

Larramendiren *Hiztegi Hirukoitzaren* digitalizazioa. Karaktereen ezagutze optikoa eta Wikitekara igotzea

Mikel Alonso, David Lindemann
UPV/EHU

Artikulu honetan Larramendiren *Hiztegi Hirukoitzaren* digitalizazioko OCR prozesua deskribatzen da, adimen artifizialaren adarra den ikasketa automatikoa baliatuz. Horretarako, eskaneatutako irudien aurreprozesamendua deskribatzen da, eta ondoren, *Kraken* erreminta baliatuz, eskuz transkribatutako laginetik abiatuta hiztegi testua ezagutuko duen ereduaren trebakuntza azaltzen da. Doitasun handiko testuaz gain, letra etzana eta testuaren posizioa gordetzen dituzten fitxategiak sortu dira, hiztegiaren egitura irudikatzeko balioko dutenak. Emaitzak prozesatu eta *Wikiteka* plataforman eskuragarri jarri direnez, auzolanez transkripzio osoa zuzendu daiteke. Zuzendutako transkripzio hori informazio-erazketa prozesutik pasako da, hiztegiaren egitura lexikografikoa ikasketa automatikoz erauzteko. Informazio horrekin RDF estandararekin bat datorren moldaketa-eredu baten lehen proposamena landuko da, *Wikidatan* integratzeko.

GAKO-HITZAK: Hiztegi historikoak · Larramendi · OCR · Ikasketa automatikoa.

The Digitization of Larramendi's *Diccionario Trilingüe*. Optical Character Recognition and Uploading to Wikisource

In this article, we describe the OCR process using machine learning, a part of artificial intelligence, in the digitization of Larramendi's *Diccionario Trilingüe*. For this purpose, we describe the pre-treatment of scanned images and the training of a tool named *Kraken*, in order to train a model from a hand-transcribed sample that will recognize the dictionary text. In addition to the highly accurate text, files containing italics and the position of the text have been created, enabling the representation of the structure of the dictionary. As the results are available on the Wikisource platform, the transcription can be corrected using crowdsourcing, so that we can carry out the information extraction process of the corrected transcription using machine learning to extract the lexicographic structure of the dictionary. With this information, the first proposal for a RDF model will be developed, in order to integrate the data in Wikidata.

KEY WORDS: Historical dictionaries · Larramendi · OCR · Machine learning.

<https://doi.org/10.26876/uztaro.120.2022.5>

Jasotze-data: 2022-12-23 *Onartze-data:* 2022-12-23

1. Sarrera eta motibazioa

Manuel Larramendiren 1745eko gaztelania-euskara-latina hiztegia, *Hiztegi Hirukoitza* (HH hemendik aurrera), mende eta erdi baino gehiagoko tartean zehar euskararako erreferentzia gailena izan da, eta euskal hiztegi gintzako periodizazioan lan lexikografiko klasikotzat har daiteke, aldaketa esanguratsua ekarri zuena (Urgell, 2002); euskal hiztegi gintzako garai modernoaren hasiera markatzen du. Hiztegia sakon aztertu da irizpide filologiko eta lexikografikoak aintzat hartuz (Urgell, 1998a; 1998b; 2000, beste batzuen artean), baina eskuz bildutako laginak baino ezin izan dira baliatu lan horietarako. Beraz, ezin izan da eduki guztia kontuan hartuko lukeen metodo kuantitatiborik erabili¹.

Hiztegi klasikoaren edizio digitalak egiteko bide berriak jorratu dira azken hamarkadetan. Humanitate Digitalen (HD) esparruan arigara, hau da, gizarte-zientziak eta konputazioa eta teknologia digitalak uztartzen dituen esparru akademikoan. Testu klasikoaren bertsio digitalak osatu eta prozesamendu konputazionalerako prestatzea da HDen ardatz nagusi bat (Lindemann eta San Vicente, 2020). Bestela esanda, testua gordetzen duen faksimile digitala eskaintzeaz gain, hiztegiaren makroegiturari (lemategiari) eta mikroegiturari (sarreraren barruko antolamenduari) dagozkien anotazioak gehitzen zaizkio, makinek testu zatiak item lexikografiko zehatz gisa interpretatu ahal izateko. Baliabidearen barruko bilaketa aurreratuak ahalbidetzeko eta baliabidetik kanporako lotura esplizituak esleitzeko baldintza teknikoak betetzen dira horrela. Lindemann eta San Vicentek (2020) laburbildutako digitalizazio-pausoei jarraituz, erreminta-kate erdiautomatikoa aplikatu dugu, eta haren emaitza neurtu. Horrek barne hartzen ditu: karaktereen ezagutze optikoa edo *Optical Character Recognition* (OCR), informazio-erazketa eta, azkenik, *Resource Description Framework* (RDF) estandarrarekin bat datorren moldaketa-eredu baten lehen proposamena, Wikidatan integratzeko aukerari begira.

