

**LENGOAIA ETA SISTEMA INFORMATIKOEN SAILA**



# **EUSKAL MORFOLOGIAREN TRATAMENDU AUTOMATIKORAKO TRESNAK**

**Euskararako prozesadore morfologiko sendo baten diseinua  
eta eraikuntza. Oinarri horrekin osatutako zuzentzaile  
ortografikoa.**

**Iñaki Alegria Loinazek**

Informatikan Doktore titulua eskuratzeko aurkezturiko

**TESI-TXOSTENA**

Donostia, 1995eko apirila.



## LENGOAIA ETA SISTEMA INFORMATIKOEN SAILA



# EUSKAL MORFOLOGIAREN TRATAMENDU AUTOMATIKORAKO TRESNAK

**Euskararako prozesadore morfologiko sendo baten diseinua  
eta eraikuntza. Oinarri horrekin osatutako zuzentzaile  
ortografikoa.**

Iñaki Alegriak Xabier Artolaren eta Kepa  
Sarasolaren zuzendaritzapean egindako  
tesiaren txostena, Euskal Herriko  
Unibertsitatean Informatikan Doktore  
titulua eskuratzeko aurkeztua.

Donostia, 1995eko apirila.



*Euskal Unibertsitatearen alde diharduen orori*

*Bereziki Joserrari eta Koldori, garai zail hauetan*



*“Maite ditut  
maite  
gure bazterrak  
lanbroak  
izkututzen dizkidanean  
zer izkututzen duen  
ez didanean ikusten  
uzten  
orduan hasten bainaiz  
izkutukoa ...”*  
(J. A. Artze)

*“Hegazti errariak  
pausatu dira  
leihoan  
argia eta itzala  
bereizten diren lekuan  
argia eta itzala  
leihoan  
pausatu dira  
hegazti errariak”*  
(J. Sarrionaindia)





## *eskerrak ematen edo zorrak kitatzen*

- Xuxenkide guztioi, guztion lana baita hau, adiskideok. Kepa, Xabier, Arantza, Xabier, Eneko, Montse, Nerea, Miriam, Itziar, Jose Mari, Izaskun, Koldo, Jon Mikel, Aitor, Alexander, ... norberaren zein taldearen laguntzarik gabe tesi hau ez litzateke existituko. Euskaraz egindako ikerketa aplikatuaren aldeko apustu honek aurrera jarrai dezan.
- Konputagailuen Arkitektura eta Teknologia nere Saileko lagunoi eta bereziki Olatzi eta Agusi; tesi honen arkitekturan erru handia izan duzue eta.
- “Kaxkarin” sindikatukideoi, eta bereziki tesi honen alde apustua egin zuenari, “tesi berdin krisi” leloari kontraadibidea bilatzearren. Gurekin batera kaxkarin izateko nahikoa “curriculum” egin duzuen guztioi.
- Irakaslego kontratatuaren “mugida”ko lankide eta borrokaideoi; unibertsitateko jauntxoan burugogorkeriaren aurrean, aurrera jarraitzeko emandako laguntzarengatik.
- Fakultateko garai zahar-zailetako ausarti guztioi.
- Lauri Karttunen-i eta Ken Bessley-ri lexiko-itzultzaileekin emandako laguntza eskuzabalarengatik.
- *awk* programaren egileei, lan asko aurreratzen laguntzeagatik.



