IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

IV. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2021eko ekainaren 9, 10 eta 11a
Gasteiz, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK ETA ARTEA

Larramendiren *Hiztegi Hirukoitza*ren digitalizazioa. Karaktereen ezagutze optikoa eta *Wikitek*ara igotzea

Mikel Alonso eta *David Lindemann*

119-126 or.
https://dx.doi.org/10.26876/ikergazte.iv.01.15

ANTOLATZAILEA: LAGUNTZAILEAK:
ueu udako euskal unibertsitatea
eman ta zabal zazu UPV EHU
upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Universidad de Navarra
Deusto Universidad de Deusto Deustuko Unibertsitatea
Mondragon Unibertsitatea
Ayuntamiento de Vitoria-Gasteiz Vitoria-Gasteizko Udala
araba álava foru aldundia diputación foral
EUSKO JAURLARITZA GOBIERNO VASCO HEZKUNTZA SAILA DEPARTAMENTO DE EDUCACIÓN
UDALBILTZA

Larramendiren *Hiztegi Hirukoitza*ren digitalizazioa. Karaktereen ezagutze optikoa eta *Wikiteka*ra igotzea

Alonso, M., Lindemann, D.
UPV/EHU
mikelalon@gmail.com, david.lindemann@ehu.eus

Laburpena

Artikulu honetan Larramendiren *Hiztegi Hirukoitza*ren digitalizazioko OCR prozesua deskribatzen da, ikasketa automatikoa baliatuz. Horretarako, irudien tratamendua deskribatu eta eskuz transkribatutako laginetik abiatuta testua ezagutuko duen ereduaren trebakuntza azaltzen da. Emaitzak Wikiteka plataforman eskuragarri daudenez, auzolanaz transkripzio osoa zuzendutakoan informazio erauzketa prozesutik pasako da hiztegiaren egitura lexikografikoa ikasketa automatikoz erauzteko.

Hitz gakoak: hiztegi historikoak, Larramendi, OCR, ikasketa automatikoa, Wikiteka

Abstract

In this article, we describe the OCR process using machine learning in the digitization of Larramendi's Diccionario Trilingüe. *For this purpose, the treatment of images is described and the training of the model from the transcribed sample that will recognize the text. 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 using machine learning to extract the lexicographic structure of the dictionary.*

Keywords: historical dictionaries, Larramendi, OCR, machine learning, Wikisource

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 hiztegigintzako periodizazioan lan lexikografiko klasikotzat har daiteke, aldaketa esanguratsua ekarri zuena (Urgell 2002); euskal hiztegigintzako 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 erabili.

Hiztegi klasikoen edizio digitalak burutzeko bide berriak jorratu dira azken hamarkadetan. Humanitate Digitalen (HD) esparruan ari gara, hau da, gizarte zientziak eta konputazioa eta teknologia digitalak uztartzen dituen esparru akademikoan. Testu klasikoen 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 (sarreren barruko antolamenduari) dagozkion 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 & San Vicentek (2020) laburbildutako digitalizazio pausoak jarraituz, erreminta-kate erdiautomatikoa aplikatu dugu, eta haren emaitza neurtu. Honek barne hartzen ditu: karaktereen ezagutze optikoa edo *Optical Character Recognition* (OCR), informazio erauzketa eta, azkenik, *Resource Description Framework* (RDF) estandarrarekin bat datorren moldaketa-eredu baten lehen proposamena, *Wikidata*n integratzeko aukerari begira.

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

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 ohiko tresna nagusiek baino emaitza hobeak lortu dituzte.

Karaktereen ezagutze optikoa burutzen 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 aztertzeko aukera ematen dute. Ohiko OCR erremintek ez dute emaitza onik testu klasikoen 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-ocr[1], ABBYY FineReader PDF[2], OCRopus[3], Transkribus[4] edo Kraken[5]. OCR tresna berrienek *Machine Learning* edo ikasketa automatikoko algoritmoak baliatzen dituzte, eskuz transkribatutako lanaren lagin bat erabiliz trebatzen direnak eta, horretan oinarrituta, loturak iragartzen dituzte pixel patroiak diren letren eta karaktere digitalen artean. OCR prozesurako ikasketa automatikoa erabiltzea berrikuntza da eta Aro Moderno hasierako inprimatuak (eta baita eskuz idatzitako testuak ere[6]) prozesatzea ahalbidetu du, egun paperean edo digitalki inprimatutako karaktereak pixel patroi uniformeekin lotu daitezkeen bitartean, hizki beraren patroiak aldakorrak baitira Aro Moderno hasierako inprimaketan. Gainera, papereko irregularitasunek, orbanek edo dagokion orriko atzeko aldeko tintak pixel patroiak alda ditzakete. Tresna hauen artean Kraken aukeratu eta trebatu da.