Artikulu honetan karaktereen ezagutze optikoa egiteko irudiak nola tratatu ikusi eta, ikasketa automatikoa baliatuz, HHko testua ezagutuko duen ereduaren trebakuntza azaldu dut. Ondoren, emaitzak Wikiteka plataformara igo ditugu eta auzolanez transkripzioa zuzentzeko aukera ematen duenez, aurrera begira prozesu horretatik lortutako emaitzetatik informazio lexikografikoa duten datuak erazteko ikasketa automatikoa baliatuko dugu, horretarako eginda ditugun frogak kontuan hartuz.

1. Artikulu hau «Hiztegi historikoaren digitalizaziorako lan-fluxua: Larramendiren *Hiztegi Hirukoitza*» lanean egindako ikerketan oinarritzen da. Proiektu hori Udako Euskal Unibertsitateak (UEU) eta Euskal Wikipediako Kultur Elkarteak (EWKE) HDen inguruko euskarazko ikerketa sustatzeko eta euskarazko datu librean sorkuntza, trukea eta zabalpena bultzatzeko asmoz 2019. urtean egin zuten ikerketa-deialdian hautatua izan zen. Bestalde, lan hau UPV/EHUko Euskal Hizkuntzalaritza eta Filologia Masterrean David Lindemann eta Blanca Urgellek zuzendutako «Larramendiren *Hiztegi Hirukoitzaren* digitalizazioa» Master Amaierako Laneko atal batetik eratorria da. Gainera, lan honek «Giza Zientziak eta Artea» ataleko komunikaziorik onenaren saria irabazi zuen UEUK antolaturiko «IkerGazte2021. Nazioarteko Ikerketa Euskaraz» Kongresuan (2021eko ekainaren 9, 10 eta 11n, Vitoria-Gasteizko Europa Biltzar Jauregian).
UZTARO 120, 83-93

2. Arloko egoera eta ikerketaren helburuak

Adimen artifizialaren adarra den ikasketa automatikoa, ingelesezko *Machine Learning* delakoa, eremu askotan aplikatzen da gaur egun, kalkulu-ahalmenaz harago, esperientziatik ikasteko gai diren konputagailu-programak baitira. Hizkuntzalaritzan ere gero eta gehiago erabiltzen dira, eta azken aldian lan ugari ari dira argitaratzen ikasketa automatikoa baliatuz testuen eta hizkuntzen prozesamenduko eremu ezberdinetan emaitza baliagarriak lortu dituztenak. Romanov, Miller, Savant eta Kiessling-ek (2017), esaterako, ikasketa automatikoko *Kraken* tresna erabili dute idazkera arabiar klasikoko testuak digitalizatzean karaktereen ezagutze optikoa egiteko, eta beste tresna nagusiek baino emaitza hobekiak lortu dituzte.

Karaktereen ezagutze optikoa egiten duten programek, labur esanda, testu baten iruditik (inprimatutako liburu baten irudi eskaneatua, esaterako) testua erauzten dute, testuaren irudia testu digital bihurtuz. Horrela, editatzeko, bilatzeko eta konputazionalki aztertzeo aukera ematen dute. Ohiko OCR erremintek ez dute emaitza onik testu klasikoaren digitalizazioan, gaur egungo karaktere moderno eta uniformeak irakurtzeko trebatu baitira, eta akats gehiegi itzultzen dituzte duela mende batzuetako dokumentuak prozesatzean.

OCR erreminta-aukera zabala aurki daiteke gaur egun, hala nola Tesseract-ocr², ABBYY FineReader PDF³, OCRopus⁴, Transkribus⁵ edo Kraken⁶. OCR tresna berriek *Machine Learning* edo ikasketa automatikoko algoritmoak baliatzen dituzte, eskuz transkribatutako lanaren lagin bat erabiliz trebatzen direnak eta, horretan oinarrituta, loturak iragartzen dituzte pixel patrioiak diren letren eta karaktere digitalen artean. OCR prozesurako ikasketa automatikoa erabiltzea berrikuntza da eta Aro Modernoaren hasierako inprimatuak (eta baita eskuz idatzitako testuak ere⁷) prozesatzea ahalbidetu du, egun paperean edo digitalki inprimatutako karaktereak pixel patrioi uniformeekin lot daitezkeen bitartean, hizki beraren patrioiak aldakorrak baitira Aro Modernoaren hasierako inprimaketan. Gainera, papereko irregulartasunek, orbanek edo dagokion orriko atzeko aldeko tintak pixel patrioiak alda ditzakete. Tresna hauen artean Kraken aukeratu eta trebatu da.

