Hizkuntz datuen formalizazio numerikorantz

Gotzon Aurrekoetxea

Aurrekoetxea1996b
Argitaratua in: *ASJU,* XXX-2, 1996, 455-467

HIZKUNTZ DATUEN FORMALIZAZIO NUMERIKORANTZ

Gotzon Aurrekoetxea
Euskal Filologia
EHU

1. Sarrera

Hizkeren arteko distantziak neurtze eta kuantifikatze bidean nire aspaldidaniko kezka bat plazaratzera nator artikulu honetan. Bere momentuan irteera bat aurkitu banuen ere oraindik orain ez dut biribiltasun osozkoa denik uste. Helburua, hizkuntz datuak prozedura automatizaturako prest uztea.

Dialektologian hizkeren arteko desberdintasuna kuantifikatzean sortzen diren arazoetan kokatzen da kezka hau. Beraz, nire kezkak jenderarteratu, eztabaida piztu eta gaiak dituzkeen ikuspuntu desberdinetarik, agian, besteren batek irtenbide egokiagoa eman diezaiokeelako esperantzan[1].

Hizkeren arteko distantziak kuantitatiboki neurtzeko beharra begi bistakoa da. Egun ez da nahikoa hizkera bat halako azpieuskalki edo euskalkikoa dela esatea, hizkera batetik bestera "jauzi"a dagoela adieraztea, alegia. Horrelakoak ia egunero entzuten ditu dialektologian ordu batzuetan pausatu den edonork. Desberdintasuna zertan mamitzen denetik, desberdintasun hori kuantifikatzera heldu behar dugu; euskalkietako hizkera guztiak hartuz euren arteko berdintasun eta desberdintasunak zenbatzeraino heldu beharra dago. Eta ez dut behinola Mitxelenak aipatu zituen hitzak errepikatu baino egin[2]

Horretarako, batetik datu multzo handia bildu da jada gure artean proiektu desberdinetan. Bestetik, bitarteko ahaltsuak ageri dira analisi estatistiko automatizatuak egiteko.

---

[1] Izatez lan hau lankidetzarako deia da; batez ere estatistikan edota inguruko disziplinetan metodo estatistiko eta sailkatzaileak erabiltzen ohituak diren ikertzaileei egindakoa.

[2] *Sobre el pasado de la lengua vasca*, 1964 (orain in *Sobre historia de la lengua vasca,* ASJUren gehigarriak, 1988, 3. or.

Hizkeren arteko distantziak kuantitatiboki neurtzeko lanak urrats desberdinetan gauzatzen dira: hizkuntz ezaugarriak aukeratu ostean, ezaugarri bakoitza hornitzeko hizkuntz datuak aukeratu, datuen formalizazioa egin, hizkuntz datuoi kodaketa numerikoa ezarri eta prozedura sailkatzaileak abiatu behar dira.

Hizkuntzalaritzan ikerketaren abiapuntua hizkuntz datuak direla esatea ezer ez esatea bezain hutsala da. Hizkuntz datuak diren bezalakoak dira eta horrela hartu behar dira, inolako manipulazio interesaturik egin barik. Hizkuntz datuak izango dira abiapuntu lan honetan ere. Euren egoeraz arduratuko naiz eta euren eraldaketan jaso daitezkeen aldaketetan zehaztasun guztiak gordetzea dute helburu lerro hauek.

Baina arazoa beste honetan datza: hizkuntz datuak ez dira aproposak azterketa sailkatzaileak edoeta estatistikoak burutzeko. Areago oraindik, ikerketan bitarteko automatizatuak erabiltzen direlarik. Hizkuntz datuak aldrebesak eta deserosoak gertatzen dira ordenadoreentzat eta gaindiezinezko arazoak sortzen dira programa informatikoak erabiltzen direnean.

Horretarako, ezinbesteko dugu hizkuntz datu horiek formalizazio numerikoen bidez zenbakietara itzultzea, bihurtzea. Hizkuntz datu bakoitza zenbaki jakin batez ordeztu behar da. Orduan izango dira prest edozelango azterketa estatistiko edo metrologikoak gauzatzeko.

Nola eraldatu, ordea, hizkuntz datuak zenbakietara? Hori da gakoa. Hizkuntzalarien ikuspuntutik begiratuz gaindiezinezko horma bat agertzen da sarri ikertzailearen mutur-muturrean hizki edo hitzen ordez zenbakiak erabili behar direnean. Eta ondorioz, horrelako ikerketak bertan behera utzi izan dira maiz, sasoi hobeagoen zain. Gutxi dira zailtasun honen aurrean aintzina egitera ausartu diren hizkuntzalariak. Filologietako ikasketetan "hizkuntz estatistika", "hizkuntz matematika" edo "hizkuntzalaritza konputerizatua" moduko eta gaur egun behar-beharrezko deritzodan ikasgaiaren hutsunea nabaria da gure artean.

Egun zientzia guztietan erabiltzen dira estatistika eta metodo sailkatzaileak, hala fisikan, nola geologia, medikuntza, eta beste hainbat eta hainbat disziplinatan. Giza-zientzietan ia bakarrik gelditu da hizkuntzalaritza metodologia hau erabiltzetik kanpo, letrak eta zenbakiak lotzeko halako ezintasuna balego moduan. Guk, ostera, hizkuntzalaritzan metodologia hori erabiltzea aldarrikatzen dugu, areago euskal hizkuntzalaritzan. Hizkuntzalaritza ezin da albo batean gelditu, hizkuntzari buruzko ikerketetan behin baino gehiagotan behar izaten baitira analisi estatistiko eta sailkatzaileak, areago dialektologian, hizkeren ikerketa konparatiboa egin gura denean. Hizkeren arteko urruntasun eta hurbiltasuna, desberdintasuna eta berdintasuna edo antzekotasuna dugunean ikerketaren xede.

