Basque in the Making: A Historical Look at a Language Isolate

2017-2021 UMR IKER 5478

Basque: a few elementary facts

- The Basque language, named by its speakers *euskara*, *euskera*, *uskara*, *eskuara*, *üskara*, is spoken by about 800.000 personnes today.
- The language is historically spoken in both sides of the Pyrenees, in the continental provinces of Labourd, Low-Navarre and Soule, and the peninsular provinces of Navarre, Gipuzkoa, Bizkaia and Araba. The area in which the language has been historically spoken is called *Euskal Herria* or country of the Basque language.

Current state of research in Basque historical linguistics

- Perhaps due to its particular status as a non-indoeuropean language in Europe and as a language isolate, most of the work devoted to the history of the Basque language have been devoted to issues related to descent.
- Objective: reconstructing earlier stages of the language (prehistoric or protohistorical ones).
- Exceptions: Mounole (2012), Manterola (2015), Padilla (2017), Krajewska (2018)

Aims of the project

- This project is oriented towards the examination of linguistic evolutions which are attested by the textual corpus. It would like to examine changes in the domains of:
- 1. Word order and its relation to discourse structure
- 2. Dependency relations in grammar (case, agreement, auxiliation)
- 3. Quantification, Determination, Number
- 4. The reference tracking system (anaphoric and pronominal systems)

The existing resources: Euskal Klasikoen Korpusa

Euskal Klasikoen Corpusa (EKC)

2005ean abiatu zuen armiarma.com-ek Klasikoen Gordailua, XX. mendea bitarteko testu klasiko ia guztien bilgunea bilakatzeko asmoz. 2007an sortu zen haren barnean Corpus Arakatzailea. Orain, corpus horretako edukiak Euskara Institutuaren webgunera ekarri nahi izan ditugu, ohi dugun kontsulta modura ekarriz.

Corpus honek XVI. mendean hasi eta 1975. urtera arteko **496 liburu** jasotzen ditu, eta denera **11,9 milioi testu hitzez** osatuta dago. Horren aurretik zen

Problems

- You can already see, from the very general issues we would like to explore, that a simple word or form-based search will not take us too far or at least not far enough.
- This is because grammatical phenomena are on the one hand relational –they require searching across (related) forms in the text-, and on the other, paradigmatic, so they are open to other related forms not present in the text.

Words

- Word search is not an unproblematic concept in Basque historical grammar.
- Take the standard negative polarity item *inor* « anyone »
- The attested forms in the Basque General Dictionary are the following:
- *Inor, nehor, nihor, ehor, ihor, yor, ñor, ñeur, ehur, nihur, ihur, inhur, igor, nigor*
- And those are the absolutive forms. This is general for each and every lexical or functional item in Basque.

Paradigmatic Issues

Optional Dative Agreement:

(1) a. Liburu bat eman dako liburuzainari (General now)

book one given aux.3sE-3sD-3sA book-keeper.dat

b. Liburu bat eman du liburuzainari (Unclear historical status)

book one given aux.3sE-3sA book-keeper.dat

« He/she gave a book to the book-keeper »

Paradigmatic Issues

- Predicates allowing optional dative agreement
  - The list of regularly alternating predicates covers verbs of giving, such as *eman* "give", *helarazi* "make-have" (cf. French *parvenir*), *eskeini* "offer", *agindu* "promise", verbs of sending/receiving, such as *eraman* "carry", *ekarri* "bring", *igorri* "send", *hedatu/zabaldu* "extend", *erosi* "buy" and *saldu* "sell", *barreatu* "spread/scatter", verbs of throwing, such as *bota* "throw"; verbs of fulfilling such as *arthamendatu* "entrust", verbs of communicated message, such as *erran* "say", *galdegin* "ask", *izkribatu* "write", *aiphatu* "mention", *kondatu* "tell". The list does not seem to be arbitrary. All the verbs consigned by **Gropen et al.** (1989) as entering the dative alternation in English show the agreement optionality above.

Paradigmatic Issues

- Cross-linguistic correlates of so-called dative alternation
  - The English dative alternation is arguably part of a wider set of phenomena involving different types of syntactic alternations in ditransitive constructions. Those may correlate also with presence/absence of clitic doubling : Spanish (Demonte, 1995; Cuervo, 2003); Romanian (Diaconescu, 2004), Bulgarian (Slavkov, 2008); Greek (Anagnostopoulou, 2003). Or with the presence/absence of an applicative construction (see Baker, 1988; Marantz, 1993; Pylkkanen, 2008).