Kraken doan eskura daitezkeen OCR tresna da, ikasketa automatikoa baliatzen duena. Pariseko PSL Unibertsitateko eScripta ikerketa-taldeak sortua da⁸, eta UNIX sistema eta Python ingurua behar ditu. Python kode irekiko lizentzia duen programazio-lengoaia bat da, askotariko plataformetan erabil daitezkeena eta sintaxi irakurgarria duena. Kraken ere kode irekikoa da eta *neural network* edo neurona-sare deituriko batean oinarritzean, gizakien ikasteko modua imitatzen daki. Horrela, HHko letra mota ezagutzeko trebatu da. Kraken tresnak, gainera, hizkiak identifikatzeaz gain, letra-tipoak bereizteko aukera ematen du. HHn letra-tipo etzana aurkitzen dugu eta sarreraren antolaketan garrantzia du. Beraz, letra-tipoa bereizi

2. Ikus <<https://github.com/tesseract-ocr>>.

3. Ikus <<https://pdf.abbyy.com>>.

4. Ikus <<https://github.com/ocropus/ocropy>>.

5. Ikus <<https://transkribus.eu/Transkribus/>>.

6. Ikus <<http://kraken.re>>.

7. Transkribus aplikazioak, adibidez, ikasketa automatikoa erabiltzen du eskuizkribuak prozesatzeko.

8. Ikus <<https://escripta.hypotheses.org/>>.

ahal izatea erabilgarria izango da hiztegia kodifikatzeko hurrengo pausoetan. Hori dela eta, Kraken hobetsi dugu antzeko ezaugarriak dituen web tresna Transkribus baino. Inprimaki eta eskuizkribuetarako diseinatutako jatedun OCR tresna nagusiek baino emaitza hobekien lortzen dituela erakutsi du; esaterako, idazkera arabiar klasikoko testuak digitalizatzean (Romanov *et al.*, 2017). Gainera, Kraken ALTO XML estandarrari jarraitzen dioten fitxategiak sortzeko gai da, eta hori ere tresna hau aukeratzeko arrazoia izan da.

Karaktereen ezagutze optikoa egiteko garaian irudien kalitatea kontuan hartzeko ezaugarri bat da. Irudietako testuak ahal bezain argiena izan behar du eta orokorrean egoera onean dagoen edizio baten irudiak erabiltzea komeni da. Beraz, faktore horiek ere kontuan hartu dira digitalizatuko den edizioa aukeratzeko, baita irudi digitalek duten kalitatea ere. Hau dpi edo dot per inch, hazbetean sartzen den puntu kopurua, erabiliz zehazten da, eta Krakeneko jarraibideetan gutxienez 300 dpi-ko irudiak behar direla zehazten den arren⁹, Romanov *et al.* (2017: Table 1) lanak irudien kalitateak espero baino garrantzi gutxiago zuela iradokitzen du.

3. Ikerketaren muina

Kraken erremintaren trebatze-prozesua, laburbilduz, honako pausoek osatzen dute: (1) irudien aurreprozesamendua programak behar dituen ezaugarrietara doitzeko; (2) erreminta trebatzeko erabiliko diren testuaren lerroen transkripzioa; (3) ereduaren trebakuntza, eta (4) emaitzak jasotzen dituzten fitxategiak sortzea.

3.1. Irudien aurreprozesamendua

Kalitateaz gain, digitalizatu nahi diren irudiek zenbait ezaugarri izan behar dituzte Krakenek prozesatu ahal izateko. Esaterako, irudiak zuri-beltzean irakurtzen ditu eta Krakenek berak zuri-beltzean jartzeko aukera eskaintzen duen arren, aurretik irudiak prozesatu dira beste ezaugarri batzuk ere doitu ahal izateko. Eskaneatutako irudi askori angelua zuzendu behar zaie, lerroak ez baitaude horizontalean, edota liburuetakako orrialdeak eskaneatzean alboetan sortzen den kurbadura zuzendu behar da. Orbanak kentzea ere komenigarria da, zuri-beltzean jarritakoan hizkiekin nahas daitezkeelako. Orrialdearen diseinua konplexua bada, egunkarietakoa bezalakoa adibidez, zutabeak bakantzea gomendagarria da, lerroak zatitzen dituen algoritmoa ez baita ondo moldatzen diseinu konplexu horietara. Irudiak aurreprozesatzeko Scan Tailor¹⁰ programa erabili da, Kraken tresnaren egileek dokumentazioan gomendatzen dutenaren arabera¹¹.

