Morfologia Konputazionala Euskaraz, 35 urte[1]

Itziar Aduriz,*[1,3] Manex Agirrezabal,[2] Eneko Agirre,[3] Iñaki Alegria,[3]
Xabier Arregi,[3] Jose M. Arriola,[3] Xabier Artola,[3]
Arantza Diaz de Ilarraza,[3] Ainara Estarrona,[3]
Izaskun Etxeberria,[3] Nerea Ezeiza,[3] Kepa Sarasola[3]

[1] Universitat de Barcelona
[2] Centre for Language Technology, Kopenhage (CST), University of Copenhagen (KU)
[3] HiTZ Zentroa - Ixa. Euskal Herriko Unibertsitatea UPV/EHU

Laburpena: Artikulu honetan morfologia konputazionalaren garapena azalduko dugu, Ixa taldeak euskararako egindako aplikazioa azpimarratuz. Bilakaera historikoa jaso nahi izan dugu, teknologiaren bilakaera eta aplikazioen bilakaera uztartuz, beti euskararen gainean egindakoa adibidetzat hartuta. Teknologiaren bilakaerari dagokionez, ezagutzan oinarritutako ereduak ohikoak eta nagusi izan dira morfologian, baina hurbilpen estatistikoan, eta, batez ere, ikasketa sakonean oinarritutakoak, ugalduz joan dira azken urteetan, horietako asko datuak eta gramatikak eskuragarriak ez dituzten hizkuntzetarako (edo dialektoetarako). Aplikazioak direla eta, ezagunena eta erabiliena zuzentzaile ortografikoa da, baina irakaskuntzarekin edo bertsogintzarekin lotutakoak ere aipagarriak dira.

Hitz gakoak: morfologia konputazionala, zuzenketa ortografikoa, hizkuntz teknologia.

1. SARRERA

Miren Azkarateren omenezko liburuari gure ekarpena egin nahi diogu Mirenek landu duen arlo nagusietako batean, morfologian hain zuzen. 1988 urtean Ixa taldea sortu zenean, erabaki estrategiko garrantzitsua izan zen euskararen morfologiaren tratamenduari lehentasuna ematea. Lanari ekin genionean gazteak eta berriak ginen arloan, geneuzkan konputagailuak oso

* Itziar Aduriz. Filologia Catalana i Lingüística General. Facultat de Filologia i Comunicació (Universitat de Barcelona). Gran Via de les Corts Catalanes, 585 (08007 Barcelona). itziar.aduriz@ub.edu. https://orcid.org/0000-0001-8681-9778.

[1] Lan hau bi ikerketa-proiektu hauei atxikitzen zaie: Ixa Taldea. A motako Talde Finkatua (Eusko Jaurlaritza: IT-1570-22) eta HARTAvas (Zientzia eta Berrikuntza Ministerioa, MCIN: PID2019-109683GB-C22).

oinarrizkoak ziren, baliabideak urri, nazioarteko zientzia kontsultatzeko aukerak, zailak; baina bagenuen urteekin baloratzen ikasi dugun altxor bat: euskara batuaren definizioa, eta bere barruan guretzat ezinbestekoak ziren bi osagai: lexikoa eta morfologia. Horrez gain, arauari jarraitzen zioten testuak eta lankidetzarako prest zegoen komunitate zabal bat. Hori gabe nekez lortuko genuen urte hauetan lortutakoa, Xuxen zuzentzaile ortografikoa edo artearen egoeran dagoen itzultzaile automatikoa, esaterako. Miren Azkarate komunitate horren partaide eta bidelagun izan dugu hasieratik. Ez da kasualitatea, Ixa taldean egindako hizkuntzalaritzako lehen tesian (Urkia 1997) Miren Azkarate zuzendari izatea. Lehen kolaborazioa estua izan zen, baina ez zen azkena izan. Izan ere, Ixa taldeko hainbat eta hainbat tesitako epaimahaietan partaide izan da eta ikerketa-proiekturen batean ere lankide.

Horrela, morfologia konputazionalaren funtsa eta bilakaera, eta bereziki euskararen gainean egindako aplikazioak, joango gara azaltzen artikulu honetan. Atalen banaketa bi ardatzen gainean dago eraikita: teknologiaren bilakaera eta izaera, batetik, eta eratorritako aplikazioak, bestetik. Teknologiari dagokionez, ezagutza linguistikoan oinarritutako ereduak ohikoak eta nagusi (izan) dira morfologian, baina hurbilpen estatistikoan, eta, batez ere, ikasketa sakonean oinarritutakoak ugalduz joan dira azken urteotan, horietako asko datuak eta gramatikak eskuragarriak ez dituzten hizkuntzetarako (edo dialektoetarako). Aplikazioei dagokienez, ezagunena eta erabiliena zuzentzaile ortografikoa da, baina irakaskuntzarekin, itzulpen automatikoarekin edo bertsogintzarekin lotutakoak ere aipagarriak dira.

Horrela, euskal morfologiaren modelizazio konputazionala azalduko dugu bigarren atalean, modelizazio horretatik abiatutako zenbait aplikazio ikusiko ditugu hirugarrenean, eta azken urteotan arloko ikergai den morfologiaren inferentzia, datuetatik informazio morfologikoa erauztea, hain zuzen ere, laugarren kapituluan. Amaieran dator ondorioak eta etorkizunari begirako gakoak biltzen dituen atala.

2. EUSKAL MORFOLOGIAREN MODELIZAZIOA

Esan bezala, Ixa taldeak, sorreratik, estrategia bat definitu zuen euskararen prozesaketa modu iraunkorrean garatu ahal izateko (Aduriz *et al*. 2011), eta estrategia horren barruan lexikoak eta morfologiak erabateko lehentasuna zuten. Artikuluan azaltzen denez, estrategiaren puntu nagusiak hauek ziren:

— Hasieran oinarrizko baliabide eta tresna sendoak sortu behar dira.
— Lexiko-morfologia-sintaxia-semantika progresioa aplikatu behar da.
— Formatu estandarrak erabili behar dira.
— Ahal den guztietan saiatu behar da software librea erabiltzen eta sortzen.

