Dialectologia 1 (2008), 107-119.

BASQUE LINGUISTIC ATLAS-EHHA: FROM SPEECH TO AUTOMATIC MAPS[1]

Gotzon Aurrekoetxea
Eukaltzaindia-Academy of the Basque Language
gotzon.aurrekoetxea@ehu.es

Abstract

The largely desired Basque linguistic atlas-EHHA is to be published nearly. The most important features of this atlas are shown in this paper. The EHHA is a linguistic project of the Academy of the Basque Language-Euskaltzaindia. We present the structure and the most important features of the data base and its different modules (module for the computerisation of data, module of lemmatisation and module of cartography). The module of lemmatisation is the necessary step to link the answers in phonetic characters and the linguistic maps based in orthographic ones. In this module the researcher must group answers, according to etymon, phonetic or morphological features. Furthermore, the maps of EHHA are based on Thiesen polygonation: each point has a polygon, in which the research puts the linguistic features using a colour's palette. As well as being coloured, the largest linguistic areas have also labels to make the interpretation of the map easier.

Key words

Dialectology, geolinguistics, database, automatic mapping

1. Speech Corpus

The most important characteristics of the corpus are the following:[2]

- The questionnaire of EHHA, which has 2.875 questions (Aurrekoetxea & Videgain (1993), was applied in 145 localities.
- In this survey we have gathered around 4.000 hours of sound; that is an average of 27-28 hours per locality.

[1] This paper was presented in the 119th meeting of the Variation Theory Forum of Japan on June 30 2006, held at Seisen University, Tokyo, thanks to the invitation of Asahi Yoshiyuki and Danny Long.
[2] For a general presentation of the Basque linguistic atlas see Aurrekoetxea & Videgain 1989 and 1994, or Aurrekoetxea 2002b and 2004.

- All transcriptions were made by hand (manuscripts) and stocked in five answer notebooks for each locality.
- We have gathered around 830.000 answers (answers and accepted proposals), and lots of linguistic texts: definitions, proverbs, tales, stories, ethnographic narrations... (Aurrekoetxea 1986, 2002a).
- The gathering started in 1987 and finished in 1992.
- The computerisation of data started in 1992 (once the gathering finished).
- Edition work: edition works started in 1998; now we have 5 volumes ready to publish. At first, we thought about two formats of publishing: format book and CD-ROM. But now we are thinking to publish it also in the Internet.

2. Structure of the work

The data were transcribed by hand during the gathering, and stocked in five notebooks by each locality.

The magnetic bands were computerised and stocked in CD-ROMs. And later, a sample of about 2 min. for each locality was published with their transcriptions in CD-ROM format and in book format (Euskaltzaindia 1999). This product can be found in the Web page www. Euskaltzaindia.net.

We started the computerisation of transcriptions from notebooks of localities, locality by locality (two members of the group work in this way).

Once the data computerised, edition work started by a horizontal reading of data, question by question and mapping the data after their lemmatisation.

In figure 1 there is a copy of one page of the 3th notebook of the Dima (117) locality. A part from the information concerning question, locality, record, etc., we gathered the general word for "bred" ***ogi*** (without any suffix). In addition to that, we also gathered *pamitxe, tremes,* and *etxekogi,* three kinds of bred. In this question we didn't suggest any word. For that reason the place of proposals is empty.