Kraken doan eskuratu daitekeen OCR tresna da, ikasketa automatikoa baliatzen duena. Pariseko PSL Unibertsitateko "eScripta" ikerketa taldeak sortua da[7], eta UNIX sistema eta Python ingurua behar ditu. Python kode irekiko lizentzia duen programazio lengoaia bat da, askotariko plataformetan erabil daitekeena 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 sarreren antolaketan garrantzia du. Beraz, letra-tipoa bereizi 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 jabedun OCR tresna nagusiek baino emaitza hobeak lortzen dituela erakutsi du; esaterako, idazkera arabiar klasikoko testuak digitalizatzean (Romanov et al. 2017). Gainera, Kraken ALTO XML estandarra jarraitzen duten fitxategiak sortzeko gai da eta hori ere tresna hau aukeratzeko arrazoia izan da.

Karaktereen ezagutze optikoa burutzeko 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 aukeratzean, eta 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

1 Ikus https://github.com/tesseract-ocr
2 Ikus https://pdf.abbyy.com
3 Ikus https://github.com/ocropus/ocropy
4 Ikus https://transkribus.eu/Transkribus/
5 Ikus http://kraken.re
6 Transkribus aplikazioak, adibidez, ikasketa automatikoa erabiltzen du eskuizkribuak prozesatzeko.
7 Ikus https://escripta.hypotheses.org/

behar direla zehazten den arren[8], Romanov et al. (2017: Table 1) lanak irudien kalitateak espero baino garrantzi gutxiago zuela iradokitzen du.

3. Ikerketaren muina

Kraken erremintaren trebatze prozesua, laburbilduz, ondorengo 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 liburuetako 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 Tailor[9] programa erabili da, Kraken tresnaren egileek dokumentazioan gomendatzen dutenaren arabera[10].

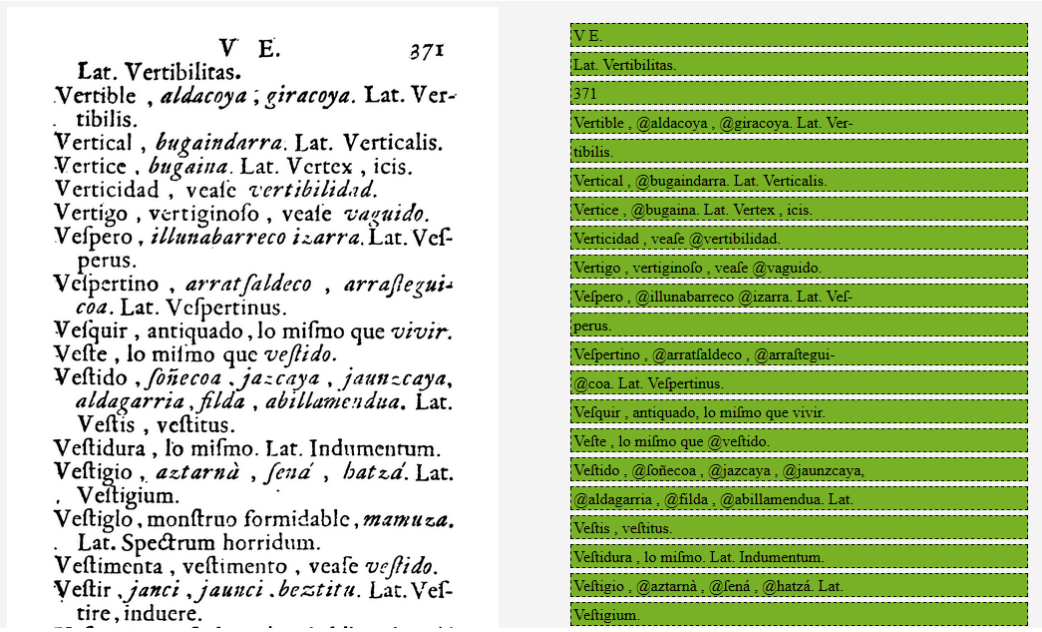
HHren irudi eskaneatuak prozesatu dira, diseinuaren eta orrialdeen egoera dela eta zenbait egokitzapen beharrezkoak baitziren. Lehenik eta behin, orrialde bakoitzean bi zutabe zeudenez, zutabeak orrialde banatan jarri dira, aipatu bezala Krakenek diseinu konplexuak zatitzeko algoritmo berezirik ez dakarrelako. Ondoren, testuaren angelua zuzendu da eta testua daukan eremua zehaztu da orrialdeetan. Eremu horri 5 mm-tako marjinak gehitu zaizkio alde bakoitzean. Azken urratsean 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 erabilita, 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 burutzeko beharrezkoak diren fitxategiak sortzeko. Irudiak lerrotan banatzen ditu eta, ondoan, editatu daitezkeen testu eremuak jartzen ditu (ikus 1. irudia). HTML fitxategiak dira eta edozein web nabigatzaile erabiliz editatu 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 eredugarriak. Hortaz, transkripzioak diplomatikoa izan behar du beti, hau da, lerroaren irudiko karaktereen sekuentzia zehatza izan behar du, jatorrizko grafia erabat mantenduz, ezer gehitu edo baztertu gabe, *Ground Truth Transcription* delakoaren jarraibideak segituz[11].