Horrela, bada, lexikoari eta morfologiari ekin genion. Euskara batuaren lexikoa eta gramatika aztertuta, lexikoaren datu-basea osatzera jo genuen eta morfologia modelizatzeko formalismo baten bila hasi ginen, euskal morfologia konputazionalki landu ahal izateko.

Eredu baten bila, arloko antzeko lanak bilatzen hasi eta ageriko kontu batekin egin genuen topo, alegia, ikerketa gehiena ingeleserako egiten zela eta ingeleserako aplikatzen ziren tresnak ez zirela egokiak euskara bezalako hizkuntza eranskari bati aplikatzeko. Eranskaria izanda, ingelesak baino morfologia aberatsagoa du euskarak.[2] Hizkuntzen tipologia morfologikoaren inguruko sailkapenetik dator eranskari terminoa, morfemen fusioaren edo uztartze mailako irizpidea aplikatuta. Horrela, hizkuntza eranskarietako hitzen morfemek esanahi bakarra dute, bakoitzak bat, eta ez dago fusiorik, latinean topatzen ahal dugun moduan, non morfo bakarrak morfema bati baino gehiagori egiten baitio erreferentzia. Espero izatekoa da, beraz, hizkuntza eranskarietan morfema kopuru altuagoa hitzeko, beste hizkuntza malgukariekin edo flexiboekin konparatuta. Are gehiago, hizkuntzen sailkapen morfologikorako tipologian erabiltzen den beste irizpidea aplikatuta, hots, sintesiarena. Irizpide honi jarraituz, kontuan hartzen da hitz batek dituen morfema kopurua. Sailkapen honetan daude hizkuntza analitikoak, isolatzaileak ere deituak, morfema bakarrekoak, eta beste muturrean, polisintetikoak, hiru morfema baino gehiago dituztenak hitzeko. Tartean eranskariak daude. Hizkuntza bakartzaileen eredurik argiena txinera da eta ingelesa ere bai, neurri handi batean. Esan dezagun, hala ere, sailkapen honetako kategoria hauek guztiak idealak direla, hizkuntza-tipologia puruei dagozkielako. Hizkuntzak, ordea, ez dira tipologikoki puruak, baina hurbiltzen dira ideal horietara. Hori dio, adibidez, Manterolak bere artikuluan euskararen eranskaritasunari buruz, nahiz eta ikuspegi diakronikoa izan (Manterola 2008).

Gauzak horrela, behin informazio lexiko eta morfologikoa ordenagailuan txertatua izan genuenean erraza izan zitzaigun frogatzea euskal morfologiaren aberastasuna. Eta frogatuta geratu zen euskararen aberastasun hori forma flexionatuak sortzeko duen ahalmen izugarrian dagoela. Aberastasun horren adibideak ditugu honako hauek (Alegria 1995: 70):

> Adibidez, izen-sarrera batetik abiatuz 135 forma flexionatu lor daitezke gutxienez. Horietako 77 determinazioa, numeroa eta deklinabide-kasua konbinatuz lortutako forma ez-emankorrak diren bitartean, gainerako beste 58ak forma emankorrak dira bi genitiboetako batez bukatutako forma sinple edo deklinatuak baitira. Genitiboen atzetik teorikoki hasierako emankortasun-ahalmen guztia dago, genitiboaren atzetik flexiorik agertzean elipsi bat dago eta. Elipsi bat baino gehiago posible izanik, atzizki-hartzea

[2] Horrela aipatzen da, adibidez, SEG gramatikan: «Egia da euskaraz morfologiak garrantzi handia duela, hizkuntza eranskaria baita (izen sintagmari dagokionez) eta aditza bera ere morfologia aldetik oso aberatsa baita». (Salaburu, Sareko Euskal Gramatika, SEG).

> errekurtsiboa izan liteke, maila teorikoan behintzat, eta ondorioz, emankortasun-ahalmena infinitua litzateke. Izan ere, elipsi bat baino gehiago agertzea ohizkoa ez bada ere oso arraro ez diren forma batzuek bi elipsi edo gehiago dute. Aurrekoaren ondorioz eta bi elipsi kontuan hartuz izen bati dagozkion forma flexionatuak honako hauek lirateke : 77 + 58 (77 + 58 (77 + 58)) = 458683 (Agirre *et al*. 92). Izen bakoitzeko horiek baino gehiago ezagutzeko eta sortzeko gai izan behar du euskararako prozesadore morfologiko batek (Alegria 1995; Alegria *et al*. 1996). Emankortasun morfologiko hori adjektiboen kasuan are handiagoa da, gradu-flexioa dela eta lau aldiz handiagoa baita.

Egoera honen aurrean, esan bezala, ez zen erraza izan aurkitzea morfologia konplexu bati aurre egiteko gaitasuna izango lukeen formalismo bat. Zorionez, Lauri Karttunen eta Kimmo Koskenniemmi finlandiarrek proposamen interesgarriak eginak zituzten, eta bigarrengoaren tesian oinarritu ginen deskribapen morfologikoaren osagaiak modelizatzeko eta prozesatzeko programak idazteko orduan.

Argitaratutako lehen artikuluan (Agirre *et al*. 1991), Elhuyar aldizkari historikoan, formalismo horren ezaugarri nagusiak azaltzen ziren:

— Eredu orokorra da, edozein hizkuntzari aplika dakiokeena.
— Ezagutza linguistikoa eta algoritmoa bereizi egiten ditu eta, ondorioz, programa berak edozein hizkuntzatarako balio dezake.
— Baliagarria da hitzen analisi morfologikorako zein hitz-sorkuntzarako.
— Analizatu edo sortuko den hitzaren azaleko maila eta hiztegiko sisteman (sistema lexikoan) errepresentatzen den maila lexiko edo sakonekoa argi eta garbi bereizten ditu.