HHren irudi eskaneatuak prozesatu dira, diseinuaren eta orrialdeen egoera dela-eta zenbait egokitzapen beharrezkoak baitziren. Lehenik eta behin, orrialde bakoitzean bi zutabe zeudenez, zutabeak orrialde banatan jarri dira, aipatu bezala Krakenek diseinu konplexuak zatitzeko algoritmo berezirik ez dakarrelako. Ondoren, testuaren angelua zuzendu da eta testua daukan eremua zehaztu da orrialdeetan. Eremu horri 5 mm-ko marjinak gehitu zaizkio alde bakoitzean. Azken urratsean

9. Ikus <<http://kraken.re/training.html#image-acquisition-and-preprocessing>>.

10. Ikus <<http://scantailor.org/>>. ScanTailor aplikazioa software librekoa da.

11. Ikus <<http://kraken.re/training.html>>.

irudiaren kalitatea aukeratu (600 dpi da lehenetsia), zuri-beltzean jarri eta letren lodiera aldatu da (parametroa +10 jarri da zenbait hizki edo hizkiren zatiak bestela «desagertu» egiten baitira). Hala ere, froga ezberdinak eginez letren lodieran balio negatiboak aukeratuz atzeko orrialdeko letrak gutxiago nabaritzen direla ikusi da. Urrats berean orbanak kentzeko aukera ematen du programak eta honetan bigarren indartsuena aukeratu da. Ordenagailu pribatu arruntak erabilia, urrats guztietatik pasatzeko orrialdeko batez beste 1,2 minutu behar direla neurtu dugu, baina ahalmen handiagoko zerbitzari edo ordenagailuekin denbora gutxiago behar izatea espero daiteke.

3.2. Transkripzioa

Irudiak aurreprozesatu ondoren, Krakenen diseinua ezagutzeko modulua abiatzen da, irudietatik abiatuz transkripzioa egiteko beharrezkoak diren fitxategiak sortzeko. Irudiak lerrotan banatzen ditu, eta ondoan, edita daitezkeen testu-eremuak jartzen ditu (ikus 1. irudia). HTML fitxategiak dira, eta edozein web nabigatzaile erabiliz edita daitezke testuari dagozkion eremuak. Lerro hauek dira programak erabiltzen dituen oinarrizko unitateak; lerroen irudiak eta testu-lerroak parekatzen dira eta hauek izango dira trebakuntza sorta osatuko dutenak, hots, programak aurreikuspenak balioztatzeko erabiliko dituen lerro eredarriak. Hortaz, transkripzioak diplomatikoa izan behar du beti, hau da, lerroaren irudiko karaktereen sekuentziak zehatza izan behar du, jatorrizko grafia erabat mantenduz, ezer gehitu edo baztertu gabe, *Ground Truth Transcription* delakoaren jarraibideak segituz¹².

V E. 371
 Lat. Vertibilitas.
 Vertible , *aldacoya* ; *giracoya*. Lat. Vertibilis.
 Vertical , *bugaindarra*. Lat. Verticalis.
 Vertice , *bugaina*. Lat. Vertex , icis.
 Verticidad , veafe *vertibilidad*.
 Vertigo , vertiginoso , veafe *vaguido*.
 Vefpero , *illunabarreco izarra*. Lat. Vefperus.
 Velpertino , *arratsaldeco* , *arraleguicoa*. Lat. Velpertinus.
 Vefquir , antiquado , lo mismo que *vivir*.
 Vefte , lo mismo que *vestido*.
 Vestido , *foñecoa* , *jazcaya* , *jauncaya* , *aldagarria* , *filda* , *abillamendua*. Lat. Vestis , vestitus.
 Vestidura , lo mismo. Lat. Indumentum.
 Vestigio , *aztarnà* , *fenà* , *batzá*. Lat. Vestigium.
 Vestiglo , monstruo formidable , *mamusa*. Lat. Spectrum horridum.
 Vestimenta , vestimento , veafe *vestido*.
 Vestir , *janci* , *jaunci* . *bestitu*. Lat. Vestire , induere.

VE
Lat Vertibilitas.
371
Vertible ,@aldacoya ,@giracoya Lat Vertibilis
Vertical ,@bugaindarra Lat Verticalis
Vertice ,@bugaina Lat Vertex ,icis
Verticidad ,veafe @vertibilidad
Vertigo ,vertiginoso ,veafe @vaguido
Vefpero ,@illunabarreco @izarra Lat Vefperus
Velpertino ,@arratsaldeco ,@arraleguicoa Lat Velpertinus
Vefquir ,antiquado ,lo mismo que vivir
Vefte ,lo mismo que @vestido
Vestido ,@foñecoa ,@jazcaya ,@jauncaya ,@aldagarria ,@filda ,@abillamendua Lat Vestis ,vestitus
Vestidura ,lo mismo Lat Indumentum
Vestigio ,@aztarnà ,@fenà ,@batzá Lat Vestigium

1. irudia. Transkripzioa egiteko fitxategia.