Hari honi ekin nahi zaio artikulu honetan. Hizkuntz datuak formulazio matematikoetara eroateko hizkuntzalariak aurkitzen dituen oztopoak aipatu gura dira; edo bestela esanda, hizkuntz datuak azterketa kuantitatiboak jasan arazteko zelan prestatu daitezkeen aztertu gura da.

Jo dezagun badugula aukeraturik hizkerak bereizteko ezaugarri kopurua eta hauetarako beharrezko den hizkuntz datuteria ere. Eta hizkuntz datuak, beharrezko orrazketa guztiak eginda, prest direla. Eta eman dezagun ere, datuen analisia egite orduan eskuz eta banaka egin ordez azterketa automatizatuak egitea deliberatu dugula. Datu kopuru handia izango da eskuartean erabiliko dena. Hartzen duten kopurua dela eta ezinezko gertatzen da hizkeraz hizkera berdintasun /desberdintasunak eskuz eta banaka banaka egitea.

Prozedura automatizatuak egin ordez eskuz egitea deliberatzen bada soberan leudeke hemendik aurrerako urratsak. Baina hauek erabiltzea onartuz gero, zein dira eman beharreko pausuak? Zelan pasatu ahal dira hizkuntz datuak zenbakietara?

2. Hizkuntz datuak

Zelan agertzen da errealitatean hizkuntz masa? Hizkuntz inkestagintzan, datu-bilketan, ibili den edonork ezagutzen du hizkeretan zehar sarri egiten den galdera batek erantzun bakarra izan ordez bi eta hiru ere izan ditzakeela. Hizkuntz errealitatea uste eta, askotan, gura baino nabarragoa da, izatez.

Begi bistakoa izaten da hori lexikoan; ez hain argia beharbada, ezaugarri gramatikala denean. Berba bakoitzak bere eremua izaten ohi du, baina bi eremuren arteko hizkeretan sarri bietako hitzak dira ezagutzen. Horrelako zerbait gertatzen da *arkakuso* eta *ardi* hitzekin, adibidez. Hitz bakoitzak eremu zehatza du, baina bi eremuon mugetan bi hitzak erabiltzen dira, ezagutzen behintzat. Eremu handiak dira zeinetan *amaitu* eta *bukatu* ezagutzen diren; edo *igande/domeka* 'domingo/dimanche', kasu batzuk baino ez ipintzearren. Lexikotik ateratzen bagara berdin gertatuko zaigu: /h/ fonema ezagutzen den eremuko hizkera batzuetan maiztasun handiagoz agertzen dela besteetan baino. Beste horrenbeste, azken kasua aipatzeko, mugagabearen erabilera edo mugatu/mugagabearen arteko oposaketa aztertzen bada; gerta liteke aukeratu diren datu guztiak kontuan hartuz hizkera batean beti betetzea oposaketa eta ondokoan behin ere ez; tarteko aukerak ere izango direla pentsatzea bidezkoa da (batzuetan oposaketak irautea, baina beste batzuetan ez), zalantzarik gabe.

Datuak jasotzean geolinguistikan une bateko erantzuna jasotzen da eta ez etengabeko iharduerakoa. Une horretako burutzapena eta pertsona horrena hartzen da hizkeraren ispilu zuzentzat. Badakit monografietan ibiliak direnak

nekez onartzen dutela prozedura hau edo zailtasunak, bederen, izaten dituztela onartzeko; baina geolinguistikan ohiko jokamoldea izatez gain, ezinbesteko da horrela izatea eremu handiko datuak bildu gura badira.

Hizkuntz errealitatearen beste izaera bat polimorfismoa da. Hizkuntz errealitatea maiz egoera polimorfikoan da, eta honek garrantzi handia dauka hizkeren arteko bereizketak eta desberdintasunak zehaztean, fenomeno jakin bat ezagutzen dutenen artean bi aukera izaten direlako baten ordez *(indar / inder; igon / igo; bezela / bezala* eta antzeko bikoteak ezagutzen diren lekuetan arau fonologiko baten agertzea eta eza aldi berean eta lekuko berarengan leudeke). Hizkuntz aldaketa gertatzen ari den unea izaten baita gehienetan. Aldaketa espazial/horizontala, edo sozial/bertikala izan daiteke (adina, sexua, maila soziala, ikasketak...). Eta agian biak batera.

Masa linguistikoa bildurik denean eta ezaugarriak zehazturik, noiz esan liteke hizkera bat ezaugarri baten jabe dela? Beharbada hala uste arren ez da, ez, txantxetako itauna. Zalantza gutxien agertzen den saila lexikoa da; datu lexikalak aztertzean ez legoke arazo handiegirik hitzak ezagutu ala ez egiten direlako. Ezaugarri gramatikaletan, ostera, ez da hain argi. Inork ez du berbarik esango "-gaz" atzizkia Bizkaiko mendebaldeko edozein hizkeraren ezaugarri bat dela adieraziko banu. Baina (Leintz-)Gatzagako hizkeraren ezaugarri bat dela noraino esan daiteke?

Egia esan, honelako arazoetan datu bilketak itzelezko eragina du. Zenbat eta datuak zehaztasun handiagoz bildu hainbat eta segurantza handiagoa izango da datu horien erabileran. Inkestagileak ezaugarri horren onartze maila eta erabilera (herrikotzat joten den; herriz kanpokoa, baina ezaguna herrian, hango herrian erabiltzen delaren ezagutza...) era zehatzean bildu beharko lituzke.

