A workflow for historical dictionary digitisation: Larramendi's Trilingual Dictionary

David Lindemann, Mikel Alonso

UPV/EHU University of the Basque Country, Vitoria-Gasteiz, Spain
E-mail: david.lindemann@ehu.eus, mikelalon@gmail.com

Abstract

In this paper, we present a workflow for historical dictionary digitisation, with a 1745 Spanish-Basque-Latin dictionary as the use case. We start with scanned facsimile images, and get to represent attestations of modern standard Basque lexemes as Linked Data, in the form they appear in the dictionary. We are also able to produce an index of the dictionary, i. e. a Basque-Spanish version, and to map extracted Spanish and Basque lexical items to reference dictionary lemma list entries. The workflow is entirely based on freely available software. OCR and information extraction are performed using Machine Learning algorithms; data exhibits and the transcription curation environment are provided using Wikisource and Wikidata. Our evaluation of a first iteration of the workflow suggests its capability to deal with early modern printed dictionary text, and to reduce manual effort in the different stages significantly.

Keywords: Historical Lexicography; Digitisation; OCR; information extraction; Linked Data

1. Introduction

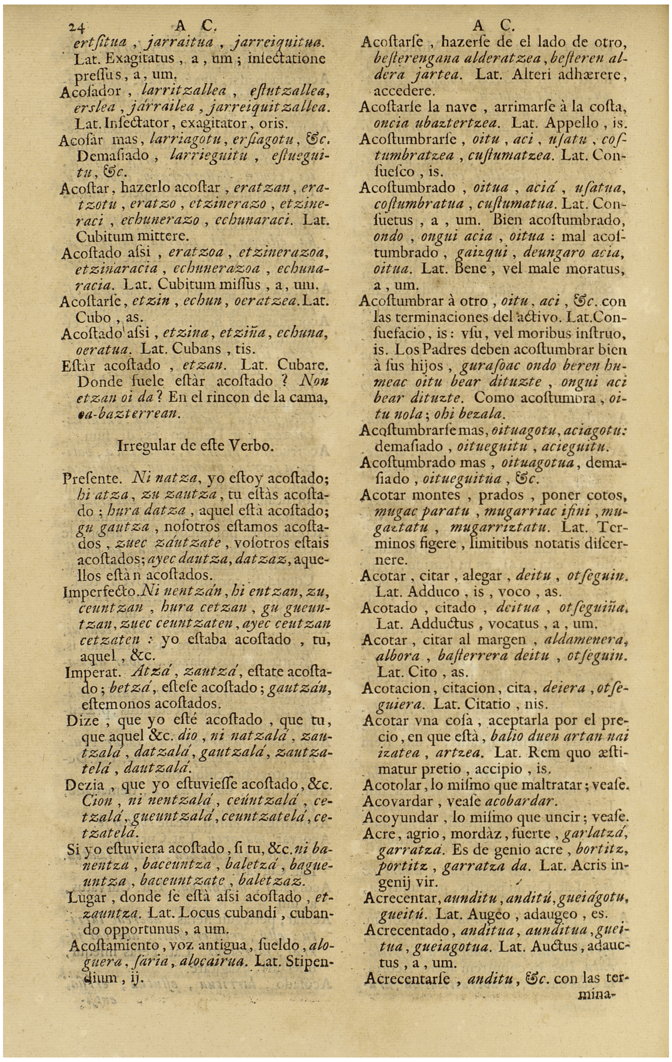
Manuel de Larramendi's Spanish-Basque-Latin Trilingual Dictionary in two volumes Larramendi (1745), henceforth LAR, for more than a century and a half has been the outstanding reference resource for Basque, and can be regarded the classic lexicographic work that brought a significant shift in the periodisation of Basque Lexicography (Urgell, 2002); it represents the beginning of modern Basque Lexicography. Nevertheless, this important classic is still available only as print dictionary, the digitisation of which has not overcome the stage of scanned images. The dictionary has been subject to in-depth philological and lexicographical research (Urgell, 1998a,b), which had to resort to manually compiled sets of examples, and thus was not able to include full-fledged quantitative methods that would take into consideration the content as a whole. For example, we do not have anything else than approximate estimations regarding the overall amount of headwords and distinct lemmata, and regarding the relation to the headword list of the 1725-1739 Spanish-Latin Diccionario de Autoridades (Real Academia Española, 2013), henceforth DA, the outstanding lexicographic work for Spanish at that time, which Larramendi used as primary reference for his dictionary.

In this project report, we reach out to propose and evaluate a workflow for digitisation, using the cited early modern print dictionary as showcase. Starting point is a collection of scanned images of both volumes of LAR dictionary, produced and provided by Koldo Mitxelena public library.[1] Following the digitisation stages outlined in Lindemann & San Vicente (2020), we apply a semi-automatic toolchain, and measure its rendering. This includes Optical Character Recognition (OCR), information extraction, and a first proposal for modeling attestations according to the Resource Description Framework (RDF), having in mind its integration in Wikidata.

Our main goal is the evaluation of the tested workflow, which includes an assessment of the precision reached by the employed tools, in order to make predictions concerning

[1] The item's first volume is available at https://www.kmliburutegia.eus/Record/26577, the second at http://www.kmliburutegia.eus/Record/203133.

manual validation and editing effort regarded necessary for a complete and accurate digitisation. We want to point out that the dictionary on hand is doubtlessly one of the harder nuts to crack, due to the early modern typefont, and lexicographic features. One working hypothesis therefore is the following: If we are able to get acceptable results for this dictionary with a predictable and limited manual workload, printed lexical resources published later than 1745 should require less effort to get digitised.



24 A C.

ertſitua , *jarraitua* , *jarreiquitua.* Lat. Exagitatus , a , um ; infectatione preſſus , a , um.

Acoſador , *larritzallea* , *eſtutzallea*, *erslea* , *járrailea* , *jarreiquitzallea.* Lat. Infectator , exagitator , oris.

Acoſàr mas , *larriagotu* , *erſiagotu* , *&c.* Demaſiado , *larrieguitu* , *eſtueguitu* , *&c.*

Acoſtar , hazerlo acoſtar , *eratzan* , *eratzotu* , *eratzo* , *etzinerazo* , *etzineraci* , *echunerazo* , *echunaraci.* Lat. Cubitum mittere.

Acoſtado aſsi , *eratzoa* , *etzinerazoa*, *etzinaracia* , *echunerazoa* , *echunaracia.* Lat. Cubitum miſſus , a , um.

Acoſtarſe , *etzin* , *echun* , *oeratzea.* Lat. Cubo , as.

Acoſtado aſsi , *etzina* , *etziña* , *echuna*, *oeratua.* Lat. Cubans , tis.

Eſtàr acoſtado , *etzan.* Lat. Cubare. Donde ſuele eſtàr acoſtado ? *Non etzan oi da* ? En el rincon de la cama, *oa-bazterrean.*

Irregular de eſte Verbo.

Preſente. *Ni natza*, yo eſtoy acoſtado; *hi atza* , *zu zautza* , tu eſtàs acoſtado ; *hura datza* , aquel eſtà acoſtado; *gu gautza* , noſotros eſtamos acoſtados , *zuec záutzate* , voſotros eſtais acoſtados; *ayec dautza*, *datzaz*, aquellos eſtàn acoſtados.

Imperfecto. *Ni nentzán* , *hi entzan*, *zu*, *ceuntzan* , *hura cetzan* , *gu gueuntzan*, *zuec ceuntzaten* , *ayec ceutzan cetzaten* : yo eſtaba acoſtado , tu, aquel , &c.

Imperat. *Atzá* , *zautzá*, eſtate acoſtado ; *betzá*, eſteſe acoſtado ; *gautzán*, eſtemonos acoſtados.

Dize , que yo eſté acoſtado , que tu, que aquel &c. *dio* , *ni natzalá* , *zautzalá* , *datzalá*, *gautzalá*, *zautzatelá* , *dautzalá.*

Dezia , que yo eſtuvieſſe acoſtado , &c. *Cion* , *ni nentzalá* , *ceúntzalá* , *cetzalá*, *gueuntzalá*, *ceuntzatelá*, *cetzatelá.*

Si yo eſtuviera acoſtado , ſi tu, &c. *ni banentza* , *baceuntza* , *baletzá* , *bagueuntza* , *baceuntzate* , *baletzaz.*

Lugar , donde ſe eſtà aſsi acoſtado , *etzauntza.* Lat. Locus cubandi , cubando opportunus , a um.

Acoſtamiento , voz antigua , ſueldo , *aloguera* , *ſaria* , *alocairua.* Lat. Stipendium , ij.

