1

Hizkuntzen Heriotza digitala*

Andras Kornai
Computer and Automation Research Institute, Hungariako Zientzien Akademia, Budapest, Hungaria
andras@kornai.com

Itzultzaile eta editoreak: Iñaki Alegria, Beñat Garaio, Belen Uranga
Sarrera-data: 2018-11-10 / Onartze-data: 2018-12-15

Laburpena. Gaur egun munduan hitz egiten diren 7.000 inguruko hizkuntzen artean, 2.500 gutxi gorabehera galzorian daudela jotzen da. Adostutako zenbaki hori oso zalantzan jarri dugu artikulu honetan, uste baitugu heriotza-arriskuan dauden hizkuntzen kopurua behetik kalkulatuta dagoela, zeren-eta soilik hizkuntzen %5 baino gutxiago igo daitezke esparru digitalera. Artikulu honetan arrakala digitalak sortutako hizkuntzen heriotza masiboaren ebidentzia aurkeztuko dugu. • ***Hitz gakoak:*** *Bizindar linguistikoa, heriotza digitala, identitate linguistikoa*

Abstract. Of the 7,000 or so languages spoken in the world today, it is estimated that around 2,500 are at risk of extinction. In this paper, we question this generally-accepted figure, since we believe that the actual number of languages at risk of dying out is much higher, given that less than 5% of all languages can be uploaded to the digital environment. We present evidence of the mass language death generated by the digital divide. • ***Key words:*** *Linguistic vibrancy, digital death, linguistic identity.*

* Testu hau 2013an argitaraturiko "Digital Language Death" artikuluaren egokitzapena da. Jatorrizko bertsioa (en) hemen topa daiteke: https://doi.org/10.1371/journal.pone.0077056. Ohartu behar dugu gainera itzulpenean jatorrizko hainbat aipu, pasarte eta zehaztasun matematiko laburtu edo kendu direla, artikuluaren ulermena eta estaldura kaltetu gabe (Itzultzaile eta editoreen oharra).

SARRERA

Hizkuntza izaki bizidun gisa ikusten duen metafora biologikoa Herder garaitik dator gutxienez (Morpurgo-Davies A, 1998) eta argi adierazia izan zen, esate baterako, *The Descent of Man* Darwin-en liburuan:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. (…) We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when others change is very like correlated growth. (…) Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters. Dominant languages and dialects spread widely, and lead to the gradual extinction of other tongues. (Darwin C, 1871, 423 or).

Aurkakoak izan dituen arren, metafora biologikoa modu zabalean onartua izan da bai hizkuntzen heriotzen inguruko ikerkuntzan, eta baita ekintza politikoaren gidaritzan ere. Artikulu honetan, bizindar linguistikoaren (vitality) inguruko ohiko metodoak eremu digitalera eraman ditugu.

Aurkakoak izan dituen arren (Frank RM, 2008), metafora biologikoa modu zabalean onartua izan da bai hizkuntzen heriotzen inguruko ikerkuntzan (Nettle D, Romaine S, 2000, 243 pp. eta Crystal D, 2002; 210), eta baita ekintza politikoaren gidaritzan ere (Ad hoc technical expert group, 2004). Artikulu honetan *igoera digital* (*digital ascent*) fenomenoa aztertuko dugu, hain zuzen, hizkuntzak komunikazioaren espazio digitalean sartzen direnean gertatzen den fenomenoa. Metafora biologikoa zabal dezakegu, hizkuntzen eklosioaz eta metamorfosiaz mintzatzeko, baina gutxi aurreratuko genuke horrela, izan ere, hizkuntzen bizitza-zikloan, hizkuntzen etorkizuneko egoera post-digitalaz espekulatu besterik ezin baita egin.

Artikulu honetan, bizindar linguistikoaren (*vitality*) inguruko ohiko metodoak eremu digitalera eraman ditugu. Lehen atalean, irizpideen transferentzia egin dugu, hiztunen populazioari begiratu ordez, online populazioari begiratuz, eta ahozko erabilera indartsuaren ordez, online erabilera indartsua behatuz… (ikus *Oinarria* atala). Bigarrenean, online iturburuetatik datu-bilketa egin dugu, aldagai esanguratsuak antzemateko, edo gutxienez aldagai horietarako hurbilpen onargarriak (Ikus *Materialak*). Hirugarrenean, lau kategoriako sailkapen-sistema bat ezarri dugu: *thriving-oparo* (T), *vital-bizi* (V), *heritage-ondare* (H), and *still-motel* (S). Kategoria horiek bat datoz nagusiki hizkuntzari dagozkion erabilera digitalaren neurriekin, eta *hazi* prototipikoak ezarri dira azalpenerako (ikus *Metodoak*). Azken atalean, hainbat sailkatzaile automatiko nola eraiki diren aztertu da, hazietan oinarrituta eta gainontzeko datuetara aplikatuta (ikus *Emaitzak*). Lau urratseko metodo hau sendoa dela frogatu

BAT Soziolinguistika Aldizkaria 111, 2019 (2) | 11-32 | Andoain (Gipuzkoa)
ISSN 2659-5176 / ISSN-L 1130-8435 | Lege gordailua: SS120/90

Garai digitalean, hizkuntzaren heriotzaren seinale goiztiar horiekin batera beste ezaugarri batzuk ere batzen dira. Mundu digitaleko funtzio-galerak eguneroko komunikazio-funtzioetako galera ukitzen du.

da, eta hautatutako *hazietatik* independentea (ikus *Eztabaida*). Ondorioetan, emaitza nagusiak interpretatuz adierazi da hizkuntzen artean gehiengo nagusia (8.000 baino gehiago), egoera *motelean* daudela, hau da, ezin dutela *igoera digitala* burutu.

OINARRIA

Hizkuntza bat ez da erabat hilko azken hiztunaren heriotza gertatu arte, baina badira hiru seinale argi, haren berehalako heriotza aurretik antzemateko. Lehena, hizkuntza jakin horren funtzio-galera da, beste hizkuntzek eremu funtzional osoak -merkataritza, esaterako- hartzen dituztenean behatu daitekeena. Bigarrena prestigio-galera da, bereziki argia dena belaunaldi gazteen hizkuntza-jarreretan antzematen bada. Eta hirugarrena, konpetentzia-galera da, sasi-hiztunen agerpenarekin islatzen dena; hiztun horiek aurreko belaunaldiaren hizkera uler dezakete oraindik, baina gramatikaren bertsio izugarri sinplifikatua darabilte. Fenomeno hori luze eta zabal dokumentatu da, adibidez menomini hizkuntzan (Bloomfield L, 1927), Eskoziako gaeleran (Dorian NC, 1981; 202pp) eta dyrbal hizkuntzan (Schmidt A., 1985).

Garai digitalean, hizkuntzaren heriotzaren seinale goiztiar horiekin batera beste ezaugarri batzuk ere batzen dira. Mundu digitaleko funtzio-galerak eguneroko komunikazio-funtzioetako galera ukitzen du (e-posta, testumezularitza), esate baterako, merkataritzan edo tramite ofizialetan. Prestigio-galera argi antzematen da "webean ez badago ez da existitzen" leloaren bitartez, eta konpetentzia-galerak hizkuntza horretan natibo digitalen kopurua mugatzen du (Prensky M, 2001). *Igoera digitala* honen alderantzizko prozesua da, eta ondoren zehazten den kasuetan gertatzen da: hizkuntza jakin batek gero eta funtzio berri gehiago eskuratzen duenean, alde batetik, eta, bestetik, haren hiztunek gero eta konpetentzia digital handiagoa eskuratzen dutenean.

Hizkuntza baten galtze-arrisku larria edo heriotza bere ohiko adieran oso ikertua izan da, eta haren aurkako borroka ere zabalduta dago. Gaur egungo EGIDS sailkapenak (Lewis MP, Simons GF, 2010), Fishman-en GIDS sailkapenak (Fishman JA., 1991; 431pp) 13 kategoria[1] ezartzen ditu: (0) Nazioartekoa, (1) Nazionala, (2) Eskualdekoa, (3) Komunikazio zabalekoa, (4) Hezkuntzakoa, (5) Garapenean, (6a) Indartsu, (6b) Arriskuan, (7) Galeran, (8a) Hilzorian, Ia iraungia (8b), Lozorroan (9) eta Iraungia (10). 7-8b arteko hizkuntzak galzorian sailkatzen ditu Unescok (Mosley C., 2010), eta 9-10 taldekoak desagertutzat ematen ditu. Sailkapen horren arabera munduko hizkuntzen artean soilik %17 desagertuta daude, eta %20a arriskuan (6b kategoria); beraz, geratzen den %63a osatzen duten hizkuntzak egoera onean daude, ziur asko go-

rabeherekin (zenbaki hauek Simons GF, Lewis MP, 2013, erreferentziatik hartuak dira). Datu horiek ohiko adieran zuzenak izan badaiteke ere (tradiziozko bizindar linguistikoa), gure ikerketaren aurkikuntza nagusiak ezeztatzen ditu, hizkuntzen gehiengo zabalak (%95etik gora) *igoera digitala* lortzeko ahalmena galdua baitu.

Hizkuntzen funtzionalitateen artean komunikazio digitala ez da gai garrantzitsua izan 1970eko hamarkada arte, orduan hedatu baitzen dokumentu digitalen sorrera.

