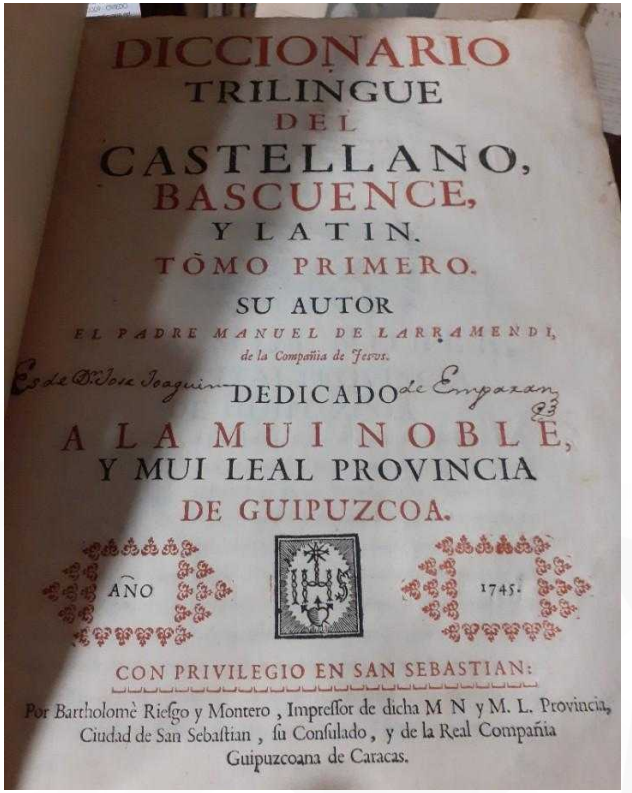
A workflow for historical dictionary digitization: Larramendi's Trilingual Dictionary

David Lindemann *david.lindemann@ehu.eus*
Mikel Alonso *mikelalon@gmail.com*

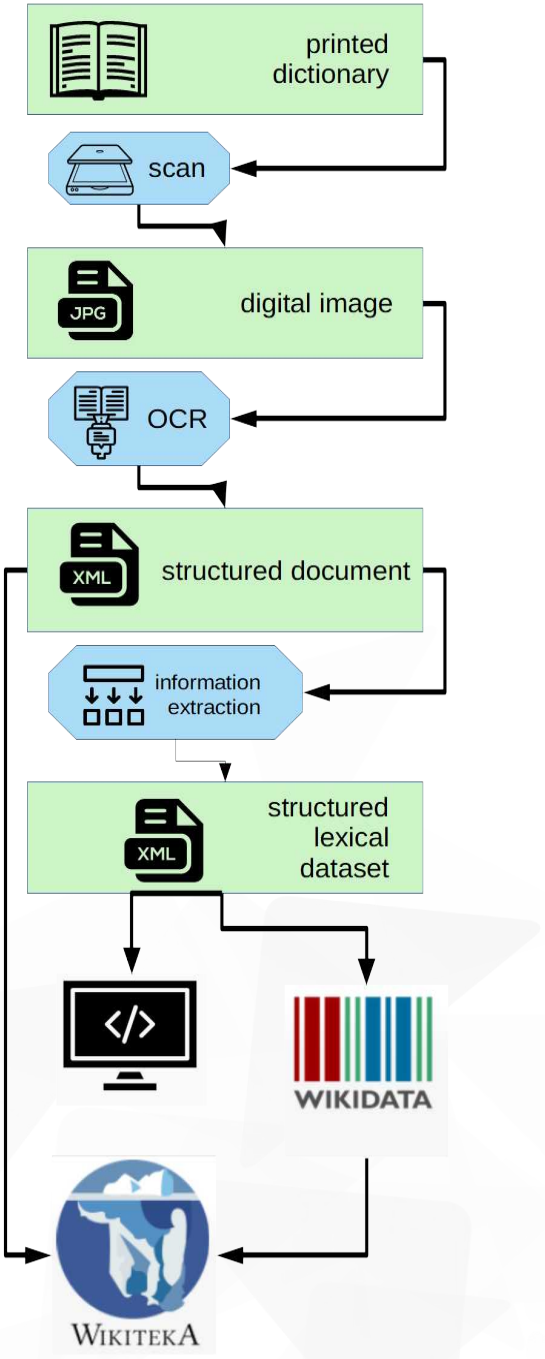
Introduction

- Larramendi's *Diccionario Trilingüe del Castellano, Bascuence y Latín* (LAR, 1745)
  - In Basque, *Hiztegi Hirukoitza* (the 'triple dictionary')
  - First published general Basque dictionary.
  - Brings significant shift in Basque Lexicography periodization (Urgell 2002).
  - During a century and a half main reference work for Basque
  - Subject to in-depth philological analysis (Urgell 1998a; 1998b, among others).
  - Until today only available on paper, and as collection of scanned images.
- UEU/Wikimedia Basque Country short-term research grant (2020)
  - Workflow design using Wikimedia platforms and freely available tools


DICCIONARIO
TRILINGUE
DEL
CASTELLANO,
BASCUENCE,
Y LATIN.
TOMO PRIMERO.
SU AUTOR
EL PADRE MANUEL DE LARRAMENDI,
de la Compañia de Jesvs.
DEDICADO
A LA MUI NOBLE,
Y MUI LEAL PROVINCIA
DE GUIPUZCOA.
AÑO
1745.
CON PRIVILEGIO EN SAN SEBASTIAN:
Por Bartholomè Riesgo y Montero, Impressor de dicha M N y M. L. Provincia, Ciudad de San Sebastian, su Consulado, y de la Real Compañia Guipuzcoana de Caracas.

