azentu-markaren bat daukaten bokalei azentua kendu zaie, esaterako: `á,à,ä,â` → `a`

- *ç* karakterearen atzetik bokalen bat ageri bada, *ç* hori *z* bihurtu da, esaterako: `ça` → `za`

[19] `www.ehu.eus/ehg/literatura/?p=480` (2016-02-14an atzitua)

[20] Zitan ematen den erreferentzia Euskaltzaindiako atarian kontsulta daiteke: `www.euskaltzaindia.net/dok/iker_jagon_tegiak/villasante/dokumentuak/332.pdf` (2016-02-14an atzitua)

| **Egilea** | **Mendea** | **Tokenak** | **OOV tokenak** | **Formak** | **OOV formak** |
|---|---|---|---|---|---|
| Etxepare | XVI | 6.604 | 3.296 | 2.592 | 1.691 |
| | | | % 49 (-17) | | % 65 |
| Axular | XVII | 97.309 | 27.338 | 16.880 | 8.690 |
| | | | % 28 (=) | | % 51 |
| Pouvreau (1) | XVII | 71.471 | 24.123 | 13.780 | 7.242 |
| | | | % 33(-24) | | % 52 |
| Pouvreau (2) | XVII | 35.509 | 12.083 | 7.472 | 3.778 |
| | | | % 34 (-24) | | % 50 |
| Haraneder | XVIII | 73.873 | 30.514 | 19.280 | 13.034 |
| | | | % 41 (-21) | | % 67 |
| Duvoisin | XIX | 25.766 | 6.418 | 6.459 | 2.421 |
| | | | % 24 (=) | | % 37 |
| Laphitz | XIX | 32.436 | 8.669 | 7.051 | 2.780 |
| | | | % 26 (-12) | | % 39 |
| Mogel | XIX | 23.080 | 12.973 | 7.347 | 5.066 |
| | | | % 56 (-13) | | % 68 |
| Agirre | XX | 33.883 | 13.362 | 8.983 | 4.327 |
| | | | % 39 (=) | | % 48 |

II.3 Taula: Analizatutako obren OOV kopuruak transliterazio-prozesua eta gero. Obra zaharrenetan, XVIII. mende artekoetan, OOV hitz kopurua 20 puntu inguru jaitsi da.

- *c* karakterearen atzetik *a*, *o* edo *u* bokalak ageri badira, *c* hori *k* bihurtu da, esaterako: `ca` → `ka`
- *c* karakterearen atzetik *e* edo *i* bokalak ageri badira, *c* hori *z* bihurtu da, esaterako: `ce` → `ze`
- hitz baten azken karakterea *c* bada, *k* bihurtu da
- *v* karakterea, *b* bihurtu da: `v` → `b`
- *qu* karaktereak, *k* bihurtu dira: `qu` → `k`

Erregela horiek aplikatu ondoren lortutako kopuru berriak II.3 taulan ageri dira. Aurreko taulakoekin konparatzen badira, ondorioa garbia da: transliterazio-prozesu sinple horrek 20 puntu inguruko jaitsiera lortu du

OOV asko dituzten obretan. Berriro ere, Mogelen *Peru Abarka* salbuespena izan da eta obra horretan jaitsiera % 13an geratu da.

Bestalde, espero zitekeen moduan, obra batzuetan transliterazio-prozesu horrek ez du inongo jaitsierarik lortu. Axularren, Duvoisinen eta Agirreren obretan, esaterako, portzentajeak bere horretan geratu dira, dagoeneko lan hori eginda zegoela adierazten duen seinale argia.

Azterketa horiek guztiak egin eta gero, analizatutako klasiko horien artean oso ezagunak diren bi obra aukeratu ditugu azkenean esperimentuetan erabiltzeko: Axularren *Gero* eta Mogelen *Peru Abarka*. *Gero* obra aukeratzeko arrazoi nagusiak honako hauek izan dira: ospe handiko obra klasikoa da, lapurteraz idatzia (euskalki nagusiena idatzizko obra zaharrenetan), eta, tamainari dagokionez, daukagun handiena da; gainera, badirudi grafiari dagokion normalizazioa egina dagoela zati handi batean, eta interesgarria izan daiteke jakitea zer nolako emaitzak lor daitezkeen oraindik estandarra ez dena automatikoki normalizatuz gero. *Peru Abarka* obra aukeratzeko arrazoi nagusia izan da zerbait osagarria aukeratzea, hau da, ahalik eta diferenteena: euskalkia aldatzea (bizkaieraz idatzita dago), gaia aldatzea (ez dago erlijioari hain lotua) eta ahalik eta eguneraketa gutxien dituen edizio batetik abiatzea (estandarretik urrutiago egotea). Bi obrak osagarriak dira, nolabait, eta bi corpus prestatu ditugu ondorengo ataletan deskribatzen den moduan.

### II.3.4 Corpusaren prestaketa: *Gero*

*Gero* obraren testu osoa formatu elektronikoan daukagu, eta testu horri eskuz jarri zaizkio etiketa pare bat: hasieran zein bukaeran `<front>` eta `<back>` etiketak daude, erraz kendu ahal izateko analisitik kanpo utzi nahi diren zenbait atal. Atal horien artean daude hasierako "Gomendiozko karta", "Approbationes" eta "Irakurtzailleari" atalak, eta bukaerako kapituluen zerrenda.

Aipatu diren atal horiek kendu ondoren geratzen den testuan, garbiketa batzuk egin dira corpusa prestatu baino lehen. Gogora ditzagun obraren lehen analisia egin denean lortutako emaitzak: 97.309 token analizatu dira obra horretan eta % 28 ez-estandarrak dira. Forma kopurua kontuan hartuz gero, 16.880 forma daude eta horietatik % 51 ez-estandarrak dira.

Obra osoaren corpusa prestatzeak luze joko lukeenez, obraren zati batzuk aukeratu dira, zoriz, bertan dauden hitz ez-estandarrak detektatzeko eta markatzeko. Markatutako horiek anotatzaile batek etiketatu ditu gero, izan ere, OOV guztiak ez dira zertan mota berekoak izan: batzuk aldaera diakronikoak izango dira (horiek dira normalizatu nahi ditugunak), baina

beste batzuk ez. Dena den, etiketatzeko testua prestatzen hasi baino lehen beste garbiketa hauek egin dira obra honetan:

- Lehenik eta behin testuan ageri diren hainbat lerro "berezi" kentzea erabaki dugu: letra xeherik ez duten lerroak. Halako lerroak ageri dira testuaren antolaketa dela eta (kapituluak eta horien barruan atalak): letrarik gabeko lerroak dira. Horietaz gain, kapitulu bakoitzaren hasierako lerroak letra larriz soilik osatuak dira (kapituluaren izenburua edo gaia adierazten dute), eta horiek ere kendu egin dira. Lerro berezi horietan guztietan komuna dena da ez dutela letra xeherik, eta ezaugarri hori erabili da programa baten bitartez horiek iragazteko.

- Testuan zehar parentesien arteko erreferentzia asko daude: (Gen. 2), (Casian. lib. 18 cap. 14), (Exod. 5) eta abar. II.1 irudian halako adibide batzuk ageri dira obrako paragrafo batean. Ez bada ezer egiten, halako termino asko ez-estandar gisa detektatuko dira testua analizatzean, eta hori ez zaigu interesatzen. Beraz, programa sinple bat idatzi dugu parentesi arteko testua automatikoki ordezteko `(...)` markarekin. Programak ordeztutako zatiak gainbegiratuta, ikusi da badagoela parentesien arteko kasuren bat ez dena adibideen parekoa eta testu arrunta duena, baina oso gutxi direnez, aurrera egin da.

  Nahiz eta parentesiekin zerikusirik ez izan, aipatu berri den programak, parentesiak kentzeaz gain, beste akats txiki bat ere konpontzen du automatikoki. Jatorrizko testuan, batzuetan, errore tipografiko sinple hau gertatzen da: zuriunerik ez izatea puntuazio-ikur baten ondoren, bi hitz lotuz. Horren adibideak dira *dela,plazentziaz*, *iartzea,eta* edota *Hirurgarrena:Eleemosyna*. Zuriunerik ez izatean, forma horietako bakoitza hitz baten moduan tratatzen da, horrek dituen ondorioekin. Konponketa erraza denez, programa arduratzen da zuriune horiek sartzeaz: puntuazio-ikur bat topatzen bada 'hitz' baten karaktereen artean, horren aurreko karakterea letra xehea bada, eta atzeko karakterea edozein letra bada, zuriune bat sartzen du puntuazio-ikurraren ondoren, hitza bitan banatuz.

Aurreko "iragazkiak" aplikatu ondoren, obraren % 10 eta % 5 inguru jasotzen duten bi zati aukeratu dira zoriz (ikasteko eta testerako erabiliko dira, hurrenez hurren), eta zatiketa hori egiteko paragrafoa erabili da unitate gisa. Obra osoak 1.175 paragrafo ditu, eta horien artean zoriz aukeratu dira 118 eta 61 paragrafoz osatutako bi zati.

Hitz ez-estandarrak detektatzeko analisiarekin hasi baino lehen, ikusi da obra horretan latinez idatzitako zita ugari dagoela (II.1 irudian adibide bat

**15** Erraiten du Aristotelek, on dela, alferkeriaren herritik khentzeko, eta desterratzeko: eta herrien ere bere erregeren edo bertzeren kontra iaikitzetik begiratzeko, zenbait obra handiren hastea, zenbait dorreren edo gazteluren egitea, eta hetan iendearen enplegatzea (Arist. lib. 5 Politic. cap. 11). Nola ageri baita Ejiptoko Piramidetan, zein eragin baitzituen errege Faraonek, iendeak alfer etzeudezin amoreakgatik. Iduriturik errege hari, ezen baldin Israeleko seme gathibu bezala bere azpian zedutzan hek (iragaiten baitziren seietan ehun milla presunatan) utzten bazituen bere plazerera eta aisiara bizitzera, urguillutzeko eta nabusitzeko bidean iarriko zirela, eta handik behar etzena sorthuko zela, egin zuen Piramide batzuen egiteko gogoeta, asmua eta pensua. Eta Piramide hetzaz mintzo dela, erraiten du San Isidorok: *Pyramides est genus sepulchrorum quadratum fastigiatum ultra omnem celsitudinem, ut a lato incipiant et in augusto finiantur* (Isidor. lib. 5 Ethimol. cap. 11). Piramidea edo Piramideak ziren sepultura suerte batzuk, pilare, harroin, edo thonba laur kantoinetako gora ailtxatu batzuk, egin ahal zitezkeien gorenak, ondoan zabal eta puntan mehar. Eta hetan trabailla arazitzen zituen Faraonek bere azpiko iende hek, seiñalaturik bat bederari, bere eguneko lana eta sailla. Eta eskuaren ibentzea bera asko bazuketen ere, ordea lanhabesak, tresnak, eta obraren egiteko gai guztiak ere, berèk bilhatu eta hornitu behar zituzten. Eta halarik ere, ezin ausart zitezkeien arrenkuratzera: halako moldez ezen hartako lehenbiziko hitza ahotik itzuri zeienean, erran baitzerauen berehala Faraonek: *Vacatis otio* (Exod. 5). Asti duzue, zeuen ongiegiak, alferkeriak, aisetasunak iratxekitzen deratzue, hark horrela kilikatzen eta mintza arazitzen zaituztezate. Eta halatan aitzinerat kargatuago zituen, lana berretu zerauen. Eta hura guztia egiten zuen, baldin bat ere astirik bazuten, edo alfer bazeuden, handik zerbait ethor zekion gogan beharrez eta beldurrez.

II.1 Irudia: *Gero* obrako paragrafo bat agerian uzteko obran zehar ageri diren parentesi arteko erreferentziak (berdez azpimarratuta) eta latinezko zitak (gorriz azpimarratuta).

ikus daiteke gorriz azpimarratua). Ezer ez bada egiten, latinezko termino asko euskara ez-estandar gisa sailkatuko dira, eta horiek ez dira interesgarriak gure lanerako. Beraz, anotazio-prozesua ahalik eta arinena izan dadin, eta anotazioaren informazioa ahalik eta zehatzena gure beharretarako, saio bat egin da latinezko zitak automatikoki detektatzeko testuan eta beste marka batez ordezteko. Horretarako, *textcat*[21] izeneko aplikazioa erabili da (van Noord, 2001).

Latinezko zitak testuan zehar tartekatuak daudenez, horiek detektatu ahal izateko beharrezkoa izan da paragrafoak esalditan banatzea[22], ondoren esaldi bakoitzaren hizkuntza detektatzeko textcat aplikazioaren bitartez[23]. Latin gisa identifikatu dituen esaldiak fitxategi batean gorde dira ziurtatu ahal izateko identifikazioa txukuna izan dela, eta jatorrizko testuan marka berezi bat gehitu da halako zita bat kendu den puntuetan (ikus A eranskina xehetasun gehiagorako).

Aurreko prozedurak ezin du ziurtatu ez dela latinezko zitarik geratuko testuan, baina helburua da ahalik eta gehienak kentzea, etiketatzeko orduan ahalik eta zarata gutxien egon dadin.

Garbiketaren urratsa beteta, bi zatietako testua prestatu beharra dago anotazioarekin hasteko, eta lan hori pixkanaka egiteko, bi zatiak berriro

[21] `http://www.let.rug.nl/~vannoord/TextCat/`

[22] Paragrafoak esalditan banatzeko kontuan hartu diren puntuazio-ikurrak izan dira: puntua (.), puntu eta koma (;), bi puntuak (:) eta galdera-marka (?).

[23] Aplikazioak bi aukera eskaintzen ditu: a) testu-fitxategi bat pasatzen bazaio parametro gisa, testuaren hizkuntza identifikatzen du; b) esaldi bat pasatzen bazaio (`-l` aukera) esaldi horren hizkuntza identifikatzen du.

| | **Paragrafoak** | **Tokenak** | **Analizatutako** tokenak | formak | **OOV** tokenak | formak |
|---|---|---|---|---|---|---|
| Guztira | 1.175 | 96.210 | | | | |
| Ikasi | 118 | 8.886 | 8.223 | 3.025 | 1.931 | 1.032 |
| Test | 61 | 4.807 | 4.386 | 1.902 | 1.015 | 636 |

II.4 Taula: *Gero* obraren zatiketa. Obra osoari dagozkion datuetan (lehenengo errenkadan) kontuan izan behar da hainbat iragazki aplikatu direla. Kendu dira: zifrak, karaktere arraroren bat duten tokenak, letra larriz soilik osatutako tokenak eta parentesien arteko testua. Ikasteko eta testeko zatietan latinezko zitak kendu dira analisirako.

banatu dira (oraingoan eskuz): 118 paragrafo dituen zatia 4 fitxategitan banatu da, eta 61 paragrafo dituen zatia, berriz, 2 fitxategitan.

Testua prestatzeko bi urrats jarraitu dira: lehendabizi testua automatikoki analizatu da bertan dauden hitz ez-estandarrak detektatzeko, eta ondoren, etiketatzeko tresnak behar duen informazioa lortu da hitz ez-estandarrak markatuta ageri daitezen anotaziorako.

- Lehen urratsari dagokionez, analisia egiteko testua tokenizatu da eta puntuazio ikurrak kendu dira. Zenbakiz osatutako tokenak eta testuan egin diren ordezkapenak adierazteko markak analisitik kanpo utzi dira. hainbat paragraforen hasieran ageri diren zenbakiak (18, 25. . . ) eta testuari gehitu zaizkion markak aurreko urratsetan egin diren ordezkapenak direla eta: `(...)` eta `[...]`.

  Hitzak analizatzeko erabili den automata espresuki sortua da lan honetarako, Alegria *et al.* (2009a) lanean oinarrituta, eta ohiko zuzentzaileak duen analizatzaile morfologikoa baino "zorrotzagoa" da[24].

  Automatak onartzen ez dituen hitzak, hots, ez-estandarrak kontsideratzen direnak, fitxategi berri batean utzi dira, gero egoki markatu ahal horiek testuan. II.4 taulan zati bakoitzean analizatutako hitz kopurua eta detektatutako ez-estandarren kopurua jasotzen dira.

[24]Sortutako automatak ez ditu onartzen forma hobetsia duten hitzak (Euskaltzaindiaren *Hiztegi Batua*n "h." marka duten sarreran dira) horiek dialektalak direlako maiz; analisi berezia besterik ez dutenak ere ("RARE" gisa identifikatzen dira analisi horiek) ez ditu onartzen. Letra larriekin arazorik ez izateko, automata gai da letra larriz hasten diren hitzak onartzeko (estandarrak badira).

- Bigarren urratsari dagokionez, *Brat* tresna (Stenetorp *et al.*, 2012) erabili da anotazioa egiteko[25]. Detektatutako hitz ez-estandarrak testuan nabarmendu dira, eta abiapuntu gisa OOV izeneko etiketa jarri zaie (ikus II.2 irudia). Nabarmendutako OOV horiek dira anotatzaileak aztertu eta anotatu behar dituenak (Brat tresnaren erabilerari buruzko zenbait argibide eta etiketei buruzko informazioa A eranskinean ematen da).

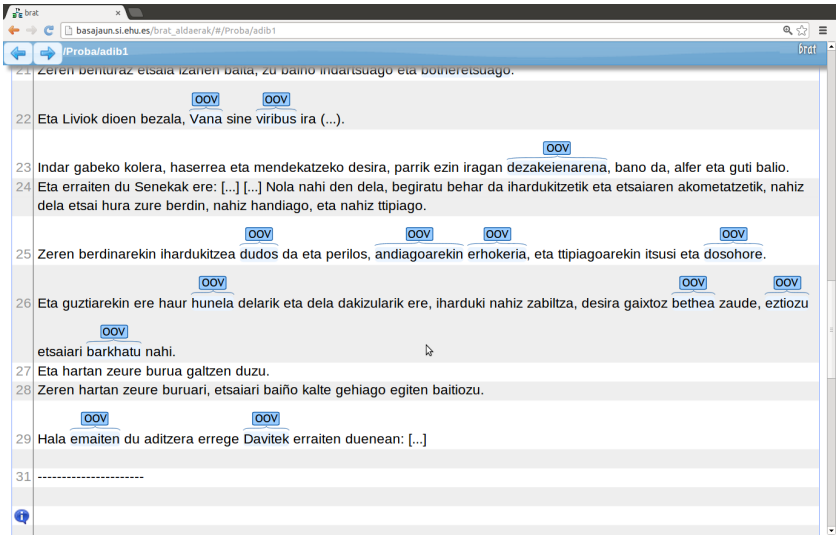
Hitzak nabarmenduta ageri daitezen testuan, formatu jakin bateko fitxategi bat sortu behar da Brat tresnarentzat, non nabarmenduta ikusi nahi dugun hitz bakoitzeko, honako informazio hau adierazi behar den: identifikadore bat (T1, T2, T3...); hitzak izango duen etiketaren izena (hasieran denak OOV) eta hitzaren hasierako zein bukaerako posizioa testuan[26].

Anotazio-lanaren adibide gisa, bi irudi ekarri ditugu hona: II.2 irudian paragrafo zati bat ageri da Brat aplikazioarekin etiketatzeko prest, non detektatutako hainbat OOV nabarmenduta ageri diren; II.3 irudian, berriz, testu zati bera ageri da anotatzaileak etiketatu ondoren. Biak konparatuta, garbi ikusten da etiketatzaileak egin dituen aldaketak.

Testua etiketatu ondoren, corpusa prest dago bertatik informazioa eskuratzeko eta normalizazioko esperimentuak hasteko.

[25] `brat.nlplab.org/` (2016-02-22an atzitua)

[26] Fitxategi hori sortzeko, programa bat idatzi dugu zeinari bi fitxategi emanda —testua bera duen fitxategi bat, eta testu horretan nabarmendu nahi den hitz zerrenda duen beste fitxategi bat—, Brat tresnak behar duen informazioa formatu egokian itzultzen duen.




II.2 Irudia: Testu zati bat Brat aplikazioaren bitartez etiketatzen hasteko prest. Detektatu diren ez-estandar guztiek OOV etiketa dute.

II.3 Irudia: Testu zati bat Brat aplikazioarekin etiketatu ondoren. Hasierako OOV hitzek anotatzaileak jarritako etiketa berria dute orain.

### II.3.5 Corpusaren prestaketa: *Peru Abarka*

Aurreko obran bezala, *Peru Abarka* obraren testu osoa formatu elektronikoan daukagu, eta hasieran zein bukaeran eskuz jarritako `<front>` eta `<back>` etiketak ditu erraz kendu ahal izateko analisitik kanpo utzi nahi diren atalak. Paragrafoen hautaketa egin baino lehen, aurreko obran egin diren antzeko iragazkiak pasa zaizkio obra honi ere:

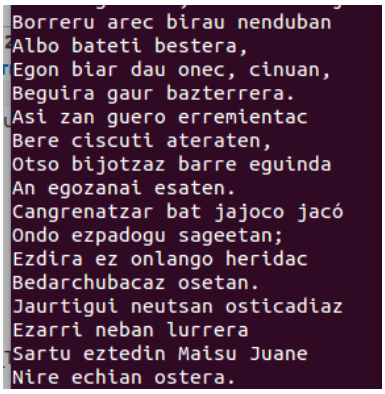
1. Hasierako eta bukaerako atalak kendu ondoren, letra larriz osatutako lerroak edo paragrafoak kendu dira. Analisirako erabiltzen den automatak ez ditu horiek onartzen eta ez dugu zarata sartzerik nahi anotazioari begira (ez dira zertan ez-estandarrak izan baina markatuta azalduko lirateke).

2. Obraren testua gainbegiratuta, ikusi da obra honetako hainbat paragrafo oso motzak direla. Hori gertatzen da, batez ere, pertsonaiaren bat zer edo zer zerrendatzen ari denean, zerrendako gai bakoitza lerro batean ematen baita; beste batzuetan, berriz, bertso moduan idatzitako zatiak ageri dira eta hor ere paragrafoak oso motzak dira. II.4 irudian testu zati txiki bat ageri da ezaugarri hori irudikatzeko. Gogoratu behar da paragrafoa dela erabili den unitatea obraren zati bat zoriz aukeratzeko, eta horren asmoa dela anotatzaileak testuinguru txikia baina nahikoa izatea etiketatzeko unean. Paragrafo motz dezente ikusi direnez obra honetan, aukeraketari ekin baino lehen, programa sinple bat idatzi da paragrafo motzak biltzeko asmoz. Lotura "berri" horiek karaktere berezi baten bitartez markatu dira, gero desegin ahal izateko.

   Paragrafo motzak lotzeko prozesu hori bete baino lehen 1.466 paragrafo daude testuan eta loturak egin eta gero, berriz, 479.

Aurreko urratsak egin ondoren, testua prest dago zoriz aukeratzeko hainbat paragrafo, horietan anotazio-lana egiteko, eta horrela corpusa prestatzeko. Bigarren corpus hau eta aurrekoa, *Gero* obrarekin prestatutakoa, antzeko tamainakoak izatea nahi da, bien esperimentuetan lortzen diren emaitzak konparatu nahi baitira.

*Gero* corpusean, obraren % 10 aukeratu da ikasteko, eta % 5 ebaluatzeko. Gogoan izanik *Peru Abarka* obraren tamaina txikiagoa dela, eta OOV tokenen portzentajea altuagoa dela, antzeko kopuruak lortzeko, 10 paragrafo aukeratu dira zoriz eta hitz ez-estandarren zein forma ez-estandarren kopuruak aztertu dira bertan. Kopuru horiek kontuan izanik, estimatu da 60–70 paragrafo inguru behar direla aurreko corpusaren antzeko tamaina lortzeko,


Borreru arec birau nenduban
Albo bateti bestera,
Egon biar dau onec, cinuan,
Beguira gaur bazterrera.
Asi zan guero erremientac
Bere ciscuti ateraten,
Otso bijotzaz barre eguinda
An egozanai esaten.
Cangrenatzar bat jajoco jacó
Ondo ezpadogu sageetan;
Ezdira ez onlango heridac
Bedarchubacaz osetan.
Jaurtigui neutsan osticadiaz
Ezarri neban lurrera
Sartu eztedin Maisu Juane
Nire echian ostera.

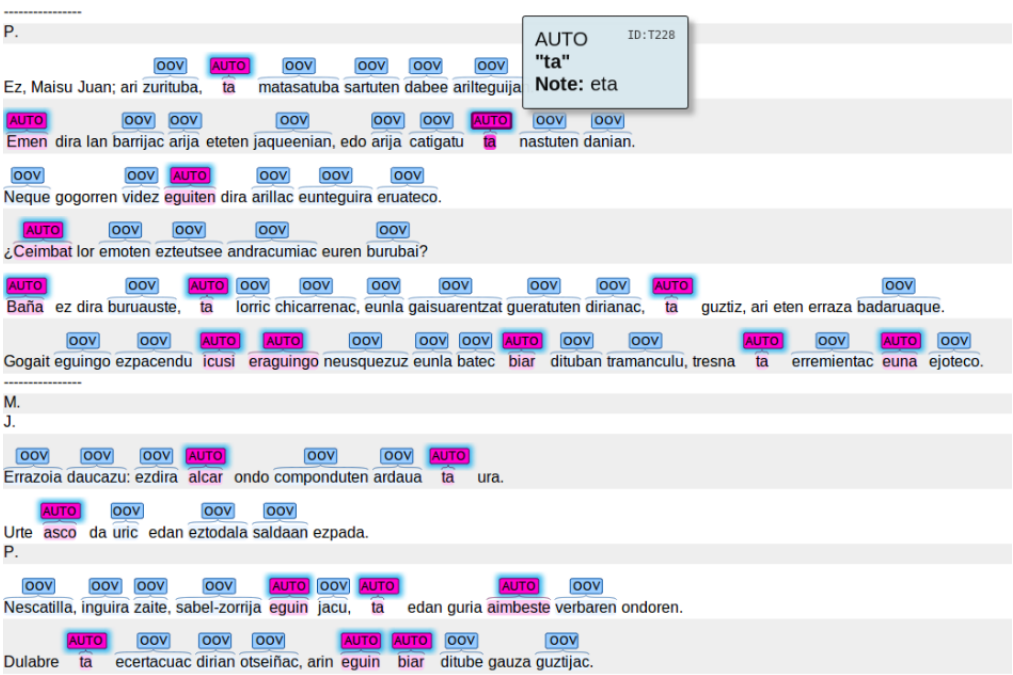
II.4 Irudia: *Peru Abarka* obraren testu zati bat obran zehar jarraian ageri diren paragrafo motzen adibide gisa.

| | | | **Analizatutako** | | **OOV** | |
|---|---|---|---|---|---|---|
| | **Paragrafoak** | **Tokenak** | tokenak | formak | tokenak | formak |
| Guztira | 479 | 24.025 | | | | |
| Ikasi | 50 | 2.078 | 1.987 | 1.199 | 1.404 | 927 |
| Test | 20 | 1.108 | 1.073 | 725 | 733 | 539 |

II.5 Taula: *Peru Abarka* obraren zatiketa. Obra osoari dagozkion datuetan (lehenengo errenkadan) paragrafo motzak "elkartuta" daude, eta letra larriz soilik osatutako paragrafoak kendu egin dira. Ikasteko eta testeko zatietan letra larriz soilik osatutako tokenak ez dira analizatu.

eta, ondorioz, 10 paragrafoz osatutako 7 fitxategi prestatu dira Brat tresnarekin etiketatzeko: 5 erabili dira ikasteko, eta 2 ebaluatzeko. Ikasteko zein ebaluatzeko zati bi horien ezaugarriak II.5 taulan ageri dira.

*Gero* obrarekin konparatuta, *Peru Abarka*n askoz OOV token gehiago daude, eta ondorioz, askoz etiketa gehiago jarri behar ditu anotatzaileak. Gainera, OOV horietako batzuk maiztasun handikoak dira: *ta* (eta), *baiña* (baina), *biar* (behar) eta abar. Oso aspergarria izan daiteke hitz arrunt bera behin eta berriro etiketatzea, eta hala, lan hori arintzeko, anotatzaileari eskatu zaio lehendabizi fitxategi bat soilik etiketatzea. Gero, fitxategi hori prozesatu da hainbat etiketa automatikoki jartzeko gainontzeko fitxategietan. Ezin denez ziurtatu etiketatze automatiko hori zuzena denik, hala




II.5 Irudia: Etiketatzaileak Brat tresnaren bitartez landu behar duen fitxategi baten zatia. Bertan, automatikoki jarritako hainbat "AUTO" etiketa ageri dira, eta etiketarekin batera aldaerari esleitutako forma estandarra (bat datorrena anotatzaileak aurretik etiketatutakoarekin).

etiketatutako hitzetan etiketa berri bat erabili da, AUTO izenekoa, eta ustez baliokidea den hitz estandarra esleitu zaio jatorrizko hitzari (ikus II.5 irudia). Etiketa automatikoki jarriak dituzten fitxategietan, anotatzaileak AUTO motako etiketak zuzenak diren aztertu behar du: AUTO etiketa zuzena bada ez du ezer aldatu behar, eta zuzena ez bada egoki aldatu behar du, etiketa berria edota baliokide berria emanez.

Beste bi fitxategi berri etiketatu eta gero, prozesu bera errepikatu da: etiketatutako lehen hiru fitxategiak prozesatu dira eta informazioa erauzi da hainbat etiketa automatikoki jartzeko gainontzeko lau fitxategietan.

## II.4 Euskarazko corpus dialektala

Gure lehenengo esperimentuei ekiteko corpusa lortzeko, Baionako IKER UMR 5478 taldearen kolaborazioa baliatu genuen, eta haien bitartez osatu genuen corpus paralelo bat lapurteraz eta estandarrez idatzitako esaldi-bikotez osatua. IKER taldeak "TSABL, *Towards a Syntactic Atlas of the Basque Language*"[27] izeneko proiektua garatu zuen 2007-2011 artean, Iparraldeko hizkeren aldakortasun sintaktikoa aztertzeko eta proiektu horren barruan "BASYQUE"[28] izeneko aplikazioa garatu zuen (Uria eta Etxepare, 2011). Aplikazioaren ezaugarriak eta aldakortasun sintaktikoa aztertzeko metodologia Uria eta Etxepare (2012) artikuluan azaltzen da, eta hortik ekarri dugu hona aplikazioa labur deskribatzeko esaldi bat:

*BASYQUE aplikazioak euskal hizkeren arteko aldakortasun sintaktikoaren inguruko informazioa gordetzeko, kudeatzeko eta kontsultatzeko askotariko aukerak eskaintzen ditu, baita izaera dialektala duten corpusak (testu-bildumak) eskuratzeko eta aztertzeko ere, besteak beste.*

Aipatzen duten informazioa datu-base sendo batean gordeta dago, hainbat informazio-iturritatik bildutako adibideak jasotzen dituena. Guretzat garrantzitsuena da adibide bakoitza (esaldi bat) bi eratan gordeta dagoela: alde batetik, "idazkera formala" deitu dutena dago, dialektoaren formari dagokiona, eta, beste aldetik, "testu normalizatua" deitu dutena, euskara estandarrari dagokiona. Corpus paraleloa osatzeko, beraz, nahikoa da datu-basean dauden adibideen bi atal horiek hartzea.

Oraindik datu-basea osatzen ari zirela, corpus paralelo txiki bat osatu genuen hainbat adibiderekin gure lehendabiziko esperimentuak bideratzeko. Corpus paraleloa, beraz, esalditan antolatua da: alde batetik, esaldia lapurteraz idatzia dago, eta, beste aldetik, esaldi baliokidea euskara estandarrean. II.6 taulan esaldi-bikote pare bat ageri dira corpus paraleloaren adibide gisa, bi aldeen arteko ezberdintasunak nolakoak diren islatzeko.

Corpusaren xehetasunak II.7 taulan ageri dira. Corpusa 2.117 esaldi-bikotez osatua da, eta esaldi horietan 12.150 hitz inguru jasotzen dira alde bakoitzean; formak kontatuz gero (hitz bakoitza behin bakarrik, nahiz xehez edo larriz idatzia egon), alde bakoitzean 3.600 forma inguru daude (zehazki, 3.830 lapurterazko esaldietan eta 3.553 euskara estandarrean idatzitako esaldietan).

Esperimentuak egiteko, bi zatitan banatu da corpusa eta ikasketa automatikoan ohikoa den moduan, zati bat ikasteko erabili da eta bestea, berriz,

[27] `http://www.iker.cnrs.fr/-tsabl-towards-a-syntactic-atlas-of-.html` (2016-02-12an atzitua)

[28] `http://ixa2.si.ehu.es/atlas2` (2016-02-12an atzitua)

| |
|---|
| (a) Ez gero uste izan **nexkatxa guziek** tu egiten **dautatela** |
| (b) Ez gero uste izan **neskatxa guztiek** tu egiten **didatela** |
| (a) Zortzi egunen mihi **phazka** emanen **baitiotet heien** kideko **guzieri** |
| (b) Zortzi egunen mihi **bazka** emanen **baitiet haien** kideko **guztiei** |

II.6 Taula: Lapurtera/Estandarra corpus paraleloaren bi adibide: (a) lapurteraz idatzitako esaldiaren bertsioari dagokio, (b) estandarrez idatzitakoari.

sistemak ikasitakoa ebaluatzeko[29]. Kasu honetan, corpusaren % 80 (1.694 esaldi) ikasteko erabili da, eta gainontzekoa, % 20 (423 esaldi), ebaluazioa egiteko. Banaketa hori zoriz egin da.

| | **Corpusa** | **% 80** | **% 20** |
|---|---|---|---|
| Esaldiak | 2.117 | 1.694 | 423 |
| Tokenak | 12.150 | 9.734 | 2.417 |
| Formak | | | |
| Lapurteraz | 3.830 | 3.292 | 1.239 |
| Estandarrez | 3.553 | 3.080 | 1.192 |

II.7 Taula: Lapurtera/Estandarra corpusa: ikasteko eta testeko zatiak.

## II.5 Beste hizkuntza batzuetako corpusak

Kapitulu honen sarreran esan dugunez, esperimentuak egiteko gaztelaniazko zein eslovenierazkoa testuekin, hizkuntza horietako zenbait ikerlariren kolaborazioa baliatu dugu haiek sortutako edota erabilitako datuak eskuratzeko. Atal honetan datu horiei buruzko ezaugarriak emango ditugu.

[29]Corpusaren tamaina txikia da eta horregatik ez da garapen-corpusik erabili. Doikuntza beharrezkoa izan denean, balidazio gurutzatua (*cross-validation*) izeneko teknika erabili da ikasteko corpusarekin.

### II.5.1 Gaztelaniazko corpusa

Gaztelaniarekin esperimentuak egiteko datuak lortzeko, Jordi Porta ikerlariaren laguntza izan dugu, eta bi datu-multzo partekatu ditu gurekin ikerlariak: (1) Freeling[30] tresnatik (Carreras *et al.*, 2004) lortutako FL-EM izeneko datu-multzoa; (2) IMPACT[31] proiektutik lortutako datu-multzoa.

#### FL-EM datu-multzoa

Porta *et al.* (2013) lanean gaztelania zaharraren inguruko esperimentuak egin dituzte normalizazioaren bidetik, eta bertan erabili dituzten datuen artean FL-EM datu-multzoa dago. FL-EM datu-multzo hori sortzeko, Freeling tresnak duen lexikoi berezi bat erabili dute, gaztelania zaharra analizatzeko duen lexikoi berezi bat, hain zuzen. Lexikoi hori nondik eta nola eraiki den Sánchez-Marco *et al.* (2011) lanean azaltzen da. *Hispanic Seminary of Medieval Studies* (HSMS) erakundeak Erdi Aroko testuz osatutako corpus bat dauka, non XII. eta XVI. mende arteko hainbat obra jasotzen diren (20 milioitik gorako hitz eta 470 mila inguru forma), eta corpus horretan ageri diren gaztelaniazko aldaerak hartu dira kontuan Freelingeko lexikoi berezi hori osatzeko. Freeling deskonposaketa morfologikoan oinarritzen denez hainbat hitz analizatzeko, lexikoi horretan ez dira ageri *-mente* bukaera duten aditzondoak, enklitikoak dituzten aditz formak, diminutiboak eta aumentatiboak, superlatiboak eta abar. Horrez gain, FL-EM datu-multzoa osatzeko ikerlariek kendu egin dituzte lexikoian ageri diren izen propioak, zenbaki erromatarrak eta hitz anitzeko edo lotutako hitzak dituzten sarrerak Porta *et al.* (2013).

| **Kategoria** | | **Kategoria** | |
|---|---|---|---|
| Adjektiboak | 4.048 | Izenordainak | 292 |
| Izenak | 11.257 | Aditzondoak | 254 |
| Aditzak | 20.339 | Konjuntzioak | 160 |
| Preposizioak | 64 | Interjekzioak | 117 |
| Determinatzaileak | 172 | Bestelakoak | 6 |
| Guztira | | | 36.709 |

II.8 Taula: FL-EM multzoa: kategoria bakoitzaren sarrera kopurua.

[30] `nlp.lsi.upc.edu/freeling/` (2016-02-22an atzitua)
[31] `www.impact-project.eu`, (2016-02-23an atzitua)

| Forma zaharra | Lema_kategoria |
|---|---|
| aalguna | alguno_D |
| aalguna | alguno_P |
| aamigo | amigar_V |
| aamigo | amigo_A |
| aamigo | amigo_N |

II.9 Taula: FL-EM multzoko adibide batzuk: egungo lema eta kategoria.

FL-EM datu-multzoak 36.709 sarrera ditu, eta II.8 taulan ikus daiteke nola banatzen diren sarrera horiek kategorien arabera.

Sarrera bakoitzak duen informazioa honako hau da: aldaera edo forma zaharra, horri dagokion egungo lema eta kategoria. II.9 taulan adibide batzuk ageri dira.

Tesi-lan honetan planteatzen diren esperimentuetarako behar den informazioa forma zaharra eta egungo formaren arteko erlazioa da, eta hala, gaztelaniazko egungo formak sortu behar izan ditugu. Horretarako, FL-EM multzoaren beste bertsio bat erabili dugu, non egungo lemaz eta kategoriaz gain, sarrera bakoitzari dagokion analisi morfosintaktiko osoa ageri den (ikus II.10 taula).

| Forma zaharra | Lema_analisia |
|---|---|
| aalguna | alguno_DI0FS0 |
| aalguna | alguno_PI0FS000 |
| aamigo | amigar_VMIP1S0 |
| aamigo | amigo_AQ0MS0 |
| aamigo | amigo_NCMS000 |

II.10 Taula: FL-EM multzoko adibide batzuk: egungo lema eta analisi morfosintaktikoa.

Informazio morfosintaktikoa duen bertsioak, hasierakoak baino sarrera gehiago ditu, 45.948 sarrera hain zuzen (9.239 sarrera gehiago aurrekoak baino). Hori gertatzen da forma zahar batek, kategoria berean, analisi morfosintaktiko bat baino gehiago izan dezakeelako. Aditzekin erraz ikusten da hori gerta daitekeela, lehen eta hirugarren pertsona singularrek forma bera

| Forma zaharra | Egungo forma | Lema_analisia |
|---|---|---|
| aalguna | alguna | alguno_DI0FS0 |
| aalguna | alguna | alguno_PI0FS000 |
| aamigo | amigo | amigar_VMIP1S0 |
| aamigo | amigo | amigo_AQ0MS0 |
| aamigo | amigo | amigo_NCMS000 |

II.11 Taula: FL-EM datu-multzoko adibide batzuk: egungo forma sortu da analisi morfosintaktikoa eta Freeling tresna erabiliz.

| Forma zaharra | Egungo forma |
|---|---|
| amassasse | amasara\|amasase |
| abuerat | hubiera\|hubiese |
| algunt | algún\|alguno |

II.12 Taula: FL-EM datu-multzoko adibide batzuk: egungo forma bat baino gehiago sortzen duten adibideak.

hartzen baitute askotan: `amassasse` aditz-forma zaharrak bi analisi morfologiko ditu, `amasar_VMSI1S0` eta `amasar_VMSI3S0`, lehen zein hirugarren pertsona singularra izan baitaiteke.

Informazio morfosintaktikoa baliatuz, egungo formak sortu ditugu Freeling tresnaren bitartez, eta fitxategi berri bat lortu dugu, non forma zaharraren ondoan forma berria ageri den. II.11 taulan adibide batzuk ageri dira.

Egungo formak sortzean badira kasu bereziak, eta gerta daiteke analisi morfologiko bati azaleko forma bat baino gehiago egokitzea. Gertaera horren adibidea aditzen artean aurki daiteke berriro, zehazki, subjuntiboko lehenaldian (nahiz eta hori ez den kasu bakarra). Esaterako, `amasar` aditzaren subjuntiboko lehenaldiaren lehenengo pertsonak (eta hirugarrenak) bi forma posible ditu: `amasara` eta `amasase`.

Egungo forma bat baino gehiago sor daitezkeen kasuetan, forma horiek `|` karakterearen bitartez banatuak ageri dira, II.12 taulako adibideetan ikusten den moduan. Ebaluazioari begira, egungo forma bakarra nahi genuen eta, beraz, kasu horietan aukeraketa bat egin behar izan da formen artean. Aukeraketa egiteko erabili den irizpidea forma zaharretik "gertuen" dagoen egungo forma uztea izan da.

| **FL-EM** | |
|---|---|
| Bikote kopurua | 31.046 |
| Zaharra=Egungoa | 1.248 (% 4,02) |
| Hitz anitzekoak | 0 |

II.13 Taula: FL-EM datu-multzoarekin sortu den bikote kopurua esperimentuetarako.

Eta gertutasuna neurtzeko, "Sørensen-Dice" izeneko koefizientea erabili dugu[32]. Bi formen arteko bigrama komunak eta osora dituzten bigrama kopuruak kontuan hartuta, koefiziente horrek bi formen arteko gertutasunaren edo antzekotasunaren neurria ematen du[33]. Beraz, forma zaharra eta berri posible bakoitzaren artean koefiziente hori kalkulatu dugu[34] eta koefiziente altuena duen bikotea utzi dugu. Esaterako, II.12 taulako lehenengo adibideari dagokionez, `amassasse - amasara` bikotearen koefizientea 0,57 da, eta `amassasse - amasase` bikotearena, berriz, 0,86. Ondorioz, bigarren bikotea utzi da: `amassasse - amasase`.

Egungo forma bat baino gehiago dituzten sarrerak prozesatu eta gero, eta forma-bikote bakoitza behin soilik utzita (errepikatuak egon zitezkeen), azkenean 31.046 bikoteko zerrenda lortu dugu esperimentuak egin ahal izateko. Esperimentuei ekin baino lehen, bi ezaugarri analizatu dira zerrendan (II.13 taulan bildu dira emaitzak):

- Analizatu da zenbat sarreratan gertatzen den forma zaharra eta berria berdinak direla, eta 1.248 sarreratan gertatzen da hori (% 4).
- Aztertu da ea forma zaharra edo berria hitz anitzekoa den, eta ez da horrelakorik gertatzen.

[32]Hainbat izenez ezagutzen da koefiziente hori, baina erabilienak dira *Sørensen index* eta *Dice's coefficient*.

[33]Xehetasun gehiagorako: `https://en.wikipedia.org/wiki/S\%C3\%B8rensen\%E2\%80\%93Dice_coefficient`

[34]Kalkulua egiten duen programa Github-etik eskuratu dugu, `https://github.com/wooorm/dice-coefficient` helbidean.

### IMPACT datu-multzoa

IMPACT, *Improving Access to Text*, proiektua Europako Komisioak babestua da eta haren helburu nagusia da OCR sistemetako doitasuna eta testu historikoen atzipena hobetzea. Honela definitzen dute proiektua haren webgunean:

*The IMPACT project will remove many of these barriers. The project will push innovation in OCR technology and language technology for historical document processing and retrieval, and share expertise to build capacity in digitisation across Europe. During the project a Centre of Competence will be set up in order to provide a central service entry point for all libraries, archives and museums involved in the digitisation of textual material.*

Gaztelaniaren kasuan, corpus bat eta lexikoi bat sortu dira proiektu horren barruan:

- Corpusak 107 obra jasotzen ditu 1481 eta 1748 urte artekoak, autore eta genero askotarikoak (prosa, antzerkia eta olerkiak). Corpus hori bi ataletan banatuta dago: GT (*Ground-Truth*) atala, IMPACTek sortutakoa eta 21 dokumentu jasotzen dituena; eta BVC atala, *Biblioteca Virtual Miguel de Cervantes* liburutegi digitaletik jasotako 86 testu jasotzen dituena. Corpusaren xehetasunak zein anotatzeko erabili diren irizpideak eta formatuak Sánchez-Martínez *et al.* (2013) lanean deskribatzen dira.

- Lexikoiari dagokionez, bi bertsio sortu dituzte, eta horietako bat "konpaktua" da: bertan ez dira corpuseko zitak ageri.

IMPACT proiektuan gaztelaniarako sortutako baliabide horiek eskura daude *IMPACT Centre of Competence* erakundearen atarian [35].

Jordi Porta ikerlariak IMPACT proiektuko lexikoi konpaktu hori prozesatua zuen, eta gure esku jarri du 24.009 sarrera dituen fitxategi bat, non sarrera bakoitzeko forma zaharra eta forma berria ageri diren. II.14 taulan adibide batzuk ageri dira.

Beraz, IMPACTeko datu-multzoan ez da beharrezkoa izan inolako prestaketarik esperimentuetan erabili ahal izateko, forma zaharra eta egungoaren arteko erlazioa jadanik adierazita baitago. Dena den, FL-EM multzoarekin konparatu ahal izateko, hari egin zaizkion azterketak egin dira IMPACT multzoan ere: aztertu da zenbat sarreratan berdinak diren forma zaharra

[35] `www.digitisation.eu/tools-resources/language-resources/impact-es/` (2016-02-23an atzitua)

| Forma zaharra | Egungo forma |
|---|---|
| abraçandola | abrazandola |
| bruxo | brujo |
| cauallerescas | caballerescas |

II.14 Taula: IMPACT datu-multzoko adibide batzuk.

| IMPACT | |
|---|---|
| Bikote kopurua | 24.009 |
| Zaharra = Egungoa | 15.337 (% 63,88) |
| Hitz anitzekoak | 206 (% 0,86) |

II.15 Taula: IMPACT datu-multzoa. Kopuruak eta aztertutako ezaugarriak.

eta forma berria, eta aztertu da ea sarreretan hitz anitzeko formak dauden, zaharrak ala berriak.

Azterketa horren emaitzak II.15 taulan jaso dira, eta konparatzen badira II.13 taulako kopuruekin, agerian geratzen da bi datu-multzoak ezaugarri ezberdinak dituztela: FL-EM multzoko % 4 sarreratan soilik gertatzen da bi formak berdinak direla, eta IMPACT multzoan, berriz, % 64 sarreratan gertatzen da hori. Bestalde, kontuan hartu behar da bata eta bestea osatzeko erabili diren obren garaia: FL-EM datu-multzoa sortzeko XII. eta XVI. mende arteko obraz osatutako corpusa erabili da, eta IMPACT datu-multzoa sortzeko, berriz, XV. eta XVIII. mende artekoak.

### II.5.2 Eslovenierazko corpusa

Eslovenierarekin esperimentuak egiteko corpusa Scherrer eta Erjavec (2015) lanean deskribatzen da. Lan horretan eslovenierazko hitz historikoen modernizazioa edo normalizazioa planteatzen da metodo jakin bat aplikatuta, eta bai erabili dituzten datuak, bai aplikatu duten metodoa, zehatz deskribatzen dira aipatutako lanean. Horrez gain, esperimentuak egiteko prestatu dituzten datu guztiak ikertzaile ororen eskura jarri dituzte egileek Interneten[36], eta hala eskuratu ahal izan ditugu guk. Datu horien artean daude esloveniera historikoaren bi lexikoi, bata entrenatzeko eta bestea testeatzeko, eta egun-

[36] http://nl.ijs.si/imp/experiments/jnle-dataset/ (2016-03-4an atzitua)

go eslovenieraren erreferentziazko hitz zerrenda bat, hitzen maiztasunarekin anotatua.

### Eslovenierazko datu-multzoa

Eslovenieraren ortografia berandu estandarizatu zen, XIX. mendearen bukaera aldera. Egun erabiltzen duten alfabeto modernoa, Gaj izeneko alfabetoa, 1840ko hamarkadan definitu zen, eta hori baino lehen Bohorič izeneko alfabetoa erabiltzen zen. Bi alfabetoen arteko diferentzia nagusia sei soinuri dagokie, eta alfabeto zaharrean idatzitako testuak irakurtzea lan zaila bilakatzen da egungo hiztunentzat. Izan ere, egungo alfabetoan alfabeto zaharreko ia letra guztiak erabiltzen dira baina beste soinu batekin. Dena den, testu zaharretan aurkitzen diren aldaerak ez dira suertatzen alfabeto aldaketagatik soilik eta beste zenbait aldaeraren oinarrian aldaketa fonologikoak, morfologikoak edota ortografikoak daude.

Scherrer eta Erjavec (2015) lanean erabiltzen diren datuak esloveniera historikoaren IMP baliabideetatik[37] erauziak dira (Erjavec, 2015), *IMP goo300k* corpusetik (300.000 hitz inguru) eta *IMP foo3M* corpusetik (3 milioi hitz inguru), hain zuzen ere. IMP bi corpus horietatik ohiz kanpoko lau testu ezabatu dituzte ikerlariek, eta hala lortu dituzte *goo* eta *foo* deitu dituzten corpusak. Lehenengoak eslovenieraz idatzitako 85 testu historikotako hainbat orrialde jasotzen ditu eta eskuz anotatuta dago, guztiz; bigarrenak, berriz, 321 testutako hainbat orri jasotzen ditu (aurreko 85 testuak multzo horretan daude) baina ez dago guztiz anotatua. Izan ere, aurreko corpusa zabaltzeko asmoz sortu zen, eta hala, aurrekoan anotatutako hitzak ez ziren berriro anotatu.

Bi corpus horietako testuak hiru multzotan banatu dituzte kontuan izanik testuaren garaia eta bertan erabiltzen den alfabetoa:

- **18B** multzoan XVIII. mendeko bigarren erdiko testuak daude, guztiak Bohorič alfabetoan idatziak;
- **19A** multzoan XIX. mendeko testuak daude (batez ere lehen erdikoak), Bohorič alfabetoan idatziak;
- **19B** multzoan XIX. mendeko testuak daude (batez ere bigarren erdikoak), Gaj alfabetoan idatziak.

Bi corpusetako datu zehatzak II.16 taulan ageri dira.

[37] `nl.ijs.si/imp/` (2016-II-23an atzitua)

| | | | | |
|---|---|---|---|---|
| | | ***goo* corpusa** | | |
| | Obrak | Orriak | Hitzak | Anotatuak |
| 18B | 8 | 155 | 22.100 | 21.807 |
| 19A | 9 | 122 | 41.861 | 41.468 |
| 19B | 70 | 751 | 203.163 | 202.020 |
| | | ***foo* corpusa** | | |
| | Obrak | Orriak | Hitzak | Anotatuak |
| 18B | 11 | 1.000 | 146.060 | 15.353 |
| 19A | 18 | 697 | 401.423 | 14.682 |
| 19B | 297 | 2.873 | 2.358.792 | 66.393 |

II.16 Taula: Eslovenierazko *goo* eta *foo* corpusen tamaina: testuak, orriak eta hitzak. "Hitzak" zutabea token guztiei dagokie eta "Anotatuak" zutabea, berriz, eskuz anotatutako tokenei (hitz anitzetako tokenak ez dira kontuan hartu eta horregatik bi zutabe horiek ez datoz bat *goo* corpusean).

*Goo* eta *foo* corpusetatik lexikoi bana erauzi dute, $L_{goo}$ eta $L_{foo}$, eta erauzketa horretan hainbat token iragazi dituzte, esaterako zenbakiak, beste hizkuntzatako hitzak, hitz anitzeko tokenak eta antzekoak. Horrez gain, $L_{foo}$ corpusetik $L_{goo}$ lexikoan ageri diren hitzak ezabatu dituzte, eta hala, osatutako bi lexikoiak disjuntuak dira.

Lexikoietako sarrera bakoitza lau eremuz osatua da:

- *wform*, corpusean ageri den forma historikoa da, letra xehez idatzita;
- *nform*, Gaj alfabeto modernoaren araberako hitz normalizatua da, transliterazio-arauak aplikatu eta gero;
- *mform*, eskuz anotatutako egungo hitza da;
- *freq*, sarreraren maiztasuna corpusean adierazten duen balioa da.

Bi lexikoien zenbait ezaugarri ekarri ditugu II.17 taulara, baina Scherrer eta Erjavec (2015) lanean deskribapen zehatza egiten dute egileek eta xehetasun gehiago ematen dira. "*Unique wforms*" zutabeak formen anbiguotasuna adierazten du, forma historiko bat egungo forma bat baino gehiagorekin lotuta egon baitaiteke. Zutabe horren balioak, normalizazio-sistemak lortu ahal izango duen doitasunaren goiko muga bat adierazten du: sistema hitz isolatuak modernizatzen saiatzen denez, inongo testuingururik gabe, ez

| | | $L_{goo}$ **lexikoia** | |
|---|---|---|---|
| | Sarrerak | *Unique wforms* | *wform=mform* |
| 18B | 6.644 | 6.494 (% 97,7) | 1.181 (% 17,8) |
| 19A | 11.600 | 11.352 (% 97,9) | 2.755 (% 23,8) |
| 19B | 28.011 | 27.252 (% 97,3) | 19.635 (% 70,1) |
| | | $L_{foo}$ **lexikoia** | |
| | Sarrerak | *Unique wforms* | *wform=mform* |
| 18B | 4.774 | 4.641 (% 97,2) | 340 (% 7,1) |
| 19A | 5.907 | 5.801 (% 98,2) | 890 (% 15,1) |
| 19B | 10.673 | 10.470 (% 98,1) | 8.120 (% 76,1) |

II.17 Taula: $L_{goo}$ eta $L_{foo}$ lexikoien ezaugarriak.

du forma moderno bat baino gehiago emango forma zahar bakoitzeko, eta hala, doitasunean ezin izango du lortu % 100. Azken zutabeari dagokionez,"*wform=mform*", zutabe horrek adierazten du zenbat sarreratan gertatzen den forma historikoa eta anotatutako egungo forma berdinak direla (testu zaharrenetan kopuru hori askoz txikiagoa da, oraindik alfabeto zaharra erabiltzen zelako).

$L_{goo}$ eta $L_{foo}$ lexikoiak sei fitxategitan banatuta daude, tauletan adierazten den moduan, eta bertatik erauzi dugu esperimentuetarako informazioa, hau da, forma historikoa eta egungo forma jasotzen duten bikoteak.

II.5.2 atalaren hasieran esan dugunez, esperimentuak egiteko datuen artean, lexikoiez gain egungo eslovenierazko hitz zerrenda bat dago, *Sloleks* izeneko lexikoiaren bitartez lortu dutena ikerlariek. Egungo eslovenieraren erreferentziazko lexikoi flexionatua da *Sloleks*, 100.000 lema inguru ditu, eta esperimentuetarako xehez idatzitako formak eta haien maiztasuna erauzi dituzte bertatik ikerlariek (920.794 forma guztira).

Hitz zerrenda hori normalizazio-sistemak proposatzen dituen formak iragazteko erabil daiteke esperimentuetan, hau da, alboratzeko egungo formak ez diren proposamenak. Horretarako erabili behar bada, ordea, interesgarria da jakitea lexikoietan ageri diren egungo formak (*mform*) zerrenda horretan dauden ala ez; eta analisi hori egin dute, hain zuzen, Scherrer eta Erjavec ikerlariek. Hala, anotatutako egungo forma ez badago zerrendan '*' karakterea erantsi diote, *out-of-vocabulary* dela adierazteko (OOV). Horrekin batera, forma arkaikoak identifikatu dituzte eta halakoei '!' karakterea erantsi diete (forma arkaikoek bi karaktereak izango dituzte erantsiak). Normali-

zazio-esperimentuak egiteko egungo forma "garbia" interesatzen zaigunez, forma historikoa eta egungo forma erauztean erantsitako karaktere horiek kendu dizkiegu egungo formei.

