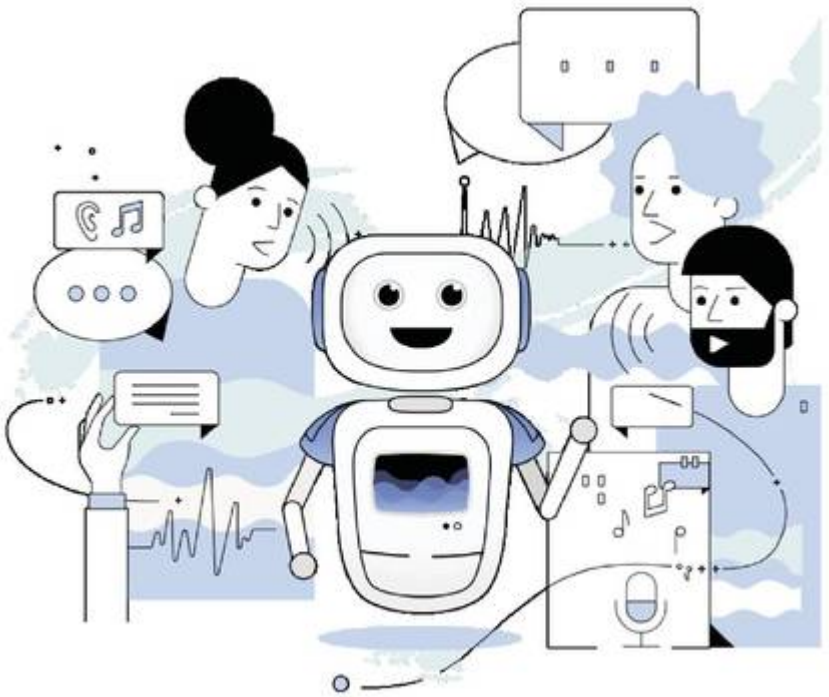
TEKNOLOGIA INTERNET SOFTWAREA

Auzolan digitala euskararen alde

Leturia Azkarate, Igor
Informatikaria eta ikertzailea
ELHUYAR HIZKUNTZA ETA TEKNOLOGIA

Gure bizitzetan, gero eta txertatuago daude hizkuntza- eta hizketa-teknologiak baliatzen dituzten tresnak eta zerbitzuak: laguntzaile birtualak, bozgorailu adimendunak, itzultzaile automatikoak... Teknologia horiek garatzeko, baliabideak behar dira, baina ez bakarrik ekonomikoak; bereziki, sistemak entrenatzeko hizkuntza-baliabideak dira guztiz-guztiz beharrezko: audio-grabazioak, elkarrizketen adibideak, itzulpenak... Beste hizkuntza hedatuago batzuetan baino urriagoak dira euskaran, eta, horregatik, berriki abiatutako ekimen batzuek erakusten duten bezala, baliabideok sortzeko, crowdsourcing-era jotzen ari gara azkenaldian. Anglizismo hotsandiko horren atzean, finean, gure artean hain errotua dagoen auzolana besterik ez dago; auzolan digitala, kasu

honetan.



Arg. Common Voice

Nork ez du noizbait erabili laguntzaile birtual edo elkarrizketa-agente bat? Gure telefono mugikor eta ordenagailuetan lehenespen gisa instalatuta etortzen dira Siri, Google Assistant, Cortana eta enparauak, eta nik neuk, adibidez, proba egiteko besterik erabili ez badut ere, oso ohikoa da belaunaldi gazteek halakoak erabiltzea. Testu bidezko elkarrizketa-sistemak ere (txatbot izenez ere ezagutzen ditugunak) gero eta ohikoagoak dira webguneetan, app-etan eta mezularitza-programetan, hala nola Whatsapp-en. Itzulpen automatikoa eguneroko baliabide bihurtu zaigu ia, menperatzen ez dugun hizkuntza batean dagoen testu bat ulertzeko edo testu bat beste hizkuntza batean sortu behar dugunean, behintzat, zuzentzeko lehen bertsio bat izateko. Zerbitzu eta webgune anitz daude horretarako, eta app eta webguneetan integratuta ere etortzen dira itzultzaile automatikoak. Audio eta bideoak automatikoki transkribatu edota azpititulatu ere egiten dira.