Har dezagun, adibide gisa, /h/ fonemaren ezagutza aztertzen dihardugula. Eta ezaugarri hau lantzeko datu andana bat bildu dela. "H"dun formak ugariago izan daitezke herri batetik bestera. Batzuk ia posible diren guztietan agertuko dute fonema hori, zaleagoak izango dira, beste batzuk gutxiago. Gauzak horrela, noiz esan liteke hizkera batek ezaugarri hori ezagutzen duela? Ba ote da nahitaezko gutxiengo kopururik edo maiztasunik hizkera batean ezaugarri bat ezagutzen dela onartzeko?

Bestalde, bada arazo bat, beste inon aztertu badut ere, hemen zeharka bederen, aipatu behar dena: ezaugarrien arteko hierarkia edo ponderazioa. Hhizkerak elkarrekin konparatzeko aukeratzen diren ezaugarrien artean hierarkiarik behar ote? Ezaugarri batzuk pisu edo garrantzi handiagoa ote dute beste batzuk baino? Itaun bera egiten genuen duela hiru urte idatzi eta orain argia ikusten duen "Hizkerak banatzeko ezaugarriez" lanean eta orduan emandako erantzuna oraindik baliagarri dela deritzot.

Eta ezaugarri askotan ez badu arazo handirik plazaratzen ere, batzuetan eztabaidagarria dela ere onartzeko moduan nengoke. Demagun arau morfonologikoak ezaugarritzat hartu direla. Ezaguna da arauotako batzuk eman daitezen aurretik beste batzuk eman behar direla. Arau bat gertatzeko aurretik beste bat gertatu behar dela. Holakoetan batak bestea baldintzatzen duen neurrian ez ote duen pisu gehiago ere galdetu izan da inoiz.

Ezaugarrien kopuruaz ere ez da behar beste eztabaidatu gure artean. Eta puntu honetan gutxiengo adostasunik ez den bitartean ondorengo urratsak ematea ez du merezi. Egia esan jadanik eman dira lehen proposamenak arlo honetan[3]. Ez naiz gehiago luzatuko.

Azkenik, non edo non agertu behar da hizkerak banatzeko aukeratu diren ezaugarriak aurkezteko bi bide behinik behin badirela: denak multzo berean nahasturik ematea edo hizkuntz parametroka sailkatzea (ezaugarri fonologikoak, izen morfologiazkoak, aditz morfologiari buruzkoak, sintaktikoak...). Nire aurreko proposamenak hizkuntz parametroka sailkatzearen bidetik izan dira. Horrela parametro bakoitzeko hizkuntz distantzia neurtzeko aukera izango genuke.

3. Kodaketa numerikoa

Nola eraldatu hizkuntz datuak zenbakietara? Hizkuntz datuak kodaketa numerikoetara eraldatzean jarraitu behar den irizpide nagusia ondokoa da: hizkuntz datuetan azaltzen diren zehaztasun eta ñabardura guztiak gorde behar dira, informaziorik gal ez dadin. Irizpide hau oso garrantzitsua da: hizkuntz datuetatik datu numerikoetarako eraldaketan ez dadila xehetasunik gal. Jaso diren hizkuntz datuei dagokien sistema aukeratzeak garrantzi handia dauka. Zeregin honetan joera desberdinak izan dira nire eskuetan eta bat baino gehiago ere erabili izan da.

Joera bat ezaugarren ezagutza, besterik gabe, adieraztea da. Honen arabera /h/ fonemaren araberako hizkeren sailkapena burutzen ari bagara, fonema hori ezagutzen duten hizkera eta ez dutenen arteko bereizketa egingo litzateke. Séguy-ren bidera jarraituz, haiek "1" kodearekin ezagutuko lirateke eta hauek "0"rekin.

---

[3] Ikus G. Aurrekoetxea, 1995, *Bizkaieraren egituraketa geolinguistikoa,* EHU, Leioa; Idem, "Hizkerak banatzeko ezaugarriez" aldizkari honetan; K. ZUAZO, 1998, "Euskalkiak, gaur", *FLV* 78, 191-233.

Baina nahikoa ote da egiaztapen honekin? Ados egongo ote ginateke gauzak horrela utzita? Baliteke bat baino gehiago ados egotea, azken baten hizkera batzuetan, gehiago edo gutxiago, erabiltzen den bitartean, beste batzuetan ez da behin ere erabiltzen eta.

Nire ustez, alabaina, kontuak era zehatzago eta modu sakonagoan landu behar dira. Ezaguna da eta ez dago esan beharrik ere, fonema honen agerrera edo maiztasuna aldatuz doala hizkera batetik bestera edo eremu batetik bestera. Ezagutzen den eremuan maizago agertzen dela ekialdean mendebaldean baino, alegia. Ezagutza enpiriko hori ez ote da aplikatu beharko helburu hizkerak bereiztea duen ikerlanean? Baietz deritzot eta ondorioz datuok kuantifikatzea proposatzen.

Bide honetn, datuen errealitateari hurbilagotik begiratuz, aztertzen dihardugun hizkeretan zehar jasotako erantzunen arabera honako multzoak egin genitzake. Pentsa dezagun lau galderetako datuak erabiltzen direla ezaugarri hau hornitzeko:

a) aukeratu diren berben artetik batean baino ez agertzea delako fonema;
b) bitan agertzea;
c) hirutan agertzea;
d) lauetan agertzea.

Jakina, "h"ren ezagutza-eza aukera alde batera utzi dela hemen, nahiz bere momentuan kontuan izan behar den.

Datuak horrela sailkatuz hizkerak sailkatzeko eta multzokatzeko irizpide zehatzagoak izango genituzke. Ehunekoetan egiteko ere ez legoke arazorik: batean %25ean beteko litzateke, bestean %50ean ...