A C.

Acoſtarſe , hazerſe de el lado de otro, *beſterengana alderatzea* , *beſteren aldera jartea.* Lat. Alteri adhærere, accedere.

Acoſtarſe la nave , arrimarſe à la coſta, *oncia ubaztertzea.* Lat. Appello , is.

Acoſtumbrarſe , *oitu* , *aci* , *uſatu* , *coſtumbratzea* , *cuſtumatzea.* Lat. Conſueſco , is.

Acoſtumbrado , *oitua* , *aciá* , *uſatua*, *coſtumbratua* , *cuſtumatua.* Lat. Conſuetus , a , um. Bien acoſtumbrado, *ondo* , *ongui acia* , *oitua* : mal acoſtumbrado , *gaizqui* , *deungaro acia*, *oitua.* Lat. Bene , vel male moratus, a , um.

Acoſtumbrar à otro , *oitu* , *aci* , *&c.* con las terminaciones del activo. Lat. Conſuefacio , is : vſu , vel moribus inſtruo, is. Los Padres deben acoſtumbrar bien à ſus hijos , *guraſoac ondo beren humeac oitu bear dituzte* , *ongui aci bear dituzte.* Como acoſtumbra , *oitu nola* ; *ohi bezala.*

Acoſtumbrarſe mas , *oituagotu*, *aciagotu:* demaſiado , *oitueguitu* , *acieguitu.*

Acoſtumbrado mas , *oituagotua* , demaſiado , *oitueguitua* , *&c.*

Acotar montes , prados , poner cotos, *mugac paratu* , *mugarriac iſini* , *mugaztatu* , *mugarriztatu.* Lat. Terminos figere , limitibus notatis diſcernere.

Acotar , citar , alegar , *deitu* , *otſeguin.* Lat. Adduco , is , voco , as.

Acotado , citado , *deitua* , *otſeguiña.* Lat. Adductus , vocatus , a , um.

Acotar , citar al margen , *aldamenera*, *albora* , *baſterrera deitu* , *otſeguin.* Lat. Cito , as.

Acotacion , citacion , cita , *deiera* , *otſeguiera.* Lat. Citatio , nis.

Acotar vna coſa , aceptarla por el precio , en que eſtà , *balio duen artan nai izatea* , *artzea.* Lat. Rem quo æſtimatur pretio , accipio , is.

Acotolar , lo miſmo que maltratar ; veaſe.

Acovardar , veaſe *acobardar.*

Acoyundar , lo miſmo que uncir ; veaſe.

Acre , agrio , mordàz , fuerte , *garlatzá*, *garratzá.* Es de genio acre , *bortitz*, *portitz* , *garratza da.* Lat. Acris ingenij vir.

Acrecentar , *aunditu* , *anditú* , *gueiágotu*, *gueitú.* Lat. Augeo , adaugeo , es.

Acrecentado , *anditua* , *aunditua* , *gueitua* , *gueiagotua.* Lat. Auctus , adauctus , a , um.

Acrecentarſe , *anditu* , *&c.* con las termina-

Figure 1: LAR, vol. 1, page 24, scanned image

LAR presents several severe deviations from an up-to-date standard in print Lexicography. First, the early modern typefont, and the scanned images made from stained and half-transparent paper are to be mentioned as strong handicaps for OCR, which is the reason for the poor quality of LAR digital text versions available today. Second, the lexicographic structure is not consistently mirrored in structural markup and layout. That is true on macrostructural level (i. e. the segmentation of the dictionary text into entries), and concerning the lexicographical microstructure, in other words, the inner organisation of entries. This makes it evident that a rule-based segmentation of the dictionary text into labelled lexicographic components like "entry", "headword" and "translation equivalent", i.e. to "extract" the information to a format that can be interpreted by machines employing fixed rules, would not lead to satisfying results. Therefore, it becomes interesting to look at applications that use neural networks for the these tasks, since algorithms based on such technology are able to predict a result also in cases where a strict rule would fail. Applications for OCR and dictionary segmentation that use such technologies are available today, and we are witnessing their consolidation in the very recent past and present.

In the following, we present our experiments, for which we have employed tools that are freely available for research purposes, so that they are fully reproducible by anybody interested in this use case, or in similar endeavours.

2. Optical Character Recognition

An OCR tool converts images of characters to digital characters, i.e. it associates pixel patterns on an image with letters. The result, a digital text (txt), unlike the pure image (a collection of pixels), enables editing, searching, and computational processing of the textual content. State-of-the-art OCR tools rely on Machine Learning (ML) algorithms, that are trained on a manually transcribed subset of the work, and predict the mappings between letters as pixel patterns and as digital characters on that basis. The advent of ML in OCR technology has made the processing of early modern printing (and even hand-written text)[2] feasible: While in modern or even digital print characters can be mapped to uniform pixel patterns, in early modern printing, the patterns for the same letter may differ from each other in a significant way. In addition, pixel patterns may be disturbed by irregularities or stains on the paper, or ink from the reverse side of the pertaining page shining through. Similar to the flexibility in human reasoning, the ML algorithm tries to associate each pixel pattern it identifies to the most probable candidate letter, which means it can resolve doubts. The shortcomings of OCR tools developed for standard (modern) print become clear if we look at the text versions of LAR offered at the moment.[3] These can be roughly classified as follows: (a) characters that do not belong to a modern standard typeset, (b) characters that match to different pixel patterns, including the impact of stains or colour changes on the paper, and (c), errors due to wrong layout identification, i. e. errors in column and line segmentation.

Kraken[4] is a freely available OCR tool that relies on ML. It requires scanned images with a minimum resolution of 300 DPI, although authors of related work argue that even lower resolutions may serve. Kraken has shown that it outperforms leading proprietary OCR solutions designed for printed and manuscript documents, for example, digitising classical Arabic-script text ((Romanov et al., 2017). In addition, the fact that Kraken produces output following ALTO XML standard (see section 3) has been a reason for choosing this tool. We have favoured Kraken over Transkribus,[5] a web-based tool with similar features, because of its ability to be flexible towards font styles, i. e. that it is able to learn not only the character but also to discriminate font styles such as italics (see section 2.2 below).[6]

2.1 Pre-processing

As input, the Kraken tool needs scanned images, which are preprocessed following the guidelines,[7] i. e., the images are converted to black-and-white binary, and, if needed, their angle is corrected, so that lines appear horizontally, misalignments due to paper curvations

[2] For example, the Transkribus software uses ML for processing hand-written text. It also offers a graphical user interface for the creation of training sets, and the manual correction of the output, see https://transkribus.eu/Transkribus/. For a use case, see (Lindemann et al., 2018).

[3] Spanish National Library BNE (http://bdh-rd.bne.es/viewer.vm?id=0000015622), Google Books (https://play.google.com/books/reader?id=whdf0XXf6gwC), and Bavarian State Library BSB (https://opacplus.bsb-muenchen.de/title/BV035479582) offer image and txt versions of LAR.

[4] See http://kraken.re. This tool is built upon OCRropus (https://github.com/ocropus/ocropy) and features a user interface for ML training set creation. Kraken has been developed in the framework of the eScripta project at Université Paris Sciences et Lettres (cf. https://escripta.hypotheses.org/tag/kraken).

[5] See note 6.