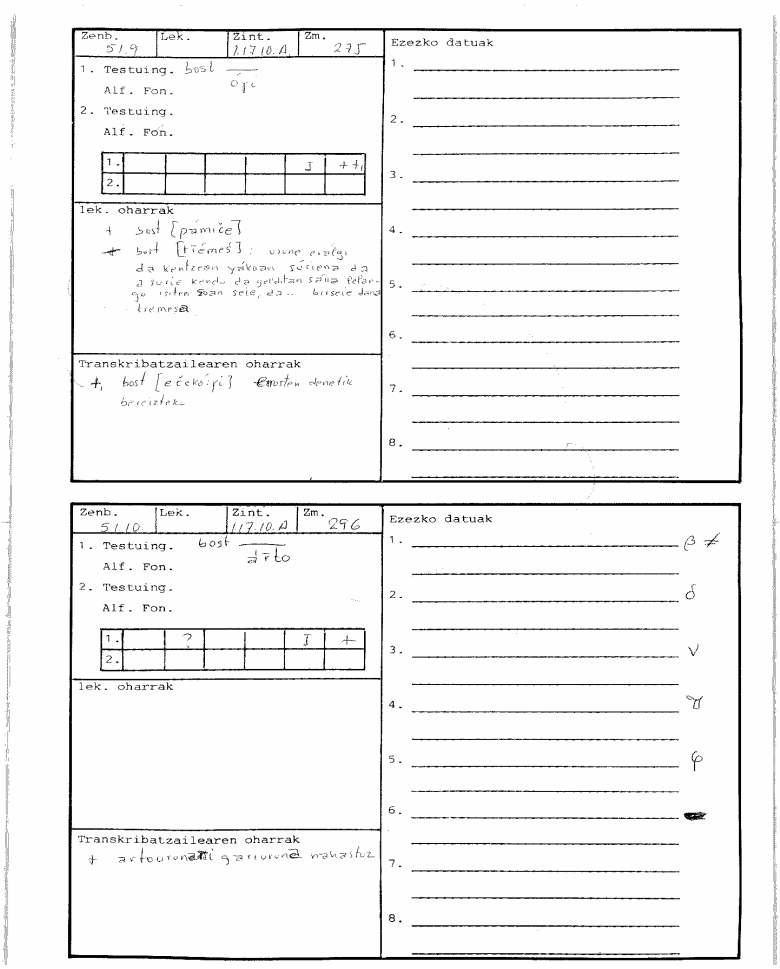

Figure 1. A page of the 3rd notebook of the Dima (117) locality.

But in the next question (below) we gathered ***arto*** for "maize" or "corn" and we suggested six words. The informant only knew the first one, but with a different meaning.

3. Features of EHHA Computer program

This program has been developed by Euskaltzaindia-Academy of the Basque Language. It is an application in Java. The EHHA program is a shared out program. It can have more than one user at the same time. And it is placed (inserted) in the corporative net of Euskaltzaindia.

The different members of the group work in different places: three members in Bilbao, in the head office; one works in Bayonne (France), two in Saint Sebastian and the last one in Pamplona. We all work connected to the server, located in the head office of Euskaltzaindia.

It takes on:

- The data base
- The module for the computerisation of data
- The lemmatisation program
- The automated cartography program

3.1. Data base

The data base is a relational data base (SQL). This data base is composed by different fields and we classify relative information in each of them. The different fields are the following:

- Answers and their grammatical information (grammatical category, reliability of the answer...).
- Proposals and their grammatical and reaction information.
- Texts (different types of texts: definitions of answers, tells, ethno linguistic information...).
- Complementary information (words gathered but not included in the questionnaire).

The answers, proposals and complementary words are always written in phonetic alphabet. We use UNICODE IPA extensions for the codification, with SIL Doulos fonts.

3.2. The module for the computerisation

In figure 2 we can see the way of operation of the computerisation module. The word *pamitxe* has been computerised in the text, whereas ***ogi*** has been computerized in the place of the answers.

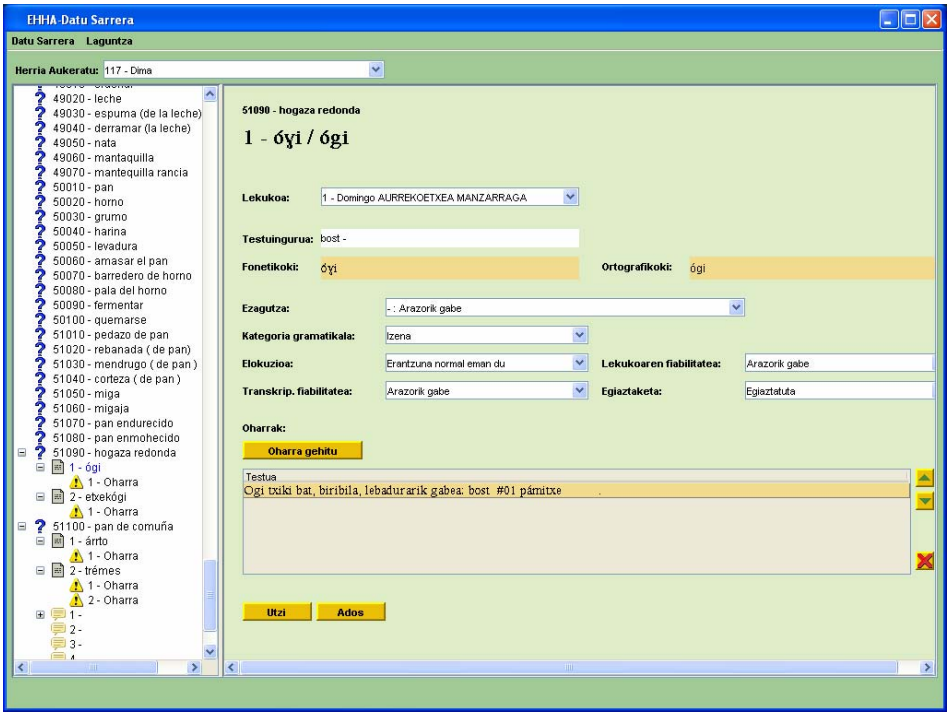

Figure 2. The module for the computerisation.