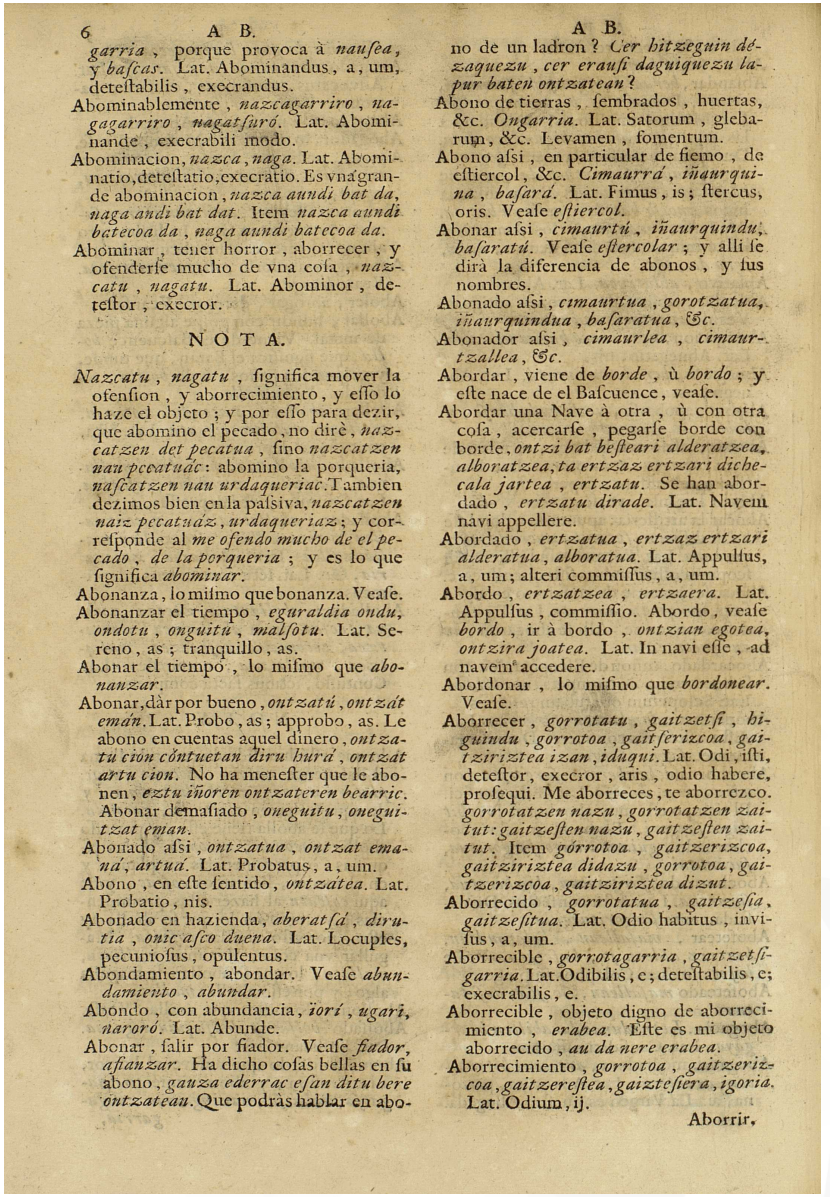
Goals

- Propose a workflow for historical dictionary digitization using Wikimedia platforms and free tools
- Predict workload...
  - for training set compilation (OCR, information extraction)
  - for result validation (covering the whole dictionary)
- Publish LAR contents on Wikisource and Wikidata platforms



Larramendi dictionary: Short overview

- Two volumes
- Dictionary body (A-Z), 828 pages.
- Entries: 40-45,000 (Urgell 1998a)
- Entry layout: First line with negative indent
- Alphabet order is not thoroughly respected (order follows first two letters; with exceptions)
- Subentries, and sometimes, additional notes, share main entry layout



6 A B.

garria, porque provoca à *nauſea*, y *baſcas*. Lat. Abominandus, a, um, deteſtabilis, execrandus.
Abominablemente, *nazcagarriro*, *nagagarriro*, *nagatſuró*. Lat. Abominande, execrabili modo.
Abominacion, *nazca*, *naga*. Lat. Abominatio, deteſtatio, execratio. Es vna grande abominacion, *nazca aundi bat da*, *naga andi bat dat*. Item *nazca aundi batecoa da*, *naga aundi batecoa da*.
Abominar, tener horror, aborrecer, y ofenderſe mucho de vna coſa, *nazcatu*, *nagatu*. Lat. Abominor, deteſtor, execror.

NOTA.

Nazcatu, *nagatu*, ſignifica mover la ofenſion, y aborrecimiento, y eſſo lo haze el objeto; y por eſſo para dezir, que abomino el pecado, no dirè, *nazcatzen det pecatua*, ſino *nazcatzen nau pecatuác*: abomino la porqueria, *naſcatzen nau urdaqueriac*. Tambien dezimos bien en la paſsiva, *nazcatzen naiz pecatuáz*, *urdaqueriaz*; y correſponde al *me ofendo mucho de el pecado*, *de la porqueria*; y es lo que ſignifica *abominar*.
Abonanza, lo miſmo que bonanza. Veaſe.
Abonanzar el tiempo, *eguraldia ondu*, *ondotu*, *onguitu*, *malſotu*. Lat. Sereno, as; tranquillo, as.
Abonar el tiempo, lo miſmo que *abonanzar*.
Abonar, dàr por bueno, *ontzatú*, *ontzát emán*. Lat. Probo, as; approbo, as. Le abono en cuentas aquel dinero, *ontzatu cion cóntuetan diru hurá*, *ontzat artu cion*. No ha meneſter que le abonen, *eztu iñoren ontzateren bearric*. Abonar demaſiado, *oneguitu*, *oneguitzat éman*.
Abonado aſsi, *ontzatua*, *ontzat emaná*, *artuá*. Lat. Probatus, a, um.
Abono, en eſte ſentido, *ontzatea*. Lat. Probatio, nis.
Abonado en hazienda, *aberatſá*, *dirutia*, *onic aſco duena*. Lat. Locuples, pecunioſus, opulentus.
Abondamiento, abondar. Veaſe *abundamiento*, *abundar*.
Abondo, con abundancia, *ioró*, *ugari*, *naroró*. Lat. Abunde.
Abonar, ſalir por fiador. Veaſe *fiador*, *afianzar*. Ha dicho coſas bellas en ſu abono, *gauza ederrac eſan ditu bere ontzateau*. Que podràs hablar en abo-

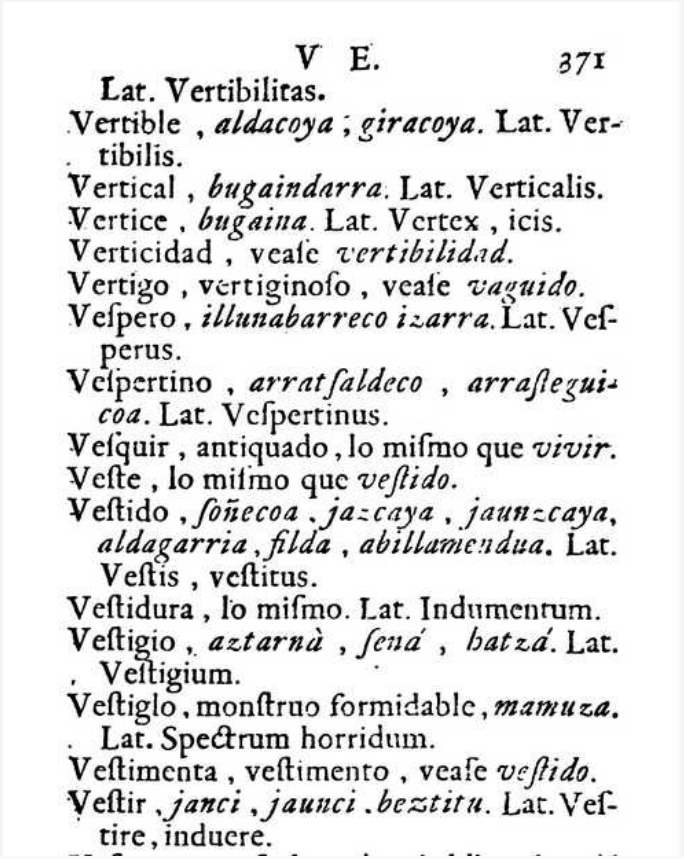
A B.