Honela, datuak nola edo hala kuantifikatzen hasi gara. Behinola gure artean Mitxelenak ere aipatu zuen bidetik[4].

Hizkeren arteko desberdintasuna neurtzeko ere prozedura landuagoa dugu era era finago batez burutzen da.

Eta datuak kuantifikatzen hasiz gero ez ote da hobe eta zehatzago lau hitzetako korpusa bildu ordez korpus handiagoa biltzea?

Errealagoa eta fidagarriago gura bada, beharbada, datu-base handiagoak ere erabili daitezke honelako ezaugarriak hornitzeko. Ondoko galderaren araberako jokabidea hartuz: zein hizkeratan eta zein maiztasunekin agertzen da

---

4 "La fragmentación dialectal: conocimientos y conjeturas", *REL* VI (1976), 309-324; orain *Lengua e historia,* Paraninfo, Madrid, 1985, 73-85, aipamena 81. or.

"h" eskuartean daukagun base osoko datuak hartuz? Emaitzak ehunekoetan emateko aukera izango litzateke, orduan. Hizkera batean %5ean bakarrik ematen da, bestean %10ean, e. a.

Aukeratu diren hizkuntz ezaugarriak hornitzeko hizkuntz datuen kopurua hartzen dena zeingura izanik, hizkuntz datuak datu numerikoetan eraldatzeko aukera bat baino gehiago dago. Hiru dira maizen aipatzen diren prozedurak: sistema hamartarra, bitarra eta binakaturikoa.

a) Sistema hamartarra

Sistema hamartarra batetik hamarrera doan kontatzeko sistema erabiltzeari deitzen zaio.

Horrela, hizkuntz datuak hartu duten egituratik abiatuz, multzo bakoitzari zenbaki bat ematen zaio. Demagun hizkerak banatzeko aukeratu diren ezaugarrietarik bat mugagabe / mugatu oposaketa bereizten den ala ez dela. Eta demagun, halaber, ezaugarri hori aztertzeko inesibo kasuko datuak hautatu direla bost amaiera desberdin hartuz *(-a, -e, -i, -o, -u* letrez amaitutako hitzena, hain zuzen ere). Eskuartean diren aukera desberdinak honela eratu daitezke:

a1) mugagabeko morfema bost aukeratatik batean ere ez bada ezagutzen (beraz, azaltzen): "0" kodea;
a2) mugagabeko morfema behin agertzen bada: 1
a3) mugagabeko morfema bi aldiz agertuz: 2
a4) mugagabeko morfema hirutan jaso bada: 3
a5) mugagabeko morfema lau bider ezagutzen bada: 4
a6) denetan agertzen bada: 5

Ezaugarri gramatikal batekin egin den azterketa lexikoan ere berdin egin daiteke: hizkera batean 3 berba erabiltzen badira nozio edo gauza batentzat hiru digito erabiliko dira: "0", "1" eta "2".

Hau da, erantzun bakoitzak kode bat darama. Zenbakizko kodaketa hau estatistika automatizaturako programentzako irakurgarria da eta ez du zailtasunik aurkitzen zenbakiak irakurtzeko. Alde horretatik ez dago arazorik.

Baina datuak horrela kontatuz gero eta programa informatikoari desberdintasunak sailkatzeko eskatzen bazaio aukera guztiak desberdintzat joko ditu. Honaino ez dago arazorik. Desberdintasuna neurtzean baina sortuko dira arazoak, euren arteko distantzia edo desberdintasuna berdintzat joko du eta:

a1-etik a2-ra (0-tik 1-era) = 1

a1-tik a3-ra (1-etik 2-ra) = 1
a1-tik a4-ra (0-tik 5-era) = 1
a1-etik a5-era (1-etik 5-era) = 1

Hau da, talde batetik bestera distantzia bera izango da beti.

Lortzen diren distantziak ikusita gorago egin den itaun bera egingo dugu: nahikoa ote da horrela kodatzea? Ez ote du kodaketa sistema honek benetako hizkuntz errealitatea ezkutatzen? ez ote da informazioa galtzen? Ez ote dago a1) aukera a2)tik, a6)tik baino hurbilago? Hau da, posible diren 6 kasutatik morfema mugagabea behin ere lortu ez den aukera ez ote dago mugagabea behin lortu den aukeratik, denetan lortu denetik baino hurbilago? Baietz esango du batek baino gehiagok.

b) Bada beste aukera bat. Kodaketa hau bestelako irakurketa batekin burutzea hurbiltasun kontzeptua erabiliz. Hots, aukera berdin eta desberdinak zenbatuz joan ordez, aukera hurbilenak biltzen joatea eta ondoko eran kontatzea:

a1-etik a2-ra = 1
a1-etik a3-ra = 2
a1-etik a4-ra = 3
a1-etik a5-era = 4
a1-etik a6-ra = 5

Oraingo honekin mugagabeko formen agerraldiaren arabera zehazten dira hizkeren arteko desberdintasunak edo distantziak.

c) Bada, azkenik, beste aukera bat. Ezaugarri honen neurketarako erabiltzen diren datuen bikoizketa egin eta amaierako bokal bakoitzean jasotzen diren datuen arabera burutzea. Horrela, ezaugarria bost azpiezaugarritan zatikaturik agertuko litzateke eta hauetako bakoitza era bitarrean aztertuko:

c1) -a amaiera: forma mugagabea lortzen bada “1” eta mugatua bada “0”
c2) -e amaiera: forma mugagabea lortzen bada “1” eta mugatua bada “0”
c3) -i amaiera: forma mugagabea lortzen bada “1” eta mugatua bada “0”
c4) -o amaiera: forma mugagabea lortzen bada “1” eta mugatua bada “0”
c5) -u amaiera: forma mugagabea lortzen bada “1” eta mugatua bada “0”

Azken honek, baina, arazo bat azaleratzen du: ezaugarri bat izan dena 5etan bihurtu dela. Eta zenbat eta gehiagotan bihurtu ezaugarri horrek beste guztiekin biltzean pisu handiagoa hartzen duela, bost aldiz kontatzen delako.