To computerise the phonetic sounds we use the phonetic board (figure 3), which is activated when we put the cursor in a position.

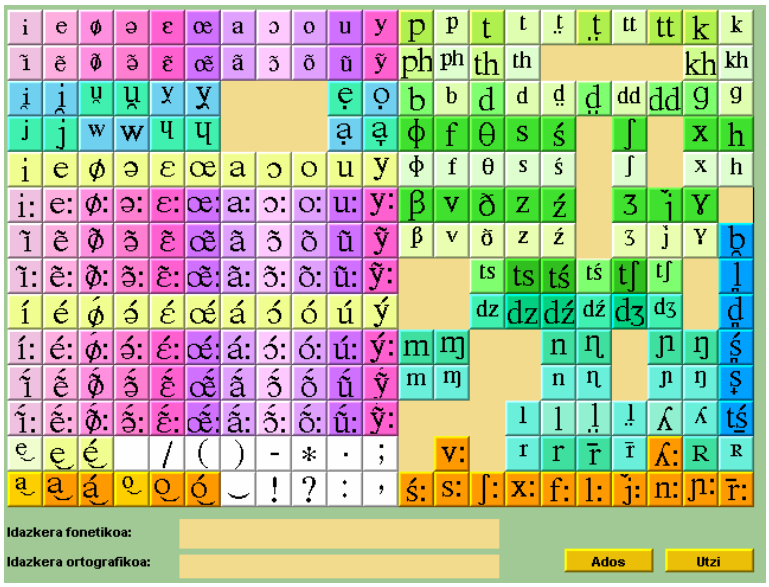

Figure 3. The phonetic board of EHHA.

Each of these characters represents one sound of the Basque language, but it represents one or more than one sign in the UNICODE system. For example: the [a:] character has two UNICODE signs "a" and two points.

For this reason and for the moment we don't have any possibility to navigate between signs, because this possibility isn't still implemented. That's why we have problems when we work with phonetic characters: when we make a mistake we can't go back one or more characters; we must start again and type all the word.

When we press one button of the phonetic board the computer program has two exits:

- On the one hand, it writes this character in phonetic alphabet
- On the other hand, it writes the same character but in orthographic alphabet.

It is because we have created this conversion board between the IPA characters and the orthographic alphabet (figure 4).

ID	Unicode (Doulos)	IPAKIEL kodea	EHHA kodea	Urto
68	s	149	194	z
69	ś	157	19C	s
70	ʃ	181	1B4	x
71	x	202	1C9	j
72	h	209	1D0	h
92	i:	2	101	i
93	e:	16	10F	e
96	ɛ:	46	12D	e
98	a:	66	141	a

ID	Unicode (Doulos)	IPAKIEL kodea	EHHA kodea	Urto
99	ɔ:	76	14B	o
100	o:	86	155	o
101	u:	96	15F	u
102	y:	110	16D	ü
103	β	127	17E	b
211	ő:	92	15B	ó
106	z	150	195	z
107	ź	158	19D	s
108	ʒ	182	1B5	j

Figure 4. The conversion board of EHHA.

This is only a part of the conversion board that we use in the EHHA computer program. In this board we have Unicode Doulos SIL characters, Ipakiel, EHHA and orthographic signs.

3.3. The lemmatisation module

Once the data are fit and before going to the maps, we have to lemmatise all of the answers that we have gathered (figure 5).