Programazioari dagokionez, zoritxarrez zeuden bi inplementazioak ez ziren libreak, eta gureari ekin genion C lengoaia erabilita (PCetan eraginkortasunez exekutatu ahal izateko) (Alegria 1995). Gaur egun martxan dagoen arren, bi inplementazio aurreratu eskuratu genituen hurrengo hamarkadetan: Xerox enpresak lizentziatutakoa garai batean, eta Mans Hulden-ek software librean sortutako Foma[3] azken urteotan (Alegria *et al*. 2009). Software librearekin posible izan da baliabide urriko hizkuntza askotarako halako prozesadoreak garatzea (gaur egun 40tik gora izango dira).

Ikuspuntu linguistikotik interesgarriena eta zailena formalismoaren osagaiak ondo definitzea izan zen. Bi osagai nagusi ditu (Agirre *et al*. 1991, Alegria 1995, Urkia 1997, Alegria & Urkia 2002): sistema lexikoa eta erregela morfofonologikoak.

---

[3] Foma Mans Hulden-ek sortutako egoera finituko automatak eta transduktoreak erabiltzeko kode ireki eta libreko tresna multzoa da.

Sistema lexikoa. Sistema lexikoan morfema-multzoa definitzen da, morfemen artean egon daitezkeen kateamenduen arabera sailkapena eginez (morfotaktika edo paradigmak deitzen zaio atal horri). Definitutako azpilexiko horiez gain, lemen eta aurrizki/atzizkien sekuentzia posibleak arautzen dituzten jarraitze-klaseak ere definitzen dira, azpilexiko multzoak direnak.

Morfemak definitzean «bi mailak» definitu behar dira, azaleko forma (*zakur*, *da* zein *Gasteiz* lemak, edo *-ri* eta *-la* atzizkiak, esaterako), eta sakoneko forma informazio morfologikoarekin. Horrela *-ri* atzizkiari dagokion informazioa datibo mugagabea da, eta sakoneko informazioa *Ri* da. *R* hori marka bat da (bi motatakoak dira, morfofonema eta hautapen-markak; aipatu den kasuan, morfofonema da), adierazteko bokal ondoren *r* hori gauzatu behar dela, baina kontsonante ondoren ez. Adibidez, *zakur* leman azken *r*-a markatzen da (*zakuR* esaterako) *r* gogorra dela adierazteko, eta *da* formaren kasuan marka bat jartzen da amaieran adierazteko *a* hori *e* bihur daitekeela *-la* eta *-n* moduko atzizkien aurrean.

Erregela morfofonologikoak. Transformazio morfofonologikoen eraginez sakoneko eta azaleko mailen artean sortzen diren diferentziak adierazteko erregelak erabiltzen dira. Erregela horiek automata edo transduktore bihurtzen dira, normalean konpiladore baten bitartez (hasiera hartan eskuz egiten genuen). Konpiladorearen arabera aldatzen dira erregelak idaztean erabiltzen diren sinboloak (espresio erregularretatik gertu, beti), eta *r*-aren kasuan, adibide sinplifikatu bat emanez, hauexek lirateke aipatutako Foma konpiladorerako erregelak:

```
R -> r r || _ «+» Bokal ; # R gogorra bikoizten da bokal aurrean
Adibidea: zakuR+a -> zakurra

R -> r || «+» _ Bokal;   # R mugagabeko marka r da bokal aurrean
Adibidea: ume+Ri -> umeri
```

Garapen hura talde-lana izan zen, lan handia eta emankorra, hainbat argitalpen, batzuk nazioartekoak (Alegria *et al*. 1996), eta bi tesi sortu baitziren, aipatutako Alegriarena eta Urkiarena. 60.000 sarrera baino gehiago sartu ziren datu-basean (EDBL, Euskararen Datu-Base Lexikala) eta hogeita hamar bat erregela konplexu sortu ziren. Urteetan zehar deskribapena, lexikoa batez ere, aberastuz eta arau berrietara egokituz joan da, eta gaur egun ere euskararen prozesaketarako oinarrizko elementua da. Gainera, taldeko hurrengo tesietarako eta aplikazioetarako oinarri garrantzitsua izan zen. Adibidez, Itziar Adurizek, hizkuntzaren aldetik eta Nerea Ezeizak, alde informatikotik, beren tesietan (2000 eta 2002, hurrenez hurren), urrats bat gehiago eman zuten anbiguotasun morfosintaktikoa ebatzi ahal izateko, hitzak analizatu ondoren analisi posibleen artean erabakitzeko (adibidez, *zuen* forma aditz laguntzailea edo izenordaina den erabakitzeko).

Bestalde, aukera izan dugu aholkulari gisa aritu eta, beste hizkuntzetako prozesadore morfologikoetako garapenean laguntzeko, morfologia konplexuko eta baliabide urriko hizkuntzetan batez ere. Zehazki, bi mailatako morfologia edota Foma honako hizkuntza hauetan baliatu dute: kitxua, aimara, guarani, nahuatl eta maputxe.