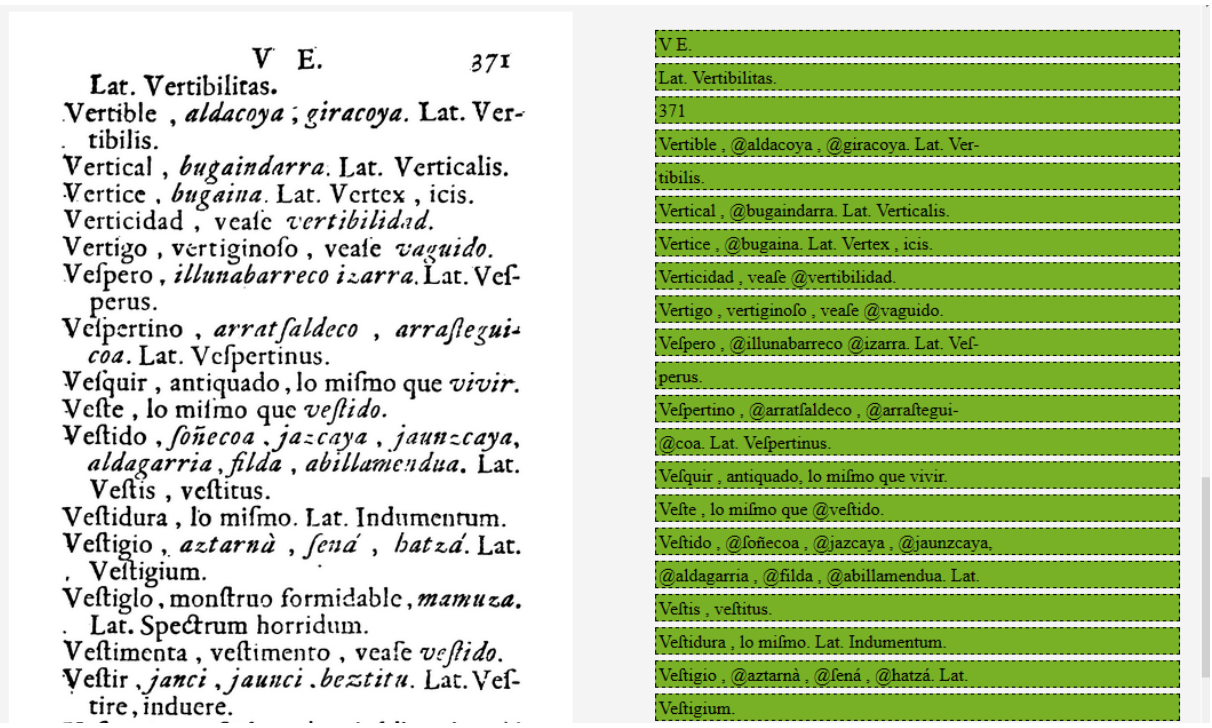
[6] This feature is not needed in hand-written text recognition, the task Transkribus was developed for.

[7] See http://kraken.re/training.html.

are eliminated, and stains are reduced. To this end, the ScanTailor application[8] has been used. We also have separated each LAR page into two, one for each of the two columns, in order to ease layout recognition to Kraken. Kraken's layout recognition module is then triggered, so that the files used in the transcription process are created.

2.2 Transcription

Kraken needs a training set, consisting of a certain amount of correctly transcribed lines. Before being given evidence from the training set, it is completely agnostic. In the guidelines, a set of 800 lines is recommended for training. It is clear that all characters that appear throughout the text to digitise have to appear in the training set. After transcribing one single two-column page of about 60 lines per column from scratch, we trained a Kraken model, and, from then on, corrected the OCR output page by page instead of transcribing from scratch. Any new corrected page was then introduced in the training set, in order to get constantly improved results which would require less corrections on the remaining pages. Despite very encouraging precision rates, that from the very first dictionary page on were clearly above the precision found in the available LAR txt versions, we realised that certain (infrequent) characters were not recognised. In the subsequent training sets, we added transcriptions of pages that contained the missing characters, mainly upper case letters that would appear massively in their corresponding alphabet sections. As the precision rates presented in Table 1 below suggest, overall precision has not significantly grown, but the infrequent characters formerly "unknown" to Kraken had now been properly identified.



V E. 371
Lat. Vertibilitas.
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* , *ſená* , *hatzá*. Lat.
Veſtigium.
Veſtiglo , monſtruo formidable , *mamuza*.
Lat. Spectrum horridum.
Veſtimenta , veſtimento , veaſe *veſtido*.
Veſtir , *janci* , *jaunci* . *beztitu*. Lat. Veſ-
tire , induere.

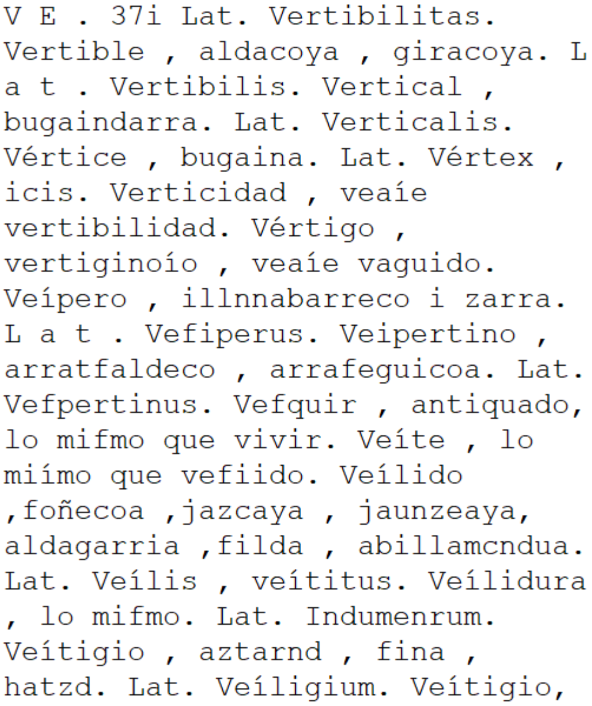
```
V E.
Lat. Vertibilitas.
371
Vertible , @aldacoya , @giracoya. Lat. Ver-
tibilis.
Vertical , @bugaindarra. Lat. Verticalis.
Vertice , @bugaina. Lat. Vertex , icis.
Verticidad , veaſe @vertibilidad.
Vertigo , vertiginoſo , veaſe @vaguido.
Veſpero , @illunabarreco @izarra. Lat. Veſ-
perus.
Veſpertino , @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 , @ſená , @hatzá. Lat.
Veſtigium.
```

Figure 2: Kraken OCR output, displayed by the transcription module

[8] Available at http://scantailor.org. ScanTailor is free software.

Transcriptions are performed inside a set of html files rendered by a web browser (see Figure 2). To each text box, which usually is a single text line on the scanned image recognised by the Kraken layout recognition module, a text field is provided. For creating the first training set, these fields are empty, and have to be filled with the text read by the user from the corresponding line. The modified page is then saved for inclusion in the training set. After a first OCR iteration, new html files are produced for a custom set of dictionary pages, and the text fields now contain the text recognised by Kraken based on the first model (derived from the first training set). From now on, the text in these fields is not typed in from scratch, but manually corrected. Corrected entire pages can be added to the training set, so that, in the next iteration, they are also considered for building the upgraded text recognition model, and so on, until the desired precision threshold is reached. Transcriptions must always reflect what is represented in typed letters in the original, without amendments or omissions, following the guidelines for Ground Truth Transcription.[9]

LAR contains two font styles, regular and italics. Kraken transcriptions are plain text without any markup, but the algorithm can deal with this using the following method: In the transcription, any word in italics is preceded by a sign not present in the whole resource, for which we chose an '@'. Kraken will learn that words written in italics, in the transcription should be preceded by this sign. As we could verify, this has worked out almost perfectly.



```
V E . 37i Lat. Vertibilitas.
Vertible , aldacoya , giracoya. L
a t . Vertibilis. Vertical ,
bugaindarra. Lat. Verticalis.
Vértice , bugaina. Lat. Vértex ,
icis. Verticidad , veaíe
vertibilidad. Vértigo ,
vertiginoío , veaíe vaguido.
Veípero , illnnabarreco i zarra.
L a t . Vefiperus. Veipertino ,
arratfaldeco , arrafeguicoa. Lat.
Vefpertinus. Vefquir , antiquado,
lo mifmo que vivir. Veíte , lo
miímo que vefiido. Veílido
,foñecoa ,jazcaya , jaunzeaya,
aldagarria ,filda , abillamcndua.
Lat. Veílis , veítitus. Veílidura
, lo mifmo. Lat. Indumenrum.
Veítigio , aztarnd , fina ,
hatzd. Lat. Veíligium. Veítigio,
```

Figure 3: BNE txt version of LAR

After transcribing 50 columns of about 60 lines each (i. e., 25 pages, cf. Figure 4), we assumed that precision would not significantly increase. In Table 1, we list the precision rates reached after each OCR iteration, with the amount of columns present in the training set. LAR, volume I and II, contains 1676 columns (two per page). This means that after manually transcribing less than 3% of the content we have gained a precision of nearly 98.5% in a txt version that covers the whole dictionary. This clearly outperforms the txt versions available before (cf. BNE version in Figure 3). Precision rates are calculated by Kraken, which uses a 10% share of the given training data as the evaluation set. Nevertheless, there is a drawback to take into account: Kraken's layout recognition module has worked out with a high precision, but still a considerable amount of lines have not been recognised. Either a line is not recognised at all, or lines are wrongly joined, so that a recognised text box range includes two real lines instead of one. Since there is no straightforward way to correct these mistakes manually, we had to leave this question for the (near) future, when our participation in a workshop related to Kraken will be possible.

[9] See https://ocr-d.de/en/gt-guidelines/.