Azken hiztuna hil ondoren ere datu digital(izatu)ek irauten dutenez gero, ezin dugu lotu, besterik gabe, datu digitalen gabezia eta igoera digitalarekiko ezintasuna. Horregatik bereiziko dugu alde batetik *ondare digital* egoera (*digital heritage*), (L1) ikerketarako eta dokumentaziorako material digitala egon arren, bertako hiztunek komunikazio digitalerako erabiltzen ez duten kasuetarako; eta *moteltasun digital* egoera (*digital still*), inolako agerpen digitalik ez dagoen kasuetarako, ezta atzerriko (L2) erabiltzaileek sortua ere. Garrantzitsua da, zalantzarik gabe, hizkuntzak egoera *motele*tik *ondare* egoerara igarotzea, eta, zentzu horretan, ekimen anitz ari dira burutzen hizkuntza horiei dagozkien datuak eta metadatuak online jartzeko, eta baliabide lexikalak eta oinarrizko testuak atzigarri jartzeko web bidez (ikus *Materialak* atala). Emaitzen atalean ikusiko dugu nola ekimen horiek goraipatzekoak izanda ere, ekarpen txikia egiten dioten galzorian dauden hizkuntzen indartze digitalari. Bere hezurdura gordetzeak edota museotan bere fosilak gordetzeak *dodo*aren desagerpena moteltzen ez duen moduan, tribuetako adinekoen herri-poesiaren errezitazioko ahots-fitxategiak online jartzeak ez du hizkuntza jakin horren igoera digitala suspertuko; *motel* zein *ondare* egoerako hizkuntzak digitalki hilda daudela esan daiteke, ez baitute hizkuntza-komunitatearen komunikazio-beharretarako balio.

Igoera digitala nahiko fenomeno berria da, bereziki hizkuntzen heriotzaz ikertzen den ehunka urtetako denbora-tartearekin alderatzen badugu. Hizkuntzen funtzionalitateen artean komunikazio digitala ez da gai garrantzitsua izan 1970eko hamarkada arte, orduan hedatu baitzen dokumentu digitalen sorrera. Ondoren etorri ziren Internet eta e-posta, 80eko hamarkadan, Weba eta blogintza 90ekoan, eta wikiak eta SMSak 2000koan. Gure hurbilpena, hala ere, kontserbadorea izango da, izan ere, euskarri eta irizpide kontzeptual estandarrak domeinu digitalera eraman ditugu. Kontserbadore jokatzeko bide horretan, aldeko ebidentziaren interpretazioa egingo dugu ahal denean, alarma faltsuak ekiditeko asmoz. Elkartzen diren bost faktore hartuko ditugu kontuan: (i) hizkuntza-komunitatearen tamaina eta osaera demografikoa, (ii) hizkuntzaren ospea, (iii) hizkuntzaren identitate-funtzioa, (iv) softwareak emandako sostengu-maila, eta (v) wikipedia. Azken biak domeinu digitalari atxikitakoak dirudite lehen begiratuan, baina geroago ikusiko dugunez, adierazle egokiak dira ohiko irizpide bat berresteko, hizkuntzaren hedapen funtzionala, hain zuzen.

Hiztun-multzo handi eta iraunkor batek ez du igoera digitala ziurtatzen; horren ondorioz, kontatu nahi duguna interakzio digitalean aritutako populazioa da.

1. Komunitatearen tamaina

Bizindarraren lehen neurria hizkuntza-komunitatearen tamaina eta belaunaldi-osaera da. Esparru digitalean gure interesa hizkuntzan dauden natibo digitalen kopuruan dago. Digitalizazioaren fenomeno berria den heinean demografiarekin lotura handia du: behin komunitate digitalaren edukia sortzen hasita (mezuak bidaliz, blogetan idatziz edo wikiak osatuz), zentzuzko itxaropena izan dezakegu hurrengo belaunaldi gazteagoek jarraituko dietela, bereziki Facebook bezalako foroetan, zeintzuk gero eta erabiliagoak baitira gurasoen eta aitona-amonen aldetik euren umeekin harremanetan egoteko. Hortaz, nahikoa izango da konektatutako komunitatearen tamaina estimatzea, uniformeki osatuta dagoela onargarritzat hartuta.

Errolda ofizialetan jaso ohi da identitate nazionala eta linguistikoa, eta, horrela, taldeetako tamainak ezagunak dira; beraz, ez da zaila hiztun-kopuruaren tamainaren estimazioa egitea, hurbilpenaz bada ere. Hala eta guztiz ere, hiztun-multzo handi eta iraunkor batek ez du igoera digitala ziurtatzen; horren ondorioz, kontatu nahi duguna *interakzio digitalean aritutako* populazioa[2] da. Material digitalaren kontsumo pasiboa, kanpoko hizkuntzetan egiten dena batez ere, ez da adierazgarria. Kaltegarria ere izan daiteke, Michael Krauss-en aipu ezaguna web mundura ere aplika daiteke:"*Television is a cultural nerve gas…odorless, painless, tasteless. And deadly*" (Cazden CB 2003: 53, 57).

Errolda ofizialetan edo bestelako ohiko behategietan interakzio digitalean aritutako populazioaren tamaina edota ospe digitala neurtzen ez direnez, beharrezkoa izan da bizindar digitaleko adierazleez baliatzea. Hizkuntza jakinetan gertatzen den komunikazio digitala neurtzea izan da benetako helburua. Metodo ezin hobea litzateke Skypez, telefonoz, Twitterrez, Facebookez eta bestelako komunikazioak eskuratzea eta dagokion hizkuntzaren proportzioa kalkulatzea; gainera, gaur egungo teknologiak ebatzi du hizkuntza identifikatzeko arazoa; adibidez, Crúbadán proiektuak (Scannell KP, 2007) halako softwarea eraikitzen du hizkuntza bakoitzerako. Teknologia horrek edukiak ulertzeari uko egiten dionez, konfidentzialtasunaren inguruko arazoak minimizatuta geratzen dira, eta bizindar digitala neurtzea erronka metodologikoa da nagusiki: testuak anonimoak izan daitezen bidea landu behar da, komunikatzen diren pertsonen baimena ere behar da eta beste hainbat lan. Halako azterketa exhaustibo bat osatu arte, eskuragarri den material publikoa erabili dugu (abantaila hori dugu: material horren egileek publiko egin dutenez materiala, konfidentzialtasunaren kezka aurrez ebatzi dugu). Online dagoen materialaren tamaina Zséder A, Recski G, Varga D, Kornai A, (2012) artikuluan dago azalduta, eta emaitzetako batzuk eskuragarri daude publikoki (http://hlt.sztaki.hu/resources/webcorpora.html).

2. Ospea

Bizindarraren bigarren neurgailu esanguratsuena ospea da. Komunikazio digitala bitarteko tradizionaleko komunikazioa baino ospe handiagokoa dela jotzen da unibertsalki, eta belaunaldien arteko disrupzioak igoera digitalaren alde jokatzen du, izan ere, belaunaldi berriek arestian aipaturiko bi baldintzak betetzen baitituzte: bitarteko digitalak izatea eta hizkuntza erabiltzeko interesa izatea. Digitalki indartsu diren hizkuntzetan hori gertatzen da ahalegin handirik gabe, ia automatikoki, baina belaunaldi berriek hizkuntza jakin hori ez badute "cool"tzat hartzen prozesua blokeatuta gera daiteke tenaza moduko batez: belaunaldi zaharrek ez dute erabilera digitala nahi, edo ez dira mundu digitalean sartzeko gai, eta gazteagoek ez dute mundu horretan erabiltzeko hizkuntza jakin hori aintzat hartzen. Adiera teknikoan gerta daiteke adin bateko hiztunak hiztun osoak izatea, gramatika eta hiztegiaren kontrol osoa mantentzen dutelako. Baina, aldi berean, eremu digitalerako hizkuntza desegokitzat jotzea. Horren adibide garbia da norvegieraren bi aldakuntza ofizialen kasua, Bokmål eta Nynorsk izenekoak. Urte-tarte luzean bi aldaeretako wikipediak neurri antzekoak izan dira, eta estimazio onenek (Rehm G, de Smedt K, 2012; 81pp) erabiltzaileen proportzioa 7:1 eskalan ezarri dute. Gaur egun, aldiz, Bokmål wikipedia lau aldiz handiagoa da, nahiz eta Nynorsk aldaerako wikipedia oraindik lehen 50en artean dagoen. Bizitza-maila handia duen populazio handi samarra du aldaera horretako komunitateak (ordenagailuen merkatu gain-kargatua eta abiadura handiko sareko zerbitzu handiarekin), eta erroldako wikipediako datuetan oinarrituta soilik Nynorsk litzateke igoera digitalerako hautagai argiena. Baina *.no* domeinua robot batez ustiatzen badugu dibergentzia handi bat agertzen da: 1.620 milioi hitz agertzen dira Bokmål aldaerarako eta 26 besterik Nynorsk-rako. Orri ofizialak (gobernuarenak eta administrazio lokalarenak) bi aldaeretan daudela kontuan hartzen badugu, orduan erabiltzaileek sortutako Nynorskerazko edukiak %1era ez dira heltzen. Nahiz eta hizkuntza-politika ofiziala ondo orekatua izan, Norvegiako herritarrek dagoeneko bozkatu dute euren blog eta txioen bitartez: soilik Bokmål aldaera ari dira aro digitalera eramaten.