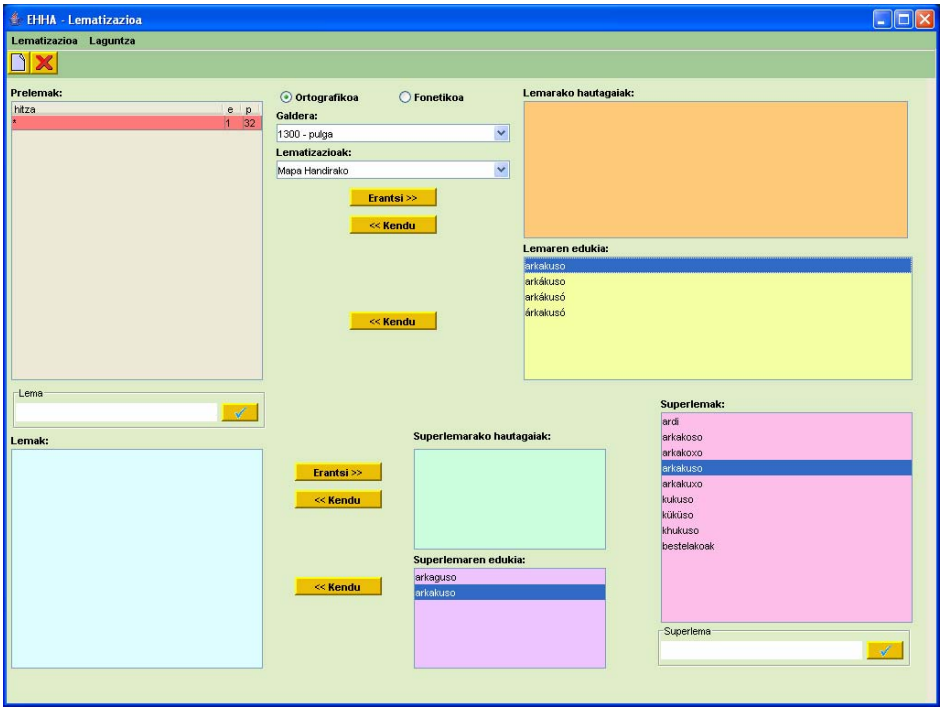

Figure 5. The lemmatisation module of EHHA.

We have four options to make the lemmatisation:

- Two orthographic lemmatisations (one for the big map and another for the small one).
- Two phonetic lemmatisations (one for the big map and another for the small one).

In the lemmatisation board, we call PRELEMA to the whole word of the square on the top right of the board. Once all of the answers in this place, we start the lemmatisation selecting and grouping similar words and creating the LEMA (in the left bottom of the board). We have the following criterion in order to group the words:

- Different etymology: for example: “ardi” and “kukuso”.
- Appearance or absence of a morpheme: *arkakuso / kukuso* (ardi > ark-).

- Appearance or absence of a phoneme: *arkakoso / arkakuso* (the phoneme /u/ vs. /o/).
- Appearance of a phonological rule: in this case we haven't got any.
- Appearance of a phonetic feature: *kukuso / küküso* ([u] and [ü] are different sounds of the phoneme /u/).

To go from the LEMA to the SUPERLEMA we work in the same way: first we select the word or words and them we choose a word or we write a different word, or part of it.

These SUPERLEMA are the legend of the maps. Once we finish creating the superlema we are ready to create a map.

This lemmatisation allows to the researcher to create maps more or less accurately. The editor can produce maps very close to the answers, or even close to the phonetic answers, and can produce also maps grouping more answers. But we put a limit: the editors can not produce more than fifteen superlemas or, in other words, they can not have more than 15 words in the legend of the map. And it is because of the limits of our sight: it is said that we are not able to distinguish more than 20 colours.

The relation between the lemmatisation and the map is very close. If we change a superlema and we open the module of the cartography, the new superlema will appear both in the legend and in the map, with its colour (Aurrekoetxea & Videgain 2006).

3.4. The automated cartography program

The maps are created by the GIS module of our computer programe. We use OpenMap application of the BBN Technologies Company. It is an open Java component.

To make the base of the map, first we drew the boundaries of the provinces. After that, we used the Thiesen polygonation[3] to give each locality a polygon. Inside this polygon, we put the feature of the locality, so that all the localities have their polygon. Therefore, we have 145 polygons in the map.

We have the possibility to create two maps: each question has at least one map. We name this map the big map. The editor is forced to make it always, in all of the

[3] For the application of Thiesen polygonation in linguistic cartography see for example Goebl 1992.

questions. Apart from that, the researcher can make another map, which is much smaller than the first one. In this map we can represent different features.