Pages	2	3	4	5	6	7	8	9	10	11	12	13
Precision	0.9614	0.9417	0.9673	0.9648	0.965	0.9767	0.9764	0.9793	0.9808	0.9775	0.9845	0.9816
Pages	14	15	16	17	18	19	20	21	22	23	24	25
Precision	0.9773	0.9847	0.986	0.9829	0.9782	0.9794	0.9792	0.9869	0.9846	0.9839	0.9802	0.9812

Table 1: OCR precision rates.

At this stage, we cannot measure the impact of mistaken line recognitions, but our revisions of the OCR output during the transcription process and beyond suggest that it is worth digging deeper at this point, towards including a layout recognition validation step in the workflow. Anyhow, if the final resource, i.e. a digitised version of LAR has to obey quality criteria as for an edited publication, despite a very high OCR precision it seems necessary to manually validate all content, and this should include the correction of any errors due to mistakes in layout recognition. By tracking that effort we will get precise information concerning the precision of automatic layout recognition.

2.3 Result export

Kraken exports OCR results in different formats, and among them, txt, and ALTO XML,[10] an OCR result representation standard used e.g. by the Library of Congress, produced by some proprietary OCR engines, and supported as input format by the Elexifier toolchain (see section 3). While plain text contains just the recognised characters of each text box, separated by line breaks, ALTO XML also preserves the exact position of each text box on the page. This means, for example, that line indents are represented, which is an essential layout feature used for entry structure representation in LAR (headwords appear with a different indent than subsequent entry lines), and therefore is information worth considering. This information is also needed for publishing a digital version of the source document that includes active links between bits of text on the image and in the transcription.

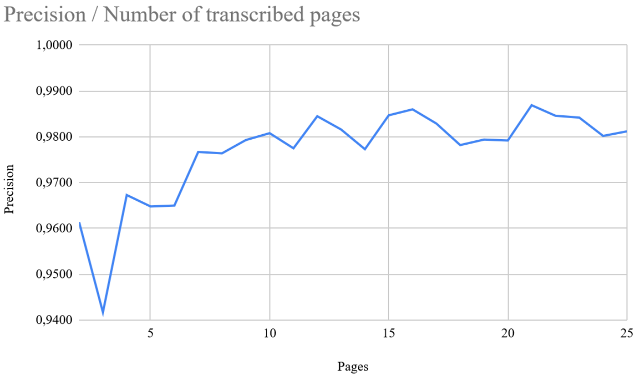


Figure 4: OCR precision rates

2.4 Wikisource

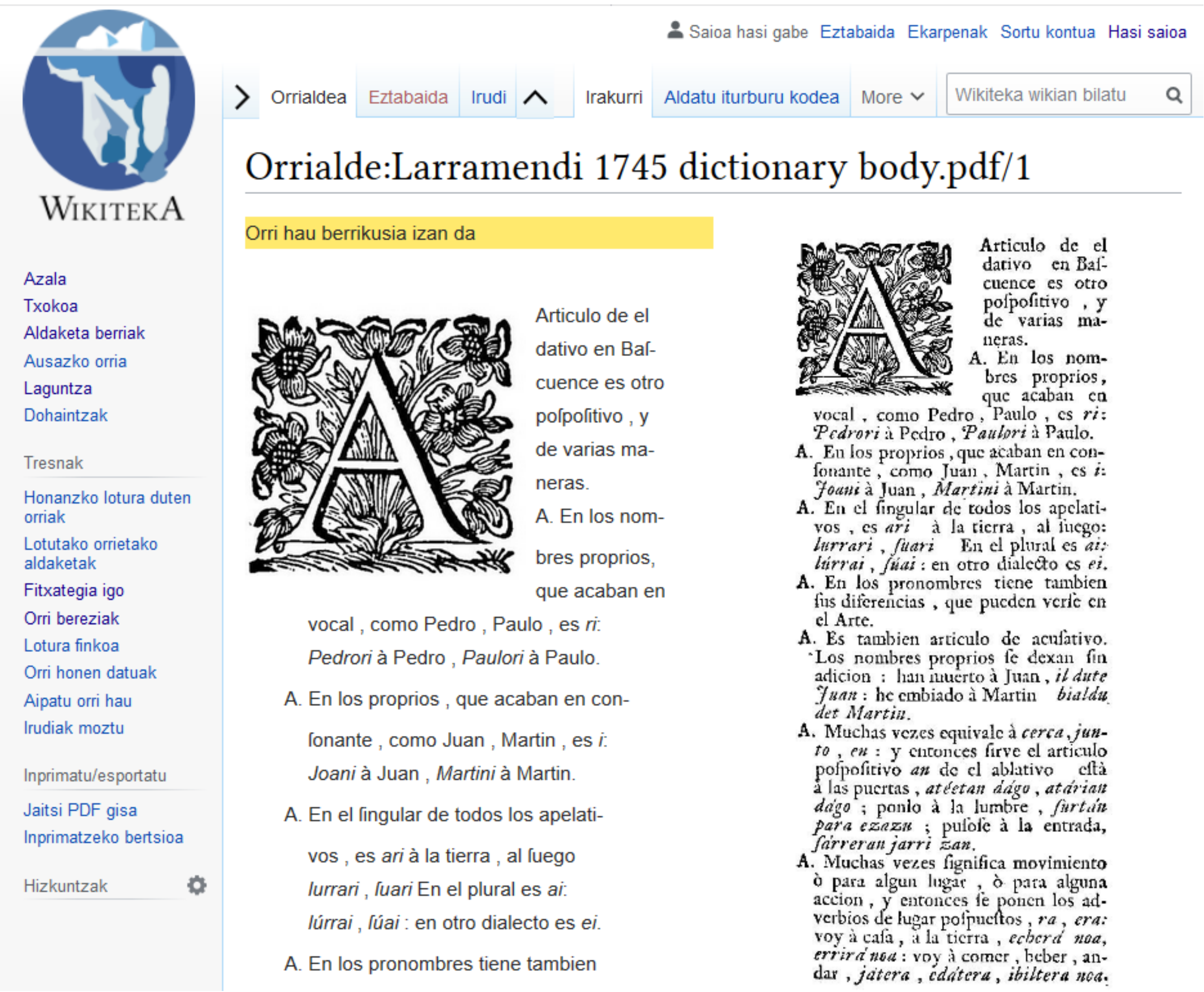
The Basque branch of Wikimedia's Wikisource platform, Wikiteka[11] (see Figure 5)[12] can be used for exhibiting and collaboratively editing OCR results: Anyone can view scanned images, along with their transcriptions, and edit the latter, in order to correct errors.

[10] See https://altoxml.github.io/.

[11] Accessible at https://eu.wikisource.org/wiki/Azala.

[12] See online at https://eu.wikisource.org/wiki/Orrialde:Larramendi_1745_dictionary_body.pdf/1.

The choice of that platform for collaborative OCR correction is due to Wikimedia Basque Country funding this small project; but a generally applying reason for that choice would be the fact that there exists an active community around Wikiteka, which has completely validated transcriptions of literary works of considerable size.[13] With this goal in mind, we have transformed the OCR result from ALTO XML format to Wikitext format.[14] Unfortunately, Wikitext format does not allow including text box position data, but nevertheless we are able to represent line indents in Wikitext, which is the layout feature used in LAR for marking up headwords (negative indent), in opposition to consequent entry lines (normal indent). Using the text box position data present in ALTO, we have defined a filter that isolates text lines with negative indent, and from these, those lines which start with a capital letter that belongs to the pertaining alphabet section. From that subset of lines we chose the first part, i.e. until the first whitespace or punctuation.[15] These headword candidates have been enriched with a Wikitext markup that allows navigation



Orrialde:Larramendi 1745 dictionary body.pdf/1

Orri hau berrikusia izan da

Articulo de el
dativo en Baſ-
cuence es otro
poſpoſitivo , y
de varias ma-
neras.
A. En los nom-
bres proprios,
que acaban en
vocal , como Pedro , Paulo , es *ri*:
Pedrori à Pedro , *Paulori* à Paulo.

A. En los proprios , que acaban en con-
ſonante , como Juan , Martin , es *i*:
Joani à Juan , *Martini* à Martin.

A. En el ſingular de todos los apelati-
vos , es *ari* à la tierra , al ſuego
lurrari , *ſuari* En el plural es *ai*:
lúrrai , *ſúai* : en otro dialecto es *ei*.

A. En los pronombres tiene tambien

Articulo de el dativo en Baſcuence es otro poſpoſitivo , y de varias maneras.