Komunikazio digitala bitarteko tradizionaleko komunikazioa baino ospe handiagokoa dela jotzen da unibertsalki, eta belaunaldien arteko disrupzioak igoera digitalaren alde jokatzen du.

Fenomeno bera topa daiteke eten digitalaren beste aldean. Adibide gisa har dezakegu mandinka hizkuntza, swahiliarekin batera AEBetan, agian Alex Hailey-ren Roots liburuari esker, hobekien ezagutzen den Afrikako hizkuntza. 1,35 milioi hiztunekin, eta Senegalen eta Gambian estatus ofizialarekin, ez da galzorian edo arriskuan dagoen hizkuntzatzat hartzen (EGIDSeko eskalan 5 kategorian-garapenean). Gainera, adin guztietako hiztunen aldetik jarrera positiboa du. Halere, igoera digitalari begira porrot-kasu gisa agertzea aurresan zitekeen: hizkuntza horretan alfabeta-

tuak %1 azpitik dira, eta wikipedia inkubagailuak (Requests for new languages/Wikipedia Mandinka, 2013) ez du erakarri bertako hiztun bakar bat ere.

Nagusiki idatzitako materialean oinarrituko garenez, garrantzi berezia du bereizteak webeko agerpen pasiboa (irakurtzeko baino ez diren lexikoak, berriak edo literatura klasikoa) eta bi norabideko erabilpen aktibo asko (sare sozialak, merkataritza eta literatura bizi).

3. Identitate-funtzioa

Nagusiki idatzitako materialean oinarrituko garenez, garrantzi berezia du bereizteak webeko agerpen *pasiboa* (irakurtzeko baino ez diren lexikoak, berriak edo literatura klasikoa) eta bi norabideko erabilpen aktibo asko (sare sozialak, merkataritza eta literatura bizi). Adibide interesgarria da txinera klasikoarena, izan ere, 3.000 artikulutik gertu eta L2 motako 30.000tik gora erabiltzaile izanik, kontuan hartzeko neurriko wikipedia du, nahiz eta ez duen hiztunik; horrela bada, soilik funtzio historikoa edo ondarearen zaintzarenak betetzen dituen hizkuntza da.

4. Domeinu funtzionalak

Hasiera batean testu-edizio digitala erakunde handien eta inprimategien esparruko teknika zen, baina PCen hedapenarekin mahaigaineko edizio digitala iritsi zen gure etxeetara. Iragarki publikoak jartzearekin batera, antzeko zerbait gertatu da: herri mailako aktoreetara murriztuta egotetik norbanakora pasa da, edonork mezuak jar baititzake hainbat iragarkitaulatan edo (mikro)blogetan. Bide horretan aro digitalaren garapenarekin atzigarri bihurtu dira orain arte soilik elite batzuen hainbat komunikazio-modu izan direnak, eta hori da, zalantzarik gabe, duten erakargarritasun nagusietako bat. Baina hizkuntza batek, bide berri eta demokratizatu horietara iritsiko bada, software apur bat beharko du (telefono adimendunak izan daitezke salbuespen nagusia, baina datu-falta dela-eta, azterketatik at geratuko da). Software-euskarria kuantifikatzeko hiru egoerako hierarkia sinple bat erabili dugu, EGIDSen alfabetazio-maila ezartzeko egiten den galdetegiarekin bat datorrena gutxi gorabehera.

5. Wikipedia

Hizkuntza jakin baten *igoera digitalaren* funtsa eremu horretan egiten den erabilera denez gero, komunikabide nagusi gisa hizkuntza hori erabiltzen duen online-komunitate aktibo bat, gutxienez, identifikatu behar da. Hainbat aukera zeuden horretarako, besteak beste, berriak, postazerrendak, Yahoo edo Google. Halere, esperientziak erakutsi du wikipedia komunitatea hizkuntza-komunitate digital aktiboenen artean dagoela. Horren ondorioz, aitzin adierazle gisa erabil daiteke arrakala digitalaren aurrean hizkuntzen egoeraz jabetzeko. Arrazoietako bat umeekin lotuta dago, izan ere, ordenagailuak jolasean ibiltzeko ez den zerbaitetarako erabiltzen hasten diren momentutik, wikipedia hor dagoela konturatzen

dira. Gainera, wikipediak, antzeko ideiak dituzten erabiltzaileekin topo egiteko ingurune erakargarri bat eskaintzen du, eta baita denon artean ezagutzarekin lotutako helburu bat lortzeko ere. Erakargarria eta instrumentala da, eremu digitalean hizkuntza eta kultura garatzeko tresna baita. Ikerketa honen emaitza bat aurreratuz, horrela laburbil daiteke ondorioa; *wikipedia gabe igoera digitalik ez*.

Wikipedia sortzeko beharra oso argi antzematen da *digitalki igotzen* ari diren hizkuntza guztietan. Hori argi geratzen da ondoko datuan: egun 533 proposamen daude inkubagailuan, indarrean dauden wikipedien kopuru bikoitza. Hala, wikipedia bat martxan edukitzeko eta sailkapenean igotzeko helburua hain da indartsua, ze hainbatetan trikimailuak erabiltzen diren gora egiteko, artikulu-kopurua baino ez baita erabiltzen sailkapenean (http://meta.wikimedia.org/wiki/List_of_Wikipedias). Potemkin wikipedia lotsagabeen adibide behinena Volapük wikipedia dugu (37.a sailkapenean), non makinaz sortutako informazio geografikoak nagusi diren. Metodoak izeneko atalean eztabaidatuko dugu nola saihets daitezkeen trikimailu horien ondorioak.

Wikipedia sortzeko beharra oso argi antzematen da digitalki igotzen ari diren hizkuntza guztietan.

Wikipedia bat martxan edukitzeko eta sailkapenean igotzeko helburua hain da indartsua, ze hainbatetan trikimailuak erabiltzen diren gora egiteko, artikulu-kopurua baino ez baita erabiltzen sailkapenean.

MATERIALAK

Erabilitako datu guztiak bilduma publikoetatik hartu dira 2012ko martxoaren eta 2013ko martxoaren artean. Bildutako datuekin egindako taularen bertsio finkatu bat, 8.426 errenkada eta 92 zutaberekin, eta hemen dago atzigarri: https://doi.org/10.1371/journal.pone.0077056.s001. Artikulu honetan horren gainbegiratze bat besterik ez dugu aurkeztuko, nahiz eta zehaztasun gehiago esteka honetan ikus daitezkeen: https://doi.org/10.1371/journal.pone.0077056.s002). Kontuan izan datuetan munduko hizkuntza guztien adierazleak bildu nahi izan ditugula, baina, halere, hainbat hutsune egon daitekeela. Halere, kontuan izan behar da datuen gainean egindako sendotasun-testak %95 baino estaldura handiagoa dela diola.

Munduko hizkuntzen inguruan bildutako lehen datua ISO 639 estandarrari dagokio, hizkuntzaren kode normalizatuaren kudeaketa jasotzen duena. Ethnologue datu-basearen azken bertsioa kontsultatu da (2012/02/28) (http://www.ethnologue.com) eta bertan 7.776 hizkuntza topatu dira, haien artean 376 dagoeneko hilak direnak 1950 urtetik gaur arte, garai hartatik kudeatzen baitu zerrenda SILek. Beste hainbat datu-iturburu kontsultatu ziren eta gure datu-basea %10 handiago da[3], 7.879 ISO kode bilduz guztira. Erabilitako guneen artean hauek azpimarra daitezke: *Open Language Archives Community* (OLAC) (http://www.language-archives.org), *Endangered Languages Project* (http://www.endangeredlanguages.com), Crúbadán Project (http://borel.slu.edu/crubadan)

Ortografia estandarizatua, wikipedia bezalako proiektuetan laguntzeaz gain, berez, bizindarra digitalaren adierazle garrantzitsua da. Eta horrekin batera beste adierazle garrantzizkoa da dokumentu luzeagoak sortzeko gaitasuna.

eta *Omniglot 'The online encyclopedia of writing systems and languages'* (http://www.omniglot.com).

Ordenagailu bidezko hizkuntza-euskarria ezagutzea oso esanguratsua da ikerketa honen helburuetarako. Horren inguruko datuak Microsoft-etik eta Apple-tik datoz, bi mailatan banatuta: datu-sarrera eta sistema eragilea. Datu-sarrerak idazketa-sistema bideratzeko metodo espezifikoetan zentratzen da, adibidez kotoeri hizkuntzak edo japonierak behar dituztenak. Sarrera-metodo egokirik gabe igoera digitala ezinezkoa da, baina alderantzizkoa ez da beti betetzen: alegia, sarrera-metodoa izateak ez du testua sortuko denik ziurtatzen, are gutxiago erabilera digital indartsua izango duenik. Sistema eragileare n mailako euskarriak bideratzen du sistemarekin interakzioak egitea, eta beraz, hizkuntza horretan izanez gero, esate baterako, menuen bidezko aukeraketak edo erroreen inguruko azalpenak hizkuntza horretan egin ahal izango dira.