Adibide gisa, II.18 taulan $L_{goo}$18B lexikoiko sarrera batzuk kopiatu dira. Bestalde, lexikoi bakoitzean ageri diren OOV kopurua eta hitz arkaikoen kopurua II.19 taulan jaso dira, eta ikusten denez, OOV hitzen portzentajea nahiko altua da: % 18 ingurukoa $L_{foo}$ lexikoian (hiru zatietan); % 15 ingurukoa, batez beste, $L_{goo}$ lexikoian.

| forma historikoa (*wform*) | forma normalizatua (*nform*) | egungo forma (*mform*) | maiztasuna (*freq*) |
|---|---|---|---|
| `bazelne` | `bacelne` | `baclje!*` | 1 |
| `belákovo` | `belakovo` | `beljakovo*` | 1 |
| `dakonzhuvavza` | `dakončuvavca` | `dokončevalca!*` | 1 |
| `dènarjov` | `denarjov` | `denarjev` | 1 |
| `sdajzi` | `zdajci` | `zdajci` | 9 |
| `tèdaj` | `tedaj` | `tedaj` | 28 |

II.18 Taula: $L_{goo}$18B lexikoiko sarrera batzuk. Sarrera bakoitzak lau gai ditu: forma historikoa, forma normalizatua, egungo forma eta maiztasuna.

| | $L_{goo}$ **lexikoia** | | |
|---|---|---|---|
| | Egungo formak | OOV | Arkaikoak |
| 18B | 5.065 | 740 (% 14,6) | 507 (% 10,0) |
| 19A | 9.594 | 1.180 (% 12,3) | 718 (% 7,5) |
| 19B | 23.888 | 4.417 (% 18,5) | 2.139 (% 9,0) |
| | $L_{foo}$ **lexikoia** | | |
| | Egungo formak | OOV | Arkaikoak |
| 18B | 3.685 | 663 (% 18,0) | 414 (% 11,2) |
| 19A | 4.830 | 909 (% 18,8) | 565 (% 11,7) |
| 19B | 9.826 | 1.799 (% 18,3) | 772 (% 7,9) |

II.19 Taula: $L_{goo}$ eta $L_{foo}$ lexikoien OOV kopurua eta hitz arkaikoen kopurua.

# III. KAPITULUA

## Metodoen aurkezpena eta hautaketa

### III.1 Sarrera

Kapitulu honetan hiru metodo ezberdin deskribatuko ditugu tesian planteatzen den ataza ebazteko, hau da, hitz ez-estandarrei edo aldaerei dagokien hitz estandarra automatikoki esleitzeko. Metodoen eraginkortasuna aztertzeko, hainbat esperimentu planteatzen dira kapitulu honetan, eta esperimentu horiek guztiak corpus bakar baten gainean egiten dira: bigarren kapituluko II.4 atalean deskribatu den lapurtera/estandarra corpus paraleloaren gainean. Hiru metodoekin ikasketa automatikoko teknikak erabili dira, hau da, corpusaren zati bat erabili da bertatik ikasteko (*train*), eta gero ikasitakoa aplikatu da corpusaren beste zatian, hots, ebaluaziorako zatian (*test*).

Esperimentuen emaitzak analizatuta, ikusiko dugu hiru metodoen artean bat gailentzen dela, WFST teknologia (*Weighted Finite State Technology*) erabiltzen duen metodoa, eta hala, IV. kapituluan metodo eta teknologia horretan sakonduko dugu gehiago.

Hiru metodoen xehetasunekin hasi baino lehen, III.2 atalean azterketa bibliografikoa egingo dugu eta antzeko ataza planteatu duten lanen berri emango dugu. Gero, III.3 atalean, erabili ditugun hiru metodoen oinarriak azalduko ditugu, eta III.4 atalean planteatutako esperimentuak zein lortutako emaitzak aztertuko ditugu. Bukatzeko, III.5 atalean, esperimentuen emaitzetatik atera ditugun ondorioak azalduko ditugu lanarekin aurrera jarraitzeko.

## III.2 Azterketa bibliografikoa

Hizkuntzen aldaerak modu gainbegiratuan ikasteko atazaren inguruan hainbat lan aurkitzen dira gaiko literaturan, eta horietan zehar ikusten da ataza hori beste zenbait arlorekin lot daitekeela, hala nola fonologia eta morfologia konputazionalarekin, ikasketa automatikoarekin eta corpusetan oinarritutako lanekin.

Jadanik esan dugun moduan, lehen kapituluko I.2 atalean, testu ez-estandarrak normalizatzeko gaur egun erabiltzen diren metodoak hiru multzotan bana daitezke:

- Erregeletan oinarritzen diren metodoak. Metodo horietan eskuz idazten dira erregela fonologikoz osatutako gramatikak, aldaeretan gertatzen diren fenomenoak aztertu ondoren. Gramatika horien bitartez aldaerei dagozkien forma estandarrak bila daitezke gero.

- Ikasketa automatikoko teknikak aplikatzen dituzten metodoak. Sistema horietan hainbat adibide ematen zaizkio sistemari bertatik ikas dezan, eta gero ikasitakoa aplikatzen da adibide berrietan. Ikasteko adibideak lortzeko eskuzko lana behar denez, teknika hauek gainbegiratuak direla esaten da.

- Teknika ez-gainbegiratuak (*unsupervised*) aplikatzen dituzten metodoak. Aldaeretatik gertuen dauden forma estandarrak bilatzen dituzten metodoak dira, eta gehienek edizio-distantzia edota distantzia fonetikoa kalkulatzen dute. Erabat automatikoak direnez, askotan oinarrilerro gisa (*baseline*) erabiltzen dira.

Esan bezala, teknika ez-gainbegiratuen multzoan gehien aplikatu izan diren bi teknikak edizio-distantzian eta distantzia fonetikoan oinarritu dira. Edizio-distantzia neurtzeko gehien erabili izan den metrika, zalantzarik gabe, Levenshtein distantzia izan da (Levenshtein, 1966). Metrika horren arabera, bi karaktere-kateren arteko distantziak adierazten du zenbat edizio-eragiketa egin behar diren kate bat bestean bihurtzeko. Hiru dira metrika horrek kontuan hartzen dituen edizio-eragiketak: karaktere bat ezabatzea, karaktere berri bat sartzea edota karaktere bat beste batez ordeztea; eta eragiketa bakoitzari egokitzen zaion kostua 1 da. Esaterako, `hunetzaz-honetaz` hitzen arteko Levenshtein distantzia 2 da, bi edizio-eragiketa egin behar baitira lehenengoa bigarrenean bihurtzeko: `u` ordez `o` jarri (`honetzaz`), eta lehenengo `z` ezabatu (`honetaz`).

Distantzia fonetikoaren ideia aurreko bera da, baina kasu horretan konparatzen diren kateak karakterez osatuak egon beharrean, fonemaz osatutako kateak dira. Distantzia fonetikoa neurtu nahi bada, oro har, beharrezkoa izango da algoritmoren bat letren eta fonemen arteko jauzia egiteko (*Soundex* algoritmoa, esaterako, oso erabilia izan da, baina ingeleseko ahoskera-arauen arabera egiten du lan).

### III.2.1 Erregeletan oinarritutako metodoak

Jurish (2010) lanean hiru hurbilpen ezberdin ebaluatzen dira aleman historikoan idatzitako hitzak kanonikalizatzeko edo normalizatzeko: ebaluatutako hiru tekniketatik bi, metodo ez-gainbegiratuetan oinarritzen dira (distantzia fonetikoan eta Levenshtein edizio-distantzian), eta hirugarrena eskuz idatzitako erregeletan oinarritzen da.

Teknikak ebaluatzeko erabiltzen den corpusa aleman historikoan idatzitako bertsoz osatua da: *gold-standard* bat prestatu da (11.242 token, 4.157 forma), non forma historiko bakoitza eskuz anotatu den egungo formarekin (edo formekin). Forma historikoak DWB (*Deutches Wöterbuch*, Bartz *et al.* (2004)) corpusetik hartu dira eta aipatzekoa da multzo horretako forma historikoen % 70,8 bat datozela egungo formekin.

Erregeletan oinarritutako metodoan, oinarri linguistikoa duten 306 erregela idatzi dira eskuz aleman historikoaren hainbat lema-forma bikote aztertuta. Lema-forma bikote horiek lehen aipatutako DWB corpusetik erauzi dira automatikoki (5,5 milioi hitz dituen corpusa da) eta idatzitako erregelak WFST transduktore batean konpilatu dira.

Hiru sistemak *gold-standard*aren bitartez ebaluatu dira, eta erregeletan oinarritutako metodoa izan da ebaluazio-emaitza onenak lortu dituena. Formen arabera egiten den ebaluazioan lortutako emaitzak dira[1]: $P = 98,5$, $R = 88,4$, $F = 93,2$. Oso emaitza onak dira, balioak altuak baitira hiru parametroetan. Erreferentzia gisa, dena den, oinarri-lerro gisa planteatzen duten sistemaren emaitzak hartu behar dira kontuan: $P = 99,9$, $R = 70,8$, $F = 82,9$; edota Levenshtein distantzian oinarritutako metodoak lortzen dituena: $P = 96,6$, $R = 78,9$, $F = 86,9$.

Pettersson *et al.* (2012) lanean eskuz idatzitako 29 erregela erabiltzen dituzte ikerlariek suedieraz idatzitako hitz historikoak normalizatzeko. Nahiz eta erregelak sortzeko XVII. mendeko obra bakar bat landu duten, emaitzek adierazten dute erregelak aplikagarriak direla 1527 eta 1812 urte bitarteko hainbat testutan. Garai horretako 15 testu dituzte eta horietatik 33.000 in-

[1] $P$, $R$ eta $F$ parametroen definizioa III.4.1 atalean ikusiko dugu.

guru token jasotzen dituen *gold-standard* bat prestatu dute ebaluaziorako. Multzo horretan dauden formen % 65,2 bat datoz egungoekin eta normalizazioa egin eta gero, berriz, % 73,0 baliora igotzen da portzentaje hori.

Porta *et al.* (2013) lanean, gaztelania zaharrean idatzitako formak analizatzeko sistema bat proposatzen da, non aurretik garatutako hainbat baliabide erabiltzen diren, hala nola egungo gaztelaniazko lexikoi bat, transkripzio fonologikoa egiten duen sistema bat, eta erregela multzo bat, modelatzen duena gaztelaniako soinuek jasan duten eboluzioa Erdi Arotik aurrera. Proposatzen duten sistemaren funtsezko osagaia transduktore bat da, forma zaharrak eta egungoak erlazionatzen dituen transduktorea, eta lanean bi transduktore konparatzen dituzte: bata Levenshtein distantzian oinarritua da (ez-gainbegiratua) eta bestea oinarri linguistikoko zenbait transduktoreren konposaketaren bitartez lortzen da, garrantzi berezia izanik transduktore horien artean denboran zehar gertatutako soinu-aldaketak gauzatzen dituen transduktoreak.

Bi aukera horiek ebaluatzeko, bost *gold-standard* erabiltzen dituzte, jatorria eta garaiaren arabera ezberdinak direnak (Erdi Aroa zein Urrezko Aroa), eta transduktore linguistikoa da bostetan emaitza onenak lortzen dituena. $F$ neurrian lortzen duen hobekuntza 30 puntutik gora dago hiru datu-multzotan, eta beste bietan aldea txikiagoa bada ere, emaitza onena lortzen duena da. FL-EM izeneko *gold-standard*ean, esaterako, transduktore linguistikoarekin lortzen dituzten emaitzak dira: $P = 69,75$, $R = 89,02$, $F = 78,22$. Hurrengo kapituluan FL-EM datu-multzoarekin egingo dugu lan gure metodoaren emaitzak konparatzeko orain aipatu ditugun horiekin.

### III.2.2 Aldaera fonologikoak ikasteko metodoak

Mann eta Yarowsky (2001) lanean, edozein bi hizkuntzaren arteko itzulpen-lexikoiak induzitzeko metodo bat proposatzen da, bi hizkuntzen artean zubi gisa erabiliz helburu-hizkuntzaren familia bereko beste hizkuntza bat. Proposamen horretan suposatzen da iturburu-hizkuntza eta zubi-hizkuntzaren arteko lexikoi bat existitzen dela, eta aztertzen duten problema da zubi-hizkuntza eta helburu-hizkuntzaren arteko 'jauzia' nola egin (biak familia bereko hizkuntzak izanik).

Proposatzen dituzten bi teknikak ikasketan oinarritzen dira: lehenengoak edizio-distantziaren pisua kalkulatzen ikasten du eta bigarrena Markov-en eredu ezkutuan oinarritzen da ikasteko. Emaitza onenak sistema hibridoak lortzen ditu, nahiz eta oso emaitza onak ez izan. Kontuan hartu behar da ataza zaila dela: ez dira aldaerak, hizkuntza desberdinak baizik.

Mann eta Yarowskyren lanean oinarrituta, Scherrer-k (2007) Suitzako

alemanaren eta aleman estandarraren arteko lexikoi bitar bat induzitzeko ataza planteatzen du. Bi urratsetan banatzen du indukzioa: (1) sortu dialektoko hitz bakoitzarentzat 500 hitz estandar dituen hautagai zerrenda bat antzekotasun metrika jakin baten arabera, eta (2) iragazi hautagaiak hitz estandarren lexikoi baten bitartez, 0–20 sarrerako hautagai zerrenda lortuz. Sorkuntzaren urratserako, Levenshtein distantzia erabiltzen du Scherrerrek oinarri-lerro gisa, eta beste bi metrika berri gauzatzen ditu: lehenengoa Mann eta Yarowsky (2001) artikuluko lehen metrika da (edizio-distantzien pisuak ikastea, *adaptative string distance*); bigarrena, berriz, eskuz idatzitako 50 erregela inguru jasotzen dituen WFST transduktore batean oinarritzen da (horrek ez du ikasketa behar).

Hiru metrikekin lortzen diren emaitzak 2.366 hitz-bikote dituen testeko corpus batekin ebaluatu dira eta bi metrika berriekin % 11 inguruko hobekuntza lortu da $F$ neurrian. Geroago ikusiko dugunez, metodo horietan erabiltzen den sorkuntza-iragazketaren ideia hori bera erabiliko dugu guk ere proposatutako metodoetan.

Kestemont *et al.* (2010) lanean lematizazioa egitea planteatzen da nederlanderako *Corpus-Gysseling* literatur corpusean. 1200 eta 1300 urte bitarteko literatur eskuizkribuak jasotzen ditu corpusak, 27 obra guztira, eta bertan ageri diren aldaera ortografikoak izugarriak dira, lematizazioaren ataza zailduz. Corpusa anotatuta dago eta bertan ageri diren 40.471 formei, 14.892 lema ezberdin esleitu zaizkie eskuz. Ataza ebazteko, hizkuntzarekiko independentea den memorian oinarritutako ikasketa-sistema bat proposatzen dute (MBL, *Memory-based learning*), gai dena lema bakoitzari lotutako formen artean gertatzen diren aldaketak 'gogoratzeko'. Horrez gain, metrika berri bat erabiltzen dute karaktere-kateen arteko distantzia neurtzeko, eta metrika berri horrekin Levenshtein distantziaren arabera proposatutako hautagaiak berrordenatzen dituzte modu erdi-gainbegiratuan.

Bollmann *et al.* (2011) lanean erregeletan oinarritzen den normalizazio-sistema bat proposatzen da XIV-XVI. mendeetako aleman historikoa normalizatzeko, baina kasu horretan erregelak automatikoki induzitzen dira corpus paralelo batetik abiatuta. Luteroren Bibliako bi bertsiorekin osatutako corpusa da erabiltzen dutena: 1545. urteko edizio bat, eta edizio moderno bat. Corpusean beharrezko lerrokatzeak egin eta gero hainbat tresna erabiliz (Gargantua eta GIZA++ aipatzen dituzte egileek), parekatutako hitz-bikotez osatutako corpus bat lortzen dute bertatik erregelak inferitzeko (parekatze horri esker ez da eskuzko lanik behar). Levenshtein distantzian oinarritzen den algoritmo bat gauzatzen dute zenbait erregela inferitzeko, eta maiztasunaren arabera ordenatzen dituzte erregela horiek. Erregelen aplikazioa ikasitako maiztasunaren arabera egiten da, eta sarrerako hitz batentzat

hitz asko proposa daitezkeenez, bakarrik onartzen dituzte Bibliako bertsio berrian dauden hitzak, besteak iragaziz. Emaitzetan ikusten denez, hitz normalizatuen portzentajea % 65etik % 91ra igotzea lortzen dute induzitutako erregelen bitartez. Ikusiko dugunez, III.3 atalean proposatuko dugun lehenengo metodoak antz handia du lan honekin.

Pettersson *et al.* (2014) lanean testu historikoen normalizazioa egiteko hiru metodo ebaluatzen dira eta esperimentuak bost hizkuntzatan egiten dira: ingelesa, alemana, hungariera, islandiera eta suediera.

Ebaluatutako hiru metodoetatik bi oinarrizkoak dira: Levenshtein distantzian oinarritzen da bata, eta bestea memorian oinarritzen da, hau da, ikasketan agertutakoa errepikatzen du. Levenshtein distantzian oinarritutako metodoa aurretik egindako lan batean deskribatzen dute xehetasunez (Pettersson *et al.*, 2013a), ebaluaziorako erabiliz Pettersson *et al.* (2012) lanean aipatutako *gold-standard* bera (erregeletan oinarritutako sistema da artikulu horretan landutakoa).

Teknika berritzaile gisa, hirugarren metodoan itzulpen automatiko estatistikoko teknikak (SMT) erabiltzen dituzte, baina karaktere-mailan (CSMT, *Character-level Statistical Machine Translation*). Teknika hori Pettersson *et al.* (2013b) lanean aurkezten dute lehenengo aldiz, eta lan horretan ebaluazioa islandiera zein suedierarekin egiten dute. 2014ko laneko ebaluaziotik ondorioztatzen da CSMT metodoa dela emaitza onenak lortzen dituena hizkuntza gehienetan (islandieran ez da hala gertatzen). Horrez gain, beste ondorio garrantzitsu bat da emaitzek erakusten dutela metodo horrekin ez dela behar datu askorik entrenatzeko, eta hori kontuan hartzekoa da datuak prestatu egin behar diren kasuetan.

Petterssonen lanarekin oso lotuta, Scherrer eta Erjavec (2015) lanean hizkuntzarekiko independentea den metodo bat proposatzen dute hitzak normalizatzeko eta eslovenierako hitz historikoak modernizatzeko atazaren bitartez ebaluatzen dute. Eskuz anotatutako bi corpus historiko erabiliz, esloveniera historikoko bi lexikoi disjuntu osatzen dituzte, non forma historikoak eta egungo formak erlazionatzen diren. Lexikoi batek 40.000 sarrera inguru ditu eta besteak 20.000 inguru.

Petterssonen lanean bezala, Scherrer eta Erjavec ikerlariek proposatzen duten normalizazio-metodoa karaktere-mailako itzulpen automatiko estatistikoan oinarritzen da (CSMT). Metodoak lortzen duen onura neurtzeko, bi oinarri-lerro planteatzen dituzte: lehenengoa eskuz idatzitako hainbat transliterazio-erregelatan oinarritzen da, eta bigarrena, Levenshtein distantzian oinarritzen da.

Esperimentuei dagokienez, bi motatako esperimentuak bideratzen dituzte bi eszenatoki ezberdin planteatuta. Lehenengo eszenatokiak plantea-

tzen du anotatutako corpusak erabilgarri daudela, eta hala, lexikoi handiena CSMT sistema eraikitzeko erabili eta gero, bigarren lexikoiarekin egiten dute ebaluazioa. Bigarren eszenatokiak, berriz, planteatzen du anotatutako informaziorik ez dagoela eskuragarri CSMT sistema eraikitzeko, eta, ondorioz, informazio hori 'sortu' beharra dago lehendabizi, gero CSMT sistema eraiki ahal izateko. Informazio hori sortzeko, corpus handieneko hitz historikoak hartzen dituzte eta antzekoak diren egungo hitzekin lotzen dituzte, antzekotasuna Levenshtein distantziaren arabera kalkulatuz.

Ebaluazioa hiru zatitan banatuta dago (ebaluatzeko lexikoia hala dagoelako banatuta, hitz historikoak ageri diren testuen garaiaren arabera) eta emaitza onenak lehenengo CSMT sistemak lortzen ditu, hau da, eskuzko anotazioan oinarritzen den sistema, beti gainditzen baititu bi oinarri-lerroak. Kalitate-parametro gisa, sistemak lortutako zehaztasuna (*accuracy*) ematen dute artikuluan. Parametro horrek adierazten du zenbatekoa den automatikoki modernizatutako hitzen portzentajea bat datorrena eskuz anotatutako forma modernoarekin. Parametro horren emaitza onenak ebaluazioko corpusaren hiru zatietarako honako hauek dira: % 68,5, % 79,3 eta % 86,6. Bigarren kapituluko II.5.2 atalean lan horretan erabiltzen duten corpusa deskribatu dugu, eta hurrengo kapituluan datu horiekin egingo dugu lan gure metodoaren emaitzak oraintxe aipatutako horiekin konparatzeko.

Bigarren eszenatokiko CSMT sistemaren emaitzek ez dituzte beti gainditzen oinarri-lerroko sistemetakoak baina gutxigatik, hiru zatietatik bitan gainditzen baitituzte. Dena den, oso eszenatoki interesgarria da, ez baitu ikasten anotatutako informazioarekin, baizik eta 'sortutako' informazioarekin.

Aurreko erreferentzia gehienak testu historikoekin erlazionatutakoak dira eta Europako hizkuntzekin egiten dute lan, baina arabieraren inguruan ere egin dira normalizazioari lotutako lanak. Horien artean dago Eskander *et al.* (2013) lana. Mundu arabiarra diglosiaren adibide prototipikoa da: ahozko hizkuntza natibo asko daude, dialekto asko, eta idatzizko ia arabiera bakarra, arabiera estandar modernoa deitutakoa izan da (MSA, *Modern Standard Arabic*). Tendentzia hori, ordea, aldatzen ari da, eta Internet bitartez gauzatzen den idatzizko komunikazio asko (posta elektronikoak, blogak eta abar) arabiera dialektalean egiten da gaur egun. Dialektoak ez direnez erabili idatzizko komunikazioan orain dela gutxi arte, ez daukate ortografia estandar bat definitua eta hizkuntza prozesatzeko tresnak garatzeko asmoz, CODA izeneko ortografia (*Conventional Orthography for Dialectal Arabic*) proposatu da arabiera dialektalez idazteko. Arazoa da idatzizko testu dialektalen ortografia oraindik espontaneoa dela eta ez duela CODA betetzen. Eskander *et al.* (2013) lanean planteatzen den ataza horixe da: arabiera dialektalaren

ortografia espontaneoa CODAra aldatzea.

Itzulpen hori egiteko hurbilpen bat baino gehiago proposatzen dute ikerlariek, eta emaitza onenak lortzeko hainbat teknika konbinatzen dituzte, hala nola karaktere-mailako sailkatzaileak, hitz-mailako aldaketak, eta analizatzaile morfologiko bat. Metodoaren berrikuntza nagusia karaktere-mailako sailkatzaileetan dago (*Character Edit Classification* deitua), non sailkatzaile bakoitzak ikasten duen posizio batean karaktere-aldaketa gerta daitekeen ala ez.

III.1 taulan aipatu berri ditugun lanen laburpena jasotzen da. Laburpen gisa esan daiteke aldaera fonologikoen ikasketa planteatzen den lanetan bi izan direla bide nagusiak oinarri-lerroez gain: (1) erregela fonologikoen indukzioa eta horien ordenaketa eta (2) kanal zaratatsuaren zenbait inplementazio: Markov-en eredu ezkutua (HMM), egoera finituko transduktore haztatuak (WFST) eta karaktere-mailako itzulpen automatiko estatistikoa (CSMT).

## III.3 Fonologiaren inferentzia

Normalizazioaren ataza ebazteko, hiru metodo aurkeztuko ditugu kapitulu honetan, hirurak egoera finituko teknologian oinarrituak (FST, *Finite-State Technology*). I.2 atalean aipatu dugu ikasketa automatikoko teknikak aplikatzea interesatzen zaigula, beraz, laneko corpusa bi zatitan banatu ondoren, ikasteko zatian gertatzen diren aldaketa fonologikoak inferitzen saiatuko gara metodo jakin bat aplikatuta, eta gero aldaketa horiek aplikatuko ditugu ebaluatzeko zatian. Aukeratutako hiru metodoen arteko ezberdintasun nagusia, beraz, ikasteko eta ebaluatzeko bi urrats horiek gauzatzen diren moduan datza, hau da, metodoak ezberdinak dira erabiltzen dituzten algoritmoetan edota tresnetan, bai ikasteko informazioa lortzeko, bai ikasitakoa aplikatzeko.

Lehenengo metodoa Almeida *et al.* (2010) lanean garatutako programa batean oinarritzen da. *Bigorna* izeneko tresna-multzoa aurkezten dute lan horretan, zein portugesez idatzitako dokumentuen ortografia automatikoki eguneratzen laguntzen duen tresna den. Akordio garrantzitsu bat hartu zen 1990ean portugesaren ortografiaren inguruan, hizkuntza hori erabiltzen duten zenbait herrialde hurbiltzeko asmoz (Angola, Brasil, Cabo Verde, Ginea Bissau, Portugal...), eta akordioaren eraginez, hainbat eta hainbat testuren ortografia "zaharkitua" geratu zen. Hori izan da Bigorna izeneko tresna-multzoa garatzeko testuingurua. Zenbait baliabide eta aplikazio hartzen ditu tresnak bere barnean, eta horien artean *lexdiff* izeneko programa dago,

| **Erreferentzia** | **Metodo nagusia(k)** | **Sailk.** | **Testu mota** | **Hizkuntza** |
|---|---|---|---|---|
| Mann & Yarowsky (2001) | Edizio-distantzien pisuak ikasi eta Eredu Markoviarra (HMM) | (1) | Beste | Gaztelania-Portugesa<br>Frantsesa-Portugesa<br>Italiera-Portugesa<br>Errumaniera-Portugesa |
| Scherrer (2007) | Edizio-distantzien pisuak ikasi eta eskuz idatzitako erregelak | (1) | Dialektala | Suitzako alemana-Egungo alemana |
| Kestemont *et al.* (2010) | MBL (Memory-based learning) | – | Historikoa | Nederlandera |
| Bollmann *et al.* (2011) | Erregelen ordenaketa (maiztasuna) | (1) | Historikoa | Alemana |
| Pettersson *et al.* (2014) | CSMT (Character-level Statistical MT) | (2) | Historikoa | Ingelesa<br>Alemana<br>Hungariera<br>Islandiera<br>Suediera |
| Scherrer & Erjavec (2015) | CSMT (Character-level Statistical MT) | (2) | Historikoa | Esloveniera |
| Eskander *et al.* (2013) | Ikasketa automatikoa karaktere-mailan (WEKA, k-NN) | – | Dialektala | Arabiera |

III.1 Taula: Aldaera fonologikoak ikasten duten zenbait lanen laburpena.

gai dena bi testu konparatzeko, eta bertan dauden karaktere-kateen arteko aldaketak detektatzeko. Programak proposatzen dituen aldaketa horiek memorizatuz gero, aukera daukagu aldaerei aplikatzeko, ea horrela lortzen den dagokien estandarra.

Proposatzen dugun lehenengo metodo horrek antza handia du Bollmann *et al.* (2011) lanarekin, baina gure metodoan corpuseko hitzak parekatzeko eta maiztasunak kalkulatzeko, Almeida *et al.* (2010) lanean proposatutako programa erabiltzen dugu, eta ez GIZA++ (Bollmann *et al.* (2011) lanean erabilitakoa).

Bigarren metodoaren oinarrian *Inductive Logic Programming* (ILP) (Muggleton eta De Raedt, 1994) motako ikasketa-algoritmo bat inplementatu dugu. Algoritmoak eraldaketa fonologikoak adierazten dituzten erregelak ikasten ditu hainbat hitz-bikote abiapuntutzat hartuta. Helburua da ikasteko datuekin bateragarria den erregela multzo minimoa lortzea, hau da, beharrezkoak zein nahikoak diren erregelak ikastea.

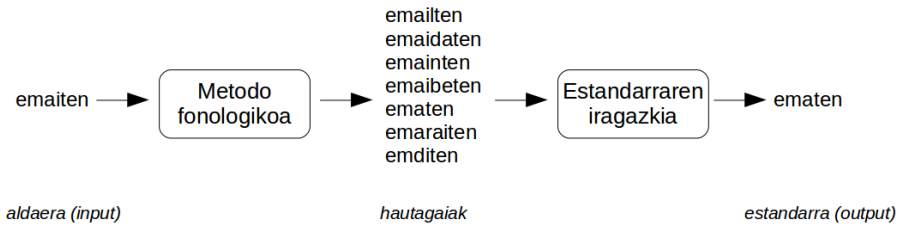
Guk dakigunaren arabera, ILP motako algoritmoa planteatzea erregelak ikasteko berria da alor honetan, eta bibliografian aurkitu ditugun lanen artean antza gehien duena Kestemont *et al.* (2010) lana da.

Hirugarren metodoak WFST teknologia erabiltzen du *Phonetisaurus* (Novak *et al.*, 2012) tresnaren bitartez. Tresna fonologikoa da Phonetisaurus, grafemen eta fonemen arteko bihurketa lortzeko garatu den tresna (G2P, *Grapheme-to-Phoneme* edo P2G). Tresna inplementatzen da WFST (*Weighted Finite-State Technology*) teknologiaren bitartez eta horretarako, OpenFst[2] liburutegia erabiltzen du (Allauzen *et al.*, 2007). Gure lanean planteatzen den bihurketa ez da grafemen eta fonemen artekoa, grafemen artekoa baizik, baina ikusiko dugun moduan, tresna erabilgarria da planteatzen dugun ataza ebazteko.

Hirugarren metodo horrek, beraz, hizketaren prozesaketan ohikoa den kanal zaratatsuaren (*noisy channel*) inplementazio batean oinarritzen da. Azterketa bibliografikoan aipatu diren CSMT eta HMM teknikak ere ideia berean oinarritzen dira, eta, esan bezala, hurrengo kapituluan alderatuko ditugu gure hirugarren metodo horren emaitzak eta Scherrer eta Erjavec CSMT bitartez lortzen dituztenak.

Proposatzen ditugun hiru metodoetan ikasten diren ereduak FST teknologian biltegiratzen dira, hau da, transduktoreen bitartez adierazten dira, eta gero, transduktore horien gainean egiten da deskodeketa, sarrerako aldaera berriak normalizatzeko helburuarekin. Lehen bietan transduktoreak itzultzaile hutsak dira, ez dute probabilitate edo kostua adierazten. Hiruga-

[2] `www.openfst.org` (2016-02-24an atzitua)




III.1 Irudia: Hautagaien lehenengo iragazkia. Erregelak aplikatuta proposatzen diren hautagaien artetik ez-estandarrak kentzen dira.

rrenean, berriz, pisuak lortzen dira eta aukera bat baino gehiago dagoenean, horiek ordenatuta eman daitezke.

Hiru metodoen helburua berbera da beti: ikasteko corpusetik informazioa lortzea, ebaluazio-corpuseko aldaerak normalizatzeko. Ebaluazioan, oro har, aplikatutako metodoak hainbat hautagai proposatuko ditu, eta beharrezkoa izango da horiek iragaztea. Iragazki ohikoena zenbait hautagai proposatzen dituzten sistemetan, helburu-hizkuntzaren eredu bat erabiltzea izaten da: hitz bateko eredua (*unigram*), bi hitzekoa (*bigram*) eta abar. Hizkuntza-eredu hori ezaugarri jakin baten arabera definitzen da, hala nola, hitzen maiztasuna, hitzen probabilitatea, hitz zuzenak izatea eta abar.

Lan honetan, metodoek proposatzen dituzten hautagaiei euskara estandarraren hitz bateko hizkuntza-eredua aplikatuko diegu iragazki gisa. Lehendabizi, iragazkiak kendu egingo ditu hautagaien artetik hitz estandarrak ez direnak, eraikitzen ari garen sistemaren helburua baita aldaerei hitz estandarrak esleitzea. III.1 irudian lehenengo iragazkiak lortzen duena islatu nahi izan da eskematikoki. Adibide horretan ikusten denez, `emaiten` aldaera normalizatzeko hainbat hautagai sortu ditu metodo batek, eta estandarrak ez direnak iragazi ondoren, hautagai bakarra geratzen da, `ematen`. Beti ez da aukera bakarra geratuko, eta horregatik, iragazki hori nahikoa ez den kasuetan, aukera izango dugu bigarren iragazki bat aplikatzeko: hautagaien artean maiztasun handienekoa soilik uzten duen iragazkia. Bigarren iragazki hori, dena den, aukerazkoa izango da.

Lehenengo iragazkia aplikatzeko, hizkuntza estandarraren analisi morfologikoa gauzatzen duen transduktore bat erabiliko dugu (Alegria *et al.*, 2009b), eta bigarren iragazkia aplikatzeko, berriz, Aduriz *et al.* (2006) lanean sortutako egungo hitz zerrenda, maiztasunaren arabera ordenatuta[3].

[3]Lan horretan *Euskaldunon Egunkaria*ren corpus handi bat prozesatu zen eta maizta-

Hautagaiak proposatzen dituzten metodoen xehetasunak eman baino lehen, lehenengo bi metodoen inguruko argibide bat eman behar dugu gero hobeto ulertzeko haien funtzionamendua. Bi metodo horien helburua bera da: karaktere-kateen arteko eraldaketak egiten dituzten erregelak ikastea datuetatik (modu batera edo bestera, berehala ikusiko dugunez), eta gero erregela horiek aplikatzea aldaerei, horrela aurkitzeko dagozkien hitz estandarrak. Ikasitako erregelak ordezkapen-erregela fonologikoen formatuaren bitartez adieraziko ditugu (Beesley eta Karttunen, 2003), eta era horretan egoera finituko transduktoreetan konpilatu ahal izango ditugu *foma* aplikazio librearen bitartez (Hulden, 2009). Erregela horiek egoera finituko transduktoretan konpilatzeko arrazoiak honako hauek dira: (1) transduktoreen bitartez erraz eta azkar prozesatzen dira sarrerako datuak (transduktoreekin lan egiteko garatuak diren zenbait aplikaziori esker); (2) sortuko dugun transduktorea beste transduktore garrantzitsu batekin konposatzeko aukera daukagu, hizkuntza estandarraren morfologia jasotzen duen transduktorearekin, hain zuzen (Alegria *et al.*, 2009b), eta horrela lehen aipatu dugun lehenengo iragazkiaren aplikazioa berehalakoa izango da; (3) aurretik egin ditugun lanetan eskarmentua hartu dugu transduktoreekin lan egiten (Etxeberria *et al.*, 2011) eta hori garrantzitsua da lana aurrera eramateko.

### III.3.1 Erregela fonologikoak

Erregela fonologikoak ordezkapen-erregelak dira, hau da, testuinguru jakin baten arabera karaktere-kateen arteko ordezkapenak adierazten dituzten erregelak. Johnson (1972) izan zen lehenengoa azaltzen erregela horiek transduktore moduan adieraz zitezkeela, baina denbora luzea igaro zen erregelak transduktoreetan konpilatzeko lehen algoritmoak garatu arte (Kaplan eta Kay, 1994).

Erregela fonologiko baten formatu sinplifikatua honako hau da:

$$A \rightarrow B \,||\, C _ \; D \tag{III.1}$$

non $A$,$B$,$C$,$D$ argumentuak sinbolo bakunak edo sinbolo-kateak diren. Halako erregela batek adierazten du $A$ katea $B$ katean bihurtzen dela baldin eta $A$ katea $C$ eta $D$ kateen artean suertatzen bada. Testuinguruko argumentuak, $C$ eta $D$, ez dira derrigorrezkoak eta ageri badira, aukera dago baldintza anitzak adierazteko. Esaterako, honako erregela honek:

```
h -> 0 || p _ , t _ , l _ , _ a s o
```
(III.2)

sunaren arabera ordenatutako 1,1 miliotik gorako forma gordetzen dituen fitxategia lortu zen.

`h` karakterearen ezabatzea adierazten du, honako testuinguru hauetan: `h`-ren aurretik `p`, `t`, edo `l` karakterea ageri bada edo `h`-ren atzetik `aso` katea ageri bada[4]. Adibidez, erregela horrek bihurtuko luke `ongiethorri` hitza (lapurteraz) `ongietorri` hitzean (estandarra).

Hainbat erregela ikas daitezke sarrerako datuetan oinarriturik, eta erregela horiek zenbait karaktere-kate ezberdinei eragin diezaiekete, beraz, erabaki beharra dago zein izango den erregelak aplikatzeko modua sarrera berrietan. Ikasitako erregelak sekuentzialki aplika daitezke, hau da, bata bestearen atzetik ordena zehatz batean, edo paraleloan, hau da, aldi berean.

Esaterako, honako bi erregela hauek:

```
u -> i || z a _
```
(III.3)

```
k -> g || z a u _
```
(III.4)

modu paraleloan aplikatzen badira, `zaukun` hitzetik `zaigun` hitza sortuko da. Aldiz, erregelak modu sekuentzialean aplikatzen badira, eta esaterako, `u` → `i` erregela aplikatzen bada lehendabizi, bigarren erregela ez da aplikatuko: lehenengo erregela aplikatuta `zaikun` hitza lortzen da, eta horren ondorioz, bigarren erregela ezin da jadanik aplikatu, ez delako haren testuingurua betetzen: `k` aurretik ez dago `zau` katea.

Erregela multzo batek lortzen dituen emaitzak aztertzeko, bi aplikazio-moduak ebaluatu beharko dira, haien artean diferentziak nabarmenak diren ala ez aztertzeko.

Erregela fonologikoen inguruan gehiago sakontzeko, Beesley eta Karttunen (2003) eta Hulden (2009) lanak kontsulta daitezke.

Erregela fonologikoak nolakoak diren ikusi ondoren, proposatzen ditugun hiru metodoen funtzionamendua aztertuko dugu ondorengo ataletan.

### III.3.2 Lehenengo metodoa: lexdiff

Lehenengo metodoaren oinarrizko ideia sinplea da: identifikatu corpus paralelo batean bat ez datozen hitzen artean gertatzen diren karaktere-kateen arteko aldaketak, eta idatzi hainbat erregela fonologiko aldaketa horiek jasotzeko.

Detekzio hori egiteko, lexdiff izeneko programa erabil daiteke, Bigorna izeneko tresna-multzoaren barruan garatua (Almeida *et al.*, 2010). Sarreran aipatu den moduan, Bigorna tresna-multzoaren helburuen artean dago portugesez idatzitako testuak automatikoki migratzea hizkuntzaren ortografia bateratura, eta hainbat programa eta baliabide garatu dituzte horretarako.

[4] Erregela bera formatu trinkoagoan: `h -> 0 || [p | t | l] _ , _ a s o`

lexdiff programa gai da corpus paralelo batean gertatzen diren diferentziak identifikatzeko. Unixeko *egrep* eta *diff* komandoetan oinarritzen da eta bi maila ezberdinetan detekta ditzake diferentziak: hitz-mailan edota hitzen barneko karaktere-mailan. Kasu batean zein bestean, programak parekatu egiten ditu hitzak edo karaktere-sekuentziak, eta hala, parekatzeak kontatuz gero, bakoitza zenbat aldiz suertatu den corpusean jakin daiteke.

Esaterako, lexdiff programari ematen badiogu sarrera gisa gure corpus paraleloaren ikasteko zatia (aldaeraz idatzitako esaldiak alde batetik, eta estandarrez idatzitakoak bestetik), bi emaitza ezberdin lortuko ditugu aukeratzen dugun parekatze-mailaren arabera:

- hitzen arteko parekatzea aukeratuz gero, eta ondoren kontaketa eginez, parekatutako hitz-bikote guztiak izango ditugu haien maiztasunarekin (bikoteko hitzak berdinak ala ezberdinak izan daitezke):

  ```
  152 bat = bat ; 104 bere = bere ;
  61 emaiten => ematen ; 15 aphez => apez ;  ...
  ```

- karaktereen arteko parekatzea testuinguruaren arabera aukeratuz gero, eta horren ondoren kontaketa eginez, karaktere-sekuentzia ezberdinen arteko parekatzeak izango ditugu haien maiztasunarekin (kasu honetan, sekuentzia ezberdinak soilik lortzen dira):

  ```
  76 ait => at ; 39 dautz => diz ; ...
  ```

Hurrengo III.3.3 eta III.3.4 ataletan ikusiko dugunez, lexdiff programarekin lortutako hitzen arteko parekatze hori, proposatuko ditugun beste bi metodoen abiapuntua izango da, beharrezkoa baita corpus paraleloa parekatzea ikasten hasi baino lehen. Azterketa bibliografikoan aipatu dugunez, Bollmann *et al.* (2011) lanean GIZA++ erabiltzen da parekatze hori egiteko, eta gure kasuan lexdiff programa erabili dugu.

Karaktere-mailako bigarren parekatzearen bitartez, berriz, bi testuen arteko aldaketa fonologiko erregularrak lortzen ditugu haien maiztasunarekin batera, eta informazio hori baliatuta, ordezkapen-erregelak idatz ditzakegu eta egoera finituko transduktore batean konpilatu foma[5] aplikazio librearen bitartez (Hulden, 2009). Transduktorea sortzeko, dena den, hainbat irizpide finkatu behar dira kontuan izanik corpusean aurkitu diren aldaketen informazioa, sortutako erregelen formatua, horien aplikazio-modua eta abar. Irizpideak aldatuz transduktore ezberdinak lortzeko aukera dago, eta aztertu egin beharko da zein den emaitza onenak lortzen dituen transduktorea. Esaterako:

[5] `code.google.com/archive/p/foma/` (2016-02-28an atzitua)

- Erregela kopurua muga daiteke kontuan hartuta aldaketen maiztasuna, hau da, aldaketa bat ez bada detektatu gutxienez $n$ aldiz corpusean, ez da sortzen horri dagokion erregela. Adibidez, muga horri 3 balioa emanez gero, aplikatuko diren ordezkapen-erregelen aldaketak, corpusean gutxienez 3 aldiz detektatukoak izango dira.

- Muga daiteke zenbat erregela aplika dakizkiokeen hitz berari. Gerta daiteke lexdiff aplikazioak bi karaktere-aldaketen bitartez adieraztea hitz-bikote batean dauden diferentziak: `agerkuntza => agerpena` bikoteari lotuta, esaterako, `rkun => rpen` eta `ntza => na` karaktere-aldaketak proposatzen ditu; hau da, bi erregela ezberdin. Ez badira bi erregelak aplikatzen, kasu horretan ez da lortuko bi hitz horien arteko aldaketa zuzena.

  Beraz, hitz berari aplika dakizkiokeen erregela kopurua mugatuz gero, irteera kopurua mugatzen da eta hori, alde batetik, interesgarria izan daiteke emaitza desegokiak ekiditeko, baina beste aldetik, gerta daiteke ez sortzea desiragarria litzatekeen irteera hori.

- Kontrola daiteke erregelak aplikatzeko modua: sekuentziala edo paraleloa. Adibideko bi erregelak, `rkun -> rpen` eta `ntza -> na`, paraleloan aplikatzen badira, `agerkuntza` hitzarako lortuko den irteerako forma ez da zuzena izango, ez baitira biak aplikatuko $n$ letrak bi erregelak gainezartzen dituelako (`agerpentza` eta `agerkuna` sortuko dira) Aldiz, bi erregela horiek sekuentzialki aplikatzen badira (berdin dio zein ordenatan) irteera zuzena izango da: lehendabizi `rkun -> rpen` aldatzen da eta gero `ntza -> na` (edo aldrebes). Zaila denez aurretik jakitea zein den onena, biak ebaluatu beharko dira esperimentuetan.

- Aukera dago lexdiff-ek proposatutako erregelak trinkotzeko testuingurua kontuan hartzen duten erregelak idatziz, eta horrela, erredundantziak kentzen dira. Esaterako, `rkun -> rpen` erregelaren ordez, testuingurua zehazten duen erregela berri bat idatz daiteke, `ku` katea `pe` katea bihurtuko duena, baldin eta ezkerrean `r` badago eta eskuinean `n`. Honela idatziko litzateke erregela berria:

  $$\texttt{k u -> p e} \;||\; \texttt{r _ n} \tag{III.5}$$

  Erregela-trinkotzeak eragina du aurretik aipatu berri den puntuan. Izan ere, erregelen arteko karaktere-gainezarpena gutxiagotan gertatuko da trinkotzearen eraginez, eta hala, hitz baten berridazketa gehiago ahalbidetuko dira erregelak paraleloan aplikatzen direnean.

Aipatu berri diren irizpide horien arabera, erregela multzo bat sortzen da, transduktore jakin batean konpilatzen dena. Transduktorearen bitartez posible da erregelak horiek sarrerako aldaera berriei aplikatzea, irteeran hainbat hautagai lortzeko.

Hurrengo urratsak iragazkiei dagozkie, III.3 atalaren sarreran esan dugunaren arabera. Lehenengo metodo honekin lortutako hautagaiei bi iragazkiak aplikatuko dizkiegu: lehendabizi hitz ez-estandarrak kenduko dira proposatutako hautagaien artetik (horretarako hizkuntza estandarraren transduktorea konposatuko da metodoak lortu duen transduktorearekin); gero, bigarren iragazki gisa, hautagai bat baino gehiago dituzten sarreren kasuan, hitzen maiztasuna kontuan hartzen duen iragazkia aplikatuko da.

### III.3.3 Bigarren metodoa: ILP motako algoritmoa

Proposatzen dugun ILP motako algoritmoa[6] erabiltzeko, hitz-bikoteak behar dira sarrera gisa, hau da, aldaerarekin eta hitz estandarrarekin osatutako hitz-bikoteak, eta bikote horiek lortzeko corpus paraleloan, lexdiff programa erabili dugu (hitz-mailako parekatzea eginez). ILP algoritmoak bikoteen arteko aldaketak aztertzen ditu, eta ordezkapen-erregela multzo minimo bat aurkitzen saiatzen da aldaketa horiek guztiak kontuan hartuta. Inplementatutako hurbilpena bat dator ILP motako ikasketa automatikoko metodoekin (Muggleton eta De Raedt, 1994), baina ILP estandarrarekin alderatuta, guk ez ditugu lehen mailako predikatu logikoak lortu irteera gisa, karakterekateren arteko ordezkapen-erregelak baizik.

Hauek dira algoritmoak jarraitzen dituen urratsak erregelak ikasteko:

(1) Lerrokatu karakterez-karaktere hitz-bikote guztiak (edizio distantzia minimoaren bitartez).

(2) Erauzi ordezkapen-erregela multzo bat.

(3) Erregela bakoitzerako, bilatu kontraadibideak.

(4) Erregela bakoitzeko, aurkitu zein den bete beharreko testuinguru minimoa ziurtatzeko erregela aplikatzen dela aldeko adibide guztietan, eta ez dela aplikatzen adibide negatiboetan. Mugatu erregela testuinguru horretan soilik aplika dadin.

[6] Lan honetan zehar lankidetza paregabea jaso dugu Mans Hulden ikerlariarengandik, eta ILP algoritmoa berak idatzitakoa da.

Ondoko adibideak algoritmoak jarraitzen duen metodoa hobeto ulertzen lagunduko digu. Demagun bi hitz-bikote besterik ez dituen corpus bat dugula:

| | |
|---|---|
| *emaiten* | *ematen* |
| *igorri* | *igorri* |

(1) urratsean lerrokatzea egiten da eta honako irteera lortzen da:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| e | m | a | i | t | e | n | | i | g | o | r | r | i |
| e | m | a | ∅ | t | e | n | | i | g | o | r | r | i |

Datu horiekin, (2) urratsean ondorioztatzen da erregela fonologiko bakarra behar dela $i$ sinboloa ezabatzeko, `i -> 0` erregela, gainontzeko sinboloak ez baitira aldatzen. Kontrako adibideak bilatzerako orduan, (3) urratsean, bi kontrako adibide aurkitzen ditugu, zeren *igorri* adibidean bi $i$ daude eta ez dira ezabatu behar. Hala, bete beharreko testuinguru motzena datuak modelatzeko eta gainsorkuntza ekiditeko, karaktere batekoa da. Adibidez:

```
i -> 0 || a _
```
(III.6)

Erregelaren testuinguruak adierazten du $i$ letraren ezkerrean $a$ letrak egon behar duela erregela aplika dadin. Garbi dago, ordea, beste aukera bat dagoela karaktere horren testuinguru minimoa adierazteko, eta hori da $i$ren eskuinean $t$ izatea, hau da:

```
i -> 0 || _ t
```
(III.7)

Bi testuinguruak luzera berekoak dira eta halako kasuetan hartutako irizpidea izan da ezker aldeko testuingurua zehaztea. Erabaki arbitrarioa da (nahiz eta oinarri fonologikoa duen) eta oso antzeko emaitzak lortzen dira alde bateko zein besteko testuingurua zehaztuta.

Laburbilduz, ILP algoritmoak datuetatik inferitu behar dituen parametroak bi dira: (1) ikasteko datuetan ageri diren karaktere-kateen arteko $X \rightarrow Y$ ordezkapenak adierazten dituzten erregelak, eta (2) erregela bakoitzaren testuinguru minimoa, ziurtatzeko ikasteko datuetan dagoen informazio guztia (ez karaktere-kateen arteko aldaketak soilik) ondo jasotzen dela. ILP metodoarekin ikasten diren erregela guztiak aldi berean (modu paraleloan) aplikatu behar dira, gero, adibide berrietan.

Lehenengo metodoan egin den bezala, ILP metodoak lortutako erregelak transduktore batean konpilatu dira foma aplikazioaren bitartez, eta transduktoreak sortutako hautagaiei aurreko bi filtro berberak aplikatu zaizkie:

hitz ez-estandarrak kentzen dituen iragazkia (hizkuntza estandarraren transduktorea eta sortutako transduktorea konposatuz gauzatzen dena) eta hitzen maiztasuna kontuan hartzen duen bigarren iragazkia.

#### Kate-kate vs. sinbolo-sinbolo erregelak

Zenbait hitz-bikotetan aldaketak ondoz ondoko sinboloen artean gertatzen dira, esaterako *daut* – *dit* bikotean:

| | | | |
|---|---|---|---|
| d | a | u | t |
| d | i | ∅ | t |

Erregelak adierazteko erabili den formalismoak ez ditu ordezkapen-erregelak murrizten sinbolo kopuru aldetik, eta beraz, bi aukera daude bikote horretan gertatzen den aldaketa adierazteko. Hala, sinbolo anitzen arteko ordezkapen erregela bakarra idatz daiteke:

`au -> i` || *testuingurua*

edo sinbolo sinpleen arteko bi erregela idatz daitezke:

`a -> i` || *testuingurua*
`u -> 0` || *testuingurua*

non *testuingurua* atalak adierazten duen gainontzeko datuen arabera bete beharreko testuinguru minimoa. Bi aukerak ebaluatu ditugu eta emaitzetan ez dago haien arteko alde esanguratsurik.

### III.3.4 Hirugarren metodoa: WFST teknologia

Lehenengo metodoan sortzen diren mugak saihestu aldera, hau da, erregelen arteko ordenaketa eta haien aplikazio-modua (sekuentziala zein paraleloa), hirugarren metodo honetan WFST teknologian oinarritu gara, transduktore haztatuak erabiltzen dituen teknologian. Egoera finituko transduktore haztatuetan (WFST transduktoreak) haztak edo kostuak hartzen dira kontuan trantsizioak egiteko orduan, eta teknologia malguagoa bilakatzen da horrela. Oso erabilia da kanal zaratatsuaren moduko aplikazioak inplementatzeko (*noisy channel*), esaterako, hizketaren prozesamenduaren arloko aplikazioetan.

Teknologia hori erabiltzeko lehen saioa Carmel[7] tresna (Graehl, 1997) baliatzea izan zen baina ez genuen emaitza egokirik lortu, transduktorea

[7] `www.isi.edu/publications/licensed-sw/carmel`

ondo modelatzeko prozesua nahiko konplexua baitzen. Abiapuntuko transduktorea inferitu behar zen lehendabizi corpusetik erauzitako hitz-bikotetan oinarrituta, eta gero transduktore hori entrenatu pisuak lortzeko. Zenbait proba egin ondoren tamaina ezberdineko zenbait transduktorerekin eta entrenatzeko datu kopuru ezberdinekin, emaitzak ez ziren onak eta beste bide bat bilatu genuen.

Irtenbide gisa, tresna berriago batekin probatzea erabaki genuen eta aukeratutakoa Phonetisaurus[8] tresna-multzoa izan zen. FSMNLP2012 (*Finite-State Methods and Natural Language Processing*) *workshop*ean aurkeztutako tresna da (Novak *et al.*, 2012), grafemen eta fonemen arteko bihurketak gauzatzeko sortua da (G2P, *Grapheme-to-Phoneme*) eta WFST teknologia erabiltzen du. Kode irekian zabaldutako tresna da, BSD lizentziapean.

Tresnaren funtzionamendua hiru urratsetan banatzen da behin ikasteko datuak prestatu ondoren (Novak *et al.*, 2012). Lehenengo biak ikasketarako urratsak dira eta hirugarrena, berriz, ikasitakoa aplikatzeko urratsa da (deskodeketa-urratsa deritzo):

1. Datuen lerrokatzea. Ikasteko hiztegian hainbat bikote daude, zeinetan lehen osagaia grafemaz osatutako adierazpena den eta bigarrena, berriz, fonemaz osatutakoa. Bi osagai horien arteko lerrokatzea egin behar da lehen urrats honetan, hau da, haien sinboloen arteko lerrokatzea, eta horretarako erabiltzen den algoritmoa Jiampojamarn *et al.* (2007) lanean oinarritzen da.

2. Ereduaren entrenamendua. Lerrokatzea eginda hiztegiaren ikuspegia aldatzen da, lerrokatzez osatutako hiztegi berri bat baita, eta horrekin $n$-gram motako eredu bat entrena daiteke. Eredua entrenatzeko, hizkuntza modelatzeko hainbat aplikazio erabil daitezke (OpenGrm Library, MITLM, SRILM eta abar). Erabiltzen dena erabiltzen dela, lortutako eredua transduktore haztatu gisa adierazten du Phonetisaurusek bukaeran.

3. Deskodeketa. WFST transduktoreari sarrera gisa hitz bat ematen badiogu, hau da grafema-kate bat, hitz horri dagokion probabilitate handieneko fonema-katea lortzen da azken urrats honetan. Aukera dago probabilitate handieneko $m$ kateak eskatzeko.

[8] `https://github.com/AdolfVonKleist/Phonetisaurus`

### Phonetisaurusen erabilpena gure testuinguruan

Nahiz eta grafema-fonema bihurketa egiteko tresna izan, guk ez dugu horretarako erabili, gure kasuan bihurketa grafemen artekoa baita: aldaerak ditugu sarrera gisa, eta horiei dagozkien hitz estandarrak lortu nahi ditugu irteeran. Erabilera horretan aldatzen den gauza bakarra karaktere-sekuentziak dira, hau da, tresnari ematen zaizkion ikasteko datuak: sarrera bakoitzaren bi osagaiak hitzak dira, eta ez hitza eta fonema-katea. Gainontzeko urratsak, tresnaren erabilerak eskatzen dituen urratsak dira, aurreko atalean aipatu ditugunak, hain zuzen. Phonetisaurus tresna-multzoaren erabiltzeko beste hainbat tresna behar badira ere (OpenFst liburutegia eta hizkuntza modelatzeko aipatutakoen arteko aplikazioren bat) erabilera aldetik tresna sinplea da erabiltzailearentzat, orain ikusiko dugun moduan.