# AURKIBIDEA

SARRERA ETA AURKEZPEN OROKORRA	9
<b>I. Lanaren nondik norakoak eta aurkezpen orokorra.</b>	<b>9</b>
I.1. Sarrera gisako aurkezpena. ....	9
I.2. Hizkuntzaren prozesaketa automatikoaren oinarria eta aplikazioak. Proiektuaren helburuak. ....	11
I.3. Euskararen ezaugarriak modu laburrean. ....	13
I.4. Zimenduak: Euskararako Datu-Base Lexikala eta corpus-ak. ....	14
I.4.1. Euskararako Datu-Base Lexikala (EDBL). ....	14
I.4.2. Corpus-ak. ....	15
I.5. Egiturazko tresna: prozesadore morfologiko automatikoa. ....	17
I.6. Prozesaketa morfologikoa hobetzen: Lexiko-itzultzaileak. ....	19
I.7. Produktu komertziala: Xuxen zuzentzaile ortografikoa. ....	20
I.8. Hurrengo urratsa: EUSLEM. ....	21
I.9. Egindakoaren aplikazio berri posibleak. ....	22
I.10 Txostenaren eskema. ....	23
 LEHEN PARTEA: ANALISI MORFOLOGIKOA	 11
<b>II. Egoera finituko morfologiaren inguruan.</b>	<b>11</b>
II.1 Analisi morfologikoa: sarrera gisakoa. ....	12
II.2 Morfologiaren eredu konputazionalak eta zenbait adibide. ....	13
II.2.1 Eredu konputazionalak: sailkapenerako irizpideak. ....	13
II.2.2 Adibideak. ....	16
II.2.2.1 DECOMP. ....	16
II.2.2.2 ATEF. ....	17
II.2.2.3 KIMMO. ....	19
II.2.2.4 Tzoukermann eta Liberman. ....	20
II.2.3 Sailkapen bat. ....	22
II.3 Bi mailatako morfologia. ....	22
II.3.1 Lexiko-sistema. ....	23
II.3.2 Bi mailatako erregelak. ....	27
II.3.2.1 Sarrera. ....	27
II.3.2.2 Osagaiak. ....	29
II.3.2.3 Erregelen formatua. ....	30
II.3.2.4 Erregelatik automatara. ....	32
II.3.3 Programa eta exekuzio-eredua. ....	34
II.3.4 Sistemaren gaineko kritikak eta proposamenak. ....	36
II.3.4.1 Deskribapen-ahalmena. ....	36
II.3.4.2 Hautapen-markak edo diakritikoak. ....	38
II.3.4.3 Morfotaktika: jarraitze-klaseak vs. baterakuntza- mekanismoak. ....	38
II.3.5 Ekarpen bat: jarraitze-klase hedatuak. ....	40
II.3.5.1 Deskripzioa. ....	40

II.3.5.2	Sintaxia .....	42
II.3.5.3	Semantika .....	42
II.4	Bi mailatako ereduaren konputazio-konplexutasuna eta azkartzeko bideak .....	43
II.4.1	Eraginkortasunaren aldetiko arazoak .....	43
II.4.2	Konputazio-konplexutasuna zehaztuz .....	44
II.4.3	Proposatutako hobekuntzak .....	45
II.4.3.1	Lexikoen fusioa .....	45
II.4.3.2	Lexiko-itzultzaileak .....	46
<b>III.</b>	<b>Prozesadore morfologiko bat euskara estandarerako.</b>	<b>51</b>
III.1	Ereduaren egokitasuna eta jarritako mugak .....	52
III.2	Euskararen morfologia laburtua .....	53
III.3	Lexikoa .....	55
III.3.1	EDBL: Euskararako datu-base lexikala .....	56
III.3.2	Lexikoko alfabetoa: morfofonemak eta hautapen-markak .....	59
III.3.3	Morfotaktika .....	61
III.3.3.1	Azpilexikoak .....	61
III.3.3.2	Jarraitze-klaseak .....	62
III.3.3.3	Izenaren eta adjektiboaren morfotaktika .....	64
III.3.3.4	Aditz-erroaren morfotaktika .....	65
III.3.3.5	Aditz jokatuaren morfotaktika .....	66
III.4	Erregelak .....	67
III.4.1	Aurredefinizioak .....	67
III.4.2	Erregela morfofonologikoak .....	69
III.4.3	Erregela ortografikoak .....	77
III.5	Programa eta emaitzak .....	78
III.5.1	Implementazioa .....	78
III.5.1.1	Programa .....	78
III.5.1.2	Token-ezagutzailea edo iragazlea .....	80
III.5.2	Analizatzailearen emaitzak eta estaldura-tasa .....	81
III.5.3	Gainsorreraren arazoak .....	84
III.5.4	Eraginkortasunari buruzko zenbait datu eta gogoeta .....	85
III.6	Erabateko hobekuntza: lexiko-itzultzaileak .....	87
III.6.1	Lexiko-itzultzaileen ezaugarriak .....	87
III.6.2	Euskararako aplikazioa .....	88
III.6.2.1	Urruneko menpekotasunak ebazteko erregelak .....	88
III.6.2.2	Ohiko diakritikoen eta erregelen berrikuntza .....	90
III.7	Morfosintaxia .....	92
<b>IV.</b>	<b>Analizatzaile sendoa osatzen.</b>	<b>95</b>
IV.1	Erabiltzailearen lexikoa .....	96
IV.1.1	Azpilexikoen ezaugarri garrantzitsuak .....	96
IV.1.2	Burutzapena .....	97
IV.1.3	Eguneratzeko prozedura .....	99
IV.2	Forma ez-estandarren analisisa .....	100
IV.2.1	Oinarria: bi mailatako mekanismo osagarria .....	100
IV.2.2	Azpilexikoak eta erregela osagarriak .....	102
IV.2.2.1	Azpilexikoak .....	102
IV.2.2.2	Erregelak .....	104
IV.2.3	Aldaera-motaren identifikazioa. Desanbiguazio lokala .....	108