Bestalde, hizkuntza askotako sarrera-metodoa estandarra izan arren ez dute idazkera baturik. Horrela, igoera digitalaren eskaileran hurrengo urratsa zuzentzaile ortografikoa sortzea da. Crúbadán proiektuaren ustez, faktore garrantzitsua denez hori, software libreko zuzentzailea dituzten hizkuntzak zerrendatzen dituzte. Ikerketa honetan *HunSpell* zuzentzaileen zerrenda aztertu da (software libreko zuzentzaile erabiliena -Németh L, Trón V, Halácsy P, Kornai A, Rung A, et al.. , 2004-), eta wikipediako edukiarekin gurutzatu da estaldura kalkulatzeko. Estaldura %50tik behera denean zuzentzailea ez-heldutzat hartu da.

Ortografia estandarizatua, wikipedia bezalako proiektuetan laguntzeaz gain, berez, bizindarra digitalaren adierazle garrantzitsua da. Eta horrekin batera beste adierazle garrantzizkoa da dokumentu luzeagoak sortzeko gaitasuna. Crúbadán proiektuak, berriz ere, esanguratsutzat hartzen du faktore hori, eta identifikatu du ea *Biblia* eta *Giza Eskubideen Aldarrikapen Unibertsala* eskuragarri dauden online hizkuntza jakinetan. Kontuan izan behar da testu luzeak biltzeko lagungarria dela idazkera estandarizatua, horixe baita hizkuntza-teknologiaren muina.

Hizkuntzaren bizindar digitalaren eta tresna linguistiko sofistikatuen arteko erlazioa eztabaidatzen da hurrengo atalean (etiketatzaile sintaktikoa, OCR tresna, hizketa-ezagutza, informazioaren erauzketa eta itzulpen automatikoa esaterako).

METODOAK

EGIDS eskala berak *igoeraren* nozioa dakar: *ahozko erabilera hutsetik* (6 kategoria) *alfabetizazio-prozesura* (5), eta hortik *ahozko erabilera indartsu* eta *alfabetizazio-prozesu iraunkorrera* (4). Hurrengo urratsak erabilerari (ofiziala

ala ez) begira daude: e*rabilera komunikabideetan baina barne mugak gainditu gabe* (3), *erabilera hezkuntzan, lan-munduan, komunikabideetan eta administrazioan eta ezagutza ofiziala lurralderen batean* (2), aurrekoa baina *ezagutza ofizial nazionala* (1), eta *nazioarteko erabilera merkataritzarako*, *ezagutza-trukerako* eta n*azioarteko politikagintzarako* (0) (Quakenbush JS, Simons GF, 2012). Alfabetizazio digitala ere ardatza da esparru digitalean, eta alfabetizazio hori nola eskuratzen den azalduz hasiko dugu atal hau.

Alfabetizazioaren lehen maila lokalizaziorako edo nazioartekotzerako (i18n adituen artean, *internationalization* kontzeptua laburtzeko) euskarri bat litzateke, bertako hizkuntzaren sarrera (idazketa) eta irteera (irakurketa) ondo definitzen duena. *Unicode* estandarraren barruan ehunka alfabeto desberdin daude jasota, eta baita berriak eransteko metodoak ezarrita ere. Horrek oinarri sendoa eskaintzen du hizkuntza berriak mundu digitalera eramateko, baita mundu horretan idatzia izateko ere. Hizkuntza bat Omniglot datu-basean (ik. aurreko atala) zerrendatuta baldin badago, maila hau gainditua duela esan daiteke. Baldintza samurragoa da online testuak izatea OLAC datu-basean, eta baldintza estuagoa testuak idazteko (sarrera) metodoren bat izatea.

Bigarren mailarako beharrezkoa litzateke hitz-mailako tresna sorta bat edukitzea, hiztegia, zuzentzaile ortografikoa eta lematizatzailea esaterako. Horretan dugun ebidentzia aipatutako *HunSpell* zerrenda hedatuena da (Németh L, Trón V, Halácsy P, Kornai A, Rung A, et al.. 2004), eta ez da hain egokia. Gainera, tresna horien kalitatea eta estaldura egiaztatzea, alegia, tresnaren heldutasuna neurtzen duten parametroak kontrolatzea eginkizun zaila da baliabide urriko hizkuntzentzat. Tresna horiek eskatzen duten hizkuntzaren estandarizazioa da zailtasun horren gakoa, tresna horiek ez baitute kontuan hartzen aldaera dialektalak eta indibidualak. Horrela, ingelesaren estandarizazioa ez zen iritsi XV. mendera arte, eta hemen aztertzen diren hizkuntza asko inoiz ez dira beharrak bultzatutako bide mingarri horretatik igaro, eta ez dute nahi kanpotik ezarritako eredu bat (Mapuche indians to Bill Gates 2018).

Bigarren mailarako beharrezkoa litzateke hitz-mailako tresna sorta bat edukitzea, hiztegia, zuzentzaile ortografikoa eta lematizatzailea esaterako. Ttresna horien kalitatea eta estaldura egiaztatzea tresnaren heldutasuna neurtzen duten parametroak kontrolatzea eginkizun zaila da baliabide urriko hizkuntzentzat.

Hirugarren mailak eskatzen ditu perpaus- edo esaldi-mailako tresnak, bigarren mailarako zehaztutako tresnetan oinarrituta soilik eraiki daitezkeenak. Tresna horien artean daude perpaus-mailako etiketatzaileak (*POS tagger* ingelesez), izendun entitateen ezagutzaileak (pertsona- edo toki-izenak identifikatzeko), hizketa-ezagutzaileak eta itzulpen automatikoa. Horri dagokionez, http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison gunean agertzen diren tauletan *bikain* kategorian ingelesa baino ez da agertzen. Gure ikerketan Google Translate ere aztertu dugu funtzio horren goranzko garrantzia islatzeko asmoz, baina maila hau lehen bi kategoriak (*oparo* (T) eta *bizi* (V)) bereiz-

Arlo digitalean hiztun batzuk esparru digitalean hizkuntza hori erabiltzen hasten direnean euren seme-alabek eta bilobek ere hala egingo dute automatikoki, eta horregatik EGIDSeko kategoria asko bakarrean kolapsatu daitezke.

teko da interesgarri, eta gure helburu nagusia beste bat da, bigarren, eta hirugarren kategoriak bereiztea, hain zuzen.

Ondorengo lerrootan, sailkapena egiteko arazoa nola ebatzi dugun azalduko dugu. GIDSen 8 kategoriak edo EGIDSen 13 kategoriak erabili ordez, hizkuntzek mundu digitalean duten aktibitate digitala mailakatzeko lau kategoria baliatu dira, digitalki *oparo* (T), *bizi* (V), *ondare* (H), eta *motel* (S) kategoriak hain zuzen. Hori dela eta, Lewis eta Simons-en (Lewis, Simons GF, 2010) erreferentzian proposatzen den zuhaitza erabat sinplifikatu dugu. Hizkuntza jakin bat kategoria batean edo bestean sailkatzeko honako datua izan dugu kontuan: ea esparru digitalean hizkuntza hori aktiboki erabiltzen den edo ez; eta, ondoren bi datu gehigarri. Datu nagusi hori kontuan izateak domeinu digitalari begira *hila*/*bizia* egoera bereiztea dakar, eta datu xeheagoek igoeraren bi graduen artean bereiztea (*bizia: bizi* (V) vs. *oparo* (T)) eta hildakoen bi graduen artean bereiztea (*hila*: *motel* (S) vs *ondare* (H)).

Hauxe izan da jarraitu dugun metodoa: sail bakoitzerako adibide prototipiko argiak hartu ditugu, eta ikasketa automatikoko teknika estandarrak erabiliz (entropia maximoan oinarritutako erregresio logistiko multinomiala zehazki -Hosmer D, Lemeshow S, 1989 eta Menard S, 2002-), hazi horiek erreproduzitzen dituen sailkatzaile bat eraiki dugu. Egindako esperimentuetan metodoaren eraginkortasuna txekeatu daiteke barneko irizpideez zein kanpokoez. Barneko irizpideetarako, sortutako sailkatzailearen kalitatea eta sendotasuna neurtzen da hazien perturbazioa eraginez, eta kanpoko irizpideetarako beste sailkapen eta multzokatze-teknikak erabiltzen dira.

EGIDSi dagokion sinplifikazioaren parte bat aipatutako balio demografikoetatik dator. Azterketa tradizionalean, EGIDSek bere sailkapenerako, bereziki, hizkuntzaren gaitasun osoa duen azken belaunaldia hartzen du kontuan: belaunaldi hori umeena bada hizkuntza *arriskuan* kategorian dago (6b), gurasoena bada *galeran* (7), aitona-amonena bada *hilzorian* (8a), bestela *ia-iraungita* (8b).

Arlo digitalean hiztun batzuk esparru digitalean hizkuntza hori erabiltzen hasten direnean euren seme-alabek eta bilobek ere hala egingo dute automatikoki, eta horregatik EGIDSeko kategoria asko bakarrean kolapsatu daitezke. Era berean, zenbaki txikiko kategoriak (0tik 3ra) kategoria bakar batean biltzea T (*oparo*) justifikatutzat jotzen dugu, estaldurak (nazioartekoa - nazionala - lurraldekoa) eta ofizialtasunak ez direlako hain garrantzizkoak esparru digitalean, nazioartekoa eta ez-ofiziala izatea definizioz datozen ezaugarriak direlako.