Hiru aukerak ikusiz nire proposamena bigarrena hautatzea izango litzateke. Lehenak hizkuntz errealitatea ezkutatzen duela uste baitut eta hirugarrenak ezaugarriari pisu gehiegi, besteak baino handiagoa ematen diolako.

Ikusten den bezala, kodaketa numerikoa ezartzean kontu handiz ibili beharra dago eta alde guztiak aztertu ondoren erabakitzeko gaia da, ez nolanahizkako aukera eginez.

b) Sistema bitarra

Erantzunak "bai/ez" edo "+/-" eran ezartzen dituen sistemari deitzen zaio sistema bitarra, ezaguna den bezala. Sistema honetan bi digito baino ez dira erabiltzen "0" eta "1". Lehena ezaugarria ematen ez denean erabiltzen da eta bigarrena jasotzen denean.

Sistema bitarra guztiz erabilgarri da dialektologian, datuak "lehen kolpearen teoria" izenaz ezagutzen den prozeduraren bidez jasotak izan direnean. Prozedura honen bidez erantzun, item, bakarra jaso ohi da galdera bakoitzean. Herri bakoitzak galdera bakoitzeko erantzun bat bakarra izango du. Ondorioz, item bat ezagutu ala ez, ez dago beste aukerarik; "0" ez bada "1" izango da, nahitanahiez.

Dialektometriaren sortzailea izan zen Jean Séguy okzitaniarrak, ALGko datuen azterketa dialektrometrikoan[5], aurretik aukeratutako hizkuntz ezaugarri bakoitzaren araberako kodaketa bitarra erabiltzea zuen helburu, nahiz ez zuen bere asmoa osoki bete.

Fonologia diakronikoan, adibidez, ez da beti "+/-" sistemara makurtzen; lehenengo ezaugarria, asimilazioa, bi mapetako (ALGko 2208 eta 2209) datuen arabera aztertu zuen: halako hizkeran erantzun bi posibleetarik batean ere ez bazen asimilaziorik gertatu "0" ezartzen zion, bietan agertzen bazen "1", batean bai eta bestean ez ematen bazen "2".

Fonologia sinkronikoan, ordea, sistema bitarra erabili zuen oso-osoan. Aditz morfologian eta morfosintaxiari buruzko datuetan ere ez zuen lortu sistema bitarra erabiltzea eta kodaketa pseudo-kuantitatiboa erabili behar izan zuen.

Kode hauek guztiak taula handi batzuetan jarri zituen eta gero hizkeren arteko desberdintasunak eskuz kontatzeari ekin zion, Séguy-k eskuz burutu baitzuen analisi estatistiko guztia Jakitun zen, halere, ordenadorez egin zitekeela

---

[5] Ikus *Atlas Linguistique de la Gascogne, vol VI + volume annexe + matrices dialectométriques (phonétique diachronique, phonologie, morpho-syntaxe, morphologie verbale, lexique),* CNRS, Paris, 1973, "Notice explicative".

eta honela dio "ces matrices dialectométriques ne sont, à vrai dire, que l'ordinateur du pauvre"[6].

D. Phips jarraitzaileak maisuaren bidea jarraitu zuen eta hark eskuz egindako lanak era automatizatuan burutu zituen[7]. Bere abiapuntua ondokoa izan zen: azterketa kuantitatiboak bakarrik jarriko zuela agirian eremuen egitura.

J. Séguy-k bezala Philps-ek ere ezin izan zuen sistema bitarra bere osotasunean erabili. Fonologian erabili bazuen ere, gainontzeko parametroetan sistema binakaturikoa aukeratu behar izan zuen.

Hain zuzen ere, gorago esan bezala, hizkuntz errealitatea nabarra baita. Inkestagintzan erabilitako sistemaren bidez hizkuntz errealitate horren aberastasuna era egokian jaso bada, erantzun edo forma bakar baten ordez erantzun eta forma aniztasuna biltzen da sarri.

Guk ez dugu sistema hau erabiltzea proposatzen "bai/ez" sistema baino zehatzagoa burutu gura delako, kuantifikazioa sartu gura delako, beste berba batzuetan.

c) Sistema binakaturikoa

Zertan datza sistema edo kodaketa hau? Bi digito baino ez dira erabiltzen, sistema bitarrean legez, binakaturikoan ere: ezaugarria agertzea edo ezagutzea "1" digitoarekin adierazten da eta ez agertzea "0" digitoarekin. Baina sistema binakaturikoan kode bakoitzean "0" edota "1" digitoak behin baino gehiagotan ager daitezke. "0" eta "1" digito multzo bat izaten da kodea.

Sistema honen abantaila nagusiena aldi bakoitzean, ezaugarri batean, hizkuntz datu bat bakarrik erabili ordez, bi, hiru, lau edo gehiago erabili ahal izatea da. Jasotako hizkuntz datuak banaka aztertzen dira, baina elkarren ondoan denak kode batean, horregatik binakaturikoa.

Sistema bitarra n aldiz errepikatua bezala uler daiteke. Kodaketa bitar multzoak osatuko luke sistema binakaturikoa.