3. APLIKAZIOAK: XUXEN, ITZULPEN AUTOMATIKOA, BERTSOGINTZA ETA BESTE

Analizatzaile morfologiko izenarekin jende gutxik irudikatzen du zenbat aplikazio sor daitezkeen. Baina oinarri horrekin oso tresna interesgarriak gara daitezke, batez ere aukeratutako formalismoarekin, non analisiaz gain sorkuntza ere lor baitaiteke. Asko dira oinarri horretatik egin diren erabilpenak. Hona hemen horietako batzuk:

— Zuzenketa ortografikoa: hitz batek analisia badu zuzentzailearen hiztegiaren parte da, bestela ez (eta azpimarratzen da). Proposamenak sortzeko modulu bat gehitu behar da, eta horretarako sorkuntza morfologikoak laguntzen du.
— OCR (Optical Character Recognition): zuzenketa ortografikoaren parekoa da, baina zuzenketa egiteko karaktereen formarekin lotutako erregelak gehitu behar dira, eta akatsak automatikoki ordezkatzeko aukera.
— Hitzen normalizazioa: testu historikoetan, dialektaletan edo sare sozialetan maiz erabiltzen diren aldaerak identifikatzea eta dagozkien forma estandarrekin lotzea funtsezkoa da halako testuetan (Etxeberria 2016). Erregela fonologiko eta lexiko gehigarriekin azkar gara daiteke halako sistema bat, nahiz eta ataza horretarako, batzuetan, inferentziak emaitza hobeak eman, gero ikusiko dugunez.
— Ikasleentzako laguntzak. Jokoak sortzeko, hiztegi-kontsultak errazteko edo ariketak automatikoki sortzeko tresna interesgarria da. *Nor-nori-nork* taula lantzeko joko bat egitea oso sinplea da, adibidez; edo *dagozkion* moduko aditz-forma jokatuak hiztegian bilatzen laguntzeko (halakoak, noski, ez baitaude zuzenean hiztegian).
— Itzulpen automatikoan laguntzeko. Ezagutzan oinarritutako ereduan analisi eta sorkuntza morfologikoak ezinbesteko elementuak dira; nahiz eta, gero ikusiko dugunez, gaur egungo sistema eraginkorrenetan beste modu batez egiten den.
— Bertsotan ikasteko edo automatikoki osatzeko elementu garrantzitsua da, errima egiaztatzea edo sortzea nahiko erraz gertatzen baita.

Azken bi aplikazioak eta zuzenketa ortografikoarena aztertuko ditugu apur bat sakonago ondoko azpiataletan.

3.1. **Xuxen**

Xuxen zuzentzaile ortografikoa (https://eu.wikipedia.org/wiki/Xuxen) izan zen analizatzaile ortografikoaren lehen aplikazioa, eta bai arrakastatsuena ere. 1994an sortutako lehen bertsioa, EHUko Ixa taldearen, UZEIren eta Baionako Hizkia enpresaren artean merkaturatua, aparteko programa bat zen, testu-editoreetatik independentea. PC edo Mac batean MS-Word edo WordPerfect-ekin testua editatu eta amaitu ondoren zuzentzea errazten zuen formatu-kontu guztiak errespetatuz. Hasieran garestia zen, eta nagusiki argitaletxeetan eta irakaskuntzan erabili zen. Euskara batua finkatzeko berebiziko tresna izan da. Izan ere, esan dezakegu Xuxen laguntza handikoa izan dela azken bi hamarkada hauetan, euskara idatziaren normalizazioan, Euskaltzaindiaren hiztegi-arauak eta gomendioak kaleratu eta gutxira normaltasunez onartuak eta erabiliak izan zitezen. Kontuan izanik, gure ustez, batuaren etorkizuna euskararen etorkizuna dela (Aduriz *et al*. 2018), aplikazio giltzarria dugu normalizazioaren bidean.

Euskarak duen konplexutasun/aberastasun morfologikoa dela eta, hitz bat ondo dagoen ala ez erabakitzeko ezin zen, beste hizkuntza batzuetarako egiten zen bezala, hitz-zerrenda batean oinarritu; lema batetik abiatuta sor daitezkeen zilegi diren formak asko direnez, zerrenda izugarria izango bailitzateke. Hori dela eta, egindako hurbilpena teknologia aldetik oso berritzailea izan zen, zuzentzaile gehien-gehienak hitzetan oinarritzen baitziren, eta ez morfologian (Aduriz *et al*. 1997). Esan bezala, analisirik ez zuen edozein hitz susmagarritzat jotzen zuen, eta, ondorioz, azpimarratu egiten zuen. Gainera, bi ezaugarri ahaltsu zituen:

— Proposamenak egiteko ohiko bideez gain (hitzaren karaktere bat gehitzea, aldatzea edo ezabatzea) «ohiko akatsak» hartzen zituen kontuan. Horretarako erregela morfofonologiko eta lexiko-sarrera berriak idatzi ziren. Adibidez, lexikoan *haundi* lema gehitu zen dagokion *handi* estandarrarekin lotuta. Erregelei dagokienez, hainbat fenomeno aurreikusten ziren: *h*-a ez jartzea, *x*/*j* aldaketa, *o*/*u* nahastea lemaren amaieran, erdararen eraginez *v* edo *c* idaztea, esate baterako.
— Erabiltzailearen hiztegia. Hitz-forma berriak gehi zitezkeen, eta, horri esker, aurrerantzean hitz-forma hori ez ezik, haren flexio osoa ezagutuko zen, beste ezer egin gabe. Erabilpen teknikorako oso garrantzitsua den ezaugarria, eta zoritxarrez, gaur egungo sistemetan integratzen ez dena.

Arrakastatsua izan arren, ortografia testua amaitu ondoren egiaztatu behar zen, ez baitzegoen integratuta testu-editoretan, Word eta Word Perfect-eko formatua errespetatzen bazuen ere. Hori dela eta, eskaera handia zegoen integratuta egon zedin ohiko testu-editoreetan, batez ere argitaletxeetan eta administrazioan.