Tresna erabili ahal izateko, lehendabizi ikasteko datuak prestatu behar dira, hau da, ikasteko hiztegi bat prestatu behar da, gure kasuan aldaera–estandar hitz-bikotez osatua. ILP metodoan bezala (gure bigarren metodoa) hitz-bikote horiek lortzeko corpus paralelotik, lexdiff programa erabili dugu. Gogoratu III.3.2 atalean esan duguna: hitz-mailako parekatzea eskatuz gero, lexdiff programak hitz berdinez zein ezberdinez osatutako bikoteak itzultzen ditu:

```
bat = bat ; bere = bere ; emaiten => ematen ; aphez => apez ...
```

Ikasteko hiztegia prestatu eta gero, lehen aipatutako bi urratsak jarraitu behar dira ikasteko eta WFST transduktorea lortzeko:

1. Lehenengo urratsean hiztegiko bikoteen arteko lerrokatzea egin behar da (`phonetisaurus-align` programa dago horretarako).

2. Bigarren urratsean $n$-gram eredua entrenatu behar da hizkuntza modelatzeko edozein tresnaren bitartez, eta gure aukera OpenGrm Library[9] erabiltzea izan da (Roark *et al.*, 2012). Tresna horrek lortutako eredua ARPA formatuan dagoenez, bukaeran formatuen arteko bihurketa bat egin behar da WFST transduktorea lortzeko (bihurketa egiteko programa `phonetisaurus-arpa2fst` da).

Transduktorea lortu ondoren, deskodeketa-urratsa dator, hau da, aldaera berriak ematen zaizkio transduktoreari, dagokien erantzun onena eman dezan (`phonetisaurus-g2p` programa erabili behar da horretarako). Horrela lortzen da ikasitako kostuen araberako erantzun edo hipotesi onena. Aukera dago $m$ erantzun onenak eskatzeko programari, eta hala eginez gero, kostuaren arabera ordenatzen ditu erantzunak (onenetik, txarrenera).

[9] `http://www.openfst.org/twiki/bin/view/GRM/NGramLibrary` (2016-03-15ean atzitua

Hautagaiak lortu ondoren, iragazkiak aplikatzeko unea da eta aurreko bi metodoetan egin den antzera, hitz bateko hizkuntza-eredua erabili da horretarako. Lehenengo iragazkia aurreko bi metodoetan aplikatutako berbera da, hau da, hitz ez-estandarrak ez diren erantzunak kendu dira hizkuntza estandarraren transduktorea erabiliz. Bigarren iragazkia, berriz, aldatu egin dugu metodo honetan: transduktore haztatuak ematen dituen erantzunak haztaren arabera ordenatuta daudela baliatuta, ordena horren araberako erantzun onena aukeratu da (ez da maiztasunaren iragazkia erabili).

## III.4 Esperimentuak

Kapituluaren sarreran esan den moduan, aurkeztutako hiru metodoen eraginkortasuna aztertzeko hainbat esperimentu egin dira II. kapituluko II.4 atalean deskribatutako lapurtera/estandarra corpus paraleloaren gainean. Corpus historikoak prestatzen ari ginen bitartean, hori zen eskura genuen corpusa eta horrekin egin genituen atal honetan deskribatu behar ditugun esperimentuak.

### III.4.1 Esperimentuen diseinua

III.3 atalean deskribatu diren hiru metodoak lapurtera/estandarra euskarazko corpus paraleloan aplikatu ditugu, lortzen dituzten emaitzak konparatu ahal izateko. Gogora ditzagun corpus paralelo horren ezaugarri nagusienak (bigarren kapituluan deskribatu da corpusa II.4 atalean): 2.117 esaldi paralelo dituen corpusa da, eta bi zatitan banatu dugu, zoriz, esperimentuak egiteko: corpusaren % 80 (ia 1.700 esaldi) ikasteko erabili da, eta gainontzekoa, % 20 (423 esaldi), ebaluaziorako. Corpusaren bi aldeetako hitzak parekatu dira lehendabizi III.3.2 atalean aipatutako lexdiff programarekin, eta horrek itzuli dituen bikote batzuk kendu dira metodoak aplikatu baino lehen (esaterako, hitz anitzeko parekatzeak dituzten bikoteak[10]). Gero, hitz guztiak letra xehez idatzi dira eta bikote guztien hitz estandarra analizatu da transduktore estandarraren bitartez: hitz estandarra onartua izan ez den kasuetan, bikote hori ere kendu egin da prozesutik[11].

III.2 taulan corpusaren ezaugarriak laburbildu dira[12]. Ikasteko zatian

[10] Oso gutxi dira hitz anitzeko parekatzeak dituzten bikoteak, 5 bikote corpus osoan.

[11] Hitzak letra xehez idatzi direnez, iragazki horrek izen propioei dagozkien parekatzeak kendu ditu, batez ere.

[12] Gogoratu *token* terminoaren bitartez adierazten dugula hitz bakoitza kontatzen dela testuan ageri den bezainbeste aldiz, eta *forma* terminoaren bitartez, berriz, hitz bakoitza behin bakarrik kontatu dugula.

| | Corpus osoa | Ikasi % 80 | Test % 20 |
|---|---|---|---|
| **Esaldiak** | 2.117 | 1.694 | 423 |
| **Tokenak** | 12.150 | 9.734 | 2.417 |
| **Formak** | | | |
| Lapurtera | 3.830 | 3.292 | 1.239 |
| Estandarra | 3.553 | 3.080 | 1.192 |
| **Hitz-bikoteak** | 3.610 | 3.108 | 1.172 |
| Bikote berdinak | 2.532 | 2.200 | 871 |
| Bikote ezberdinak | 1.078 | 908 | 301 |

III.2 Taula: Lapurtera/Estandarra corpus paraleloaren hainbat kopuru. Hitz-bikoteak lexdiff programaren bitartez lortzen dira eta gero hainbat bikote iragazten dira. Iragazitakoen artean, gehienak izen propioak dira.

3.108 hitz-bikote geratu dira (2.200 bikotetan hitzak berdinak dira eta 908tan ezberdinak) eta testeko zatian 1.172 bikote (871 bikotetan hitzak berdinak, eta 301etan ezberdinak). Testeko zatiari dagokionez, metodo guztien ebaluazioa egiteko zati horren bikote ezberdinak soilik erabili dira, hau da, 301 bikote ezberdin horien gainean egin da ebaluazioa.

Proposatutako hiru metodoen kalitatea neurtzeko, ohikoak diren hiru parametro erabili ditugu ebaluazioan: doitasuna ($P$, *precision*), estaldura ($R$, *recall*) eta aurreko bien batezbesteko harmonikoa, $F$ neurria ($F_1$-*score*). Parametro horiek kalkulatzeko, informazio-berreskuratzearen arloan (IR) erabili ohi den definizioa hartu dugu kontuan eta hori azalduko dugu ondoren, metodoen emaitzak ematen hasi baino lehen.

Testeko corpusetik erauzitako bikoteak erabili behar ditugu ebaluaziorako (kasu honetan 301 bikote) eta horiek dira erantzun "zuzenak", hau da, lortu nahiko genituzkeen erantzunak. Horiei *erantzun zuzenak* esango diegu (IR arloan *relevant documents*). Eraiki dugun sistema bakoitzari bikote horien aldaerak (lapurterazko hitzak) eman dizkiogu sarrera gisa, eta sistemak itzuli dituen erantzunak jaso ditugu. Oro har, hitz bakoitzeko gerta daiteke erantzun bat jasotzea (edo agian gehiago), edo erantzunik ez jasotzea. Sistemak itzulitakoei *jasotako erantzunak* esango diegu (IR arloan *retrieved documents*).

Doitasunak ($P$) adierazten du jasotako erantzun zuzenen eta jasotako erantzun guztien arteko proportzioa:

$$P = \frac{\{jasotako_erantzunak\} \cap \{erantzun_zuzenak\}}{\{jasotako_erantzunak\}} \tag{III.8}$$

eta estaldurak ($R$) adierazten du jasotako erantzun zuzenen eta erantzun zuzen guztien arteko proportzioa:

$$R = \frac{\{jasotako_erantzunak\} \cap \{erantzun_zuzenak\}}{\{erantzun_zuzenak\}} \tag{III.9}$$

Esan bezala, $F$ neurria aurreko bien batezbesteko harmonikoa da:

$$F_1 = 2 \cdot \frac{P \times R}{P + R} \tag{III.10}$$

### III.4.2 Oinarri-lerroak

Oinarri-lerroko bi sistema planteatu ditugu aldaerako hitzak normalizatzeko: lehenengo sistema planteatzeko, kontuan hartu dugu egin behar ditugun esperimentuetako testuingurua, eta bigarrenerako, berriz, hainbat sistematan ohikoa den Levenshtein distantzia erabili dugu. Azkenik, aurreko bi sistema horiek konbinatu ditugu oinarri-lerro berri batean.

#### Lehenengo oinarri-lerroa (OL1): memoria

Lehenengo oinarri-lerroak kontuan hartzen du zein testuingurutan planteatzen den normalizazioa gure esperimentuetan. Testuinguru horren arabera, esperokoa da ebaluazioan izatea zenbait kasu jadanik ikasteko datuetan ageri izan direnak, eta hortaz, oinarri-lerro honek egiten duen gauza bakarra zera da: gordetzea (memorizatzea) ikasteko ematen zaizkion bikoteak, gero informazio hori erabiltzeko ebaluazioan. Hala, ebaluazioan ematen den sarrerako aldaera gordetako bikote baten aldaerarekin bat badator, sistemak bikote horren hitz estandarra itzultzen du erantzun gisa. Aldiz, galdetutako sarrera ez badago gordetako bikoteen artean, sistemak ez du erantzunik ematen. Gerta daiteke galdetutako sarrera bikote bat behin baino gehiagotan azaltzea, eta hori gertatuz gero, erantzun bat baino gehiago ematen du sistemak

sarrera horretarako, bikote horien guztien bigarren osagaia itzultzen baitu erantzun gisa. Kasu horiek tratatzeko aukera gehiago daude, baina oinarri-lerro honetan hori hartu da. Dena den, ez da espero horrelako kasu askorik izatea.

Deskribatu berri dugun oinarri-lerroaren emaitzak III.3 taulako lehen errenkadan ageri dira (OL1). Espero zitekeen moduan, sistemaren doitasuna altua da, % 95,62; hau da, sistemak erantzuten duenean, zuzena izan ohi da haren erantzuna. Halere, doitasuna ez da iritsi % 100era, eta hori gertatzen da lehen aipatutako arrazoiarengatik: corpusean egin diren hitz-parekatzeen artean, badira zenbait kasu, non, esaldiaren arabera, lapurterazko hitz bera estandar ezberdinekin parekatu den. Erantzun posible guztiak eman direnez, doitasuna ezin da maximoa izan. Sistema erantzun bakar bat ematera behartzen badugu, ez du beti asmatuko aukeratutakoarekin, eta doitasuna, berriro ere, ez da maximoa izango[13].

Estaldurari buruz, argi dago baxua dela lortutako balioa eta hori ere esperokoa zen, ez baita askotan gertatuko hitz berbera ikusi izan dela ikasteko datuetan. Estalduraren 43,52 balioak adierazten digu ebaluatzeko hitzen erdia baino gehiago ez dela ikasteko datuetan ikusi.

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| OL1 | 95,62 | 43,52 | 59,82 |
| OL2 | 40,30 | 35,22 | 37,59 |
| OL3 | 63,54 | 58,47 | 60,90 |

III.3 Taula: Oinarri-lerroko hiru sistemen emaitzak.

### Bigarren oinarri-lerroa (OL2): Levenshtein distantzia

Esan den moduan, bigarren oinarri-lerroa Levenshtein distantzian oinarritzen da, 1eko distantzian zehazki. Euskara estandarraren transduktorean oinarrituta transduktore berri bat eraiki dugu, zeinaren bitartez sarrerako hitzetik 1 distantziara dauden hitz estandarren zerrenda lor dezakegun (zerrenda hutsa ere lor daiteke). Beraz, sarrera gisa aldaera bat ematen badiogu, hainbat hautagai lortuko ditugu transduktore horren bitartez, eta horiek maiztasunaren arabera iragaziz gero, erantzun bakarra lortuko dugu.

[13]Proba egin da aipatutako iragazki hori aplikatzen, eta lortutako emaitzak oso antzekoak izan dira: $P = 97{,}73$, $R = 42{,}86$ eta $F_1 = 59{,}58$.

Oinarri-lerro berri horrekin lortutako emaitzak III.3 taulan jaso ditugu (OL2), aurreko sistemakoekin batera. Argi ikusten denez, emaitza kaskarrak dira aurrekoekin konparatuta: bai doitasuna, bai estaldura, lehen baino baxuagoak dira, eta hala, lortutako $F$ neurria ere baxua da.

#### Hirugarren oinarri-lerroa (OL3): konbinazioa

Aurreko bi sistemak bakar batean konbina daitezke emaitzak nola aldatzen diren ikusteko. Lehenengoak doitasun altuena lortzen duenez, horren erantzunei lehentasuna ematea da aukera onena, eta sistema horrek erantzuten ez duenean, bigarrenaren erantzuna hartzea, ea horrela lehenengo sistemaren estaldura igotzen den.

Konbinazio horrekin lortutako emaitza berriak ere III.3 taulan jaso ditugu (OL3). Aurreko bi sistemen emaitzekin konparatuta, ongi ikusten da konbinazioak estaldura igo duela nabarmen (15 eta 23 puntu inguru); doitasuna, berriz, asko jaitsi da lehenengo sistemarekin konparatuta(32 puntu inguru). Beraz, konbinazioak erantzun gehiago ematen ditu baina ez du beti asmatzen, eta horregatik, azkenean, ez dago alde handirik lehen eta hirugarren sistemek lortutako $F$ balioan: puntu bateko aldea besterik ez.

### III.4.3 Lehenengo metodoaren emaitzak (lexdiff)

Metodo honetan lexdiff programa erabiltzen dugu corpus paraleloko testuen arteko aldaketa fonologiko erregularrak lortzeko. Programari karakteremailako parekatzea eskatu zaio (ikus III.3.2 atala), eta informazio hori baliatuta aldaerako hitz berriak normalizatzeko hainbat transduktore lortu ditugu finkatutako parametroen arabera. Parametro horiek sortutako erregela multzoa zein transduktorea baldintzatzen dute, eta III.3.2 atalean deskribatutakoak dira: (1) zein den aldaketa bati eskatzen zaion maiztasun minimoa erregela bat sortzeko, (2) zenbat erregela aplika daitezkeen hitz batean, (3) erregelen aplikazio-modua (sekuentziala edo paraleloa) eta (4) erregelen trinkotzea bai ala ez.

Lau parametro direnez, esperimentu asko bideratu dira balioak aldatuz eta, kasu guztietan, lortutako transduktorea konposatu egin da hizkuntza estandarra onartzen duen transduktorearekin, hau da, lehenengo iragazkia beti aplikatu da. Bigarren iragazkiari dagokionez, hizkuntza estandarraren maiztasunaren araberako hizkuntza-eredua aplikatzen duen iragazkia hain zuzen, ez da esperimentu guztietan automatikoki aplikatu haren efektua zenbaterainokoa den analizatu ahal izateko.

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| $f = 1$ | 38,95 | 66,78 | 49,20 |
| $f = 2$ | 46,99 | 57,14 | 51,57 |
| $f = 3$ | 49,39 | 53,82 | 51,51 |
| **2. iragazkia aplikatuta** | | | |
| | $P$ | $R$ | $F_1$ |
| $f = 1$ | 70,28 | 58,13 | 63,64 |
| $f = 2$ | 70,18 | 53,16 | 60,49 |
| $f = 3$ | 71,76 | 51,50 | 59,96 |

III.4 Taula: Lexdiff metodoak lortutako emaitzak lehenengo esperimentuan. Aldatzen den parametro bakarra da erregelak sortzeko eskatzen den maiztasun minimoaren balioa ($f$).

Esan bezala, metodoaren ebaluazioa testeko zatiaren 301 bikoteren gainean egin da (hitz ezberdinez osatutako bikoteen gainean) eta ebaluazioa III.4.1 atalean deskribatutako hiru parametroen bidez egin dugu: doitasuna, estaldura eta $F$ neurria.

Lehenengo esperimentuan parametro bakar baten balioa aldatu dugu, adierazten duena zenbat aldiz ikusi behar den, gutxienez, karaktereen arteko aldaketa jakin bat, hari dagokion erregela fonologikoa sor dadin (zerrendako lehenengo parametroa da). Gainontzeko parametroak finko utzi dira esperimentu honetan: erregela bakarra aplikatzen da eta ez dira erregelak trinkotzen. Emaitzak III.4 taulan jasotzen dira, lehen hiru errenkadetan, eta garbi ikusten denez, maiztasunaren balioak kontrako efektua du doitasunean eta estalduran: zenbat eta baxuagoa izan eskatutako maiztasuna, orduan eta altuagoa da lortutako estaldura, baina aldi berean, gero eta baxuagoa da doitasuna. Arrazoizkoa da emaitza hori, zeren erregela kopurua handituz gero, aukera gehiago daude erantzun zuzenak sortzeko (estaldurak gora egiten du), baina, aldi berean, erantzun oker gehiago sortzen dira (doitasuna jaisten da).

Bigarren iragazkia aplikatzen bada, hau da, maiztasun handieneko erantzuna ematen bada soilik, nabarmena da lortzen den onura. Emaitza horiek III.4 taulan bertan jaso dira (azken hiru errenkadetan), balioak erraz konparatzeko. Doitasuna 20 puntu baino gehiago handitu da hiru kasuetan, eta

nahiz eta estaldura gutxitu, azkenean $F$ neurria 10 puntu inguru handitu da kasu guztietan.

Transduktoreak sortzeko unean finkatu beharreko gainontzeko parametroei dagokienez, esperimentu asko egin ditugu, eta ez ditugu hona ekarri emaitza guztiak. Lortutako emaitzetatik ateratako ondorioak honako hauek dira:

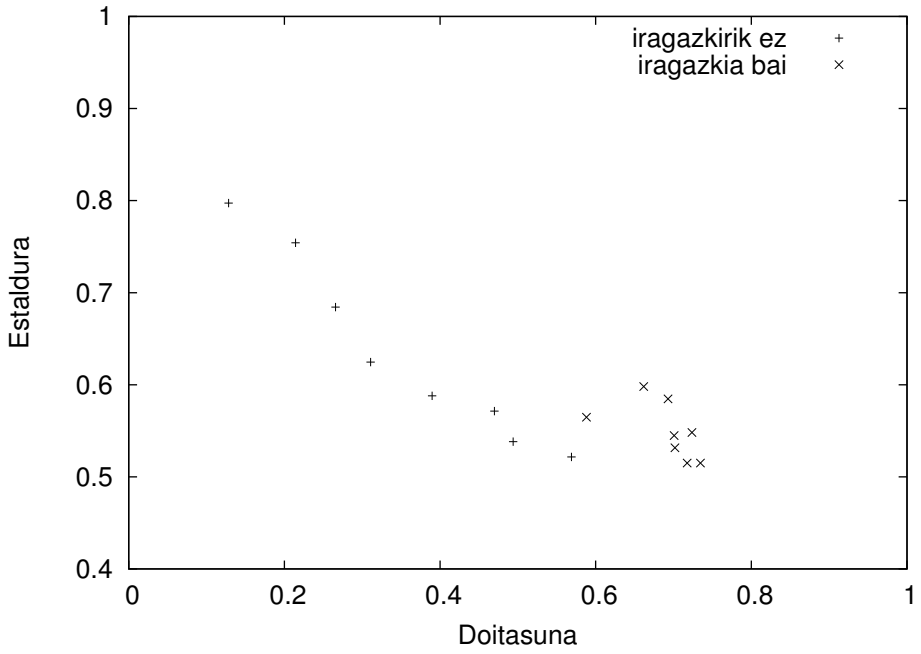
- erregelak trinkotzea eta testuingurua zehazten duten erregelak idaztea aukera onena da beti;
- erregela bat baino gehiago aplikatu ahal izateak efektu negatiboa izaten du doitasunean, eta estaldura ez da asko hobetzen;
- erregelen aplikazio-moduak, sekuentzialak zein paraleloak, ez du eragin handirik emaitzetan.

Horrez gain, erantzunak iragazteak beti handitzen du $F$ neurria, batzuetan nabarmen eta beste batzuetan ez hainbeste. Lortutako emaitzen adibide gisa, III.5 taulan esperimentuetan lortutako hiru emaitza onenak jaso ditugu ($F$ neurriaren ikuspuntutik ordenatu ditugu):

1. Lehenengo esperimentuaren parametroak (Esp1) honako hauek izan dira: erregela sortzeko aldaketaren maiztasun minimoa, 2; aplika daitezkeen erregelak, 2; aplikatzeko modua, paraleloa; erregelen trinkotzea, ez.
2. Bigarren esperimentuaren parametroak (Esp2) honako hauek izan dira: erregela sortzeko aldaketaren maiztasun minimoa, 1; aplika daitezkeen erregelak, 1; erregela bakarra aplikatzen denez, aplikazio-moduak ez du eraginik; erregelen trinkotzea, bai.
3. Hirugarren esperimentuaren parametroak (Esp3) honako hauek izan dira: erregela sortzeko aldaketaren maiztasun minimoa, 2; aplika daitezkeen erregelak, 2; aplikatzeko modua, paraleloa; erregelen trinkotzea, bai. Hau izan da emaitza onena lortu duen esperimentua.

| | $\boldsymbol{P}$ | $\boldsymbol{R}$ | $\boldsymbol{F}_1$ |
|---|---|---|---|
| Esp1 | 72,20 | 57,81 | 64,21 |
| Esp2 | 72,13 | 58,47 | 64,59 |
| Esp3 | 75,10 | 60,13 | 66,79 |

III.5 Taula: Lexdiff metodoarekin lortutako hiru emaitza onenak.




III.2 Irudia: Doitasunaren eta estalduraren arteko konpentsazioa agerian uzten duen irudia. lexdiff metodoarekin zenbait esperimentutan lortutako emaitzak irudikatu dira.

Azkenik, III.2 irudiak agerian uzten du nola konpentsatzen diren doitasuna eta estaldura lexdiff metodoarekin egindako zenbait esperimentutan. Horrekin batera iragazkiaren efektua ikus daiteke. Grafikoan bi motatako puntuak ageri dira iragazkia aplikatu den ala ez adierazteko. Iragazkia aplikatzen denean, doitasuna, oro har, altua da baina estaldura galtzen da. Iragazkia aplikatzen ez bada, estaldura handiagoa lor daiteke baina doitasuna galduz.

### III.4.4 Bigarren metodoaren emaitzak (ILP)

ILP metodoa aplikatzeko, hitz-bikoteak behar dira horietatik inferitzen baititu algoritmoak aldaketa fonologikoak adierazteko erregelak. Corpus paralelotik hitz-bikoteak lortzeko lexdiff programaren erabili dugu eta horiek filtratu ondoren 3.108 hitz-bikote lortu dira ILP ikasketa-algoritmoa exekutatzeko horien gainean (ikus III.2 taula). Algoritmoari $f$ parametroa zehaztu behar zaio, hau da, adierazi behar zaio zenbat aldiz agertu behar izan den

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| $f = 1$ | 85,02 | 58,47 | 69,29 |
| $f = 2$ | 82,33 | 54,15 | 65,33 |
| $f = 3$ | 80,53 | 50,83 | 62,32 |
| $f = 4$ | 81,19 | 50,17 | 62,01 |
| **2. iragazkia aplikatuta** | | | |
| | $P$ | $R$ | $F_1$ |
| $f = 1$ | 86,13 | 57,80 | 69,18 |
| $f = 2$ | 83,42 | 53,49 | 65,18 |
| $f = 3$ | 82,07 | 50,17 | 62,26 |
| $f = 4$ | 82,32 | 49,50 | 61,83 |

III.6 Taula: ILP metodoaren esperimentuen emaitzak maiztasunaren parametroa ($f$) aldatuz.

hitz-bikote bat corpusean, ematen duen informazioa kontuan har dezan algoritmoak ordezkapen-erregela bat sortzeko unean.

Parametro horren arabera esperimentu batzuk egin ditugu. Aurreko metodoan bezala, kasu bakoitzean lortutako erregela multzoa transduktore batean konpilatu dugu foma bitartez, eta transduktore hori hizkuntza estandarra onartzen duen transduktorearekin konposatu dugu. Beraz, aurreko metodoan bezala, lehenengo iragazkia beti aplikatu dugu. Bigarren iragazkia, ordea, ez da beti automatikoki aplikatu, haren efektua neurtzeko asmoz.

ILP metodoan lexdiff metodoan gertatutakoa errepikatzen da, baina oraingoan diferentziak handiagoak dira. Aurrekoan bezala, emaitza onenak hitz-bikote guztiak kontuan hartuta lortu dira, hau da, $f$ parametroari 1 balioa emanez. III.6 taulan ageri diren emaitzetan (lehenengo lau errenkadetan) ongi ikusten da nola okertzen diren emaitzak parametro horren balioa handitzen den heinean: bai doitasuna, bai estaldura, biak gutxitzen dira.

Bestalde, aurreko metodoaren emaitzekin konparatuta, ILP metodoaren emaitzak hobeak dira, oro har. Lexdiff metodoak estaldura handiena lortu du, % 60,13 (ikus III.5 taula), baina ez doitasuna, % 75,10ean geratu dena. ILP metodoarekin, berriz, estaldura txikixeagoa da, % 58,47, baina doitasuna ia 10 puntu handiagoa da, % 85,02, eta, hala, $F$ neurriaren ikuspuntutik, ILP metodoak lortu du emaitzarik onena: % 69,29.

ILP metodoaren doitasuna 10 puntu inguru altuagoa da, eta hori gertatzen da, hein batean behintzat, ebidentzia negatiboak hartzen dituelako

kontuan erregelak sortzeko unean: erregelen testuingurua hobeto zehazten da modu horretan.

Beste ondorio interesgarria bigarren iragazkiaren efektuarena da, hots, hitzen maiztasunaren iragazkiarena. ILP metodoan ez da ia onurarik lortzen iragazki hori aplikatuta, eta oso ondo ikusten da hori III.6 taulako emaitzei begira. Taulan bi emaitzak ageri dira, iragazki hori aplikatu gabekoa eta aplikatuta, eta ongi ikusten da haren efektua eskasa dela: doitasuna handitzen du kasu guztietan (puntu bat inguru), baina estaldura gutxitu (puntu bat baino gutxiago), eta azkenean $F$ neurria pixka bat jaisten da.

Horrez gain, kontuan hartzekoa da ILP metodoan ez dela lehenengo metodoan bezain kritikoa hitz ez-estandarrak iragazten dituen filtroaren aplikazioa. III.6 taulan ageri diren emaitzetan, iragazki hori aplikatuta dago beti, baina probatu dugu iragazkia kentzera, eta lortutako $F$ baliorik onena % 69,29tik % 56,14ra jaitsi da. Lehenengo metodoan, berriz, iragazki hori funtsezkoa da eta ez bada aplikatzen, kasu onenaren $F$ balioa 55 puntu jaisten da: % 66,79 baliotik, % 11,53 balioraino.

### III.4.5 Hirugarren metodoaren emaitzak (WFST)

Hirugarren metodo hau aplikatzeko behar den informazioa, ILP metodoa aplikatzeko behar den berbera da, hau da, lexdiff programaren bitartez ikasteko corpus paralelotik lortutako 3.108 hitz-bikoteak. Informazio hori baliatuta, Phonetisaurus tresna-multzoa erabil daiteke bikote horietatik ikasteko eta transduktore haztatu bat lortzeko. Gero, transduktore horren bitartez, aldaera berriei dagozkien "berridazketak" lor daitezke.

Transduktore haztatuak itzultzen dituen erantzunak, aurreko metodoetan bezala, hizkuntza estandarra onartzen duen transduktorearen bitartez iragaziko ditugu, baina hori egin ondoren, metodo honetan ez dugu erabiliko aurreko bi metodoetan aplikatu dugun bigarren iragazkia. Metodo honetan baliatuko dugu transduktore haztatuaren erantzunak ordenatuta daudela kostuen arabera, eta hala, estandarrak ez diren erantzunak kendu ondoren, hitz bakoitzeko emango dugun erantzun bakarra izango da transduktorearen arabera erantzun onena dena (kostu txikienekoa). Aurreko metodoetan ez zegoen inongo ordenarik erantzunen artean, eta horregatik erabili behar izan da hitzen maiztasuna kontuan hartzen duen bigarren iragazkia. Ikusi dugu zein zen iragazki horren efektua: doitasuna handitzen zuen, oro har, baina estaldura gutxitu. Metodo honetan, transduktore haztatuak itzultzen dituen erantzunen ordena ikasteko datuetan oinarrituta dago, hau da, tresnak ikusi duen informazioan, eta hori baliatuko dugu bigarren iragazki gisa.

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| $n = 1$ | 87,79 | 63,73 | 73,81 |
| $n = 3$ | 82,88 | 71,59 | 77,30 |
| $n = 5$ | 82,88 | 72,81 | **77,50** |
| $n = 10$ | 81,78 | 73,46 | 77,39 |
| $n = 20$ | 80,78 | 73,87 | 77,16 |
| $n = 30$ | 79,95 | 74,35 | 77,04 |

III.7 Taula: WFST metodoak lortu dituen emaitzen batezbestekoak doikuntzako esperimentuetan (lau ataleko balidazio gurutzatua). Aldatutako parametro bakarra eskatutako erantzun kopurua da ($n$). Esperimentu guztietan *beam* parametroa 5.000 da.

### Doikuntza

Phonetisaurus tresna-multzoaren barneko programek hainbat parametro dituzte, eta bi parametroren efektua aztertu eta doitu nahi izan dugu ebaluazioa planteatu baino lehen. Bi parametro horiek dira: (1) transduktoreari eskatutako erantzun kopurua deskodeketan eta (2) deskodeketa-urratseko bilaketaren "sakonera" (*beam*) finkatzen duen parametroa. Azken parametro horri dagokionez, sakonera handitzeak hipotesi gehiago ebaluatzen ditu eta, ohikoa denez, horrek eragina du exekuzio-abiaduran. Tresnak erabiltzen duen lehenetsitako balioak mugatu egiten du hipotesi kopurua, eta hori handitzera behartu dugu *beam* delako parametroaren bitartez.

Beraz, parametro bi horien balio onenak finkatzeko, **doikuntzarako** esperimentu batzuk egitea erabaki dugu, eta hala, ikasteko corpusa (corpus osoaren % 80 jasotzen duena) 4 zati osagarrietan banatu dugu balidazio gurutzatuaren teknika (*cross-validation*) aplikatzeko parametroen balio onenen bila. Ikasteko corpusa lau zatitan banatuz gero, lau esperimentu egin ditzakegu, non, ebaluaziorako zatiaren tamaina bat datorren azken ebaluazioko zatiaren tamainarekin. Erantzun kopurua adierazten duen parametroarekin probatu diren balioak izan dira 1, 3, 5, 10, 20 eta 30; bilaketaren sakonera finkatzen duen parametroarekin bi balio probatu dira: 500 (lehenetsitakoa) eta 5.000.

III.7 taulak doikuntzako esperimentuen emaitzak biltzen ditu. Esperimentu bakoitza lau ataletan errepikatzen denez, taulan ageri diren balioak lau ataletako batezbestekoak dira. Aipatu beharra dago taulako balio horiek guztiak *beam* parametroa 5.000 izanik lortu direla, balio horrek beti lortu

| Metodoa | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| OL1 | 95,62 | 43,52 | 59,82 |
| OL3 | 63,54 | 58,47 | 60,90 |
| lexdiff | 75,10 | 60,13 | 66,79 |
| ILP | 85,02 | 58,47 | 69,29 |
| WFST | 83,46 | 75,42 | 79,23 |

III.8 Taula: Oinarri-lerroen emaitzak eta hiru metodo fonologikoekin lortutako emaitza onenak ebaluazioko corpusaren gainean.

baititu emaitza hobeak. Erantzun kopuruari dagokionez, emaitzek adierazten dute erantzun kopurua handitzeak $F$ balio hobea lortzen duela baina goiko muga jakin batera iritsi arte, kasu honetan 5 balioan kokatzen dena.

Emaitza horiek islatzen duten beste ondorio agerikoa da doitasuna eta estaldura $n$ balioaren arabera orekatzen direla: bata handitzen bada, bestea gutxitzen da.

#### Ebaluazioa

Doikuntzako esperimentuen emaitzak kontuan izanik, metodoari dagokion azken ebaluazioa egin dugu: ikasteko corpus osoa erabili da WFST transduktorea lortzeko, eta ebaluazioa testeko corpusarekin egin da. Ebaluazio horretan 5 erantzun eskatu dira deskodeketa-urratsean, eta 5.000 balioko *beam* parametroa finkatu da. Lortutako emaitzak honako hauek izan dira: $P = 83{,}46$, $R = 75{,}42$ eta $F_1 = 79{,}23$. Aurreko bi metodoek lortutako emaitzekin konparatuta hobekuntza nabarmena da, eta konparazioa errazteko III.8 taulan jaso ditugu metodo guztien emaitzak[14]: lexdiff eta ILP metodoek lortutako emaitza onenak $F$ neurrirako % 66,79 eta % 69,29 ziren, hurrenez hurren, eta WFST metodoak 10 puntuko aldea lortu du, % 79,23.

### III.4.6 Informazio morfologikoaren erabilera

Hiru metodoen emaitzak aztertuta, garbi geratzen da WFST metodoa gailentzen dela beste biekiko, eta urrats bat aurrera eman nahian, esperimentu berri bat planteatu dugu aztertu ahal izateko ea emaitzak hobe daitezkeen beste motatako informazioa emanez gero ikasteko urratsean.

[14] Oinarri-lerroko hiru sistemen emaitzetatik bi onenak jaso dira taulan: antzeko $F$ neurria dute, baina doitasunaren eta estalduraren balioak oso diferenteak dira.

Gogoratu behar da Phonetisaurus tresnari eman behar zaion informazio ia bakarra ikasteko hiztegia dela. Aurreko esperimentuetan hitzen arteko erlazioa eman zaio hiztegi horretan, hau da, hiztegiko sarrera bakoitza hitz-bikote bat izan da: aldaera zein estandarra jasotzen dituen hitz-bikote bat. Halaber, hiztegian parekatutako hitz-bikote guztiak sartu dira, berdinak zein ezberdinak. Bikote horien artean daude, esaterako, honako bi hauek:

**emaiten → e m a t e n**
**nehori → n e h o r i**

Orain azalduko dugun esperimentu berrian, kontuan izan dugu aldaeraren eta hitz estandarraren artean gertatzen diren zenbait aldaketa, morfema-mugan gertatzen direla, esaterako:
*emaiten* → *ematen* ; *hortan* → *horretan* ; *deneri* → *denei* ; ...

Gure hipotesia da morfema-mugako informazio hori ematea lagungarria izan daitekeela ataza ebazteko, hau da, agian emaitza hobeak lortuko direla informazio hori emanez gero, tresnak hobeto ikasiko duelako non egin behar diren aldaketak. Hori dela eta, esperimentu berri bat planteatu dugu: estandarrari dagokion informazioa hitza bera izan beharrean, hitzaren informazio morfologikoa eman diogu tresnari ikasteko.

Informazio morfologiko horren arabera, bi esperimentu bideratu ditugu:

1. Lehenengo esperimentuan, hitz estandarraren analisi morfologikoa[15] gure analizatzaileak itzultzen duen moduan utzi da, morfofonemak barne[16] (Beesley eta Karttunen, 2003). Esaterako, lehen azaldu diren bi adibideetan honako analisi hau ematen da:

   **emaiten → e m a N + t e n**

   **nehori → n e h o Q + R i**

   Adibide horietan, **N**, **Q** eta **R** morfofonemak dira: lehenengo biek **n** zein **r** epentetikoak adierazten dituzte lemetan, eta hirugarrenak **r** epentetikoa adierazten du atzizkian.

2. Bigarren esperimentuan, analisi morfologikoa sinplifikatu dugu: morfofonemak haiei dagokien grafeman bihurtu ditugu eta horrela morfemen adierazpen kanonikoen kateaketa lortzen da. Hipotesia da hainbat morfofonemak agertzeko probabilitate txikia dutela, eta zarata sor dezaketela ikasteko prozesuan.

[15] *Analisi morfologiko* zein *segmentazio morfologiko* terminoak sinonimo gisa erabili ditugu lan honetan.

[16] `http://www.ehu.eus/seg/hizk/1/3` (2016-03-31n atzitua)

Aurreko adibideen informazioa honako hau izan da esperimentu honetan:

**emaiten** → **e m a n + t e n**

**nehori** → **n e h o r + r i**

Kontuan izan behar da analisi morfologikoa erabiliz gero ikasteko informazio gisa, sarrerako aldaera bakoitzarentzat deskodeketa-fasean lortzen den erantzuna ere (bat edo gehiago), analisi morfologiko bat dela, hori baita ikasitakoa. Hortaz, prozesaketa berri bat egin behar da analisi horri (edo horiei) dagokion hitz estandarra lortzeko (analisia ez bada zuzena, ez da hitz estandarrik lortzen). Nahiz eta ohikoa ez izan, urrats horretan gerta daiteke analisi batek hitz estandar bat baino gehiago izatea aukeran, eta hori gertatzen denean maiztasun handieneko hitz estandarra aukeratu dugu (berriro erabili dugu lexdiff eta ILP metodoetan bigarren iragazki gisa erabili den hizkuntza-eredua). Adibidez, `har+k` analisiari dagokion hitz estandar gisa, bi aukera itzultzen dizkigu daukagun transduktoreak: `hark` eta `harrek`, eta bi horien artean maiztasun handienekoa aukeratzen bada, `hark` izango da emango den azken erantzuna.

Bukatzeko, sarrerako aldaera bati hitz estandar bat baino gehiago egokitzen bazaizkio (analisi zuzen bat baino gehiago itzuli duelako WFST transduktoreak), azken erantzunaren aukeraketa kostuaren arabera egin da (III.4.5 ataleko esperimentuetan egin den antzera), WFST transduktoreak ordenatuta itzultzen baititu erantzunak.

### Doikuntza

Ikasteko informazioa aldatzen denez esperimentu berrietan, berriro balidazio gurutzatuaren teknika erabili dugu ikasteko corpusarekin, parametroak ahalik eta ondoen doitzeko. Gogoratu bi parametro direla doitu nahi ditugunak: (1) transduktoreari eskatutako erantzun kopurua deskodeketa-fasean, eta (2) bilaketaren sakonera (*beam*) finkatzen duen parametroaren balioa.

III.9 taulak balidazio gurutzatuko lau ataletan lortutako emaitzen batezbestekoak biltzen ditu. Emaitza horietan guztietan *beam* parametroaren balioa 5.000 da, balio horrek lortu baititu beti emaitza onenak. III.7 taulako emaitzekin konparatuta, garbi ikusten da analisi morfologikoa erabiltzen duten bi transduktoreekin erantzun gehiago eskatu behar direla emaitza onenak lortzeko, 20 erantzun eskatu behar baitira bi kasuetan, eta lehen nahikoak ziren 5.

| | $\boldsymbol{P}$ | $\boldsymbol{R}$ | $\boldsymbol{F}_1$ |
|---|---|---|---|
| WFST2 | | | |
| $n = 1$ | 91,46 | 60,48 | 72,78 |
| $n = 3$ | 89,09 | 68,19 | 77,24 |
| $n = 5$ | 88,15 | 70,46 | 78,30 |
| $n = 10$ | 86,54 | 72,00 | 78,58 |
| $n = 20$ | 85,27 | 73,70 | **79,05** |
| $n = 30$ | 84,34 | 73,86 | 78,74 |
| WFST3 | | | |
| $n = 1$ | 91,14 | 60,24 | 72,49 |
| $n = 3$ | 88,23 | 68,16 | 77,52 |
| $n = 5$ | 86,63 | 70,87 | 77,94 |
| $n = 10$ | 85,36 | 72,25 | 78,24 |
| $n = 20$ | 83,74 | 73,87 | **78,48** |
| $n = 30$ | 83,19 | 74,19 | 78,42 |

III.9 Taula: WFST bi transduktore berriekin lortutako emaitzak (balidazio gurutzatuaren batezbestekoak). WFST2: *hitza–analisi_morfologikoa*. WFST3: *hitza–analisi_morfologiko_sinplifikatua*. $n$ eskatutako erantzun kopurua da; *beam* parametroa beti 5.000.

### Ebaluazioa

Doikuntzako esperimentuen emaitzak kontuan izanik, azken ebaluazioan 20 erantzun eskatu zaizkie bi transduktoreei. Emaitzak III.10 taulan jaso dira eta konparazioa errazteko,III.4.5 ataleko WFST sistemaren emaitza ere taulara ekarri dugu (WFST1 deitu diogu). Zenbakietan ikusten denez, emaitza onenak azken transduktoreari dagozkio (WFST3), hau da, hitza eta analisi morfologiko sinplifikatutik ikasten duen transduktoreari. Dena den, hiru WFST transduktoreen arteko diferentziak txikiak dira eta ez dira estatistikoki esanguratsuak ($p$ balioa $> 0{,}1$ izan da Bhapkar testean, (Bhapkar, 1966)).

| | $\boldsymbol{P}$ | $\boldsymbol{R}$ | $\boldsymbol{F}_1$ |
|---|---|---|---|
| WFST1 | 83,46 | 75,42 | 79,23 |
| WFST2 | 85,39 | 75,75 | 80,28 |
| WFST3 | 85,56 | 76,74 | **80,91** |

III.10 Taula: Hiru WFST transduktoreen emaitzak ebaluazio-corpusaren gainean.

## III.5 Esperimentuen ondorioak eta erabakiak

Kapitulu honetan hiru metodo aplikatu ditugu ataza bera ebazteko: aldaerei dagozkien hitz estandarrak automatikoki esleitzeko. Hiru metodoekin datu berberak erabili ditugu: hainbat esaldiz osatutako lapurtera/estandarra corpus paralelo txiki bat, bi zatitan banatu dena esperimentuak bideratzeko: % 80 ikasteko erabili da eta % 20 ebaluatzeko.

Hainbat esperimentu egin ondoren, metodo bakoitzak lortu duen emaitzarik onena hartuz gero (ikus III.11 taula), nabaria da WFST teknologian oinarria duen metodoa, Phonetisaurus tresnaren bitartez gauzatua, beste bien artean gailentzen dela: metodoaren arrakasta lortutako estalduran datza, inolako zalantzarik gabe, alde handia ateratzen baitie beste bi metodoei parametro horretan (16 puntu ingurukoa), eta horren eraginez, $F$ neurrian ere (10 puntu inguruko aldea). Doitasunari dagokionez, baliorik onena WFST metodoak lortzen du, baina ILP metodoa ere oso gertu dago.

Beraz, corpus dialektalarekin lortutako emaitza horiek kontuan harturik, hurrengo kapituluan hirugarren metodoan sakonduko dugu eta corpus historikoen gainean planteatuko ditugu esperimentu berriak. Esperimentu

| Metodoa | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| OL1 | 95,62 | 43,52 | 59,82 |
| OL3 | 63,54 | 58,47 | 60,90 |
| lexdiff | 75,10 | 60,13 | 66,79 |
| ILP | 85,02 | 58,47 | 69,29 |
| WFST1 | 83,46 | 75,42 | 79,23 |
| WFST3 | **85,56** | **76,74** | **80,91** |

III.11 Taula: Metodo bakoitzarekin lortutako emaitza onenak: bi oinarrilerroko onenak, lexdiff eta ILP metodoen emaitza onenak, WFST metodoaren bi aukera onenak.

horietan aztertu nahi dugu: (1) ea metodoa aproposa den normalizazioaren ataza corpus historikoetan ebazteko, eta (2) zein den metodologia egokia normalizazio-lan hori bideratzeko.

Horrekin batera, esperimentu berrietan argitu nahi dugu zer informazio baliatuta lortzen den normalizazio-emaitzarik onena, puntu hori ez baita guztiz garbi geratu orain arte egin ditugun esperimentuetan: badirudi informazio morfologikoa erabiltzea lagungarria izan daitekeela emaitzak hobetzeko (ikus III.10 taula), baina hiru WFST sistemen emaitzak gertuegi daude bakar baten aldeko hautua egiteko. Hori dela eta, hurrengo kapituluan honako bi aukera hauekin jarraitu behar dugu aurrera: alde batetik, aldaera eta hitz estandarra erabiliko dugu ikasteko informazio gisa (WFST1 deitu duguna), hori baita aukera sinpleena; beste aldetik, analisi morfologikoa erabiltzen duten bi aukeren artean (WFST2 eta WFST3), azkena erabiliko dugu, hau da, ikasteko aldaera eta hitz estandarraren analisi morfologiko sinplifikatua erabiltzen duena.

Euskarazko corpus historikoekin esperimentatzeaz gain, erdarazko zenbait corpusekin lan egin nahi dugu hurrengo kapituluan, gure sistemaren emaitzak alderatzeko bibliografian azaldutako hainbat sistemaren emaitzekin.

# IV. KAPITULUA

## Testu historikoen normalizazioa WFST erabiliz

### IV.1 Sarrera

Kapitulu honetan esperimentu berriak egin ditugu zenbait corpus historikotan, modu horretan aztertzeko aukeratu dugun normalizazio-metodoaren egokitasuna, ezaugarriak eta mugak. Corpusei dagokienez, II. kapituluan deskribatutako eta prestatutako euskarazko bi corpusak erabili ditugu, *Gero* corpusa (II.3.4 atala) eta *Peru Abarka* corpusa (II.3.5 atala). Bi horiez gain, eskuratu ahal izan ditugun gaztelaniazko corpusa eta eslovenierazko corpusa (II.5.1 eta II.5.2 atalak) erabili ditugu metodoaren egokitasuna frogatzeko.

Normalizazio-metodoari dagokionez, III. kapituluan ikusi dugu Phonetisaurus tresnaren bitartez lortu direla emaitza onenak corpus dialektalean, eta hori izango da kapitulu honetan erabiliko dugun metodo bakarra.

Normalizazioaren inguruko esperimentu berriekin hasi baino lehen, Phonetisaurus tresnaren oinarrian dagoen teknologian sakonduko dugu. Hala, IV.2 atalean, WFST teknologiaren laburpena egingo dugu eta aipatuko dugu hizkuntzaren prozesamenduko zein aplikaziotan erabili izan den teknologia hori. Gero, IV.3 atalean, Phonetisaurus tresnan sakonduko dugu gehiago ongi ulertzeko tresnak nola funtzionatzen duen, eta argi uzteko nola erabiltzen dugun gure atazaren barruan.

Teknologian sakondu ondoren, euskaraz idatzitako testu historikoak normalizatzeko egin ditugun esperimentuak azalduko ditugu. Lehendabizi, *Gero* obraren corpusarekin egin diren hainbat esperimentu deskribatuko dira IV.4 atalean: corpusaren ikasteko zatiarekin doikuntza esperimentuak egin dira

erabakitzeko zein den metodoa aplikatzeko modurik egokiena, eta gero azken ebaluazioa egin da testeko zatiarekin.

Corpus horretan lortu diren emaitzak berresteko, IV.5 atalean *Peru Abarka* obraren corpusarekin errepikatu ditugu esperimentu berdinak, eta, azkenik, IV.6.1 eta IV.6.2 ataletan, gaztelaniarekin zein eslovenierarekin egin ditugu esperimentu berriak, bi helburu hauek lortzeko: (1) frogatzeko proposatzen dugun normalizazio-metodoa hizkuntzarekiko independentea dela, eta (2) konparatzeko gure metodoak lortzen dituen emaitzak beste sistema batzuek lortutakoekin.

## IV.2 WFST teknologia

III. kapituluan azaldu dugun moduan, transduktore arruntekin lan egitean aurkitu ditugun mugak saiheste aldera, WFST teknologiarekin egin dugu lan, hau da, transduktore haztatuekin, eta bide horretatik lortu dira emaitza onenak corpus dialektalarekin egin diren esperimentuetan.

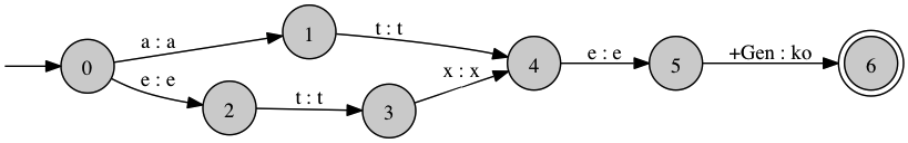
Transduktore haztatuen teknologia oso erabilia da hizkuntzaren prozesamenduan, eta atal honetan haren oinarriak azalduko ditugu labur. Horren ondoren, teknologia hori erabili duten zenbait aplikazio-arlo zerrendatuko ditugu.

### IV.2.1 Teknologiaren oinarriak

Egoera finituko automatei eta transduktoreei buruzko bibliografia asko dago kontsultatzeko, baina atal honetan egin dugun deskribapen laburra, batez ere, Hulden (in print) lanean oinarrituta dago, eta esan beharra dago bertan egiten den hurbilpena hizkuntzalaritza konputazionalaren ikuspegitik egiten dela gaia errazago ulertzeko (ikuspegi matematikoa zurrunagoa eta formalagoa litzateke).

Egoera finituko automata arrunt bat (FSA, *finite-state automaton*) lengoaia bat definitzen duen egitura konputazionala da, zeinaren arabera sarrerako karaktere-kateak bi multzotan sailkatzen dira: onartuak eta ez-onartuak. Hiru dira automata finitu baten osagaiak: (1) egoera multzo finitu bat, non egoera batzuk (gutxienez bat) bukaerako egoerak diren; (2) trantsizioak, hots, karaktere-kateekin etiketatutako egoeren arteko noranzkodun bideak, eta (3) hasierako egoera bat.

Automataren egiturak onartutako kateak definitzen ditu, hau da, sarrerako kate bat onartzeko, bide bat egon behar du automatan hasierako egoeran hasten dena, egoeraz egoera joaten dena trantsizioak eginez katearen karaktereen arabera eta trantsizioen etiketen arabera, eta bukaerako egoera




IV.1 Irudia: Egoera finituko transduktore baten irudia.

baten batera iristen dena sarrerako katearen karaktere guztiak onartuta. Ez badago horrelako biderik, automatak ez du katea onartzen.

Trantsizioetako etiketak bikoteak badira, hau da, sarrerako zein irteerako karaktere-kateak dituzten bikoteak, transduktore bat definitzen da (FST, *finite-state transductor*): automatak bezala, sarrerako lengoaia bat onartzen du transduktoreak, baina horretaz gain, itzuli egiten du sarrerako kate bakoitza trantsizioetan adierazitako bikoteen arabera. Transduktoreak bi lengoaiaren arteko erlazioa definitzen duenez, ohikoa da 'sarrerako' edota 'irteerako' lengoaiei buruz hitz egitea.

IV.1 irudian zazpi egoerako transduktore bat ageri da. Hasierako egoera 0 zenbakiduna da (gezi batez markatua dago) eta bukaerakoa 6 zenbakiduna. Ikusten denez, egoeren arteko trantsizioek bikoteekin etiketatu dira: 5 eta 6 egoeren arteko etiketak adierazten du `+Gen` katea `ko` katean bihurtu behar dela; gainontzeko trantsizioetako etiketak sinpleak dira irteerako katea sarrerako bera delako. Transduktore horrek `ate+Gen` eta `etxe+Gen` sarrerako kateak onartzen ditu, eta irteeran `ateko` zein `etxeko` itzultzen du, hurrenez hurren.

Trantsizioei hazta edo kostu bat eransten bazaie, automata haztatuak zein transduktore haztatuak ditugu. Ohikoa da kostu horiek probabilitate gisa estimatzea, eta hala eginez gero, automata haztatu batek karaktere-kateen distribuzio probabilistikoa modelatzen du: sarrerako karaktere-kate baten probabilitatea kalkulatzeko, trantsizioetako eta bukaerako egoeraren probabilitateak biderkatzen dira. Sarrerako kateak bide bat baino gehiago baldin badu automatan zehar (bide *paraleloak* dituela esaten da) bide guztien probabilitateak batu behar dira katearen probabilitatea lortzeko.

Dena den, probabilitatea ez da kostuak interpretatzeko aukera bakarra, eta hizkuntzaren prozesamenduan gutxitan erabiltzen den interpretazioa da, haren kostu konputazionala dela eta (biderketa eragiketa motela da batuketarekin konparatuta, zenbakiak oso txikiak direnean arazoak sortzen dira eta abar). Hala, oso ohikoak dira beste bi kostu-egitura hauek transduktore haztatuekin lan egiteko: *probabilitateen logaritmo negatiboen* egitura eta

*tropical* izeneko egitura.

Ez dugu egitura bakoitzaren xehetasunak azalduko hemen, ez baita hori atal honen helburua; argi utzi nahi dugu, ordea, edozein egiturak zehaztu behar duena: nola interpretatzen diren kostuak bideetan zehar eta bideen artean. Hala, kostu-egitura bat definitzeko, 5 parametroko aljebra bat definitu behar da, $(S, \oplus, \otimes, \bar{0}, \bar{1})$, *eraztunerdi* izeneko egitura aljebraikoaren ezaugarriak betetzen dituena. $S$ multzo bat da, eta $\otimes$ (biderketa abstraktua) eta $\oplus$ (batuketa abstraktua) multzoaren gainean definitzen diren eragiketak dira. Lehenengoak definitzen du nola konbinatu behar diren bideko trantsizioen kostuak bideari dagokion kostua kalkulatzeko, eta bigarrenak, berriz, nola konbinatu behar diren bide paraleloen kostuak sarrerako kate bati dagokion kostua kalkulatzeko. $\bar{0}$ eta $\bar{1}$ elementuak $S$ multzokoak dira eta bi eragiketen elementu neutroak dira, hurrenez hurren[1].

Transduktoreen ezaugarrien artean bada bat oso erabilgarria suertatzen dena: transduktoreen arteko *konposaketa* deritzon eragiketa itxia da (beste transduktore bat sortzen du), eta eragiketa horrek sortzen duen transduktore berria kalkulatzeko algoritmoa eraginkorra da. Bi transduktoreren arteko konposaketa definitzea sinplea da: $A$ transduktoreak X lengoaia formaletik Y lengoaiarako itzulpena egiten badu, eta $B$ transduktoreak Y lengoaiatik Z lengoaiarakoa, bi transduktore horien konposaketa, $C = A \circ B$ transduktore berri bat da, X eta Z lengoaien arteko itzulpena egiten duena. Beraz, lehenengo transduktorearen irteerako lengoaia bigarrenaren sarrerakoa bilakatzen da konposaketaren bitartez.

Konposaketa asko erabiltzen da hizkuntzaren prozesamenduko hainbat aplikazio garatzeko. Izan ere, halako aplikazio askoren diseinua hainbat transduktore sinpleren konposaketan oinarritzen da. Eragiketa horrekin batera, oso erabilgarria izaten da maiz transduktorea alderantziz erabili ahal izatea.

Algoritmo asko daude transduktore haztatuak inferitzeko (alegia, corpusetik transduktorearen topologia eta pisuak lortzeko), konbinatzeko, manipulatzeko edota optimizatzeko (Mohri, 2009), eta liburutegi bat baino gehiago eraiki da denboran zehar transduktore horiekin lan egin ahal izateko. Gaur egun maizen aurkitzen den erreferentzia OpenFst[2] liburutegiarena da (Allauzen *et al.*, 2007), kode irekian zabaldua. Bere aitzindaria AT&T FSM liburutegia izan zen (Mohri *et al.*, 1997). Transduktore haztatuekin lan egiteko beste aukera bat Carmel[3] izeneko softwarea da (Graehl, 1997).

[1] Beste hainbat lege bete behar dira eraztunerdi egitura izateko, baina ez ditugu guztiak zerrendatu. Xehetasun gehiagorako Mohri (2009) kontsulta daiteke.

[2] `www.openfst.org` (2016-02-24an atzitua)

[3] `www.isi.edu/publications/licensed-sw/carmel`

### IV.2.2 Aplikazioak

Knight eta May (2009) lanean, hizkuntzaren prozesamenduko hainbat aplikazio zerrendatzen dira egoera finituko transduktore haztatuetan oinarritzen direnak, eta agerian uzteko nola erabiltzen diren transduktore horiek hizkuntzaren arloan, horietako bi aplikazio deskribatzen dituzte sakonago lan horretan: izenen eta termino teknikoen transliterazioa egiteko aplikazioa, eta esaldiak itzultzeko aplikazioa.