Gorago aztertu den adibidea berrartuko da sistema hau nola iraikitzen den adierazteko: mugagabe/mugatu oposaketa zehazteko deklinabideko inesibo mugagabeko kasua bost amaiera desberdinekin (bost bokalez amaitutako

---

6 Idem, 23. or.

7 *Atlas dialectométrique des Pyrenées Centrales,* Thèse pour le Doctorat d'état sous la direction de J.-L. Fossat, UTM, 1985 (argitaragabea).

datuekin), hain zuzen ere. Bokal bakoitzeko amaiera bakoitzean bi aukera daude: edo mugagabez burutzen da edo mugatuz. Bost datutan oinarritzea erabakiko balitz (amaiera bakoitzeko datu bat) bost digitotako kodaketa eraikiko litzateke: kasu bakoitzari kodaketaren gune edo posizio bat egokituko zaio eta bi aukera izango ditu (0 eta 1). Oposaketa bizirik bada kodea "1" izango da eta ez bada jadanik bizirik "0". Kodaketa honek bost digito izango lituzke elkarren segidan, gorago esan bezala, "00000" adibidez. Bost zeroko kodeak mugagabea behin ere ez dela jaso esan gura izango du eta "11111" kodeak oposaketa bost kasuetan egiten dela. Hona hemen adibide batzuk (lehenik kodaketa agertzen da eta ondoren bost amaierak):

- 00000: -an, -ean, -ian, -oan, -uan.
- 10000: -atan, -ean, -ian, -oan, -uan.
- 11000: -atan, -etan, -ian, -oan, -uan.
- 11100: -atan, -etan, -itan, -oan, -uan.
- 11110: -atan, -etan, -itan, -otan, -uan.
- 11111: -atan, -etan, -itan, -otan, -utan.
- 01000: -an, -etan, -ian, -oan, -uan.
- .....
- 00100: -an, -ean, -itan, -oan, -uan.

...

- 00010: -an, -ean, -ian, -otan, -uan.

...

- 00001: -an, -ean, -ian, -oan, -utan.

Era honetan lortutako guarismoak benetako zenbakiak baino etiketak direla esan daiteke. Ezker-eskuin irakurri ordez, goitik behera irakurri behar izaten baitira "zenbakiok".

Bost digitoko kodeotan lehen zutabeak *-a* hizkiz amaitzen den deklinagaiaren datuak ordezkatzen ditu *(-an* denean "0" digitoa eta *-tan* denean "1"), bigarrenak *-e* hizkiz amaitutakoarenak, e. a.

Distantzien neurketan sistema hamartarretik urrundu egiten da guztiz sistema hau. Hartan desberdintasuna nabarmentzen zen, baina kuantifikatzeko zailtasunak genituen; honetan, aldiz, goitik beherako irakurketaren bidez desberdintasunak modu erraz batez kuantifika daitezke.

Nola zenbatu edo kuantifikatu desberdintasunak? Goitik beherako irakurketaren bidez digito desberdintasuna zenbatu egingo da zutabez zutabe. Adibdez "00000" aukeratik "10000" aukerara dagoen distantzia kuantitatiboa "1" izango da:

00000 (kasu guztietan forma mugatua jaso da)
10000 (lehenengoan forma mugagabea eta besteetan mugatua)

Hau da, lehen zutabean baino ez dago desberdintasuna, beste guztietan digito bera agertzen baita. Ondorioz, distantzia "1" izango da.

Distantzia bera egongo da, halaber, 00000-tik 01000-ra nahiz 00001-ra: hau da, bigarren kodean bigarren digitoa da desberdina eta hirugarren kodean bostgarren digitoa; bietan ere behin baino ez da desberdinki gauzatu (-a amaieradun erantzunak hiru kodeotan mugatuak dira, -e amaieradunetan bitan forma mugatua jaso da, baina batean mugagabea, -i amaieradunetan hiru hizkeratan forma mugatua, -o amaieradunekin ere berdin, eta -u amaieradunetan bitan mugatua eta batean mugagabea). Hiru hizkerak binaka harturik digito batean baino ez dira desberdinak; beraz, euren arteko distantzia "1" izango da.

Aldiz, 00000-tik 00011-ra distantzia "2" izango da: lehenengo hizkeran ez da mugagabeko morfema bat bera ere agertu (horregatik dira digito guztiak "0"), baina bigarrenean bi kasutan agertzen da mugagabeko morfemari dagokion digitoa. Kodeok bitan direnez desberdin distantzia "2" markatuko da.

Sistema binakaturikoa, gainera, ezaugarri baterako aukeratu diren datuei egokitu egiten zaie. Bost datu desberdin aukeratu badira bost digitotako sistema ezarriko da; hiru datu aukeratu badira, hiru digito. Guztiz sistema erosoa da lan egiteko.

4. Erantzun aniztasuna

Gertatzen da hizkera batzuetan erantzun bat baino gehiago jaso izatea. Hizkuntz datuen egoera aberatsa da. Inkesta metodologia egokia erabili behar da errealitate hori bere osotasunean jasotzeko. EHHAko inkestagintzan beti begi aurrean izan dugun errealitatea da eta inkestero ahal denik eta zehatzen jasotzen saiatu gara; inkestagilea ez da gelditzen izan, kasu askotan, erantzun bakarrean beste erantzun batzuk posible zirela aurrikusten bazen, behintzat. Erantzun hutsak ez du asetu behar inkestagilea, erantzun horren inguruan induskatu behar da datu osoagoen bila.

Hiztunen gaitasunean kontzeptu bat adierazteko item bat baino gehiago izatea edo ezaugarri gramatikal bat modu batean baino gehiagotan formulatzea ez da azken urteotan asmatutako bereiztasuna, nahiz geolinguistikako lehen aroko inkestagintzan ez zen kontuan hartzen, "lehen kolpearen teoria"k izan zuen indarra dela eta.