no de un ladron? *Cer hitzeguin dézaquezu*, *cer eraúſi daguiquezu lapur baten ontzatean*?
Abono de tierras, ſembrados, huertas, &c. *Ongarria*. Lat. Satorum, glebarum, &c. Levamen, fomentum.
Abono aſsi, en particular de ſiemo, de eſtiercol, &c. *Cimaurrá*, *iñaurquina*, *baſará*. Lat. Fimus, is; ſtercus, oris. Veaſe *eſtiercol*.
Abonar aſsi, *cimaurtú*, *iñaurquindu*, *baſaratú*. Veaſe *eſtercolar*; y alli ſe dirà la diferencia de abonos, y ſus nombres.
Abonado aſsi, *cimaurtua*, *gorotzatua*, *iñaurquindua*, *baſaratua*, &c.
Abonador aſsi, *cimaurlea*, *cimaurtzallea*, &c.
Abordar, viene de *borde*, ù *bordo*; y eſte nace de el Baſcuence, veaſe.
Abordar una Nave à otra, ù con otra coſa, acercarſe, pegarſe borde con borde, *ontzi bat beſteari alderatzea*, *alboratzea*, *ta ertzaz ertzari dichecala jartea*, *ertzatu*. Se han abordado, *ertzatu dirade*. Lat. Navem navi appellere.
Abordado, *ertzatua*, *ertzaz ertzari alderatua*, *alboratua*. Lat. Appulſus, a, um; alteri commiſſus, a, um.
Abordo, *ertzatzea*, *ertzaera*. Lat. Appulſus, commiſſio. Abordo, veaſe *bordo*, ir à bordo, *ontzian egotea*, *ontzira joatea*. Lat. In navi eſſe, ad navem accedere.
Abordonar, lo miſmo que *bordonear*. Veaſe.
Aborrecer, *gorrotatu*, *gaitzetſi*, *higuindu*, *gorrotoa*, *gaitſerizcoa*, *gaitziriztea izan*, *iduqui*. Lat. Odi, iſti, deteſtor, execror, aris, odio habere, proſequi. Me aborreces, te aborrezco. *gorrotatzen nazu*, *gorrotatzen zaitut*: *gaitzeſten nazu*, *gaitzeſten zaitut*. Item *gorrotoa*, *gaitzerizcoa*, *gaitziriztea didazu*, *gorrotoa*, *gaitzerizcoa*, *gaitziriztea dizut*.
Aborrecido, *gorrotatua*, *gaitzeſia*, *gaitzeſitua*. Lat. Odio habitus, inviſus, a, um.
Aborrecible, *gorrotagarria*, *gaitzetſigarria*. Lat. Odibilis, e; deteſtabilis, e; execrabilis, e.
Aborrecible, objeto digno de aborrecimiento, *erabea*. Eſte es mi objeto aborrecido, *au da nere erabea*.
Aborrecimiento, *gorrotoa*, *gaitzerizcoa*, *gaitzereſtea*, *gaizteſiera*, *igoria*. Lat. Odium, ij.

Aborrir.

Optical Character Recognition (OCR)

- Kraken
  - machine-learning based
  - freely available
  - deals with font style recognition
  - result export in ALTO XML format
- Steps
  - image pre-processing
  - transcription of a training set
  - application of the trained model to the whole content



V E. 371

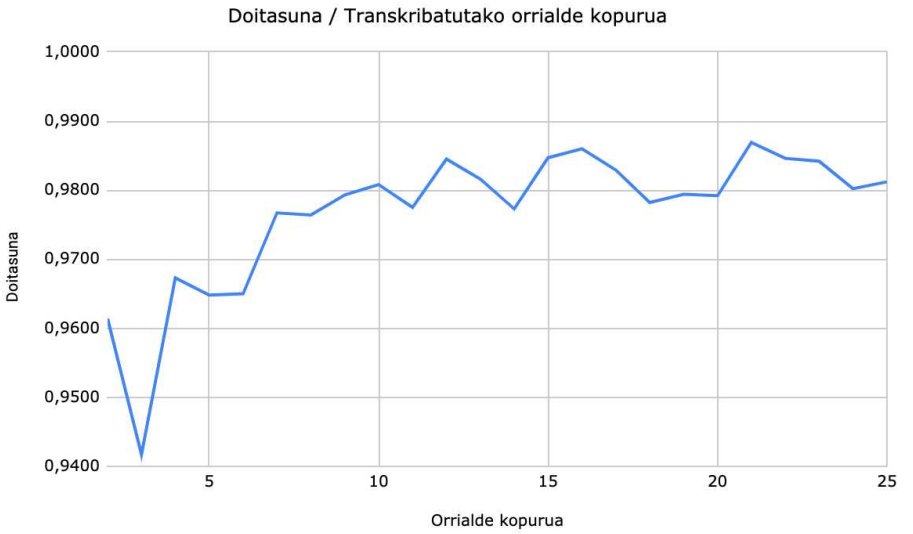
Lat. Vertibilitas.

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* , *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* , *ſenà* , *hatzà*. 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. Vertibilis.
Vertical , @bugaindarra. Lat. Verticalis.
Vertice , @bugaina. Lat. Vertex , icis.
Verticidad , veaſe @vertibilidad.
Vertigo , vertiginoſo , veaſe @vaguido.
Veſpero , @illunabarreco @izarra. Lat. Veſperus.
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 , @ſenà , @hatzà. Lat.
Veſtigium.

Optical Character Recognition (OCR)

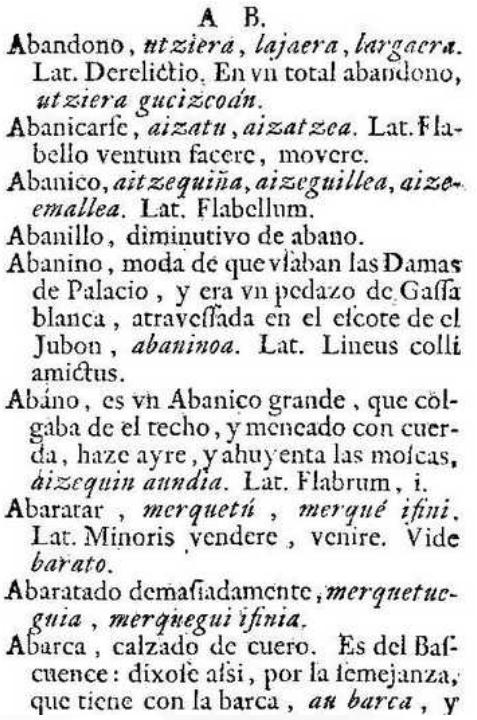
- Model training: Inclusion of more transcribed dictionary columns into the training set, until desired precision rate reached
  - After 25 pages (i.e. 50 columns), we estimated that precision would not significantly grow
  - Infrequent characters not properly recognized: Inclusion of additional 14 columns (where these appear)
  - With a semi-automatic transcription of less than 4% of the content, OCRed version of the whole dictionary with 98,5% precision
  - Significantly better than TXT versions available before