IV.2.3.1. Aldaera-mota eta kopurua. ....	108
IV.2.3.2. Desanbiguazio lokala .....	110
IV.2.4. Integrazioa lexiko-itzultzaileetan.....	111
IV.2.5. Emaitzak, konplexutasuna eta erabilpenak.....	113
IV.3 Lema lexikoan ez duten hitzen analisia. ....	114
IV.3.1. Gakoa: bi mailatako erregela bereziak.....	114
IV.3.2. Emaitzak, lemaren bilaketa eta desanbiguazio lokala. ....	117
IV.3.2.1. Lemaren bilaketa .....	118
IV.3.2.2. Desanbiguazio lokala .....	119
IV.4 Analizatzaile sendoa. Emaitzak. ....	119
 <b>BIGARREN PARTEA: ZUZENKETA ORTOGRAFIKOA</b>	 <b>121</b>
<b>V. Erroreen zuzenketa.</b>	<b>121</b>
V.1. Aplikazioak, sailkapena eta irizpideak. ....	122
V.2. Egiaztatzea.....	123
V.2.1 Hitz-zerrendatan oinarritutako metodoak .....	124
V.2.2 Hitz-zatitan oinarritutako metodoak .....	125
V.2.3 Morfologian oinarritutako metodoak .....	126
V.3. Zuzenketa.....	128
V.3.1 Errore-motak eta ezaugarriak.....	128
V.3.1.1 Oinarrizko sailkapena.....	128
V.3.1.2 Aplikazioarekin lotutako ezaugarriak. ....	130
V.3.1.3 Aldaerak. ....	131
V.3.1.4 Tratamenduaren garrantzia errore-mota eta aplikazioaren arabera. ....	132
V.3.2 Antzekotasun-neurriak.....	133
V.3.3 Zuzenketa-metodoak.....	135
V.3.3.1 Oinarrizko metodoak.....	135
V.3.3.2 Metodo konbinatuak.....	137
V.4. Hizkuntza flexionatuen eta eranskarien zuzenketa.....	139
V.4.1 Lexikoaren aberasketa.....	139
V.4.2 Zuzenketa. Zenbait adibide. ....	140
V.4.3 Ondorioak. ....	142
 <b>VI. Xuxen: bi mailatako morfologian oinarritutako zuzentzaile ortografikoa.</b>	 <b>145</b>
VI.1. Sarrera.....	146
VI.2. Egiaztatzea.....	147
VI.3. Errore tipografikoen tratamendua.....	149
VI.3.1. Azkartzeko bideak. ....	150
VI.4. Gaitasun-erroreen zuzenketa. ....	153
VI.5. Sistemaren arkitektura eta ezaugarriak.....	156
VI.5.1. Proposamenen sailkapena.....	156
VI.5.2. Erabiltzailearen hiztegia. ....	158
VI.5.3. Iragazlea edo token-ezagutzailea. ....	159
VI.6. Produktu komertzialaren diseinua. ....	160
VI.6.1. Doitasuna/eraginkortasuna oreka.....	160
VI.6.2. Erabiltzailearekiko interfazea. ....	161
VI.7. Doitasuna eta eraginkortasuna. ....	163

VI.7.1. Egiaztatzea. ....	163
VI.7.2. Zuzenketa. ....	165
VI.8. Proposatutako hobekuntzak. ....	167
VI.8.1. Lexiko-itzultzaileen erabilera. ....	168
VI.8.2. Erro-hizkiaren bidezko proposamen-sistema. ....	168
 <b>ONDORIOAK ETA AURRERA BEGIRAKOAK</b>	 <b>169</b>
<b>VII. Ondorioak eta zabaldutako ikerlerroak.</b>	<b>169</b>
VII.1. Ondorioak. ....	169
VII.2. Zabaldutako ikerlerroak eta perspektibak. ....	170
VII.2.1 Prozesaketa morfologikoa hobetzen. ....	171
VII.2.2 Zuzenketa. ....	171
VII.2.3 EUSLEM. ....	172
VII.2.4 Beste aplikazioak. ....	173
 <b>BIBLIOGRAFIA</b>	 <b>175</b>
<b>Morfologia</b>	<b>175</b>
<b>Egiaztapen/zuzenketa ortografikoa.</b>	<b>179</b>
<b>Etiketatzeta.</b>	<b>182</b>
<b>Euskararen deskribapena.</b>	<b>183</b>