Relational Issues: Word order

- A notion of « clausal edge » can be defined in Basque from the position of the finite verb. It is typically preceded by the lexical verb and the focus:

- (1) a. Jonek bi liburu leitu ditu
  - Jon.erg two books read has

  « Jon has read two books/TWO BOOKS » (Standard/Dialectal)

  b. Jonek leitu ditu bi liburu

  Jon.erg read has two books

  « JON read two books »(Standard/Dialectal)

  « Jon read two books » (Dialectal)

Relational Issues: Word order

But in fact outside the standard, things are more complicated for (1b).

(1) Jonek leitu ditu bi liburu

Jon.erg read has two books

« JON read two books »(Standard/Dialecta/Historicall)

« Jon read two books » (Dialectal/Historical)

- We will not annotate discourse roles, but if we can ask the search engine for sequences of grammatical categories, we will have saved an enormous quantity of work. We are interested in the relative order of grammatical categories, not words.

Other stuff: Constructional Notions

- Non-finite relatives (unclear historical status)
- (1) a. Atzo jin gizona
  Yesterday arrive.partc man.det
  « The man who arrived yesterday »

  b. Atzo **ikusi** **gizona**
  Yesterday see.partc man.det
  « The man who was seen yesterday »

  Cf. English *The singing man* « The man who is singing »

The objectives of the project

- To create a grammatically annotated database, that will allow for systematic searches at different levels of grammatical complexity.
- The database will be based on a corpus of about 750.000 words. It will include annotation over lexical categories, morphosyntactic categories and syntactic structures.
- The corpus at the source of the database includes the most representative texts issued or produced between the XVth and the mid XVIIIth century. This period corresponds to Archaic and Ancient Basque, according to the periodization proposed by Lakarra (1997).

Discarded Sources for the Database

Roman inscriptions (Gorrotxategi, 1984):

- **VM . ME . SA . HAR**FI NAR . HVN . GE
- SI . A . BI / SVN . HA . RI . F. LIO / ANN
- XXV . T . P. S . S
- Vmme Sahar fi(lius) / Narhungesi Abi / sunhari. filio / ann(orum) XXV t(itulum) p(osuit) s(umptu) s(uo). Cf. **Ume zahar** «old child »

Medieval Sources

- The Codex Calixtinus (XIIth Century)
- Si illos comedere uideres, canibus edentibus uel porcis eos computares. Sique illos loqui audires, canum atrancium memorares. Barbara enim lingua penitus habentur. Deum uocant *Urcia*, Dei genitricem *Andrea Maria*, panem *ogui*, uinum *ardum*, carnem *aragui*, piscem *araign*, domum *echea*, dominum domus *iaona*, dominam *andrea*, ecclesiam *elicera*, presbiterum *belaterra*, quod interpretatur pulcra terra, tricticum *gari*, aquam *uric*, regem *ereguia*, sanctum Iacobum *laona domne Iacue* ...

The Included Sources

- The first long texts in Basque correspond to a language stage that historical linguists have located in the XVth century (letter by Machin de Zalba, epic verse, Old Sayings)

First published book: Detchepare, *Linguae Vasconum Primitiae* (1545)

- Calvinist translation of the New Testament (1571)
- The corpus is continuous from the XVIth century on

Building the Database

- A four step process:
- 1. The corpus: the entire corpus at the origin of the database has to be checked and updated according to the latest philological work
- 2. The corpus has to be annotated for text normalization
- 3. It has to be annotated for the syntactic categories and structures that will allow search
- 4. A search interface must be developed