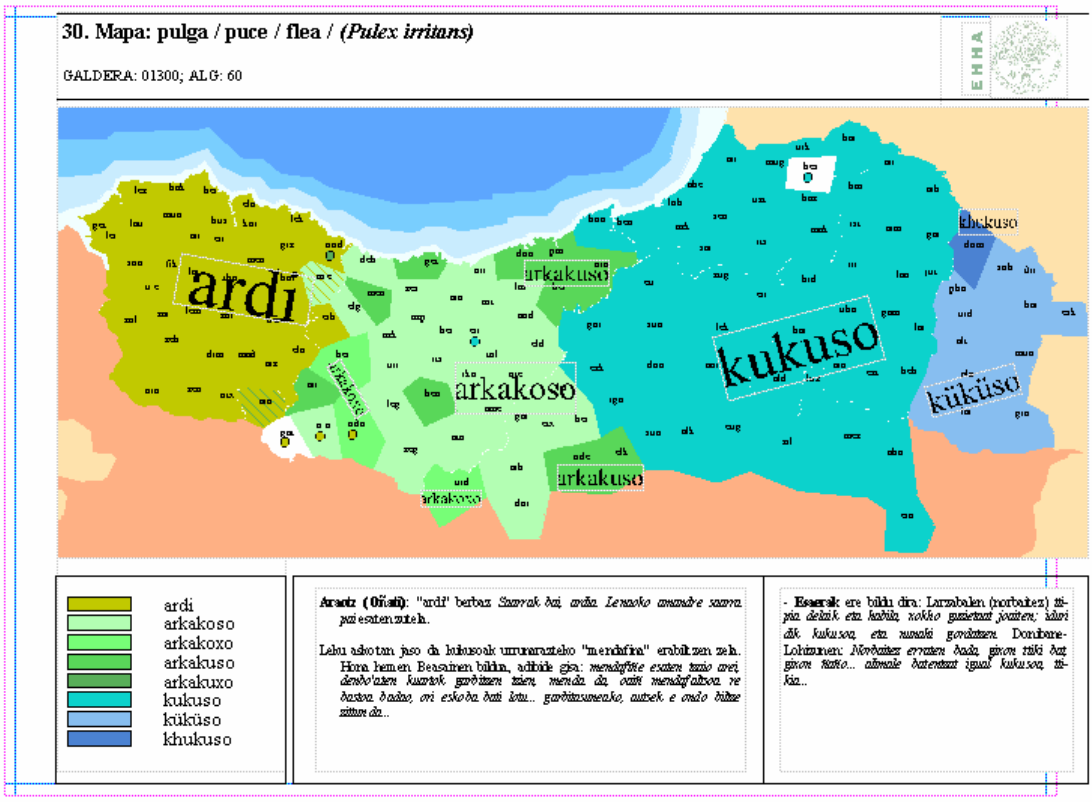

Figure 6. A map of the EHHA.

3.4.1. The big map

Thinking in the publication of this work in book format, we have designed a map with A4 dimensions. In spite of this, we can put three linguistic characteristics in the map: two answers and one proposal. For that purpose, we use three visual features (fig. 6):

- The colour of the background of the map for the first answer: for example, the blue for the “kukuso”, or the yellow for “ardi”.
- When we have two answers in the same place we use a weft with the colour of the corresponding word over the colour of the background for the second answer; for example the green for the word “arkakuso”, or the yellow in Exebarria.

- If we have gathered one or more than one proposal, we can draw one of them inside a circle with his corresponding colour: for example, there are yellow circles in the left bottom of the map, where the informant accepted the word "ardi" when the researcher asked him whether he knew it or not.

The localities are designed by codes in the maps. These codes are abbreviations of the name, in most of the cases. When two or more abbreviations are identical, we change one of then; i.e. **lem** (Lemoa) and **lez** (Lemoiz) in the Biscayan province. These abbreviations are included in the data base, next to the whole name of the localities.

The researcher can choose the colour of the superlema of the legend and he has done it in almost all the cases. The linguistic affinity or proximity of the superlemas is represented by a similar colour or a range of colours, so that the linguistic difference is reflected by colour difference.

The editor adds labels over the map. These labels can have the same text as the word of the legend or can be different, especially when the surface of the label is very small. In those cases we have a tendency to approach to the realisation of the answer.

These maps can be "read" or be consulted in three levels:

- The first level of consulting is seeing the colour of the biggest areas of the map and reading the labels.
- The second level is reading the small areas, the wefts and the circles of the map.
- The third level is reading the notes bellow the map.

3.4.2. The small map

The editor of the Linguistic Atlas of the Basque Language has the opportunity to create a second map for each question. It is him who decides: he can make it or not. It is his judgement.