ALTO to Wikitext

- ALTO XML format produced by Kraken:
  - Represents characters, font style, and text box position on page.
  - The latter is used for a rule-based approach of segmentation of dictionary text into entries (indented lines). We allow only indented lines starting with capital letter of the corresponding alphabet section.
- ALTO to Wikitext:
  - Preserves font style
  - Line indent represented by colon(s)
  - Anchor template for headwords (in draft: first word of negative indented line starting with correct letter)

```
:::::A B.
::bandono, ''utziera'', ''lajaera'', ''lagaera''.
::Lat. Derelictio. Envn total abandono,
::''utziera'' ''gucizcoán''.
:{{sarrera|abanicarse}}Abanicarfe, ''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.

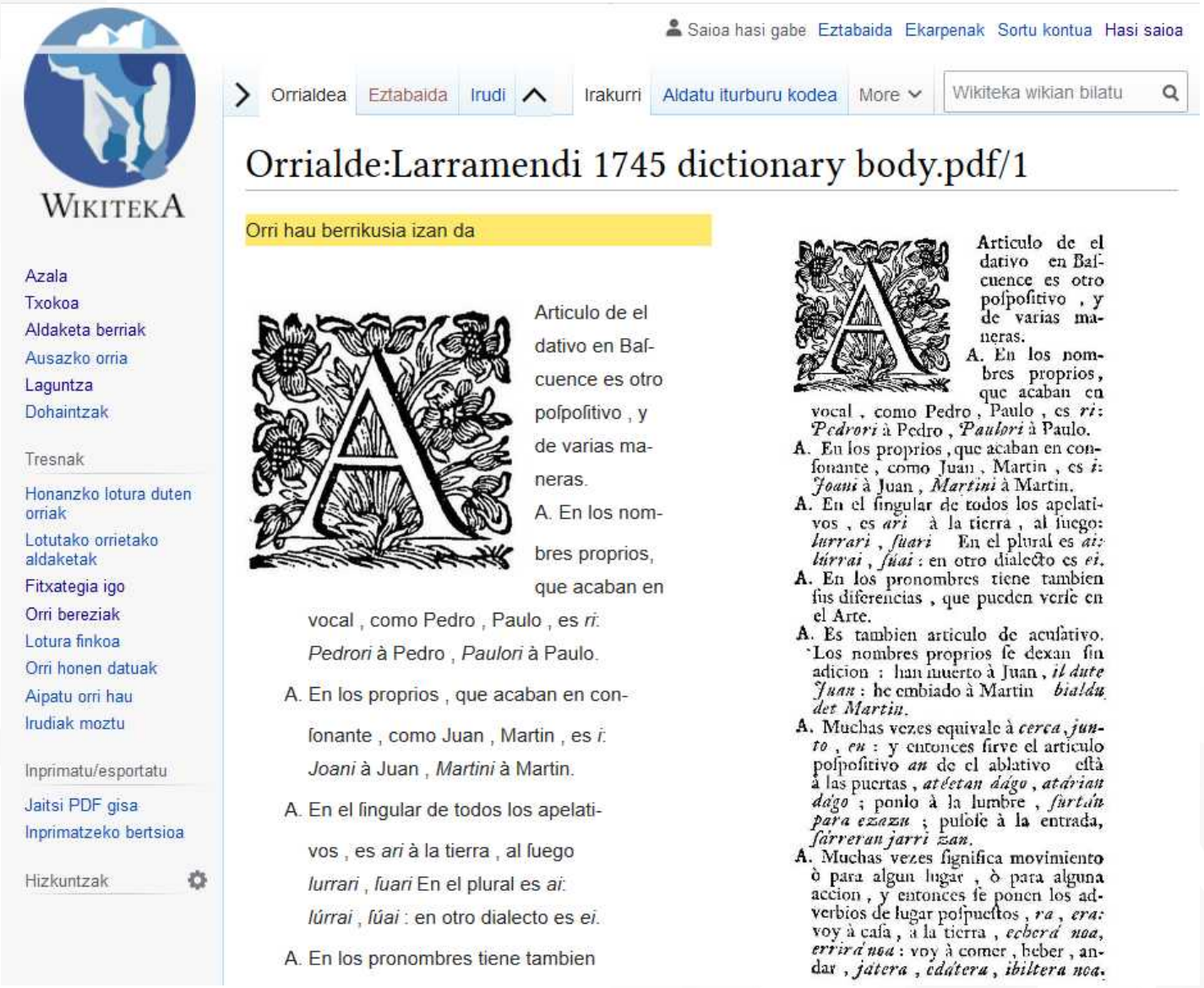
Abaratar, *merquetú*, *merqué ifini*. Lat. Minoris vendere, venire. Vide *barato*.

Abaratado demaſiadamente, *merquetuequia*, *merquequi ifinia*.

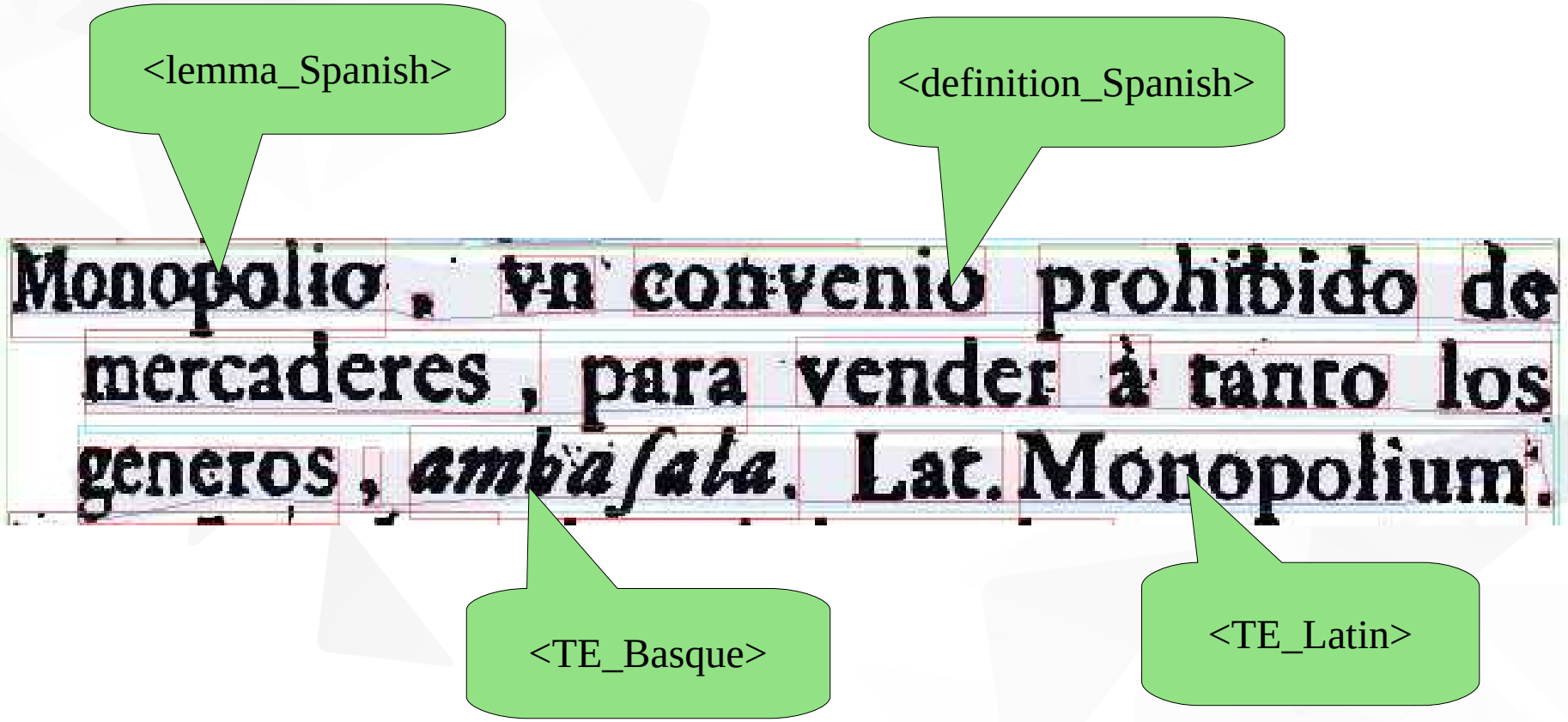
Abarca, calzado de cuero. Es del Baſcuence: dixoſe aſsi, por la ſemejanza, que tiene con la barca, *an barca*, y

Wikiteka (Basque Wikisource platform)

- Collaborative validation of OCR results
- Active community
- Two-step validation

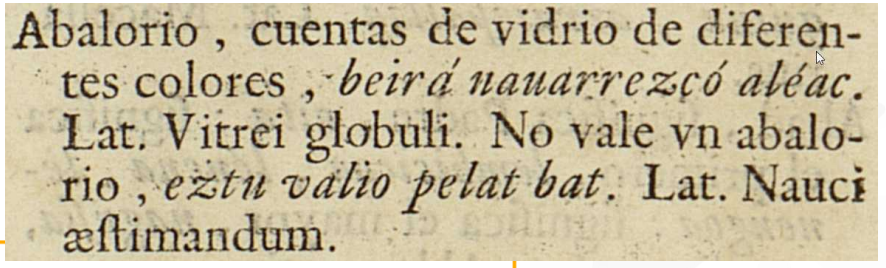


Information extraction



Information extraction

- ELEXIFIER is a ML-based tool for information extraction from dictionary plain text. As Elexis observer institution, we take part in Elexifier beta evaluation, as early adopters.
  - TEI-Lex0, with (at the time of preparing the dataset) a still limited tagset
  - Training set: Manual annotation of entry structure on 10 pages.
  - Annotated microstructure elements: headword, Basque equivlent, latin equivalent, cross-reference, definition (Spanish), "cit" element for other content