A. En los nombres proprios , que acaban en vocal , como Pedro , Paulo , es *ri*: *Pedrori* à Pedro , *Paulori* à Paulo.

A. En los proprios , que acaban en conſonante , como Juan , Martin , es *i*: *Joani* à Juan , *Martini* à Martin.

A. En el ſingular de todos los apelativos , es *ari* à la tierra , al ſuego: *lurrari* , *ſuari* En el plural es *ai*: *lúrrai* , *ſúai* : en otro dialecto es *ei*.

A. En los pronombres tiene tambien ſus diferencias , que pueden verſe en el Arte.

A. Es tambien articulo de acuſativo. Los nombres proprios ſe dexan ſin adicion : han muerto à Juan , *il dute Juan* : he embiado à Martin *bialdu det Martin*.

A. Muchas vezes equivale à *cerca , junto* , *en* : y entonces ſirve el articulo poſpoſitivo *an* de el ablativo eſtà à las puertas , *atéetan dágo* , *atárian dágo* ; ponlo à la lumbre , *ſurtán para ezazu* ; puſoſe à la entrada, *ſárreran jarri zan*.

A. Muchas vezes ſignifica movimiento ò para algun lugar , ò para alguna accion , y entonces ſe ponen los adverbios de lugar poſpueſtos , *ra* , *era*: voy à caſa , à la tierra , *echerá noa*, *errirá noa* : voy à comer , beber , andar , *jatera* , *edátera* , *ibiltera noa*.

Figure 5: LAR sample on Basque Wikisource platform

[13] See e.g. https://eu.wikisource.org/wiki/Gero for a Basque literature classic, or https://eu.wikisource.org/w/index.php?title=Berezi:OrrialdeGuztiak for a list of transcriptions.

[14] See documentation at https://en.wikipedia.org/wiki/Help:Wikitext.

[15] See code at https://github.com/dlindem/LBLR/blob/master/Larramendi/wikisource/wssarreraanchor.py.

inside the dictionary text (see Figure 6 below.) We have uploaded the plain text enriched in the described way to Wikiteka, together with the corresponding scanned images.[16] The task of correcting any errors, aiming to increase the transcription precision to 100%, is thus delegated to the community of Wikiteka users, which is open to anybody. General guidelines for transcription are given on the platform.[17] To that we add here some points to have in mind in this particular case, and as explanation of the sample shown in Figure 6:[18]

- Centred text (like the "A B." running title in Figure 6) will be preceded by five colons (":::::").
- Negative indent lines will be preceded by one colon (":").
- Other lines will be preceded by two colons ("::").
- Words in italics will be enclosed in pairs of single quotes (i.e. two "’", before and after the word).
- Line breaks and end-of-line hyphenations will be kept as in the scanned original.
- If the anchor markup element is not properly set, like in the second line of the example page in Figure 6, that shall be corrected. In this case, where the OCR tool has missed to identify the first capital letter ‘A’, the corrected line will start "{{sarrera|abandono}}Abandono".
- The anchor markup element that encloses headword candidates contains a single word. In the case of homograph headword candidates, that anchor includes a disambiguation number. If the anchor, instead of a single word, should enclose a multiword unit, the anchor shall be manually adapted, like for the entry with headword "abaratado demasiadamente" where "merquetueguia" and "merquequi ifinia" are listed as Basque equivalents: the anchor’s scope will be widened to two words, so that "{{sarrera|abaratado}}" (Figure 6) will be corrected to "{{sarrera|abaratado demasiadamente}}",[19] while leaving the following text as it is.

3. Information Extraction

Our method for isolating Spanish headword candidates described in the preceding section is entirely rule-based; it takes into account the text line position data present in ALTO format, and the correspondence of the first capital letter in that line to the pertaining alphabet section. We have defined as the headword candidate what precedes a whitespace or punctuation sign. Another method for defining headword candidates is to manually annotate headwords in a sample, and train a ML tool for predicting headword candidates in the whole dictionary text. Such a method can provide results that may be complementary to the rule-based approach.

Very recently, the ELEXIS project[20] launched Elexifier,[21] a toolchain supported by graphical user interfaces for information extraction from dictionaries. Dictionary content

[16] Accessible at https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza. The scanned images are a processed version (see section 2.1) of the image collection distributed by Koldo Mitxelena public library.

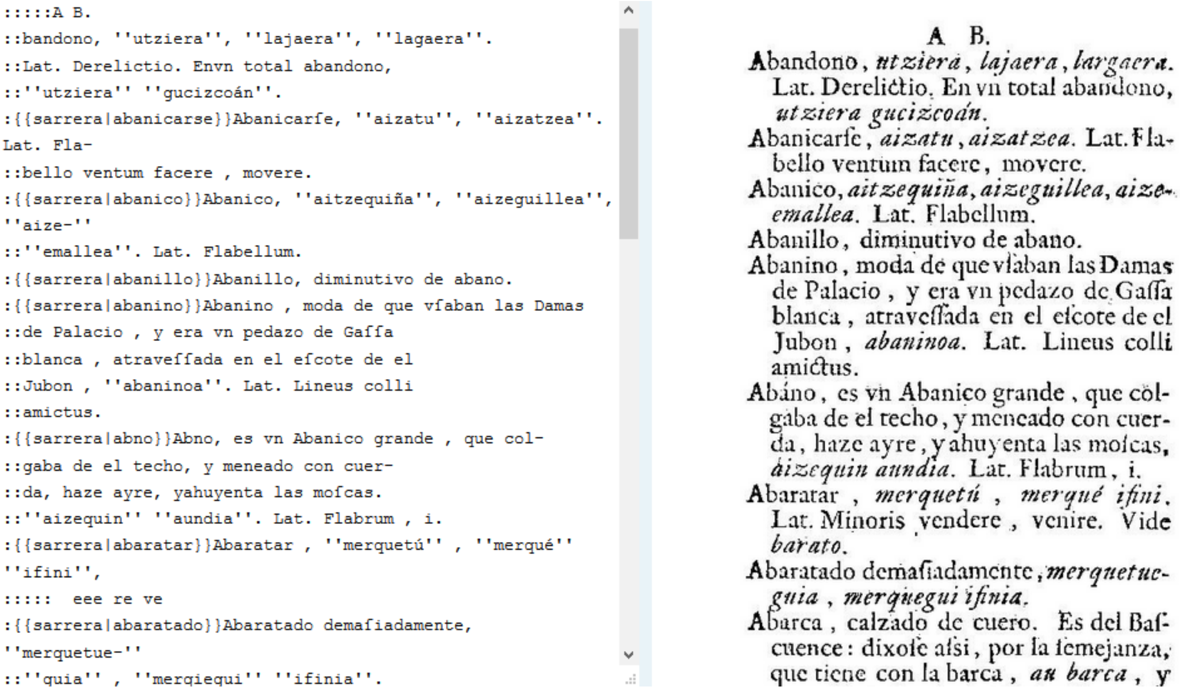
[17] For an English version, see https://en.wikisource.org/wiki/Help:Page_status.

[18] See online at https://eu.wikisource.org/w/index.php?title=Orrialde:Larramendi_1745_dictionary_body.pdf/4&action=edit.

[19] Note: "sarrera" is the Basque equivalent for "entry".

[20] See project homepage at http://elex.is.

[21] See https://elexifier.elex.is/. UPV/EHU has an observer status in the ELEXIS project, and among other activities, it is early adopter of the Elexifier toolchain, being this project a first use case. Other



```
::::A B.
::bandono, ''utziera'', ''lajaera'', ''lagaera''.
::Lat. Derelictio. Envn total abandono,
::''utziera'' ''gucizcoán''.
:{{sarrera|abanicarse}}Abanicarſe, ''aizatu'', ''aizatzea''.
Lat. Fla-
::bello ventum facere , movere.
:{{sarrera|abanico}}Abanico, ''aitzequiña'', ''aizeguillea'',
''aize-''
::''emallea''. Lat. Flabellum.
:{{sarrera|abanillo}}Abanillo, diminutivo de abano.
:{{sarrera|abanino}}Abanino , moda de que vſaban las Damas
::de Palacio , y era vn pedazo de Gaſſa
::blanca , atraveſſada en el eſcote de el
::Jubon , ''abaninoa''. Lat. Lineus colli
::amictus.
:{{sarrera|abno}}Abno, es vn Abanico grande , que col-
::gaba de el techo, y meneado con cuer-
::da, haze ayre, yahuyenta las moſcas.
::''aizequin'' ''aundia''. Lat. Flabrum , i.
:{{sarrera|abaratar}}Abaratar , ''merquetú'' , ''merqué''
''ifini'',
::::  eee re ve
:{{sarrera|abaratado}}Abaratado demaſiadamente,
''merquetue-''
::''quia'' , ''merqiequi'' ''ifinia''.
```

A B.