Bi aplikazio horietan egoera finituko automata haztatuak (WFSA) eta transduktore haztatuak (WFST) erabiltzen dira kanal zaratatsuaren eredua (*noisy channel*) inplementatzeko, eta egileek diote teknika berdintsuak erabili direla hizkuntzaren prozesamenduko beste zenbait aplikaziotan: antzeko egiturak eta diseinuak erabiltzen dira guztietan, eta aldatzen dena da modelatu beharreko datuak. Honako aplikazio hauek aipatzen dituzte lanean (aplikazio bakoitzaren erreferentzia bibliografiko zehatzak Knight eta May (2009) lanean kontsulta daitezke):

- Hizketaren ezagutza eta sintesia. Hizketaren seinalea akustikoa izanik, ahoskatutako hitzak lortu behar dira hizketaren ezagutzan. Hori lortzeko, $n$-gram hizkuntza-eredu estandarra eta kanal zaratatsuaren eredua inplementatzen duten transduktore-kate bat erabil daitezke. Hizkuntza-ereduaren zein transduktoreen kostuak doitzeko, beharrezkoak dira entrenatzeko datuak. Kapitulu honetan sakondu behar dugun Phonetisaurus tresna arlo honi dagokio, grafemen eta fonemen arteko bihurketa egiten duten sistemak eraikitzeko tresna baita.

- Lexikoaren prozesaketa. Egoera finituko teknologiaren erabilera arrakastatsua izan da morfologiaren arloan, bereziki flexio aberatseko hizkuntzetan (euskara, turkiera, suomiera), non hitza banatu behar den esanahia duten hainbat morfematan. Beste zenbait hizkuntzatan, txineran esaterako, ez da zuriunerik erabiltzen eta ebatzi beharreko ataza da informazioa hitzetan banatzea.

- Etiketatzea. Hizkuntzaren prozesamenduko zenbait aplikaziotan beharrezkoa suertatzen da hitzak etiketatzea, hau da, hitz bakoitzari esleitu behar zaio etiketa jakin bat aurretik adostutako etiketa multzo batetik. Horren adibide klasikoa kategoria gramatikalen etiketatzearena da (*part-of-speech*); baina gehiago daude, esaterako, entitateen identifikazioa (pertsonak, tokiak, erakundeak). Horrelako aplikazioetan, automata zein transduktore haztatuak erabil daitezke etiketa-sekuentziak modelatzeko, zein hitzen eta etiketen arteko ordezkapenak egiteko.

- Laburpen automatikoa. Arlo honen helburua da dokumentu baten edo multzo baten testua laburtzea erabilgarria den informazioaren azpimultzoa lortzeko. Albisteetako tituluak automatikoki sortzeko aplikazioak arlo honetakoak dira, eta WFST teknologia erabili izan da halako aplikazioren batean.

- Karaktereen ezagutza optikoa (OCR). Inprimatutako testua formatu elektronikora bihurtzeko OCR teknikak erabiltzen dira maiz, eta aurrerapen handiak egin dira horietan ezagutza-teknika probabilistikoak erabiltzean. Kanal zaratatsuaren eredua egokia da teknika horiek modelatzeko eta erabilia izan da arlo honetan ere.

- Itzulpen automatikoa. Itzulpen automatikoan ere erabili izan da WFST teknologia. Atal honen hasieran esan den moduan, Knight eta May (2009) lanean esaldiak itzultzeko eredua deskribatzen da, eta eredu hori WFST teknologiaren bitartez gauzatuta dago.

## IV.3 Phonetisaurus tresnaren deskribapen zehatza

Phonetisaurus tresnaren deskribapen laburra egin dugu honez gero III.3.4 atalean: Josef Novak-ek garatutako tresna-multzoa da[4] eta grafema-fonema (G2P, *Grapheme-to-Phoneme*) bihurketa egiten duten sistemak eraikitzeko balio du. Tresnak egoera finituko transduktore haztatuak erabiltzen ditu (WFST) eta OpenFst liburutegian oinarritzen da (Allauzen *et al.*, 2007). Kode irekian zabaldutako tresna da, BSD lizentziapean.

Grafema-fonema bihurketa oso ataza garrantzitsua da hizkuntzaren prozesamenduko hainbat arlotan, esaterako hizketaren ezagutza automatikoan (ASR, *Automatic Speech Recognition*) eta testutik abiatutako hizketaren sintesian (TTS, *Text-to-Speech synthesis*). Dinamikoak izan behar dute bi arlo horietan garatutako sistemek, hizketak berezkoa baitu dinamikoa izatea, eta dinamikotasun horrek eskatzen du sistemek gai izan behar dutela hitz berriak ahoskatzeko edota ahoskera berriak proposatzeko jadanik ezagutzen diren hitzetarako. Bi gaitasun horiek izan behar ditu, oro har, grafema-fonema bihurketa-sistema batek, eta behar-beharrezkoa da lan hori zehatz betetzen duten sistemak eraikitzea (Novak *et al.*, 2015). Zenbait hizkuntzatan lan hori zaila suertatzen da eta baliabide asko eskatzen ditu, ahoskera-arauek salbuespen asko baitituzte. Ingelesez edo frantsesez, esaterako, grafemetatik fonemetara igarotzeko arauen artean arau berezi asko daude eta euren arteko gatazkak maiz gertatzen dira. Hurrengo ataletan ikusiko dugunez,

[4] `https://github.com/AdolfVonKleist/Phonetisaurus`

Phonetisaurus gai da datuetatik ikasteko, hau da, grafemen eta fonemen arteko bihurketa egiteko arauak ez dira aurretik definitu behar. Horren ordez, tresnari hainbat adibide ematen zaizkio arauak ikas ditzan, eta gero aukera dago ikasitakoa datu berrien gainean aplikatzeko. Tresnaren emaitzak onak izateaz gain, bere egileak garrantzia ematen dio entrenatze-fasea luzea ez izateari eta erabilpen errazeko tresna izateari.

Nahiz eta grafema-fonema bihurketa ebazteko tresna izan, aurreko kapituluko III.3.4 atalean ikusi dugu guk ez dugula horretarako erabili. Gure lana ez da hizkuntzaren ahoskeraren inguruan kokatzen, hizkuntzaren idazkeran baizik, zehazki, idazkera horren barruan aurki ditzakegun aldaeren inguruan. Phonetisaurus datuetatik ikasten duen tresna denez, datu horietan aldaeraren eta estandarraren arteko hainbat erlazio eman dakizkioke ikas dezan, eta gero ikasitakoa erabil daiteke aldaera berriak normalizatzeko, eta hala erabili dugu gure lanean.

Honela definitzen du Novak-ek Phonetisaurus tresna Novak *et al.* (2015) artikuluaren sarreran: *The Phonetisaurus G2P approach, which is the subject of this work, is another variation on the well-known joint multigram approach, and can be summarised in four steps. The first step is data preparation, which in involves collecting a suitable pronunciation lexicon for training. This should include a list of known words and their corresponding pronunciations. The second step is to align the training lexicon, so as to approximate a mapping between the graphemes and phonemes in the lexicon. In the third step the aligned corpus is utilised as the input to estimate a standard n-gram model, which is subsequently converted into a Weighted Finite-State Transducer (WFST). In the fourth and final step, pronunciations for previously unseen words are predicted by using weighted composition (Mohri and Pereira and Riley 2002) to compute the intersection of the WFST representation of the target word and the joint N-gram model. The most likely pronunciation is determined by extracting the shortest path through the combined machine.*

Sarrerako definizio hori eman ondoren, aipatutako lanean azalpen asko ematen dira grafema-fonema bihurketa egiten duten sistemen teoriari buruz eta haien inplementazioari buruz WFST transduktoreen bitartez. Ez dugu artikulu horretan ageri den informazio guztia hona ekartzeko asmoa, gure helburua ez baita tresnaren funtzionamendua egokitzea edo aldatzea, baina tesi-lanean planteatu ditugun helburuak lortzeko oinarrizko osagaia bihurtu denez Phonetisaurus, beharrezkoa iruditzen zaigu aurreko definizioan ageri diren terminoak argitzea.

Alde batetik, WFST teknologia aipatzen da definizioan, eta honez gero ikusi ditugu teknologia horren oinarriak IV.2 atalean. Orain, IV.3.1 atalean,

definizioan ageri diren eredu estatistikoen ezaugarriak argitzen saiatuko gara (*joint multigram*, *n-gram model*, *joint n-gram* model), eta ikusiko dugu noiz eta non erabiltzen dituen tresnak eredu horiek. Gero, IV.3.2 atalean, definizioan adierazitako azken urratsa egiteko eragiketak argituko ditugu (*weighted composition*, *shortest path*).

### IV.3.1 Eredu estatistikoak

Ikusi berri dugun definizioan (aurreko atalean kopiatutako zitan), *joint multigram* hurbilpenaren bariazio gisa deskribatzen da lehendabizi Phonetisaurus, baina gero, haren urratsak zertan diren azaltzerakoan, *n-gram* eredua aipatzen da hirugarren urratsean, eta *joint n-gram* eredua azken urratsean. Jarraian eredu horiek banan-banan deskribatuko ditugu.

#### N-gram eta multigram ereduak

*Multigram* eredu estatistikoa hizkuntza modelatzeko aukera berri gisa proposatu zen 1995ean (Bimbot *et al.*, 1995). Ordura arte, *n-gram* eredu klasikoa erabiltzen zen hizkuntza modelatzeko, zeinaren arabera hitz jakin baten probabilitatea hitz-segida batean (esaldi bat izan daiteke) aurreko $n$-1 hitzen araberakoa den; hala, segidaren probabilitatea estimatzeko, hitz guztien probabilitatea biderkatu behar da[5]. Aldiz, *multigram* ereduaren arabera, hitz-segida baten probabilitatea bat dator probabilitate handieneko segidaren segmentaziokoarekin, kontuan izanik segida barneko segmentuak luzera aldakorrekoak direla maximo bat arte, eta independenteak haien artean (Bimbot *et al.*, 1995).

Esaterako, **3-gram eredu klasikoaren arabera**, $[h_1h_2h_3h_4]$ 4 hitzeko katearen egiantza estimatzeko ($\mathcal{L}(h_1h_2h_3h_4)$ bitartez adieraziko duguna), honako kalkulu hau egiten da:

$$\mathcal{L}(h_1h_2h_3h_4) = P(h_1)P(h_2|h_1)P(h_3|h_2h_1)P(h_4|h_3h_2) \tag{IV.1}$$

[5]Hizkuntzaren modelatzeari buruz ari gara eta *hitza* hartu dugu segidaren unitate gisa, nahiz eta hori ez den aukera bakarra: segidaren unitatea grafema, silaba edota fonema ere izan daitezke.

eta **3-multigram ereduaren arabera**, berriz, beste hau:

$$\begin{aligned}\mathcal{L}(h_1h_2h_3h_4) = max\{&P([h_1h_2h_3])P([h_4]),\\ &P([h_1])P([h_2h_3h_4]),\\ &P([h_1h_2])P([h_3h_4]),\\ &P([h_1h_2])P([h_3])P([h_4]),\\ &P([h_1])P([h_2h_3])P([h_4]),\\ &P([h_1])P([h_2])P([h_3h_4]),\\ &P([h_1])P([h_2])P([h_3])P([h_4])\}\end{aligned} \tag{IV.2}$$

### *Joint multigram* eredu estatistikoa

*Multigram* eredua formulatu zen urte berean, Deligne *et al.* ikerlariek (1995) eredua zabaldu zuten, eta *joint multigram* izeneko eredua formulatu zuten, bi sinbolo-sekuentziaren arteko lerrokatze-prozesua modelatzeko. Eredua proposatzeaz gain, lan horretan ikerlariek aztertu zuten ea grafemen eta fonemen arteko lerrokatze-prozesua eredu horren bitartez modela zitekeen. Phonetisaurus eredu honen bariazioa denez, ereduaren formulazioa eta parametroak estimatzeko prozedura ulertzeko argibideak emango ditugu.

### Formulazioa

*Joint multigram* ereduak kontsideratzen du askotariko sinboloz osatutako bi kate $\begin{pmatrix} O = o_1 \dots o_T \\ \Omega = \omega_1 \dots \omega_\Theta \end{pmatrix}$ lortzen direla hainbat bikote-sekuentzia independente $\begin{bmatrix} s_t \\ \sigma_t \end{bmatrix}$ lotuz.

Hala, $(n, v)$ *joint multigram* ereduan $s_t$ sekuentzia luzeenak $n$ sinbolo izango ditu eta $\sigma_t$ luzeenak, berriz, $v$ sinbolo. Gerta daitekeenez bikote bateko $s_t$ eta $\sigma_t$ osagaiak luzera berekoak ez izatea, ereduak bi kateren arteko anitz-anitz (*many-to-many*) motako lerrokatzea onartzen duela esaten da.

Hasierako kateak, $O$ eta $\Omega$, hainbat sekuentziatan segmenta daitezke, besteak beste: $L_O$ eta $L_\Omega$. Halaber, $L = (L_O, L_\Omega)$ bi segmentazioren arteko baterako-segmentazio bateragarria da (kosegmentazioa), $O$ eta $\Omega$ uztartu eta kosekuentziatan segmentatuko dituena. Esan beharrik ez dago baterako-segmentazio bat baino gehiago egon daitekeela, eta hala, baterako-segmentazio bateragarri guztien multzoa $\{L\}$ denotatuko dugu. Kontuan izanik kosegmentazio baten barruan lotzen diren bikote-sekuentziak independenteak direla, kosegmentazio baten egiantza estima daiteke haren bikote-sekuentzien probabilitateak biderkatuz, hau da:

$$\mathcal{L}(O, \Omega, L) = \prod_t p \left[ \begin{array}{c} s_t \\ \sigma_t \end{array} \right] \tag{IV.3}$$

Eredua erabili nahi bada kosegmentazioen artean bat aukeratzeko, onartzen da $O$ eta $\Omega$ arteko kosegmentaziorik onena dela egiantza handiena lortzen duena. Egiantza onena hurbilpen honen bitartez kalkulatzen da:

$$\mathcal{L}^*(O, \Omega) = \max_{L \in \{L\}} \mathcal{L}(O, \Omega, L) \tag{IV.4}$$

Beraz, $(s_i, \sigma_j)$ bikote-sekuentzien probabilitateak izanez gero, eredua transkripzio automatikoak egiteko erabil daiteke, hau da, deskodetzeko zein den $O$ sarrerako kate bati dagokion $\widehat{\Omega}$ irteerako katerik onena:

$$\widehat{\Omega} = \arg\max_{\Omega} \mathcal{L}(\Omega|O) = \arg\max_{\Omega} \mathcal{L}(O, \Omega) \tag{IV.5}$$

### Parametroen estimazioa

*Joint multigram* eredua definitzeko, beraz, parametroen estimazioa egin behar da, hau da $(s_i, \sigma_j)$ bikote-sekuentzien probabilitateak lortu behar dira.

Demagun $\mathcal{D}$ dela $(n, v)$ *joint multigram* ereduaren hiztegia, non $(s_i, \sigma_j)$ motako bikote-sekuentzia posible guztiak ageri diren. Parametroen estimazioaren bitartez bikote guztien probabilitateak zehazten dira, $\{p(s_i, \sigma_j)\}_{i,j}$, kontuan izanik $\sum_{i,j} p(s_i, \sigma_j) = 1$ izan behar duela.

Beraz, $(O, \Omega)$ motako hainbat kate-bikotez osatutako corpus batetik abiatuta, non ezezaguna den bikote bakoitzaren kosegmentazioa, probabilitate maximo horien estimazioa lortu behar da, eta hori *Expectation-Maximization* (EM) algoritmoa (Dempster *et al.*, 1977) aplikatuz lortzen da.

#### Eredu estatistikoen erabilera Phonetisaurus tresnan

Phonetisaurusek bi eredu estatistiko ezberdin erabiltzen ditu, bi une zehatz eta zeharo ezberdinetan. Lehendabizi *joint multigram* eredua erabiltzen du ikasteko dituen bikoteak lerrokatzeko, hau da, bikote bakoitzeko grafemak eta fonemak lerrokatzeko; eta gero, bikoteak lerrokatuta, *n-gram* eredu klasikoa erabiltzen du lerrokatze bakoitzaren probabilitatea estimatzeko.

Tresnaren egileek azaltzen dutenez Novak *et al.* (2015) lanean, lerrokatze-urratsaren *joint multigram* eredu estatistikoa inplementatzeko, Jiampojamarn *et al.* lanean (2007) proposatutako algoritmoa jarraitu dute (gero zabaldua Jiampojamarn eta Kondrak (2010) lanean) baina aldaketa batzuekin. Esaterako, murriztapen bat ezarri dute, zeinaren bitartez entrenatze

prozesuan bat-bat (*one-to-one*) erlazioaz gain, kontuan hartzen diren soilik anitz-bat (*multiple-to-one*) eta bat-anitz (*one-to-multiple*) erlazioak.

Ikus dezagun lerrokatze-algoritmoak egin behar duena, ingelesezko adibide baten bitartez (erreferentziako lanean ematen duten adibideetako bat da). Demagun sarrerako informazioan honako grafema/fonema bikote ageri dela:

$$\text{TEXTBOOK} \rightarrow \text{T EH K S T B UH K}$$

Lerrokatze posible bat, sinpleena, honako hau da (bat-bat lerrokatzea eginez lortua):

$$\begin{array}{cccccccc} \text{T} & \text{E} & \text{X} & \text{T} & \text{B} & \text{O} & \text{O} & \text{K} \\ | & | & | & | & | & | & | & | \\ \text{T} & \text{EH} & \text{K} & \text{S} & \text{T} & \text{B} & \text{UH} & \text{K} \end{array}$$

baina naturalagoa eta hobea dirudi beste lerrokatze hau (bat-anitz lerrokatzea erabiliz lortua):

$$\begin{array}{ccccccc} \text{T} & \text{E} & \text{X} & \text{T} & \text{B} & \text{O,O} & \text{K} \\ | & | & | & | & | & | & | \\ \text{T} & \text{EH} & \text{K,S} & \text{T} & \text{B} & \text{UH} & \text{K} \end{array}$$

Lerrokatze-algoritmoak, beraz, bikoteetako sinboloen arteko lerrokatze onena lortzea du helburu. Algoritmo hori zehatz gauzatzeko eta adierazteko, WFST teknologia erabiltzen du Phonetisaurus tresnak. Lehenengo urratsean, lerrokatze-grafo bat eraikitzen du lexikoiko hitza-ahoskera bikote bakoitzarentzat eta lerrokatze posible guztiak probabilitate berarekin hasieratzen ditu. Hori egin ondoren, probabilitateak doitzeko *Expectation-Maximization* (EM) algoritmoa errepikatzen da aurretik ezarritako iterazio kopuru maximora iritsi arte, edo bi iterazioren arteko diferentzia aurretik ezarritako muga bat gainditzen ez duen arte. Prozesua amaitzen denean, probabilitate handieneko lerrokatzea lortu da hiztegiko bikote bakoitzeko, eta horrela osatzen da lerrokatutako sekuentziaz osatutako hiztegi edo corpus bat, ondorengo urratsekin jarraitu ahal izateko.

Bigarren ereduari dagokionez, $n$-gram eredua, Phonetisaurusek ez du ereduaren inplementazio berezirik egiten, dagoeneko garatuak diren tresnak erabiltzeko aukera ematen baitu (tresnaren tutorialean hiru tresna zerrendatzen dira urrats hori emateko: OpenGrm[6], SRILM[7] eta MITLM[8]). Aipatu beharra dago tresna horiek hizkuntza modelatzeko erabiltzen direla oro

[6] `http://www.openfst.org/twiki/bin/view/GRM/NGramLibrary`
[7] `https://www.sri.com/engage/products-solutions/sri-language-modeling-toolkit`
[8] `https://github.com/mit-nlp/mitlmetaCMU-CambridgeSLM`

har, hau da, hitz-sekuentziak modelatzeko (edo silaba-sekuentziak edo letra-sekuentziak), baina oraingo honetan eredua entrenatzeko corpusa ez dago hitzez osatua, $(s_i, \sigma_j)$ bikote-sekuentziaz baizik, hots, lerrokatze-urratsean lortutako grafema↔fonema sekuentziaz (G↔P). Hori dela eta, eredu horri "*joint n-gram*" ere esaten zaio.

### IV.3.2 Azken urratsa: deskodeketa

Aurreko atalean ikusi dugunez, Phonetisaurusek eredu estatistikoak erabiltzen ditu ikasteko ematen zaizkion datuen modelizazioa egiteko, eta modu horretan eredu bat lortzen du WFST transduktore batean jasota geratzen dena. Behin hori lortuta ikasketa-prozesua amaituta dago, eta hurrengo urratsean, deskodeketa deritzon urratsean, lortutako transduktorea sarrerako hitz berriei dagokien hipotesi onena emateko erabiltzen da, (ahoskera "onena" emateko). Hipotesi onenaren bilaketa inplementatzeko ($p_{best}$), transduktore haztatuen arteko hainbat eragiketa egin behar dira, honako adierazpen honen arabera (Novak *et al.*, 2015):

$$p_{best} = shortestpath(project_o(w \circ M)) \tag{IV.6}$$

(IV.6) adierazpenean "$w$" sarrerako hitza da eta "$M$" ikasketa-urratsetan lortutako WFST transduktorea. Halaber, $\circ$ eragileak $M$ transduktorean $w$ sarrera aplikatzen dela adierazten du, eta $project_o$ funtzioak, irteerako sinboloen proiekzioa egin behar dela (alegia, transduktorearen sarrera-irteera kate bikotetik, irteerarekin baino ez geratu). Azkenik, *shortestpath* funtzioak bide motzenaren aukeraketa adierazten du, hau da, hasierako egoeratik bukaerako egoera bateraino automatan zehar egin daitekeen kostu txikieneko bidearen aukeraketa. Aukera dago hipotesi bat baino gehiago lortzeko, hau da, $k$ hipotesi onenak lor daitezke antzeko moduan.

Deskodeketa-urratsa egiteko, gaur egun Phonetisaurus tresnak aukera bat baino gehiago eskaintzen du, baina tesi-lan honetan erabili dugun 2012ko bertsioan, aukera bakarra zegoen inplementatuta, LMBR *Lattice Minimum Bayes-Risk* deskodeketa, hori izan baitzen tresnaren egileen lehen aukera urrats hau egiteko (Novak *et al.*, 2012). Azken lanean (Novak *et al.*, 2015), ordea, aukera berri bat eskaintzen da deskodeketa egiteko, RNNLM *Recurrent Neural Network Language Model* hurbilpenean oinarrituta, baina ez dugu lan egin horrekin. Etorkizuneko lanetarako geratu da aukera berri hori probatzea, emaitzak aldatzen diren aztertzeko.

### IV.3.3 Phonetisaurusen erabilera beste lan batzuetan

Phonetisaurus tresnaren deskribapenean esaten den moduan, grafemen eta fonemen arteko bihurketak egiteko tresna gisa garatua izan da, eta logikoa denez, aplikazioa non erabili izan den bilatzen bada, batez ere hizketaren arloko aplikaziotan erabili dela aurkitzen da (Schlippe *et al.* (2014), Leidig *et al.* (2014), Su *et al.* (2014), Chen *et al.* (2015), Chen *et al.* (2016)).

Dena den, aurkitu ahal izan dugu zenbait lanetan, gurean bezala, tresna edo tresnaren zati bat erabili dela beste motatako aplikazioen barruan. Hala gertatzen da, esaterako, Eger (2015) lematizazioari buruzko lanean edota Rajan (2014) lanean, non Indiako Konkani hizkuntzarako transliterazio-sistema bat eraikitzen duten.

### IV.3.4 Phonetisaurusen erabilera aldaeren normalizazioan

Esan bezala, Phonetisaurus tresnaren funtzionamendu arrunta lau urratsetan laburbiltzen da:

1. Datuak prestatu behar dira, hau da, lexikoi bat osatu behar da ikasteko informazioarekin, tresna gai baita ematen zaizkion datuetatik "ikasteko". Oro har, sarrerako datu horiek hitzez (grafemak) eta haien ahoskeraz (fonemak) osatutako bikoteak izango dira.

2. Lerrokatu egin behar dira lexikoiko bikoteak, hau da, mapatu edo lotu behar dira bikote bakoitzaren bi alderdietako sinboloak (oro har, grafemak eta fonemak). Hori egiteko, azaldutako *joint multigram* eredua erabiltzen da eta hala, lexikoiari dagokion $(s_i, \sigma_j)$ bikote-sekuentziaz osatutako corpus bat lortzen da. Lerrokatze horretan 1-anitz, anitz-1 eta 1-1 motatako sekuentziak lor daitezke.

3. Entrenatu behar da *joint n-gram* eredu bat aurreko urratsean lortutako corpus berriarekin. Eredu hori entrenatzeko hainbat tresna daude eskuragarri (OpenGrm, SRILM, MITLM) eta urratsa amaitzeko, horietako edozeinekin lortutako eredua WFST transduktore gisa adierazi behar da dagokion bihurketa eginez.

4. Deskodeketa-urratsean lortu berri den WFST transduktorea erabiltzen da sarrerako kate berriei (oro har, hitzei) dagokien irteerako katerik onena (oro har, ahoskera onena) lortzeko.

Ikus dezagun nola aplikatu diren urrats horiek gure sisteman aldaeren normalizazioa lortzeko.

| Aldaera | Estandarra | Estandarra "banatuta" |
|---|---|---|
| akhabatzeko | akabatzeko | a k a b a t z e k o |
| arteiño | arteraino | a r t e r a i n o |
| baitzeien | baitzitzaien | b a i t z i t z a i e n |
| darabillatela | darabiltela | d a r a b i l t e l a |
| derautzatzunak | dizkiozunak | d i z k i o z u n a k |
| egoiteak | egoteak | e g o t e a k |

IV.1 Taula: Lexikoiaren adibide batzuk. Phonetisaurusi ematen zaion informazioa da lehenengo eta hirugarren zutabeek osatzen dutena (erdikoa ez). Tresnak fonema-kate gisa ulertzen du bikotearen bigarren zatia, eta fonemak zuriunez banatuta egotea espero du (hirugarren zutabean adierazten diren moduan).

#### IV.3.4.1 Datu-prestaketa

Gure kasuan prestatu behar dugun lexikoiko sarrera bakoitza bi hitzez osatzen da: alde batetik, aldaerako hitza (dialektoa, euskara zaharra...) eta beste aldetik, anotazio-prozesuan hitz horri esleitu zaion euskarazko hitz estandarra. Adibide batzuk IV.1. taulan ageri dira (*Gero* obraren ikasteko zatian anotatutako adibideak dira).

#### IV.3.4.2 Datuen lerrokatzea

Datuen arteko lerrokatzea funtsezkoa da eta azaldu ditugu lerrokatze hori egiteko erabiltzen den eredu estatistikoaren oinarriak. Bikoteetako bi aldeetako sinbolo kopuruak (gure kasuan grafemak-grafemak) bat ez badatoz, beharrezkoa izan daiteke grafema nulua ($\epsilon$) erabiltzea lerrokatzean.

Lerrokatze-prozesua egin ondoren, hainbat bikote-sekuentziaz osatutako hiztegi edo corpus berri bat lortzen da ondorengo urratsekin jarraitu ahal izateko. IV.2 taulan corpus berri horretan ageri diren lerrokatze batzuk ikusten dira, IV.1 taulako hiztegiko sarrerei dagozkien lerrokatzeak, hain zuzen. Kontuan izan behar da, dena den, lerrokatze-algoritmoak askoz sarrera gehiago dituen hiztegiarekin egiten duela lan (horiek adibide batzuk besterik ez dira).

Lerrokatze-algoritmoaren inplementazioak urrats asko ditu eta konplexua izan daiteke horien xehetasunak ulertzea. Halere, prozesu hori guztiz gardena da tresnaren erabiltzailearentzat, zeinak Phonetisaurusek duen pro-

| Lerrokatutako sarrera |
|---|
| a:a k,h:k a:a b:b a:a t:t z:z e:e k:k o:o |
| a:a r:r t:t e:e i:r,a ñ:i,n o:o |
| b:b a:a i:i t:t z:z e:i,t $\epsilon$:z,a i:i e:e n:n |
| d:d a:a r:r a:a b:b i:i l,l:l a:$\epsilon$ t:t e:e l:l a:a |
| d,e:d r,a:i u:z,k t,z:$\epsilon$ a:i,o t,z:z u:u n:n a:a k:k |
| e:e g:g o:o i,t:t e:e a:a k:k |

IV.2 Taula: Lexikoiko sarrerak lerrokatu ondoren bikote-sekuentziaz osatutako corpus berri bat lortzen da. Bikote horietan "," ikurrak adierazten du sekuentzia sinbolo anitzekoa dela (1-anitz, anitz-1 eta 1-1 sekuentziak ageri daitezke) eta "$\epsilon$" ikurrak sinbolo nulua adierazten du.

grama jakin bat besterik ez duen exekutatu behar bikoteen arteko lerrokatzea lortzeko[9].

#### IV.3.4.3 *Joint n-gram* ereduaren entrenatzea

Aurreko ataleko lerrokatze-urratsaren ondorioz, corpuseko informazioaren ikuspegia aldatu da eta IV.2 taulan ageri diren adibideen itxura du, hau da, bikote-sekuentziaz osatutako corpusa da orain. Corpus horrekin $n$-gram eredu bat entrenatu behar da urrats honetan, eredu horrek ahalbidetuko baitu ikasi ez diren datu berriei dagozkien irteerako hipotesi onenak bilatzea gero.

Hainbat tresna daude eskura *joint n-gram* eredu hori entrenatzeko kapitulu honen IV.3.1 atalean esan dugun moduan, eta gure kasuan OpenGrm erabili dugu[10].

#### IV.3.4.4 Deskodeketa

Deskodeketa-urratsean aurretik entrenatutako eredua erabiltzen da sarrerako hitz berriei dagokien hipotesi onena lortzeko, hau da, gure kasuan sarrerako aldaera berriei dagokien idazkera "berri" onena lortzeko, beti ere ikasitakoaren arabera.

[9]Programa `phonetisaurus-align` da eta B eranskinean ematen dira bere erabilerari buruzko xehetasunak.

[10]OpenGrm tresna erabiltzeko komandoen xehetasunak B eranskinean ematen dira.

| Sarrera | Kostua | Hipotesia |
|---|---|---|
| aiphatzen | 9.29055 | `aipatzen` |
| aiphatzen | 17.0441 | `aipetzen` |
| aiphatzen | 17.9131 | `haipatzen` |
| aiphatzen | 18.6881 | `aipatzen_di` |
| aiphatzen | 19.0471 | `aipezatzen` |
| etzaituztet | 21.3908 | `ez_zaituztet` |
| etzaituztet | 22.882 | `ez_zaituzte` |
| etzaituztet | 23.7063 | `ez_zaituzkiet` |
| etzaituztet | 24.7063 | `ez_zaituztat` |
| etzaituztet | 25.3413 | `ez_zaituzdt` |

IV.3 Taula: Deskodeketa-urratsean lortutako erantzunen adibideak. Sarrera bakoitzeko 5 hipotesi onenak eskatu dira eta horiek kostuaren arabera ordenatuta lortzen dira (lehen posizioan hipotesi onena).

Ikusi dugu hainbat eragiketa egin behar direla hori lortzeko, baina erabiltzailearen ikuspuntutik oso sinplea da deskodeketa-urratsa betetzea. Phonetisaurusek duen programa jakin bat exekutatu behar da, zeinari sarrera gisa deskodetu nahi diren hitzen zerrenda ematen zaion, eta irteera gisa hitz bakoitzari dagokion hipotesirik onena (edo onenak, bat baino gehiago eskatu bazaio) itzultzen duen[11]. Esaterako, IV.3 taulan *aiphatzen* eta *etzaituztet* bi aldaerei dagozkien hipotesiak ageri dira. Adibide horretan argi ikusten denez hipotesi bat baino gehiago eskatu da (5 zehazki), eta emaitzak kostuen arabera ordenatuta irteten dira: zenbat eta kostu txikiagoa, orduan eta hipotesi hobea[12].

## IV.4 *Gero* corpusa: esperimentuak eta emaitzak

Atal honetan *Gero* obraren corpusarekin egin ditugun esperimentuen berri emango dugu. II. kapituluko II.3.4 atalean, corpusaren prestaketa azaldu da xehetasunez, eta orain, corpusaren ezaugarriak laburbilduko ditugu esperimentuekin hasi baino lehen.

[11] Programa zehatza `phonetisaurus-g2p` da eta erabiltzeko xehetasunak B eranskinean ematen dira.

[12] Adibidean gertatzen ez bada ere, suerta daiteke eskatutakoa baino emaitza gutxiago lortzea irteeran, beti ezin baita ziurtatu bide kopurua automatan zehar.

| | Analizatutako | | OOV | |
|---|---|---|---|---|
| | tokenak | formak | tokenak | formak |
| Ikasi | 8.223 | 3.025 | 1.931 | 1.032 |
| Test | 4.386 | 1.902 | 1.015 | 636 |

IV.4 Taula: *Gero* corpusa. Ikasteko zein testeko zatien ezaugarriak.

Gogora dezagun, labur-labur, corpusa prestatzeko prozesua: obraren bi zati aukeratu dira zoriz (bat ikasteko eta bestea ebaluatzeko), eta horietan hainbat OOV detektatu dira automatikoki, gero eskuz anotatuak izan direnak Brat anotazio-tresnaren bitartez (ikus A eranskina anotazioari buruzko argibideetarako). IV.4 taulara corpusaren tamainari buruzko datuak ekarri ditugu berriro.

Phonetisaurus tresna erabili behar dugu esperimentuetan, baina corpus dialektalarekin egin dugun antzera hirugarren kapituluko III.4.5 eta III.4.6 ataletan, lehendabizi hainbat proba egin ditugu sistema doitzeko, horrela jakiteko zein den ikasteko informaziorik egokiena eta zein diren deskodeketarako parametro onenak.

Doikuntzako esperimentuak ikasteko corpusarekin egin ditugu, eta corpus hori jadanik lau fitxategitan banatuta dagoenez (anotazio-lana egiteko banatu da), lau zati horiek baliatu ditugu doikuntzako esperimentuetan lau ataleko balidazio gurutzatua egiteko.

Dena den, esperimentuekin hasteko, anotatuta dugun ikasteko corpusetik erauzi behar dugu informazioa, eta aurrera egin baino lehen argi utzi nahi dugu zein izan den erauzitako informazioa, eta zer nolako kasu bereziak aurkitu ditugun informazio horren barruan, normalizazioari begira.

Anotatu diren OOVen artean interesatzen zaizkigun bakarrak *Aldaera* eta *Zuzena* etiketak dituztenak dira (latinezko hitzak edo beste etiketak dituztenak ez ditugu normalizazio-lanean sartuko), eta horiek soilik erauzi ditugu anotatutako fitxategietatik (anotazio-lanaren xehetasunak A eranskinean). Hurrengo urratsa konprobaketa-urratsa izan da, eta hala, *Aldaera* gisa etiketatutako kasuetan aztertu egin da ea anotatzaileak eman duen estandarra onartzen duen gure automatak[13]. Azterketan ikusi denez, zenbait kasutan ez da onartzen "estandar" hori:

1. Aldaera baten baliokide gisa hitz bat baino gehiago eman behar denean, hitzak "lotuta" eman ditu anotatzaileak azpimarra baten bitar-

[13]Konprobaketa horren bitartez anotazio-errore batzuk zuzendu ahal izan dira.

tez, adibidez: *etzaigu* → *ez_zaigu*. Estandarraren automata prestatu dugu halako loturak onartzeko, baina ez edozein bi hitzen artekoa, honako murriztapen hau jarri baita transduktore hori sortzeko unean: lotura onartzen du baldin eta bigarren hitza aditz laguntzailea bada. Murriztapen hori betetzen ez duten kasuak ez ditu onartzen, eta horietako batzuk daude, esaterako: *diozunorrek* → *diozun_horrek*, *esperantzaturik* → *itxaropena_izanik* eta *Eztagoela* → *Ez_dagoela*.

2. OOV hitzak detektatzen dituen automatak ez ditu onartzen forma hobetsia duten hitzak[14], beraz, anotatzaileak horrelakoren bat eman badu baliokide estandar gisa, automatak ez du onartuko. Ikasteko zatian ez da horrelako kasurik geratu azkenean erroreak zuzendu eta gero, baina kontuan hartu behar da gerta daitezkeela horrelako kasuak testeko corpusean.

Aipatu berri ditugun kasu horiek ez dira datuen artetik kendu esperimentuetarako[15], baina ziur dakigu horiek ezingo ditugula ongi ebatzi gure sistemarekin, transduktore estandarrak ez dituelako onartzen (eta hori derrigorrezko baldintza da ematen diren erantzun guztietarako).

Gure sistemak izango duen beste muga *Zuzen* gisa etiketatutako kasuekin gertatzen da. Mota horretakoak dira, batez ere, izen propioak: *Krisostomo*, *Kalef* eta abar. Zenbait izen propio ez daudenez euskara estandarraren lexikoaren barruan, gure automatak ez ditu onartzen (horregatik, hain zuzen, markatu dira OOV gisa anotazio-prozesua baino lehen) baina anotatzaileak zuzenak kontsideratu ditu. Euskara estandarraren iragazkia aplikatuz gero, kasu horiek ere ezin izango ditugu ongi ebatzi lehen aipatutako arrazoi berberagatik: transduktore estandarrak ez dituelako onartzen hitz horiek. Argitu beharra dago, dena den, izen propio guztiak ez direla beti zuzen gisa kontsideratuak izan, eta badirela aldaera gisa anotatutakoak ere, esaterako: *Augustin* → *Agustin*, *Aristotelek* → *Aristotelesek* eta abar.

Informazioan aurkitu ditugun mugak azalduta, doikuntzako esperimentuekin hasiko gara. Hiru esperimentu planteatu ditugu sistemaren doikuntzarako:

- Lehenengo esperimentua oinarrizkoa izan da: anotatutako bikoteak

[14] Euskaltzaindiaren *Hiztegi Batu*ko hainbat formak "h." marka dute beste forma estandar bat hobesten delako. Esaterako, *bertze* sarreraren forma hobetsia *beste* da.

[15] Alboratu den kasu bakarra hitz anitzekoa izan da, hau da, aldaera hitz bat baino gehiagoz osatuta dagoenean, esaterako"*gogan behar*" bi hitzak, "*goganbehar*" estandarrarekin etiketatu dira eta kasu hori ez da sartu esperimentuetan. Salbuespena da, dena den.

soilik erabili dira ikasteko, eta ikasketa zein testa formen arabera egin da (bikote bakoitza behin).

- Bigarren esperimentuan aztertu da zenbateraino aldatzen diren emaitzak formekin lan egin beharrean, tokenekin egiten bada lan (bikote bakoitza ageri den bezainbeste aldiz).
- Hirugarren eta azken esperimentuan analizatu da ea emaitzak aldatzen diren ikasteko informazio gisa erabiliz gero anotatutako bikoteak ez ezik, hitz estandarrak ere.

### IV.4.1 Oinarrizko doikuntza (1. esperimentua)

Ikasteko ditugun lau fitxategietatik *Aldaera* zein *Zuzena* etiketadun bikoteak erauzi eta filtratu ondoren, anotatutako formen zerrendak prestatu dira (forma bakoitza behin bakarrik utzi da fitxategi bakoitzean, nahiz xehez edo larriz ageri), eta bikote guztiak xehez idatzi dira. Hori egin ondoren, balidazio gurutzatua egiteko 4 atalak prestatu dira: atal bakoitzean 3 fitxategi ikasteko erabiltzen dira, eta laugarrena ebaluatzeko. Ikasteko 3 fitxategiak elkartzen direnean, ziurtatu behar da berriro bikote bakoitza behin bakarrik ageri dela. IV.5 taulan atal bakoitzaren tamainak edo kopuruak ageri dira.

| | A1 | A2 | A3 | A4 |
|---|---|---|---|---|
| Ikasi | 780 | 785 | 748 | 744 |
| Test | 285 | 294 | 332 | 337 |

IV.5 Taula: *Gero*. Lehenengo esperimenturako balidazio gurutzatuko atalek dituzten bikote kopuruak. Bikote bakoitza behin fitxategi bakoitzean.

Ikasteko zein ebaluatzeko fitxategiak prest ditugula, hurrengo urratsa da ikasteko metodoak aplikatzea eta bakoitzaren ebaluazioa egitea.

#### Oinarri-lerroa

III. kapituluan corpus dialektalarekin egin diren esperimentuetan oinarri-lerroko sistema bat baino gehiago planteatu dugu. Kontuan izanik Levenshtein distantzian oinarritutako sistemak emaitza kaskarrak lortu dituela corpus horretan, eta oinarri-lerroko bi sistemen arteko konbinazioak ere, ez duela kuantitatiboki asko hobetu $F$ parametroaren balioa, oraingo honetan oinarri-lerro bakarra inplementatu dugu: memoria duen oinarri-lerroa.

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| A1 | 98,18 | 37,89 | 54,68 |
| A2 | 95,28 | 41,16 | 57,48 |
| A3 | 96,83 | 36,75 | 53,28 |
| A4 | 96,85 | 36,39 | 53,02 |
| bb. | 96,78 | 38,05 | **54,61** |
| d.e. | 1,19 | 2,17 | 2,05 |

IV.6 Taula: *Gero*. Oinarri-lerroaren emaitzak lehenengo esperimentuan.

Sistema horren funtsa zein den III.4.2 atalean ikusi dugu: ikasteko bikoteak memorizatzen dira, eta gero testean ikasitakoren bat ageri bada, gordetatako informazioa ematen da; ez bada ikasi, ez da erantzunik ematen. Gerta daiteke *Aldaera* gisa anotatutako hitz bati ez esleitzea beti estandar bera, eta halako kasuetan, sistemak memorizatutako "lehenengo" estandarra ematen du erantzun gisa, ez baitauka informazio gehiago kontuan hartzeko (kasu guztiak behin bakarrik ageri dira).

Metodo honen bitartez lortzen diren emaitzak IV.6 taulan ageri dira. Bertan, atal bakoitzean lortutako balioak zein atalen arteko batezbestekoak (bb.) eta desbideratze estandarra (d.e.) ageri dira. Ikusten denez, oinarri-lerro honek lortzen duen doitasuna handia da, batez beste % 96,78, baina estaldura txikia da, batez beste % 38,05, eta ondorioz, lortzen den $F$ neurriaren balioa ere nahiko txikia da, % 54,61 batez beste.

### WFST metodoa

Aurreko kapituluan egin ditugun esperimentuen arabera eta III.5 atalean ondorioztatu duguna jarraituta, bi esperimentu nagusi egin nahi ditugu Phonetisaurus tresnaren bitartez *Gero* obraren corpus berrian. Bi esperimentu horiek ezberdinak dira ikasteko informazioaren ikuspuntutik: lehenengoan *hitza-hitza* informazioa eman zaio aplikazioari ikas dezan, hau da, aldaerako hitza eta horri dagokion estandarra; eta bigarrenean, *hitza-analisia* informazioa eman zaio, hau da, aldaerako hitza eta anotazioan horri esleitu zaion hitz estandarraren analisi morfologiko sinplifikatua.

Analisi sinplifikatu hori lortzeko transduktore bat daukagu, zeinari hitz estandar bat emanda, hitz horri dagokion analisi sinplifikatua itzultzen duen. Baina bi kasu berezi gerta daitezke: 1) zenbait kasutan bikoteko alde estandarra ez du onartzen transduktoreak (azpimarraz lotutako kasuren bat, izen

propio batzuk), eta beraz, ez du analisirik itzultzen; 2) transduktoreak analisi bat baino gehiago itzul dezake, eta kasu horietan zer egin erabaki behar da.

Analisirik izango ez dutenen artean daude, esaterako, zuzentzat etiketatu diren hitzak (horietako gehienak izen propioak). Kasu horietan erabaki da hitza bera ematea analisi gisa (adibidez, *Krisostomok* → *Krisostomok*).

Bigarren kasuari dagokionez, transduktoreak analisi bat baino gehiago itzultzen duenean, analisi horietako bat aukeratzea erabaki da: lema luzeena duen analisia. Baldintza hori ez bada nahikoa, atzizki gutxien duen analisia aukeratuko da (analisi-kate motzena izatearekin parekatu dena). Adibidez, anotazio-prozesuan *arazitzen* aldaera *arazten* estandarrarekin parekatu du anotatzaileak eta hitz estandar hori analizatzean, sei analisi posible hauek itzultzen ditu automatak (parentesi artean ageri da egin den programak lematzat hartuko duena analisi horretan):

| | |
|---|---|
| `aratz+te+n` | (lema `aratz`) |
| `aratz+ten` | (lema `aratz`) |
| `ara+araz+te+n` | (lema `ara`) |
| `ara+araz+ten` | (lema `ara`) |
| `araz+ten` | (lema `araz`) |

Lema luzeena lehenengo biek dute, `aratz`, eta horien artean bigarrenak du kate motzena, beraz, hori aukeratzen da ikasten emateko, eta hiztegirako osatutako sarrera da: *arazitzen* → `aratz+ten`.

### Hitza-hitza

Egin dugun lehenengo proban, erantzun bakarra eskatu zaio Phonetisaurusi eta itzuli duen erantzuna "ontzat" hartu da inongo azterketarik gabe, hau da, hitz estandar zuzen bat den aztertu gabe. Horrela eginez gero, sistemak beti ematen du erantzun bat eta bakarra, eta ondorioz, $P$, $R$ eta $F$ balioak berdinak dira (zenbait testuingurutan *Accuracy* esaten zaio balio horri). Lortutako balioak IV.7 taulan ageri dira, eta, argi ikusten denez, oinarrilerroarekiko lortutako hobekuntza 21 puntukoa da batez beste. Oso emaitza ona da kontuan izanik ez dela inolako iragazkirik aplikatu, eta, gainera, corpus dialektalarekin lortutako emaitzetatik gertu dago (ikus hirugarren kapituluko III.7 taulan ageri diren $F$ parametroaren balioak).

Dena den, hirugarren kapituluko esperimentuetan ikusi dugunez, erantzun gehiago eskatzea Phonetisaurusi eta gero erantzunak iragaztea, oro har,

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $P = R = F_1$ | 76,14 | 78,91 | 74,70 | 75,07 | **76,21** | 1,90 |

IV.7 Taula: *Gero*. WFST metodoaren emaitzak lehenengo esperimentuan. Ikasteko *hitza-hitza* erabili da eta eskatutako erantzun kopurua 1 izan da (iragazkirik ez).

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $n = 5$ | 85,34 | 86,69 | 86,04 | 84,94 | **85,75** | 0,77 |
| $n = 10$ | 85,07 | 86,83 | 85,81 | 84,95 | 85,67 | 0,86 |
| $n = 20$ | 84,92 | 86,68 | 85,49 | 83,90 | 85,25 | 1,16 |
| $n = 30$ | 84,66 | 86,42 | 85,67 | 84,08 | 85,21 | 1,04 |

IV.8 Taula: WFST metodoaren emaitzak lehenengo esperimentuan ($F$ neurriaren balioak). Ikasteko *hitza-hitza* erabili da eta eskatutako erantzun kopurua aldatzen doa ($n$).

emaitzak hobetzen ditu, eta hori izan da hurrengo proba. Hirugarren kapituluko III.4.5 atalean egin ditugun antzeko probak errepikatu ditugu, oraingoan 5 erantzunetatik aurrera (5, 10, 20 eta 30 erantzun eskatzen). Lortutako erantzunetatik automatak onartzen ez dituenak iragazi dira (ez dira estandarrak), eta geratzen direnen artean, lehen posizioan dagoena aukeratu da (WFST transduktorearen arabera, onena). Iragazkia dela eta, orain ezin da ziurtatu sarrerako aldaerak beti izango duenik erantzun bat eta beraz, $P$, $R$ eta $F$ ez dira zertan berdinak izan.

IV.8 taulan $F$ neurriaren balioak besterik ez ditugu jarri taula gehiegi ez luzatzearren, 4 atal baitira ditugunak. Bukaerako bi zutabeetan, errenkada bakoitzeko batezbestekoa (bb.) eta desbideratze estandarra (d.e.) adierazi dira. Erantzun kopurua handitu ahala, $F$ balioak ez dira ia aldatzen, zeren estaldura pixka bat handitzen da baina doitasuna gutxitu egiten da, eta, hala, $F$ ez da ia aldatzen (doitasunaren eta estalduraren balioak ez dira taulan ageri). Dena den, emaitza onena 5 erantzun eskatuta lortu da, corpus dialektalarekin gertatu den bezala (ikus III.7 taula).

Beraz, garbi geratzen da erantzun gehiago eskatzea eta gero horiek iragaztea onuragarria dela. Izan ere, IV.7 eta IV.8 tauletako emaitzak konparatuta 9 puntu inguruko hobekuntza ikusten da (nahiz eta erantzun bakarra eskatu denean ez den iragazkirik aplikatu).

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $n = 5$ | 85,38 | 83,96 | 81,65 | 82,76 | 83,44 | 1,60 |
| $n = 10$ | 85,50 | 84,25 | 82,22 | 83,23 | 83,80 | 1,40 |
| $n = 20$ | 85,12 | 84,57 | 83,55 | 83,12 | **84,09** | 0,92 |
| $n = 30$ | 85,02 | 84,38 | 83,28 | 83,12 | 83,95 | 0,91 |

IV.9 Taula: WFST metodoaren emaitzak lehenengo esperimentuan ($F$ neurriaren balioak). Ikasteko *hitza-analisia* erabili da eta eskatutako erantzun kopurua aldatzen doa ($n$).

### Hitza-analisia

Aurrekoan bezala, erantzun kopuru ezberdinak eskatu dira eta IV.9 taulan jaso dira emaitzak. Emaitza horiek konparatzen badira IV.8 taulakoekin, argi ikusten da lortutako $F$ balioa, oro har, baxuagoa dela (1,7 puntu inguru) ikasteko informazio gisa hitz estandarraren analisia erabiltzen bada hitzaren ordez. Horrez gain, argi geratzen da analisi morfologikoa erabiliz gero ikasteko, erantzun gehiago eskatu behar zaizkiola Phonetisaurusi emaitza onena lortzeko: hitzarekin ikasiz gero nahikoa zen 5 erantzun eskatzea, eta orain, ordea, 20 eskatu behar dira.

#### Lehenengo esperimentuaren ondorioak

Lehenengo esperimentuaren emaitzek argi eta garbi uzten dute WFST metodoa oinarri-lerroa baino askoz hobea dela: $F$ neurriaren balioa oinarri-lerroko sistemarekin % 54 inguruan dago eta WFST metodoarekin % 85 edo % 84 inguruan.

Ikasteko informazioari dagokionez, *hitza-hitza* erabiltzeak *hitza-analisia* baino emaitza hobeak lortu ditu, nahiz eta diferentzia ez den oso handia: 1,7 puntu inguru $F$ neurrian. Dena den, emaitza hori berria da, zeren corpus dialektalarekin egindako esperimentuetan emaitzak oso antzekoak ziren bi kasuetan (gogoratu III.11 taula). Kontuan izan behar dugu *Gero* corpusean fenomeno berri bat gertatzen ari dela analisiak erabiltzeko unean, hainbat hitzek ez dutelako analisirik. Kasu horietan hitza bera eman da analisi gisa, eta horrek, agian, "zarata" sortu du ikasteko unean eta emaitza okertzeko arrazoi bat bilakatu da.

Azkenik, Phonetisaurusi eskatu behar zaion erantzun kopuruari dagokionez, lehenengo esperimentu honetan aurreko kapituluan ikusi duguna errepikatu da (III.4.5 eta III.4.6 ataletan): *hitza-hitza* informazioa erabiliz gero

ikasteko, 5 erantzun eskatu behar dira emaitza onenak lortzeko eta *hitza-analisia* informazioa erabiliz gero, 20 erantzun eskatu behar dira.

### IV.4.2 Maiztasunaren eragina (2. esperimentua)

Bigarren esperimentuan aztertu da formen maiztasunaren eragina prozesuan. Hori dela eta, bai ikasteko eta bai ebaluatzeko, anotatuak dauden bikoteak bere horretan utziko dira, errepikatuak kendu gabe. Beraz, tokenak hartu ditugu kontuan lan egiteko bigarren esperimentu honetan.

Datuak prestatzeko, anotatutako lau fitxategietatik *Aldaera* zein *Zuzena* etiketa duten bikote guztiak hartu dira, oraingo honetan errepikatuak kendu gabe. Lehen bezala guztia letra xehez jarri da eta balidazio gurutzatua egiteko fitxategi berriak prestatu dira. Errepikapenak direla eta, bikote kopuruak handiagoak dira orain. IV.10 taulan atal bakoitzaren kopuruak ageri dira eta aurreko esperimentukoak baino handiagoak dira (ikus IV.5 taula).

| | A1 | A2 | A3 | A4 |
|---|---|---|---|---|
| Ikasi | 1.426 | 1.415 | 1.343 | 1.294 |
| Test | 400 | 411 | 483 | 532 |

IV.10 Taula: *Gero*. Bigarren esperimenturako kopuruak atal bakoitzean. Bikote errepikatuak ez dira kendu.

Fitxategi berriak prest, berriro ikasteko metodoak aplikatu eta ebaluatu behar dira.

#### Oinarri-lerroa

Oinarri-lerroaren funtsezko ideia ez da aldatzen: ikasteko bikoteak memorizatu, eta ebaluazioan ikasitako hitzen bat ageri bada, bere baliokidea eman. Oraingoan, dena den, bikote bat behin baino gehiagotan ageri daitekeenez, maiztasunaren informazioa ere gorde da. Hala, beharrezkoa izanez gero, ebaluazioan informazio hori erabil daiteke maiztasun handieneko erantzuna emateko.

Oinarri-lerroak lortzen dituen emaitzak IV.11 taulan ageri dira (batezbestekoa eta desbideratze estandarra eman dira soilik). Konparatzen badira emaitza horiek aurreko esperimentukoekin (IV.6 taulako azken bi errenkadak), argi ikusten da maiztasuna kontuan izatea onuragarria dela zenbakiak

| | ***P*** | ***R*** | $\boldsymbol{F_1}$ |
|---|---|---|---|
| bb. | 98,61 | 50,87 | **67,09** |
| d.e. | 1,19 | 2,11 | 0,75 |

IV.11 Taula: *Gero*. Oinarri-lerroaren emaitzak bigarren esperimentuan. Bikoteak behin baino gehiago ageri daitezke, bai ikasteko unean, bai ebaluatzeko unean.

| | ***P*** | ***R*** | $\boldsymbol{F_1}$ |
|---|---|---|---|
| bb. | 98,00 | 38,55 | **55,29** |
| d.e. | 1,33 | 2,03 | 1,86 |

IV.12 Taula: *Gero*. Oinarri-lerroaren emaitzak bigarren esperimentuan. Bikoteak behin baino gehiago ageri daitezke ikasteko unean, baina ez test egiteko unean (1. esperimentuko test zerrendak).

emateko unean, batez besteko $F$ neurriaren balioa hamahiru puntu inguru igo baita: % 54,61 baliotik % 67,09 baliora. Diferentzia hori estalduraren handitzeari dagokio batez ere, doitasuna altua baita bi esperimentuetan.

Dena den, ez da ahaztu behar bi esperimentuetako test zerrendak ez direla berdinak, test berri honetan hitz bat behin baino gehiago ageri baitaiteke (ikus IV.10 taulako kopuruak). Beraz, konparazioa garbiagoa izan dadin, beste test bat egin dugu baina aurreko esperimentuko test zerrendak mantenduz, hau da, ikasteko unean kontuan hartu dugu hitzen maiztasuna, baina test egiteko unean ez. Lortutako emaitzak IV.12 taulan ageri dira. Konparatzen badira horiek eta IV.6 taulakoak, garbi dago emaitzak ez direla ia aldatzen: batez beste, bai doitasuna eta bai estaldura zertxobait igotzen dira (doitasuna gehiago estaldura baino) eta ondorioz $F$ neurria ere pixka bat igotzen da, 0,6 puntu inguru.