Txinera klasikoa, sanskritoa eta latinaren kasuek adierazten dutenez, desagertuta dauden hainbat hizkuntza ikuspegi digitalean hornituago

egon daitezke adiera tradizionaleko egoera *oparo*an baina digitalki *txiro* diren beste hizkuntza batzuk baino. Artxibo digitaletan egonik ere hiztunik ez duten hizkuntzak H kategorian (*heritage-ondare*) kokatzen ditugu. Aspaldian egoera horretan dauden halako hizkuntza batzuk Wikipediara heldu dira, baina soilik dokumentatzeko asmoa duten kasu berriak "Wikia" izeneko proiektura mugitu dira egun. Artxibo digitalak irauteko helburuarekin sortzen direnez, hizkuntza jakin bat behin *ondare* (H) egoerara iritsita ezin du egoera hori galdu; digitalizazioaren uholdeak ekarriko du hizkuntza asko *motel*-egoeratik (S) *ondare*-egoerara igarotzera, ageriko presentzia digitalik ez dutelako. Baina mugimendu hori ez da biziberritze moduan ikusi behar, bi norabideetako komunikazioaren ikuspuntutik behintzat, bi kategoria horietako hizkuntzak digitalki hilda baitaude. Hizkuntzen heriotzaren inguruko azterketa guztiek apurtu ezina den oinarri bera dute: *komunitaterik gabe biziraupenik ez*. Darwinek zioenez, Lyell aipatuz, "*A language, like a species, when once extinct, never (…) reappears*".

Txinera klasikoa, sanskritoa eta latinaren kasuek adierazten dutenez, desagertuta dauden hainbat hizkuntza ikuspegi digitalean hornituago egon daitezke adiera tradizionaleko egoera oparoan baina digitalki txiro diren beste hizkuntza batzuk baino.

Ondare hizkuntzak ez bezala, hizkuntza indartsuak bilatzea erraza da. Orokorrean T (*oparo)* eta V (*bizi*) kategorietako hizkuntzetarako bilioika (mila milioi) hitz topa ditzakegu, egunero milioika hitz berrirekin. Igoera digitalaren azken muturrean *oparo* motako hizkuntza bat L1 zein L2 komunitateetako hiztunek erabiltzen dute. Inplikazio hierarkiko zuzen bat agertzen da baliabide anitzeko hizkuntzetan: halako hizkuntza batek Apple sistema eragilerako euskarria baldin badu -Appleko sarrerako euskarria-, Microsofterako hizkuntza-paketea eta zuzentzaile ortografiko librea ere izan ohi ditu. Hizkuntza bakoitzari dagozkion mota horretako baliabideak zenbatzea (kontagailuari R deituko diogu) oinarrizko metodo bat izan daiteke eta Rren balioa merezimenduarekin lotu da. Bada, aurkitu dugunaren arabera 244 hizkuntzek bakarrik dute R>0; eta horien artean ehun inguru dira zalantzarik gabe bideragarriak.

Metodoan aipatutako haziak ezartzeko R erabili da. R maximoa (5) duten 16 hizkuntzek T_0 osatzen dute: ingelesa, japoniera, frantsesa, alemana, gaztelania, italiera, portugesa (Brasilgoa zein Europakoa), nederlandera, suediera, norvegiera (Bokmål), daniera, suomiera, errusiera, poloniera, txinera (tradizionala eta sinplifikatua) eta koreera. Hain zuzen ere, hauek dira Applek sistema eragilean onartzen dituen hizkuntzak eta irizpide hori prestigioaren isla da, aipatutako (ii) irizpidea (hizkuntzaren ospea), Apple beraren prestigioa handia baita komunitate digitalean. Wikipediako sailkapenean dauden lehen 16 hizkuntzen artean (v) irizpidea (wikipedia) hartuko bagenu T_1 hazi alternatibo gisa, helmuga berera iritsiko ginateke. Beste zentzuzko irizpide bat litzateke baliabide urriko hizkuntzen lehiakide nagusiak bilatzea. Scannell KP (2007) erreferentziaren arabera, lehiakide nagusiak dira ingelesa, gaztelania, frantsesa, errusiera,

Arrakala digitalaren beheko muturrerako (S kategoria) zerrenda guztietatik kanpo dauden hizkuntzak hautatu dira. Eta honako ezaugarriak dituzte: ez dute wikipediarik, ez dute Bibliarik ez zuzentzaile ortografikorik, ezta Apple-ko zein Microsoft-eko euskarririk ere; ez dute aipamenik Omniglot-en ez eta daturik ere ez Crúbadán proiektuan.

italiera, alemana, holandesa eta portugesa aipatutako sekuentzian; arabiera eta poloniera behin bakarrik aipatuta agertzen dira. Hori ere izan daiteke egoera *oparo*rako (T) hazirako irizpidea, eta ez litzateke oso desberdina izango.

Bizi (V) kategoria eskuz osatu dugu zalantzarik gabe egoera horretan dauden 84 hizkuntza hautatuz. Horien artean, zoriz, 40 osagaiko bi hazimultzo eskuratu dira, V_0 eta V_1. Horien artean hizkuntza hauek ditugu: banjar (bjn), eslovakiera (slk), guarania (gug), assamera (asm), bielorrusiera (bel), kirgizera (kir), chichewa (nya), armeniera (hye), hausa (hau), eta letoniera (lvs).

H_0 hazia ezartzeko *ondare* hizkuntzetarako, zalantzarik gabe egoera horretan dauden hizkuntza-multzo txiki bat hautatu da eskuz: aramera (arc), antzinako eliza-eslaviera (chu), koptoa (cop), manxera (glv), antzinako hebreera (hbo), txinera klasikoa (lzh), sanskritoa (san), and asiriera (syc). H1 hazi alternatibo gisa beste hauek aukeratu dira: antzinako ingelesa (ang), avestera (ave), kornubiera (cor), ge'ez (gez), latina (lat), mandaic (myz), paliera (pli), armeniera klasikoa (xcl), eta anglo-normandiera (xno). Aurreko kasuetan bezala, zalantzarik gabeko hizkuntzak baino ez ziren sartu multzo horietan.

Arrakala digitalaren beheko muturrerako (S kategoria) zerrenda guztietatik kanpo dauden hizkuntzak hautatu dira. Eta honako ezaugarriak dituzte: ez dute wikipediarik (ezta inkubagailu moduan ere), ez daude UDHR zerrendan, ez dute Bibliarik ez zuzentzaile ortografikorik, ezta Apple-ko zein Microsoft-eko euskarririk ere; ez dute aipamenik Omniglot-en ez eta daturik ere ez Crúbadán proiektuan. Aipatu behar da, nolanahi ere, horrek ez duela esan nahi ez dutela aktibitate digitalik, baizik eta egin diren saiakeretan ez dela haien arrastorik agertu. 6.541 hizkuntza dira eta haien artean bi azpimultzo sortu dira zoriz, S_0 eta S_1, bakoitzean 75 hizkuntza sartuz. Adibide tipikoak honakoak dira: renau (rea), terik (tec), ekialdeko limba (lma), naami (bzv), hegoko puget sound salish (slh), abure (abu), lavukaleve (lvk), tarao (tro), korupun.sela (kpq) eta lachi (lbt).

EGIDSeko sailkapena zenbakietara eramanez, BALIO altuko datuak (hiztunak, wikipediaren tamaina…) eskala logaritmikora eraman dira, eta bi datu-transformazio egin dira. Batetik, alfabetoaren arabera karaktere bakoitzean informazio kopuru desberdina metatzen denez gero, karaktere kopurua normalizatu da hizkuntzaren karaktereen entropiaren bidez. Adibidez, txinerako karaktere bati nederlanderaren lau karaktere dagozkio. Bestetik, wikipedian makinaz sortutako artikuluen eragina minimizatzeko, alemaneko 450 karaktereren baliokidea den kopurua gainditzen dituzten artikuluak baino ez dira kontuan hartu. Doikuntza horiekin wikipedia gehienen luzera heren bat laburrago geratzen da,

baina kasu batzuetan handitu egiten da, txekieraren kasuan esaterako. Volapük hizkuntzarena muturreko adibidea da, sailkapen ofizialean duen 37. postutik 163.era iragan baita.

EMAITZAK

S_0, H_0, V_0 eta T_0 lau hazietan oinarrituta entrenatzen dira hainbat sailkatzaile entropia maximoaren teknika erabiliz. Lehen sailkatzaileak 4 kategoriatan bereizten ditu (S-*motel*, H-*ondare*, T-*oparo* eta V-*bizi* kategoriekin); bigarrenak 3 kategoriatan (S-motel, H-*ondare*, VT-*bizioparo*) digitalki bizirik dauden hizkuntzak, V-*bizi* eta T-*oparo*, bilduz; hirugarrenak 3 kategoriatan (SH-*motelondare*, V-*bizi*, T-*oparo*), baina kasu horretan digitalki hildakoak bilduz; eta azkenak 2 kategoriatan (SH-*motelondare*, VT-*bizioparo*), aurretik bildutako bi multzoak bakarrik utziz.