Abalorio , cuentas de vidrio de diferen-
tes colores , *beirá nauarrezçó aléac.*
Lat. Vitrei globuli. No vale vn abalo-
rio , *eztu valio pelat bat.* Lat. Nauci
æſtimandum.

entry

headword: Abalorio , def: cuentas de vidrio de diferen- tes colores , translation: *beirá nauarrezçó aléac* . Lat . sense: Vitrei globuli .

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

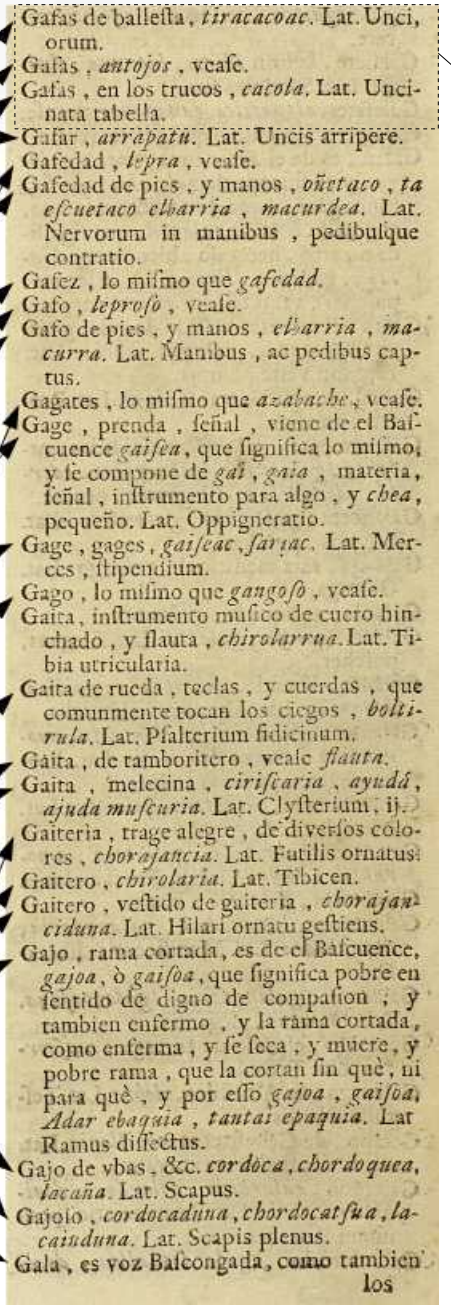
Information extraction

- ELEXIFIER results:
  - Spanish headwords identified with very high precision.
  - Basque equivalents also identified with high precision, but many false positives (other items in italics font style also taken as Basque equivalents, such as cross-references)
  - Latin equivalents identified with low precision.

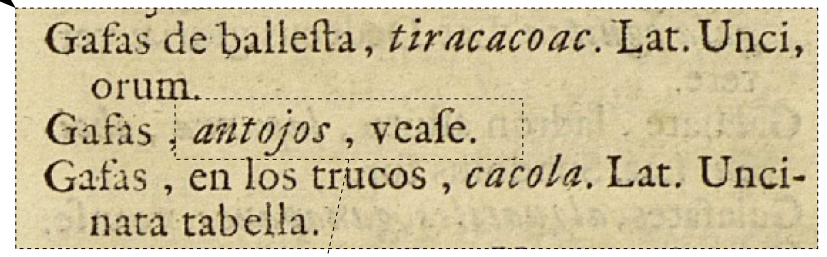
- Strategy for improving results:
  - Use a fully validated transcription (once wikisource community has finished).
  - Use a full TEI-Lex0 tagset
  - Manually annotate a larger training set.

DA / LAR (+ OEH)

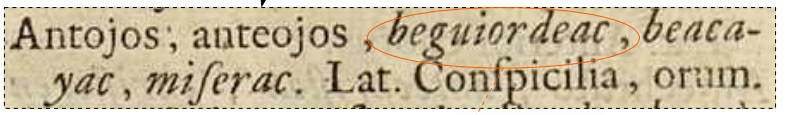
21693 gafar
21694 gafas
21695 gafedad
21696 gafete
21697 gafez
21698 gafo
21699 gagates
21700 gage
21701 gages
21702 gago
21703 gaita
21704 gaiteria
21705 gaitero
21706 gajo
21707 gajoso
21708 gala



Gafas de balleſta, *tiracacoac*. Lat. Unci, orum.
Gafas, *antojos*, veaſe.
Gafas, en los trucos, *cacola*. Lat. Uncinata tabella.
Gafar, *arrapatu*. Lat. Uncis arripere.
Gafedad, *lepra*, veaſe.
Gafedad de pies, y manos, *oñetaco, ta eſcuetaco elbarria, macurdea*. Lat. Nervorum in manibus, pedibulque contratio.
Gafez, lo miſmo que *gafedad*.
Gafo, *leproſo*, veaſe.
Gafo de pies, y manos, *elbarria, macurra*. Lat. Manibus, ac pedibus captus.