Figura 7. The small map of EHHA.

The emplacement of this map in the book is between the answers and the big map, at the right bottom of the first page of each question (figure 7).

This map has a smaller size than the big map in the book format and shows two or three characteristics that appear in the whole data of this question. For example, the small map of the 01300 question shows the partition of the space in three parts: *ardi, arkakuso* and *kukuso/küküso*. It is a simplification of the big map.

The directors of the project encourage the editors to create those maps when the big map can not show a relevant feature. For example:

- A linguistic loan vs. an autochthon word: for 'butterfly' we have *mariposa* or *papillon* (two linguistic loans) and *mitxeleta, pitxeleta, jinkoaren oilo, pinpilinpauxa, txilipotadea...* (autochthon words).
- The different realisations of a phoneme.

3.4.3. Other answers

In some cases we have gathered more than two answers and more than one proposal. When that happens we show this data, which don't have a place in the map, in a appropriate place near of the list of answers.

The program is prepared to show the whole map, or only a part of it: for example, the Navarre part or the Biscayan part of the map. But in the actual version it is not implemented.

4. Conclusions

We have shown the software structure of the Basque Linguistic Atlas-EHHA. This data base will be appropriated for automatic on-line search in the nearby future. Apart from the data base we have shown the module of automatic cartography. The linguistic maps of EHHA are based on Thiesen polygonation and each point has a polygon in which the researcher puts the linguistic features of the point using a colour's palette. We think that this kind of map is a new contribution to linguistic cartography.

The lemmatisation we propose and the cartography used are essential to attract a wider public than usual.

5. References

AURREKOETXEA, Gotzon (1986) "Euskal Herriko Hizkuntza Atlasa (EHHA): inkesta metodologia eta ezezko datuak", *Euskera* XXXI, 413-424.

AURREKOETXEA, Gotzon (2002a) "Algunas consideraciones sobre la contrapregunta en las encuestas lingüísticas", in L. RABASSA (ed.) *Mélanges offerts à Jean-Louis Fossat,* Toulouse: Université de Toulouse II-Le Mirail, 57-65.

AURREKOETXEA, Gotzon (2002b) "El atlas lingüístico vasco (EHHA)", in M Aurnague / M. Roché (dir.), *Hommage à Jacques Allières-1 Domaines basque et Pyrénéen,* Biarritz: Atlantica, 63-71.

AURREKOETXEA, Gotzon (2004) "El atlas lingüístico vasco: 20 años de innovación tecnológica", in Maria Pilar PEREA (ed.) *Dialectologia i recursos informàtics,* Barcelona: Universitat de Barcelona, 15-41.

AURREKOETXEA, Gotzon & Charles VIDEGAIN (1989) «L'atlas linguistique du Pays Basque. Euskal Herriko Hizkuntza Atlasa», *Bulletin du Musée Basque,*1. trim., 495-510.

AURREKOETXEA, Gotzon & Charles VIDEGAIN (1993) "Euskal Herriko Hizkuntza Atlasa-EHHA. Galdesorta / cuestionario / questionnaire", *Euskera* XXXVIII-3, 529-647.

AURREKOETXEA, Gotzon & Charles VIDEGAIN (1994) "Historia y futuro del Atlas Lingüístico Vasco (EHHA)", in P. GARCÍA MOUTON (ed.) *Geolingüística. Trabajos europeos,* Madrid: Biblioteca de Filología Hispánica, CSIC, 79-96.

AURREKOETXEA, Gotzon & Charles VIDEGAIN (2006) "L'interprétation dans les atlas linguistiques? le cas de l'Atlas linguistique basque (EHHA)", in A. TIMUSKA (ed.) *Proceedings of the 4th International Congress of Dialectologists and Geolinguistis, Riga, 2003*, Riga: University of Latvia.

EUSKALTZAINDIA-The Royal Academy of the Basque Language (1999) *Euskal Herriko Hizkuntza Atlasa: ohiko euskal mintzamoldeen antologia* [Anthology of the traditional varieties of the Basque language], book and CD, Bilbao: Euskaltzaindia [ftp://www2.euskaltzaindia.net/Antologia/Antologia.pdf].

GOEBL, Hans (1992) "Problèmes et méthodes de la dialectométrie actuelle (avec application à l'AIS)", in G. AURREKOETXEA & Ch. VIDEGAIN (eds.) *Proceedings of International Congress on Dialectology,* Bilbao: Euskaltzaindia, 429-476.