Microsoftekin akordioa bideratu eta gero, hala gertatu zen; eta hurrengo urteetan, Eleka-Elhuyarrek egindako lanari esker, beste enpresa batzuetako aplikazioetan integratzen joan da (http://xuxen.eus/eu/deskargatu). Software librearen bertsioa ere lortu zen, nahi baino beranduago, morfologikoki konplexu diren hizkuntzetarako zeuden zuzentzaile estandarrak desegokiak zirelako (Alegria *et al*. 2011). Hungariatik proposatutako *Hunspell* estandartzat hartu zenean, berriz, dena erraztu zen.

Hizkuntzaren aldetik, osatuz eta eguneratuz joan da, eta zuzentzaile ortografiko berriak sortu dira euskararako: Hobelex (UZEIk kaleratua eta ezagutza berean oinarritua) edo multinazionalek gaur egun, zorionez, eskaintzen dituztenak (Microsoftena dagoeneko ez da Xuxen, eta Google Docs-ek berea du).

Bizkaierarako ere bertsio bat egin genuen enkarguz Eleka-Elhuyarrekin batera (Alegria *et al*. 2010), Labayrurekin lankidetzan. Datu-basean forma berriak sartu ziren lehenetsitako estandarrekin lotuta (*deutso-dio*, *uri-hiri*, *gitxi-gutxi*...) eta hautazko erregela morfofonologiko berriak gehitu ziren morfemetan eta azalean aldaketak eragiteko (*-sino ->sio*, *-e->-a*, *-ea->-a*, esate baterako, *telebisino*/*telebisio*, *laba*/*labe* eta *alabea*/*alaba* onartu eta proposatu ahal izateko). Arrakasta mugatua izan da, bizkaiera «estandarizatua» ez dagoelako ondo definituta, besteak beste.

3.2. **Beste aplikazioak: itzulpen automatikoa eta bertsolaritza**

Morfologia funtsezko osagaia dugu itzulpen automatikoan ere, ezagutzan oinarritutako teknologian erabiltzen denean, batez ere. Gaztelania-euskara (es-eu) bikoterako garatu genuen Matxin itzultzailean (Mayor 2007), deskribatu dugun prozesadore morfologikoa berrerabili genuen, baina analisiari baino gehiago sorkuntzari atera genion etekina.

Gaur egun, itzulpen automatikoan emaitza onenak sistema neuronalekin lortzen dira, baina horretarako testu-masa handiak behar dira. Egoera horretan ez dauden hizkuntzen kasuetan, ezagutzan oinarritutako hurbilpena erabiltzen da oraindik.

Bertsolaritzari dagokionez, prozesadore morfologikoak erraztu zigun bertsogintzako hainbat ataletan errima egiaztatzea eta hitz errimatuak sortzea, batez ere, eta neurria kontrolatzea.

Lehen urrats batean (Arrieta *et al*. 2001), hitz errimatuak gure inplementazioa erabiliz sortzea izan zen helburua, hitzak eta morfema-sistema osoa atzekoz aurrera bihurtu genituen, horrela posible eginez sorkuntza abiatzea amaieratik hasita.

Bigarren urrats batean (Agirrezabal 2012), aurreko elementua asko sinplifikatu genuen, garaiko teknologiak ahalbidetzen zuelako konpilazio automatikoan

reverse eragilea erabiltzea. Gero, eta teknologia berarekin silaba-banatzaile edo -kontatzaile batekin osatuta, bertso-bilketa bat egin zen eta bertso-eskoletako ikasleei begirako aplikazio bat garatu zen (arbel digitala). Eta azkenik, Bertsobot izeneko proiektuari hasiera eman genion, non epe luzeko helburua zen adimen artifizialeko erronka handi bati ekitea: bertso-sorkuntza.

2012rako eginak zeuden hurbilpen oinarrizkoenak: 1) Memoria hutseko lana, 2) Corpusen gaineko bilaketa arrunta, eta 3) Corpusen gaineko bilaketa adimenduna. Horrez gain, sorkuntza librean urrats batzuk eman ziren.

Hortik aurrera, Bertsobotek garapen interesgarria izan du EHUko beste talde bitan: robotikan (Astigarraga 2019) eta kantaeraren sorkuntzan (Sarasola 2020). Aurrerapen horiekin guztiekin, eta adimen artifizialean testu-sorkuntzan gertatzen ari diren aurrerapen izugarriekin ez litzateke harritzekoa laster Bertsobot errealitatea izatea.

4. MORFOLOGIAREN INFERENTZIA

Aipatutako Koskenniemmiren eta Karttunenen ekarpenetatik abiatuta egoera finituko morfologia asko garatu zen, eta horrek euskarari eta morfologikoki konplexu diren beste hizkuntza askori lagundu zien irtenbide konputazionala bideratzen. Ildo horretatik, Karttunen-ek esan zuen «from a computational point of view morphology was a solved problem» (Karttunen 2007). Hizkuntza semitikoetarako ere irtenbidea bilatu zitzaion bi mailako ereduari mailak gehituz.

Baina, zer gertatzen da hizkuntza baten lexikoa eta arau morfologikoak ezezagunak direnean? Edo, ezagutu arren, garapen konputazional azkar bat egin nahi dugunean? Hor aukera interesgarria izan daiteke morfologia inferitzea. Azken bi-hiru hamarkadetan funtsezko ikergaia bihurtu da, bi urratsetatik igaroz: eredu estatistikoa eta ikasketa sakona.

4.1. **Eredu estatistikoa**

Testu hutsetatik abiatuta hitzak osatzen dituzten morfemak inferitzea proposatu da. Metodoen oinarria testuetako hitzen hasierak eta, batez ere, amaierak estatistikoki aztertzean datza, aurrizkiak, artizkiak eta atzizkiak identifikatuz, eta bide batez stem-ak (azaleko lemak edo sasi-lemak). Azaleko mailan lan egiten duten metodo hauek, beraz, forma kanonikoen ordez hitz-zatiak identifikatzen dituzte, aldaketa morfofonologikoak kontuan hartu gabe (horregatik lemak baino stem-ak identifikatzen dira).