Lehen sasilkatzaileak 4 kategoriatan bereizten ditu (S-motel, H-ondare, T-oparo eta V-bizi); bigarrenak 3 kategoriatan (S-motel, H-ondare, VT-bizioparo); hirugarrenk 3 kategoriatan (SH-motelondare, V-bizi, T-oparo) eta azkenak 2 kategoriatan (SH-motelondare, VT-bizioparo)

Aurretiazko emaitzak etsigarriak izan ziren, zeren eta, 10 karpetako balidazio gurutzatua[4] erabiliz emaitzen %40a baino ez ziren zuzenak. Baina, esan bezala, kopuru handiak eskala logaritmikoetara eramanda, emaitzen hobekuntza nabarmena lortu zen, doitasuna % 85-100 tartean geratuz (ikus 1. taula).

1.taula[5]: sailkapenaren doitasuna 10 karpetako balidazio gurutzatuz

	Seed 0				Seed 1			
# feat	SH-VT	S-H-VT	SH-V-T	S-H-V-T	SH-VT	S-H-VT	SH-V-T	S-H-V-T
33	95.0	99.3	92.3	90.7	99.3	98.6	94.3	87.9
18	97.2	99.3	91.4	96.4	99.3	98.6	95.0	89.3
10	97.9	99.3	92.9	95.7	100.0	99.3	93.6	90.0
8	97.1	99.3	92.9	97.1	100.0	96.4	94.3	85.7
6	97.1	99.3	92.1	93.6	100.0	96.4	95.7	89.3

doi:10.1371/journal.pone.0077056.t001

Entropia maximoaren ereduetan atributuen pisuak erabiltzen dira. Sailkatzen asko laguntzen duten atributuek pisu handiago dute besteek baino. Ondokoa hartu behar da kontuan: hasierako 33 atributuekin (# feat taulan) baino hobeto sailkatzen da pisu txikien duten atributuak kentzen badira. Horri atributu-hautapena deritzo, eta klasikoa da ikasketa automatikoan, emaitzak orokortzeko eta sailkapena azkartzeko (Pajkossy K, (2013). (Zehaztasunetarako ik. https://doi.org/10.1371/journal.pone.0077056.s002)

Atriburuei begira, espero zitekeenez, hizkuntzaren egungo egoera, EGIDS balioa da SIL erakundearen arabera aurresale onena (atributu onena), hautapen guztietan aukeratua izan zen-eta. Bigarren onena *Endangered*

170 hizkuntza inguruko multzo txiki bat identifikatu da (%2) esparru digitalari dagokionean igoera egoeran edo egoera sendoan dauden hizkuntzak biltzen dituena, eta 140 bat hizkuntza (%1,4) mugan daudenak.

Languages proiektuan kodetutako egoera izan zen (batean izan ezik guztietan hautatua) eta hurrengoak sekuentzia honetan: wikipediaren kalitatea, L1 hiztunen kopurua, Crúbadán bildumaren tamaina, zuzentzaile librearen presentzia eta OLAC datu-baseko online testu-kopurua. Azken hori, *ondare* gisa gordetzeko ahaleginaren adierazle onena gure ustez, %5ean baino ez da aukeratu, eta horrela argi geratu da igoera digitalari begira ez dela oso garrantzitsua.

Hamar sailkatzaile onenekin bozketa bat aurrera eramanez (*bagging* izeneko teknikaz (Breiman L., 1996)), sailkatzaileek *galzorian* egoera digitala esleitzen diete hizkuntzen %96ri (SH-VT kategoriak erabiliz). Azpimarratu nahi dugu heriotza masiboaren aurresan hori ez dela etorkizunerako aurreikuspena, politika egokien bidez saihestu edo murriztu baitaiteke, baizik eta dagoeneko gertatu den zerbait. Talde horretatik kanpo 170 hizkuntza inguruko multzo txiki bat identifikatu da (%2) esparru digitalari dagokionean *igoera* egoeran edo egoera *sendoan* dauden hizkuntzak biltzen dituena, eta 140 bat hizkuntza (%1,4) *mugan* daudenak. Hurrengo atalean eztabaidatuko dugu horretaz.

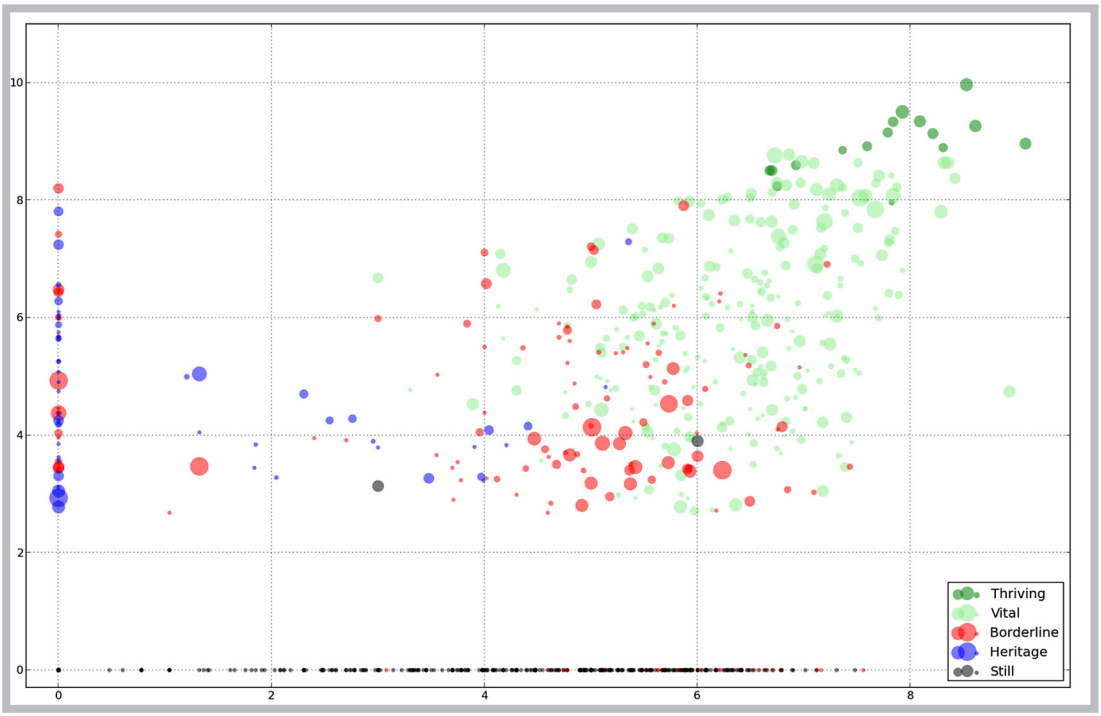
EZTABAIDA

Igoera digitalean gertatzen den arazoen inguruko epai argia lortzen den bitartean, egoera *motel* edo *biziaren* arteko mugan dauden hizkuntzen arteko bereizketa egiteko ez da nahikoa sailkatzaile baten epaia. Ziur asko, hizkuntza bakoitzerako gure datuetan kodetu gabe dauden gertaerak kontuan hartu beharko dira, eta horien arabera aurreikusi mugan dauden hizkuntzen etorkizun digitala. Hizkuntza bakoitzaren etorkizuna irekia da, baina panorama orokor sendoa marraztu dugu, metodoan hainbat aldaketa txiki eginda ere.

Egoera *motelean* dauden hizkuntzen bizi-iraupena bide aldapatsua den bitartean, egoera *bizi*ko hizkuntzek bere iraupena ziurtatuta dute. Hori dela eta, aipatutako 3 kategoriako sailkatzaileekin ere egin dugu proba, S-H-VT kategoriekin, egoera *motel* eta *ondare* egoerak ondo bereizteko. Modu horretan trebatutako sailkatzaileak 6-8 atribututan oinarrituta dabiltza hobekien, %97,1-%100 tarteko doitasunarekin. Sendotasuna handitzeko esperimentu kopuru handi bat eraman da aurrera, S egoerako hizkuntzen hautapen desberdin askorekin. Esperimentu guzti horien emaitzak bilduta ondorio argi batera iristen da: 8.000 hizkuntza baino gehiago digitalki hilda daude, laurden bat gehiago arrasto digitalik gabeko 6.541 hizkuntzak baino. H (*ondare*) egoeran 300 bat (aldakuntza-marjina handiarekin) hizkuntza leudeke, eta VT kategorian (*bizioparo)* 420 gehienez. Labur esanda, heriotza-tasa %95,5 litzateke, %0,4ko akats-tartearekin.

Lehen irudian S-H-VT sailkatzaile biren emaitza islatzen da. Wikipedian dauden hizkuntzak, inkubagailu egoerakoak barne, bakarrik marrazten dira, abzisetan (x) agertuz hiztun kopurua eskala logaritmikoan eta ordenatuetan (y) wikipediaren tamaina normalizatua, eskala logaritmikoan ere. T (*Thriving-oparo)* egoerako hizkuntzak azaltzen dira, V (*vital-bizi*) egoerakoak (T-V banaketa egin da gero), Hkoak (*heritage-ondare),* Skoak (*still-motel)* eta *borderline-mugan* daudenak. Mugan daudenak sailkatzaile bietan sailkapen desberdina zuten.

1. irudia. Sailkapenaren grafikoa, Wikipediaren tamainan eta hiztun kopuruan oinarrituta[6].



T (*thriving-oparo*) kategoriako 16 hizkuntzek etorkizun digital oparoa dute, 100 urtera begira behinik behin. EGIDS sailkapenean 0 balioa duen hizkuntza bat, arabiera, ez da T kategorian sailkatu, digitalki indartsu izan arren, ez duelako euskarririk Apple-en aldetik eta datuak bildu zirenean ez zegoelako lehen 20 wikipedien artean.