The Basic Corpus

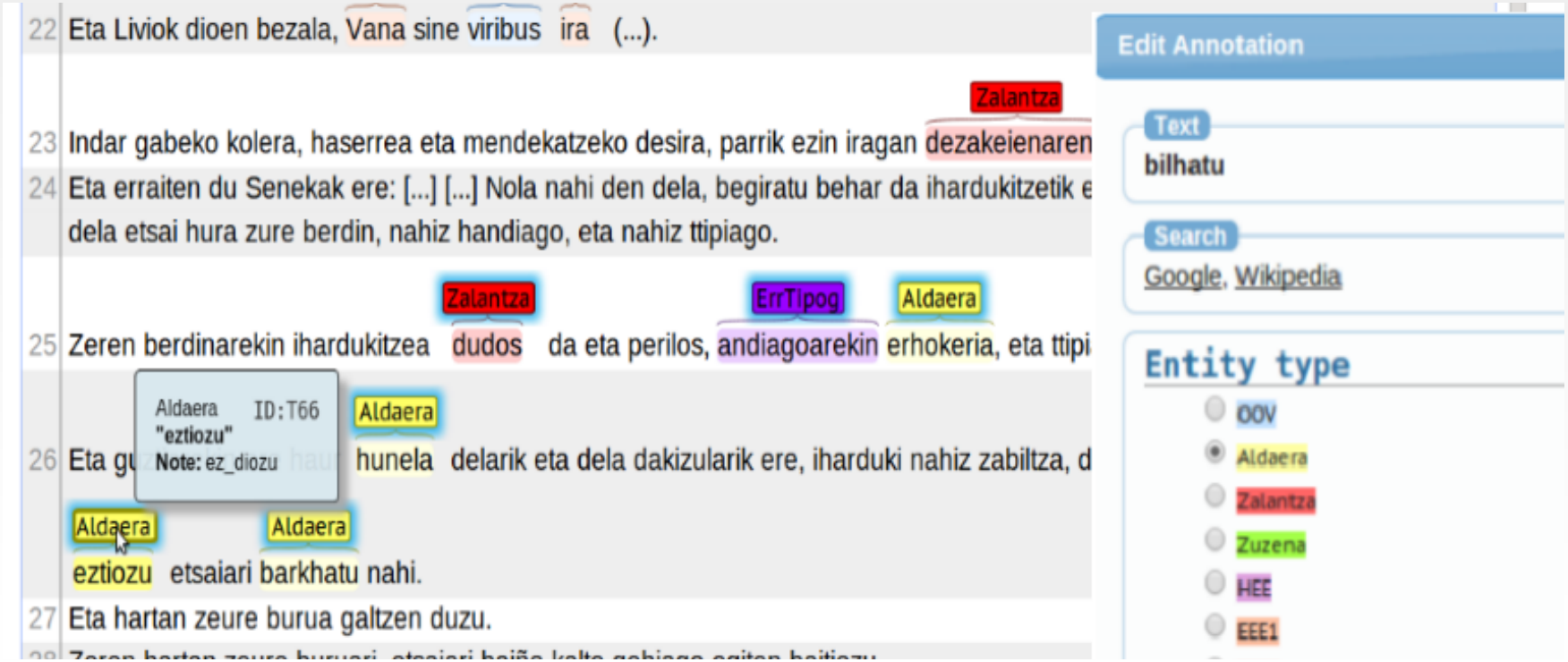
- The Institute of the Basque Language, a public institution of the University of the Basque Country, hosts a 11.9 million transcribed corpus in xml format, consisting of most of the literary production in Basque from the XVth century to the mid XXth century. This corpus, compiled during the last 30 years under the early impulsion of the editing house Susa and the *Repository of Basque Texts* initiative led by professor Patxi Salaberri, and augmented by the free contribution of different authors, computational linguists and philologists, is available under a creative commons licence. The most important part of the corpus, corresponding to the literary production, is transcribed and standardized in modern orthography (keeping the original morphological forms).

The Basic Corpus

- The basic corpus is the result of an amateur enterprise, spanning across a period of 30 years, and showing varying degree of coherence and reliability in the transcription and the texts that have been chosen.

- Example: the edition that has been chosen for the text *Acto para la Nochebuena* (Barrutia, first half of the XVIIIth century) is the one prepared by Juan Carlos Guerra in the beginning of the XXth century, not the latest one of the eighties (Lakarra and Knorr, 1981). Several mistakes in the comprehension of the text have been inherited from that edition.