Arrakasta handienetako metodoen artean Morfessor aipa daiteke, flexio handiko hizkuntzetarako egokia delako. Ixa taldeko tesi batean (Etxe-

berria 2016) tresnaren funtsa azaltzen da eta euskararen testu historikoei aplikatzen zaie (garaiko osagai morfologikoak identifikatu nahian). Etxeberriak (2016: 141) dioenez:

> ...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 (M), beraz, helburua da lexikoi onena bilatzea sarrerako corpusa segmentatzeko.

Eta adibideak ere ematen ditu:

> Horrela lortu ditugu, esaterako, *amoreakgatik* → amore + ak + gatik edota *elkharri* → elkhar + ri moduko segmentazioak...

Itzulpen automatiko neuronalean oso hedatua dagoen BPE metodoa (Senrich *et al*. 2015) ere saiatzen da atzizki adierazgarriak lortzen hitz ezezagunetan, eta bide batez hitz horiei dagozkien lemak asmatzen. Euskararako ere erabili izan da eta emaitza onak ematen ditu.

Baina Etxeberriaren tesian helburu nagusia testu historikoen normalizazioa da, eta horretarako forma historiko bakoitza gaur egungo idazkeran dagokion forma estandarrekin lotzen da, testu historikoen kontsulta errazteko asmoz (Estarrona *et al*. 2020). Hori lortzeko, erregela morfofonologikoak inferitzen dira, baina ezagutzan oinarritutakoetan ez bezala sortzen diren transduktoreak eredu probabilistiko batean konbinatzen dira. Horrela, *amoreakgatik* lotzen da *amoreengatik* forma estandarrarekin eta *elkharri, elkarri* formarekin, eta horretarako *ak*/*en* aldaketa ikasten da, baita *h*-ren galera ere zenbait testuingurutan. Tesian frogatzen da helburu horretarako ezagutzan oinarritutako erregeletan baino emaitza hobeak lortzen direla eredu estatistikoekin, betiere eskuz prestatutako bikote (aldaera/estandarra) zerrenda batetik abiatuta.

Teknika estatistiko horiek testu historikoak edo dialektalak prozesatzeko asko erabiltzen dira, baita OCR (*Optical Character Recognition*) akatsak zuzentzeko edo sare sozialetako testuak normalizatzeko ere.

4.2. **Ikasketa sakona morfologian aplikatuta**

Aipatu den bezala, azken urteotan morfologia konputazionalaren ikerkuntzan lerro nagusia morfologia inferitzea izan da, hau da, adibideetatik paradigmak eta aldaera fonologikoak orokortzea, salbuespenak ere bereiziz. Horretarako bide emankorrenetako bat *shared task* edo partekatutako atazak izan dira. Horietan erronka bat planteatzen da modu irekian, eta mundu osoko edozein ikertzaile edo ikertzaile taldek parte har dezake solu-

zio onena bilatzen. Horrelako erronketan ohikoak izaten dira ikasketa sakonean oinarritutako sistemak, eta morfologia konputazionala ez da salbuespena izan.

Horren adibidea SIGMORPHON ataza partekatua da. Hasieran helburua datuetatik abiatuta sortzaile morfologikoak garatzea izan zen eta aurrerago bestelako atazak gehitu zituzten. Morfologiara aplikatutako ikasketa sakonaren gorakada nagusia LMU-MED (Kann & Schütze 2016) sistemarekin hasi zen, 2016an (Cotterell *et al*. 2016), non Kann-en sistema hoberena izan zen beste guztiekin alderatuta, eta sistema neuronalen egokitasuna agerian geratu zen ataza honetarako. Hurrengo urteetan ataza partekatuetan erabili ziren sistema gehienek modu batean edo beste batean ikasketa sakoneko ereduren bat erabili dute, dela LSTMdun sare errekurrenteak edo gaur egun puri-purian dauden *Transformer* ereduak. Ataza horietan frogatuta geratu da datu nahikoa eskura izanez gero, sortzaile morfologikoak sortzea erlatiboki erraza dela.

Azken urteotan, baina, interesa piztu dute datu gutxirekin sortutako sistemek eta honako galderei erantzuten saiatu dira: zenbat aditz jokatu behar ditugu aditzetarako sortzaile morfologiko bat sortzeko? Horrelako galderak planteatu zituzten 2018ko sorkuntza morfologikoko ataza partekatuan, non 100 hizkuntzatarako sortzaile morfologikoak sortzea zen ataza. Datu gutxi zuten hizkuntzetan enfasia egitearren, hiru modalitatetan lan egiteko aukera eskaini zuten: txikia, ertaina eta handia; entrenamendurako, berriz, 100, 1.000 edo 10.000 datu-instantzia ematen ziren. Gai hauetan egiten diren aurrerapenak baliabide urriko hizkuntzetarako baliagarriak izan daitezke. 100 hizkuntza horien artean euskara zegoen eta sistema onenak, % 98,90ko asmatze-tasa izan zuen datu-multzo handienarekin, % 88,10ekoa datu-multzo ertainarekin eta % 13,30ekoa txikienarekin.

Orokorrean, urtez urte emaitzak hobetuz badoaz ere, badaude oraindik mugak. Egoera finituko makinetan (FSM) oinarritutako sistemak oso eraginkorrak dira, teorikoki modu oso konpaktuan sistema oso konplexuak kodetzen baitituzte. Gainera, jakintza linguistikoa eskuz kodetuta dagoenez, arazo bat baldin badago, nahiko erraza da arazoaren muina aurkitu eta konpontzea. Sistema neuronalen arazo bat, berriz, gaur egun puri-purian dagoena, irakurgarritasun edo interpretabilitatea da. Sare neuronaletan oinarritutako sistema guztiek sarrerako edozein datu emanik emaitza bat itzultzen dute, eta askotan irteera hori nola eta zergatik itzuli duten interpretatzea ez da erraza. Horregatik, maiz sare neuronalok kutxa beltzak direla esan ohi da.