Abandono, *utziera*, *lajaera*, *largaera*. Lat. Derelictio. En vn total abandono, *utziera gucizcoán*.

Abanicarſe, *aizatu*, *aizatzea*. Lat. Flabello ventum facere, movere.

Abanico, *aitzequiña*, *aizeguillea*, *aizeemallea*. Lat. Flabellum.

Abanillo, diminutivo de abano.

Abanino, moda de que vſaban las Damas de Palacio, y era vn pedazo de Gaſſa blanca, atraveſſada en el eſcote de el Jubon, *abaninoa*. Lat. Lineus colli amictus.

Abáno, es vn Abanico grande, que colgaba de el techo, y meneado con cuerda, haze ayre, y ahuyenta las moſcas, *aizequin aundia*. Lat. Flabrum, i.

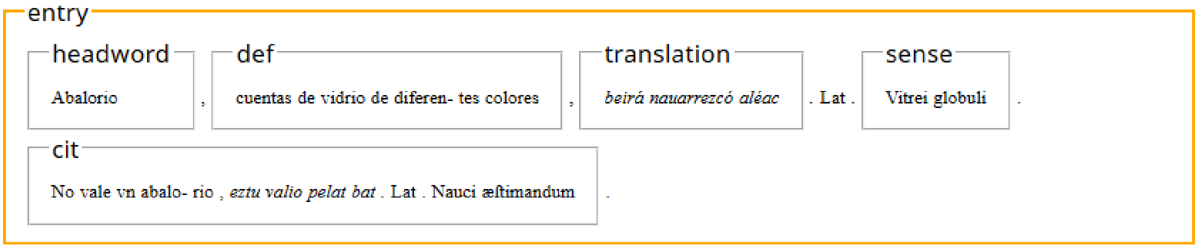
Abaratar, *merquetú*, *merqué iſini*. Lat. Minoris vendere, venire. Vide *barato*.

Abaratado demaſiadamente, *merquetueguia*, *merquegui iſinia*.

Abarca, calzado de cuero. Es del Baſcuence: dixoſe aſsi, por la ſemejanza, que tiene con la barca, *au barca*, y

Figure 6: Wikitext "source code" editable view on Wikiteka platform

available as text or rich text (in PDF format) or ALTO XML is parsed into an XML structured format that represents the structure of the dictionary, i.e. the division between entries, and inside the entry, the division into lexicographic items such as headwords, definitions, and translation equivalents. For this task, a ML application is trained by providing manually annotated training material.



Abalorio , cuentas de vidrio de diferen- tes colores , *beirá nauarrezcó aléac* . Lat . Vitrei globuli .

No vale vn abalo- rio , *eztu valio pelat bat* . Lat . Nauci æſtimandum .

Figure 7: Elexifier annotation module, graphical interface

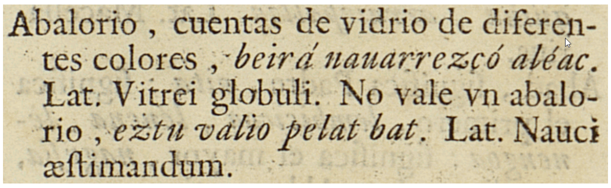
The Elexifier toolchain is currently in beta stage, and still subject to some feature restrictions. In particular, a limited tagset for the representation of microstructural items is available as for the current version: Entry, and as child elements of Entry: Headword, Translation, Sense, Part of Speech, Definition, and Example. The XML element tags that correspond to these lexicographical items are defined according to TEI-Lex0.[22]

use cases are planned. Hence, we were interested to test Elexifier in the workflow presented here. Another tool with similar features (that supports a more complete TEI tagset, but lacks a graphical interface), is GROBID-dictionaries, see https://github.com/MedKhem/grobid-dictionaries.

[22] In the framework of the Text Encoding Initiative (TEI), and DARIAH-EU working group "Lexical Resources", co-funded by the ELEXIS project, a tagset for representation of dictionary content has been developed and proposed as standard, in order to ensure interoperability of lexical datasets, see https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

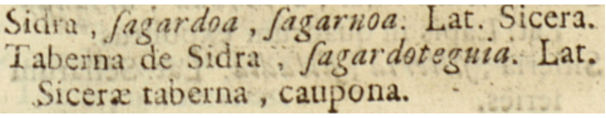
We have annotated a sample of LAR assigning tags to entries ("entry"), headwords ("headword"), definitions ("def"), Basque translations ("translation"), examples and notes ("cit"), and Latin translations (due to the limitations in the available tagset, "sense").[23] This can be observed in the screenshot image from the Elexifier annotation module reproduced as Figure 7, together with an image of the original entry (Figure 8).

Following the recommendations given in Elexifier documentation, the annotation has been carried out for twenty columns (ten dictionary pages), and then used as training set for the Elexifier segmentation ("information extraction") module, which structures the content of the whole dictionary according to what it has been given as training set.



Abalorio , cuentas de vidrio de diferentes colores , *beirá nauarrezçó aléac.* Lat. Vitrei globuli. No vale vn abalorio , *eztu valio pelat bat.* Lat. Nauci æſtimandum.

Figure 8: LAR entry example, scanned image



Sidra , *ſagardoa* , *ſagarnoa.* Lat. Sicera. Taberna de Sidra , *ſagardoteguia.* Lat. Siceræ taberna , caupona.

Figure 9: LAR entry example, scanned image

A first evaluation of the information extraction results suggests that Spanish headwords and Basque translation equivalents have been recognised by the software with high precision. Latin equivalents, the third category we have looked at, have been recognised with much lower precision. Headwords seem to be recognised seamlessly, which should be due to the fact that headwords are positioned in the entry layout in a first negatively indented line, and followed by a comma. This has been the case in all annotated entries, and thus is a very straightforward criterion for the ML algorithm. On the other hand, also items that do not describe headwords are placed in a negatively indented line, and subsequently, have been identified as headwords. This is the case for the items listed in the example shown in Figure 1 above, between "acostar" and "acostamiento", where non-canonical inflected and combined forms of the preceding headword (such as "estar acostado"), and even items representing grammatical information that serve for introducing a list of inflected forms appear in that position. Figure 9 also contains an example for a sub-entry that appears just as headwords appear, but in this case, not only breaking alphabetical order but totally out of the scope of the current alphabet section. Latin equivalents, as can be observed in Fig. 8, will appear for headwords, but also as translation of usage examples; here we have contradictory evidence that makes the algorithm unable to predict the correct annotation for Latin items in many cases.

Abondamiento , abondar. Veaſe *abundamiento* , *abundar.*

Figure 10: LAR entry example, scanned image

Basque equivalents are not that clearly identifiable, since their layout feature (italics font style) is also present in examples (Basque translation of the idiomatic usage example, see Fig. 8), and also in Spanish to Spanish

[23] According to TEI-Lex0, "sense" is not at all defined as adequate for annotating translation equivalents; it groups word senses within entries. Due to the lack of an appropriate tag in the current version of Elexifier, we have nevertheless chosen this workaround.

cross-references, as e. g. in "Acovardar" in Fig. 1 above,[24] and in the example shown in Fig. 10. In fact, "abundamiento" and "abundar", which in that example entry, correctly identified, would represent Spanish headwords where a cross-reference leads to, have been identified by the software as Basque equivalents. This should be solved by annotating cross-references (that in the dictionary text are preceded or followed by the structure markers "vease" or "lo mismo que") with a special tag, not available in the present version of Elexifier,[25] so that the segmentation algorithm gains evidence for identifying words preceded by such structure marker as cross-reference, regardless of their font style.[26]

Compiling the training set for Elexifier, in cases of multiword items as headwords, we have annotated it accordingly as multiword headword (i.e., for example, "abaratado demasiadamente", with "merquetueguia" as equivalent). Using that evidence, Elexifier has identified multiword headwords in 1,925 cases in the whole dictionary text. If we compare the results of both methods (rule-based and ML-based) regarding the whole headword list, we gain the figures shown in Table 2:

	LAR, rule-based	**LAR, ML-based**
Spanish Headwords	36,451	29,932
- of which appear in DA	33,015	25,057
- of which are multiword items	0	1,925

Table 2: Items identified as headword

The ruleset for spelling normalisation and comparison will be explained in the following.

4. Merging historical lemma lists

In order to compare the Spanish lemma list extracted from LAR to DA lemma list,[27] we have performed a normalisation of lemma-signs found in both resources. This step is necessary for defining pairs of matching lemmata that from resource to resource show different written representations. For the purpose of achieving mappings such as those represented in Table 3, we processed all lemma-signs of both dictionaries with the unidecode Python module,[28] which removes all diacritics and replaces non-canonical (non-ASCII) characters with the most similar canonical one. We excepted the "ñ" letter, canonical in Spanish, from that replacement, preventing it from being converted to "n".