V (*vital-bizi*) kategorian sailkatutako 252 hizkuntzen artean hazi moduan erabilitako 83ak digitalki indartsuak dira zalantzarik gabe. Esperientziak erakusten duenez, guztien artean gaur egun ez dira 150 baino gehiago egoera *bizi*an daudenak. Hizkuntza horiek esparru digitalean *oparo* bihurtzeko lan handia dagoen bitartean (adibidez, hungarierak badu PCtan euskarria, baina Mac-etan ez), zalantza gutxi dugu horietako batzuez;

BAT Soziolinguistika Aldizkaria 111, 2019 (2) | 11-32 | Andoain (Gipuzkoa)
ISSN 2659-5176 / ISSN-L 1130-8435 | Lege gordailua: SS120/90

Munduko hizkuntzak igoera digitalari begira automatikoki sailkatu dira, eta gure ondorioa da igoera digitalera iristen diren hizkuntzak gehienez %5 direla.

digitalizazioaren itsasgoraren eraginez, gutxienez erdia, gehiago agian, egoera onera pasako dira. Kategoria horretan daude EGIDSeko 1 kategorian daudenen arteko bi heren (%66), 2 kategoriako ia erdia (%46), 3 kategoriako %13, 4ko, %8 eta 5eko % 2. Gainontzekoen artean %1a bakarrik agertzen da.

Borderline-muga modura marraztutako 162 hizkuntzen kasua azpimarratu nahi dugu, sailkapen automatikoari dagokionean zalantzak sortzen dituztelako (sailkatzailearen arabera kategoria batean edo bestean kokatu daitezke).

H (*heritage-ondare*) kategorian wikipedia duten 51 hizkuntza biltzen dira. Horietako batzuk, aipatu bezala, hiztunik gabeko hizkuntzak dira. EGIDS eskalan duten batez-bestekoa 7,83koa da, dagokien *ondare* egoera baieztatuz. Adibide gisa hazietan ez zeuden batzuk zehaztuko ditugu ondoren: creera (cre), dalmaziera (dlm), erdiko nederlandera (dum), ido (ido), gotikoa (got), antzinako eskandinaviera (non), pipil, prusiera zaharra (prg), romagnol-a (rgn) eta samogitian (sgs).

Beltzez marraztutako 307 hizkuntza daude *natibo* digitalen arrastorik gabe. Hiztunen kopurua, batez-beste, 0,7 milioikoa da, eta wikipedia gehienak inkubagailu egoeran daude (hutsak normalizazio-kalkuluak egin ondoren). Kanuri (kau) hizkuntza da horren adibidea, bere hiru dialektoekin, tumari (krt), manga (kby) eta beriberi (knc) EGIDS sailkapenean 6a, 5 eta 3 egoerekin hurrenez hurren. Erabilera indartsua du, 3,76 milioi hiztun, irratia eta telebista, eta erdiko dialektoa behintzat ez dager inolako zerrendetan *galzoriko* egoeran. Baina wikipedia itxi zen bertako hiztunen eduki-ezagatik eta komunitate birtualik ez edukitzeagatik. Eta Crúbadán proiektuan 5.000 hitz biltzen duten hiru dokumentu baino ez ziren bildu.

ONDORIOAK

Munduko hizkuntzak igoera digitalari begira automatikoki sailkatu dira, eta gure ondorioa da igoera digitalera iristen diren hizkuntzak gehienez %5 direla. Hizkuntzen eta hizkuntza-familien arabera egoera konplexu samarra da, eta ez da erraza ondorioak laburtzea. Gure estimazio subjektiboaren arabera *inkubagailu* egoeran dauden hizkuntzen artean heren batek baino gutxiagok burutuko dute aro digitalerako trantsizioa. Klingon hizkuntza zaharra adibide gisa jar dezakegu, ikusteko lagun-talde sutsu bat saiatu daitekeen arren, zailagoa dela komunitate trinko bat sortzea. Wikipediaren hizkuntzen gaineko politikak (https://meta.wikimedia.org/wiki/Language_proposal_policy) eskatzen du "hizkuntza horretan erregularki editatuko duten bost erabiltzaile aktibo izatea gu-

txienez arrakastatsutzat hartzeko egitasmo hori (*at least five active users must edit that language regularly before a test project will be considered successfu*l)"; gure iritziz, ezin bihozberagoa da baldintza hori, izan ere, benetako langa askoz altuagoa da. Wikipedia toki egokia da digitalki aktiboak diren hiztunak biltzeko, baina ahalegin horretan sortzen den ohiko emaitza *ondare* motakoa izan ohi da.

Wikipedian lankidetzaz aritzen diren editorez osatutako komunitate bat beharrezkoa da hizkuntza eta kultura web sarera ekartzeko, baina ez da nahikoa hizkuntzaren benetako iraupenerako; testuinguru digital aberatsa behar du *igoera digitalak*, definizioz. Honek ez du ukatzen ondarearen preserbazioaren balioa, baina aro digitalean, hizkuntzaren biziraupena hizkuntzaren estandarrarekin dago lotuta, ospea eta funtzio nagusiak biltzen dituen aldaerarekin, hain zuzen.

Aurrekoaren adibide tipikoa piamontera hizkuntza da; 2-3 milioi hiztun ditu Torino inguruan, eta, bertako administrazioaren aldetik badu halako estatus ofizial bat ere; baina ez du presentzia digital nabarmenik. Agian, komunitate itxiago batzuek aukera handiagoa dute, esate baterako faroera hizkuntzak. 50 mila hiztun ditu eta kalitate handiko wikipedia ere garatu dute. Itxaropenerako zantzuak ere badaude esate baterako mendebaldeko flandrieraren kasuan, jakitera eman baita 40.000 aldiz jaitsi dela hizkuntza hori ikasteko app-a. Dena den, orokorrean, aro digitalean hedadura handiko beste hizkuntza nagusi batekin lehiatzen duten hizkuntzentzat aukerak nahiko urriak dira, partikularki Erresuma Batuko hizkuntza gutxituentzat.

Kasu bakanetan, kurduerarenean esaterako, sumatzen da estandar baten agerpena, hizkuntzaren hiru bertsio (iparraldeko kurmanji, erdialdeko sorani, eta hegoaldeko kermanshahi) mantentzen diren egoeran ere. Baina "via regia" modukorik ez dago aro digitalean. Azterketa hau sinkronikoa den arren, tradiziozko alfabetaziorako eta alfabetizazio digitalerako bide diakronikoa ezaguna da. Bide horretan parte hartzen dute honako aldagaiek: argitaratzeko azpiegitura minimoak, estandarra sortzeko beharrak, urte asko hartzen duen heziketa formala zein hizkuntzalaritza konputazionaleko adituen trebakuntzak eta horiek garatzeko beharrezkoak diren tresnak. Agian adibide azpimarragarriena euskararena da, hain zuzen, Europar Batasunaren ikuspegi zabalaren babesa duena. Kontuan izan behar da halere, halako arrakasta-kasuak izatea zaila dela, bereziki ekonomikoki suntsituago eta linguistikoki aniztunago diren hizkuntza-komunitateen artean.

Itzulpen automatikoari dagokionez, badakigu zerbitzu hori gero eta garrantzitsuagoa dela hizkuntzen arteko komunikaziorako. Normalean tresna hori egoera *oparo*ko hizkuntzei lotuta agertzen da, Haitiko kreo-

Itzulpen automatikoari dagokionez, badakigu zerbitzu hori gero eta garrantzitsuagoa dela hizkuntzen arteko komunikaziorako. Normalean tresna hori egoera oparoko hizkuntzei lotuta agertzen da.

Begien aurrean duguna, ez da bakarrik munduko hizkuntzen heriotza masibo bat, iraultza neolitikoaren azken ekitaldia baizik.

leraren salbuespenarekin (frantsesera itzultzeko) (Spice B., 2012). Halako teknologiarako Google-k neurri handiko corpus elebakarrak eta elebidunak izatea baldintza modura jartzen du. Eta jakina, egoerako *bizi*ko hizkuntzentzat, akaso, betebehar hori ez da arazo izango, baina *mugan* dauden hizkuntzentzat hori benetako oztopoa da.

Egoera digitala askoz okerragoa da kontsentsuko zenbakiek (2.500-3.000 hizkuntza galzorian) adierazten dutena baino. Ikerketa pesimistenek ere hizkuntzen %10, 600 bat, arriskutik kanpotzat jotzen dute (Krauss M., 1992), baina adituak gero eta gehiago mintzatzen dira ideia horren aurka (Poser W., 2012). Errealitatean, gaur egun 250 hizkuntza baino gutxiago daude digitalki goiko aldean, eta *mugan* dauden erdiak baino gehiagok Marokoko arabieraren (ary) antza dute. Alegia, hizkuntza nagusi baten dialekto mintzatuak dira, ospe txikia dutenak, baina *akroleto*aren[7] ospearekin batera bizindarra erakusten ari direnak. Horretan oinarrituta iradoki daiteke heren bat baino gutxiago izango direla etorkizunean *bizi*ak.