8 Ikus http://kraken.re/training.html#image-acquisition-and-preprocessing
9 Ikus http://scantailor.org/. ScanTailor aplikazioa software librea da.
10 Ikus http://kraken.re/training.html
11 Ikus https://ocr-d.de/en/gt-guidelines/

1. irudia: Transkripzioa egiteko fitxategia



V E. 371
Lat. Vertibilitas.
Vertible , *aldacoya* ; *giracoya*. Lat. Vertibilis.
Vertical , *bugaindarra*. Lat. Verticalis.
Vertice , *bugaina*. Lat. Vertex , icis.
Verticidad , veaſe *vertibilidad*.
Vertigo , vertiginoſo , veaſe *vaguido*.
Veſpero , *illunabarreco izarra*. Lat. Veſperus.
Veſpertino , *arratſaldeco* , *arraſteguicoa*. Lat. Veſpertinus.
Veſquir , antiquado , lo miſmo que *vivir*.
Veſte , lo miſmo que *veſtido*.
Veſtido , *ſoñecoa* , *jazcaya* , *jaunzcaya*, *aldagarria* , *filda* , *abillamendua*. Lat. Veſtis , veſtitus.
Veſtidura , lo miſmo. Lat. Indumentum.
Veſtigio , *aztarnà* , *ſeñá* , *hatzá*. Lat. Veſtigium.
Veſtiglo , monſtruo formidable , *mamuza*. Lat. Spectrum horridum.
Veſtimenta , veſtimento , veaſe *veſtido*.
Veſtir , *janci* , *jaunci* , *beztitu*. Lat. Veſtire , induere.

```
V E.
Lat. Vertibilitas.
371
Vertible , @aldacoya , @giracoya. Lat. Ver-
tibilis.
Vertical , @bugaindarra. Lat. Verticalis.
Vertice , @bugaina. Lat. Vertex , icis.
Verticidad , veaſe @vertibilidad.
Vertigo , vertiginoſo , veaſe @vaguido.
Veſpero , @illunabarreco @izarra. Lat. Veſ-
perus.
Veſpertino , @arratſaldeco , @arraſtegui-
@coa. Lat. Veſpertinus.
Veſquir , antiquado, lo miſmo que vivir.
Veſte , lo miſmo que @veſtido.
Veſtido , @ſoñecoa , @jazcaya , @jaunzcaya,
@aldagarria , @filda , @abillamendua. Lat.
Veſtis , veſtitus.
Veſtidura , lo miſmo. Lat. Indumentum.
Veſtigio , @aztarnà , @ſeñá , @hatzá. Lat.
Veſtigium.
```

Krakenek lerro hauetan transkribatutakoa baino ez du ezagutuko, hau da, 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 lerrotako bi zutabeko orrialde bakarra hutsetik transkribatu ondoren, Krakeni eredu bat sortarazi zaio eta, lehen OCR iterazioa 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.