Annotation


22 Eta Liviok dioen bezala, Vana sine viribus ira (...).
Zalantza
23 Indar gabeko kolera, haserrea eta mendekatzeko desira, parrik ezin iragan dezakeienaren
24 Eta erraiten du Senekak ere: [...] [...] Nola nahi den dela, begiratu behar da ihardukitzetik e
dela etsai hura zure berdin, nahiz handiago, eta nahiz ttipiago.
Zalantza
ErrTipog
Aldaera
25 Zeren berdinarekin ihardukitzea dudos da eta perilos, andiagoarekin erhokeria, eta ttipi
Aldaera ID:T66
"eztiozu"
Note: ez_diozu
Aldaera
26 Eta gu hunela delarik eta dela dakizularik ere, iharduki nahiz zabiltza, d
Aldaera
Aldaera
eztiozu etsaiari barkhatu nahi.
27 Eta hartan zeure burua galtzen duzu.
Edit Annotation
Text
bilhatu
Search
Google, Wikipedia
Entity type
OOV
Aldaera
Zalantza
Zuzena
HEE
EEE1

Issues in Annotation

- The « ask a linguist » part of the annotation involves some problematic decisions. For instance, something like:
- (1) *Eros zezan*

  buy aux
- The auxiliary could be a Past Tense in the XVIth century (« he/she bought it », but it is only a subjunctive form these days (« so that he/she may buy », ocurring in embedded sentences.

The Search Interface: Previous Models

- TSABL project:
  - Towards a Syntactic Atlas of the Basque Language 2007-2011
- IHAP project:
- Iparraldeko Hizkeren azterketa eta Prozesamendua 2011-2013 (Larraitz Uria)
- *Basyque* database: http://ixa2.si.ehu.es/atlas2/index.php?lang=en