12. Ikus <<https://ocr-d.de/en/gt-guidelines/>>.

Krakenek lerro hauetan transkribatutakoa baino ez du ezagutuko, hots, trebatze sortatik frogak jaso aurretik, tresna erabat agnostikoa da. Beraz, argi dago lerro hauetan testuan zehar azaltzen diren karaktere ezberdin guztiak bildu behar direla. Jarraibideetan azaltzen denez, arabierazko edo hebreerazko idazkerak gutxienez 800 lerroren transkripzioa eskatzen du, eta Mendebaldeko idazkeretarako ere antzeko kopurua iradokitzen da, orrialdeko 25 eta 40 lerro arteko kopuruarekin gutxienez 30 orrialde transkribatu behar direla aipatzen baita. Hirurogeina lerroko bi zutabeko orrialde bakarra hutsetik transkribatu ondoren, Krakeni eredu bat sorrarazi zaio, eta, lehen OCR iterazioa egin eta gero, HTML fitxategi berriak sortzen dira aukeratutako hiztegiko orrialde sortarako. Testu-eremuek oraingo honetan Krakenek lehen ereduan oinarrituta ezagututako testua edukiko dute (lehen trebatze sortatik sortua). Hortik aurrera, eremuetako testua ez da hutsetik idazten, baizik eta eskuz zuzentzen. Zuzendutako orrialde osoak gehi dakizkioke trebatze sortari, hurrengo eredu berritua sortzeko erabiliko direlarik ondorengo iterazioan, eta prozedura nahi den doitasun-maila lortu arte errepika daiteke.

Gaur egun ez bezala, hizkien ondoren eta koma eta puntuaren aurretik nahiko hutsune handia dago eta transkribatzeko garaian hutsune horiek errespetatu dira. Gaztelaniaz, *una* bezalakoak batzuetan *vna* bezala azaltzen dira, eta horrelaxe transkribatu dira. Gaur egun erabiltzen ez den ese altu arkaikoa ere, «*ʃ*», mantendu egin da¹³.

HHk bi letra mota ditu, borobila eta etzana. Lehenago aipatu bezala, transkripzioan letra etzana gordetzeak garrantzi handia du, letra etzana hiztegiko sarreraren egituran aurki daitekeen ezaugarri tipografiko bakarra baita. Krakeneko transkripzioak testu hutsa dira, hau da, ez dute letra-tipoak zehazteko berariazko aukerarik, ezagutu ditzakeen arren. Horregatik, letra etzanez dagoen hitz oro markatuz gero, algoritmoak ebatz dezake zein hitzek duten letra etzana. Hasieran letra etzanean dauden hitzak HTML lengoaiako <i> etiketarekin markatu ziren, baina letra etzana horrela markatzeak transkripzioa dezente zailtzen du, hitz edo esaldi etzan bakoitza markatzeko zazpi karaktere behar baitira. Markaketa sinpleago bat erabiltzearen, letra etzanean dauden hitzei baliabide guztian zehar azaltzen ez den ikur bat jarri zaie aurretik, @ ikurra. Transkribatzea errazagoa da eta eredia trebatzerakoan doitasun handiagoa ematen duela egiaztatu ahal izan dugu, letra etzanez idatzita dauden hitzek aurretik ikur hori daramatela ikasiko baitu Krakenek. Frogatu dugun bezala, ia akatsik gabe ebatzi du hori.

Bestalde, gerta daiteke orrialdea lerrotan banatzen duen algoritmoak zuzen ez jokatzeko eta bi lerro bakarra balitz bezala interpretatzea (ikus 2. irudia) edota lerroak bitan banatzea.

13. Unicode karakterea erabili dugu: U+017F LATIN SMALL LETTER LONG S «*ʃ*».

En atencion à eſto , *onelacoren beguiru-
nez.*

Atender , *arretatu , oartu , arreta eman.*
Lat. Attendere.

@beguirineac.
En atencion à eſto , @onelacoren @beguiru-
@nez.
NOTA.

NOTA.

2. irudia. Krakenek bi lerro bakarria balitz bezala interpretatzea.

Akats hauei aurre egiteko modurik eraginkorrena lerro horiei dagozkien transkripzio-eremuak hutsik uztea da, programak kontuan har ez ditzan.