### WFST metodoa

Ikusita zer gertatu den oinarri-lerroko sisteman hitzen maiztasuna kontuan hartu denean, WFST metodoarekin ere espero daiteke zenbaki altuagoak lortzea testeko zerrendan hitz errepikatuak badaude. Ez dagoena batere garbi da zer gertatuko den testeko zerrenda lehenengo esperimentukoa bada, hau da, maiztasuna kontuan hartzen bada ikasteko soilik.

Beraz, WFST metodoaren emaitzak konparagarriak izan daitezen oinarri-lerroko sistemakoekin, horretan egin diren bi testak errepikatu dira: lehenengo testean zerrenden kopuruak IV.10 taulakoak izan dira (bikote bat behin baino gehiago ageri daiteke), eta bigarrenean, berriz, kopuruak IV.5 taulakoak dira. Ikasteko beti erabili dira IV.10 taulako kopuruak.

Horrez gain, lehenengo esperimentuan garbi geratu denez hobe dela erantzun bat baino gehiago eskatzea eta gero horiek filtratzea (IV.4.1 atalean), bigarren esperimentu honetan hori egin da hasieratik.

### Hitza-hitza

Phonetisaurusi *hitza-hitza* eman diogu ikasteko eta bi testak egin ditugu. Lehenengoaren emaitzak IV.13 taulan ageri dira eta, espero zitekeen moduan, $F$ neurriaren balioa altuagoa da testean errepikatutako bikoteak badaude, nahiz eta diferentzia ez den oinarri-lerroko sisteman suertatutakoa bezain handia: 13 puntu inguruko aldea lortu da oinarri-lerroan eta WFST metodoan, berriz, 3 puntu ingurukoa.

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $n = 5$ | 88,68 | 89,14 | 89,59 | 88,69 | **89,03** | 0,43 |
| $n = 10$ | 88,60 | 88,95 | 89,54 | 88,30 | 88,85 | 0,53 |
| $n = 20$ | 88,51 | 89,23 | 89,48 | 88,05 | 88,82 | 0,66 |
| $n = 30$ | 88,08 | 88,89 | 89,63 | 87,96 | 88,64 | 0,78 |

IV.13 Taula: *Gero*. WFST metodoaren emaitzak bigarren esperimentuan ($F$ neurriaren balioak). Ikasteko *hitza-hitza* erabili da eta bikote kopuruak IV.10 taulakoak dira. Eskatutako erantzun kopurua ($n$) aldatzen doa.

Dena den, lehen esaten genuenaren ildotik, konparazio hori egiten denean testeko zerrenda ezberdinak hartzen ari dira kontuan. Interesgarriagoa da aztertzea zer gertatzen den lehenengo esperimentuko testeko zerrendak erabiliz gero, eta hori da IV.14 taulako emaitzetan jaso dena. Ikusten denez, anotatutako bikoteen maiztasuna kontuan hartzeak soilik ikasteko ez ditu emaitzak hobetzen (IV.8 taulako emaitzekin konparatu behar dira). Izan ere, batez besteko $F$ onenaren balioa pixka bat jaitsi da: % 85,75etik % 85,41era.

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $n = 5$ | 85,34 | 86,69 | 85,07 | 84,54 | **85,41** | 0,92 |
| $n = 10$ | 85,23 | 86,43 | 85,03 | 84,11 | 85,20 | 0,95 |
| $n = 20$ | 85,13 | 86,63 | 84,99 | 83,72 | 85,12 | 1,19 |
| $n = 30$ | 84,56 | 86,17 | 85,22 | 83,59 | 84,89 | 1,09 |

IV.14 Taula: *Gero*. WFST metodoaren emaitzak bigarren esperimentuan ($F$ neurriaren balioak). Ikasteko *hitza-hitza* erabili da. Ikasteko bikoteak behin baino gehiago ageri daitezke baina testekoak ez (lehenengo esperimentuko test zerrendak dira).

### Hitza-analisia

Phonetisaurusi *hitza-analisia* eman diogu ikasteko eta bi testak egin ditugu berriro. Lehenengoaren emaitzak IV.15 taulan ageri dira. Oraingo honetan ere, $F$ neurriaren balioa altuagoa da testean errepikatutako bikoteak badaude eta aldea antzekoa da berriro, 3,8 puntu ingurukoa.

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $n = 5$ | 88,29 | 87,35 | 87,24 | 86,68 | 87,39 | 0,67 |
| $n = 10$ | 88,95 | 87,66 | 87,79 | 86,93 | **87,83** | 0,83 |
| $n = 20$ | 88,45 | 87,84 | 87,77 | 87,15 | **87,80** | 0,53 |
| $n = 30$ | 88,37 | 87,87 | 87,61 | 87,00 | 87,71 | 0,57 |

IV.15 Taula: *Gero*. WFST metodoaren emaitzak bigarren esperimentuan ($F$ neurriaren balioak). Ikasteko *hitza-analisia* erabili da eta bikote kopuruak IV.10 taulakoak dira.

Lehenengo esperimentuko testeko zerrendak erabilita lortutako emaitzak IV.16 taulan jaso dira, eta lehen egin dugun analisiak berriro balio du: bikoteen maiztasuna kontuan hartzeak soilik ikasteko, ez ditu emaitzak hobetzen (orain IV.9 taularekin konparatu behar da) ia berdina baita batez besteko $F$ onenaren balioa : lehen % 84,09 eta orain % 84,02.

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| $n = 5$ | 84,66 | 84,29 | 81,59 | 82,95 | 83,37 | 1,40 |
| $n = 10$ | 85,66 | 84,57 | 82,49 | 83,36 | **84,02** | 1,39 |
| $n = 20$ | 85,02 | 84,84 | 82,52 | 83,71 | **84,02** | 1,15 |
| $n = 30$ | 84,92 | 84,89 | 82,32 | 83,49 | 83,90 | 1,25 |

IV.16 Taula: *Gero*. WFST metodoaren emaitzak bigarren esperimentuan ($F$ neurriaren balioak). Ikasteko *hitza-analisia* erabili da. Ikasteko bikoteak behin baino gehiago ageri daitezke baina testekoak ez (lehenengo esperimentuko test zerrendak dira).

### Bigarren esperimentuaren ondorioak

Lehenengo esperimentuan gertatu den moduan, bigarren esperimentuaren emaitzek garbi adierazten dute WFST metodoa oinarri-lerroa baino hobea dela. Nahiz eta hitzen maiztasuna kontuan hartu testean (horrek onurarik handiena oinarri-lerroari ekarri dio), bien arteko diferentzia nabarmena da oraindik eta 20 puntu ingurukoa da: lortutako $F$ neurria % 67 ingurukoa da oinarri-lerroarekin, eta % 89 edo % 87 ingurukoa WFST metodoarekin.

Ikasteko informazioari dagokionez, gauzak ez dira aldatu eta, lehenengo esperimentuan bezala, *hitza-hitza* informazioarekin lortutako emaitzak *hitza-analisia* informazioarekin lortutakoak baino hobeak izan dira, nahiz eta diferentzia, berriro, txikia izan den: 1,2 puntu inguru $F$ neurrian.

Phonetisaurusi eskatu behar zaion erantzun kopuruari dagokionez, errepikatzen da *hitza-hitza* informazioa erabiliz gero ikasteko, 5 dela erantzun kopuru onena, baina *hitza-analisia* informazioa erabiltzen bada, 10 eta 20 balioak oso parean geratu dira.

Azkenik, kontuan hartzekoa da tokenekin ikasteak, hau da, bikote bat behin baino gehiagotan erabiltzea soilik ikasteko unean, ez duela hobekuntzarik lortu ebaluazioa formekin egin denean. Beraz, ez du merezi tokenekin ikastea.

Hurrengo corpusarekin esperimentuak planteatzen ditugunean, ez dugu esperimentu hau errepikatuko: tokenak edo formak erabiltzearen ondorioak aztertu nahi genituen, eta horregatik egin dugu esperimentu hau, baina hemendik aurrera, beti egingo dugu lan formen zerrendekin.

### IV.4.3 Hitz estandarren eragina (3. esperimentua)

Esperimentu honetan probatu nahi da ea metodoak lortzen dituen emaitzak aldatzen diren ikasteko prozesuan *Aldaera* edo *Zuzena* etiketa duten hitz bikoteak ez ezik, jatorrizko testuan ageri diren hitz estandarrak erabiltzen badira. Hipotesia da informazio hori edukita hobeto ikas dezakeela tresnak aldaketek behar duten testuingurua, horrela hobeto egokituz trantsizioetako kostuak.

Hitz estandarren kopurua anotatutako bikoteen kopurua baino askoz handiagoa denez, hiru proba egin ditugu:

1. Gehitu ikasteko informaziora testuan (ikasteko zatian) ageri diren hitz estandar guztiak.

2. Gehitu ikasteko informaziora testuko hitz estandarren erdia.

3. Gehitu ikasteko informaziora testuko hitz estandarren laurdena.

Erdiaren eta laurdenaren aukeraketa egiteko, hitzen maiztasuna testuan hartu da kontuan, hau da, maiztasun handieneko erdia eta maiztasun handieneko laurdena gehitu dira, hurrenez hurren.

Datuak prestatzeko, bikote guztiak (estandarrak edo anotatuak) behin bakarrik sartu dira ikasteko zein testeko zerrendetan, eta aurreko bi esperimentuetan bezala, dena letra xehez jarri da. Ikasteko bikote kopuruak IV.17 taulan ageri dira hiru probetarako, eta testeko zerrendak lehenengo esperimentuko berberak dira.

| | A1 | A2 | A3 | A4 |
|---|---|---|---|---|
| Ikasi | | | | |
| Estandar guztiak | 2.382 | 2.449 | 2.379 | 2.363 |
| Estandar erdia | 1.555 | 1.583 | 1.534 | 1.520 |
| Estandar laurdena | 1.141 | 1.155 | 1.120 | 1.105 |
| Test | 285 | 294 | 332 | 337 |

IV.17 Taula: *Gero*. Hirugarren esperimenturako balidazio gurutzatuko atalek dituzten bikote kopuruak. Hitz estandarrak erabiltzen dira ikasteko: guztiak, erdia edo laurdena. Testeko kopuruak lehenengo esperimentukoak dira (IV.5 taula).

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| Est. guztiak | 83,93 | 87,34 | 85,90 | 84,86 | 85,51 | 1,46 |
| Est. erdia | 84,91 | 87,86 | 85,90 | 85,17 | **85,96** | 1,33 |
| Est. laurdena | 85,66 | 86,85 | 85,44 | 84,81 | 85,69 | 0,85 |
| Lehenengo esperimentuko emaitzak | | | | | | |
| Oinarri-lerroa | 54,68 | 57,48 | 53,28 | 53,02 | 54,61 | 2,05 |
| Phonetisaurus *hitza-hitza* | 85,34 | 86,69 | 86,04 | 84,94 | **85,75** | 0,77 |

IV.18 Taula: *Gero*. WFST metodoaren emaitzak hirugarren esperimentuan. Ikasteko *hitza-hitza* erabili da, hitz estandarrak barne. Eskatutako erantzun kopurua 5 izan da beti.

### Hitza-hitza

Aurreko esperimentuetan ikusi dugunez, *hitza-hitza* informazioarekin ikasiz gero emaitza onenak 5 erantzun eskatuta lortzen dira, eta horiek bakarrik eman ditugu IV.18 taulan.

Konparazioa errazteko, lehen esperimentuko emaitza batzuk ere ekarri ditugu taulara: oinarri-lerroaren emaitzak (IV.6 taulakoak) eta *hitza-hitza* informazioarekin $n$=5 probaren emaitzak (IV.8 taula). Emaitzei begira, ez da garbi geratzen onuragarria den hitz estandarrak gehitzea ikasteko prozesuan. Horiek erabili gabe batez besteko balioa $F$ neurriarentzat % 85,75 zen (1. esperimentua), eta oraingo hiru probetatik bitan pixka bat jaitsi da balio hori. Kasu bakar batean igo da balioa (hitz estandarren erdia gehitu denean) eta igoera txikia izan da: % 85,96 balioraino.

### Hitza-analisia

Ikasteko informazio gisa *hitza-analisia* erabili dugunez, 20 erantzun eskatu zaizkio Phonetisaurusi eta lortutako emaitzak IV.19 taulan bildu dira.

Aurrekoan bezala, lehen esperimentuko emaitzak ekarri ditugu taulara konparazioak errazteko: oinarri-lerroaren emaitzak berriro (IV.6 taulakoak) eta *hitza-analisia* informazioaren n=20 probaren emaitzak (IV.9 taula). Oraingo honetan, hiru kasuetan hobetu dira lehenengo esperimentuko emaitzak, $F$ neurriaren balioa % 84,09 baino altuagoa izan baita hiruretan. Balio onena, % 84,93, hitz estandar guztiak gehituta lortu da.

| | A1 | A2 | A3 | A4 | bb. | d.e. |
|---|---|---|---|---|---|---|
| Est. guztiak | 84,75 | 87,10 | 83,89 | 83,99 | **84,93** | 1,49 |
| Est. erdia | 83,55 | 86,54 | 83,15 | 85,31 | 84,64 | 1,58 |
| Est. laurdena | 85,34 | 84,59 | 83,74 | 83,81 | 84,37 | 0,75 |
| Lehenengo esperimentuko emaitzak | | | | | | |
| Oinarri-lerroa | 54,68 | 57,48 | 53,28 | 53,02 | 54,61 | 2,05 |
| Phonetisaurus *hitza-analisia* | 85,12 | 84,57 | 83,55 | 83,12 | **84,09** | 0,92 |

IV.19 Taula: *Gero*. WFST metodoaren emaitzak hirugarren esperimentuan. Ikasteko *hitza-analisia* erabili da, hitz estandarrak barne. Eskatutako erantzun kopurua 20 izan da beti.

#### Hirugarren esperimentuaren ondorioak

Ikasteko informazioari dagokionez, hirugarren esperimentuak lehenengo esperimentuaren ondorioa berresten du: *hitza-hitza* erabiltzeak *hitza-analisia* baino emaitza hobeak lortu ditu. Dena den, bi aukeren arteko diferentzia txikiagoa izan da orain: puntu bat ingurukoa, bi aukeretako $F$ neurriaren emaitza onenak kontuan hartuta (% 85,96 eta % 84,93).

Bestalde, hitz estandarrak kontuan izatea ikasteko prozesuan lagungarria suerta daiteke, baina ezin da esan beti hala denik. Izan ere, *hitza-hitza* kasuan ez da oso portaera erregularra lortu, eta hitz estandarren erdia gehituz gero emaitza bi hamarren hobetu da, baina hitz guztiak edo laurdena gehituta, emaitza bi hamarren eta hamarren bat jaitsi da, hurrenez hurren. *Hitza-analisia* kasuan, berriz, estandarrak gehitzeak beti lortu du emaitza hobea, emaitza onena lortuz estandar guztiak gehituta.

### IV.4.4 Azken ebaluazioa: test-corpusa

Doikuntzako hiru esperimentuak egin ondoren, azken ebaluazioa egin da: *Gero* corpusaren ikasteko zati osoa erabili dugu ikasketa-prozesuan, eta corpusaren testeko zatiarekin egin dugu gero ebaluazioa. Dena den, ebaluazioa egin baino lehen erabaki beharra dago zenbat informazio erabili behar den ikasteko (hitz anotatuak soilik edo anotatuak eta estandarrak), zein informaziorekin egin behar den ikasketa (*hitza-hitza* edo *hitza-analisia*) eta zenbat erantzun eskatu behar diren deskodeketa-urratsean. Izan ere, erabaki horien arabera erauzi eta prestatu behar dira azken ebaluaziorako datuak.

Doikuntzako hiru esperimentuen emaitzak analizatuta, honako hauek izan dira hartutako erabakiak:

- Bigarren esperimentuak argi utzi du hitzen maiztasuna kontuan hartzea testa egiteko unean emaitzen alde dagoela, zenbaki altuagoak lortzen baitira modu horretan (3-4 puntu altuagoak $F$ neurrian). Dena den, azken ebaluazioa egiteko ez gara eszenatoki horretan jarri, eta testa egin dugu bikote bakoitza behin bakarrik kontuan hartuta.

- Ikasteko informazioari dagokionez, bi ebaluazio egin dira: horietako batean *hitza-hitza* informazioa erabili da ikasteko eta 5 erantzun eskatu dira deskodeketan; bestean *hitza-analisia* informazioa erabili da ikasteko eta 20 erantzun eskatu dira deskodeketan.

- Azkenik, ez denez guztiz garbi geratu hitz estandarrak sartzea ikasteko prozesuan onuragarria denik, azken ebaluazioa bi eratara egitea erabaki dugu: (1) Hitz estandarrak kontuan hartu gabe ikasteko prozesuan; (2) Hitz estandarrak kontuan hartuta ikasteko prozesuan, baina kopuru ezberdin erabiliz ditugun bi kasuetarako: *hitza-hitza* kasuan hitz estandarren erdia gehitu dira soilik, eta *hitza-analisia* kasuan hitz estandar guztiak gehitu dira.

Aurreko erabakiak hartu ondoren, datu zerrendak prestatu ditugu ebaluazioa egin ahal izateko. Datuak prestatzeko urratsak ez dira berriak (doikuntzako esperimentuetan aipatu dira) baina merezi du horiek laburbiltzea:

- Ikasteko erabili behar diren datuak ikasteko corpusetik (obraren % 10) erauzi dira. Hartu berri ditugun erabakien arabera, hiru zerrenda ezberdin prestatu dira ikasteko:

    1. Eskuz anotatutako formak dituen zerrenda, hau da, *Aldaera* zein *Zuzena* etiketekin anotatutako bikoteak dituen zerrenda (bikote bakoitza behin bakarrik).
    2. Eskuz anotatutako formak (aurreko zerrendakoak) eta ikasteko corpusean dauden hitz estandarren erdia jasotzen duen zerrenda (maizen ageri direnak). Zerrenda berri hori *hitza-hitza* informazioarekin ikasteko erabili dugu.
    3. Eskuz anotatutako formak (lehenengo zerrendakoak) eta ikasteko corpusean dauden hitz estandar guztiak jasotzen dituen zerrenda. Azken zerrenda hori *hitza-analisia* informazioarekin ikasteko erabili dugu.

- Testeko zerrenda bakarra da eta testeko corpusetik (obraren % 5) *Aldaera* zein *Zuzena* etiketekin anotatutako bikoteek osatzen dute zerrenda hori (bikote bakoitza behin).

Zerrenda horien guztien tamainak IV.20 taulan ageri dira. Logikoa denez, ikasteko zerrenda motzena anotatutako bikoteak soilik dituena da (taulan *anotatuak* esan diegu bikote horiei), eta beste biek zenbait hitz estandar dituzte aurreko bikoteez gain.

| **Azken ebaluazioa** | |
|---|---|
| Ikasi | |
| anotatuak | 956 |
| anotatuak + estand. erdia | 1.953 |
| anotatuak + estand. guztiak | 2.949 |
| Test | 566 |

IV.20 Taula: Azken ebaluazioko kopuruak ikasteko zein testerako.

WFST metodoaren onura zenbatekoa den ikusteko, lehendabizi oinarri-lerroarekin egin da ebaluazioa. Sistema horretan ez du inongo eraginik hitz estandarrak izatea ala ez izatea, beraz, nahikoa da ebaluazioa behin egitea.

WFST metodoarekin berriz, lau ebaluazio egin dira, ikasteko informazioaren formatuaren arabera (*hitza-hitza* edota *hitza-analisia*) eta ikasteko hitz zerrendaren arabera (estandarrak kontuan hartu gabe edo kontuan hartuta). Ebaluazio horien guztien emaitzak IV.21 taulan ageri dira.

| Metodoa | $\boldsymbol{P}$ | $\boldsymbol{R}$ | $\boldsymbol{F}_1$ |
|---|---|---|---|
| Oinarri-lerroa | 94,87 | 39,22 | 55,50 |
| WFST *hitza-hitza* | | | |
| anotatuak | 91,53 | 78,27 | 84,38 |
| anotatuak + estand. erdia | 91,84 | 79,51 | **85,23** |
| WFST *hitza-analisia* | | | |
| anotatuak | 91,08 | 77,56 | 83,78 |
| anotatuak + estand. guztiak | 90,91 | 77,74 | 83,81 |

IV.21 Taula: *Gero*. Azken ebaluazioko emaitzak metodo bakoitzarekin.

### IV.4.5 Ondorioak

Azken ebaluazioan lortu diren emaitzak (IV.21 taula) nahiko koherenteak dira doikuntzako esperimentuetan ikusi dugunarekin:

- Oinarri-lerroarekin konparatuta, WFST metodoaren edozein aukera askoz hobea da. Egia da doitasun handiena oinarri-lerroak lortzen duela, baina estaldura ia bikoizten da WFST metodoarekin, eta hortik dator $F$ neurrian lortzen den aldea.

- Ikasteko informazioari dagokionez, azken ebaluazioan berriro gertatu da emaitza hobea lortu dela beti *hitza-hitza* informazioarekin ikastean. Informazio hori erabiliz, emaitza onena lortu da hitz estandarren erdia gehituta ikasteko prozesuan. Ikasketa *hitza-analisia* informazioarekin egin denean emaitza pixka bat okerragoa izan da, diferentzia 0,6 eta 1,4 puntu ingurukoa izanik aurreko informazioarekiko. Beraz, badirudi *hitza-analisia* informazioak ez duela onurarik ekartzen, nahiz eta ezin den esan askoz okerragoa denik.

Azken ebaluazioan egin diren normalizazio okerrak aztertu ditugu, eta ikusi dugu izen propioek tratamendu berezi bat beharko luketela normalizatuak izateko, eraiki dugun sistemak ez baitu ongi ebatziko inoiz izen propio baten normalizazioa. Horren arrazoia da Phonetisaurusi ikasteko ematen zaizkion bikote guztiak letra xehez idatzi direla, izen propioak barne. Modu horretan ikasita, deskodeketan lortzen diren aukerazko erantzunak ere beti daude letra xehez idatziak, eta ez dute pasako hizkuntza-ereduaren iragazkia izen propioei dagozkien normalizazioak badira. Esaterako, estandarraren automatak ez ditu onartuko *aristotelesek* edo *davidek* moduko normalizazioak, ez direlako estandar zuzenak (ez dira letra larriz hasten). C eranskinean azaltzen da proposamen sinple bat arazo horri aurre egiteko, bi izanik proposamen horren helburu nagusiak: (1) neurtu izen propioen arazoa garrantzitsua den ikuspegi kuantitatibotik, eta (2) aztertu tratamendu sinple horrek alboko efektu negatiborik duen.

Beste analisi interesgarria da *Gero* corpus historikoan lortutako emaitzak (IV.21 taula) konparatzea III. kapituluan corpus dialektalarekin lortutako emaitzekin (III.11 taula). Analisi hori eginez gero, bi ondorio nagusi ateratzen ditugu:

- WFST metodoak lortzen dituen emaitzak oso onak dira oinarri-lerroko sistemekin konparatuta bi corpusetan: $F$ neurriaren diferentzia sistema horien artean 20 puntu ingurukoa zen corpus dialektalean (% 60

oinarri-lerroekin eta % 80 WFST onenarekin), eta *Gero* corpusean, diferentzia hori oraindik handiagoa izan da, 30 puntu ingurura iritsi baita (% 55 oinarri-lerroarekin eta % 85 WFST onenarekin).

- Corpus dialektalean *hitza-analisia* informazioaren bitartez lortutako emaitzak hobeak ziren *hitza-hitza* informazioarekin lortutakoak baino, nahiz eta diferentzia oso handia ez izan (1,7 puntu inguru $F$ neurrian). *Gero* corpusean ez da horrela izan eta *hitza-hitza* informazioak lortu ditu beti emaitza hobeak, diferentzia antzekoa izanik (1,4 puntu).

## IV.5 *Peru Abarka* corpusa: esperimentuak eta emaitzak

Atal honetan corpus berri batekin egin ditugun esperimentuak azalduko ditugu. Esperimentu horien bitartez aztertu nahi da ea *Gero* corpusarekin lortu ditugun emaitzak berresten diren euskarazko beste corpus batean. Beraz, aurretik egin ditugun esperimentuak errepikatu ditugu euskarazko beste klasiko batekin, Mogelen *Peru Abarka* obrarekin, hain zuzen ere.

*Peru Abarka* corpusaren prestaketa-lana bigarren kapituluko II.3.5 atalean deskribatu da, eta *Gero* corpusarekin egin dugun moduan, IV.22 taulara ekarri ditugu berriro ditugu prestatutako corpusaren oinarrizko ezaugarriak.

Gogoratu behar dugu ikasteko zatia 5 fitxategitan banatuta dagoela (testeko zatia beste 2tan banatuta dago). Beraz, *Gero* corpusarekin egin dugun moduan, ikasteko corpusaren 5 zati horiek baliatu ditugu doikuntzako esperimentuak planteatzeko 5 ataleko balidazio gurutzatua eginez.

Dena den, ez ditugu doikuntzako hiru esperimentuak errepikatuko, bi baizik. *Gero* obrarekin egin den bigarren esperimentua (tokenekin lan egiten duena) ez dugu errepikatuko, eta beti egingo dugu lan kontuan izanik bikote bakoitza behin, bai ikasteko eta bai ebaluatzeko unean.

| | **Analizatutako** | | **OOV** | |
|---|---|---|---|---|
| | tokenak | formak | tokenak | formak |
| Ikasi | 1.987 | 1.199 | 1.404 | 927 |
| Test | 1.073 | 725 | 733 | 539 |

IV.22 Taula: *Peru Abarka* corpusa. Ikasteko zein testeko zatien ezaugarriak.

### IV.5.1 Oinarrizko doikuntza (1. esperimentua)

Lehendabizi ikasteko datuak prestatu behar dira, eta hori egiteko prozesua ez da aldatzen. Lehendabizi ikasteko 5 fitxategietatik anotatutako *aldaerak* zein *zuzenak* erauzi dira, hitz anitzekoak kendu dira[16] eta bikote bakoitza behin bakarrik utzi da fitxategi bakoitzean. Gero, bikote guztiak xehez idatzi dira eta balidazio gurutzatuko atalak prestatu dira: atal bakoitzean 4 fitxategi ikasteko eta 5.a testa egiteko (ikasteko informazioa 4 fitxategi elkartuz lortzen denez, elkarketa egin ondoren berriro ziurtatu behar da bikote bakoitza behin bakarrik ageri dela).

IV.23 taulan atal bakoitzaren bi zerrenden kopuruak ageri dira eta konparagarriak dira *Gero* corpusekoekin (IV.5 taula).

| | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| Ikasi | 771 | 737 | 767 | 704 | 768 |
| Test | 204 | 227 | 203 | 274 | 199 |

IV.23 Taula: *Peru Abarka*. Lehenengo esperimenturako kopuruak balidazio gurutzatuko atal bakoitzean.

**Oinarri-lerroaren** funtzionamendua ezaguna da eta ez dugu berriro gogoratuko. Sistema horrek lortzen dituen emaitzak IV.24 taulan ageri dira (batezbestekoak soilik eman dira taula sinplifikatzeko): doitasuna oso altua da, % 98 ingurukoa, baina estaldura, berriz, oso txikia. Ondorioz, lortutako $F$ neurriaren balioa baxua da: % 46,5 batez beste. *Gero* corpusean lortutako emaitzekin konparatuta (ikus IV.6 taula), $F$ neurriaren balioa 8 puntu inguru jaitsi da corpus berrian, eta hori da estaldura nabarmen jaitsi delako. Jaitsiera hori zergatik gerta daitekeen arrazoiaren bila, ataletako zerrenda kopuruak aztertu ditugu, baina hortik ezin da justifikatu jaitsiera hori, oro har txikiagoak baitira test zerrendak corpus berrian. Beraz, arrazoia aldaera berrien ezaugarrietan dago: dirudienez, *Peru Abarka* corpusa aberatsagoa da bertan ageri diren aldaera linguistikoei dagokienez.

[16] Gutxi dira hitz anitzekoak: *eurac gana* → *eurengana*; *eguitia gaiti* → *egiteagatik*.

| | ***P*** | ***R*** | $\boldsymbol{F}_1$ |
|---|---|---|---|
| bb. | 98,27 | 30,48 | **46,43** |
| d.e. | 1,92 | 3,63 | 4,27 |

IV.24 Taula: *Peru Abarka.*Oinarri-lerroaren emaitzak lehenengo esperimentuan (batezbestekoak).

**WFST** metodoarekin bi esperimentu egin dira, *Gero* obran egin den moduan, haien arteko ezberdintasun nagusia izanik ikasteko ematen den informazioa: *hitza-hitza* edo *hitza-analisia.*

Portaera errepikatzen den ikusteko, berriro probatu da eskatutako erantzun kopurua aldatzen ($n$ parametroa tauletan), eta emaitzak errepikatu direnez, ez ditugu kasu guztien emaitzak ekarri tauletara horiek laburragoak izan daitezen. Hala, IV.25 taulan bildu ditugu interesatzen zaizkigun bi emaitzak: *hitza-hitza* informazioarekin 5 erantzun eskatuta lortutako emaitza, eta *hitza-analisia* informazioarekin 20 erantzun eskatuta lortutako emaitza. Emaitza horiek *Gero* corpusean lortutakoekin konparatzeko, IV.8 eta IV.9 taulak hartu behar dira kontuan, eta hala eginez gero ikusten da *Peru Abarka* corpusean lortutako emaitzak 6–7 puntu inguru txikiagoak direla, hau da, oinarri-lerroan ikusi dena errepikatzen da.

| | A1 | A2 | A3 | A4 | A5 | bb. | d.e. |
|---|---|---|---|---|---|---|---|
| $n = 5$ | 83,60 | 81,04 | 75,14 | 81,05 | 78,33 | **79,83** | 3,22 |
| $n = 20$ | 82,94 | 76,19 | 72,97 | 78,80 | 76,24 | **77,43** | 3,71 |

IV.25 Taula: *Peru Abarka.* WFST metodoaren emaitzak lehenengo esperimentuan ($F$ neurriaren balioak.) Lehen errenkadako emaitzak lortu dira ikasteko *hitza-hitza* erabiliz eta 5 erantzun eskatuta. Bigarren errenkadako emaitzak lortu dira *hitza-analisia* erabiliz eta 20 erantzun eskatuta.

Lehenengo esperimentuaren emaitzek aurreko corpusarekin ateratako ondorioak berresten dituzte: (1) metodoei dagokienez, WFST metodoa oinarri-lerroa baino askoz hobea da ($F$ neurria 30 puntu baino gehiago hobetzen da, batez beste); (2) ikasteko informazioari dagokionez, *hitza-hitza* informazioak emaitza onena lortzen du; (3) eskatu beharreko erantzun kopuruari dagokionez, *hitza-hitza* kasuan 5 erantzun eskatzea da aukera onena eta *hitza-analisia* kasuan, 20 eta 30 kopuruek ia emaitza bera lortu dutenez

(taulan 20 erantzunekoa soilik ageri bada ere) ikusteke dago zer gertatzen den hurrengo esperimentuetan.

Lehenengo esperimentuak bi corpusetan lortu dituen emaitzak konparatzen badira, lehen ondorioa oso nabarmena da eta jadanik aipatu dugu: *Peru Abarka* corpusean lortu den $F$ balioa *Gero* corpusean lortutakoa baino txikiagoa da beti, bi metodoekin (6–8 puntuko aldea). Kontuan izan behar da bi corpusen artean ezberdintasun nabarmenak daudela: euskalki ezberdina erabiltzen da bietan eta OOV proportzioa oso ezberdina da bi corpusetan (bi obretako lehen analisia bigarren kapituluko II.3.3 atalean egin da errepasatu nahi izanez gero). Dirudienez, beraz, *Peru Abarka* corpusean aberastasun edo bariazio gehiago dago aldaerei dagokionez, eta normalizazioa egitea zailagoa gertatzen da. Dena den, azterketa sakonagoa egin beharko litzateke hori horrela dela egiaztatzeko.

$F$ neurriaren batezbestekoak konparatzeaz gain, beste konparazio interesgarria da neurri horren desbideratzearena bi corpusetan. *Peru Abarka* corpusean desbideratze hori 3 eta 4 puntu artean dago WFST metodoan (IV.25 taulako azken zutabea), eta 4 puntutik gorakoa da oinarri-lerroan (IV.24 taula). *Gero* corpusean, berriz, puntu baten ingurukoa izan da WFST metodoan (IV.8 eta IV.9 taulak), eta 2 puntukoa oinarri-lerroan (IV.6). Horren arrazoia, hein batean, atalek dituzten zerrenden kopuruetan egon daiteke, corpus berrian ditugun kopuruak ez baitira aurreko corpuseko bezain erregularrak, batez ere, testeko zerrendetan (konparatu IV.23 eta IV.5 taulak). Baina logika horren arabera, pentsatuko genuke test zerrenda motzenak dituzten atalek lortu behar dituztela emaitza onenak, eta ez da horrela: A1, A3 eta A5 atalak dira test zerrenda motzenak dituztenak (ikus IV.23 taula) eta A3 atala da emaitza txarrenak lortzen dituena WFST metodoa aplikatzen denean (ikus IV.25 taula). Beraz, desbideratze estandarrean dagoen diferentzia nabarmen hori, berriro ere, corpusen ezaugarriei dagokiela uste dugu.

### IV.5.2 Hitz estandarren eragina (3. esperimentua)

IV.5 atalaren hasieran aipatu dugunez, *Peru Abarka* corpusean ez dugu doikuntzako bigarren esperimentua errepikatu baina bai hirugarrena, ikasteko unean kontuan hartzen dituena corpusean ageri diren hitz estandarrak.

Corpus honetan hitz estandarren kopurua *Gero* corpusekoa baino askoz txikiagoa denez, esperimentu bakarra egin dugu, eta ikasteko zatian dauden hitz estandar guztiak erabili ditugu ikasteko informazioa prestatzeko (*Gero* corpusarekin hiru esperimentu egin dira IV.4.3 atalean). Testerako, aurreko esperimentuko test zerrenda berberak erabili dira berriro. Esperimentu

| | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| Ikasi | 1.010 | 974 | 1.002 | 925 | 1.009 |
| Test | 204 | 227 | 203 | 274 | 199 |

IV.26 Taula: *Peru Abarka*. Hirugarren esperimenturako kopuruak atal bakoitzean. Hitz estandarrak erabiltzen dira ikasteko (guztiak). Testeko kopuruak lehenengo esperimentukoak dira (IV.23 taula).

honetan erabili diren kopuruak zati bakoitzean IV.26 taulan ageri dira, eta lehen esperimentukoekin alderatuta, ikasteko bikote kopuruak dira aldatzen diren bakarrak.

Hitz estandarrek ematen duten informazioak ez dituenez aldatzen oinarrilerroaren emaitzak, egin beharreko esperimentuak WFST metodoari dagozkio. Bi esperimentu egin dira ikasteko informazioa aldatuta —*hitza-hitza* eta *hitza-analisia*—, eta lehen esperimentuan bezala, portaera errepikatzen den ikusteko, berriro probatu da eskatutako erantzun kopurua aldatzen. Portaera errepikatu denez, taula sinplifikatu dugu eta emaitza onenak soilik eman ditugu IV.27 taulan.

| | A1 | A2 | A3 | A4 | A5 | bb. | d.e. |
|---|---|---|---|---|---|---|---|
| $n = 5$ | 84,21 | 81,43 | 75,68 | 81,20 | 78,12 | **80,13** | 3,29 |
| $n = 20$ | 84,29 | 77,14 | 74,87 | 78,57 | 78,02 | **78,58** | 3,49 |

IV.27 Taula: *Peru Abarka*. WFST metodoaren emaitzak hirugarren esperimentuan ($F$ neurriaren balioak). Ikasteko informazioan hitz estandarrak sartzen dira. Lehen errenkadako emaitzak lortu dira ikasteko *hitza-hitza* erabiliz eta 5 erantzun eskatuta. Bigarren errenkadako emaitzak lortu dira *hitza-analisia* erabiliz eta 20 erantzun eskatuta.

Emaitza horiek lehenengo esperimentuko emaitzekin alderatzen baditugu, hau da, IV.25 taulakoekin, garbi ikusten da *hitza-hitza* kasuan estandarrak gehitzeak ikasteko prozesuan onura txikia izan duela $F$ neurrian (3 hamarren igo da batezbestekoa); *hitza-analisia* kasuan, berriz, emaitzaren hobekuntzak puntu osoa gainditu du.

Konparazioa egiten bada *Gero* corpusaren hirugarren esperimentuan lortutako emaitzekin (IV.18 eta IV.19 taulak), *Peru Abarka* corpusean oso an-

tzeko emaitzak lortu direla ikusten da: hitz estandarrak gehitzea ikasteko prozesuan onuragarria da, batez ere, ikasteko informazioa *hitza-analisia* denean.

### IV.5.3 Azken ebaluazioa: test-corpusa

*Peru Abarka* corpusean egin ditugun doikuntzako esperimentuen ondorioak eta *Gero* corpusean ondorioztatutakoak oso antzekoak izan direnez, test corpusarekin ebaluazioa egiteko irizpide berberak erabili ditugu berriro, bai erantzun kopuruari dagokionez, bai hitz estandarren erabilerari dagokionez.

Beraz, bi zerrenda ezberdin erauzi dira ikasteko corpusetik: batean anotatutako bikoteak soilik sartu dira, eta bestean hitz estandarrak gehitu zaizkio aurreko zerrendari (estandar guztiak kasu honetan). Testeko corpusetik, berriz, test zerrenda berria erauzi da. Hiru zerrenda horien kopuruak IV.28 taulan ageri dira eta konparagarriak dira *Gero* corpuseko kopuruekin (ikus IV.20 taula). *Peru Abarka* corpusean hitz estandarren kopurua askoz txikiagoa denez, zerrenda hori nabarmen motzagoa da, baina besteak nahiko antzekoak dira.

| **Azken ebaluazioa** | |
|---|---|
| Ikasi | |
| anotatuak | 904 |
| anotatuak + estand. guztiak | 1.179 |
| Test | 504 |

IV.28 Taula: *Peru Abarka*. Azken ebaluazioko kopuruak ikasteko zein testerako.

Ebaluazioa egiteko, oinarri-lerroa aplikatu dugu lehendabizi, eta ondoren, lau ebaluazioak egin dira WFST metodoa aplikatuta: ikasteko informazioaren formatuaren arabera (*hitza-hitza* edota *hitza-analisia*), eta ikasteko hitz zerrendaren arabera (estandarrak erabili gabe edo erabilita). Ebaluazio horien guztien emaitzak IV.29 taulan jaso dira. Horiek analizatuta, bi emaitza dira bereziki arreta ematen digutenak:

1. *Hitza-analisia* informazioarekin (hitz anotatuak eta estandarrak barne) lortu dira emaitza onenak WFST metodoa aplikatuta, eta hori

| Metodoa | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| Oinarri-lerroa | 96,82 | 30,16 | 45,99 |
| WFST *hitza-hitza* | | | |
| anotatuak | 83,74 | 68,45 | 75,33 |
| anotatuak + estand. guztiak | 83,53 | 69,44 | 75,84 |
| WFST *hitza-analisia* | | | |
| anotatuak | 82,12 | 69,25 | 75,13 |
| anotatuak + estand. guztiak | 83,53 | 71,43 | **77,01** |

IV.29 Taula: *Peru Abarka*. Azken ebaluazioko emaitzak metodo bakoitzarekin.

berria da corpus historikoekin egin diren esperimentuetan. Orain arte testu historikoetan emaitza okerragoak lortu dira informazio hori erabili denean: esperimentu guztietan eta bi corpusetan.

2. Azken ebaluazioan lortutako $F$ neurriaren balioak, oro har, doikuntzako esperimentuetan lortutakoen baino nabarmen baxuagoak dira, eta hori ere, berria da. Doikuntzako esperimentuetan $F$ neurriaren balioa % 79,5 inguruan geratu da *hitza-hitza* informazioarekin eta % 77,5 – % 78,5 inguruan *hitza-analisia* informazioarekin. Azken ebaluazioan, berriz, lortutako balioa % 76,0tik behera geratu da lau ebaluazioetatik hirutan, eta laugarrenean % 77an geratu da. *Gero* corpusean ez da halako jaitsierarik gertatu ebaluazioa egin denean testeko corpusarekin, eta horregatik azterketa berezi bat egin dugu detektatzeko ea *Peru Abarka* corpuseko testean zer edo zer berezia gertatzen ari den.

   Hala, testeko corpusa bi fitxategitan banatuta dagoela baliatuta, bi zerrenda berri prestatu ditugu eta horiekin bi test egin ditugu. Test horietan metodoaren lehenengo aukera erabili dugu, hau da, *hitza-hitza* informazioa ematen da eta hitz estandarrak ez dira sartzen. Lortutako $F$ neurriaren balioak % 77,04 eta % 76,60 izan dira bi test horietan.

   Nahiko balio baxuak dira, berriro ere, eta hori ikusita, konparazioa egin dugu doikuntzako esperimentuetan atal bakoitzean lortutako balioekin (IV.25 eta IV.27 tauletako lehen errenkadan ageri dira). Azken konparazio horrek agerian utzi du lehen ere lortu direla antzeko balioak: A3 atalean lortutako $F$ neurriak lehenengo eta hirugarren esperimentuetan, orain lortutakoak baino txikiagoak izan dira, % 75,14 eta % 75,68, hurrenez hurren. Balio baxu horiek ez dira lortu testeko

zerrenda luzea delako atal horretan, ia zerrenda motzena baita (203 bikote ditu). Arrazoia ez dago kopuruetan, beste nonbaiten baizik, eta jakiteko non, azterketa sakonago bat egin beharko litzateke corpus horretan gertatzen diren aldaerak aztertuz.

Dena den, egindako zenbakien analisiak agerian utzi du bai bi test berri horietan gertatutakoa, bai azken ebaluazioko testean gertatutakoa ez dela berria izan, eta aurretik ere gertatu dela (A3 atalean). Horretaz ohartzeko, desbideratze estandarra hartu behar da kontuan, eta ez batezbestekoa soilik. Lehen esperimentuko ondorioetan esan dugu *Peru Abarka* corpuseko esperimentuetan gertatzen den desbideratzea altua dela oro har, eta funtsezkoa da ezaugarri hori kontuan izatea ebaluazioko emaitzak justifikatzeko.

Bukatu baino lehen, euskarazko bi corpus historikoetan lortu diren emaitzak berrikusita, azken ondorio gisa esan behar dugu *Peru Abarka* corpusarekin lortutako emaitzek *Gero* corpusarekin ateratako bi ondorio berresten dituztela:

- WFST metodoak oinarri-lerroa gainditzen du oso nabarmen, 30 puntu inguruko hobekuntzarekin $F$ neurrian.
- Phonetisaurusi eskatu behar zaizkion erantzunak emaitza onenak lortzeko 5 edota 20 dira, ikasteko informazioaren arabera, *hitza-hitza* zein *hitza-analisia*.

Ikasteko informazioari dagokionez, ez dira ondorio berdinak ateratzen bi corpusetan egindako ebaluaziotik:

- *Gero* corpusaren ebaluazioak adierazten du *hitza-hitza* informazioa dela aukera onena ikasteko; *Peru Abarka* corpusaren ebaluazioak, ordea, *hitza-analisia* hobesten du. Honetaz erabakitzea zaila da eta aztertu beharko litzateke ea izen propioek zarata sortu duten *hitza-analisia* erabili denean *Gero* corpusean, zenbait kasutan analisirik baitute izen horiek.
- Ikasteko prozesuan hitz estandarrak kontuan izatea onuragarria izan daiteke normalizazioaren emaitzak hobetzeko. Hala izan da bi corpusetan, diferentzia nabarmenagoa izanik *Peru Abarka* corpusean.

Azkenik, esan beharra dago corpusaren ezaugarriak erabakigarriak direla normalizazio-metodoak lor dezakeen arrakasta zenbatekoa den esateko, eta

hori garbi geratu da landu ditugun bi corpusetan. Emaitzen arabera, badirudi *Gero* obraren normalizazio-ataza *Peru Abarka* obrarena baino sinpleagoa dela, eta horregatik lortu dela arrakasta handiagoa atazan. Gogoratu behar da, beste behin, bi obretako ezaugarriak oso ezberdinak direla eta *Peru Abarka* obran OOV hitzen portzentajea oso handia dela. Badirudi korrelazio bat dagoela portzentaje horren eta atazaren zailtasunaren artean, eta, hein batean, logikoa da: zenbat eta urrutiago egon estandarretik, normalizazio-ataza zailagoa bilakatzen da.

## IV.6 Esperimentuak gaztelaniarekin eta eslovenierarekin

Kapitulu honetako sarreran esan dugunaren arabera, IV.1 atalean, gaztelania eta esloveniera hizkuntzekin bideratu ditugu esperimentuak, aukera izan baitugu hizkuntza horietako datuak eskuratzeko hainbat ikerlariren kolaborazioa baliatuz.

Esperimentu berri hauen helburua bikoitza da: alde batetik aztertu nahi dugu zer nolako emaitzak lortzen diren normalizazioan, tesi-lan honetan proposatzen den metodoa beste hizkuntza batzuetan aplikatzen bada, eta, beste aldetik, konparatu nahi ditugu proposatzen dugun metodoaren emaitzak beste ikerlariek lortu dituztenekin.

### IV.6.1 Gaztelania zaharra: esperimentuak

Bigarren kapituluko II.5.1 atalean gaztelaniazko corpusaren xehetasunak eman ditugu. Gaztelaniazko bi datu-multzo ezberdin ditugu esperimentuak egiteko, FL-EM multzoa eta IMPACT multzoa, eta bi motatako esperimentuak egin ditugu horiekin. Lehendabizi, multzo bakoitzaren barruan egin ditugu esperimentuak, eta gero, bi multzoak hartu ditugu kontuan batera esperimentu berri batean.

#### IV.6.1.1 Esperimentuak FL-EM datu-multzoarekin

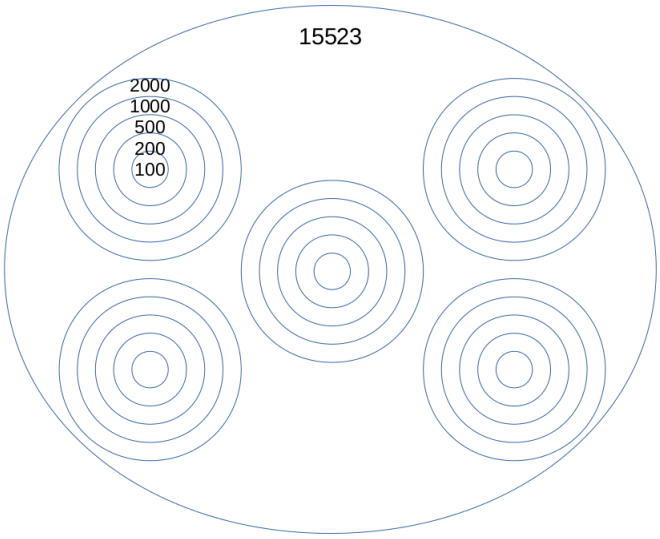
Esperimentuekin hasi baino lehen, multzoaren ezaugarriak gogoratu behar ditugu (prestaketa-prozesua bigarren kapituluan azaldu da): 31.046 bikote dituen multzoa da, gaztelaniazko forma zaharrak eta egungoak erlazionatzen dituztenak. Bikote guztiak ezberdinak dira, eta 1.248 bikotetan (% 4,02) bikotearen bi formak berdinak dira.

Euskarazko bi corpusekin konparatuta, corpus hau oso ezberdina da, batez ere, tamaina hartzen bada kontuan. Euskaraz, 1.000 eta 500 inguruko

bikotez osatutako bi zati izan ditugu ikasteko zein ebaluatzeko (anotatutako formak kontatuta), ez disjuntuak (ebaluazioko corpuseko bikote batzuk ikasteko zatian ageri dira).

FL-EM multzoak askoz bikote gehiago ditu, 31.046 bikote, eta horrek aukera ematen digu ikasketa-kurbak lortzeko, hau da, analizatzeko zenbat bikoterekin ikasi behar den, ebaluazioan kalitate minimoko emaitzak lor daitezen.

Beraz, lehenik eta behin, datu-multzoa zatitu dugu zoriz tamaina bereko bi zatitan. Horietako zati bat, 15.523 bikotez osatua, test gisa erabili dugu orain azalduko ditugun esperimentu guztietan. Beste 15.523 bikote dituen erditik, hainbat tamainatako azpimultzo ezberdin sortu dira ikasketa-kurbak lortu ahal izateko, zehazki 100, 200, 500, 1000, 2000, 5000 eta 15.523 (zati osoa) bikotez osatutako multzoak. Halaber, tamaina txikienetan, 100–2000 tartekoetan, ez da azpimultzo bakarra egin, 5 azpimultzo baizik, haien artean disjuntuak (hau da, tamaina bereko 5 multzoek ez dute bikote komunik). Multzo horiek sortzeko, lehendabizi 100 tamainakoak aukeratu dira, eta gero horiek zabaldu dira oraindik aukeratu gabe dauden bikote berriak eta ezberdinak gehituz.




IV.2 Irudia: Ikasteko zatiaren banaketa. 15.523 bikoteekin 100–2.000 tamainetako 5 azpimultzo disjuntu egin dira, guztira 10.000 bikote aukeratuz. Geratzen direnetatik, 5.000 aukeratu dira tamaina horretako esperimentua bideratzeko, eta azken esperimentuan bikote guztiak hartu dira ikasteko.

IV.2 irudiak modu grafikoan azaltzen du nola sortu ditugun tamaina txikienetako multzo horiek. Bertan ikusten denez, 100–2000 tarteko multzoekin 10.000 bikote aukeratu dira guztira, eta 5.523 geratzen dira erabili gabe. Horietatik 5.000 zoriz aukeratu dira tamaina horretako esperimentua bideratzeko, eta azken esperimenturako, berriz, bikote guztiak hartu dira, hau da, 15.523 bikoteak.

Aplikatu dugun normalizazio-metodoa WFST metodoa izan da, eta euskarazko esperimentuen emaitzak kontuan hartuta proba bakarra egin dugu: *hitza-hitza* informazioa eman zaio Phonetisaurusi ikasteko, eta deskodeketan 5 erantzun eskatu zaizkio. Euskarazko esperimentuetan ikusi dugunez, hautagaiak iragaztea beharrezkoa da gaztelaniazko egungo hitzak ez direnak alboratzeko, eta horretarako, transduktore berri bat sortu dugu gaztelaniarako Freeling tresnan oinarriturik[17]. Azkenik, sarrerako aldaera batentzat aukera bat baino gehiago geratzen den kasuetan, Phonetisaurusen arabera onena den erantzuna aukeratu da, euskarazko corpus historikoetan egin den bezala.

Lortutako emaitzak bi modutara eman ditugu: IV.30 taulan zenbakiak ageri dira, eta IV.3 irudian, berriz, parametro bakoitzaren adierazpen grafikoa ageri da. Taulari dagokionez, 100–2000 tamainetako multzoetan ez da multzo bakoitzaren emaitza eman, baizik eta 5 multzoen batezbestekoak (desbideratze estandarra parentesi artean adierazi da).

Emaitza horietan badira deigarri gertatu zaizkigun bi ezaugarri:

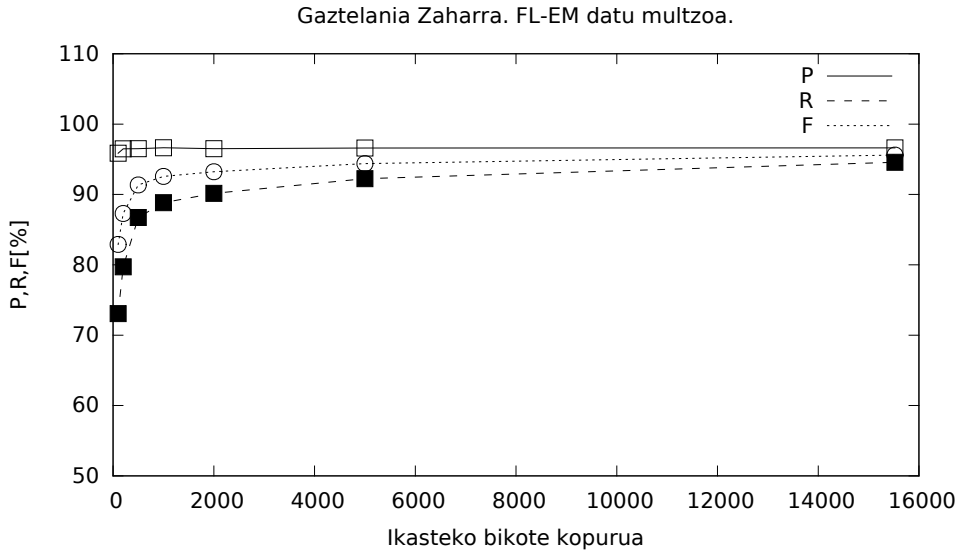
- Doitasuna ($P$) ez da ia aldatzen nahiz eta bikote kopurua handitu ikasteko. Gainera, parametro horren balioa oso altua da hasieratik, % 95,87 balioan hasten baita tamaina txikieneko multzoekin.

- Estaldura ($R$) ordea, hazi egiten da zenbat eta bikote gehiago erabili ikasteko, eta hazkuntza hori nabarmenagoa da hasierako tamainetan gerora baino. Nahiz eta estalduraren hazkuntza apalagoa izan tamaina handitzen den heinean, datuek adierazten dute estaldura ez dela asetzen eta beti doala gorantz.

[17]Tresna bera erabili da II. kapituluko II.5.1 atalean FL-EM datu-multzoko egungo hitzak sortzeko analisi morfosintaktikotik abiatuta.

| Ikasi | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 100 | 95,87 (0,92) | 73,06 (3,90) | 82,90 (2,85) |
| 200 | 96,48 (0,44) | 79,72 (1,50) | 87,30 (1,07) |
| 500 | 96,50 (0,14) | 86,72 (0,39) | 91,35 (0,22) |
| 1.000 | 96,64 (0,13) | 88,83 (0,22) | 92,57 (0,14) |
| 2.000 | 96,51 (0,13) | 90,16 (0,22) | 93,23 (0,17) |
| 5.000 | 96,61 | 92,25 | 94,38 |
| 15.523 | 96,62 | 94,58 | 95,59 |

IV.30 Taula: FL-EM multzoarekin egin diren esperimentuen emaitzak. Lehen zutabeak adierazten du zenbat bikote erabili diren ikasteko. 100–2000 tarteko errenkadetan ematen diren balioak 5 multzo ezberdinetan lortutakoen batezbestekoak dira (parentesi artean desbideratze estandarra). 5.000 eta 15.523 tamainetako esperimentuetan multzo bakarra dagoenez, balioak ez dira batezbestekoak.




IV.3 Irudia: IV.30 taulako emaitzak grafikoki adierazita.

Hizkuntza berri bati aplikatu diogu gure normalizazio-metodoa, gaztelaniari, eta euskararekin zein gaztelaniarekin lortutako emaitzak konparatzea pentsa dezakegu. Zaila da, ordea, gaztelaniazko FL-EM multzoan lortutako emaitzak zuzenean alderatzea euskarazkoekin, batez ere, oso ezberdinak direlako esperimentuetako kopuruak: euskaraz, 1.000 inguru bikoterekin ikasi da (hitz estandarrak erabili ez direnean) eta 500 bikoterekin egin da ebaluazioa. Gaztelaniaz egin diren esperimentu guztien artean, bost egin dira 1000 bikote ezberdin erabiliz ikasteko, baina kontuan hartu behar da bost esperimentu horietan ebaluazioa 15.523 bikoterekin egin dela, eta ez 500ekin. Dena den, gaztelaniaz lortutako $F$ neurria kasu horretan % 92,5 ingurukoa da, eta euskaraz lortu dugun altuena % 85,5 ingurukoa izan da (*Gero* corpusarekin). Ikuspuntu horretatik, FL-EM multzoan lortutako emaitzak izugarri onak direla esango genuke.