Hizkerak bereizteko saioan hizkuntz gertakari hau ez da ahaztu behar, polimorfismoaren kasuan gertatzen den bezala. Hizkera batean bi item ezagutzen

badira, hizkera horren ondarekoak badira, biak izan behar dira kontuan konparatze momentuan.

Egia esan, horrelakoetan ikertzaileak zailtasunak aurkitzen ditu eskuartean dituen datuak zenbakitan jartzean. Adibidez, puntukaritasuna hizkerak banatzeko ezaugarritzat hartzen bada eta berau aztertzean ("ari da" perifrasia ezagutzen den ala ez aztertzen duena), badira hizkera batzuk bi soluzioak ezagutzen dituztenak: *dago* eta *ari da*.

Arazoari holako berezitasunak dituzten ezaugarriak bikoiztuz eman behar zaio irtenbidea sistema hamartarrean: lehenengoan *dago* erantzun tipoa ezagutzen den aztertuko litzateke, bigarrenean *ari da* perifrasia ezagutzen den ala ez: *dago* erantzuna ezagutzea "1" kodatuko litzateke, bestea "0"; forma hau ezagutzen duten hizkeren artean ez litzateke desberdintasunik kontatuko. *Ari da* kasuan ere ezagutzea "1" eta ezezagutza "0" kodatuko litzateke.

Baina ezaugarria bikoiztearekin desberdintasunak handitu egingo zaizkigu. Aztertzen dihardugun kasuan ezaugarria bikoiztean (lehena *dago* aztertzeko utzi eta bigarrena *ari da* perifrasia), lehen erantzuna ezagutzen den hizkera eta bigarrena ezagutzen duenaren artean desberdintasuna "2" izango da: bat bigarrenak ez duelako *dago* ezagutzen eta "2" lehenengoak ez duelako *ari da* ezagutzen. Baina, berez, hizkera horien arteko desberdintasuna "1" izan beharko luke: batak *dago* erabiltzen duelako eta besteak *ari da*.

Sistema binakaturikoan, ordea, ez dago ezaugarria bikoiztearen beharrik. Bi digitotako kodaketa asmatuz lehen digitoa *dago* ordezkatzeko erabiliko da eta bigarrena *ari da* forma. Demagun hiru hizkera ditugula: batak *ari da* forma ezagutzen duela, bigarrenak *dago* eta hirugarrenak biak, *ari da* eta *dago*. Honako eran kodatuko lirateke:

01 *(ari da* forma erabiltzen da eta ez *dago)*
10 *(dago* forma erabiltzen eta ez *ari da)*
11 *(dago* eta *ari da* formak erabiltzen dira, biak)

Baina kodatzean arazoak sortzen dira. Honen arabera hiru aukera hauetarik lehen eta bigarren hizkeren artean distantzia "2" izango da: lehenak ez du *dago* forma ezagutzen (beraz, desberdintasun bat) eta bigarrenak *ari da* forma (bigarren desberdintasuna). Baina hizkera batean aukera bat ala beste hartzeak (hizkera batean *ari da* esaten bada eta bestean *dago)* distantzia "1" eman beharko luke, eta ez "2".

Gehiago oraindik, lehen bi hizkerak eta hirugarrenaren artean distantzia "1" da, azken honek bi formak ezagutzen dituelako, baina ez luke hizkuntz distantziarik egon behar; hau da, "01" aukeratik "11" aukerara (edo beste era batera esanda, bai *dago* bai *ari da* formak (biak, beraz) ezagutzen dituen

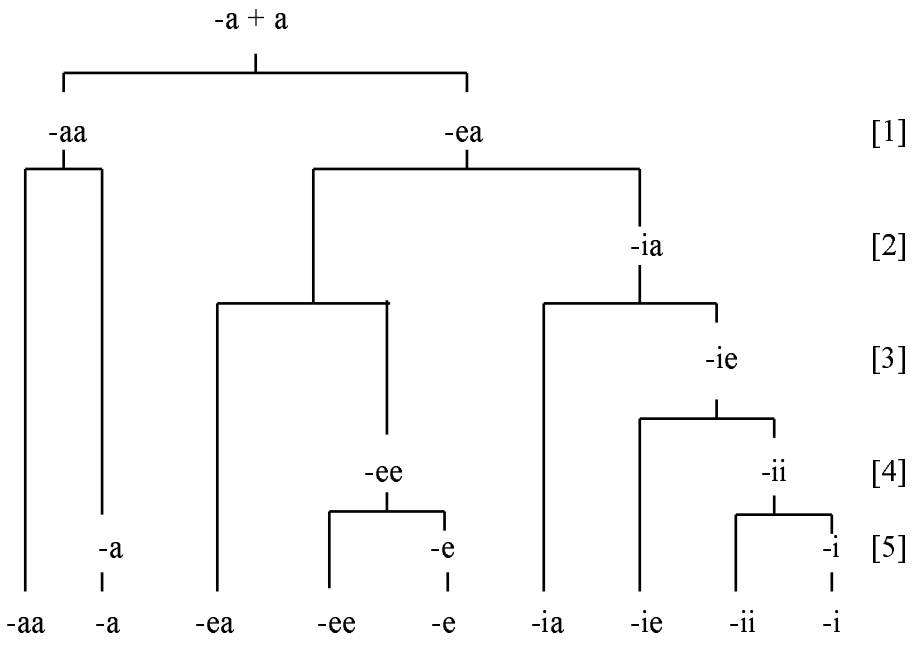
hizkeratik bietarik bat ezagutzen duen hizkerara) distantzia "0" izan beharko litzateke.