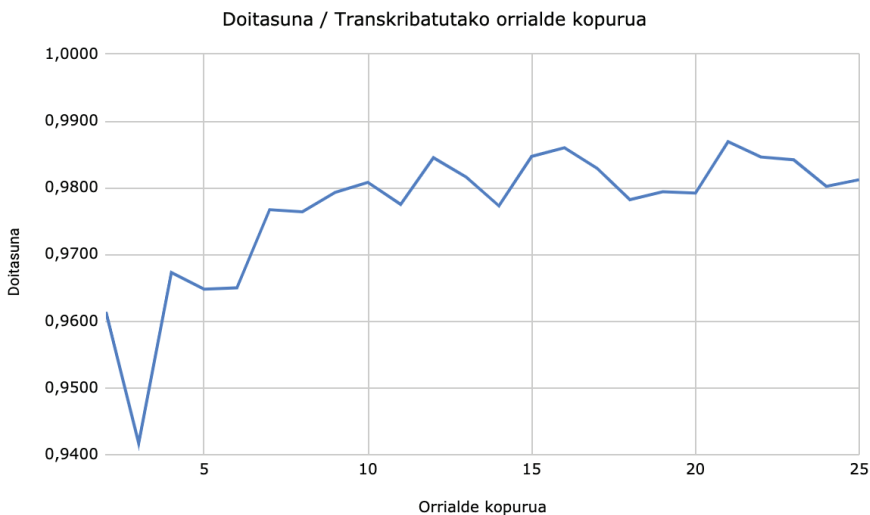
3.3. Ereduaren trebakuntza

Eredua trebatzeko adimen artifizialaren adarra den ikasketa automatikoan oinarritzen da Kraken. Lerro bat «irakurtzen» du eta ondoren akatsak konpentsatzen ditu aipaturiko neurona-sareaz baliatuz. Ondoren, lerroz lerro prozesua errepikatzen du azken lerroa iritsi arte, eta orduan berriz ere hasten da lehen lerrotik, trebakuntza-lerroak behin eta berriz irakurriz, gero eta eredu hobea lortuz. Horretarako lerro kopuru bat bereizten du, trebakuntza sorta (*training set*), eta beste alde batetik ebaluatzeko sorta (*test set*) erabiltzen du trebatutako ereduaren doitasuna kalkulatzeko (balio lehenetsia % 10 da ebaluatzeko sortarako). Eredua eraiki ondoren, ebaluatzeko sortarekin frogatzen du, eta ezagutzen dituen karaktereen eta egiten dituen akatsen arteko erlazioa izango da ereduaren doitasuna. Trebakuntza-prozesua nolabait ausazkoa denez, ezin da aurreikusi zenbat denbora beharko den eredu bat trebatzeko, eta konfigurazio lehenetsiak *early stopping* edo gelditze goiztiarra izeneko hurbilketa egiten du, hau da, ebaluatzeko sortan ez bada akats-indizea hobetzen, prozesua gelditu egiten da. Horrek *overfitting* edo gaindoitzea ekiditen du, hots, eredu trebatze-datuak bakarrik ezagutzeko doitzea, datu horietan aurkituko lituzkeen patrioiak ezagutu beharrean. Horrela lortzen da akats gutxien egiten duen eredu: doitasun handiena duen eredu, hain zuzen ere.

Transkripzioari eskainitako atalean aipatu bezala, letra etzanak bereizteko <i> markaketa erabiliz, zazpi karaktere gehitzen dira hitz edo esaldi bakoitzeko, lerroko karaktere kopurua handituz eta, oro har, ereduak karaktereak ezagutzeko zailtasuna gehituz. Horrela transkribatutako orrialde batekin trebatutako ereduak % 83,99ko doitasuna lortu zuen, eta bigarren orrialde bateko transkripzioa gehituta % 92,56koa. Markaketa sinpleago batekin, letra etzana markatzeko @ erabiliz, ordea, ereduaren doitasuna nabarmenki hobetu zen, bi orrialde erabilita % 96,14ko doitasuna lortu baitzuen hain lerro gutxirekin lanean arituta. Hori horrela, letra etzanak @rekin markatzea erabaki zen eta 1. taulan eta 3. irudian jasotako doitasun-balioak lortu ziren.

1. taula. Orrialde kopuruak eta doitasuna.

Or. kop.	2	3	4	5	6	7	8	9	10	11	12	13
Doitasuna	0,9614	0,9417	0,9673	0,9648	0,965	0,9767	0,9764	0,9793	0,9808	0,9775	0,9845	0,9816
Or. kop.	14	15	16	17	18	19	20	21	22	23	24	25
Doitasuna	0,9773	0,9847	0,9860	0,9829	0,9782	0,9794	0,9792	0,9869	0,9846	0,9839	0,9802	0,9812



3. irudia. Doitasuna eta transkribatutako orrialde kopurua.