Bestalde, ematen du ikasketa sakoneko hurbilpenean, oro har, informazio linguistikoak ez duela toki handirik eta, corpus erraldoiak izanez gero, emaitza ona itzultzen duela, askotan, informazio linguistikoa erabiliko bagenu baino hobea, itzulpen automatikoko hainbat kasutan ikusi izan dugun

bezala. Hala ere, badira ahotsak itsu-itsuan ibiltze horrekin oso ados ez, eta informazio linguistikoa oinarrian erabilita sistema neuronalak martxan jarri dituztenak, emaitzak hobetuz morfologian (Nzeyimana & Niyongabo Rubungo 2022).

5. ONDORIOAK

Hogeita hamabost urte luze joan dira, estreinakoz euskararen morfologia konputazionalaren tratamenduari ekin genionetik. Denbora joan da, beraz, eta denbora horretan zehar aurrera egin du morfologia konputazionalak, bilakaera historikoan ikusi dugun bezala, hainbat eredu desberdinen garapenarekin.

Horren guztiaren adibide gisa, Ixa taldean morfologiak izandako ibilbidea ikusi dugu, eta agerian geratu da modulu morfologikoa dagoela hainbat tresna eta aplikazioren oinarrian. Seguru asko, Xuxenek, aplikazio ezagunena izateaz gain, eragin nabarmenena izan du euskara batuaren normalizazioan.

Eta ez da kasualitatea, azkenean, izenburuan aipatzen diren 35 urte horietako estrategia eta ikerkuntza dago atzean. Ixa taldeak hizkuntzaren tratamendu automatikoaren arloan egiten du ikerketa, eta euskararen tratamendua du ardatz: hor daude, besteak beste, oinarrizko tresna eta baliabideak, hainbat eta hainbat aplikaziotan euskara presente egotea, webean kontsultagarri dauden corpus eta hiztegiak, bertako eta nazioarteko kongresu eta aldizkarietan egindako argitalpenak. Ixa taldeak mantentzen ditu Xuxenen oinarri den hiztegia —EDBL— eta Xuxenen motorra; izan ere, euskarazko zuzentzailea ez baita hiztegi huts batean oinarritzen, segmentatzaile morfologikoa behar baitu hitz bat zuzen idatzita dagoen ala ez erabakitzeko.

Ondorio nagusi gisa, puntu hau azpimarratu nahi dugu: euskararen etorkizunari begira, oro har, eta bereziki, hizkuntzak ingurune digitalean dituen erronkei begira, funtsezkoa da hizkuntza-teknologien arloak duen garrantzi estrategikoa azpimarratzea eta inplikaturiko eragile guztien indarrak biltzea.

Esker onez

Lan honen bukaerara iritsi gara xuxen-xuxen eta, esker-hitzak idazteko unea dugu. Hitz egokirik ezin aurkitu Miren Azkarateri gure esker on guztia adierazteko.

Sarreran aipatu bezala, Ixa taldearen lanen baitan, morfologia konputazionaleko lehen tesia berak zuzendu zuen: Miriam Urkia euskaltzainarena, hain zuzen ere (Urkia 1997).

Horren ondoren, halaber, hainbat eta hainbat ixakideren tesietan egon da epaimahaian. Hauexetan:

> Itziar Aduriz (Aduriz 2000), Mikel Lersundi (Lersundi 2005), Maxux Aranzabe (Aranzabe 2008), Ruben Urizar (Urizar 2012), Aitziber Atutxa (Atutxa 2014), Mikel Iruskieta (Iruskieta 2014) eta Uxoa Iñurrieta (Iñurrieta 2019).

Tesi horietako gaiak askotarikoak dira. Morfologia bai, Adurizenean, adibidez, baina sintaxia ere bai: Aranzaberenean edo Atutxarenean, adibidez. Mirenek hain gogoko duen eratorpenetik pasatuz (Lersundirenean), edo hitzkonbinazioen gaia ukituz (Urizarrenean eta Iñurrietarenean).

Honek erakusten du Miren Azkarateren ezaugarri handienetako bat: euskara maite du eta, euskaltzalea denez ezinezkoa da aurkitzea haren interesekorik ez den euskarari buruzko gairik.

Mila esker.

6. BIBLIOGRAFIA

Aduriz, Itziar, Miriam Urkia, Iñaki Alegria, Xabier Artola, Nerea Ezeiza & Kepa Sarasola. 1997. A spelling corrector for Basque based on morphology. *Literary and Linguistic Computing* 12(1). 31-38. http://ixa.ehu.eus/node/3350.

Aduriz, Itziar. 2000. *EUSMG: Morfologiatik sintaxira murriztapen gramatika erabiliz. Euskararen desanbiguazio morfologikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak*. PhD Thesis. UPV/EHU. http://edtb.euskomedia.org/id/eprint/5790.

Aduriz, Itziar, Iñaki Alegria, Xabier Artola, Arantza Díaz de Ilarraza & Kepa Sarasola. 2011. Teknologia garatzeko estrategiak baliabide urriko hizkuntzetarako: euskararen eta Ixa taldearen adibidea. *Linguamática* 3(1). 13-31. http://ixa.ehu.eus/node/3873.

Aduriz, Itziar, Iñaki Alegria, Izaskun Aldezabal, Xabier Artola, Arantza Díaz de Illarraza, Nerea Ezeiza, Kepa Sarasola & Ruben Urizar. 2018. Euskara (batua) ingurune digitalean: bidean ikasiXa eta etorkizuneko erronkak. In *Euskaltzaindiaren XVII. Biltzarra (Arantzazutik mundu zabalera)*. Arantzazu. http://ixa.ehu.eus/node/12744.