[24] As can be observed, this layout feature is not strictly applied: in the follwing entry "acoyundar", the cross-referenced headword is not printed in italics.

[25] As explained above, we have used all available tags, so that, for cross-references, in this first experimental iteration, we had no remaining option.

[26] As soon as Elexifier offers a full-fledged tagset, we will repeat the process. An interim solution for identifying such cross-reference items in the output of Basque translation equivalents, we can check for the presence of the items in the DA headword list, which for the example solves the problem, since "abundamiento" and "abundar" both are listed there as headwords. The Basque item, on the other hand, should be checked for if it is a homograph translation of a Spanish headword (the headword of the same entry), such as LAR Basque equivalent "saca" for Spanish "saca".

[27] The DA lemma list is available at http://web.frl.es/DA_Preliminares/DA_lemario.pdf. This list contains headwords only. Other parts of the digitised DA are accessible only through a graphical user interface that allows one-by-one queries by lemma (available at http://web.frl.es/DA.html). The unabridged content is not publicly accessible in any other format than on paper.

[28] Available at https://pypi.org/project/Unidecode/.

Also, we converted all upper case characters to lower case, and double s to single s (historical "f" having been converted to "s" by the unidecode module). As the examples listed in table 3 show, "ss" (which in 18th century Spanish was still frequent) and diacritics are not used in the same way, and also their use inside LAR and DA is not concise. We will evaluate the described normalisation process in detail, having in mind related work about historical Spanish, which uses an approach based on Levenshtein distance thresholds, which is more flexible, but prone to yield false-positive mappings (Porta et al., 2013). We then wrote normalised lemma-signs from LAR and DA, their original written representations, and, for LAR, also Basque equivalents, as elements into the same XML tree, so that we were able to produce the datasets[29] listed in Table 4.[30]

LAR	**DA**	**matching normalised lemma sign**
Obsession	obsessión	obsesion
Hueffo	huesso	hueso
Occiffion	occisión	occision
Atràs	atras	atras

Table 3: Lemma-sign normalisation mappings

#	**List**	**Rule or ML**	**Rule**	**ML**	**Rule and ML**
1	LAR: all lemmata	32,700	30,045	27,125	24,470
2	Union of LAR and DA: all lemmata	46,843			
3	Lemmata appearing in LAR, but not in DA	4,875	2,431	4,875	2,431
4	Lemmata appearing in DA, but not in LAR	14,143	14,354	19,718	19,929
5	LAR and DA: intersection	27,825	27,614	22,25	22,039
6	All items extracted as LAR Basque equivalent candidates	60,193	58,235	38,300	36,342
7	LAR equivalents that also appear in SAR, WD, or OEH	15,152	14,886	11,551	11,285
8	LAR equivalents that also appear in SAR with "1745" datation	3,134	3,088	2,508	2,461
9	LAR equivalents that also appear in SAR with "1745" datation, and in WD	1,478	1,456	1,201	1,179
10	LAR equivalents with attestation in Wikidata (2021-01)	1,416	1,396	1,151	1,131

Table 4: Produced datasets

Besides that, we produced an index of LAR, that is, a version where Basque lexical items point to their Spanish equivalents, the original lemmata. In Table 4 (6-10), we show the amounts of Basque items extracted using both methods. Based on rules, we got all items printed in italics, that is, as explained above, not only Basque items, but all content printed in italics. We compare these amounts with those obtained from the Elexifier tool, for which we had manually annotated a sample, as explained in section 3. We have developed a set of rules for linking historical spellings of Basque lexical items to standard

[29] These datasets are available at http://lexbib.org/larramendi.

[30] For this task, we have used the TLex Dictionary Writing System, see https://tshwanedje.com/tshwanelex/.

spelling, similar to the approach used for matching Spanish LAR and DA headwords, but that contains a total of 36 regular expressions, to be executed in a certain order.[31] The ruleset is discussed in detail in Alonso Arrospide (2021). We then mapped LAR Basque equivalents in their written representation as modified by the ruleset to the lemma lists of SAR (Sarasola, 1996), OEH (Mitxelena & Sarasola, 1988) dictionaries, and Wikidata Basque lexemes (WD), with the results listed in Table 4. These datasets now are available for further research that can also include quantitative methods, although, for this version of the datasets we must stress the fact that transcription precision is below 100%.[32] Having a closer look at the data, for example, in list (4), 457 items can be found that describe superlative inflected adjective forms (e. g., "alegrissimo", "aliviadissimo"), which apparently are referenced systematically as lemmata in DA, but not in LAR. This suggests that groups of lemmata present in DA but missing in LAR can be, at least in part, identified in groups. This list obviously also contains those LAR headwords that have not been properly converted to text in the OCR process, or that constitute an orthographical variant that has not been handled in the normalisation process. List (3), in turn, also contains headwords that due to OCR errors or failed normalisation have not been mapped to their counterpart in DA, but in addition to these, it contains those headwords that have been added by Larramendi, without having had reference in DA (e.g. "derecho natural", in Basque, "sortaraudea", "sorneurtartea").

5. Enriching Wikidata

5.1 Wikidata Lexemes

In section 4, we have shown how we have performed a merging of historical lemma lists, a process that also can be seen as a linking of lexical resources, at the lemma sign level (i.e., without regard to part of speech or word sense disambiguation). We have taken two resources into consideration, LAR and DA. In order to link a lexical dataset to more and different resources, the workflow proposed for Elexifier resorts to the already mentioned TEI-Lex0 XML annotation scheme, which has been developed for that purpose (Bański et al., 2017).

Another way to link lexical data, which can be characterised as an upcoming trend regarding Linked Open Data,[33] is to make use of Wikidata lexemes. Wikidata represents entities such as concepts, lexemes, and properties that describe relations between the former, according to the Resource Description Framework (RDF). RDF uses semantic triples consisting of subject, predicate, and object, for the representation of statements, which can be visualised through the Wikidata graphical interface[34] or retrieved through a query interface using SPARQL.[35]

If we look at how a lexeme in Wikidata is linked to the concept it denotes on one hand, and to translation equivalents on the other, we find that while in dictionaries statements about

[31] See ruleset at https://github.com/dlindem/LBLR/blob/master/Larramendi/erkaketak_eus_elexifier/rules.csv.

[32] See all results at http://lexbib.org/larramendi, including detailed merged subsets of all discussed dictionaries, and access to Wikidata attestations.

[33] For this concept and a short overview on the topic, see e.g. https://en.wikipedia.org/wiki/Linked_data#Linked_open_data.

[34] See http://wikidata.org.

[35] See https://en.wikipedia.org/wiki/SPARQL.

lexemes are encoded as lexicographic items, so that a human user can discriminate them by structural design features, here we are in front of statements coded in machine-readable semantic triples. For example, the English noun "magic"[36] is furnished with statements about its attestation (with an OED online entry ID as URI for the reference), with word senses and translations that belong to a certain sense, and with a link to the (ontological) concept denoted by a one of the senses, which is shortly defined as "type of beliefs and practices involving supernatural acts", member of class "occult" and part of "Magic and Religion", which is further described in a Wikipedia article entitled "Magic (supernatural)". Another Wikipedia article, "magic", describes another Wikidata entity, member of the classes "circus skill" and "performing arts", and that is linked to a different sense of the same lexeme "magic".

Translation equivalence is expressed in two ways in and around Wikidata. On the one side, Wikidata items that correspond to word senses,[37] i.e. not lexical but ontological items, are multilingually labelled. On the other side, translation equivalence can be expressed between lexemes, and between senses of lexemes, using a set of properties and classes related to lexical data defined in Wikidata itself,[38] but also using the linked data vocabularies developed by the OntoLex-Lexica Community group inside W3C,[39] which is a collection of RDF models that is used also in Wikidata. In the following, we describe how to link the historical lexical data on hand to lexical data contained in Wikidata.

5.2 Linking attestations

As we have mentioned, Wikidata contains not only ontological concepts (entity URI starting with a "Q"), but also lexemes (URI starting with an "L"). Senses of Lexemes can be linked to the concepts they denote using Wikidata P5137 property ("item for this sense").