Basyque

Basque syntactic database

Log in

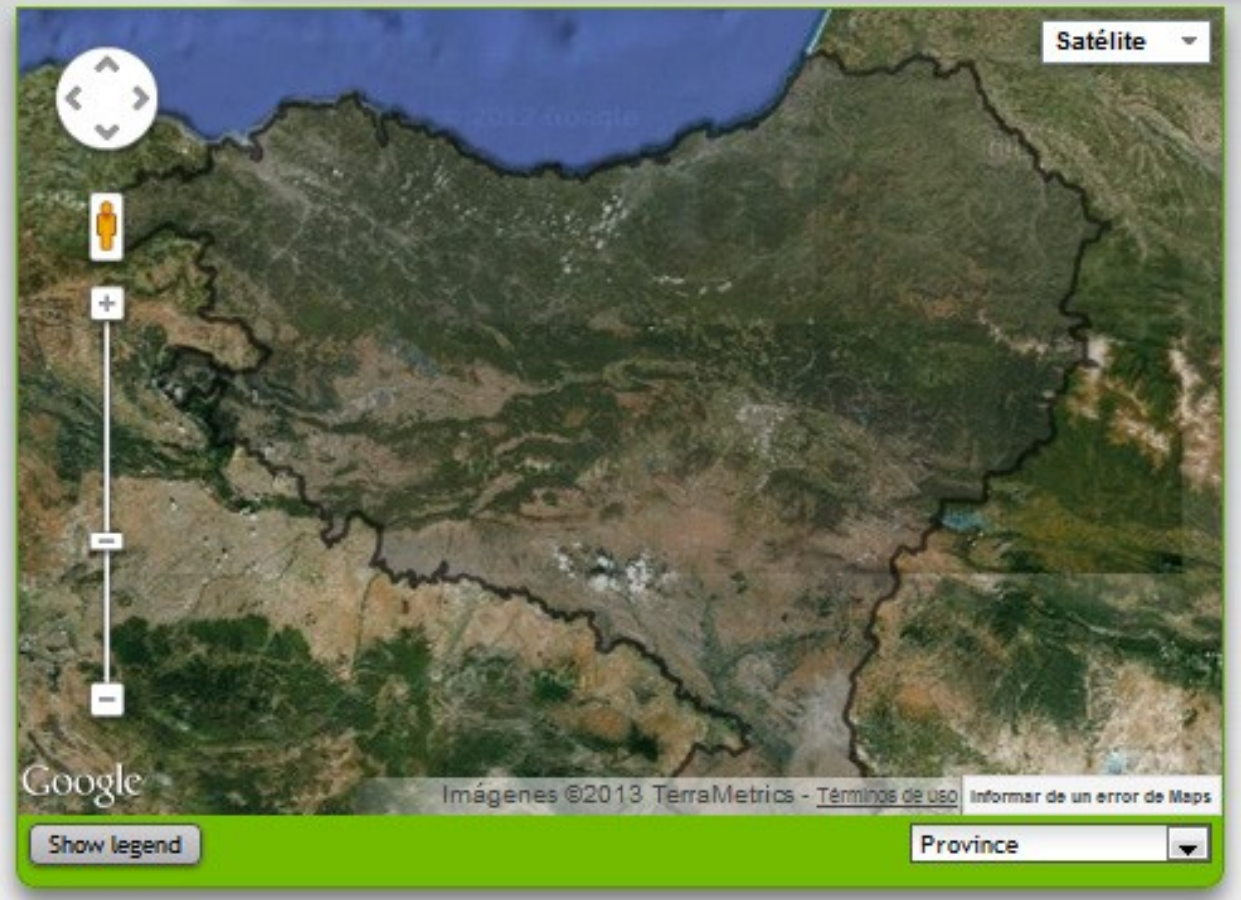
Username

Password

Log in | Cancel

Home | Information | Management | Contact | Help | English

Questionnaires | Other sources | Literary corpora



Search options

Meta-category

Meta-category
Keyword
Province
Dialectal area
Location
Linguistic property
Questionnaire section
Questionnaire
Fieldworker
Informant's birth year
Age range