Beste alderaketa interesgarria da gure sistemaren emaitzak konparatzea Porta *et al.* (2013) lanean ematen direnekin. Zenbakiak zuzenean alderatu baino lehen, ordea, kontuan hartu behar dira bi lanen arteko ezberdintasunak: (1) Porta *et al.* (2013) lanean planteatzen den ataza ez da zehatzmehatz guk planteatutakoa, lan horretan forma zaharrak analizatzeko sistema proposatzen baita, eta guk forma zaharrak normalizatzeko sistema bat eraiki dugu; (2) lan horretan egiten den ebaluazioa datu-multzo osoarekin egiten da (eskuz idatzitako erregeletan oinarritutako sistema da, ez du datuetatik ikasten), eta gure ebaluazioa, berriz, datu-multzoaren erdiarekin egiten da.

Zehaztasun horiek egin ondoren, egin dezagun zenbakien konparazioa. Porta *et al.* (2013) lanean ematen diren emaitzak FL-EM datu-multzoarentzat dira: $P = 69{,}75$, $R = 89{,}02$ eta $F = 78{,}22$. Esperimentu zehatz batekin konparatzearren, ikasketarako 1.000 bikote erabiliz lortutako emaitzak hartuko ditugu kontuan: $P = 96{,}64$, $R = 88{,}83$ eta $F = 92{,}57$. Ikusten denez, estaldura oso antzekoa da bietan, baina doitasunean 25 puntu baino gehiagoko aldea dago, eta horren eraginez $F$ neurriko diferentzia 14 puntukoa da.

Konparazio horren ondorio gisa esango genuke: (1) proposatzen dugun normalizazio-metodoa artearen egoeraren barruan aurkitzen direnekin konparagarria dela, eta (2) proposatzen dugun metodologia egokia dela gaztelaniarako ere, hau da, 1.000 forma inguru anotatuz gero, eskuz, nahiko emaitza onak lor daitezkeela normalizazioan WFST teknologiaren bitartez.

#### IV.6.1.2 Esperimentuak IMPACT datu-multzoarekin

Esperimentuekin hasi baino lehen, IMPACT multzoaren ezaugarriak gogoratu behar ditugu: 24.009 bikote dituen multzoa da, gaztelaniazko forma zaharrak eta egungoak erlazionatzen dituztenak, bikote guztiak ezberdinak dira, eta 15.337 bikotetan (% 63,38) bi formak berdinak dira.

FL-EM multzoarekin konparatuta, multzo hau txikiagoa da (nahiz eta euskarazko corpusak baino askoz handiagoa den), baina diferentzia nabarmenena beste ezaugarri batean dago: IMPACT multzoko ia % 64 bikotetan gertatzen da aldaera eta egungo formak berdinak direla. FL-EM multzoan horien portzentajea % 4 da, eta euskarazko corpusetan, ezaugarri hori duten bikoteak hitz estandarrei dagozkien bikoteak dira (ikasteko urratsean erabili dira halako bikoteak zenbait esperimentutan, baina inoiz ez ebaluatzeko).

Beraz, IMPACT multzoak ezaugarri ezberdinak dituenez, eta garai ezberdineko gaztelaniari dagokionez, interesgarria izan daiteke aztertzea nolako emaitzak lortzen dituen gure normalizazio-metodoak multzo berri horretan aplikatzen bada.

FL-EM datu-multzoarekin egin dugun bezala, IMPACT datu-multzoa bi zatitan erdibitu dugu zoriz: 12.005 bikote dituen erdia test gisa erabili dugu esperimentu guztietan, eta beste 12.004 bikote dituen erdia hainbat tamainatako azpimultzotan banatu dugu.

Testeko multzotik lau sarrera kendu behar izan ditugu arazoak ematen zizkigutelako deskodeketa-urratsean[18]. Sarrera bereziak dira, dena den, eta haien ezaugarria da forma zaharra azentu-marka duen karaktere bakar batez osatua dagoela: `á` → `a` , `à` → `a` , `ò` → `o` , `ý` → `ahi`. Beraz, testa egiteko 12.001 bikote erabili ditugu esperimentuetan.

FL-EM multzoarekin lortutako emaitzen arabera, hipotesia da emaitza onak lortzeko nahikoa dela 500 edo 1.000 bikote erabiltzea ikasteko, eta hortaz, kopuru horietako 5 multzo sortu ditugu zoriz, lehen bezala, haien artean disjuntuak.

Esperimentuen emaitzak IV.31 taulan ageri dira eta IV.30 taulako tamaina bera dutenekin konparatu behar dira. Ikusten denez, hiru parametroak jaitsi dira, baina ez asko: doitasuna 1,5-2,0 puntu inguru jaitsi da, estaldura puntu 1 inguru, eta $F$ neurria 1,3 puntu inguru. Ezaugarri ezberdinak dituzten datu-multzoak izanik, ezin da esan alde handia dagoenik bien emaitzen artean.

[18] Letra bakarreko sarrerak dira, eta Phonetisaurus geratu egin da deskodeketa-urratsean aurrera egin gabe horietako sarrera bat eman zaionean. Horren aurrean, letra bakarreko sarrerak kendu ditugu esperimentuak egiteko.

| Ikasi | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 500 | 94,75 (0,69) | 85,77 (1,00) | 90,04 (0,82) |
| 1000 | 94,98 (0,50) | 87,82 (0,80) | 91,26 (0,65) |

IV.31 Taula: IMPACT datu-multzoarekin egin diren esperimentuen emaitzak. Lehen zutabeak adierazten du zenbat bikote erabili diren ikasteko. Tamaina bakoitzeko 5 multzo erabili direnez, ematen diren balioak batezbestekoak dira (parentesi artean desbideratze estandarra).

Konparazio gehiagorik ez dugu egin, IMPACT datu-multzoa ez baita erabiltzen aipatutako Porta *et al.* (2013) lanean.

#### IV.6.1.3 Esperimentu berria: datu-multzoak gurutzatu

Gaztelaniazko FL-EM eta IMPACT datu-multzoak ezberdinak direnez, aurreko esperimentuen emaitzak ikusita, berehala planteatzen den galdera da ea zer gertatuko den biak batera erabiltzen badira, esaterako, datu-multzo batekin ikasten bada eta bestearekin egiten bada ebaluazioa. Esperimentu berri horren bitartez jakin ahal izango dugu zenbateko menpekotasuna duen gure normalizazio-sistemak datu-multzoarekiko. Logikoa da pentsatzea menpekotasuna dagoela: sistemak ikasi egiten du datuetatik, beraz, beharrezkoa da ikasteko zein ebaluatzeko zatietan antzeko kasuistika izatea emaitzak kalitate minimo bat izan dezaten. Ditugun bi datu-multzo horietan ageri den kasuistika, oso ezberdina da? FL-EM datu-multzoan ageri diren aldaerak testu zaharragoetatik ateratakoak dira. Mantentzen da kasuistika aldaera berriagoetan? Esperimentu berriak aukera emango digu galdera horiei erantzuna emateko.

Hala, egin dugun azken esperimentua hauxe izan da: FL-EM multzoa erabili dugu ikasteko (500 bikotetik aurrerako multzoak) eta IMPACT multzoarekin egin dugu ebaluazioa (12.001 bikote ebaluatzeko). Azkeneko kasu bat gehitu dugu, non FL-EM datu-multzo osoa erabili dugun ikasteko. Emaitzak IV.32 taulan jaso ditugu. Konparatzen baditugu horiek eta IV.31 taulakoak, argi dago emaitza berriak baxuagoak direla. Doitasunari dagokionez, ez da inoiz iristen aurreko % 95 inguruko baliora; % 92 ingurura iristen da kasurik onenean (FL-EM datu-multzo osoarekin ikasi denean). Estaldurari dagokionez, bi emaitzak gertuago daude: aurreko taulan % 86–87 ingurukoa izan bada, orain % 86 da lortutako onena.

| Ikasi | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 500 | 89,71 (1,03) | 74,88 (1,00) | 81,63 (1,01) |
| 1.000 | 90,35 (1,28) | 77,51 (1,94) | 83,43 (1,65) |
| 2.000 | 90,74 (1,26) | 79,81 (2,27) | 84,92 (1,82) |
| 5.000 | 89,59 | 79,54 | 84,27 |
| 15.523 | 91,96 | 85,14 | 88,42 |
| 31.046 | 91,82 | 86,00 | 88,82 |

IV.32 Taula: Datu-multzoak gurutzatuz lortutako emaitzak. Ikasketa FL-EM multzoko bikoteekin egiten da (hainbat tamaina) eta ebaluazioa IMPACT multzoaren erdiarekin (12.001 bikote).

Emaitzetatik ateratzen den beste ondorio argia da ikasteko datu kopurua handitzea estalduraren mesederako dela, hori baita gehien hazten den parametroa; doitasunari ere laguntzen dio handitze horrek, baina modu apalagoan.

Azken ondorio gisa esan daiteke IMPACT datu-multzoan gertatzen den kasuistikaren zati handi bat FL-EM datu-multzoaren bitartez ikas daitekeela. Izan ere, lortutako emaitzak nahiko altuak dira: FL-EM datu-multzo osoarekin ikasiz gero, % 88,82 balioko $F$ neurria lortu da IMPACTeko test zatiarekin ebaluazioa egitean.

### IV.6.2 Esloveniera historikoa: esperimentuak

Eslovenierazko corpusaren xehetasunak II. kapituluko II.5.2 atalean azaldu ditugu. Corpus hori Scherrer eta Erjavec (2015) lanean erabili duten bera da eta gogoan izan beharreko ezaugarriak ekarri ditugu berriro hona esperimentuekin hasi baino lehen.

Corpusa bi lexikoiz osatuta dago, $L_{goo}$ eta $L_{foo}$, eta lexikoi bakoitza hiru zatitan banatuta dago: 18B, 19A eta 19B. Banaketa hori egin da kontuan izanik testuen garaia eta alfabetoa: 18 eta 19 zenbakiek XVIII. eta XIX. mendeak adierazten dituzte, hurrenez hurren; A eta B letrek, berriz, mendearen lehenengo eta bigarren erdia.

IV.33 taulan bigarren kapituluan adierazitako ezaugarriak kopiatu ditugu berriro, lexikoien tamainak gogoratzeko. Gogoratu behar den beste ezaugarria da bi lexikoiak disjuntuak direla, hots, ez dagoela bi lexikoietan ageri den bikoterik.

Bigarren kapituluan esan dugun moduan, periodo eta lexikoi bakoitzaren

| | | $L_{goo}$ **lexikoia** | |
|---|---|---|---|
| | Sarrerak | Egungo formak | Zaharra=Egungoa |
| 18B | 6.644 | 6.494 (% 97,7) | 1.181 (% 17,8) |
| 19A | 11.600 | 11.352 (% 97,9) | 2.755 (% 23,8) |
| 19B | 28.011 | 27.252 (% 97,3) | 19.635 (% 70,1) |
| | | $L_{foo}$ **lexikoia** | |
| | Sarrerak | Egungo formak | Zaharra=Egungoa |
| 18B | 4.774 | 4.641 (% 97,2) | 340 (% 7,1) |
| 19A | 5.907 | 5.801 (% 98,2) | 890 (% 15,1) |
| 19B | 10.673 | 10.470 (% 98,1) | 8.120 (% 76,1) |

IV.33 Taula: $L_{goo}$ eta $L_{foo}$ lexikoien ezaugarriak.

arabera bikote zerrenda bat erauzi da, non bikote bakoitzak jasotzen dituen hitz historikoa eta eskuz anotatutako egungoa. Beraz, 6 zerrenda ditugu esperimentuak egiteko, lexikoi bakoitzeko hiru.

Lortzen ditugun emaitzak Scherrer eta Erjavec (2015) lanean ematen dituztenekin alderatu nahi ditugunez, lan horretan egin diren esperimentuak errepikatu ditugu eta hala, erreferentziazko lanean bezala, $L_{goo}$ lexikoia ikasteko erabili dugu eta $L_{foo}$ lexikoia, berriz, ebaluazioa egiteko.

Ebaluazioa $L_{foo}$ lexikoiaren hiru zatien arabera egin behar denez, bi esperimentu planteatu dituzte Scherrer eta Erjavec ikerlariek, biak ezberdinak ikasteko erabiltzen den informazioari dagokionez. Lehenengoan ebaluazioko periodo bereko hitzekin soilik ikasten da, eta bigarrenean, ikasketa egiten da periodo guztietako informazioarekin. Emaitzak konparatu ahal izateko, guk ere bi esperimentu horiek egin ditugu.

Gure metodoaren ezaugarriei dagokienez, *hitza-hitza* informazioa eman diogu Phonetisaurusi ikasteko, eta 5 erantzun eskatu dizkiogu deskodeketa-urratsean, gaztelaniarekin egin dugun moduan. Deskodeketan lortutako erantzunak filtratzeko, eslovenierazko egungo hitz zerrenda bat erabili dugu[19].

Ebaluazioan lortu ditugun emaitzak IV.34 taulan ageri dira. Emaitza horien arabera argi geratzen da ez dagoela ia diferentziarik modu batera edo bestera ikastean, ebaluazioko zati bakoitzean lortutako balioak oso antzekoak baitira bi moduetan. Horrez gain, aldakortasun handia ikusten da

[19]Sloleks lexikoitik erauzitako zerrenda da eta xehetasunak II. kapituluko II.5.2 atalean ematen dira.

emaitzetan zatien arabera: $F$ neurria kontuan hartuta, ia 14 puntuko aldea dago 18B eta 19B zatien artean.

| | Periodoa | $\boldsymbol{P}$ | $\boldsymbol{R}$ | $\boldsymbol{F}_1$ |
|---|---|---|---|---|
| Ikasi | 18B | 79,68 | 62,36 | 69,96 |
| periodoaren | 19A | 87,02 | 73,24 | 79,54 |
| arabera | 19B | 89,63 | 78,69 | 83,81 |
| Ikasi | 18B | 78,62 | 62,78 | 69,81 |
| hitz | 19A | 85,59 | 73,98 | 79,36 |
| guztiekin | 19B | 89,51 | 78,80 | 83,81 |

IV.34 Taula: Eslovenierarekin egindako lehen esperimentuen emaitzak.

Gure sistemaren emaitza horiek Scherrer eta Erjavec (2015) lanekoekin konparatu nahi izan ditugu, baina ikusi dugu ezin direla zuzenean konparatu. Lan horretan proposatzen duten sistemak beti ematen du erantzun bat, eta hala, sistemaren kalitatea neurtzeko ematen duten parametroa *Accuracy* da:

*"We evaluate our models on modernisation accuracy, defined as the percentage of automatically modernised words that are identical with their manually annotated mform in $L_{foo}$."*

Horrekin batera, normalizazio-sistemaren erantzuna bi modutara lortu dute. Lehenengo moduan ez diote iragazkirik aplikatu sistemak ematen duen erantzunari, eta bigarrenean, berriz, bai. Dena den, Sloleks zerrendaren araberako iragazkiaren ondorioz erantzunik gabe geratzen badira, orduan euren sistemak proposatutako lehenengo erantzuna ematen dute (nahiz eta Sloleksen ez egon).

Konparazioa egiteko, beraz, gure sistema egokitu dugu bigarren moduko baldintzen arabera. Orain arte egin dugun bezala, Phonetisaurusek proposatzen dituen erantzunak iragazi ditugu lehendabizi Sloleksen arabera, eta geratzen direnen artean lehenengo posizioan dagoena aukeratu dugu sistemaren erantzun gisa. Sistemaren egokitzapena erantzunik gabe geratzen diren kasuetan egin dugu, kasu horietan Phonetisaurusek proposatutako lehenengo erantzuna utzi baita, nahiz eta Sloleksen ez egon. Horrela, gure sistemak beti ematen du erantzun bat eta bakarra ($P$, $R$ eta $F$ parametroek balio bera hartzen dute), eta kalkulatzen dugun parametroa bat dator Scherrer eta Erjavec-ek kalkulatzen dutenarekin.

Sistema egokituaren emaitzak IV.35 taulan ageri dira, eta horiekin ba-

| | Periodoa | Gure sistema | Scherrer & Erjavec |
|---|---|---|---|
| Ikasi | 18B | 67,43 | **67,8** |
| periodoaren | 19A | **79,40** | 78,4 |
| arabera | 19B | **86,81** | 84,6 |
| Ikasi | 18B | 67,91 | **68,5** |
| hitz | 19A | **79,67** | 79,3 |
| guztiekin | 19B | **86,83** | 84,2 |

IV.35 Taula: Esperimentuen emaitzak gure sistema egokitu ondoren. Egokitzapenaren eraginez sistemak beti ematen du erantzun bat sarrerako aldaerentzat. Kalkulatutako parametroa zehaztasuna da (*Accuracy*). Azken zutabean Scherrer eta Erjavec (2015) lanean lortu dituzten emaitzak kopiatu dira konparaziorako (iragazkia erabiliz lortutako emaitzak).

tera, Scherrer eta Erjavecek lortutakoak jarri ditugu konparazioa errazteko (filtroa erabiliz lortu dituzten emaitzak).

Balio horien arabera, argi geratzen da bi sistemen emaitzak nahiko parekoak direla: aldaera zaharrenekin haien sistemak emaitza hobea lortzen du (0,4 eta 0,6 puntu txikiagoa da gure sistemaren zehaztasuna), baina beste bi multzoetan gure sistemak zehaztasun hobea lortzen du beti, nahiz eta diferentzia oso handia ez izan (0,4 eta 2,6 puntu artekoa). Emaitza onak dira kontuan izanik ez dugula inolako doikuntzarik egin gure sisteman eslovenierazko corpusarekin lan egiteko.

Beste konparazio posible bat hizkuntzen artekoa da, hau da, eslovenierarekin lortu diren emaitzak konparatzea euskararekin eta gaztelaniarekin lortu direnekin. Horretarako, lehenengo esperimentuaren emaitzak konparatu behar dira (IV.34 taulakoak), euskaraz eta gaztelaniaz egindako esperimentuetan ez baita beti erantzun bat ematen eta horregatik neurtutako hiru parametroak, $P$, $R$ eta $F_1$, ezberdinak dira.

Euskarazko ebaluazioan lortutako $F_1$ neurririk onena % 85 ingurukoa izan da *Gero* corpusaren kasuan eta % 77 ingurukoa *Peru Abarka* corpusean; gaztelaniazko esperimentuetan, berriz, $F_1$ neurriak % 90 gainditu du erabili diren bi corpusetan, FL-EM eta IMPACT corpusetan. Eslovenierarekin egin den lehenengo esperimentuaren emaitzek balio baxuagoak lortu dituzte eta alde handia ikusten da corpusaren hiru zatietan lortutako emai-

tzetan: % 70 ingurukoa da kasurik okerrenean, eta % 84 ingurukoa kasurik onenean. Aldakortasun horrek gogorarazten du euskarazko bi corpusetan ikusi duguna, eta berresten du lortutako emaitzak ez dagozkiola metodoari soilik: corpusen ezaugarriak ere garrantzitsuak dira.

Hizkuntzen arteko konparazioak ez dira sinpleak eta ezaugarri gehiago hartu behar dira kontuan, hizkuntza bakoitzak eta datu-multzo bakoitzak, bere ezaugarri propioak baititu. Dena den, esan dezakegu proposatzen dugun sistemak emaitza onak lortzen dituela hizkuntza eta periodo ezberdinetarako doikuntza berezia egin behar gabe.

# V. KAPITULUA

## Morfologiaren ekarpena

### V.1 Sarrera

Lehen kapituluko I.2 atalean aipatu dugu testu ez-estandarretan aurkitzen ditugun aldaera asko maila fonologikoan gertatzen direla, baina maila lexiko-morfologikoan ere gertatzen direla hainbat aldaketa. Egia da bi maila horien arteko muga lausoa dela zenbait adibidetan, eta, batzuetan, zaila da esatea aldaera eta estandarraren artean gertatzen den aldaketa fonologikoa edo morfologikoa den. Esaterako, *Gero* corpuseko ikasteko zatian ikusten dugu *disposizioneari* aldaerari *disposizioari* hitz estandarra esleitu diola anotatzaileak. Aldaketa hori fonologikoa edo morfologikoa da? Beste adibide bat: *amoreakgatik* → *amoreengatik* bikotean, tarteko *ak* morfeman gertatzen den aldaketa, fonologikoa ala morfologikoa da?

Aurreko bi kapituluetan, III. eta IV. kapituluetan, maila morfologikoko informazioa erabili izan dugu zenbait esperimentutan ikasteko informazio gisa. Izan ere, *hitza-analisia* informazioaren bitartez ikasi denean bikoteetako hitz estandarraren segmentazio morfologikoa erabili da. Orain urrats bat aurrera eman nahi dugu kapitulu honetan, eta saiatu nahi dugu ikasteko prozesuan bikoteetako bi aldeen segmentazio morfologikoa erabiltzen. Esaterako, aurreko adibidearen kasuan honako informazio hau erabili nahi dugu ikasteko prozesuan: $amore + ak + gatik \rightarrow amore + en + gatik$.

Segmentazio morfologikoaren informazio hori lortzeko, ordea, aldaerak morfologikoki segmentatzeko moduren bat behar dugu, euskara estandarra analizatzeko gai den transduktoreak ez baititu aldaerak segmentatzen. Hori dela eta, kapitulu honen V.2 atalean aztertuko dugu zer tresna dauden hitzen segmentazioa lortzeko modu ez-gainbegiratuan, eta nola erabil ditzakegun

tresna horiek euskararen kasuan. Behin hori lortuta, V.3 atalean *analisia-analisia* moduko informazio berria erabiliko dugu WFST metodoarekin normalizazio-ataza ebazteko, eta lortutako emaitza berriak aurreko kapituluetan lortutakoekin konparatuko ditugu. Horren ondoren, V.4 atalean beste bide bat jorratuko dugu, eta saiatuko gara aldaeraren eta estandarraren arteko baliokidetza morfologikoak lortzen anotatuako informaziotik abiatuta (modu gainbegiratua). Baliokidetza horien bitartez analizatzaile estandarra hedatu ahal izango dugu, eta horrek normalizazio-ataza ebazteko beste bide bat zabalduko digu. Bai V.3 atalean, bai V.4 atalean planteatuko ditugun esperimentuetan, *Gero* corpus historikoa erabiliko dugu soilik. Bukatu baino lehen, V.5 atalean, WFST metodoan oinarritutako hiru normalizazio-sistemen emaitzak analizatuko ditugu kuantitatiboki zein kualitatiboki, eta haien mugak azalduko ditugu. Azkenik, V.6 atalean, kapitulu honetan ateratako ondorioak azalduko ditugu.

## V.2 Segmentazio morfologiko ez-gainbegiratua

Atal honetan aztertu nahi dugu nola lor dezakegun aldaeren segmentazio morfologikoa modu ez-gainbegiratuan. Beraz, bilatu behar dugu zer tresna dauden hori egiteko, horien artean bat aukeratu behar dugu, eta aztertu behar dugu zein den modurik egokiena tresna hori erabiltzeko gure testuinguruan, hots, euskarazko aldaerekin.

### V.2.1 Bibliografia

Morfologiaren ikasketaren arloak gai asko hartzen ditu bere barne, hala nola lexikoien eraikuntza, hitzen analisi morfologikoa, hitzen segmentazio morfologikoa, paradigma morfologikoen ikasketa eta abar.

Morfologiaren ikasketa automatikoa lortzeko jarraitu diren metodoak ikasketa ez-gainbegiratuan oinarritu dira batez ere, hau da, etiketatu gabeko datuekin egindako ikasketan. Hammarström eta Borin (2011) lanean ikasketa ez-gainbegiratuari buruzko azterketa zabala eta sakona egiten da.

Kapitulu honetan bereziki interesatzen zaigun gaia segmentazio morfologikoarena da, izan ere gai horren helburu nagusia hitzak morfemetan banatzea da. Segmentazio horrek lortzen duen analisia erabilgarria da hizkuntzaren inguruko hainbat aplikaziotan (Ruokolainen *et al.*, 2016): hizketaren ezagutzan, informazio-berreskuratzean, itzulpen automatikoan eta abar (ikus aipatutako lana erreferentzia bibliografiko zehatzetarako).

Aldaerako hitzak nahi ditugu segmentatu gure kasuan, eta hori lortzeko teknika ez gainbegiratuak interesatzen zaizkigunez, bibliografiara jo dugu.

Koskenniemi-k (2013) bi tresna edo algoritmo aipatzen ditu hitzak morfologikoki segmentatzeko eta biak dira ez-gainbegiratuak: *Linguistica* (Goldsmith (2001) eta Goldsmith (2006)) eta *Morfessor* (Creutz eta Lagus (2004) eta Creutz eta Lagus (2005)). Tresna ezberdinak dira eta erabiltzen dituzten algoritmoak ere halakoak dira. Dena den, Morfessorren egileek diote bere tresna flexio aberatseko hizkuntzekin lan egiteko garatu dela (finlandierarekin lan egiten dute hainbat esperimentutan):

*Morfessor is a general model for the unsupervised induction of a simple morphology from raw text data. Morfessor has been designed to cope with languages having predominantly a concatenative morphology and where the number of morphemes per word can vary much and is not known in advance. This distinguishes Morfessor from resembling models, e.g., Goldsmith (2001), which assume that words consist of one stem possibly followed by a suffix and possibly preceded by a prefix.*

Kontuan izanik euskara, finlandiera bezala, hizkuntza eranskaria dela, Morfessor tresna egokiena dirudi lanarekin aurrera jarraitzeko. Aurreko definizioan esaten den moduan, tresnaren helburua da hizkuntza-eredu bat induzitzea modu ez-gainbegiratuan testu hutseko corpus batetik abiatuta. Oinarrizko hizkuntza-eredua (*Baseline model* deitzen diote) morfemaz osatutako lexikoi bat da ($\mathcal{M}$), beraz, helburua da lexikoi onena bilatzea sarrerako corpusa segmentatzeko. Matematikoki adierazita:

$$argmax_{\mathcal{M}} P(\mathcal{M}|corpus) = argmax_{\mathcal{M}} P(corpus|\mathcal{M}) \cdot P(\mathcal{M})) \qquad \text{(V.1)}$$

Beraz, lexikoi onena bilatzeak bi zati ditu: hizkuntza-ereduaren probabilitatea kalkulatzea ($P(\mathcal{M})$) eta corpusaren probabilitate maximoaren estimazioa eredu jakin baten arabera ($P(corpus|\mathcal{M})$).

Beraz, gure lanarekin aurrera jarraitzeko Morfessor aukeratu dugu, eta hurrengo atalean tresna hori erabiltzen ikasteko egin ditugun esperimentuak eta horien ondorioak azalduko ditugu.

Halere, aipatu beharra dago morfologia ikasteko erabili diren metodo guztiak ez direla ez-gainbegiratuak izan, eta Ruokolainen *et al.* (2016) lanean jasotzen den moduan, azken urteetan gero eta interes gehiago sortzen ari da teknika erdi-gainbegiratuetan. Oro har, metodo erdi-gainbegiratuen helburua da etekina ateratzea bi motatako datuei, etiketatuak zein etiketatu gabeak, eredu zehatzagoak lortzeko.

Aipatutako lanean segmentazio morfologikoa egiten duten hiru metodo konparatzen dira, eta hiru horietako bat Morfessor da (Linguistica tresna ez da aipatzen lan horretan). Beste bi metodoak erdi-gainbegiratuak dira: bata "*Adaptor Grammar* (AG)" hurbilpenean oinarritzen da (Sirts eta Goldwater, 2013), eta bestea "*Conditional Random Fields* (CRF)" proposamenean

(Ruokolainen *et al.* (2013) eta Ruokolainen *et al.* (2014)). Etorkizunerako lanetan kontuan hartu beharreko proposamenak dira metodo berri horiek.

### V.2.2 Hitzen segmentazioa Morfessor bitartez

Morfessor tresna baliatu nahi dugu, beraz, aldaerak segmentatzeko eta erabili dugun bertsioa Morfessor 1.0[1] izan da. Tresnak bertsio bat baino gehiago eskaintzen du haren webgunean ikus daitekeen moduan, ez baitiote tresnaren garapenari utzi eta 2005etik aurrera bertsio eta lan berriak eskaini dituzte bertan. Dena den, lan honetan zerotik abiatu behar genuenez tresna erabiltzen, lehenengo bertsioa aukeratu genuen. Gainera, bertsio horri lotuta oso lan egokia aurkitu genuen tresnarekin lanean hasteko: Creutz eta Lagus (2005) lana.

Sarrerako informazio gisa, fitxategi bat eman behar zaio Morfessorri, non lerro bakoitzeko hitz bat ageri behar den. Hitzaren aurretik, aukeran, haren maiztasuna adierazten duen zenbaki bat eman dakioke; ez bada ezer ageri maiztasun hori 1 dela suposatzen du tresnak. Dena den, maiztasuna ez bada ematen eta hitza behin baino gehiagotan ageri bada, agerpenak metatu egiten ditu tresnak berak.

Datuez gain, aukerazko hainbat parametro ditu Morfessorrek (ikus Creutz eta Lagus (2005) xehetasunetarako) eta horregatik, tresna erabili baino lehen aldaerako hitzak segmentatzeko, beharrezkoa izan da aztertzea zein diren aukera egokienak euskarazko hitzak segmentatzeko. Hala, erreferentzia bat behar dugu zeinarekin konparatu Morfessorrek egiten dituen segmentazioak, eta horretarako, nahitaez, euskara estandarrarekin egin behar izan dugu lan lehendabizi. Suposatzen dugu euskara estandarrerako parametro egokienak, ez-estandarrerako ere egokienak izango direla.

Atal honetan labur deskribatuko dugu Morfessorrekin eta euskara estandarrarekin egin ditugun esperimentuak eta lortutako emaitzak. Esperimentu horiek planteatzeko, Creutz eta Lagus egileen 2004 eta 2005eko bi lanetan deskribatzen diren esperimentuetan oinarritu gara (finlandiera zein ingelesarekin egiten dute lan). Egin ditugun esperimentu horien helburua beti bera da: konparatzea Morfessorrek (hainbat aukeraren arabera) lortzen dituen segmentazioak eta euskara estandarraren analizatzaileak ematen dituenak, horrela jakiteko zein diren aukera egokienak Morfessor erabiltzeko euskarazko hitzak segmentatzeko ahalik eta doitasun handienarekin.

[1] `http://www.cis.hut.fi/projects/morpho/` 2016-04-20an atzitua

| tokenak | **euskara formak** | finlandiera formak | ingelesa formak |
|---:|---:|---:|---:|
| 10.000 | **5 200** | 5 500 | 5 200 |
| 50.000 | **17 300** | 20 000 | 7 200 |
| 250.000 | **50 000** | 65 000 | 17 000 |
| 11.500.000 | **436 000** | - | - |
| 12.000.000 | - | - | 110 000 |
| 16.000.000 | | 1 100 000 | - |

V.1 Taula: Morfessor tresnarekin esperimentuak egiteko erabili diren multzoen tamainak. Forma kopuruak batezbestekoak dira (bost multzorekin lortutakoak) euskararen kasuan. Beste bi hizkuntzetako kopuruak Creutz eta Lagus (2005) lanean ematen dituztenak dira.

#### V.2.2.1 Corpusa

Euskara estandarreko corpus gisa *Euskaldunon Egunkaria* egunkariko corpus bat erabili dugu, 2000–2002 urteetako testuekin sortutakoa. Corpusa tokenizatu eta garbitu ondoren (karaktere arraroak eta antzekoak kentzeko) 23 milioi hitz dituen corpusa osatu dugu, 11,5 milioi hitzeko bi zatitan banatu duguna zoriz esperimentuetarako: garapen- eta test-zatiak.

Morfessorren segmentazioak ebaluatu nahi ditugu test gisa tamaina ezberdineko multzoak erabiliz, horrela ikusteko nola aldatzen diren emaitzak datu kopurua aldatzen denean. Erabili ditugun tamainak V.1 taulan ageri dira eta Creutz eta Lagus (2005) lanean erabiltzen dituzten berberak dira. Lan horretan finlandierarekin zein ingelesarekin egiten diren esperimentuak errepikatu nahi ditugu euskararekin, horrela tresnaren funtzionamendua aztertzeko eta egokitzeko, eta hala, taula berean kopiatu ditugu lan horretan ematen dituzten datuak bi hizkuntza horietarako. Tamaina txikieneko multzoen kasuan, hau da, 10.000, 50.000 eta 250.000 tamainako multzoetan, ez da multzo bakar bat egin, bost baizik, eta horregatik bigarren zutabean ageri den forma kopurua, bost multzoetan ageri direnen batezbestekoa da.

Forma kopuruari dagokionez, garbi ikusten da euskarak antza gehiago duela finlandierarekin ingelesarekin baino (biak dira flexio handiko hizkuntzak).

#### V.2.2.2 Esperimentuak

Esperimentuetan konparatu nahi dira Morfessorrek egiten dituen segmentazioak eta analizatzaile estandarrak egiten dituenak. Esperimentuen emaitzak eman baino lehen beharrezkoa jotzen dugu azaltzea, labur eta xehetasun gehiegitan sartu gabe, nola egin den bi segmentazioren arteko konparazioa, hau da, nola egin den Morfessorren segmentazioen ebaluazioa eta zein zailtasunekin topatu garen hori egiteko unean.

Segmentazio bakoitzean detektatutako morfema-mugak dira ebaluatu nahi ditugunak, beraz, segmentazio bakoitzeko hainbat neurri kalkulatuko ditugu: zenbat morfema-muga detektatzen dituen Morfessorrek ($M$), zenbat analizatzaile estandarrak ($E$), eta zenbat datozen bat, hau da, detektatutakoetatik zenbat diren "zuzenak" ($Z$). Kalkulu horiek segmentatutako sarrera bakoitzeko egin behar dira, eta zehaztasun batzuk hartu behar dira kontuan:

1. Morfema-muga zuzenak kalkulatzeko, bi segmentazioak konparatu behar dira, Morfessorrena eta analizatzaile estandarrarena. Konparazio hori, ordea, segmentazio-kateen eskuinetik zein ezkerretik egin daiteke, eta kontua da emaitza ez dela zertan berdina izan alde batetik zein bestetik konparatuta. Beraz, bi konparazioak egin dira, eta emaitza onena eman duena hartu da ebaluaziorako.

   Esaterako, *adierazpenean* hitza segmentatzen badugu analizatzaile estandarraren bitartez, honako segmentazio hau lortzen dugu:
   *adierazpenean* → `adierazpen + an`
   Demagun Morfessorrek egin duen segmentazioa hitz horretarako beste hau dela:
   *adierazpenean* → `adieraz + pen + ean`
   Konparazioa eskuinetik eginez gero, emaitza da ez dagoela morfema-muga zuzenik, `+an` eta `+ean` morfemak ez datozelako bat. Baina konparazioa ezkerretik egiten bada, emaitza da morfema-muga zuzen bat dagoela. Izan ere, ezkerretik hasita ondorioztatzen da lehenengo muga okerra dela (`adieraz` zatiaren ondoren dagoena), baina bigarrena zuzena kontsideratzen da, bi segmentazioetan dauden kateak berdinak baitira: `adierazpen+` (aurreko muga ez da kontuan hartzen bigarren konparazioan).

2. Analizatzaile estandarrak segmentazio bat baino gehiago ematen du askotan. Kasu horietan Morfessorren segmentazioa segmentazio estandar guztiekin konparatu da, eta emaitza onena eman duen kasua hartu da ebaluaziorako.

Adibide gisa, *antolatzaileek* hitza segmentatzen bada analizatzaile estandarraren bitartez, bi segmentazio lortzen dira:
*antolatzaileek* → `antola + tzaile + ek`
*antolatzaileek* → `antolatzaile + ek`
Demagun Morfessorrek egin duen segmentazioa hitz horretarako honako hau izan dela:
*antolatzaileek* → `an + to + la + tzaile + ek`
Morfessorren segmentazioak lau morfema-muga ditu. Lehenengo segmentazio estandarrarekin konparatzen badugu esango dugu lau horietatik bi morfema-muga zuzenak direla. Bigarren segmentazio estandarrarekin konparatuz gero, ordea, morfema-muga bakar bat da zuzena.

Beraz, segmentazio bakoitzeko aipatutako hiru balioak kalkulatuz gero —$M$, $E$ eta $Z$—, segmentazio guztiak har daitezke kontuan multzo osoan lortutako doitasuna eta estaldura kalkulatzeko: $P = \Sigma Z/\Sigma M$ eta $R = \Sigma Z/\Sigma E$.

Bi esperimentu egin ditugu Morfessorrekin, bat jakiteko zein den corpuseko informazio egokiena emaitzei begira, eta bestea Morfessorren parametro bat doitzeko:

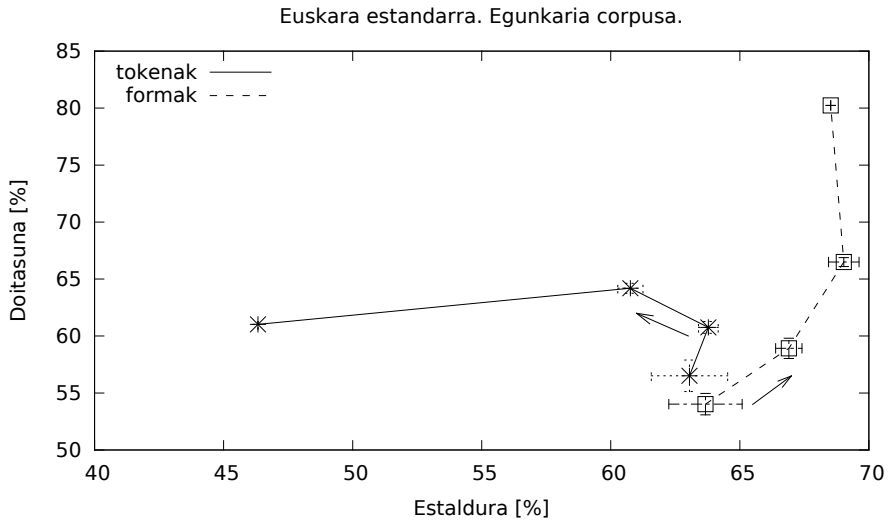
1. Lehenengo esperimentuan aztertzen da nola lortzen diren segmentazio egokienak, sarreran tokenak emanez (hitz bakoitza corpusean ageri den bezainbeste aldiz) edo formak emanez (hitz bakoitza behin). Horrez gain, aztertu nahi da ea corpusaren tamainak garrantzia duen sarrera gisa eman behar den informazio egokiena aukeratzeko.

2. Bigarren esperimentuan aztertu nahi da Morfessorren sarrerako parametro baten eragina emaitzetan. Parametro hori `gammalendistr` izeneko parametroa da, eta gero ikusiko dugun moduan, aldatu egiten du lexikoiaren probabilitatea ($P(\mathcal{M})$) kalkulatzeko modua.

### Lehenengo esperimentua: token vs forma

Esperimentu honetan Morfessorren oinarrizko eredua erabili dugu (*Baseline model*), hau da, ez dugu sarrerako parametrorik aldatu eta lehenetsitako algoritmoarekin egin dugu lan. Morfessorri eman diogun sarrerako informazioa izan da corpusaren testeko zatian egin ditugun azpimultzoak (gogoratu lehenengo 10.000–250.000 tamaina bakoitzeko 5 azpimultzo ditugula). Gero, azpimultzo bakoitzarekin lortutako segmentazioak ebaluatu ditugu.

V.1 irudian tamaina bakoitzarekin lortutako doitasuna eta estaldura ageri dira, eta ikusten da tokenekin zein formekin lortutako grafikoen joerak oso




V.1 Irudia: Morfessorren oinarrizko ereduaren emaitzak tokenak zein formak emanez datuetan. Grafikoetako balioak batezbestekoak dira (5 multzo ezberdinekin lortutakoak) eta desbideratzea inguruko tartea adieraziz ematen da. Geziek datuen hazkuntza adierazten dute.

ezberdinak direla. Formak erabiliz gero, emaitza hobeak lortzen dira multzoaren tamaina hazten den heinean: doitasuna eta estaldura, biak, handitzen dira, azken tamainan izan ezik, horretan estaldura pixka bat jaisten baita. Tokenak erabiliz gero, ez dago tendentzia garbirik: doitasuna hazten da hasieran, baina ez estaldura, eta tamaina handieneko multzoarekin estaldura izugarri jaisten da.

Oro har, grafiko horretatik ateratzen dugun ondorioa da segmentazio hobeak lortzen direla Morfessorri formak emanez, eta ez tokenak. Tamaina txikieneko multzoekin (10.000 token) ez dago hain garbi hala denik, tokenekin zein formekin lortutako emaitzak parean baitaude; hurrengo tamainarekin, ordea, estalduran dagoen diferentzia garrantzitsuagoa da doitasunean dagoena baino, eta hortik aurrerako tamainetan, emaitzak askoz hobeak dira formak erabiltzen direnean.

Konparatzen badugu lortutako grafikoa eta Creutz eta Lagus (2005) artikuluan ageri direnak finlandierarako eta ingeleserako hauek dira ondorioak:

- Bi hizkuntza horietan ere, emaitza onenak sarreran formak emanez lortzen dira: estaldura nabarmen handiagoa token ordez formak era-

biliz. Gure kasuan ez dago halako diferentziarik tamaina txikienean, baina gero bai.

- Finlandierarekin konparatuta, euskararekin lortutako estalduraren balioak hobeak dira kasu guztietan. Aldiz doitasunean, ez da hori gertatzen.

- Ingelesarekin konparatuta, euskarazko emaitzak alderantzizkoak dira: doitasun aldetik, oro har, hobeak dira, baina estalduran ez.

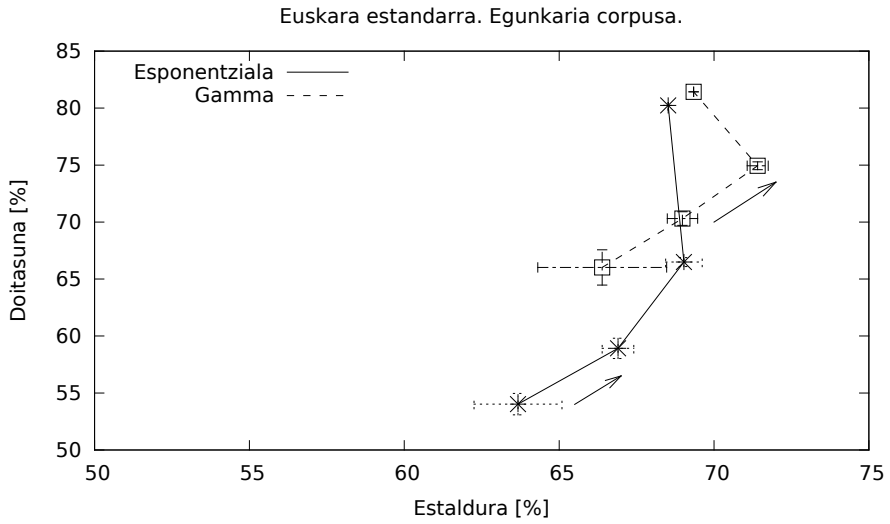
### Bigarren esperimentua: gamma banaketa

Esan dugunez, bigarren esperimentuan `gammalendistr` izeneko parametroaren efektua neurtu nahi dugu. Parametro horren arabera aldatu egiten da lexikoiaren probabilitatea ($P(\mathcal{M})$) kalkulatzeko modua. Izan ere, Morfessorren oinarrizko ereduak banaketa esponentzial bat erabiltzen du lexikoiko morfemen luzerei probabilitateak esleitzeko. Aktibatzen bada `gammalendistr` parametroa, ordea, gamma banaketa bat erabiltzen du probabilitate horiek esleitzeko, eta lexikoiaren probabilitatearen kalkulua aldatu egiten da.

Aurreko esperimentuaren emaitzetan oinarrituta, esperimentu berri honetan formak erabili dira beti sarrera gisa, baina ez dakigunez gamma banaketaren zein baliorekin lortuko diren emaitza onenak (bi zenbaki erreal adierazi behar dira parametro horretan), lehendabizi doikuntzarako hainbat proba egin ditugu corpuseko garapen zatiarekin informazio hori lortzeko. Hala, zenbait esperimentu egin ditugu sarrera gisa forma kopurua aldatuz, eta balio ezberdinak probatuz gamma banaketarako, eta horrela jakin dugu zein baliorekin lortzen diren emaitza onenak. Gero, doikuntza horretan finkatutako balioak horiek erabili ditugu testeko multzoekin, eta lortutako emaitzak V.2 irudian ageri dira.

Gamma banaketarekin lortutako emaitzekin batera, grafikoan banaketa esponentzialarekin lortutakoak ageri dira (aurreko V.1 irudian azaldutako berberak dira). Biak konparatuz gero, argi dago gamma banaketa erabiltzea oso onuragarria dela hasierako tamainetan, hau da, 250.000 token arteko multzoetan. Aldiz, multzoa oso handia denean, gure esperimentuetan 11,5 milioi token, bi banaketek lortutako emaitza antzekoa da.

Creutz eta Lagus (2005) lanean portaera hori bera islatzen da finlandierarako eta ingeleserako.




V.2 Irudia: Emaitzen konparazioa gamma banaketa bat zein banaketa esponentziala erabilita morfemen luzerei probabilitateak esleitzeko. Grafikoetako balioak batezbestekoak dira (5 multzo ezberdinekin lortutakoak) eta desbideratzea inguruko tarteak adierazten du. Geziek datuen hazkuntza adierazten dute.

## V.3 Normalizazioa morfologiaren bitartez

Aurreko atalean aztertu dugu zein den modurik egokiena Morfessor erabiltzeko euskara estandarrarekin, eta atal honetan ikusiko dugu nola erabili tresna hori aldaerako hitzak segmentatzeko.

### V.3.1 Ikasteko informazio berria: *analisia-analisia*

*Gero* corpuseko ikasteko zatian ditugun bikoteen *analisia-analisia* motako informazioa lortu nahi dugu WFST metodoa aplikatzeko berriro. Beraz, lehenengo urratsa aldaerak segmentatzea da Morfessor bitartez (estandarrak segmentatu ditugu jadanik estandarraren analizatzailearen bitartez).

Lehenengo urratsa Morfessorren araberako *segmentazio-eredu* bat lortzea izan da tresnari testua emanez. Aldaeraren eredua lortu nahi dugunez, *Gero* obra osoaren testua erabiliko dugu urrats honetan, eta aurreko ataleko bi esperimentuen emaitzak hartuko ditugu kontuan segmentazio-eredu hori lortzeko. Hau da, aurreko esperimentuak euskara estandarrarekin egin

dira lortutako segmentazioak ebaluatu ahal izateko, eta orain, esperimentu horietatik ateratako ondorioak aldaera segmentatzeko erabiliko dira. Gure hipotesia da hizkuntza berarekin ari garela, eta ondorioak bietarako balio dutela, hots, estandarrerako eta aldaerarako.

*Gero* obraren analisia egin dugunean II. kapituluan ikusi dugu 97.000 token inguru dituela obra horrek (ikus bigarren kapituluko II.2 taula). Lehen esperimentuko ondorioen arabera, beraz, obrako formak eman dizkiogu Morfessorri sarrera gisa (ondorioztatu dugu 50.000 tokenetik gora hobe dela formak ematea). Horrez gain, bigarren esperimentuaren arabera, gamma banaketa erabili da segmentazio-eredu hobea lortzeko.

*Gero* obrari dagokion segmentazio-eredua modu ez-gainbegiratuan lortu ondoren, bigarren urratsean eredu hori erabili dugu ikasteko bikoteetako aldaerako zatia segmentatzeko. Horrela lortu ditugu, esaterako, $amoreakgatik \rightarrow amore + ak + gatik$ edota $elkharri \rightarrow elkhar + ri$ moduko segmentazioak, eta aldaera segmentatu ondoren, prestatu ahal izan dugu *analisia-analisia* motako informazio berria bikote guztietarako ($amore + ak + gatik \rightarrow amore + en + gatik$ eta abar) .

Informazio hori izanik, gure WFST metodoa aplikatu dugu (IV. kapituluko esperimentuetako metodoa) normalizazioa egiteko. Ikasteko corpusarekin informazio berria prestatu da, eta testeko corpusarekin egin da ebaluazioa (oraingo honetan ez dugu inolako doikuntza-esperimenturik planteatu). Beraz, Phonetisaurus tresnari *analisia-analisia* informazioa eman diogu sarrerako hiztegian, eta gero, ebaluazioa egiteko test zatiarekin, 20 erantzun eskatu dizkiogu deskodeketa-urratsean (20 izan delako *hitza-analisia* informazioarekin erabilitako balioa). Ikasteko bikoteei dagokienez, bi ebaluazio egin ditugu berriro, IV. kapituluko IV.4.4 atalean egin dugun moduan: batean anotatutako bikoteak besterik ez dira erabili ikasteko; bestean ikasteko corpusean ageri diren hitz estandarrak ere sartu dira ikasteko informazioan.

Emaitzak V.2 taularen hasieran bildu ditugu, eta konparazioa errazteko asmoz, aurreko kapituluko IV.21 taulan ageri diren emaitzak kopiatu ditugu berriro taula horretan.

Ikusten denez, emaitzak ez dira hobetu. Konparatzen badira *hitza-analisia* informazioak lortu dituenekin, oso antzekoak dira (ia berdinak ikasteko hitz estandarrak erabili direnean), baina *hitza-hitza* da emaitza onenak lortzen dituen ikasteko aukera.

| Metodoa | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| WFST *analisia-analisia* | | | |
| anotatuak | 90,68 | 75,62 | **82,47** |
| anotatuak + estand. guztiak | 91,63 | 77,39 | **83,91** |
| WFST *hitza-analisia* | | | |
| anotatuak | 91,08 | 77,56 | 83,78 |
| anotatuak + estand. guztiak | 90,91 | 77,74 | 83,81 |
| WFST *hitza-hitza* | | | |
| anotatuak | 91,53 | 78,27 | 84,38 |
| anotatuak + estand. erdia | 91,84 | 79,51 | 85,23 |
| Oinarri-lerroa | 94,87 | 39,22 | 55,50 |

V.2 Taula: *Gero* corpusa. Azken ebaluazioko emaitzak ikasteko informazioaren arabera.

### V.3.2 Segmentazio-eredua hobetzeko saioak

Informazio morfologiko berria erabiliz lortutako emaitzak ez dira izan guk espero bezain onak, eta saiatu gara Morfessorrek lortzen duen segmentazioa nolabait egokitzen, horretarako kontuan izanik analizatzaile estandarrak ematen duen informazioa.

Aurreko atalean ikusi dugu Morfessor erabili dugula segmentazio-eredu bat lortzeko *Gero* obraren testutik abiatuta, eta gero eredu hori erabili dugula ikasteko corpusaren aldaerak segmentatzeko. Segmentazio-eredu hori testu hutsa da, ez du inolako formatu berezirik, V.3 irudian ikus daitekeen moduan. Irudi horretan Morfessorrek itzuli duen ereduaren zati txiki bat ikus daiteke (sarrera gisa *Gero* obrako forma guztiak eman zaizkio): sarrerako hitzen segmentazioak ageri dira, eta segmentazio bakoitzaren aurretik beti 1 balioa. Aurreko balio horrek adierazten du zenbat aldiz azaldu den hitz hori sarreran, eta formak eman dizkiogunez, beti da 1.

Formatu sinple hori ikusita bururatu zaigu Morfessorrek lortutako eredu hori "osatzea" edota "egokitzea", eta bi saio egin ditugu helburu hori lortzeko:

- Lehenengo saioan obrako hitz estandarren segmentazio estandarra gehitu diogu Morfessorren ereduari haren formatuaren arabera. Hau da, obrako hitz estandarrak segmentatu dira analizatzaile estandarrarekin, eta analisi horiek gehitu dizkiogu eredu ez-estandarrari. Ideia horrekin

```
1 bekhatu + ak + gatik
1 bekhatu + an
1 bekhatu + aren
1 bekhatu + aren + a
1 bekhatu + ari
1 bekhatu + az
1 bekhatu + ei
1 bekhatu + ek
1 bekhatu + en
1 bekhatu + entzat
1 bekhatu + etan
1 bekhatu + etara
1 bekhatu + etarat
1 bekhatu + etarik
1 bekhatu + ez
1 bekhatu + gatik
1 bekhatu + ra
1 bekhatu + rekin
1 bekhatu + ren
1 bekhatu + rik
1 bekhatu + tan
1 bekhatu + tarik
1 bekhatu + tik
1 bekhatu + z
```

V.3 Irudia: Morfessorrek itzulitako ereduaren zati bat: lerro bakoitzean sarrerako formari dagokion segmentazioa itzultzen du.

bi proba egin dira. Lehenengoan analisi bakarra gehitu da hitz estandar bakoitzeko (lema luzeena duen analisia; luzera bera izanez gero, analisi-kate motzena duena). Bigarrenean, berriz, analizatzaile estandarrak ematen dituen analisi guztiak gehitu dira. Ebaluazioa egin dugu berriro (ikasteko bikote anotatuak soilik erabiliz) eta emaitzak ez dira ia aldatu (V.2 taulako lehen errenkadakoekin konparatu behar dira): analisi bakarra gehituta lortutako balioak izan dira $P = 90{,}89$, $R = 75{,}80$, $F = 82{,}66$, eta analisi posible guztiak gehituta, berriz, $P = 91{,}08$, $R = 75{,}80$ eta $F = 82{,}74$.

- Bigarren saioan Morfessorren ereduari informazioa gehitu ordez, ereduaren zati bat ordeztu dugu: hitz estandarrei dagokien zatia izan da, hain zuzen, ordeztu duguna.

  Morfessorren ereduak hitz guztien segmentazioak hartzen ditu barne, aldaerak zein estandarrak. Aldaeren segmentazioa dagoen moduan utzi da eredu horretan, baina estandarren segmentazio gisa analizatzaile estandarrak emandakoa jarri da (hau da, hitz estandarretan Morfessorrek egindako segmentazioa ordeztu da). Ebaluazioa eginda, berriro ere, emaitzak ez dira ia aldatu (eta aldaketa txikia okertze aldera izan da): $P = 90{,}62$, $R = 75{,}09$ eta $F = 82{,}13$.

Beraz, saiatu arren, ez dugu lortu *analisia-analisia* informazioarekin lortutako emaitzak hobetzea (V.2 taulakoak).

## V.4 Morfologiaren ikasketa modu gainbegiratuan

Kapitulu honen sarreran esan dugun moduan, atal honetan bide berri bat jorratu dugu aldaeraren eta estandarraren arteko baliokidetza morfologikoak lortzeko, horretarako anotazioko informazioa baliatuta.

### V.4.1 Morfemen lerrokatzea

Baliokidetza morfologiko horiek lortzeko estrategiak antza du Ahlberg *et al.* (2014) lanean erabiltzen dutenarekin paradigmak aurkitzeko. Egin duguna izan da *hitza-analisia* motako bikoteetan dagoen informazioa baliatzea bi aldeetan dauden morfemak parekatzeko, modu horretan automatikoki detektatzeko aldaerari dagozkion lemak, adizkiak edota atzizkiak.

Argi dezagun ideia hori adibide pare batekin. Corpusaren ikasteko zatian honako bikote hauek ditugu beste hainbatekin batera:

`akhusatuak` → `akusatuak`
`bertzetik` → `bestetik`
`dadukanak` → `daukanak`
`derakunari` → `digunari`

Alde estandarreko hitzaren segmentazioa lortuz gero estandarraren analizatzailearekin, ezaguna dugun *hitza-analisia* motako informazioa daukagu:

`akhusatuak` → `akusatu + ak`
`bertzetik` → `beste + tik`
`dadukanak` → `dauka + n + ak`
`derakunari` → `digu + n + ari`

Informazio hori aztertuta, erraza da programa bat idaztea bi aldeetako lemak edo adizkiak parekatzen saiatzeko: aurretik ez dakigu zein den aldaerei dagokiena, baina saia gaitezke hori inferitzen. Hori lortzeko aukera bat izan daiteke bi aldeak letraz letra konparatzen hastea eskuinetik. Konparazioa amaitzen da azken morfema-mugara iristean (bat baino gehiago badago, eskuinekoak ez dira kontuan hartzen) edota bat ez datozen letrak topatu arte.