Zer ezaugarri komun dute aipatutako adibide horiek guztiek? Bi gauza bederen, bai: bata, den-denak hizkuntza- eta hizketa-teknologietan oinarrituta daude; bestea, euskaraz ez dago horrelako sistemarik edo, oro har, beste hizkuntzetan baino okerrago funtzionatzen dute.

Azken horren zergatietako bat, jakina, ekonomikoa da. Diru eta giza baliabide askoz gehiago esleitzen zaizkio halako teknologiak hizkuntza handietan ikertu eta garatzeari,

hizkuntza handien tamaina, botere eta hedapenagatik, eta askoz gutxiago euskaraz garatzeari. Baina beste arrazoi bat ere bada: alde handia dago grabazioen, itzulpenen, elkarrizketa-adibideen eta horrelako hizkuntza-baliabideen eskuragarritasunean. Hizkuntza hegemonikoek askoz baliabide gehiago dituzte eskura euskarak baino.

Izan ere, gaur egun, hizkuntza- eta hizketa-teknologiak garatzeko gehien erabiltzen diren eta emaitza hoberenak ematen dituzten metodoak adibideetan oinarritzen dira. Bereziki, sare neuronal sakonen teknologia (*deep neural networks*) erabiltzen da, gaur egun, teknologia horietan, frogatu baita kalitaterik onena horiekin lortzen dela. Eta sistema horiek adibide mordoa behar izaten dituzte, haietatik nolabait ikasi eta funtzionatu ahal izateko. Sare neuronal bidezko itzulpen automatikoko sistema batek itzulpen-adibide asko behar ditu, entrenatu eta ongi ibiltzeko; elkarrizketa-sistema batek, elkarrizketen adibide asko, eta transkribapen-sistema batek, transkribatutako audioen adibide asko. Horregatik dira hain beharrezko aipatutako hizkuntza-baliabide horiek, eta horregatik dabiltza okerrago horrelako baliabide gutxiago dituzten hizkuntzetako sistemak.

Euskaldunok tematiak izaki, gure hizkuntzan ere izan nahi ditugu beste hizkuntzek dituzten tresnak eta zerbitzuak, eta horretarako hizkuntza-baliabideak sortu behar direnez, hainbat ekimen jarri dira martxan berriki, horiek *crowdsourcing* bidez sortzeko. Zerbait lortzeko pertsona askoren lankidetza baliatzea esan nahi du *crowdsourcing*ak; bereziki, Interneten garapenarekin garatu da jarduera, jende-multzoen komunikazioa eta koordinazioa errazten baitu. Baina izen horren atzean, azken finean, gurean aspalditik baliatzen dugun auzolana besterik ez dago; kasu honetan, auzolan digitala (Librezale elkartekoek erabiltzen dute termino hori, jarraian azalduko dugun Common Voice ekimena izendatzeko).

Euskarazko Common Voice ekimena

Euskararentzat baliabideak sortzeko azken proiektuetako bat da Common Voice. Izatez, ez da Euskal Herrian bertan sorturiko ekimena, Mozilla Fundazioak abiarazitakoa baizik. Mozilla Fundazioak, zeina Firefox nabigatzaile librearen atzean dagoen erakundea baita, web ireki eta libre bat lortzea du helburu, eta jende guztiarentzat irisgarriago egin nahi ditu, besteak beste, Firefox nabigatzailea bera eta bestelako gailu eta tresnak. Horretarako, hizketa-ezagutzarako teknologia librea sortu nahi du ahalik eta hizkuntza gehienentzat.

Common Voice proiektuaren bidez, jendeak ahots-grabaketak ematen ditu dohaintzan, gero hizketa-ezagutzako sistemak garatu ahal izateko. Grabaketa horiek libreak dira; beraz ez Mozillak bakarrik, beste edozeinek ere balia ditzake hizketa-ezagutzako teknologia garatzeko. Mundu osoko jende ugari ari da hainbat hizkuntzatan grabaketak egiten, Common Voice proiektuan: 2.000 ordu inguru grabatu dira 28 hizkuntzatan, eta beste hizkuntza batzuk bidean daude.