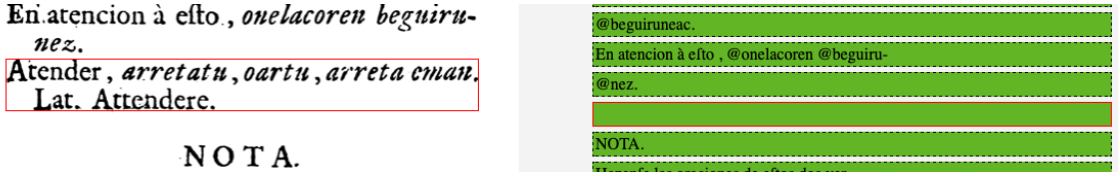
Gaur egun ez bezala, hizkien ondoren eta koma eta puntuaren aurretik nahiko hutsune handia dago eta transkribatzeko garaian hutsune horiek errespetatu dira. Gaztelaniaz, una bezalakoak batzuetan *vna* bezala azaltzen dira, eta horrelaxe transkribatu dira. Gaur egun erabiltzen ez den ese altu arkaikoa ere, “ſ”, mantendu egin da[12].

HHk bi letra mota ditu, borobila eta etzana. Lehenago aipatu bezala, transkripzioan letra etzana gordetzeak garrantzi handia du, letra etzana hiztegiko sarreren egituran aurkitu daitekeen ezaugarri tipografiko bakarra baita. Krakeneko transkripzioak testu hutsa dira, hau da, ez dute letra-tipoak zehazteko berariazko aukerarik, ezagutu ditzakeen arren. Horregatik, letra etzanaz dagoen hitz oro markatuz gero, algoritmoak ebatzi 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 erabiltzearren, letra etzanean dauden hitzei baliabide guztian zehar azaltzen ez den ikur bat jarri zaio aurretik, @ ikurra. Transkribatzea errazagoa da eta eredua trebatzerakoan doitasun handiagoa ematen duela egiaztatu ahal izan dugu, letra etzanaz idatzita dauden hitzek aurretik ikur hau daramatela ikasiko baitu Krakenek. Frogatu dugun bezala, ia akatsik gabe ebatzi du hori.

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

12 Unicode karakterea erabili dugu: U+017F LATIN SMALL LETTER LONG S “ſ”.

2. irudia: Krakenek bi lerro bakarra balitz bezala interpretatzea



Aipatu bezala, lerroak dira programak erabiltzen dituen oinarrizko unitateak eta transkripzioa lerroka egin behar denez, bi lerrotan dagoen testua ezin da lerro bakarrean idatzi. Beraz, akats hauei aurre egiteko modurik eraginkorrena lerro horiei dagozkien transkripzio eremuak hutsik uztea da, programak kontuan har ez ditzan.