Esaterako, azken bikotearen kasuan, `derakunari` → `digunari`, letraz letra konparatzen ditugu eskuinetik `i`, `r` eta `a`. Puntu horretan estandarrean morfema-muga batekin egiten dugu topo, baina azkena ez denez, aurrera jarraitzen dugu eta `n` konparatzen dugu. Berriro morfema-mugara iritsi gara

estandarrean, orain bai azkenekora, eta ondorioz suposatzen dugu bi aldeetan geratzen dena baliokidetza bat dela (kasu honetan adizkien artekoa): `deraku` → `digu`.

Algoritmo bera aplikatuta gainontzeko adibideetan, honako baliokidetza hauek lortzen dira: `akhusatu` → `akusatu`, `bertze` → `beste`, `daduka` → `dauka`.

Garbi dago horrelako baliokidetzarik ez dela bikote guztietan lortuko, eta hainbat kasutan, azken morfema-mugara iritsi baino lehen amaituko da konparazioa bi aldeetako letrak bat ez datozelako. Hori gertatzen da, adibidez, `beranduraiño` → `berandu+raino` bikotean. Konparazioa eskuinetik hasiz gero, `o` bat dator bietan, baina gero `ñ` eta `n` ez, beraz, konparazioa amaitzen da baliokidetzarik detektatu gabe. Beste hainbeste gertatzen da bikote honetan: `arazitzea` → `aratz+te+a`. Eskuinetik hasita, lehenengo bi karaktereak, `a` eta `e`, bat datoz bi kateetan, baina gero, `z` eta `t` ez, eta konparazioa amaitzen da lema-baliokidetzarik proposatu gabe.

Badira beste zenbait kasu non ez den konparaziorik behar: hitz estandarraren analisiari dagokion osagaia hitz bakar batez osatuta badago (ez dago morfema-mugarik), zuzenean hartzen da lemen edo adizkien arteko baliokidetzat.

Lemekin eta adizkiekin egin den moduan, atzizkien arteko baliokidetzak detekta daitezke antzeko algoritmoa aplikatuta, baina oraingoan konparazioa ezkerretik hasita. Esaterako, `beranduraiño` → `berandu+raino` bikoteko osagaiak ezkerretik konparatzen baditugu, morfema-mugaraino iritsiko gara eta lortuko dugu `+raiño` → `+raino` atzizki-baliokidetza; `nitzaz` → `ni+taz` adibidean `+tzaz` → `+taz` baliokidetza eta abar.

Atzizkien arteko baliokidetzen bila topo egin dezakegu estandarraren aldean atzizki bat baino gehiago duten kasuekin. Esaterako, `kalteakgatik` → `kalte+en+gatik` bikotea ezkerretik konparatuz gero lehen morfema-mugaraino, `+akgatik` → `+en+gatik` baliokidetza lortzen da, baina estandarraren aldean atzizki bat baino gehiago dagoenez oraindik, bigarren konparazioa egiten da atzizkiak parekatzeko, oraingoan eskuinetik hasita. Hala, adibide horretan lortzen den atzizkien arteko parekatzea da `+ak` → `+en`.

Zenbait kasutan, bi motatako baliokidetzak lor daitezke *hitza-analisia* bikote berean. Esaterako, `disposizioneari` → `disposizio+ari` adibidean, lemen zein atzizkien arteko baliokidetza lortzen da azaldutako algoritmoen arabera: `disposizione` → `disposizio` eta `+neari` → `+ari`.
Halako kasuetan, ezker aldean lortutako baliokidetzari eman diogu "lehentasuna" eta atzizkiena alboratu dugu.

### V.4.2 Estandarraren hedapena eta ebaluazioa

Azaldu berri dugun algoritmoak jarraituta aukera daukagu, beraz, eskuzko anotaziotik (modu gainbegiratuan) aldaerako eta estandarreko lemen, adizkien eta atzizkien arteko baliokidetzak proposatzeko: `akhusatu` → `akusatu` edota `bertze` → `beste` moduko baliokidetzak. Eta baliokidetza horiekin analizatzaile estandarra heda daiteke, hau da, euskarazko lexikoa zabal daiteke morfema berriak sartuta estandarrekin batera, eta horrela, teorikoki, analizatzaile hedatu hori gai izango da onartzeko (edo analizatzeko) zenbait aldaera eta gainera estandar batekin lotuko ditu. Esaterako, hedatzen badugu lexiko estandarra *akusatu* sarrerarekin batera *akhusatu* jarriz, analizatzaile hedatu berria gai izango da *akhusatuarekin* hitza onartzeko eta *akusatuarekin* estandarrarekin lotzeko.

Baina analizatzailea hedatzeko unean galdera asko sortzen dira: Zenbat informazio erabili hedapena egiteko? Inferitutako baliokidetza guztiak edo behin baino gehiagotan ageri izan direnak anotazioan? Merezi du hedapenean ezberdin tratatzea lemei edota adizkiei dagozkien baliokidetzat eta atzizkiei dagozkienak? Intuizioa da asko hedatzeak estaldura hobetuko duela, baina doitasuna gutxitu. Bestalde, zuhurra izateak hedapena egiteko unean, doitasuna handituko du baina, seguru asko, estaldura gutxitu.

Kontuan izanik eskuzko anotazioa ez dela oso handia (1.000 bikote inguru), hauek izan dira egin ditugun probak:

- Analizatzailea hedatu dugu eskuzko anotaziotik erauzitako morfema berri guztiekin. Analizatzaile hedatu horri *hed-lem1atz1* deitu diogu.

- Atzizkiak laburragoak izan ohi direnez (eta horrek "zarata" sor dezakeenez), analizatzailea hedatu dugu eskuzko anotaziotik erauzitako lema eta adizki guztiekin, baina atzizkiei dagokienez, gutxienez 2 aldiz topatu behar izan da atzizki berri bat analizatzailean sartzeko. Analizatzaile hedatu horri *hed-lem1atz2* deitu diogu.

- Analizatzailea hedatu dugu lema, adizki edo atzizki berri batekin, 2 aldiz gutxienez aurkitu izan bada eskuzko anotazioan morfema hori. Analizatzaile hedatu horri *hed-lem2atz2* deitu diogu.

Beraz, hiru analizatzaile berri ditugu eta hirurek hedatzen dute estandarra, hau da, hirurak dira gai zenbait aldaera onartzeko estandar batekin lotura eginez. Horren arabera, analizatzaile hedatu horiek planteatu dugun normalizazio-ataza ebazteko erabil daitezke.

Analizatzaile berri bakoitzaren egokitasuna neurtzeko, ebaluazioa egin dugu *Gero* corpuseko test zatiarekin eta emaitzak V.3 taulan jaso ditugu.

| Analizatzailea | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| *hed-lem1atz1* | 95,34 | 57,77 | 71,95 |
| *hed-lem1atz2* | 95,60 | 57,60 | 71,89 |
| *hed-lem2atz2* | 98,17 | 37,99 | 54,78 |

V.3 Taula: Analizatzaile hedatuekin lortutako emaitzak test-corpusean.

Bertan ikusten denez, *hed-lem2atz2* analizatzaileak (gutxien hedatu dena) doitasun handia lortzen du, % 98 ingurukoa, baina estaldura txikia, % 38 ingurukoa. Lortu dituen balio horiek konparagarriak dira oinarri-lerroak lortu dituenekin (ikus V.2 taulako azken errenkada). Beste bi analizatzaileek lortutako emaitzen arabera –oso antzekoak biak–, argi dago oinarri-lerroa gainditzeko beharrezkoa dela anotaziotik erauzi diren lema berri guztiekin hedatzea analizatzailea: doitasuna pixka bat jaisten bada ere, 2–3 puntu, estaldura 20 puntu inguru igotzen da, eta hala, bi analizatzaile horiek lortzen duten $F$ neurria asko hobetzen da bai lehendabizikoarekin konparatuta, bai oinarri-lerroarekin konparatuta.

Hiru analizatzaile hedatuen doitasuna altua dela eta, aztertu dugu ea abantailaren bat lor dezakegun analizatzaile horien erantzuna konbinatzen badugu IV. kapituluan landu dugun WFST metodoak ematen duenarekin. Sistema bat baino gehiago eraiki dugunez WFST metodoarekin, horietako bat aukeratu dugu konbinazioaren proba egiteko: anotatutako bikoteak soilik eta *hitza-hitza* informazioa erabiltzen duen sistema.

Erantzunen konbinazioa egiteko, analizatzaile hedatuen erantzunari ematen zaio lehentasuna (doitasun handiagoa lortzen dute eta). Analizatzaile hedatuak erantzuna ematen ez duen kasuetan, WFST sistemaren erantzuna ematen da.

V.4 taulan konbinazioen bitartez lortutako emaitzak bildu dira, eta konparazioa errazteko, lehenengo errenkadan WFST sistema soilak lortutako emaitzak kopiatu dira. Hiru konbinazioetan hobetu da emaitza, gutxi bada ere, eta balio onenak doitasun altueneko analizatzailearekin konbinatuta lortu dira (*hed-lem2atz2*): 0,5 puntu inguru hobetzen dira parametro guztiak konbinazio horretan ($P$, $R$ eta $F$).

| | ***P*** | ***R*** | $\boldsymbol{F}_1$ |
|---|---|---|---|
| WFST *hitza-hitza* | 91,53 | 78,27 | 84,38 |
| *hed-lem1atz1* + WFST | 91,36 | 78,45 | 84,41 |
| *hed-lem1atz2* + WFST | 91,56 | 78,62 | 84,60 |
| *hed-lem2atz2* + WFST | **91,94** | **78,62** | **84,76** |

V.4 Taula: Ebaluazioaren emaitzak sistemen erantzunak konbinatuta. Analizatzaile hedatuen erantzuna WFST *hitza-hitza* sistemaren erantzunarekin konbinatu da.

## V.5 WFST sistemak: emaitzen analisia

Kapitulu honen V.3 atalean ikusi dugunez, aldaeren informazio morfologikoa erabiltzeak ikasteko unean (*analisia-analisia*) ez ditu emaitza hobeak lortu normalizazio-atazan. Horrez gain, V.2 taulan laburbiltzen diren emaitzak analizatzen baditugu ikuspegi kuantitatibotik, badirudi normalizazioa egiteko sistema onena dela guztietatik sinpleena dena, hots, ikasteko *hitza-hitza* informazioa erabiltzen duen sistema: hori izan da emaitza onenak lortu dituena (bai hitz estandarrak kontuan hartuta, bai hartu gabe).

Ondorio hori, ordea, ezin da ondorio orokor gisa hartu, sistemen arteko konparazioa egiten ari baikara corpus bakar batean gertatutakoa ikusita, *Gero* corpusean, hain zuzen. Ezin dugu ziurtatu beste corpus batzuetan ere halakoa izango denik sistemen portaera. Izan ere, IV. kapituluan ikusi dugu *Peru Abarka* corpusean lortutako emaitza onenak ez direla *hitza-hitza* informazioarekin lortu (ikus IV.29 taula), *hitza-analisia* informazioarekin baizik. Horrek agerian uzten du ezin dugula baztertu morfologiak izan dezakeen ekarpena normalizazioaren ataza ebazteko.

Gure ustez, aukera dago kalitate handiagoko *analisia-analisia* motako informazioa lortzeko, bai sistema berriak probatuz (bibliografian aipatzen dira erdi-gainbegiratutako sistemak), bai Morfessor bera hobeto doituz. Etorkizunean gehiago landu beharreko bidea dela uste dugu.

Halere, hiru sistemen emaitzen analisiarekin jarraituz, analisi kualitatibo bat egin dugu osagarriak ote diren ikusteko. Aztertzen badugu zer den sistema bakoitzak ongi normalizatu duena, hiru sistemetan aurkitzen dugu adibideren bat, zein sistema horrek bakarrik ebatzi duen ongi eta beste bi sistemek ez.

Ikus ditzagun adibide zehatzak argi uzteko esandakoa. Ikasteko hitz es-

tandarrak erabiltzen ez duten hiru sistemen emaitzak analizatu ditugu eta hauxe ikusi dugu:

1. WFST *hitza-hitza* sistemak ongi normalizatu ditu honako aldaera hauek: *arintkiago*, *autsikizetik*, *baillezakete*, *bereganik*, *dathorrenean*, *etzedilla*, *fintkiago*, *lothu*, *zeikan* eta *zuetzaz*. Beste bi sistemek, ordea, ez.

2. WFST *hitza-analisia* sistemak ongi normalizatu ditu *baiteraku*, *erraxten*, *fariseoek*, *hilzaileak* eta *lekhukok* aldaerak, baina beste bi sistemek ez.

3. WFST *analisia-analisia* sistemak ongi normalizatu ditu *ezterauet*, *konsideratzeak* eta *malizia*, baina beste bi sistemek ez.

Beraz, hiru sistemek egiten diote ekarpenaren bat normalizazio-atazari, eta osagarriak izan litezke. Hori dela eta, hiruren erantzunak konbinatzen saiatu gara bozketa eginez. Hiru aukera izanik, bozketa-algoritmoa sinplea da: edozein bi sistemak erantzun bera ematen badute (erantzun "hutsa" izan daiteke), erantzun hori ematen da bozketan. Gainontzeko kasuetan, gure aukera izan da lehenengo sistemaren erantzuna ematea, zenbakiek adierazi baitute hori dela emaitza onenak lortu dituena. V.5 taulan bozketa eginez lortutako emaitzak ageri dira (aurretik sistema bakoitzak lortutakoak daude konparazioa errazteko). Ikusten denez, bozketa eginez lortzen den $F$ neurria hobea da, baina aldea ez da handia.

### 1. orakulua

Erantzunen osagarritasuna aztertuta bururatzen zaigun beste galdera bat da ea zenbatekoa izango litzatekeen hiru sistemekin lor litekeen emaitzarik onena. Balizko sistema horri *orakulua* esaten zaio, eta suposatzen da hiru sistemen erantzunen artean erantzun zuzena baldin badago, gai dela erantzun hori aukeratzeko. Ez daukagu algoritmorik aukeraketa hori egiteko, baina bai kalkula dezakegula zenbateko estaldura lortuko lukeen orakuluak corpus horretan: erantzun on guztiak biltzen dituenez, bere estaldura sistema bakoitzak lortutakoa baino handiagoa da. Orakuluaren doitasuna kalkulatzeko, berriz, erabaki beharra dago noiz geratuko den orakulua erantzuna eman gabe. Hau izan da erabakia: erantzun zuzena ez dagoenean erantzun posibleen artean, orakuluak ez du erantzunik ematen baldin eta sistema baten batek ez badu erantzunik eman.

Horrelako orakuluak lortuko lituzkeen emaitzak V.6 taulan jaso dira. Ikusten denez, estalduraren igoera 2,6–3,0 puntu ingurukoa da estaldura altuenarekiko, eta doitasuna ere 3,5 puntu inguru igotzen da.

| | ***P*** | ***R*** | $\boldsymbol{F_1}$ |
|---|---|---|---|
| Oinarri-lerroa | 94,87 | 39,22 | 55,50 |
| hitz anotatuak soilik | | | |
| WFST *hitza-hitza* | 91,53 | 78,27 | 84,38 |
| WFST *hitza-analisia* | 91,08 | 77,56 | 83,78 |
| WFST *analisia-analisia* | 90,68 | 75,62 | 82,47 |
| **Bozketa** | 91,94 | 78,62 | **84,76** |
| hitz anotatuak + estandarrak | | | |
| WFST *hitza-hitza* | 91,84 | 79,51 | 85,23 |
| WFST *hitza-analisia* | 90,91 | 77,74 | 83,81 |
| WFST *analisia-analisia* | 91,63 | 77,39 | 83,91 |
| **Bozketa** | 92,75 | 79,15 | **85,41** |

V.5 Taula: Hiru WFST sistemen erantzunak konbinatuta lortutako emaitzak. Konbinaziorako bozketaren algoritmoa erabili da.

| | ***P*** | ***R*** | $\boldsymbol{F_1}$ |
|---|---|---|---|
| Oinarri-lerroa | 94,87 | 39,22 | 55,50 |
| hitz anotatuak soilik | | | |
| WFST *hitza-hitza* | 91,53 | 78,27 | 84,38 |
| WFST *hitza-analisia* | 91,08 | 77,56 | 83,78 |
| WFST *analisia-analisia* | 90,68 | 75,62 | 82,47 |
| **Orakulua** | **94,85** | **81,27** | **87,54** |
| hitz anotatuak + estandarrak | | | |
| WFST *hitza-hitza* | 91,84 | 79,51 | 85,23 |
| WFST *hitza-analisia* | 90,91 | 77,74 | 83,81 |
| WFST *analisia-analisia* | 91,63 | 77,39 | 83,91 |
| **Orakulua** | **95,29** | **82,16** | **88,24** |

V.6 Taula: Hiru WFST sistemen erantzunen konbinaziorik onena: orakulua.

### 2. orakulua

Aurreko atalean orakuluak lortuko lituzkeen kalitate-parametroak kalkulatu ditugu, eta kalkulu horrek beste kalkulu berri bat egitera bultzatu gaitu. Jakin nahi duguna da zenbaterainoko emaitzak lortuko genituzkeen gai izanez gero transduktoreek sortzen dituzten proposamen posible guztien artean egokiena aukeratzeko.

Esperimentuetan zehar ikusi dugunez, proposatzen dugun normalizazio--sistemak erantzun bakar bat ematen du gehienez sarrera bakoitzeko (batzuetan ez du erantzunik ematen). Erantzun hori, oro har, aukeratu egin da hainbat proposamenaren artean, WFST transduktoreari 5 edo 20 erantzun eskatu baitzaizkio aldaera bakoitzari dagokion estandarra bilatzeko. Gero erantzunak iragazi dira estandarrak ez direnak kenduz, eta azkenean, erantzun bakar bat aukeratu da geratu direnen artean.

Azken aukeraketa horrek, agian, ez du asmatuko eta gerta daiteke erantzun zuzena galtzea urrats horretan. Kalkulatu dugu, beraz, zein izango litzatekeen balizko estaldura onena gai izanez gero WFST sistema bakoitzak proposatutako estandar guztien artean erantzun zuzena aukeratzeko (zuzenik balego). Hau da, orakulu berri honek hiru erantzun baino gehiago izan ahalko lituzke aukeran, sistema bakoitzak erantzun estandar bat baino gehiago sor lezakeelako. Modu horretan kalkulatutako estaldura izango litzateke WFST sistemekin lor litekeen maximoa. Jakin nahi dugu bada, noraino iristen den maximo hori, eta ea oso urruti geratu garen. Aurreko orakuluan bezala, erantzun egokia ez dagoen kasuetan orakulu berriak ez du erantzunik ematen baldin eta sistemaren batek ez badu ematen.

Sistemak egokitu ditugu, beraz, azken aukeraketa ez egiteko, eta erantzun posible guztiak bilduta orakulu onenaren emaitzak kalkulatu ditugu: V.7 taula. Ikusten denez, balizko orakulu onena ez da aurrekoa baino askoz hobea: estaldura puntu bat inguru igo da soilik.

Interesgarria da ikustea non geratu den gure sistema maximo horrekiko, eta analisi horretan estaldura soilik izango dugu kontuan. Gure hiru sistemen erantzunak bozketa bitartez konbinatuta lortu dugun estaldura handiena % 79,15 (edo % 78,62) izan da; orakulu onenaren estaldura maximoa, ordea, % 83,04 (edo % 82,16) da. Bien arteko diferentzia, beraz, 4,0 (edo 3,5) puntu ingurukoa da. Diferentzia horren garrantzia baloratzea zaila da, baina garrantzitsua da jakitea noraino irits liteke gure sistema sortu duen informazioa modurik eraginkorrenean kudeatuko bagenu.

| | ***P*** | ***R*** | $\boldsymbol{F}_1$ |
|---|---|---|---|
| Oinarri-lerroa | 94,87 | 39,22 | 55,50 |
| hitz anotatuak soilik | | | |
| WFST *hitza-hitza* | 91,53 | 78,27 | 84,38 |
| WFST *hitza-analisia* | 91,08 | 77,56 | 83,78 |
| WFST *analisia-analisia* | 90,68 | 75,62 | 82,47 |
| Orakulu onena | 95,48 | **82,16** | 88,32 |
| hitz anotatuak + estandarrak | | | |
| WFST *hitza-hitza* | 91,84 | 79,51 | 85,23 |
| WFST *hitza-analisia* | 90,91 | 77,74 | 83,81 |
| WFST *analisia-analisia* | 91,63 | 77,39 | 83,91 |
| Orakulu onena | 95,92 | **83,04** | 89,02 |

V.7 Taula: Balizko orakulu onenaren emaitzak. Sistema bakoitzak proposamen bat baino gehiago egin dezake (aurrekoan bakarra egiten zuten) eta horien guztien artean aukeratzen da onena.

### Atazaren zailtasunaz

Eraikitako WFST sistemen arteko osagarritasunaren azterketa egin dugu V.5 atalean eta bukatu baino lehen interesgarria iruditzen zaigu planteatu dugun normalizazio-atazaren zailtasuna nolabait neurtzea. Izan ere, ataza planteatu den moduan, aldaeraren normalizazioa bat etorri behar du zehatz-mehatz anotatutakoarekin normalizazio hori ontzat har dadin. Nolakoak dira normalizazio-sistemaren okerrak? Berezitasunen bat ikusten da horien azterketa eginez gero?

Ildo horretan, eraikitako hiru WFST sistemen okerreko erantzunak aztertuta ikusi dugu, esaterako, hirurek lotu dutela *konparazino* aldaera *konparaketa* estandarrarekin. Erantzun hori ez da ontzat hartu, ordea, testeko corpusean anotatutako hitza *konparazio* delako eta ez *konparaketa*. Idazkerari begira gertuago dago anotatutako hitza sistemek eman dutena baino, eta ez dago hori eztabaidatzerik. Hiru sistemek proposatutako estandarra, ordea, ontzat har liteke zenbait erabileratan.

Beste kasu batzuetan, sistemak ez dira gai izan ezta erantzun bat proposatzeko. Hala gertatu da adibidez *absolbazinoarekin* eta *asirioen* aldaeretan. Anotatutako estandarrak *absoluzioarekin* eta *asiriarren* dira, hurrenez hu-

rren, baina bi kasu horietan hiru WFST sistemak erantzunik eman gabe geratu dira, ez direlako gai izan hitz horiek estandar batekin normalizatzeko. Dirudienez, horietan egin behar diren aldaketak ez dira ikasi aurretik.

Aurreko adibideen ildotik bi hausnarketa egiten ditugu:

- Ikasteko datuen koherentzia eta zehaztasuna ziurtatzea oso garrantzitsua da sistemek ahalik eta ondoen ikas dezaten eta gerora ematen dituzten erantzunak koherenteak izan daitezen.

- Ikasteko datuak mugatuak direnez, zaila da aldaeretan gerta daitezkeen fenomeno guztiak ikasteko datuetan islatuta egotea, eta, beraz, ezin izango da guztia normalizatu: beti egongo dira automatikoki normalizatu ezingo diren kasuak.

## V.6 Ondorioak

Kapitulu honetan saiatu gara informazio morfologiko berria erabiltzen normalizazioaren ataza ebazteko, eta saiakera hori bi bidetatik egiten ahalegindu gara.

Lehenengo bidetik aldaeren segmentazio morfologikoa lortzen saiatu gara modu ez-gainbegiratuan. Informazio hori lortzeko, beharrezkoa izan da lehendabizi euskara estandarraren morfologia erauztea modu ez-gainbegiratuan, eta lortutako erauzketa hori ebaluatzea. Guk dakigula, lehenengo aldia da horrelako lana egiten dela euskararekin.

Aldaeren analisia lortu ondoren, saio berri bat egin ahal izan dugu normalizazio-ataza ebazteko, *analisia-analisia* motako informazioa erabiliz ikasteko prozesuan. *Gero* corpusarekin lortutako emaitzak ez dira hobetu, nahiz eta aurreko emaitzetatik gertu egon. Gure ustez, morfema-mugen informazioa baliagarria izan daiteke normalizazioan egin behar diren aldaketak identifikatzeko, baina lortutako analisiek, agian, ez dute kalitate nahikoa izan atazan laguntzeko. Ezin dugu modu automatikoan egiaztatu hala den, ezin baitugu neurtu analisi horien kalitatea erreferentzia batekin konparatuz. Bidea ikusten dugu, dena den, lanean jarraitzeko analisi zehatzagoen bila, bai eskura ditugun tresnak egokituz, bai bibliografian aipatzen direnak probatuz. Ildo horretatik interesgarria dirudi morfologia ikasteko modu gainbegiratua eta ez-gainbegiratua nolabait konbinatzea emaitzak hobetzeko. Egungo bibliografian badira zenbait lan hizkuntzen paradigmak modu erdi-gainbegiratuan ikasteko (Ahlberg *et al.*, 2014), eta aipatutako konbinazioa etorkizunean landu beharreko beste bide bat izan daiteke.

Morfologiari etekina ateratzeko bigarren bidea, eskuz anotatutako informazioa baliatzea izan da aldaera eta estandarraren arteko baliokidetza morfologikoak lortzeko. Nahiz eta eskuz anotatutako informazioa erabili, baliokidetza horiek modu automatikoak lortu dira, eta horiekin, gero, euskara estandarraren analizatzailea hedatu dugu. Hedapen horretan aldaera dagokion informazioa estandarrarekin lotzen da, eta horrek aukera ematen du aldaerako hitz bat onartzeko loturak adierazten duen estandar batekin nolabait "baliokidetuz". Hedapenaren bidetik lortu diren emaitzak interesgarriak izan dira, baina, oro har, WFST metodoak lortutako emaitzetatik urruti geratu dira.

Morfologiak egin dezakeen ekarpena bi bidetatik egin bada ere, *Gero* corpusean soilik ebaluatu da, eta beharrezkoa iruditzen zaigu corpus gehiagotan egiaztatzea zer nolako hobekuntzak lor ditzakeen informazio morfologiko horrek. Esperimentuetan zehar ikusi dugu corpusek oso ezaugarri ezberdinak izan ditzaketela, eta emaitzak alda daitezkeela nahiz eta metodo berdinak aplikatu.

Azkenik, orakuluetan lortutako emaitzek erakutsi dute sistemen arteko osagarritasuna dagoela. Agian, emaitza horietara hurbil gintezke hitzetako hizkuntza-eredu konplexuago bat erabiliz gero, bi edo hiru hitzetako hizkuntza-eredua alegia. Etorkizuneko lanetarako geratzen da aukera hori lantzea.

# VI. KAPITULUA

## Ondorioak, ekarpenak eta etorkizuneko lanak

### VI.1 Sarrera

Tesi-lan hau testu ez-estandarren prozesaketaren arloan kokatzen da. Interes handia sortu da azken urteetan hizkuntzaren prozesamenduko ohiko tresnak aplikatzeko testu ez-estandarretan (testu historikoetan zein sare sozialetako egungo testuetan), baina ikusi da tresna horien errendimendua asko jaisten dela horrelako testuekin erabiltzen direnean. Hori dela eta, testu horiek prozesatzea erronka berria bilakatu da hizkuntzaren prozesamenduaren arloan, eta erronka horri aurre egiteko planteatu diren irtenbideen artean, testu ez-estandarren normalizazioa dago.

Tesi-lan honek planteatu duen galdera nagusia izan da ea metodo bat ikas daitekeen, morfofonologia konputazionaleko tresnak erabiliz, euskarazko aldaerak normalizatzeko, hau da, euskarazko aldaerei —diakronikoei zein dialektalei— automatikoki esleitzeko dagozkien forma estandarrak.

Galdera nagusi horri erantzuteko, hiru metodo fonologiko probatu dira euskarazko corpus dialektal batean, eta emaitza onenak lortu dituena aukeratu da esperimentu berriak planteatzeko beste corpus batzuetan. Gero, aukeratutako metodoan sakondu ondoren, tesi-lan honetan zehar prestatu diren euskarazko bi corpus historikoetan aplikatu, egokitu eta ebaluatu da metodoa. Euskararekin lan egiteaz gain, gaztelaniarekin zein eslovenierarekin egin dira esperimentuak, metodoa hizkuntzarekiko independentea dela egiaztatzeko, eta lortutako emaitzak konparatzeko beste metodo batzuek lortzen dituztenekin. Azkenik, normalizazio-atazari lagundu nahian, aldaerei dagokien informazio morfologikoa lortzeko saioa egin da, informazio berri hori baliatzeko atazaren ebazpenean.

## VI.2 Ondorio nagusiak

Normalizazio-ataza ebazteko planteatu diren metodoei eta haien kalitateari dagokionez, tesi-lan honen III. kapituluan hiru metodo fonologiko proposatu dira ataza hori ebazteko. Hiru metodoak ikasketa automatikoko tekniketan oinarritu dira, hau da, metodoek ematen zaizkien datuetatik ikasten dute, gero ikasitakoa aplikatzeko datu berrien gainean. Metodo onena aukeratzeko, hirurak aplikatu dira lapurtera/estandarra corpus paralelo batean, eta emaitzetan nabarmen gailendu den metodoa transduktore haztatuen teknologian (WFST) oinarritutakoa izan da, Phonetisaurus aplikazioaren bitartez inplementatu duguna. Beste bi metodoak —lexdiff eta ILP deitu ditugun metodoak, hurrenez hurren— transduktore arruntetan oinarritu dira: ikasteko datuetatik erregela fonologikoak inferitu dira (metodo bakoitzak bere modura), foma aplikazioaren bitartez konpilatu dira erregela horiek transduktore arruntetan, eta azkenik, aldaera berriei aplikatu zaizkie erregela horiek transduktoreen bitartez. Erregelak sortzeko irizpideak zein horiek aplikatzeko biderik egokiena topatzea, ordea, ez da erraza suertatu metodo horietan, eta lortutako emaitzak mugatuak izan dira. Arazo hori saihestu du, neurri handi batean, WFST teknologian oinarritutako metodoak. Metodo hori gai da, ikasitakoaren arabera, ematen dituen erantzunak ordenatzeko, eta informazio hori baliatuta askoz emaitza hobeak lortu dira.

III. kapituluko III.11 taulan, corpus dialektalarekin ebaluazioan lortutako emaitzen laburpena ematen da. Emaitzetatik ateratzen den ondorioa argia da. WFST teknologian oinarritutako metodoa egokiena da normalizazioren ataza ebazteko, balio altuenak lortzen dituen metodoa baita ebaluatu diren hiru parametroetan, doitasuna, estaldura eta $F$ neurria. Metodoaren arrakasta, batez ere, estalduran lortutako balioan dago, beste bi metodoak baino 16 puntu inguruko altuagoa den estaldura lortzen baitu.

Corpus dialektalean lortutako emaitzak ikusita, berretsi nahi izan dugu WFST metodoa egokia den testu historikoetan ageri diren aldaerak normalizatzeko eta hori izan da IV. kapituluaren helburu nagusia. Lehenik eta behin, euskarazko bi corpus historikotan planteatu dira esperimentuak, baina esperimentu horiek bideratu ahal izateko, beharrezkoa izan da lehendabizi corpus historiko horiek prestatzea (II. kapituluan deskribatu da prestaketa-lan hori).

Hala, tesi-lan honetan bi corpus prestatu dira euskal literaturako bi klasikorekin: *Gero* (Pedro Agerre *Axular*) eta *Peru Abarka* (J. A. Mogel) obrekin, hain zuzen. Bi obra horietako corpusak prestatzeko, obra bakoitzean bi zati aukeratu dira zoriz, eta zati horietan ageri diren aldaerak anotatu dira eskuz. Zati bat ikasteko erabili da (eta doikuntzako hainbat esperimentu egiteko)

eta bestea metodoa ebaluatzeko. Ikasteko zatian eskuz anotatutako aldaera kopurua, gutxi gora behera, 1.000 izan da, eta testeko zatian kopuru hori 500 ingurukoa izan da bi corpusetan.

Ebaluazioaren emaitzei dagokienez, bi corpus horietan lortu diren emaitzak ez dira berdin-berdinak izan. Alde batetik, lortutako balioak ezberdinak izan dira: *Gero* corpusean egindako ebaluazioan WFST metodoak lortu duen $F$ neurriaren balio onena % 85 ingurukoa izan da (oinarri-lerroak % 55, ikus IV.21 taula); *Peru Abarka* corpusean, berriz, % 77 ingurukoa izan da (oinarri-lerroak % 46, ikus IV.29 taula). Halaber, emaitza horiek ez dira lortu ikasteko informazio bera erabiliz bi kasuetan: *Gero* corpusean ikasketan *hitza-hitza* informazioa erabiliz lortu da emaitza hori, eta *Peru Abarka* corpusean *hitza-analisia* erabiliz.

Modu batera edo bestera, WFST metodoa aplikatuz lortu diren emaitzak normalizazio-atazan, oinarri-lerroak lortutakoak baino 30 puntu handiagoak izan dira $F$ neurriaren ikuspuntutik. Gainera, doitasuna estaldura baino handiagoa izan da bi corpusetan (10–12 puntu handiagoa), eta hori oso ezaugarri garrantzitsua izan daiteke normalizazioa automatikoki egitea planteatzen bada (liburutegi digitalen testuinguruan, esaterako): hobe normalizatu gabe uztea, gaizki normalizatzea baino.

Emaitzen arteko aldea bi corpus horietan, beraz, 8 puntu ingurukoa izan da, baina hasieratik izan da horrela, baita oinarri-lerroan ere. Diferentzia hori corpusen ezaugarriekin lotuta dago, oso ezaugarri ezberdinak dituzten obrak baitira (euskalkia, gaia..., ikus bigarren kapituluko II.1 eta II.2 taulak), eta *Peru Abarka*, *Gero* baino askoz urrutiago dago gaur egungo estandarretik. Dirudienez, eta logikoa denez, horrek eragin zuzena du lor daitezkeen emaitzen kalitatean.

Egindako esperimentuen emaitzetan oinarrituta, tesi-lan honen hasierako galderari erantzun diezaiokegu. Gogora dezagun lanaren funtsa izan dela aztertzea ea ikas daitekeen metodo bat euskarazko aldaerei —diakronikoei zein dialektalei— automatikoki esleitzeko dagokien forma estandarra. Euskarazko corpusetan lortu diren emaitzetatik ondorioztatzen da baietz, ikas daitekeela metodo bat ataza hori automatikoki ebazteko, baina kontuan izan behar dela metodoak ezin izango duela aldaera guztiak normalizatu, ikasteko datuak mugatuak direlako.

Beste ondorio garrantzitsua da ez dela behar oso corpus handia ikasteko, beste normalizazio-sistema batzuen pare geratzeko: eskuz anotatutako 1.000 inguru aldaera erabiliz lortu diren emaitzak kalitate onekoak izan dira.

Euskarazko corpusetan aplikatu, egokitu eta ebaluatu dugun normalizazio-metodoa, WFST metodoa, hain zuzen, gaztelaniarekin zein eslovenierarekin aplikatzeko aukera izan dugu hainbat ikerlariren kolaborazioari esker.

Gaztelaniaren kasuan, bi datu-multzorekin egin dira esperimentuak, FL-EM eta IMPACT multzoekin, eta normalizazio-atazan lortutako emaitzak oso onak izan dira bietan: $F$ neurriak % 90 baino altuagoa izan da, oro har, bi kasuetan (ikus IV.30 eta IV.31 taulak). Emaitza horietatik ateratzen den lehenengo ondorioa da proposatzen dugun metodoa hizkuntzarekiko independentea dela, emaitza onak lortu baititu doikuntza berezirik egin behar izan gabe hizkuntzaren arabera. Horrez gain, emaitza horiek berresten dute planteatu dugun metodologia egokia dela: eskuz anotatutako 1.000 forma ingururekin ikastea, nahikoa da emaitza onak lortzeko normalizazioan, bi multzo horietan % 90etik gorako $F$ neurria lortu baita.

Gaztelaniaren kasuan, gure metodoak lortutako emaitzak alderatu nahi izan ditugu beste sistema batzuek lortutakoekin. Alderaketa hori, ordea, kontuz egin behar da sistemek planteatzen duten ataza eta hori ebaluatzeko datuak ez badira berberak. Argi utzi dugu hori IV.6.1.1 ataleko ondorioetan, gure sistemak lortutako emaitzak alderatzen saiatu garenean Porta *et al.* (2013) lanean ematen direnekin. Lan horretan proposatzen den sistema erregeletan oinarritzen da, eta ebazten duen ataza ez da zuzenean normalizazioarena, aldaeren analisiarena baizik. Ataza horretan lortzen duten estaldura gure sistemak lortutakoaren oso antzekoa da; doitasuna, aldiz, nabarmen baxuagoa da (25 puntu baino gehiago). Ondorioa da, beraz, gure WFST metodoaren emaitzak erregela-sistema horrek lortutakoa gainditzen duela, hein batean behintzat.

Eslovenierarekin ere esperimentuak egiteko aukera izan dugu Scherrer eta Erjavec ikerlarien lanari esker (Scherrer eta Erjavec, 2015). Ikerlari horiek proposatzen duten sistema ikasketa automatikoko tekniketan oinarritzen da, gurea bezala, eta karaktere-mailako itzulpen automatiko estatistikoaren metodoak erabiltzen ditu (CSMT). Sistema hori eslovenierazko hitz historikoak modernizatzeko atazan erabili dute aipatutako lanean, eta emaitzak argitaratzearekin batera, ikerlari ororen eskura jarri dituzte esperimentuetan erabili dituzten datu guztiak. Hori horrela, aukera paregabea izan dugu haiek egindako esperimentuak errepikatzeko gure metodoa aplikatuz, eta ondorioz, emaitzen konparaketa bidezkoa izan da kasu honetan. IV. kapituluko IV.35 taulan bi metodoek lortutako emaitzak ageri dira, ondorioa izanik bi sistemak parean geratzen direla. Hori oso positiboa da kontuan hartzen bada gure sistema ez dugula batere doitu eslovenierari aplikatzeko, eta gainera, artearen egoeraren adierazgarria den sistema batekin konparatzen ari garela, haien sistema 2015ekoa baita.

Beraz, gaztelaniarekin eta eslovenierarekin lortutako emaitzetatik ateratzen diren ondorio nagusiak bi dira: (1) tesi-lan honetan proposatzen dugun WFST metodoa normalizazioaren ataza ebazteko hizkuntzarekiko indepen-

dentea da; (2) metodoak lortzen dituen emaitzak parekoak dira artearen egoeran dauden beste sistemekin konparatuta.

Euskarazko bi corpusekin egin diren esperimentuetan, IV. kapituluan, hitz estandarren informazio morfologikoa erabiltzen saiatu gara normalizazio-ataza hobeto ebazteko, gure hipotesia izan baita informazio hori lagungarria izan daitekeela egin behar diren aldaketen testuingurua hobeto zehazteko. Hala, *hitza-analisia* motako informazioa erabili dugu ikasteko unean, baina bide horretatik lortutako emaitzak gazi-gozoak izan dira. Ezin da esan informazio morfologikoa lagundu ez duenik, izan ere, *Peru Abarka* corpuseko emaitza onenak informazio hori baliatuta lortu dira, lehen aipatu dugun moduan. *Gero* corpusean, aldiz, ez da horrela izan, eta horregatik diogu emaitza ez dela izan espero bezain ona.

Morfologiaren bidetik, eta urrats bat aurrera eman nahian, V. kapituluan bide berri bat jorratu dugu aldaeren informazio morfologikoa lortzeko. Bi estrategia ezberdin probatu dira informazio morfologiko berria lortzeko.

Lehenengo estrategiak aldaeren segmentazioa lortzea izan du helburu. Horretarako Morfessor izeneko tresna ez-gainbegiratua erabili da, eta horrekin lortutako segmentazioak erabili dira informazio berri gisa WFST metodoa aplikatzeko. Ebaluazioa *Gero* corpusean soilik egin bada ere, lortutako emaitzak ez dira ia aldatu (ikus V.2 taula).

Bigarren estrategiak eskuz anotatutako informazioa baliatu du aldaera eta estandarraren arteko baliokidetza morfologikoak automatikoki inferitzeko (morfemen arteko baliokidetzak). Baliokidetza horiekin euskara estandarra "hedatu" da (aldaera eta estandarraren arteko lotura eginez), eta horrek aukera berri bat zabaldu du normalizazioaren ataza ebazteko. Analizatzaile hedatuen bitartez lortutako emaitzak interesgarriak izan dira (ikus V.3 taula), eta emaitzak apur bat hobetu dituzte oinarrizko WFST sistemarekin konbinatuta (ikus V.4 taula).

Morfologiaren ildotik egin diren esperimentuen ondorioa da etorkizunean gehiago landu beharreko bidea dela. Informazio morfologikoak egin dezakeen ekarpena gai interesgarria da, dudarik gabe, eta aukerak daude bide horretan gehiago ikertzeko normalizazioaren atazari laguntzeko.

## VI.3 Ekarpenak

Tesi-lan honen garapenaren bitartez hainbat ekarpen egin dira testu ez-estandarren normalizazioaren inguruan:

- Euskarazko corpus historiko anotatu bat eraiki dugu normalizazio-atazaren inguruko esperimentuak bideratzeko, eta corpus hori lagungarria

izan daiteke etorkizunean antzeko ataza planteatzen bada beste obra batzuetan.

- Metodologia erdi-gainbegiratu bat proposatzen dugu testu historikoak normalizatzeko (ondorengo azpiatalean egiten da metodologiaren laburpena).
- Proposatzen dugun normalizazio-metodoaren egokitasuna frogatzeko gaztelaniazko eta eslovenierazko zenbait ikerlariren lankidetza izan dugu beharrezko datuak eskuratzeko. Erlazio hori probetxugarria izan daiteke etorkizuneko lanei begira.
- Eraiki dugun normalizazio-sistemak aplikazio esperimental bat izan dezake liburutegi digitalei begira. Izan ere, testu ez-estandarrak normalizatuz gero, aukera dago haien testua indexatzeko indize normalizatu berriekin ere, eta ez testu hutsarekin soilik. Modu horretan lor daiteke bilaketetan termino ez-estandarrak aurkitzea nahiz eta galdera hizkuntza estandarrean egin.
- Proposatutako normalizazio-metodoa erabilgarria da estandar ez den edozein testurekin, bai historikorekin bai bestelakorekin. Hala frogatzen du 2013ko SEPLN kongresuko TweetNorm tailerrean aurkeztutako sistemak (Alegria *et al.*, 2013).
- Anotatutako informaziotik abiatuta, euskarazko analizatzailea hedatu dugu modu erdi-gainbegiratuan, eta aldaerako zenbait morfema berri gehitu dizkiogu lexikoari, morfema estandarrekin lotura eginez. Hedapena mugatua izan da, anotatutako informazioa gutxi baita, baina aukera interesgarria izan daiteke anotatutako informazioa handitzen den heinean.
- Morfessor tresna ebaluatu dugu euskara estandarrarekin lehenengo aldiz. Tresna horrek modu ez-gainbegiratuan segmentatzen ditu testu-hitzak eta egin dituen segmentazioak ebaluatu ditugu analizatzaile estandarra erreferentzia gisa hartuta. Ebaluazioaren emaitzak finlandierarekin lortutakoen antzekoak izan dira.

### Testu historikoak normalizatzeko metodologia

Esan bezala, testu historikoak normalizatzeko metodologia zehatz bat proposatzen dugu tesi-lan honetan. Metodologia horren arabera eraikitzen den sistema WFST teknologian oinarritzen da eta Phonetisaurus tresnaren bitartez inplementatzen da. Hona hemen metodologiaren laburpena:

1. Testu historikoaren corpusa bildu testu elektroniko gisa. Urrats hau betetzeko aukerak ez dira beti berdinak eta estu lotzen zaizkio bai testu motari, bai garatutako baliabideei. Zenbait kasutan OCR teknikak erabil daitezke baina kontuan izanik testu historikoek ezaugarri bereziak dituztela, beharrezkoa izaten da ondorengo prozesaketa bat erroreak zuzentzeko[1] (Piotrowski, 2012).

2. Testua analizatu OOV hitzak detektatzeko, hots, hizkuntza estandarretik kanpo dauden hitzak detektatzeko. Urrats hori betetzeko beharrezkoa da zuzentzaile ortografiko bat edo egungo hizkuntza estaltzen duen hitz zerrenda bat.

3. Detektatutako OOV horien artean eskuz anotatu 1000 inguru OOV desberdin. Egin ditugun esperimentuetan Brat anotazio-tresna erabili dugu urrats horretan, eta tresna erosoa izan da, bai anotazioa egiteko, bai gero anotatutako informazioa erauzteko.

4. Anotatutako informazioa erabili bertatik ikasteko. Forma estandarrak ere (OOV ez direnak) sar daitezke ikasteko prozesuan, orokorrean hobekuntza txiki bat lortzen baita. Phonetisaurus tresnaren ikasteko urratsak betez gero, transduktore haztatu bat lortzen da ikasteko informaziotik abiatuta, gero hurrengo urratsetan erabiliko dena aldaera berriak normalizatzeko. Ikasketa-urrats horiek betetzeko, Phonetisaurus tresnaz gain $n$-gram eredua modelatzeko tresna bat behar da.

   Aukera ezberdinak daude ikasteko informazioa adierazteko. Sinpleena da hitzen arteko erlazioa adieraztea (*hitza-hitza* deitu duguna), baina ikusi dugu informazio morfologikoaren erabilpena lagungarria izan daitekeela. Alderdi horretatik bi aukera ikusi ditugu: hitz estandarraren segmentazio morfologikoa erabiltzea (*hitza-analisia* deitu dugun aukeran) edota segmentazio morfologikoa erabiltzea bi aldeetan (*analisia-analisia* deitu duguna). Azken kasu horretan, Morfessor tresna erabili dugu aldaerak segmentatzeko (corpus historikoaren testu osoa eman zaio Morfessorri aldaeraren segmentazio-eredua lortzeko).

5. Normalizazio-proposamenak lortu aldaera berrientzat Phonetisaurus tresna eta ikasketa-urratsetan eraikitako transduktorearen bitartez. Urrats hau transduktore bakoitzarekin egin behar da baldin eta transduktore bat baino gehiago eraiki badira informazio ezberdina erabili delako ikasteko (hiru aukera aipatu ditugu aurreko urratsean).

[1] Lan honetan ez dugu lehenengo urrats hau landu eta kontsideratu dugu testua jadanik digitalizatuta dagoela.

6. Normalizazio-proposamenak iragazi "onena" aukeratzeko. Berriro ere, urrats hau transduktore bakoitzak eman dituen proposamenekin egin behar da. Ondoren egin behar den erabilpenaren arabera, interesgarria izan daiteke aukeratzen den normalizazioa lematizatzea.

   Metodologia honetan proposatzen dugun iragazkia da estandarrak ez diren hautagaiak alboratzea eta geratzen direnen artean transduktoreak onena kontsideratzen duena aukeratzea.

7. Transduktore bakarra eraiki bada urrats hau ez da egin behar, baina bat baino gehiago eraiki badira, horien erantzunak konbinatzeko proposatzen dugun algoritmoa bozketa da.

8. Aldaeren normalizazioa lortuta, aukeran geratzen da informazioa berreskuratzeko sistema batean (IR) integratzea. Testu historikoa indexa daiteke forma historikoekin ez ezik, normalizatutako formekin edota lemekin ere.

## VI.4 Etorkizuneko lanak

Hauek dira etorkizun hurbilean aurreikusten ditugun lanak:

- Tesi-lan honetan planteatu den ataza ebazteko, funtsezkoa izan da Josef Novakek garatutako Phonetisaurus tresna. Hainbat urratsetan egiten du lan tresna horrek, eta urrats bakoitzean zenbait parametro erabili behar dira. 2015eko azken publikazioan aukera berri bat proposatzen du tresnaren egileak deskodeketa-urratsean sare neuronalak erabiltzeko, eta esperimentu berriak egin behar dira aukera hori aztertzeko.

- Antzeko normalizazio-ataza planteatu duten bi lanetan, Pettersson *et al.* (2014) eta Scherrer eta Erjavec (2015), itzulpen automatiko estatistikoaren teknikak erabili dituzte karaktere-mailan (CSMT), eta lortu dituzten emaitzak onak izan dira. Interesgarria izan daiteke teknika horiek aplikatzea euskararekin ikusteko zenbaterainoko aldea dagoen teknika hori eta tesi-lan honetan proposatutako teknikaren artean. Halere, ez dago alde handia bi tekniken artean; biek inplementatzen dute kanal zaratatsuaren eredua metodo estatistikoetan oinarriturik.

- Morfologiak egin dezakeen ekarpena testuak normalizatzeko ez da garbi geratu egin diren esperimentuetan, eta lan gehiago egin behar da

bide horretatik. Alde batetik, corpus bakar batean egin dira esperimentuak, eta aztertu beharra dago zer gertatzen den beste corpus batzuetan. Bestalde, aztertzeke dago nola konbinatu morfologia ikasteko modu ez-gainbegiratua eta gainbegiratua emaitzak hobetzeko, eta bide horretatik, bibliografian aipatzen diren metodo erdi-gainbegiratu berriek aukera berri bat zabaltzen digute morfologiaren ildotik lanean jarraitzeko.

- Normalizazioa hitz-mailan planteatu da tesi-lan honetan, hau da, hitza isolaturik kontsideratuta. Orakuluaren emaitzak lortu nahi badira sistemek proposatzen dituzten aukeren artean egokiena aukeratzeko, beharrezkoa ikusten dugu beste maila batzuetako informazioa kontuan izatea. Bi edo hiru hitzetako hizkuntza-eredua aplikatzea lagungarria izan daiteke hautagaien arteko aukeraketa egokiena egiteko, eta landu beharreko beste aukera da hori.

- Lortu ditugun emaitzek adierazten dute normalizazio-sistemaren erabilpena bideragarria dela aplikazio erreal batean, esaterako, liburu digitaletan inplementatzeko informazio-berreskuratze (IR) moduko aplikazio "zabalagoak". Etorkizunerako geratzen da gure sistema integratzea halako aplikazio erreal batean.

# Bibliografia


Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., de Ilarraza D.A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15, 2006.

Ahlberg M., Forsberg M., eta Hulden M. Semi-supervised learning of morphological paradigms and lexicons. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 569–578, 2014.

Alegria I., Etxeberria I., Ezeiza N., eta Maritxalar M. Morfología de estados finitos en software libre: aplicación al euskera. Finite state morphology using free software: application to Basque. *Procesamiento del lenguaje natural*, 43:359–360, 2009a.

Alegria I., Aranbarri G., Ceberio K., Labaka G., Laskurain B., eta Urizar R. A Morphological Processor Based on Foma for Biscayan (a Basque dialect). *LREC*, 2010.

Alegria I., Etxeberria I., Hulden M., eta Maritxalar M. Porting basque morphological grammars to foma, an open-source tool. *Finite-State Methods and Natural Language Processing*, 105–113. Springer, 2009b.

Alegria I., Etxeberria I., eta Labaka G. Una cascada de transductores simples para normalizar tweets. *Tweet-Norm@ SEPLN*, 15–19, 2013.

Allauzen C., Riley M., Schalkwyk J., Skut W., eta Mohri M. Openfst: A general and efficient weighted finite-state transducer library. *Implementation and Application of Automata*, 11–23. Springer, 2007.

Almeida J.J., Santos A., eta Simoes A. Bigorna–a toolkit for orthography migration challenges. *Seventh International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta*, 2010.

Bartz H.W., Burch T., Christmann R., Gärtner K., Hildenbrandt V., Schares T., eta Wegge K. Der Digitale Grimm. Deutsches Wörterbuch von Jacob und Wilhelm Grimm, 2004.

Beesley K.R. eta Karttunen L. *Finite-State Morphology.* CSLI Publications, 2003.

Bhapkar V.P. A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association*, 61 (313):228–235, 1966.

Bilbao Telletxea G. eta Gómez López R. Textos antiguos vascos en Internet. *Janus: estudios sobre el Siglo de Oro*, 1:111–121, 2014.

Bimbot F., Pieraccini R., Levin E., eta Atal B. Variable-length sequence modeling: Multigrams. *Signal Processing Letters, IEEE*, 2(6):111–113, 1995.

Bollmann M., Petran F., eta Dipper S. Rule-based normalization of historical texts. *Proceedings of the International Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 34–42, 2011.

Carreras X., Chao I., Padró L., eta Padró M. FreeLing: An Open-Source Suite of Language Analyzers. *LREC*, 2004.

Chen I.F., Ni C., Lim B.P., Chen N.F., eta Lee C.H. A keyword-aware grammar framework for lvcsr-based spoken keyword search. *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 5196–5200. IEEE, 2015.

Chen I.F., Ni C., Lim B.P., Chen N.F., eta Lee C.H. A keyword-aware language modeling approach to spoken keyword search. *Journal of Signal Processing Systems*, 82(2):197–206, 2016.

Creutz M. eta Lagus K. Induction of a simple morphology for highly-inflecting languages. *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, 43–51. Association for Computational Linguistics, 2004.

Creutz M. eta Lagus K. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology, 2005.

Deligne S., Yvon F., eta Bimbot F. Variable-length sequence matching for phonetic transcription using joint multigrams. *Fourth European Conference on Speech Communication and Technology*, 1995.

Dempster A., Laird N., eta Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Eger S. Designing and comparing g2p-type lemmatizers for a morphology-rich language. *Systems and Frameworks for Computational Morphology*, 27–40. Springer, 2015.

Erjavec T. The IMP historical Slovene language resources. *Language Resources and Evaluation*, 49(3):753–775, 2015.

Eskander R., Habash N., Rambow O., eta Tomeh N. Processing Spontaneous Orthography. *HLT-NAACL*, 585–595, 2013.

Etxeberria I., Alegria I., Hulden M., eta Uria L. Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural*, 52:13–20, 2014.

Etxeberria I., Alegria I., eta Leturia I. Ortografia-erroreak eta konpetentzia-erroreak Webeko euskarazko testuetan. *EKAIA Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria*, 24:219–236, 2011.

Etxeberria I., Alegria I., Uria L., eta Hulden M. Evaluating the Noisy Channel Model for the Normalization of Historical Texts: Basque, Spanish and Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1064–1069, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Goldsmith J. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198, 2001.

Goldsmith J. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(04):353–371, 2006.

Graehl J. Carmel finite-state toolkit. *ISI/USC*, 1997.

Hammarström H. eta Borin L. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350, 2011.

Hulden M. Foma: a finite-state compiler and library. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, 29–32, Athens, Greece, 2009. Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=1609057`.

Hulden M. Finite-State Technology. *Oxford Handbook of Computational Linguistics, 2nd ed.* Ruslan Mitkov (ed.). Oxford University Press., in print.

Hulden M., Alegria I., Etxeberria I., eta Maritxalar M. Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, 39–48. Association for Computational Linguistics, 2011.

Jiampojamarn S. eta Kondrak G. Letter-phoneme alignment: An exploration. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 780–788. Association for Computational Linguistics, 2010.

Jiampojamarn S., Kondrak G., eta Sherif T. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. *HLT-NAACL*, 7 lib., 372–379, 2007.

Johnson C.D. *Formal aspects of phonological description*, 3 lib. Mouton, The Hague, 1972.

Jurish B. Comparing canonicalizations of historical German text. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 72–77. Association for Computational Linguistics, 2010.

Kaplan R.M. eta Kay M. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378, 1994.

Kestemont M., Daelemans W., eta Pauw G.D. Weigh your words—memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301, 2010. ISSN 0268-1145.

Knight K. eta May J. Applications of weighted automata in natural language processing. *Handbook of Weighted Automata*, 571–596. Springer, 2009.

Koskenniemi K. An informal discovery procedure for two-level rules. *Journal of Language Modelling*, 1(1):155–188, 2013.

Leidig S., Schlippe T., eta Schultz T. Automatic detection of anglicisms for the pronunciation dictionary generation: A case study on our german it corpus. *Spoken Language Technologies for Under-Resourced Languages*, 2014.

Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10 lib., 707–710, 1966.

Mann G.S. eta Yarowsky D. Multipath translation lexicon induction via bridge languages. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, 1–8. Association for Computational Linguistics, 2001. URL `http://dx.doi.org/10.3115/1073336.1073356`.

Mohri M. Weighted automata algorithms. *Handbook of weighted automata*, 213–254. Springer, 2009.

Mohri M., Pereira F., Riley M., eta Allauzen C. At&t fsm library-finite state machine library. *AT&T Labs-Research*, 1997.