Gagates, lo miſmo que *azabache*, veaſe.
Gage, prenda, ſeñal, viene de el Baſcuence *gaiſea*, que ſignifica lo miſmo, y ſe compone de *gal, gaia*, materia, ſeñal, inſtrumento para algo, y *chea*, pequeño. Lat. Oppigneratio.
Gage, gages, *gaiſeac, ſariac*. Lat. Merces, ſtipendium.
Gago, lo miſmo que *gangoſo*, veaſe.
Gaita, inſtrumento muſico de cuero hinchado, y flauta, *chirolarrua*. Lat. Tibia utricularia.
Gaita de rueda, teclas, y cuerdas, que comunmente tocan los ciegos, *boltirula*. Lat. Pſalterium fidicinum.
Gaita, de tamboritero, veaſe *flauta*.
Gaita, melecina, *ciriſcaria, ayudá, ajuda muſcuria*. Lat. Clyſterium, ij.
Gaiteria, trage alegre, de diverſos colores, *chorajancia*. Lat. Futilis ornatus.
Gaitero, *chirolaria*. Lat. Tibicen.
Gaitero, veſtido de gaiteria, *chorajanciduna*. Lat. Hilari ornatu geſtiens.
Gajo, rama cortada, es de el Baſcuence, *gajoa*, ò *gaiſoa*, que ſignifica pobre en ſentido de digno de compaſion; y tambien enfermo, y la rama cortada, como enferma, y ſe ſeca, y muere, y pobre rama, que la cortan ſin què, ni para què, y por eſſo *gajoa, gaiſoa, Adar ebaquia, tantai epaquia*. Lat Ramus diſſectus.
Gajo de vbas, &c. *cordoca, chordoquea, lacaña*. Lat. Scapus.
Gajoſo, *cordocaduna, chordocatſua, lacainduna*. Lat. Scapis plenus.
Gala, es voz Baſcongada, como tambien los



Gafas de balleſta, *tiracacoac*. Lat. Unci, orum.
Gafas, *antojos*, veaſe.
Gafas, en los trucos, *cacola*. Lat. Uncinata tabella.



Antojos, anteojos, *beguiordeac, beacayac, miſerac*. Lat. Conſpicilia, orum.

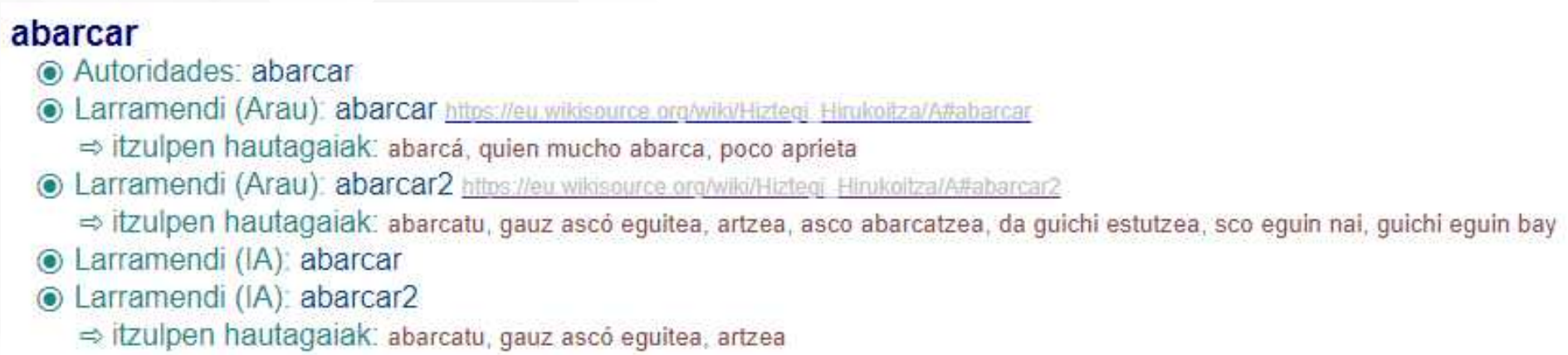
BEGI-ORDE (L ap. A; SP, Lar, Hb, H). "Lunettes, biserak (Ax)" SP. "**Antojos, anteojos, begiordeak, beakaiak, miserak**" **Lar.** "Begietakoak ou begiordeak (SP), lunettes" Dv. "Begi ordeak, lunettes; id. begi-orde thailluak (Ax)" H.