3.3. Ereduaren trebakuntza

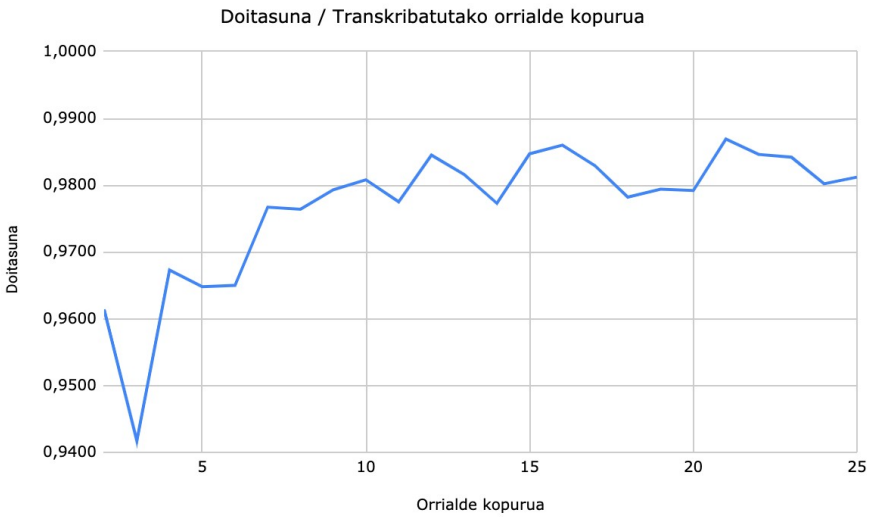
Eredua trebatzeko adimen artifizialaren adarra den ikasketa automatikoan oinarritzen da Kraken. Lerro bat 'irakurtzen' du eta ondoren akatsak konpentsatzen ditu aipaturiko neurona-sareaz baliatuz. Ondoren, lerroz lerro prozesua errepikatzen du azken lerrora iritsi arte, eta orduan berriz ere hasten da 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. Honek *overfitting* edo gaindoitzea ekiditen du, hots, eredua trebatze datuak bakarrik ezagutzeko doitzea, datu horietan aurkituko lituzkeen patroiak ezagutu beharrean. Horrela lortzen da akats gutxien egiten dituen eredua: doitasun handiena duen eredua, 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 grafikoan 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. Horrela, ondoriozta daiteke transkripzio lerro kopuru batetik gora ez duela merezi eredu berriak trebatzea, doitasuna ez baita esanguratsuki hobetuko. Hirurogeina 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 eredua lortu genuen, eta doitasun orokorra ez zen esanguratsuki handitu (% 98.5), baina ezohiko karaktereak, hasieran Krakenentzako “ezezagunak”, orduan ondo ezagutu zituen.

3.3. Emaitzen fitxategiak sortzea eta Wikitekara igotzea

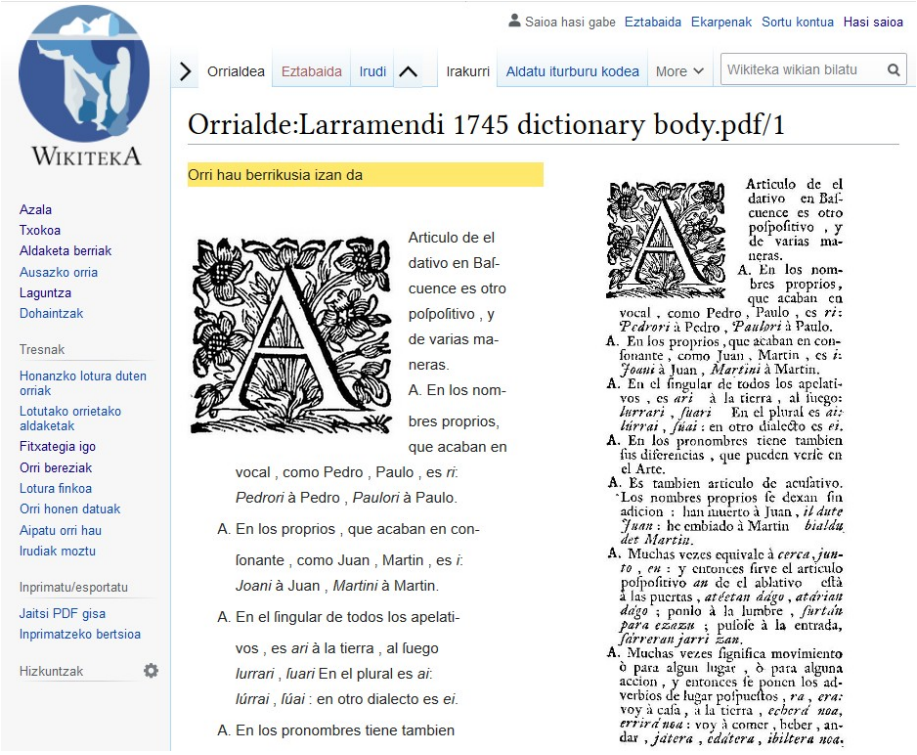
Emaitzak fitxategi formatu ezberdinetan ematen ditu Krakenek, hala nola txt edo ALTO XML[13], OCR emaitzak gordetzeko estandar bat. AEBtako Kongresuaren Liburutegiak, esaterako, erabiltzen du ALTO XML eta informazioa erauzteko Elexifier[14] erreminta-kateak input formatu bezala onartzen du. Testu fitxategi 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 berreraikitzeko aukera eskainiz. Testuko koskak irudikatzeko gai denez, HHren sarreren egitura irudika dezake, sarrera-buruak koska 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, *Wikiteka*k[15], OCR emaitzak erakusteko eta editatzeko aukera ematen du eta bertan *Wikitext* formatura[16] egokitu dugun ALTO fitxategien OCR emaitza[17] jarri dugu eskuragarri (ikus 4. irudia). Wikitekako edukiak irekiak direnez, transkripzioak zuzentzeko erabil daiteke. Are gehiago, badago zehazki transkripzio proiektuei eskainitako atal bat[18], jarduera honek gune horretan duen garrantziaren seinale. Beraz,