Muggleton S. eta De Raedt L. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

Nguyen D., Doğruöz A.S., Rosé C.P., eta de Jong F. Computational Sociolinguistics: A Survey. *arXiv preprint arXiv:1508.07544*, 2015.

Novak J.R., Minematsu N., eta Hirose K. WFST-Based Grapheme-to-Phoneme Conversion: Open Source tools for Alignment, Model-Building and Decoding. *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, 45–49, Donostia–San Sebastian, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-6208`.

Novak J.R., Minematsu N., eta Hirose K. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 1–32, 2015.

Pang B. eta Lee L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Pettersson E., Megyesi B., eta Nivre J. Rule-Based Normalisation of Historical Text–a Diachronic Study. *LThist 2012–First International Workshop on Language Technology for Historical Text (s), 11th Conference on Natural Language Processing (KONVENS 2012), September 19-21, 2012, Vienna, Austria*, 333–341. Österreichische Gesellschaft für Artificial Intelligence (ÖGAI), 2012.

Pettersson E., Megyesi B., eta Nivre J. Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting. *NODALIDA*, 163–179, 2013a.

Pettersson E., Megyesi B., eta Nivre J. A multilingual evaluation of three spelling normalisation methods for historical text. *Proceedings of LaTeCH*, 32–41, 2014.

Pettersson E., Megyesi B., eta Tiedemann J. An SMT approach to automatic annotation of historical text. *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, 18 lib., 54–69, 2013b.

Pettersson E. eta Nivre J. Automatic verb extraction from historical Swedish texts. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 87–95. Association for Computational Linguistics, 2011.

Piotrowski M. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157, 2012.

Porta J., Sancho J.L., eta Gómez J. Edit transducers for spelling variation in Old Spanish. *Proc. of the workshop on computational historical linguistics at NODALIDA 2013. NEALT Proc. Series*, 18 lib., 70–79, 2013.

Rajan V. Konkanverter-A Finite State Transducer based Statistical Machine Transliteration Engine for the Konkani Language. *Proc. 4-th Workshop on South and Southeast Asian Natual Language Processing of COLING*, 2014.

Reynaert M. Non-interactive OCR post-correction for giga-scale digitization projects. *Computational Linguistics and Intelligent Text Processing*, 617–630. Springer, 2008.

Roark B., Sproat R., Allauzen C., Riley M., Sorensen J., eta Tai T. The OpenGrm open-source finite-state grammar software libraries. *Proceedings of the ACL 2012 System Demonstrations*, 61–66. Association for Computational Linguistics, 2012.

Ruokolainen T., Kohonen O., Sirts K., Grönroos S.A., Kurimo M., eta Virpioja S. A Comparative Study on Minimally Supervised Morphological Segmentation. *Computational Linguistics*, 2016.

Ruokolainen T., Kohonen O., Virpioja S., eta Kurimo M. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. *CoNLL*, 29–37, 2013.

Ruokolainen T., Kohonen O., Virpioja S., eta Kurimo M. Painless semi-supervised morphological segmentation using conditional random fields. *EACL*, 84–89, 2014.

Sánchez-Marco C., Boleda G., eta Padró L. Extending the tool, or how to annotate historical language varieties. *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, 1–9. Association for Computational Linguistics, 2011.

Sánchez-Martínez F., Martínez-Sempere I., Ivars-Ribes X., eta Carrasco R.C. An open diachronic corpus of historical Spanish. *Language resources and evaluation*, 47(4):1327–1342, 2013.

Scherrer Y. Adaptive string distance measures for bilingual dialect lexicon induction. *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, ACL '07, 55–60. Association for Computational Linguistics, 2007. URL `http://portal.acm.org/citation.cfm?id=1557835.1557847`.

Scherrer Y. eta Erjavec T. Modernising historical Slovene words. *Natural Language Engineering*, FirstView:1–25, 8 2015. ISSN 1469-8110. URL `http://journals.cambridge.org/article\_S1351324915000236`.

Schlippe T., Quaschningk W., eta Schultz T. Combining grapheme-to-phoneme converter outputs for enhanced pronunciation generation in low-resource scenarios. *The 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014), St. Petersburg, Russia*, 14–16, 2014.

Sirts K. eta Goldwater S. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266, 2013.

Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., eta Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107. Association for Computational Linguistics, 2012.

Su H., Hieronymus J., He Y., Fosler-Lussier E., eta Wegmann S. Syllable based keyword search: Transducing syllable lattices to word lattices. *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 489–494. IEEE, 2014.

Uria L. eta Etxepare R. BASYQUE: Aplicación para el estudio de la variación sintáctica. *Linguamática*, 3(1):35–44, 2011.

Uria L. eta Etxepare R. Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: BASYQUE aplikazioa. *Lapurdum. Euskal ikerketen aldizkaria— Revue d'études basques— Revista de estudios vascos— Basque studies review*, 16:117–135, 2012.

van Noord G. Textcat language guesser. *World Wide Web, http://odur. let. rug. nl/˜ vannoord/TextCat/. Downloaded*, 09–04, 2001.

Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

Informatika Fakultatea

# Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz

Eranskinak

**Izaskun Etxeberria Uztarrozek**

Informatikan Doktore titulua eskuratzeko aurkezturiko
TESI-TXOSTENA

Donostia, 2016ko ekaina

# A. ERANSKINA

## Brat anotazio-tresna erabiltzeko gida laburra

### A.1 Sarrera

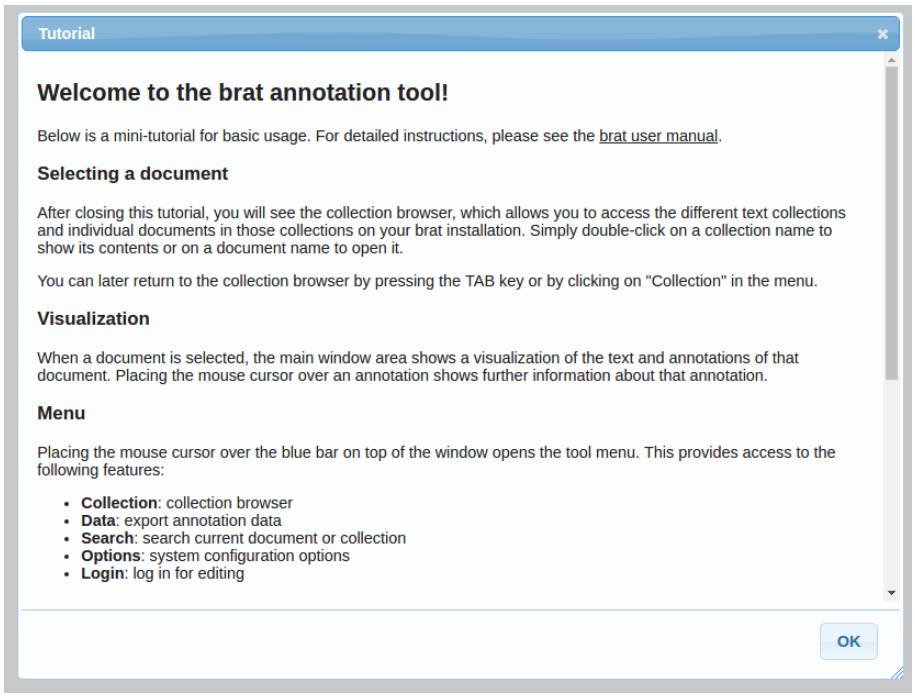
Eranskin honetan *Brat* anotazio-tresna erabiltzeko gida laburra aurkezten da, hori izan baita tesi-lan honetan sortu diren corpus historikoak etiketatzeko erabili den tresna. Nahiz eta bi obra etiketatu, eranskin honetan ageri diren erreferentzia guztiak etiketatutako lehen obrari buruzkoak dira, alegia, Axularren *Gero* obrari buruzkoak.

Brat tresna erosoa da, erabiltzeko erraza eta interfaze sinplekoa. Erabiltzeko unean sumatu den eragozpena ia bakarra izan da motel samarra suertatzen dela etiketatu behar den testua handi samarra bada. Hori dela eta, *Gero* obran zoriz aukeratu diren bi zatiak (testuaren %10 eta % 5 dituztenak) beste hainbat zatitan banatu dira Brat tresnarekin etiketatzeko. Zati handiena, % 10 duena, 4 fitxategitan banatu da, eta txikiena, % 5 duena, beste 2 fitxategitan[1]. Banaketa horiek eskuz egin dira paragrafoak kontuan hartuta (zorizko aukeraketan ere paragrafoa erabili da unitate gisa).

### A.2 Anotazio-tresna abiatzeko urratsak

Brat anotazio-tresna erabiltzeko, IXA taldeko *basajaun* makinara egin behar da konexioa Chrome nabigatzailea erabiliz. Horretarako, helbide egokia idatzi behar da nabigatzailearen barran eta aldaerekin lan egiteko prestatu den katalogoaren helbide zehatza `basajaun.si.ehu.es/brat_aldaerak` da.

[1] Lehenengo 4 fitxategiek Gero_10B_n izena dute, non $n$ hori 1, 2, 3 edo 4 den. Beste 2 fitxategiek Gero_10A_n izena dute, eta horietan $n$ren balioa 3 edo 4 da.

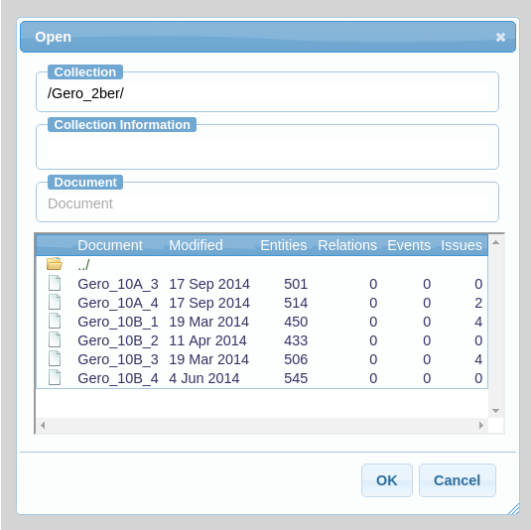



A.1 Irudia: Brat anotazio-tresnaren ongietorriko leihoa.

Behin hor kokatuta, ongietorriko leiho bat irekitzen da (ikus A.1 irudia), eta beheko aldean dagoen "OK" botoia sakatuz gero, katalogoan dagoen informazioa ageri da beste leiho batean (`brat_aldaerak` katalogoan). Leiho berrian katalogo zerrenda bat azalduko da eta *Gero* corpusari dagokion katalogoa "Gero_2ber" izenekoa da. Karpeta hori irekiz gero lehen aipatutako 6 fitxategien zerrenda ikusi ahal izango da (ikus A.2 irudia).

## A.3 Etiketatzea

*Gero* obrako zoriz aukeratutako testua daukagu, beraz, 6 fitxategitan banatuta eta egoki markatuta, etiketatzaileak bere lana egin ahal izan dezan. Horietako edozein fitxategi irekiz gero, A.3 irudian ageri den antzeko leihoa zabalduko da, non fitxategi horrek jasotzen duen testua ikusi ahal izango den. Testu horretan hainbat etiketa eta marka berezi ageriko dira, eta horiek argituko ditugu ondorengo lerroetan, etiketatzaileak informazio guztia izan dezan etiketatze-lanari ekin baino lehen.




A.2 Irudia: Brat tresna. Irekitako katalogoan dauden fitxategien zerrenda.

### A.3.1 Testuan ageri diren "markak"

Etiketatze-lana hasi baino lehen, honako marka berezi hauek aurkitu ahal izango dira testuan (ikus A.3 irudia):

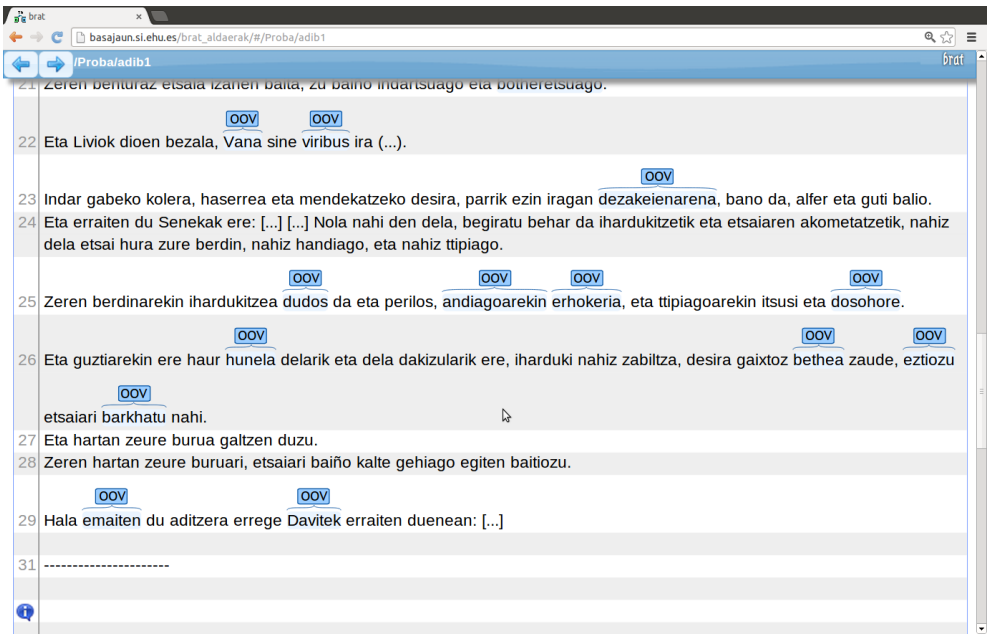
1. OOV etiketak.

   Testuko hainbat hitz etiketatuta daude jadanik OOV izeneko etiketa batez: *Out Of Vocabulary*. Izan ere, hitzak analizatu egin dira aurretik, eta estandarretik kanpo daudenei[2] etiketa hori jarri zaie abiapuntu gisa. Etiketatzailearen lana da OOV hitzei dagokien etiketa egokia jartzea aurrerago azalduko diren etiketen artean bat aukeratuta.

2. Marratxoz osatutako lerroak: - - - - - - - -.

   Jatorrizko testuko paragrafo-muga markatzen dute halako lerroek. *Gero* obraren hainbat paragrafo aukeratu dira zoriz etiketatu behar den testua osatzeko, eta horrek esan nahi du ez duela zertan izan lotura semantikorik testuan ageri diren paragrafoen artean, ez direlako ondoz ondoko paragrafoak jatorrizko testuan. Anotatzaileak garbi izan dezan

[2] II. kapituluko II.3.4 atalean esan den moduan, ohiko ez-estandarrez gain, beste bi kasu hauek ere ez-estandar gisa kontsideratzen dira : (1) forma hobetsia duten hitzak (h. gisa markatuak *Hiztegi Batua*n) eta (2) "RARE" motako analisi besterik ez duten hitzak.




A.3 Irudia: Etiketatu behar den testuaren hasierako itxura. Zenbait hitz etiketatuta OOV gisa eta testuan zehar beste bi marka ageri dira: (...) eta [...].

non bukatzen den jatorrizko paragrafo bakoitza, marratxoz osatutako lerroak gehitu dira muga horiek adierazteko.

3. Parentesien arteko puntuak: (...).

   Jatorrizko testuan erreferentzia asko daude parentesi artean: (Gen. 2), (Casian. lib. 18 cap. 14), (Exod. 5) eta abar. Horietan ageri diren izen asko ez-estandar gisa hartuko dira, baina horiek ez dira interesgarriak lan honetan. Beraz, zarata ez sortzeko etiketatzeko unean, parentesien arteko testua kendu egin da, eta horren ordez hiru puntuko marka jarri da.

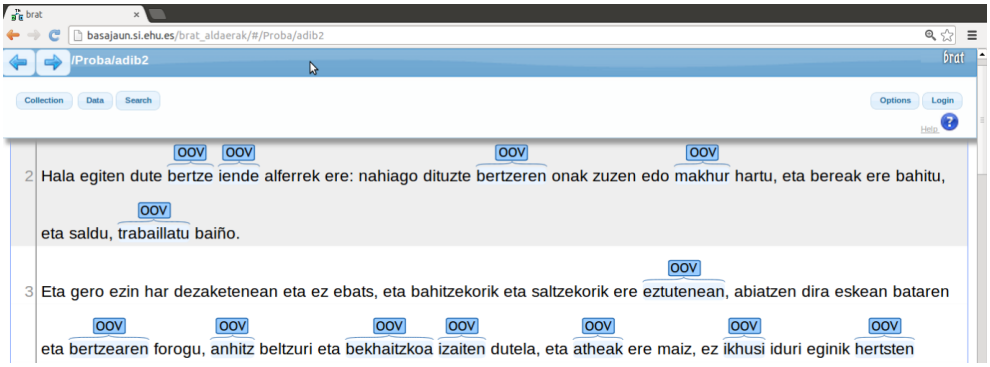
4. Kortxeteen arteko puntuak: [...].

   Jatorrizko testuan latinezko zita ugari ageri dira, eta logikoa denez latinez idatzitako hitz gehienek ez-estandarrak izango dira euskararen ikuspuntutik. Dena den, latinezko hitzak ez dira gure lanerako interesgarriak eta ez ditugu etiketatzerik nahi. Beraz, etiketatzeko testua ahalik eta garbiena izateko, latinez idatzitako esaldiak[3] automatikoki ezabatu dira testutik. Dena den, anotatzailearentzat interesgarria izan daitekeenez jakitea puntu jakin batean esaldi bat ezabatua izan dela, esaldia zegoen tokian kortxeteen arteko marka jarri da hori ohartarazteko.

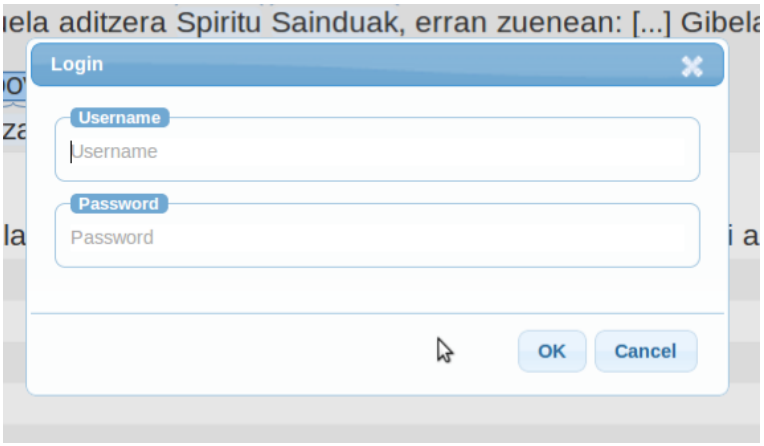
### A.3.2 Etiketak

Esan dugunez, etiketatzailearen lana da OOV gisa identifikatutako formak analizatzea eta egoki etiketatzea, gero hortik informazioa atera ahal izan dadin testuaren normalizaziorako. Testua etiketatu ahal izateko, hau da, anotazioak egiteko, beharrezkoa da Brat aplikazioan "sartzea" edo *login* egitea administratzaileak aurredefinitu duen erabiltzaile-izen eta pasahitza baten batekin (erabiltzaile bat baino gehiago egon daiteke). Aplikazioan sartzeko, beraz, fitxategi bat eduki behar da zabalik eta leihoaren goiko aldean ageri den barra urdinaren gainean jarri behar da sagua, horrela menu bat zabalduko baita zeinaren eskuineko aldean "Login" egiteko aukera ikusiko den (ikus A.4 irudia). Bertan sakatuta, erabiltzailea eta pasahitza eskatuko ditu aplikazioak (ikus A.5 irudia), eta horiek zuzen emanez gero, anotatzailea aldaketak egin ahal izango ditu testuan (etiketak jarri eta abar).

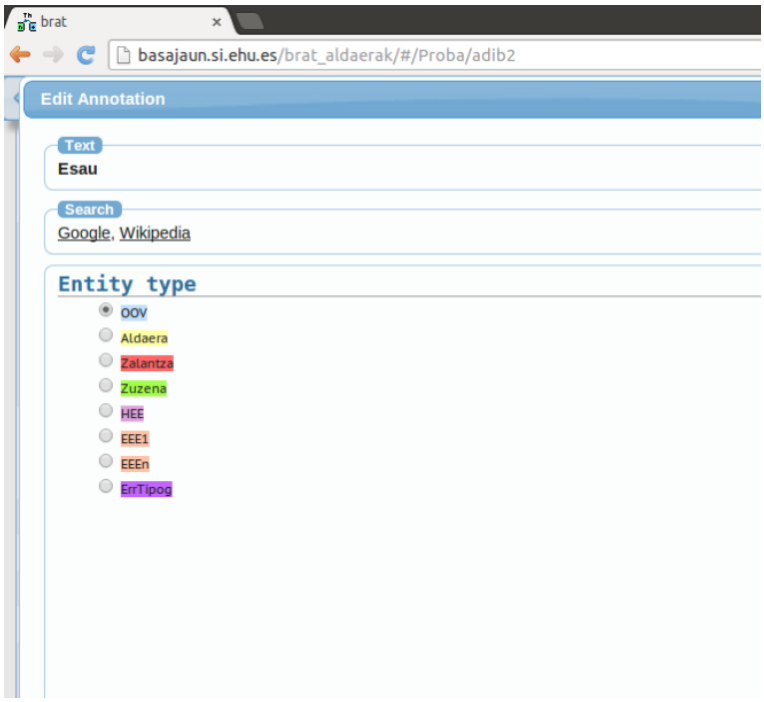
[3] Latinez idatzitako esaldiak detektatzeko textcat izeneko tresna erabili da.




A.4 Irudia: Brat aplikazioan sartzea beharrekoa da aldaketak egin ahal izateko. Goiko barra urdinaren gainean sakatuz gero, aukera batzuk zabaltzen dira eta eskuin aldean **"Login"** egiteko aukera ageri da.




A.5 Irudia: Erabiltzailea eta pasahitza eskatzen dituen pantaila.

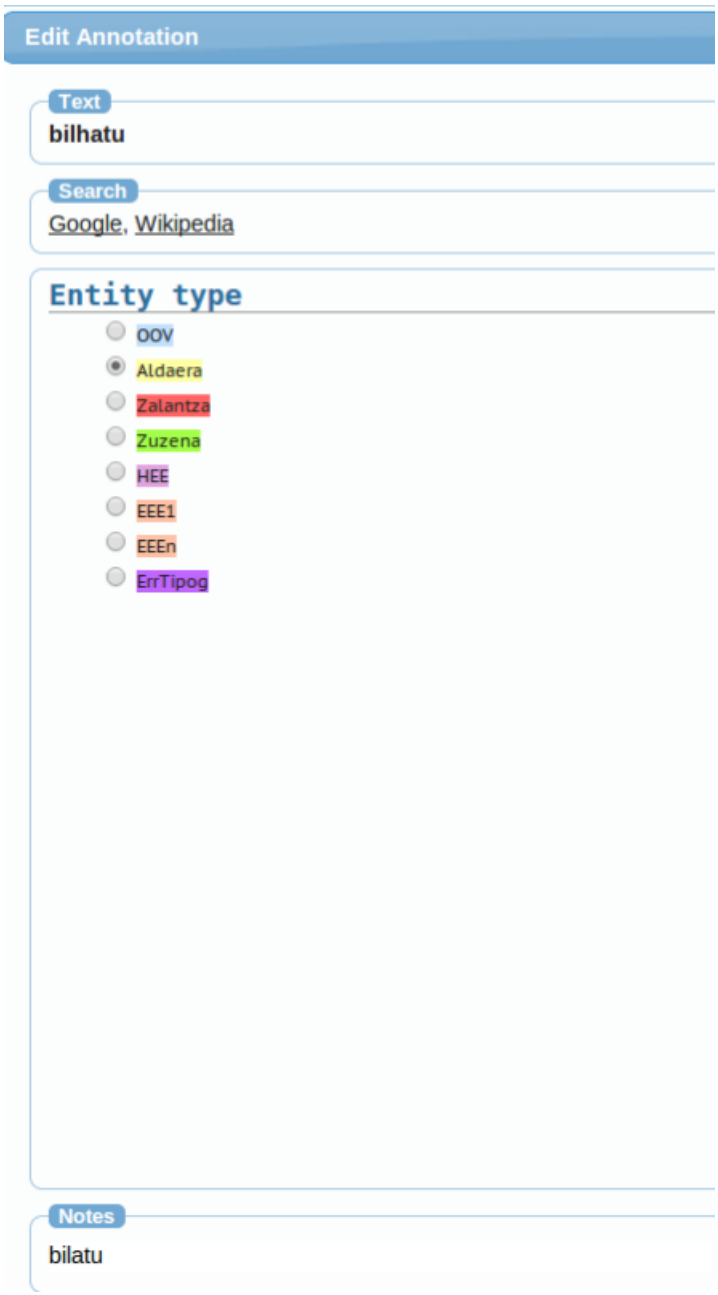



A.6 Irudia: *Gero* obraren corpusa etiketatzeko aurredefinitu diren etiketak.

Hitz bati OOV etiketa aldatzeko, OOV etiketaren gainean jarri behar da sagua eta bi aldiz sakatu ezkerreko botoia. Hala, A.6 irudian ageri den leihoa irekiko da, eta bertan azalduko dira aukera daitezkeen etiketak. Etiketa horiek aurredefinituta daude, hau da, ezin dira zuzenean asmatu anotazioa egiteko unean (beharrezkoa da etiketak aurretik erazagutzea Brat aplikazioaren administrazioko fitxategietan).

Honako hauek dira *Gero* obrako testua etiketatzeko aurredefinitu diren etiketak eta haien esanahia:

- **Aldaera**. Etiketa hori aukeratu behar du anotatzaileak baldin eta hitza aldaera bada, eta baliokide bat badu euskara estandarrean. Etiketa berria aukeratzeaz gain, baliokidea idatzi behar du anotatzaileak pantailaren beheko aldean ageri den "Notes" atalean (ikus A.7 irudia). Hori egin eta gero, etiketa aldatuta ageriko da hitza, eta sagua haren gainean jarriz gero, anotatzaileak emandako baliokidea irakurri ahal izango da (ikus A.8 irudian *eztiozu* aldaera).




A.7 Irudia: **Aldaera** etiketa aukeratuz gero, beheko 'Notes" atalean formaren baliokide estandarra adierazi behar da. Irudian *bilhatu* formari *bilatu* estandarra esleitu dio anotatzaileak.

- **Zalantza**. Etiketa hori aukeratu behar da ziurtasunik ez dagoenean, edozein izanik horren arrazoia: ez jakitea zein den baliokidea, ez jakitea aldaera den ala ez, edo egotea guztiz ziur emandako baliokideaz eta abar.

  Zalantzazkoen artean, **ZalantzaAUX** motako kasuak bereizi dira. Etiketa horrekin markatzen dira zalantzazkoak diren aditz laguntzaileak, anotatzaileak ez duelako garbi ikusten dagokion baliokide estandarra. Horrelako adibideak bereiztea interesgarria izan daitekeela aurreikusi da, etorkizunean aditz laguntzailearen inguruko azterketa egin nahi bada. Izan ere, etiketa berezitu horren bitartez erraz identifikatu ahal izango dira kasu horiek.

- **Zuzena**. Etiketa hori aukeratu behar da baldin eta testuko hitza zuzena bada, nahiz eta aurretik egin den analisian hitza ez-estandartzat hartu den (horregatik dago OOV gisa etiketatuta). Ikusi denaren arabera, halako kasuak dira, batez ere, izen propioak edo toki-izenak, analizatzaileak onartzen ez dituetank bere lexikoan ez daudelako (*Krisostomo*, *Szipion* eta abar).

- **HEE**, *Hitza Ez Euskara*. Siglaren esanahiak adierazten duen moduan, etiketa hori aukeratu behar da baldin eta hitza beste hizkuntza batekoa bada. A.3.1 atalean esan denez, latinezko zita ugari daude *Gero* obran, eta nahiz eta horiek automatikoki kentzeko saioa egin den, gerta daiteke tartean latinezko hitzen bat geratu izana.

- **EEE1** eta **EEEn**: *Esaldia Ez Euskara*. Aurreko etiketan azaldutako arrazoi beragatik, suerta daiteke latinezko esaldi bat edo hitz segida bat tartekatuta egotea testuan. Hitz horiek banan-banan etiketatzen egon beharrean, bi etiketa hauek proposatzen dira, esaldiaren lehenengo (EEE1) eta azkeneko (EEEn) hitzei soilik aldatzeko etiketa. Kasu horretan, erdiko hitzak aldatu gabe utziko dira, OOV gisa.

- **ErrTipog**. Azken etiketa hau proposatzen da etiketatzeko jatorrizko obrak izan ditzakeen errore tipografikoak. Kontuan izan behar da abiapuntu gisa daukagun *Gero* obraren testu-fitxategia bertsio elektroniko bat dela. Bertsio hori lortzeko bidea edozein izanik (OCR tresnen bidez edota norbaitek eskuz transkribatuta), jatorrizko testuari ez dagozkion hitzak azal daitezke, eta gerta daiteke anotatzaileak horrelako kasurik detektatzea OOV artean. Esaterako, testuan *bigarrren* hitza ageri da, non hiru *r* idatzi diren, eta errore tipografiko gisa etiketatu da.

A.8 Irudia: Testu zati baten itxura hainbat etiketa aldatu ondoren. Irudian ikusten denez, sagua etiketa baten gainean baldin badago, etiketa horren informazio osagarria ematen da ondoan azaltzen den laukitxoan.

Esan bezala, aurrekoak dira *Gero* obrako testua etiketatzeko aurredefinitu diren etiketak. Horrek ez du esan nahi etiketa gehiago definitu ezin direnik, eta agian beharrezkoa izango da hori egitea beste corpus batzuk lantzen badira. Azaldu berri diren etiketen adibide batzuk A.8 irudian ageri dira.

## A.4 Etiketatzearen inguruko oharrak

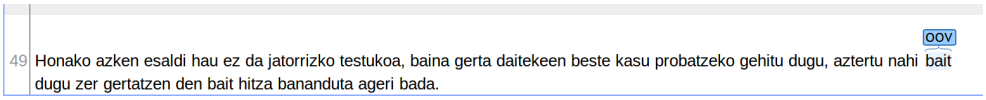
Etiketak deskribatu baino lehen (A.3.2 atala) aipatu da hitz bati OOV etiketa aldatzeko prozesua sinplea dela: (1) OOV etiketaren gainean sagua jarri eta bi aldiz sakatu ezkerreko botoia etiketen leihoa ireki dadin, eta (2) etiketa berria aukeratu, eta baliokide estandarra eman "Notes" atalean, hala badagokio. Hori bezain sinpleak izango dira etiketatu beharreko kasu gehienak, baina badira kontuan izan beharreko ohar batzuk etiketatzea zuzena izan dadin, eta horiek azpimarratu nahi ditugu hurrengo azpiataletan.

### Aldaera bati dagokion estandarra hitz anitzekoa denean

Etiketatu beharreko zenbait kasutan gertatuko da aldaerari dagokion baliokide estandarra hitz anitzekoa dela. Kasu horietan, estandarrari dagozkion hitz horiek idazterakoan "Notes" atalean, azpimarra batez lotuak idatziko dira. A.8 irudian halako kasu bat ageri da: *eztiozu* aldaerari bi hitz dagozkio, eta horiek lotuta idatzi dira laukitxoan ongi ikusten den moduan: *ez_diozu*. Gerora egin behar den prozesaketarako oso garrantzitsua da estandarra hitz anitzekoa denean horrela idaztea.

### Testuko hitz bat baino gehiago "elkartu" nahi denean

Kasu hau azaltzeko modurik errazena adibide bat jartzea da. Demagun A.9 irudiko testua etiketatu behar dela. Bertan *bait* hitza azaltzen da OOV gisa etiketatua, baina ez besterik. Anotatzaileak etiketa aldatu nahi dio hitz horri, baina ondorengoarekin lotuta, hau da, adierazi nahi du *bait dugu* dela aldaera eta *baitugu* estandarra dagokiola.



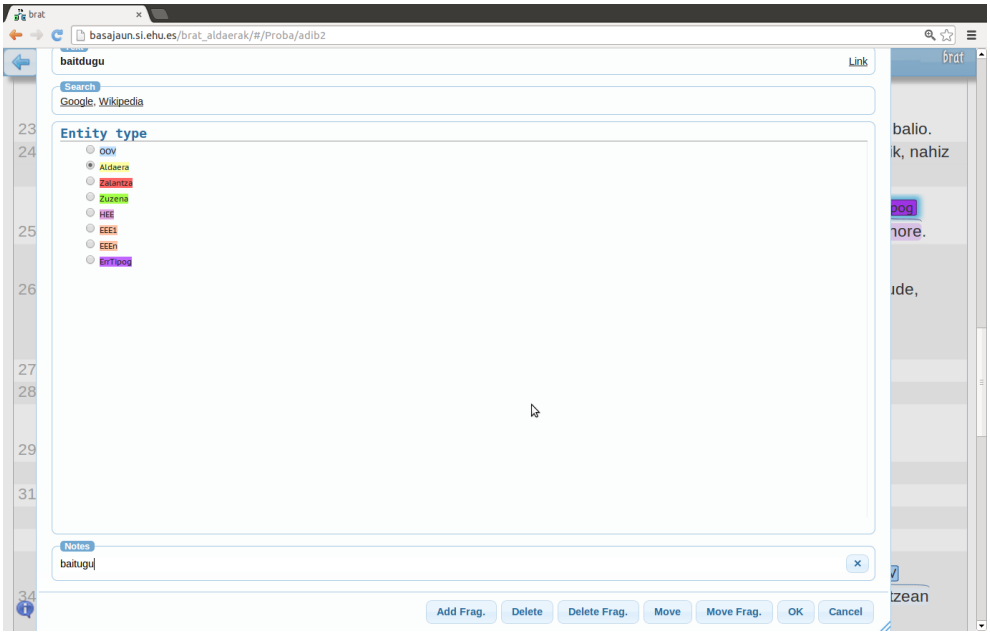

A.9 Irudia: Bi hitz elkartu anotazio bakar batean. Adibideko *bait* hitza OOV gisa markatuta dago eta *dugu* hitzarekin elkartu nahi da (*baitugu*).

Testuko *bait* eta *dugu* hitzak lotzeko egin behar dena hauxe da:

1. Bi aldiz sakatu *bait* hitzaren OOV etiketaren gainean etiketen leihoa ireki dadin.

2. Beheko aldearen eskuinean ageri diren aukeren artean "Add Frag." aukeran sakatu. Hori egin orduko aplikazioa testuko leihora itzuliko da. Leiho horretan *dugu* hitza aukeratu beharko da, sagua haren gainean jarriz eta ezkerreko botoia bi aldiz sakatuz, eta hori egin ondoren, bi hitzak "lotuta" azalduko dira etiketa berarekin.

3. Azken urratsean etiketa aldatzen da, sinple balitz bezala. Horretarako, elkartutako hitz baten etiketa gainean sakatuko da (berdin dio zeinen gainean) etiketen leihoa zabal dadin (ikus A.10 irudia), hitzei dagokien etiketa berria aukeratuko da (kasu honetan *Aldaera*), eta, hala



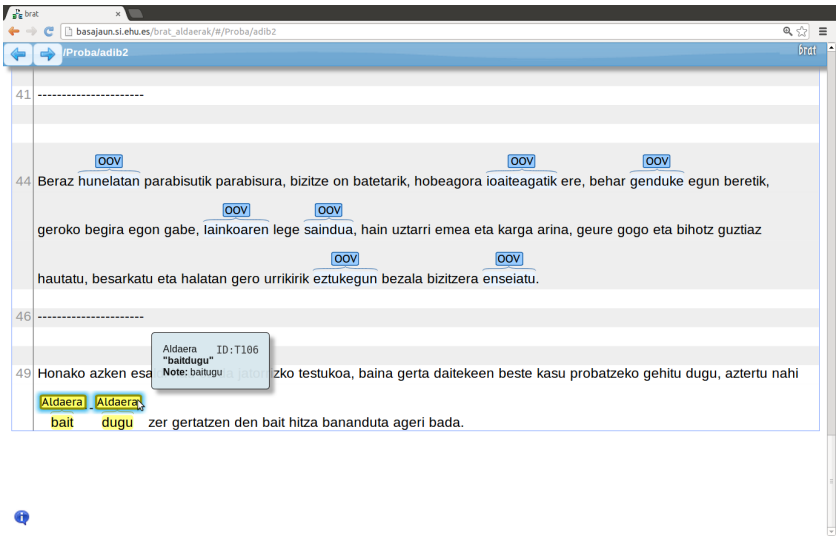

A.10 Irudia: Bi hitzak elkartu dira jadanik (hala ikusten da goiko aldean) eta leiho honetan *Aldaera* etiketa aukeratu dagokien baliokidea emanez: *baitugu*.

badagokie, baliokide estandarra idatziko da "Notes" atalean (*baitugu*). Azkenean, bi hitzak elkartuta geratuko dira etiketa bakar batez eta baliokide bakar batez, A.11 irudian ikusten den moduan.

### Etiketarik ez duen hitz bat etiketatu nahi denean

Corpusa etiketatzeko planteatzen den lana da OOV etiketak aztertzea eta egoki aldatzea aurreikusitako etiketen arabera. Baina horrek ez du esan nahi ezin dela beste hitzik etiketatu. Testuko edozein hitz etiketatu daiteke, eta horretarako egin behar den gauza bakarra da sagua hitzaren gainean jarri, ezkerreko botoia bi aldiz sakatu etiketen leihoa irekitzeko, eta etiketa aukeratu.

Anotazioaren helburua ez da OOV markarik ez duten hitzak etiketatzea, baina sor daiteke horren beharra. Esaterako, A.8 irudiko testuan latinez idatzitako esaldi zati bat ageri da etiketatuta: *Vana sine viribus ira (...)*. Baina testu zati hori analizatua izan denean OOV hitzak markatzeko, hitz horiek guztiak ez dira OOV gisa markatu, eta *sine* eta *ira* hitzei ez zaie



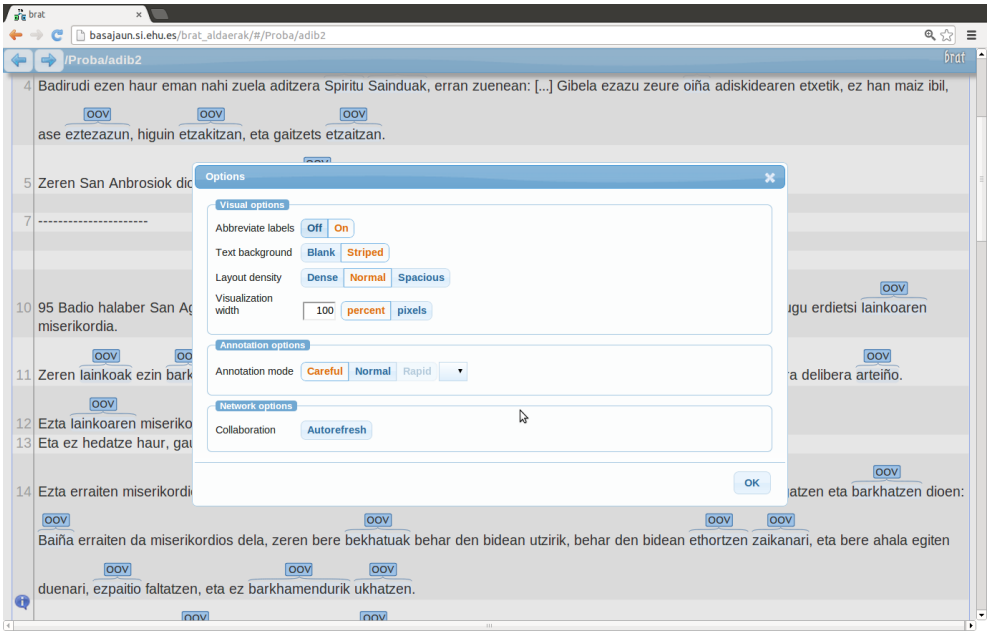

A.11 Irudia: Irudian ikusten da *bait dugu* bi hitzak elkartuta daudela etiketa bakar batekin, eta anotatu den baliokidea *baitugu* dela (laukitxoan ageri da biei dagokien informazio osagarria).

OOV etiketarik jarri. Esaldi osoa markatzeko, *Vana* hitzari EEE1 etiketa jarri behar zaio eta *ira* hitzari EEEn etiketa. Lehenengoak OOV etiketa du eta aldatu besterik ez da egin behar, baina azkenak ez du etiketarik, eta beraz, etiketarik ez duen hitz bat etiketatuko da kasu horretan.

## A.5 Anotazio-prozesua azkartzeko aukera

Brat tresna erosoa da anotazioak egiteko baina motel samarra suerta daiteke, eta anotatzaileak aukera du prozesu hori pixka bat azkartzeko, tresnak eskaintzen duen aukera bat aldatuta.

Tresnaren aukerak ikusteko, fitxategi bat ireki behar da eta goiko barra urdinaren gainean jarri behar da sagua ("login" egiteko egin den moduan). Zabaltzen den menuaren eskuin aldean "Options" botoia ageri da, eta hori sakatuz gero A.12 irudian ageri den leihoa zabaltzen da. Leiho horren erdian *Annotation options* aukera dago eta, ikusten denez, posible da *Careful* edo *Normal* aukeratzea. Lehenetsitako aukera *Careful* da, eta hori da gomendatzen dena, baina etiketatzaileak aukera du hori aldatzeko. *Careful* aukerarekin, hitz baten etiketatzea bukatzeko, beharrezkoa da beheko aldeko "OK" botoian sakatzea. Aldiz, *Normal* aukeratzen bada, hitzari dagokion




A.12 Irudia: Goiko barra urdinaren gainean “Options” sakatuta, tresnaren aukerak azaltzen dira leiho batean: *Visual options* eta *Annotation options*. *Careful* aukeratu beharrean *Normal* aukeratzen bada, sakatze kopurua txikiagoa da eta prozesua zertxobait azkartzen da.

etiketa aukeratu ahala leihoa ixten da eta testura itzultzen da aplikazioa, etiketatzaileak “OK” sakatu behar izan gabe. Hori dela eta, etiketatzea pittin bat azkarragoa izango da. Dena den, kontuan hartu behar da kasu horretan ordena inportantea dela, hau da, “Notes” atala bete behar bada baliokidea emateko, lehendabizi hori egin behar da eta gero etiketa aldatu. Aldrebes eginez gero, “Notes” bete gabe itzuliko da aplikazioa testuko leihora, eta denbora aurreztu beharrean, galdu egingo da berriro itzuli behar delako aurreko leihora.

# B. ERANSKINA

## Phonetisaurus tresnaren erabilera

### B.1 Sarrera

Phonetisaurus tresnaren funtzionamenduari buruz sakondu da IV. kapituluko IV.3 atalean, eta eranskin honen xedea da tresnaren tutorial moduko bat egitea, erabili diren komandoen inguruko xehetasunak jasotzeko.

### B.2 Tresnaren urratsak eta komandoak

Tresnaren funtzionamendua lau urratsetan banatuta ikus daiteke ,eta ondorengo lerroetan urrats horietako bakoitza zein programa eta komandoren bitartez bete dugun zehaztuko dugu.

#### B.2.1 Datu-prestaketa

Phonetisaurus tresna erabiltzen hasteko lehenengo urratsa datuak prestatzea da, hau da, entrenatzeko emango zaizkion datuak prestatzea formatu egokian. Informazio hori hainbat bikotez osatzen da, non lehenengo osagaia sarrerako informazioari dagokion, eta bigarrena irteeran lortu nahi den informazioari. Bikote horiek bi zutabeko testu-fitxategi batean eman behar dira, eta zutabe horiek banatzeko karakterea tabuladorea da.

Tresnaren helburua grafema-fonema bihurketa egitea denez, lehenengo zutabeko informazioa grafema-kate gisa ulertzen du Phonetisaurusek, eta ez da ezer berezirik egin behar, baina bigarren zutabeko informazioa fonema kate gisa ulertzen duenez, zuriunez banatuta espero ditu bigarren zutabeko

```
derauzkigute    d i z k i g u t e
derauzkitzu     d i z k i z u
derauzkitzutenak        d i z k i z u t e n a k
derauztegun     d i z k i e g u n
derragun        e s a n _ d e z a g u n
derrakegun      e s a n _ d e z a k e g u n
desiros g o g o t s u
deskonsolatuak  n a h i g a b e t u a k
dezakeielarik   d e z a k e e l a r i k
diatzatzun      d i e z a z k i z u n
diazoten        d i e z a i o t e n
diferentki      d e s b e r d i n k i
diferentziak    d e s b e r d i n t a s u n a k
diferentzien    d e s b e r d i n t a s u n e n
diozunorrek     d i o z u n _ h o r r e k
dirateken       d i r a t e k e e n
direiño d i r e n e r a i n o
disposizionean  d i s p o s i z i o a n
disposizionearen        d i s p o s i z i o a r e n
disposizioneari d i s p o s i z i o a r i
diteke  d a i t e k e
ditekeien       d a i t e k e e n
ditekeiena      d a i t e k e e n a
```

B.1 Irudia: Phonetisaurusek ikasteko behar duen informazioaren formatua erakusten duten adibideak. Lerro bakoitza adibide bat da: lehen osagaia aldaera da eta bigarrena hitz estandar bat, zeinaren karaktereak zuriunez banatuta dauden.

sinboloak. Gure kasuan sinbolo horiek ez dira fonemak, baina berdin dio: karaktereak zuriunez banatu behar dira hasi baino lehen. Aldaera/estandarra bikoteak baldin baditugu prest, nahikoa da programa sinple bat idaztea datuak formatu horretan adierazteko. B.1 irudian sarrerako hainbat bikote ageri dira oraintxe azaldutako formatuan.

### B.2.2 Datuen lerrokatzea

Entrenatzeko datuak lerrokatzeko `phonetisaurus-align` izeneko programa exekutatu behar da:

```
$ phonetisaurus-align --input=fi.train --ofile=fi.corpus
```

Programari aurreko urratsean prestatutako informazioa ematen zaio sarrera gisa `fi.train` fitxategian, eta programak datu horiek lerrokatuta itzul-

```
d|e}d r|a}i u}_ z}z k}k i}i g}g u}u t}t e}e
d|e}d r|a}i u}_ z}z k}k i}i t|z}z u}u
d|e}d r|a}i u}_ z}z k}k i}i t|z}z u}u t}t e}e n}n a}a k}k
d|e}d r|a}i u}_ z}z t}k|i e}e g}g u}u n}n
d}e|s e}a _}n|_ r}d r}e|z a}a g}g u}u n}n
d}e|s e}a _}n|_ r}d r}e|z a}a k}k e}e g}g u}u n}n
d}g e|s}o i|r}g o}o s}t|s _}u
d|e}n s}a|h k|o}i n|s}g o|l}a a}b|e t}t u}u a}a k}k
d}d e}e z}z a}a k}k e|i}e e}e l}l a}a r}r i}i k}k
d}d i}i _}e|z a}a t|z}z a}k|i t|z}z u}u n}n
d}d i}i a}e|z z}a|i o}o t}t e}e n}n
d}d i}e f}s|b e}e r}r e}d|i n}n t|k}k i}i
d}d i}e f}s|b e}e r}r e}d|i n}n t|z}t i}a|s a}u|n k}a|k
d}d i}e f}s|b e}e r}r e}d|i n}n t|z}t i}a|s _}u e}n|e n}n
d}d i}i o}o z}z u}u n}n|_ o}h|o r}r r}r e}e k}k
d}d i}i r}r a}a t}t e}e k}k e}e n}e|n
d}d i}i r}r|e e}n|e i}r|a ñ}i|n o}o
d}d i}i s}s p}p o}o s}s i}i z}z i}i o}o n|e}_ a}a n}n
d}d i}i s}s p}p o}o s}s i}i z}z i}i o}o n|e}_ a}a r}r e}e n}n
d}d i}i s}s p}p o}o s}s i}i z}z i}i o}o n|e}_ a}a r}r i}i
d}d|a i}i t}t e}e k}k e}e
d}d|a i}i t}t e}e k}k e|i}e e}e n}n
d}d|a i}i t}t e}e k}k e|i}e e}e n}n a}a
```

B.2 Irudia: Lerrokatze-prozesuak ikasteko datuen formatua aldatzen du, irudiko adibideetan ikusten den moduan.

tzen ditu `fi.corpus` fitxategian[1]. Irteerako fitxategia testu-fitxategi bat da eta B.2 irudian ikus daiteke zer nolako informazioa jasotzen duen (B.1 irudiko adibide berberak ageri dira B.2 irudian).

### B.2.3 *Joint n-gram* ereduaren entrenatzea

Urrats honetan erabili dugun tresna OpenGrm izan da (beste batzuk erabil daitezke), eta hauek izan dira erabilitako komandoak:

```
$ ngramsymbols < fi.corpus > fi.syms
$ farcompilestrings --symbols=fi.syms \
 --keep_symbols=1 fi.corpus > fi.far
$ ngramcount --order=7 fi.far > fi.cnts
$ ngrammake --method=kneser_ney fi.cnts > fi.mod
$ ngramprint --ARPA fi.mod > fi.arpa
```

Komando horietan ikusten da $7-gram$ eredu bat entrenatu dela (–order=7) eta Kneser_Ney estrategia erabili dela hizkuntza-eredu normalizatua lortzeko. Horiek dira Phonetisaurusen tutorialean proposatutako balioak eta horiekin egin dugu lan.

[1] Programa egikaritzen ari den bitartean, pantailan *Expectation-Maximization* prozesuari buruzko informazioa ageri da.

Aurreko komandoen bitartez lortutako eredua, fi.arpa, ARPA formatuan dago, eta WFST formatuan adierazteko beste programa bat exekutatu behar da:

```
$ phonetisaurus-arpa2fst --input=fi.arpa --prefix="fi"
```

Programa horren bitartez, `fi.fst` WFST transduktorea lortzen da eta ikasketa-prozesua amaitutzat ematen da.

### B.2.4 Deskodeketa

Deskodeketa-urratsean aurretik lortutako eredua erabiltzen da sarrera berriei dagokien hipotesi onena lortzeko.

Hori egiten duen programa `phonetisaurus-g2p` da, eta hainbat parametro eman behar zaizkio:

```
$ phonetisaurus-g2p --model=fi.fst --input=test --isfile
--words --nbest=5 --beam=5000 >test.out
```

Komando horretan programari adierazten zaio eredua inplementatzen duen transduktorea `fi.fst` dela, deskodetu beharreko sarrera `test` izeneko fitxategian dagoela, hipotesiarekin batera sarrerako hitza emateko irteeran eta 5 hipotesi onenak bilatzeko sarrera bakoitzeko. Azken parametroak, *beam* izenekoa, bilaketaren sakonera mugatzen du. Lehenetsitako balioa parametro horretan 500 da, eta suposatzen da horrekin azkarragoa izango dela deskodeketa. Azkartasuna ez da gure esperimentuetan kontuan hartu dugun irizpidea eta probak bi balioekin egin ditugu. Ez dugu alde handirik sumatu abiadurari dagokionez eta hipotesi hobeak lortu dira beti 5000 balioarekin (espero zitekeen moduan).

# C. ERANSKINA

## Izen propioak *Gero* corpusean

Eranskin honetan azalduko dugu izen propioek sortzen duten arazoa tratatzeko egin dugun saioa. *Gero* obraren corpusean, bai ikasteko zatian eta bai testeko zatian, OOV gisa detektatutako hainbat hitz izen propioak dira, euskara estandarreko automatak onartzen ez dituen izen propioak. Horien artean daude, esaterako: *Aristotelek*, *Davitek*, *Krisostomok* eta abar. Horietako batzuk *Aldaera* moduan etiketatu ditu anotatzaileak, eta haien estandarrarekin lotu ditu, esaterako:*Aristotelek* → *Aristotelesek*, eta *Davitek* → *Davidek*; baina beste batzuk zuzenak kontsideratu ditu eta *Zuzena* etiketa esleitu die, esaterako: *Krisostomok*.

Etiketa bat edo beste jarri arren, eraiki dugun WFST sistemak ez du inoiz ongi ebatziko izen propio baten normalizazioa. Zergatik gertatzen da hori? Kontuan hartu behar da Phonetisaurusi ikasteko ematen zaizkion bikote guztiak letra xehez idatzita daudela, izen propioak barne, eta beraz, tresnak emango dituen erantzunak ere, beti izango dira letra xehez idatzitakoak. Baina erantzun horiek ontzat eman daitezen, iragazki bat pasa behar dute, hau da, estandarraren automatak onartu behar ditu, eta logikoa denez, automata horrek ez ditu onartuko *aristotelesek** edo *davidek** bezalako erantzunak, ez baitira letra larriz hasten izen propioei dagokien moduan.

Arazo hori konpondu nahian, eta arazoaren garrantzia analizatu ahal izateko aldi berean, proba bat egin dugu zuzenean testeko corpusarekin, eta sistemak jarraitzen duen algoritmoan aldaketa bat egin dugu: Phonetisaurusek itzultzen dituen erantzunek iragazkia pasako dute automatak bere horretan onartzen baditu, hau da, xehez idatzita, edo lehenengo letra larriz jarrita onartzen baditu. Hala, Phonetisaurusek *davidek* proposatzen badu, erantzun hori onartuko da *Davidek* bai onartzen duelako automatak.

Aldaketa hori egin ondoren, ebaluazioa errepikatu dugu kasu bakar batekin: IV.21 taulan emaitza onena lortu duen kasuarekin, hau da, ikasteko *hitza-hitza* informazioa erabiliz, eta hitz estandarren erdia erabiliz.

Algoritmo berriaren emaitzak C.1 taulan ageri dira, eta konparazioa errazteko, aurreko taulako kasu beraren emaitza ere taula horretan kopiatu da.

| | $\boldsymbol{P}$ | $\boldsymbol{R}$ | $\boldsymbol{F}_1$ |
|---|---|---|---|
| letra larria tratatu gabe | 91,84 | 79,51 | 85,23 |
| letra larria tratatuta | 91,53 | 80,21 | **85,50** |

C.1 Taula: Izen propioak tratatzeko egin den saioaren emaitzak testeko corpusean. Ikasteko *hitza-hitza* informazioa erabili da (hitz estandarren erdia barne) eta 5 erantzun eskatu dira.

Ikusten denez, estaldura handiagoa da orain, hots, erantzun zuzen gehiago ematen ditugu eta hori espero genuen testean *Aldaera* motako izen propioren bat egonez gero, baina horrekin batera, emaitzetan ikusten da doitasuna jaitsi dela, eta hori ez zaigu interesatzen. Zergatik jaitsi da doitasuna? Horrek esan nahi du letra xehe/larri tratamendu horrekin erantzun gehiago eman direla, baina ez beti onak, zeren erantzun okerren proportzioa handiagoa da orain.

Hori non eta zergatik gertatu den jakiteko, erantzun berrien azterketa kualitatibo bat egin dugu adibideren bitartez ulertzeko gertatu dena. Testeko aldaeren artean dago, esaterako, *uriak* aldaera, eta anotatzaileak *uriak* → *urak* esleipena egin du. Phonetisaurusek proposatutako erantzun guztiak okerrak kontsideratu dira sistemaren lehenengo bertsioan, guztiak ez-estandarrak direlako (automatak ez ditu onartzen). Bost erantzun horiek dira, ordenan: *uriak*, *iak*, *uiak*, *udiak* eta *oriak*.

Atal honetan proposatutako algoritmo-aldaketarekin, ordea, sistemak *uriak* eman du erantzun gisa, izan ere hitz horren lehen letra larriz jarrita, *Uriak*, automatak onartzen du. Beraz, erantzun bat gehiago eman du orain sistemak, baina okerra, ez baitator bat anotatzaileak esleitutakoarekin.

Beraz, posible da, bai, letra xehe/larri tratamendu horrekin bi motatako kasuak izatea: normaliza daitezke izen propioei dagozkien kasuak (lehen ezinezkoa zena), baina, aldi berean, erantzun oker gehiago eman daitezke, lehen onartzen ez zen erantzunen bat orain bai onartzen delako, eta ez da erantzun zuzena.

Dena den, lortutako emaitzek adierazten dutena da proportzioan ez direla asko testeko corpusean ageri diren izen propioak: estaldura 0,7 igo da soilik.