Mapping historical to modern standard spelling (ES, EU)

LAR	*DA*	*matching normalized lemma-sign*
Obsession	obsessión	obsesion
Hueſſo	huesso	hueso
Occiſſion	occisión	occision
Atràs	atras	atras

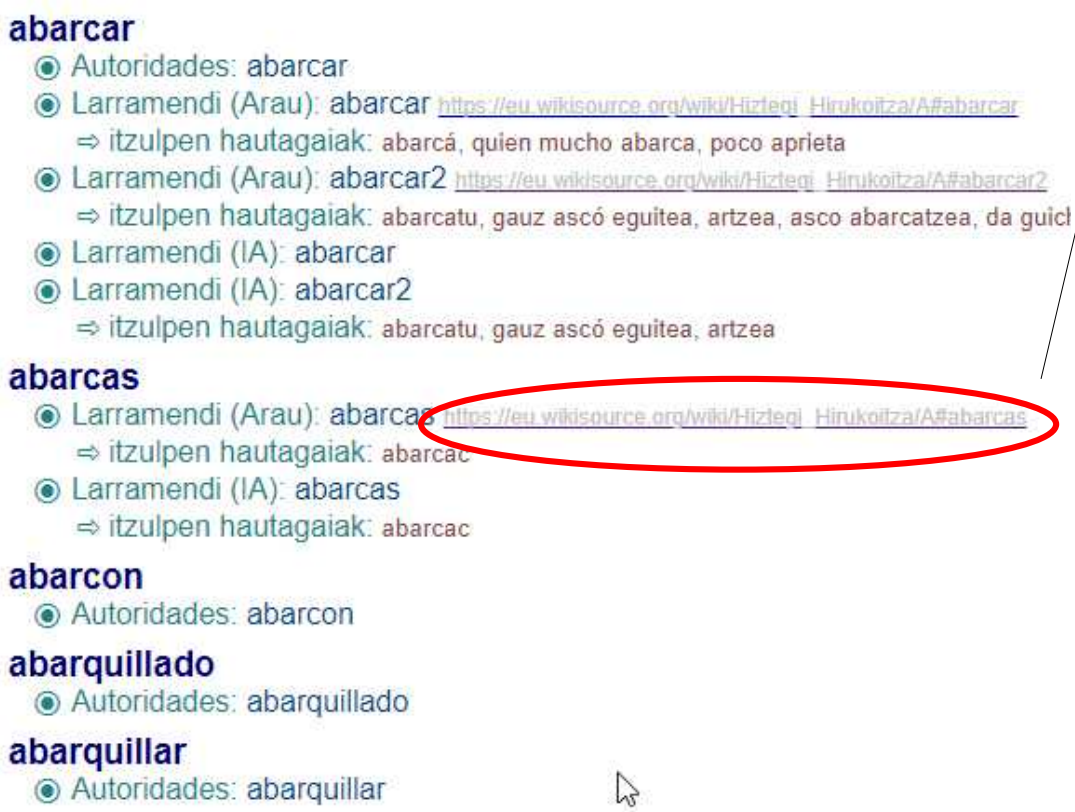
- Spanish
  - Remove diacritics, historical 'long s' and double 's' > 's'
- Basque
  - Sequence of regex operations, discussed in Alonso (2021)
  - examples:
    - *beguiordeac* > *begiordeak* > *begiorde*
    - *astigarra* > *astigar*
    - *astotua* > *astotu*

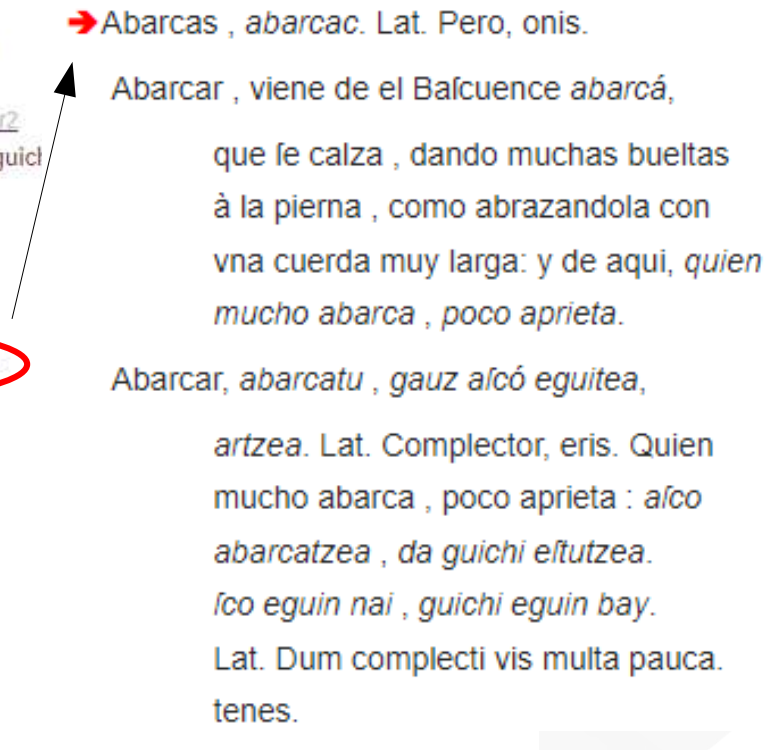
Index of LAR: ES ⇨ EU



- DA (Diccionario de Autoridades) lemmata:
  - Out of 41,968, 27,825 found in LAR; 4,875 LAR lemmata not found in DA
- LAR lemmata and Basque equivalent candidates:
  - Rule-based extraction:
    - 30,045 unique equivalent candidates for 32,044 Spanish headwords
  - Machine-learning-based extraction:
    - 27,125 unique equivalent candidates for 29,932 Spanish headwords

Index of LAR: ES ⇨ EU





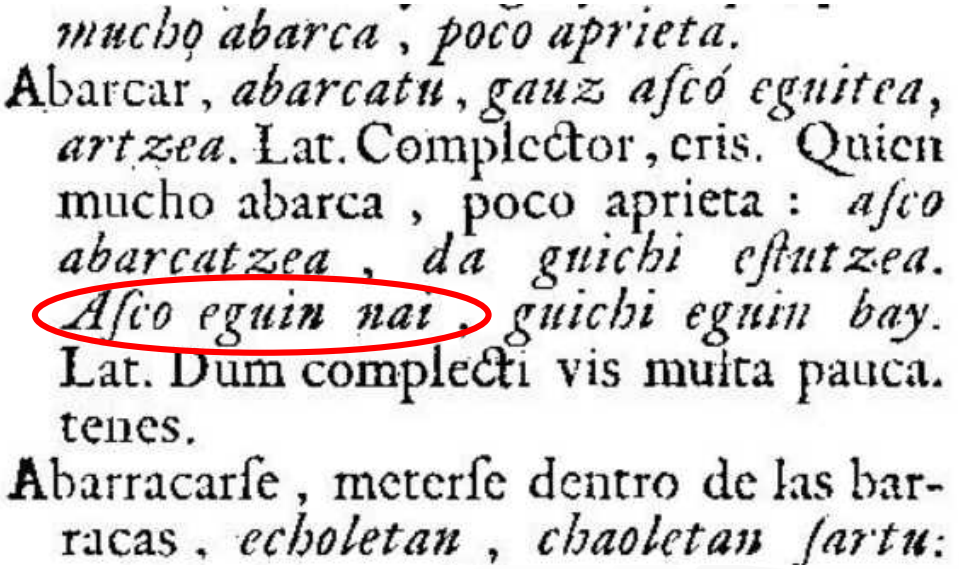
- Rule-based headword extraction: Headwords connected to wikisource navigation anchors

Index of LAR: ES ⇨ EU

➔Abarcas , *abarcac*. Lat. Pero, onis.

Abarcar , viene de el Baſcuence *abarcá*, que ſe calza , dando muchas bueltas à la pierna , como abrazandola con vna cuerda muy larga: y de aqui, *quien mucho abarca , poco aprieta.*

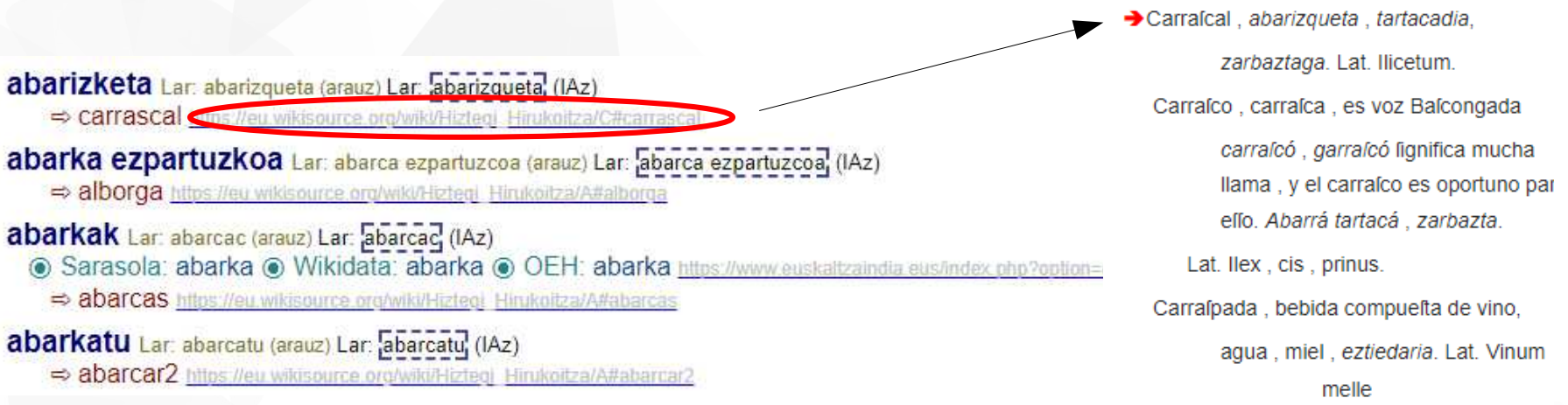
Abarcar, *abarcatu , gauz aſcó eguitea, artzea*. Lat. Complector, eris. Quien mucho abarca , poco aprieta : *aſco abarcatzea , da guichi eſtutzea. ſco eguin nai , guichi eguin bay.* Lat. Dum complecti vis multa pauca. tenes.



mucho abarca , poco aprieta.