Agirre, Eneko, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Illarraza, Patxi Goenaga, Montse Maritxalar, Kepa Sarasola & Miriam Urkia. 1991. Bi mailatako morfologiaren euskararako egokitzapena. *Elhuyar* 17. 6-14. https://www.ixa.eus/node/3365.

Agirre, Eneko, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Montse Maritxalar, Kepa Sarasola & Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. In *Third Conference on Applied Natural Language Processing*, 119-125. https://aclanthology.org/A92-1016/.

Agirrezabal, Manex. 2012. *BertsoBot: lehen urratsak*. Master Amaierako Lana. UPV/EHU. http://hdl.handle.net/10810/11253.

Alegria, Iñaki. 1995. *Euskal morfologiaren tratamendu automatikorako tresnak*. Doktore Tesia. UPV/EHU. https://www.euskadi.eus/appcont/tesisDoctoral/PDFak/Inaki_Alegria_TESI.pdf.

Alegria, Iñaki & Miriam Urkia. 2002. *Morfologia konputazionala: euskararen morfologiaren deskribapena*. Udako Euskal Unibertsitatea. http://ixa.si.ehu.es/node/3341.

Alegria, Iñaki, Xabier Artola, Kepa Sarasola & Miriam Urkia. 1996. Automatic morphological analysis of Basque. *Literary and Linguistic Computing* 11(4). 193-203. http://ixa.si.ehu.es/node/3355.

Alegria, Iñaki, Garbiñe Aranbarri, Klara Ceberio, Gorka Labaka, Bittor Laskurain & Ruben Urizar. 2010. A morphological processor based on foma for Biscayan (a Basque dialect). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. https://aclanthology.org/L10-1096/.

Alegria, Iñaki, Izaskun Etxeberria, Mans Hulden & Montse Maritxalar. 2009. Porting Basque morphological grammars to foma, an open-source tool. In *International Workshop on Finite-State Methods and Natural Language Processing*, 105-113. Springer. https://doi.org/10.1007/978-3-642-14684-8_11.

Aranzabe, Maxux. 2008. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. PhD Thesis. http://edtb.euskomedia.org/id/eprint/5304.

Arrieta, Bertol, Iñaki Alegria & Xabier Arregi. 2001. An assistant tool for verse-making in Basque based on two-level morphology. *Literary and linguistic computing* 16(1). 29-43. http://www.ixa.eus/node/3343.

Astigarraga, Aitzol. 2019. *Bertsobot: gizaki-robot arteko komunikazio eta elkarrekintzarako portaerak*. Doktore Tesia. UPV/EHU. http://hdl.handle.net/10810/24049.

Atutxa, Aitziber. 2014. *Aditzen inguruko informazio lexikala eta anbiguotasun sintaktikoen ebazpena*. PhD Thesis. http://www.ixa.eus/node/4130.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner & Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology,* 10-22. https://aclanthology.org/W16-2002/.

Estarrona, Ainara, Izaskun Etxeberria, Ricardo Echepare, Manuel Padilla & Ander Soraluze. 2020. Sintaktikoki etiketatutako euskarazko corpus historikoa eraikitzen. In *Fontes Linguae Vasconum 50 urte: ekarpen berriak euskararen ikerketari*, 237-251. Gobierno de Navarra. http://www.ixa.eus/node/13186.

Etxeberria, Izaskun. 2016. *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz*. Doktore Tesia. UPV/EHU. https://www.ixa.eus/node/8336.

Ezeiza, Nerea. 2002. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua*. Doktore Tesia. UPV/EHU. http://www.ixa.eus/node/4123.

Iñurrieta, Uxoa. 2019. *Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala*. Doktore Tesia. UPV/EHU. http://hdl.handle.net/10810/41483.

Iruskieta, Mikel. 2014. *Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean*. Doktore Tesia. UPV/EHU. http://ixa.si.ehu.es/node/4126.

Katharina, Kann & Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 62-70. https://aclanthology.org/W16-2010/.

Karttunen, Lauri. 2007. Word play. *Computational Linguistics* 33(4). 443-467. https://direct.mit.edu/coli/article-abstract/33/4/443/1958.

Lersundi, Mikel. 2005. *Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak*. Doktore Tesia. EUPV/EHU. https://www.ixa.eus/node/4119.

Manterola, Julen. 2008. Is Basque an agglutinative language? A proposal for the diachrony of nominal morphology. In *Proceedings of the Basque Studies Symposium May 14th, 2008*. TINTA, Research Journal of Hispanic and Lusophone Studies, Department of Spanish and Portuguese, UCSB. https://artxiker.ccsd.cnrs.fr/artxibo-00350108.

Mayor, Aingeru. 2007. *Matxin, erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Doktore Tesia. UPV/EHU. http://ixa.si.ehu.es/node/4103.

Nzeyimana, Antoine & Andre Niyongabo. 2022. KinyaBERT: a Morphology-aware Kinyarwanda Language Model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 5347-5363. Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2203.08459.

Salaburu, Pello. 2011. Morfologia, hitzaren gramatika. In *Sareko Euskal Gramatika (SEG)*. https://www.ehu.eus/seg/morf/4.

Sarasola, Xabier. 2020. *Application of singing synthesis techniquest to bertsolaritza*. Doktore Tesia. UPV/EHU. http://hdl.handle.net/10810/50503.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2015. *Neural machine translation of rare words with subword units*. https://arxiv.org/abs/1508.07909.

Urizar, Ruben. 2012. *Euskal lokuzioen tratamendu konputazionala*. Doktore Tesia. UPV/EHU. http://ixa.si.ehu.es/node/4134.

Urkia, Miriam. 1997. *Euskal morfologiaren tratamendu informatikorantz*. Doktore Tesia. UPV/EHU. http://www.ixa.eus/node/4110.