Doitasunaren grafikoa ikus daitekeen bezala, orrialde kopurua handitu ahala doitasuna handitzen doa, baina ez dago korrelazio zuzenik eta gero eta orrialde gehiago behar dira akats-indizea jaisteko. Hori horrela, ondoriozta daiteke transkripzio-lerro kopuru batetik gora ez duela merezi eredu berriak trebatzea, doitasuna ez baita esanguratsuki hobetuko. 60na lerro inguruko 25 orrialde transkribatu ondoren doitasuna ez zela esanguratsuki handituko ikusi genuen. Hala ere, doitasun-indize aipagarriak lortu arren, zenbait karaktere (ezohiko) ez zituela ezagutzen ohartu ginen. Hori dela eta, trebatze sortan falta ziren karaktere horiek zituzten orrialdeen transkripzioak gehitu ziren, batez ere zegokien alfabeto atalean ugari agertuko ziren hizki larriak zituztenak. Horrela, 32 orrialderekin trebatutako ereduak lortu genuen, eta doitasun orokorra ez zen esanguratsuki handitu (% 98,5), baina ezohiko karaktereak, hasieran Krakenentzat «ezezagunak», orduan ondo ezagutu zituen.

3.4. Emaizten fitxategiak sortzea eta Wikitekara igotzea

Emaizak fitxategi-formatu ezberdinetan ematen ditu Krakenek, hala nola txt edo ALTO XML¹⁴, OCR emaitzak gordetzeko estandar bat. AEBko Kongresuaren Liburutegiak, esaterako, erabiltzen du ALTO XML eta informazioa erazteko Elexifier¹⁵ erreminta-kateak input formatu bezala onartzen du. Testu-fitxategi soilak (txt formatukoak) testu-kutxa bakoitzean ezagututako karaktereak lerroka gordetzen ditu; ALTO XML formatuak, berriz, testua ez ezik, karaktereen posizioa ere gordetzen du, orrialdearen diseinua berreraikitzeak aukera eskainiz. Testuko koskak irudikatzeak gai denez, HHren sarreren egitura irudika dezake, sarrera-buruak koska

14. Ikus <<https://altotml.github.io/>>.

15. Ikus <<https://elexifier.elex.is/>>.

ezberdinarekin agertzen baitira hurrengo lerroen aldean. Testu-kutxen kokapena, beraz, kontuan hartzea merezi duen informazioa da, irudiko eta transkripzioko testu zatien artean esteka edo lotura aktiboak eskaintzen dituen edizio digitala prestatzeko garaian berebizikoa izan daitekeena.

Wikisource plataformako euskal atariak, Wikitekak¹⁶, OCR emaitzak erakusteko eta editatzeko aukera ematen du eta bertan Wikitext formatura¹⁷ egokitu dugun ALTO fitxategien OCR emaitza¹⁸ jarri dugu eskuragarri (ikus 4. irudia). Wikitekako edukia irekiak direnez, transkripzioak zuzentzeko erabil daiteke. Are gehiago, badago zehazki transkripzio-proiektuei eskainitako atal bat¹⁹, jarduera honek gune horretan duen garrantziaren seinale. Beraz, komunitateak HHren transkripzioak zuzentzeko aukera izango du, beste herrialdeetako proiektu askotan egin den bezala.

The screenshot shows the Wikiteka website interface. At the top, there is a navigation bar with links like 'Orrialdea', 'Eztabaida', 'Irudi', 'Irakurri', 'Aldatu iturburu kodea', and a search bar. The main heading is 'Orrialde:Larramendi 1745 dictionary body.pdf/1'. Below this, there is a list of categories on the left, including 'Azala', 'Txokoa', 'Aldaketa berriak', etc. The main content area features a large decorative initial 'A' and a list of entries from the dictionary, including 'Artículo de el dativo en Bafcuence es otro polpofitivo...' and 'A. En los propios, que acaban en fonante...'.

4. irudia. HHren 1. orrialdea Wikiteka plataforman²⁰.

16. Euskarazko ataria:< <https://eu.wikisource.org/wiki/Azala>>.

17. Ikus <<https://en.wikipedia.org/wiki/Help:Wikitext>>.

18. Ikus <https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza>.

19. Ikus <https://en.wikisource.org/wiki/Wikisource:Transcription_Projects>.

20. Ikus <https://eu.wikisource.org/wiki/Orrialde%3ALarramendi_1745_dictionary_body.pdf/1>.

4. Ondorioak

HHren A-Z atalak, lehen eta bigarren liburukiak kontuan hartuta, 828 orrialde ditu; beraz, edukiaren % 4 baino gutxiago eskuz transkribatuta (32 orrialde), hiztegi guztia hartzen duen txt bertsioan ia % 98,5eko zehaztasuna lortu da. Doitasunak lehenago eskuragarri zeuden txt bertsioak gailentzen ditu argi eta garbi.