Konpondu gabeko korapiloa da eskuartean duguna. Eta bide onetik askatu beharrekoa hizkuntz datuen errealitatea modu fidagarriz eta ahal denik eta estuen formulazio numerikoetan eraldatzea gura badugu.

5. Arau fonologikoen kodaketa numerikoa

Lexemak artikulua jasotzen gertzen diren arau morfonologikoen araberako hizkuntz distantziak zehaztea bideragarri egitea guztiz egokia da. Jo dezagun "-a + a" dugula (hots, -a amaieradun hitzari artikulua erantsi zaiola). Eta jaso diren erantzunen artean ondoko amaierak ditugula: *-aa, -a, -ea, -e, -ia, -ie* eta *-i*.

Nola egiten da bereizketa eta distantzien azterketa? Nahikoa ote da *-aa* forma eta *-ea* desberdinak direla esatea? Distantzia berdina ote dago *-a* formatik *-ia* formara edo *-i* formara? Ezetz uste dugu. Desberdintasun horiek modu desberdinez kuantifikatu egin behar direlako ustea dugu. Uste hau jokoan parte hartzen duten arau morfonologiko desberdinak direlakoan datza. Kasu honetan ematen diren arau morfonologikoak hierarkia batean sartu bada ondoko egitura hartuko luke:



---	---	1	1	1	1	1	1	1	[1]
---	---	---	---	---	2	2	2	2	[2]
---	---	---	4	4	---	3	3	3	[3]
---	---	---	---	---	---	---	4	4	[4]
---	5	---	---	5	---	---	---	5	[5]

Egitura honetan arau morfologikoak mailakatu egin dira. Maila bakoitzean arau bat agertzen da. Arauon hierarkia gauzatu da horrela. Zertarako? Batetik arauon hurrenkera zehazteko; bestetik kodaketa numerikoa errazteko.

Ematen diren arau morfonologikoak hauek dira:

[1] a > e / ___ a #

[2] e > i / ___ a #

[3] a > e / (i, u) ___

[4] V > V_1 / V_1 __

[5] V_1 > Ø / V_1 __

Arau bakoitzari kodaketa numerikoa:ezartzean maila bakoitzean agertzen den arau morfonologikoaren ezagutza "1" kodearekin adieraziko da sistema binakaturikoan; ezezagutza "0" kodearekin. Sistema hamartarrean, ostera, hori egitea ez da posible. Aukera bakoitzak bere zenbakia izango du, besterik gabe. Beraz,

a) Sistema binakaturikoa

- aa: 00000
- a: 00001
- ea: 10000
- ee: 10010
- e: 10101
- ia: 11000
- ie: 11100
- ii: 11110
- i: 11111

b) Sistema hamartarra

- aa: 1
- a: 2
- ea: 3
- ee: 4
- e: 5
- ia: 6
- ie: 7
- ii: 8
- i: 9

Kodaketa numerikorako sistema binakaturikoan distantzia neurtzeko oinarria ondokoa da: ematen den arau morfonologiko bakoitzak distantzia unitate bat markatuko du. Ondorioz *-aa* formatik *-ea* formara distantzia bat kontatuko da, arau fonologiko batek baino ez duelako parte hartzen (-aa > -ea = 1); *-a* formatik *-ea* formara distantzia "2" izango da (-a > -ea = 2) bi arau fonologiko direlako tartean: *-a* formatik *-aa* formara bat eta honetatik *-ea* formara bigarrena. Azkenik, *-aa* formatik - *i* formara "5" izango da hizkuntz distantzia (-aa > -ea > -ia > -ie > -ii = 5), bost direlako aldaera batetik bestera joateko beharrezko diren arau fonologikoak.

Sistema hamartarrean, berriz, ez da arau bakoitza kontatzen, arau guztien gehiketak beti distantzia bera emango du: "1".

6. Ondorioak

Artikuluan hizkuntz datuen analisian estatistika automatizatuaren ontasuna aldarrikatzen da, inguruko disziplinetan aplikatzen den eran.

Estatistika automatizatuak hizkuntz datu kopuru handiak ahal denik eta era egokienean aztertzeko eta sailkatzeko. Bide horretan, artikuluaren helburua hizkuntz datuak formulazio matematiko edo numerikora eraldatzeko behar diren urratsak ematea da. Datuen eraldatze hau behar-beharrezkoa da analisi estatistikoak egiteko.

Eraldaketa honetan hizkuntz datuek errealitatean ezagutzen duten aberastasuna oso-osorik gorde gura da.

Lehen-lehenik datuak errealitatean nola bizi diren zehaztu da, eta errealitate horri ondoen egokitzen zaion formulazioa asmatzen ahalegindu da.

Hiru formulazio desberdin aztertu dira: sistema hamartarra, bitarra eta binakaturikoa. Sistema bakoitzaren abantaila eta sortzen dituen arazoak aipatu dira eta egokiena zein izan daitekeen ere proposatzen da.

Dena den, badira arazo batzuk batak ere konpontzen ez dituenak. Horietarik bat hizkera batean galdera batentzat bi erantzun edo gehiago direnean gertatzen dena da. Aukera desberdinak aztertu arren ez zaio irtenbide egokia eman, oraingoz.

LABURPENA

Hizkeren arteko hizkuntz desberdinketa eta distantzia gaia jorratzen da artikuluan zehar. Desberdintasun edo distantzia horiek nola edo hala kunatifikatzea proposatu ostean ikertzaileak lan horretan aurkitzen dituen eragozpenak ditu aztertzen. Lanok sistema informatiko integratuen bidez burutzeko hizkuntz datuak datu numerikoetara eraldatu behar dira. Eraldaketa honetan informazio galtzerik gerta ez dadin kodaketa egokiena zein izan daitekeen galdetzen da.