Litekeena da beste 20 bat hizkuntza mintzatu ere igotzea, gaur egun wikipediako inkubagailu egoeran edo aurrekoan izan arren. Baina edozein kasutan aldapa gorako borroka egin beharko dute. Adiera klasikoaren arabera, bizirik dauden 7.000 hizkuntzen artean, agian 2.500ek iraungo dute bizirik beste mende bat. Baina 250 hizkuntza baino ez dute digitalki iraungo, eta besteak, ezinbestean, *ondare* egoerara (Nynorsk) edo desagerpen digitalera (Mandinka) mugituko dira. Horren ondorioz babeste lanak (http://www.endangeredlanguages.com) oso garrantzitsuak dira.

Zoritxarrez, ikuspuntu praktikotik, ondareari begirako proiektuak (inkubagailu moduko wikipediak barne) planifikaziorik gabekoak izan ohi dira, ez baitute dokumentazio-programa sistematikorik. Baliabideak maiz alferrik xahutzen dira, aurretik gertatutako galera funtzionalak eta hizkuntza-aniztasunaren aurka doazen pizgarri ekonomikoak kontuan hartuta (Ginsburgh V, Weber S (2011).

Begien aurrean duguna, ez da bakarrik munduko hizkuntzen heriotza masibo bat, iraultza neolitikoaren azken ekitaldia baizik. Adibide gisa har dezagun komiera hizkuntza, wikipedia duten bi aldaerekin (permyak, 94.000 hiztun, eta zyrian, 293.000). Aldaera bakoitzean hainbat dialekto daudenez, batzuk dagoeneko desagertuta eta beste batzuk argi eta garbi egoera *motel*ean daude. Esperantza argiena hiri nagusian dago, Syktyvkar-en, bertako aldaera estandar bihurtu delako. Behin ortografia estandarizatuta, unibertsitateak (hezkuntzarako hizkuntza nagusia errusiera duena) hizkuntzalari konputazionalengana jo dezake zuzentzaile ortografikoa sortzeko eta alfabetazio digitalerako lehen urratsak emateko (Prószéky G, Novák A., 2005). Baina horrek, alde batetik, onura ekarriko die estandarraren hiztunei, baina, bestetik, ziur asko, mendi aldeko eta

ospe txikiko zyryan aldaera atzean uztea ekarriko du seguruena. Gainera, komierarako azaldutako jokalekua optimista dela hartu behar dugu kontuan; ehunka mila hiztun ditu, oraindik populazioaren laurdena, eta unibertsitatea ere badago. Horrez gain, eskualdea garatzeko pizgarri ekonomikoak ere badira (petrolioa, zura). Baina munduko hizkuntzen %95ean ez dute aldeko faktore horietako bat edo gehiago, eta eten digitala saihesteko esperantza txikia da. ●

OHARRAK

1. Itzultzailearen oharra: Beñat Garaiok proposatutako terminologian oinarrituta: (http://www.soziolinguistika.eus/files/benat_garaio_bat97.pdf). Jatorrizkoa: 0. International; 1. National; 2 Provincial; 3 Wider communication; 4 Educational; 5 Developing; 6a Vigorous; 6b Threatened; 7 Shifting; 8a Moribund; 8b Nearly Extinct; 9 Dormant; 10 Extinct.

2. *Actively engaged in digitally mediated interaction* jatorrizkoan

3. Itzultzailearen oharra: jatorrizko artikuluan zehaztasunak ematen dira % 10 horretaz

4. *10-fold cross-validation* jatorrizkoan.

5. Seed -> hazia, # feat -> atributu kopura.

6. Editoreen oharra: Grafikoan agertzen diren gris koloreak ez dira behar bezain argigarriak. Jatorrizko koloretan ikusteko egin daiteke kontsulta webguneko bertsioan; http://www.soziolinguistika.eus/BAT111 edo jatorrizko bertsioan https://doi.org/10.1371/journal.pone.0077056

7. Aldakuntzen artean edo continuumean aukera prestigiodunena da akroletoa

BIBLIOGRAFIA

Ad hoc technical expert group on indicators for assessing progress towards the 2010 biodiversity target UNEPCoBD. (2004). *Indicators for assessing progress towards the 2010 target: status and trends of linguistic diversity and numbers of speakers of indigenous languages*.

Bickerton D (1981). *The Roots of Language*. Ann Arbor: Karoma, 351 pp.

Bloomfield L (1927). Literate and illiterate speech. *American speech* 2: 432–439.

Breiman L (1996). Bagging predictors. *Machine Learning* 24: 123–140.

Cazden CB (2003). Sustaining indigenous languages in cyberspace. *Nurturing Native Languages* : 53–57.

Crystal D (2002). *Language death*. Cambridge University Press, 210 pp.

Darwin C (1871). *The Descent of Man, and Selection in Relation to Sex*. London: John Murray. .

Dorian NC (1981). *Language death: The life cycle of a Scottish Gaelic dialect*. University of Pennsylvania Press Philadelphia, 202 pp.

Fishman JA (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened language*, volume 76. Multilingual Matters Ltd, 431 pp.

Frank RM (2008). The language-organism-species analogy: A complex adaptive systems approach to shifting perspective on 'language'. In:

M FR, Dirven R, Ziemke T, Bernárdez E, editors, *Body, Language and Mind*. Vol. 2. Sociocultural Situatedness, Berlin: Mouton de Gruyter. 215–262.

Ginsburgh V, Weber S (2011). *How many languages do we need?: The economics of linguistic diversity*. Princeton University Press.

Hiddinga A, Crasborn O (2011). Signed languages and globalization. *Language in Society* 40: 483–505.

Hosmer D, Lemeshow S (1989). *Applied logistic regression.* Wiley.

Izreel S (2003) The emergence of spoken Israeli Hebrew. In: Hary B, editor, *Corpus Linguistics and Modern Hebrew: Towards the Compilation of the Corpus of Spoken Israeli Hebrew,* Tel Aviv University Press. pp. 85–104.

Kegl J (1994). The Nicaraguan sign language project: An overview. *Signpost* 7: 24–31.

Krauss M (1992). The world's languages in crisis. *Language* 68: 4–10.

Lewis MP, Simons GF (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistiqu*e 2: 103–119.

Mapuche indians to Bill Gates: hands off our language. http://www.smh.com.au/news/biztech/mapuche-indians-to-bill-gates-hands-off-our-language/2006/11/[24-November-2006, accessed 27-August-2012].

Menard S (2002). *Applied logistic regression analysis.* Sage Publications, 128 pp.

Morpurgo-Davies A (1998). *History of Linguistics.* Vol. IV: Nineteenth-Century Linguistics. London and New York: Longman, 464 pp.

Mosley C (2010). *Atlas of the World's Languages in Danger.* UNESCO Publishing, http://www.unesco.org/culture/languages-atlas. [accessed 17-July-2013].

Németh L, Trón V, Halácsy P, Kornai A, Rung A, et al.. (2004). Leveraging the open source ispell codebase for minority language analysis. In: Carson-Berndsen J, editor, Proc. *SALTMIL.* pp. 56–59.

Nettle D, Romaine S (2000). *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press, 243 pp.

Pajkossy K (2013). *Studying feature selection methods applied to classification tasks in natural language processing.* MSc thesis, Eötvös Loránd University.

Poser W (2012). Personal communication.

Prensky M (2001). Digital natives, digital immigrants part 1. *On the horizon* 9: 1–6.

Prószéky G, Novák A (2005). Computational morphologies for small Uralic languages. Inquiries into Words, Constraints and Contexts Festschrift in *the Honour of Kimmo Koskenniemi on his 60th Birthday* : 116–125.

Quakenbush JS, Simons GF (2012). Looking at Austronesian language vitality through EGIDS and SUM. In: *Proc.* 12-ICAL.

Rehm G, de Smedt K (2012). The Norwegian Language in the Digital Age: Norsk i Den Digitale Tidsalderen. *Springer*, 81 pp.

Requests for new languages/Wikipedia Mandinka. http://meta.wikimedia.org/wiki/Requests _for _new _languag[14-September-2008, accessed 17-July-2013].

Scannell KP (2007). The Crúbadán Project: Corpus building for underresourced languages. In: *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*. volume 4, pp. 5–15.

Schmidt A (1985). The fate of ergativity in dying Dyirbal. *Language* : 378–396.

Senghas A, Kita S, Özyürek A (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science* 305: 1779–1782.

Simons GF, Lewis MP (2013). *A global profile of language development versus language endangerment*. http://www-01.sil.org/~simonsg/presentation/Simons%20and%20Lewis%20ICLDC%202013.pdf. [Online, accessed 27-June-2013].

Spice B (2012). *Carnegie Mellon releases data on Haitian Creole to hasten development of translation tools.* http://www.eurekalert.org/pub_ releases/2010-01/cmu-cmr012710.php. [Online, accessed 27-August-2012].

Thangali A, Nash JP, Sclaroff S, Neidle C (2011). Exploiting phonological constraints for handshape inference in ASL video. In: *Proc*. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 521–528.

Varga D, Halacsy P, Kornai A, Nagy V, Nemeth L, et al.. (2007). Parallel corpora for medium density languages. In: Nicolov N, Bontcheva K, Angelova G, Mitkov R, editors, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP*-05, Amsterdam: Benjamins. 247–258.

Zséder A, Recski G, Varga D, Kornai A (2012). Rapid creation of large-scale corpora and frequency dictionaries. In: *Proceedings to LREC 2012* pp. 1462–1465.