Abarcar , *abarcatu , gauz aſcó eguitea, artzea*. Lat. Complector , eris. Quien mucho abarca , poco aprieta : *aſco abarcatzea , da guichi eſtutzea. Aſco eguin nai , guichi eguin bay.* Lat. Dum complecti vis multa pauca. tenes.
Abarracarſe , meterſe dentro de las barracas , *echoletan , chaoletan ſartu:*

- If a transcription error is found, it can be directly corrected (logged into wikisource)

Index of LAR: EU ⇨ ES



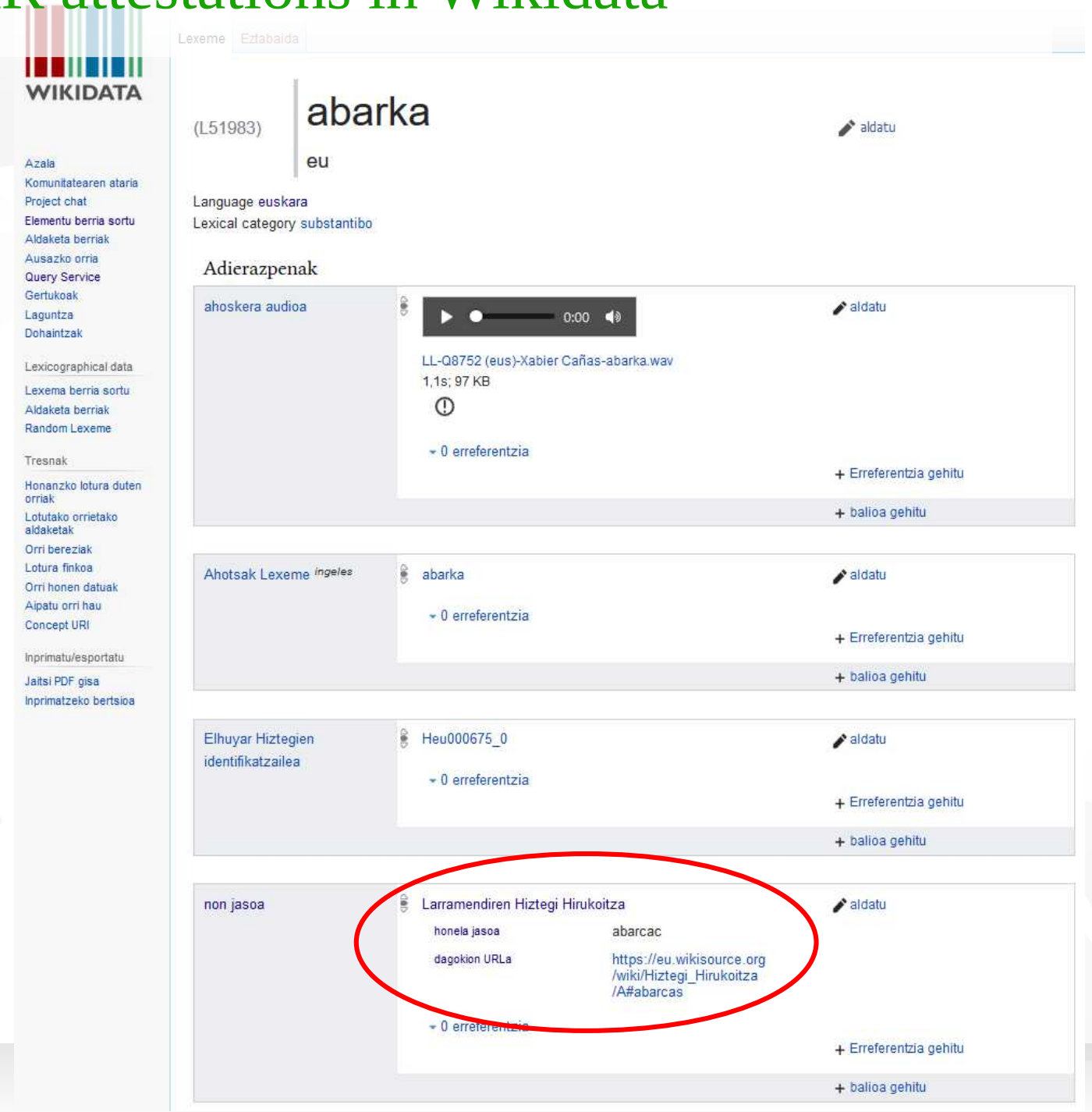
- Reversed dictionary Basque to Spanish:
  - Spanish equivalents linked to wikisource anchors
- Mapping of Basque items to modern Basque dictionaries:
  - ML extracted Basque items:
    - 38,300 candidates, 11,551 found in modern dictionaries
  - Rule-based approach:
    - 60,193 candidates, 15,152 found in modern dictionaries

LAR attestations in Wikidata

- "attested in" property, "attested as" and "described at URL" qualifiers
- Wikidata Basque lexemes: linked to senses, forms, and other lexical resources
- To start with:

  Both methods agree, and Sarasola cites as LAR item

  (1,179 lemmata)
- POS mapping
- Problems with letter H, and hyphenization
- To be continued...



Links

- LAR at Wikidata: http://www.wikidata.org/entity/Q65216433
- LAR at Wikisource: https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza
- Transcription guidelines: https://eu.wikisource.org/wiki/Laguntza:Orrialdeen_egoera
  https://en.wikisource.org/wiki/Help:Page_status
- Produced dictionaries: http://lexbib.org/larramendi
- Code: https://github.com/dlindem/LBLR/tree/master/Larramendi

Eskerrik asko / Thank you!

Funding acknowledgements:
UEU/EWKE Digital Humanities Grant (2020),
Basque Government (IT 1344-19)

David Lindemann *david.lindemann@ehu.eus*
Mikel Alonso *mikelalon@gmail.com*