Diseinua ezagutzeko moduluak zehaztasun handiz egin du lan, baina, hala ere, kontuan hartu beharreko lerro kopuru esanguratsu bat ez du ezagutu. Zenbait kasutan, edo ez du lerroren bat inolaz ere ezagutu, edo lerroak gaizki batu ditu, testu-kutxaren eremuak bi lerro hartzen dituelarik bat hartu beharrean, eta hori eskuz zuzentzeko modu zuzenik ez dago. Trebakuntza sorta egitean arazorik ematen ez badu ere, hiztegi guztia digitalizatzean horrek izango lukeen eragina neurtu beharko litzateke. Oraingo ezin izan dugu neurtu gaizki ezagututako lerroen eragina, baina transkripzio-prozesuan zehar eta ondoren OCR emaitzak berrikusteak iradokitzen du merezi duela puntu hau sakonago aztertzea, diseinuaren ezagutzaren balidazio-pausoa lan-fluxuan sartzeko asmoz. Bestalde, horrelakoak konpontzeko, baliteke irudien aurreprozesamenduko parametroak aldatu eta prozesua errepikatzeak doitasun handiagoa ekartzea, lerroak irudian identifikatzeko orduan. Nolanahi ere, azken baliabideak, hots, HHren bertsio digitalizatuak, argitalpen baten kalitate-irizpideei jarraitu behar badie, OCR prozesuan zehaztasun handia lortu arren, badirudi beharrezkoa izango dela eskuz eduki guztia balioztatzea, diseinuaren ezagutze-prozesuko hutsegiteek sortutako akatsen zuzenketa barne.

5. Etorrizunarako planteatzen den norabidea

Wikitekan jarritako testuen *ground truth* irizpideen araberako transkripzioa lortzea izango da behin-behineko helburua, digitalizazioaren lan-fluxuko hurrengo urratsak doitasun erabateko datuekin errepikatu ahal izateko. Hurrengo urratsean, informazio-erazketa prozesuan, transkripzio horretatik informazioa erauzi eta hiztegiaren sarrerak eta sarreren barruko item lexikografikoak irudikatuko dituen egitura lortzea da helburua. Horretarako hiztegiko orrialdeen diseinua baliatuko da hiztegiko item lexikografikoak sailkatu eta identifikatzeko, hau da, testu digitaletik informazio lexikala erauziko da. Urrats hori honela deskriba daiteke: hiztegiaren sarrerak eta sarreren barruko antolamendua zatitu, eta item lexikografikoak irudikatzen dituen markaketa semantikoaz osatzen dira. Horretarako, lagin bat eskuz anotatu eta gero, ikasketa automatikoko algoritmo batek zatiketa eta anotazioa ikasi eta testu osoan aplikatzen ditu.

Dagoeneko beta fasean dagoen Elexifier tresna erabili dugu informazioa erauzteko. Lortutako emaitzak oinarri hartuta eta HHren balioztatutako transkripzioarekin prozesua errepikatu ahal izango dugu; batetik, Elexifier trebatze sorta handiago batez elikatuta eta, bestetik, markaketa konplexuagoa erabilita, oraingo honetan beta fasean dagoen tresna horren garapenak aurrera egiten duen heinean.

6. Erreferentziak

- Alonso, Mikel (2021): *Larramendiren Hiztegi Hirukoitzaren digitalizazioa* (Master Amaierako Lana). DOI: 10.13140/RG.2.2.27926.68169
- Larramendi, Manuel (1745): *Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa*, Bartholomé Riesgo y Montero, Donostia.
- Lindemann, David eta San Vicente, Iñaki (2020): «Baliabide lexikoen sarea: Baldintza filologiko eta tekniko zenbait», in Itziar Laka (koor.), *Hitzak sarean: Pello Salabururi esker onez*, UPV/EHU Argitalpen Zerbitzua, Bilbo, 79-96.
- Lindemann, David eta Alonso, Mikel (prestatzen): «Hiztegi historikoen digitalizaziorako lan-fluxua: Larramendiren 'Hiztegi Hirukoitza'».
- Romanov, Maxim G.; Miller, Matthew Thomas; Savant, Sarah Bowen eta Kiessling, Benjamin (2017): «Important New Developments in Arabographic Optical Character Recognition (OCR)». DOI: 10.7916/alusur.v25i1.6996
- Urgell, Blanca (1998a): «'Hiztegi Hirukoitza' eta 'Diccionario de Autoridades' erkatuaz (I): Oinarrizko ezaugarri zenbait», *Anuario del Seminario de Filología Vasca 'Julio de Urquijo'*, 32(1), 109-163.
- _____, (1998b): «'Hiztegi Hirukoitza' eta 'Diccionario de Autoridades' erkatuaz (II): Sarreraren edukia», *Anuario del Seminario de Filología Vasca 'Julio de Urquijo'*, 32(2), 365-414.
- _____, (2000): *Larramendiren Hiztegi Hirukoitza-ren osagaiez* (Doktorego-tesia), EHU, Gasteiz.
- _____, (2002): «'Hiztegi Hirukoitza'-ren kanpoko eta barruko historiaz», *Anuario del Seminario de Filología Vasca 'Julio de Urquijo'*, 629-649.