Basyque

Basque syntactic database

Log in

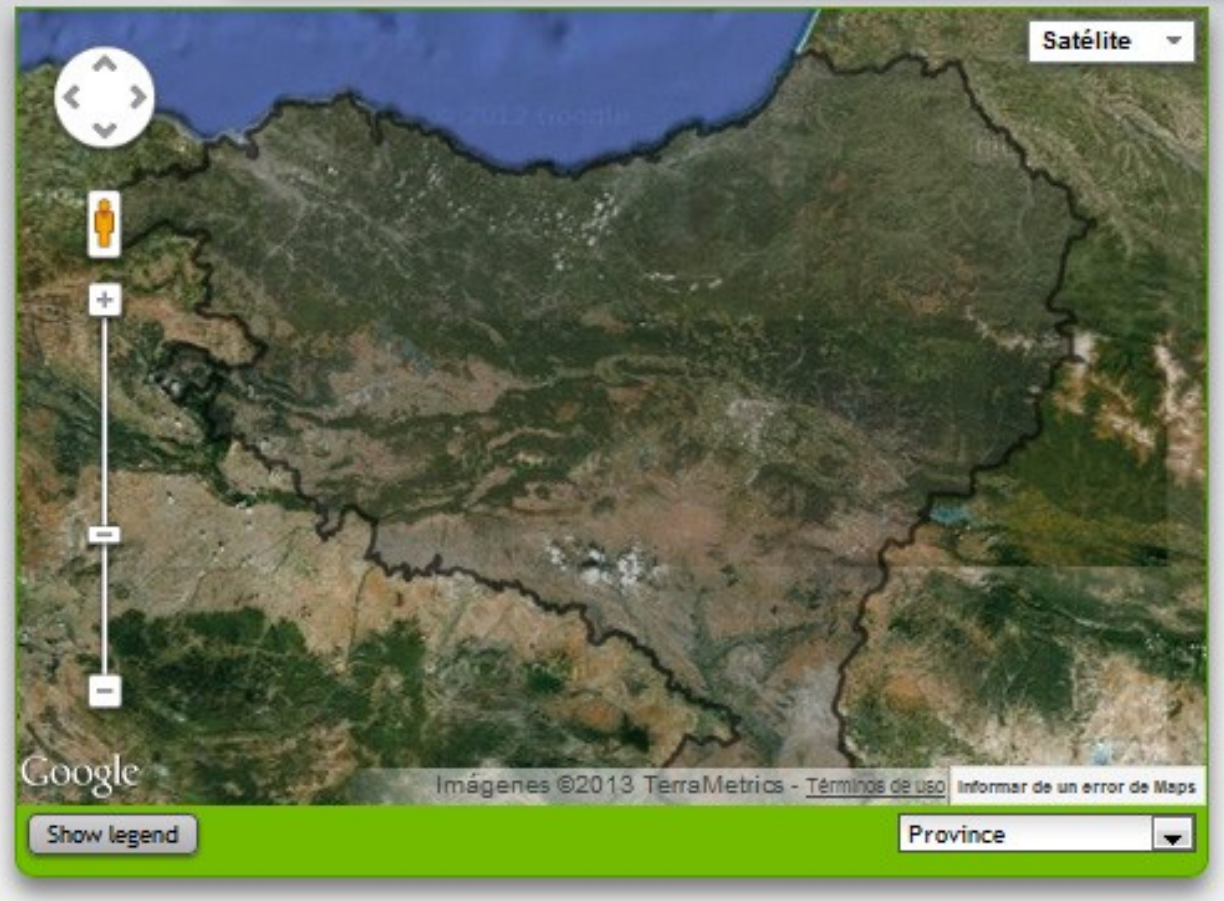
Username

Password

Log in | Cancel

Home | Information | Management | Contact | Help | English

Questionnaires | Other sources | Literary corpora



Search options

Linguistic property

>ABS/ERG

>ABS/ERG
ERG/ABS
Ablative
Absolutive
Agreement
Agreement mismatch
Agreement-PLUR
Agreement-SING
Allative
Allative-animate
Allative-telic
Case mismatch
Causative
Comparative
Dative
Dative argument: demonstrative
Dative argument: DP definite
Dative argument: DP indefinite
Dative argument: elided
Dative argument: indefinite pronoun