Euskara IKTen munduan bultzatzeko helburua du Librezale elkarteak, eta software librea lehenesten du. Otsailean abiarazi zuen Common Voice proiektuaren barruan euskarazko grabaketak egiteko ekimena. Hasierako lanak egin zituen Librezalek (webgunea itzultzea, grabatzeko esaldien bilduma osatzea...), eta behin martxan jarrita, ekimena sustatzen, maratoiak antolatzen eta beste hainbat jarduera aurrera eramaten aritu da, hainbat eragileren laguntzarekin: Argia, iAmetza, EHUko IXA eta Aholab taldeak, Garabide, Elhuyar Fundazioa... Lan eskerga egin da, eta fruituak ematen ari da: proiektua abiarazi eta lau hilabetera, 508 erabiltzaileri esker, 83 ordu zeuden grabatuta, eta horietako 45, baliozkotuta. Ez dago batere gaizki, kontuan izanik, garai berean eta lehenago hasita, gaztelaniaz, adibidez, 32 ordu zeudela eginda; italieraz, 35 ordu; nederlanderaz, 21 ordu... Lortu nahi diren 1.200 orduetatik urruti gaude oraindik, baina bide onean doa, zalantzarik gabe. Ekimenean lagundu nahi baduzu, sartu https://voice.mozilla.org/eu helbidean, eta esaldiak grabatu edo daudenak baliozkotu.

IXA taldearen elkarrizketa-bilketa

Euskal Herriko Unibertsitateko IXA taldean ere auzolan digitalaren bidea hartu dute, euskararako txatbot edo elkarrizketa-sistema bat garatzeko. Zehazki, txatbot bat garatu nahi da, erabiltzailearen informazio-eskaerei informazioa Interneten bilatuz erantzungo diena, elkarrizketa ahalik eta modu naturalenean izanik. Ekimena ikerketa-proiektu baten barruan garatuko da: Eneko Agirre eta Aitor Soroa irakasleek gidatzen dute, eta Jon Ander Campos eta Arantxa Otegi ikertzaileek eta Aitor Agirre master-ikasleak hartzen dute parte. Gainera, Google erakundeak urtero ematen dituen ikerketa-sarietako bat jaso du (Google Faculty Research Awards). Proiektua ingelesez egindako elkarrizketetan oinarrituta dago, baina beste hizkuntza batzuetan ere garatzeko baliatuko da.

Esan bezala, horrelako sistema bat garatzeko, elkarrizketa errealen adibide asko behar da,

eta adibide horiek euskaldun boluntarioen ekarpenaren bidez osatu nahi izan dituzte . Horretarako prestatu zuten webgunea k binaka jarri zituen erabiltzaileak; batek Wikipediako artikulu baten inguruko galderak egiten zituen, eta besteak erantzunak ematen zizkion, 10 minutu inguruko saioetan. Hau litzateke horrelako elkarrizketa baten adibide bat, Wikipediako Korrikari buruzko artikuluan oinarritua:

- Zer da *Korrika*?

- *Korrika* euskararen alde Euskal Herrian zehar lasterka egiten den martxa da.

- Zer luzera du?

- Ibilbidea aldatu egiten da, baina beti 2.300 kilometro inguru izan ohi da.

- Zenbat denbora?

- Bi aste inguru.

- Gelditu gabe?

- Bai, martxa ez da inoiz geratzen, ez gauez, ezta egoera klimatologiko txarrengatik ere.

Arg. IXA

Adibide-bilketa ekainean egin zen, 400 elkarrizketa jasotzeko asmoa zuten, eta 356 elkarrizketa jaso zituzten. Ez da gutxi! Jasotako elkarrizketak libre jartzeko asmoa dago, beste edonork beste edozein proiektutan erabili ahal izan ditzan.

Argi dago halako ekimenak oso interesgarriak eta beharrezkoak direla etorkizunerako. Hain gurea den auzolana mundu digitalean ere behar bezala aurrera eramaten asmatzen

badugu euskaldunok, lortuko dugu makinekin euskaraz jardutea, ziur.

Proiektu honen laguntzaile izan nahi baduzu, **harpidetu.** Urtean 20 € baino ez.