The Elhuyar Foundation, a major dictionary publisher in the Basque Country,[40] has recently shared the Basque lemmata contained in their Basque-Spanish bilingual dictionaries on Wikidata. In any case, we shall not propose creating new Wikidata lexemes for the (historical) Basque lexical forms extracted from LAR, but rather link them to existing lexemes, as attestation. This is not trivial, since we have to deal with historical spelling, as discussed in section 3, and with the part of speech, a property Wikidata lexemes are furnished with by default.

For a first iteration, we have chosen those lexical items identified as Basque by both the rule-based and the ML-based approach, that, at the same time, could be mapped to items present in Wikidata, and to items present in SAR, and that are marked in that dictionary with the attestation datum "1745". In other words, we chose those 1,179 items which are double-checked to appear in LAR by SAR dictionary. To be sure to avoid mistaken part-of-speech mappings, from those we chose the 1,131 items which do not appear with an ambiguous part of speech on Wikidata. Wikidata lexemes data model does not foresee

[36] See https://www.wikidata.org/wiki/Lexeme:L3.

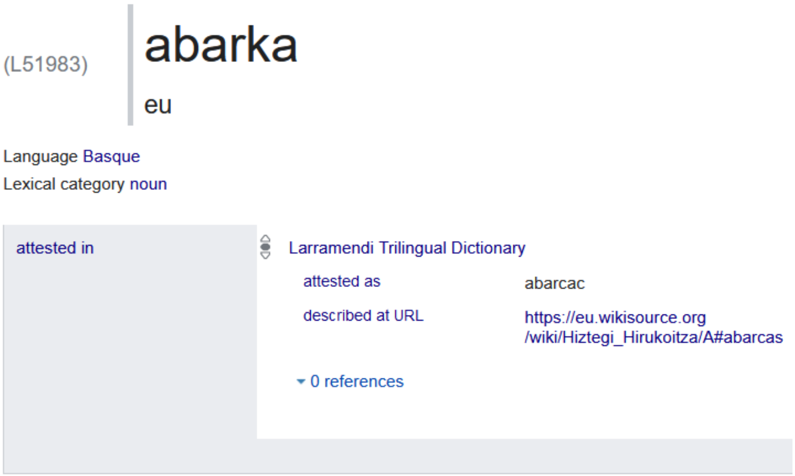
[37] For this link, property http://www.wikidata.org/entity/P5137 is used ("item for this sense").

[38] See https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation.

[39] See https://www.w3.org/community/ontolex/, and related publications (McCrae et al., 2017; Bosque-Gil et al., 2017).

[40] Elhuyar dictionary portal is accessible at https://hiztegiak.elhuyar.eus/.

more than one part of speech assigned to a lexeme, and Basque lexemes are represented according to that, so that lexemes with a different part of speech that share the same written representation (which certainly is not infrequent in Basque)[41] are represented as distinct lexemes. Since LAR does not contain part of speech data, and, on the other hand, lemma and equivalents in LAR often do not share the same part of speech, such disambiguation at the homograph level could not be carried out in this first iteration; most probably, manual work will be required here.[42]



(L51983) abarka

eu

Language Basque

Lexical category noun

attested in Larramendi Trilingual Dictionary

attested as abarcac

described at URL https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza/A#abarcas

▾ 0 references

Figure 11: Wikidata lexeme attestation

We have used Wikidata property P5323, "attested in",[43] for the attestation statement, together with P7855, "attested as",[44] and P973, "available at URL"[45] as qualifiers to that statement, that is, the claim that a lexeme is attested in LAR is further described, providing the attested written representation, and the reference to the corresponding dictionary entry, which is a hyperlink pointing to a headword anchor in the dictionary text on the Wikiteka platform (see section 2.4).[46] Since that text is aligned at page level with the facsimile image version, full reference to the attestation source is guaranteed.

6. Outlook

It was the purpose of our small study to run through the whole digitisation process for a historical dictionary, starting from scanned images, with this being one of the harder tasks to solve for texts of this age. In this paper, we have tried to make our workflow transparent. We have pointed out achievements and drawbacks encountered at each stage. Although the OCR process did not yield 100% precision, we have sent the output to the next stage in the pipeline, i.e. information extraction, which has also not brought error free results. Nevertheless, we believe that we have showed what automatic tools can do for us, and that the datasets we have been able to create with a very reduced manual validation effort already have something to offer to further research. Since we have used open software tools provided by the research community, and Wikimedia-related communities, this workflow is easily reproducible. For the near future, we propose to manually validate the ALTO XML content we have produced, using the Wikiteka platform, which allows this to be a

[41] Also in English this is not at all infrequent (cf. items like ’sound’, with three part of speech values assigned in dictionaries).

[42] Another option would be a lexical data model that foresees a part-of-speech assignation at a level lower than the lemma sign, i. e., either between lemma and sense, or inside the sense.

[43] See https://www.wikidata.org/wiki/Property:P5323.

[44] See https://www.wikidata.org/wiki/Property:P7855.

[45] See https://www.wikidata.org/wiki/Property:P973.

[46] See the statements shown in Fig. 11 online at http://www.wikidata.org/entity/L51983.

community-driven effort. Based on the tracked working time spent on transcription, we estimate an average of 15 minutes for correcting a dictionary column's transcription, that is, around 425 working hours for producing a ground truth transcription of the whole dictionary.

We then propose to take actions for improving precision in information extraction. That is, to annotate a larger training set for the Elexifier tool, and to make use of a more complex tag set. That would also mean annotating microstructural items other than translation equivalents, such as examples and cross-references.

We finally want to further develop the proposed model for integration in Wikidata. We are currently discussing the possibility to use an own instance of Wikibase,[47] i.e. the software solution that drives Wikidata, as a separate ecosystem for the development of linked (Basque) lexical datasets. Such a parallel resource would serve as infrastructure for collaborative research on converting plain dictionary text into structured datasets, its integration with other kinds of lexical resources, its representation as Linguistic Linked Data, and ultimately, regarding sufficiently validated lexical data, its transfer to the main Wikidata platform.

7. Acknowledgements

In this paper, we present methodology and findings from an eight months project started in January 2020, funded by Wikimedia Basque Country and UEU Udako Euskal Unibertsitatea. Our research has also received funding from the Basque Government Eusko Jaurlaritza (IT 1344-19).

8. References

Alonso Arrospide, M. (2021). *Larramendiren Hiztegi Hirukoitzaren digitalizazioa.* Master's thesis, UPV/EHU University of the Basque Country, Vitoria-Gasteiz. URL http://doi.org/10.13140/RG.2.2.27926.68169.

Bański, P., Bowers, J. & Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017.* Brno: Lexical Computing CZ s.r.o., pp. 485–494. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper29.pdf.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017).* Galway/Gaillimh, pp. 74–84. URL http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf.

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., Khemakhem, M. & Romary, L. (2018). Retro-Digitizing and Automatically Structuring a Large Bibliography Collection. In *European Association for Digital Humanities (EADH) Conference.* Galway/Gaillimh, Ireland: National University of Ireland. URL https://hal.archives-ouvertes.fr/hal-01941534/.

[47] See https://wikiba.se/.

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, pp. 79–96. URL http://www.ehu.eus/ehg/salaburu/liburua/HitzakSarean06.pdf.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*. Leiden: Lexical Computing, pp. 587–597. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

Mitxelena, K. & Sarasola, I. (1988). *Diccionario general vasco - Orotariko euskal hiztegia*. Euskaltzaindia; Editorial Desclée de Brouwer.

Porta, J., Sancho, J.L. & Gómez, J. (2013). Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*. Linköping University Electronic Press, pp. 70–79.

Real Academia Española (ed.) (2013). *Diccionario de autoridades: 1726 - 1739*. Boadilla del Monte (Madrid): JdeJ Editores, ed. facs. con motivo del iii centenario edition.

Romanov, M., Miller, M.T., Savant, S.B. & Kiessling, B. (2017). Important New Developments in Arabographic Optical Character Recognition (OCR). *arXiv:1703.09550 [cs]*. URL http://arxiv.org/abs/1703.09550.

Sarasola, I. (1996). *Euskal Hiztegia*. Donostia: Kutxa Gizarte-eta Kultur Fundazioa.

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), pp. 109–163. URL https://www.ehu.eus/ojs/index.php/ASJU/article/view/8709.

Urgell, B. (1998b). "Hiztegi Hirukoitza" eta "Diccionario de Autoridades" erkatuaz (II): sarreraren edukia. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 32(2), pp. 365–414. URL https://www.ehu.eus/ojs/index.php/ASJU/article/view/8717.

Urgell, B. (2002). *Euskal Lexikografia. Irakaskuntza proiektua*. UPV/EHU. URL https://www.academia.edu/3481533/Euskal_Lexikografia._Irakaskuntza_proiektua.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/