13 Ikus https://altoxml.github.io/
14 Ikus https://elexifier.elex.is
15 Euskarazko ataria: https://eu.wikisource.org/wiki/Azala.
16 Ikus https://en.wikipedia.org/wiki/Help:Wikitext.
17 Ikus https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza.
18 Ikus https://en.wikisource.org/wiki/Wikisource:Transcription_Projects.

komunitateak HHren transkripzioak zuzentzeko aukera izango du, beste herrialdetako proiektu askotan egin den bezala.

4. irudia. HHren 1. orrialdea Wikiteka plataforman[19]



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 zehaztasun teorikoa 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 hau eskuz zuzentzeko modu zuzenik ez dago. Trebakuntza sorta egiterakoan arazorik ematen ez badu ere, hiztegi guztia digitalizatzean horrek izango lukeen eragina neurtu beharko litzateke. Oraingoz 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 irizpideak jarraitu behar baditu, OCR prozesuan zehaztasun handia lortu arren, badirudi beharrezkoa izango dela eskuz eduki guztia balioztatzea, diseinuaren ezagutze prozesuko hutsegiteek sortutako akatsen zuzenketa barne.

5. Etorkizunerako 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 erauzketa prozesuan, transkripzio

19 Ikus https://eu.wikisource.org/wiki/Orrialde%3ALarramendi_1745_dictionary_body.pdf/1.

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 hau 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 automatiko algoritmo batek zatiketa eta anotazioa ikasi eta testu osoan aplikatzen ditu.

Dagoeneko beta fasean dagoen Elexifier tresna erabili dugu informazioa erauzteko (Lindemann & Alonso, prestatzen). 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. Gainera, HHtik erauzitako euskal formak Wikidatan integratu ditugu nahikoa ebidentzia izan dugun kasuetan, hau da, HHn topatu ditugun formak Wikidatan aurretik bazeuden lexemei gehitu dizkiegu, Elhuyar Fundazioak eta Euskal Wikipediako Kultur Elkarteak elkarlanean Elhuyar hiztegietako datu lexikalak Wikidata ekimenean txertatu dituztela baliatuz.

6. Erreferentziak

Alonso, M. (2021). Larramendiren *Hiztegi Hirukoitza*ren digitalizazioa. (Master Amaierako Lana)

Larramendi, M. (1745). *Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa*. San Sebastián: Bartholomé Riesgo y Montero.

Lindemann, D. & San Vicente, I. (2020). Baliabide lexikoen sarea: Baldintza filologiko eta tekniko zenbait. In *Hitzak sarean: Pello Salabururi esker onez*. Bilbo: UPV/EHU Argitalpen Zerbitzua.

Lindemann, D. eta Alonso, M. (prestatzen). A workflow for historical dictionary digitization: Larramendi's Trilingual Dictionary. *Proceedings of eLex 2021*, Brno, Txekia.

Romanov, M., Miller, M. T., Savant, S. B., & Kiessling, B. (2017). Important New Developments in Arabographic Optical Character Recognition (OCR). *arXiv preprint arXiv:1703.09550*.

Urgell, B. (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.

7. Eskerrak eta oharrak

- Artikulu "Hiztegi historikoen 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 libreen sorkuntza, trukea eta zabalpena bultzatzeko asmoz 2019. urtean egin zuten ikerketa deialdian hautatua izan zen.
- Bestalde, lan hau UPV/EHU-ko Euskal Hizkuntzalaritza eta Filologia Masterrean David Lindemann eta Blanca Urgellek zuzendutako "Larramendiren *Hiztegi Hirukoitza*ren digitalizazioa" Master Amaierako Laneko atal batetik eratorria da.