Basyque

Basque syntactic database

Log in

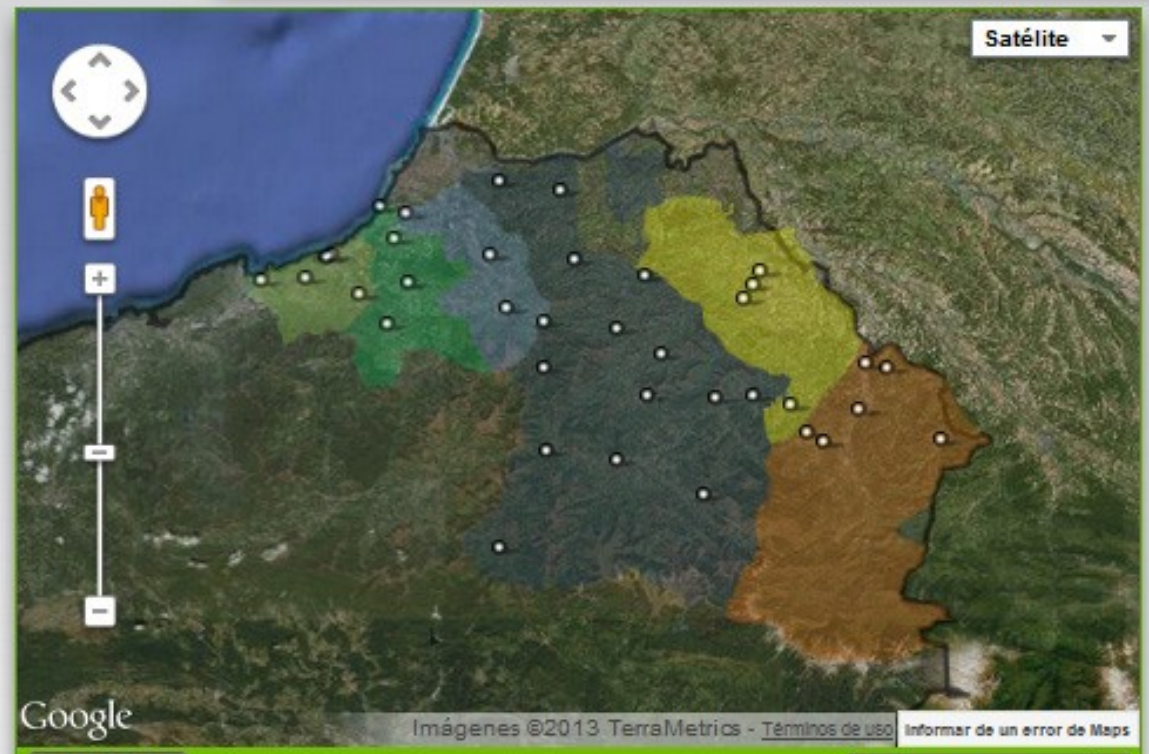
Username

Password

Log in | Cancel

Home | Information | Management | Contact | Help | English

Questionnaires | Other sources | Literary corpora



Show legend

Dialectal area

Search options

Linguistic property

Ablative

Yes

+

Linguistic property: Dative - Yes

Linguistic property: Agreement - Yes

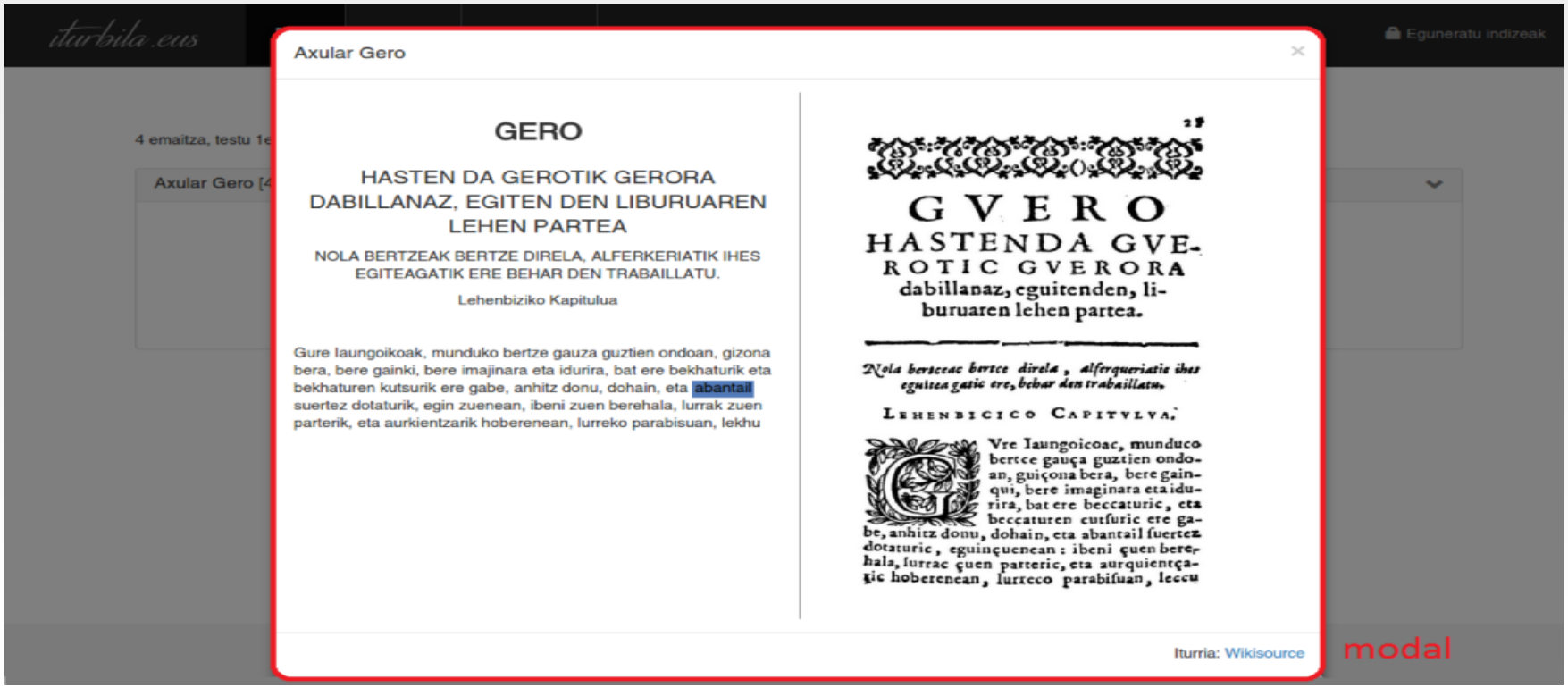
Search

Options

Page 1 Answers per page 5 Answers 1841

Advanced export | PDF | XML | TXT

The Search Interface



The Team

- IKER UMR 5478 (CNRS, UBM, UPPA), IXA (UPV-EHU), Monumenta Linguae Vasconum (UPV-EHU), HiTT (UPV-EHU), Basdisyn (UPV-EHU), University of Deusto.
- French partners: SFL UMR 7023 (Lea Nash), LACITO (Georges Rebuschi)
- University of Cambridge (Ian Roberts), University of Surrey (Greville Corbett), Hungarian Academy of Sciences (Katalin Kiss)
- Three postdocs: Manuel Padilla (IKER, UPPA), Ainara Estarrona (IKER, CNRS), and Ander Soraluze (IKER, CNRS).

Fin

r.etxepare@iker.cnrs.fr

Milesker

Thank You