DiG 24 (2016), 21–41 *DOI 10.1515/dialect-2016-0002*

Analysis of the morphological variation of Basque

Gotzon Aurrekoetxea

Abstract

The lack of homogeneous data recorded by similar methodologies has been a handicap to the development of more advanced studies on the dialectal variation of Basque. Now that the first five volumes of the Linguistic Atlas of the Basque Language have been published, researchers have access to a great amount of data and the possibility to use more sophisticated procedures to analyze the variation of the Basque language from a geolinguistic point of view.

In this contribution, we use data taken from the fifth volume of this atlas, which is devoted to noun morphology. First, these data will be analyzed linguistically, and instead of using the phonetic representation, we will use the phonological or underlying representation. In order to do this, we will analyze the inventory of the morphological suffixes used in nominal inflexion cases and the phonological rules which appear in these cases (mainly assimilation, dissimilation, insertion and deletion), using the classical view of generative phonology.

As far as the cartography of the data is concerned, we will use the recently created *Diatech* tool, instead of other dialectometrical tools, because of the multiple responses (MR) which that data show. In effect, the *Diatech* tool provides a more accurate processing of the MR than other tools. As well as conceptual maps, synthetic maps will be presented; these maps will show different geographical organizations of the suffixes and phonological rules.

1 Background

Studies on geolinguistic variation in the Basque language have been fruitful since the beginning of the dialectal literature (Alvarez & Aurrekoetxea 1987; Martínez Areta 2013), and have contributed to our knowledge of the dialectal variation of Basque, but the lack of homogeneous data has been a handicap to the development of more advanced studies. Now that the first five volumes of the *Eskararen Herri Hizkeren Atlasa* 'Linguistic Atlas of the Basque Language' (henceforth EHHA) have been published, researchers have access to a great amount of data (for the first five volumes, Euskaltzaindia 2010–2013), never before even thought of.

In this contribution, we will use data taken from the fifth volume of the EHHA, which is devoted to noun morphology; we will not take all of the data, only the nominal inflexion ending in the *-o* vowel. Firstly, we will use

phonetic data to carry out a statistical analysis, and then these data will be analyzed linguistically, and instead of using the phonetic representation, we will use the phonological or underlying representation to carry out a second statistical analysis. In order to do this, we will analyze the inventory of the morphological suffixes used in nominal inflexion cases and the phonological rules which appear in these cases (mainly assimilation, dissimilation, insertion and deletion), using the classical view of generative phonology.

As far as the cartography of the data is concerned, we will use the recently created *Diatech+* tool, instead of other dialectometrical tools, because of the multiple responses (MR) which that data show, and in order to analyze linguistic distance from the phonetic and phonological points of view. In effect, the *Diatech+* tool provides a more accurate processing of the MR (Aurrekoetxea et al. 2013) than other tools. As well as conceptual maps, synthetic maps will be shown; these maps will show different geographical organizations of the suffixes and phonological rules.

2 Methodology

Following I. Laka (1994), the Basque language, Euskara, is an agglutinative language and uses postpositions instead of prepositions. These postpositions are attached to the last word of the complement noun phrase. For example, using the noun *leiho* 'window', we give the following phrase as an example:

[zazpi leiho]	-tatik >>> zazpi leihotatik
[seven window]	from
'from seven windows'	

The Basque language is rich in morphology; it is very abundant in both noun and verb morphology, and these can not be compared with the morphology of its neighboring languages. One of the most remarkable features of Basque noun morphology is its inflection. The grammar of Basque published by the Euskaltzaindia (1993) proposes 16 cases of inflexion. According to their function in a sentence we can gather them into three groups[1]: four of them (absolutive, ergative, dative and partitive) are grammatical cases, which have agreement with the verb. The absolutive case (morpheme -Ø) can be the subject of an intransitive verb or the direct object of a transitive verb. The ergative (-k) can only be the subject of transitive verbs, and the dative (-*i*) can only be the indirect complement of the verb. The partitive (-*ik*) has similar functions to the absolutive case, but in indefinite concepts.

[1] For the English names of the cases we follow Hualde & Ortiz de Urbina 2003.

Two cases form noun complements: the possessive genitive (-en) and the relational or locative genitive (*-ko*). In the third group we have the cases which form verb complements: locative cases, such as the locative or 'inessive' (*-n* /*-engan*), ablative (*-tik* / *-engandik*), allative (*-ra* or *-ra arte* /*-engana*), directional (*-rantz* / *-enganantz*) 'towards' and terminative (*-raino* / *enganaino*) 'up to'. And finally, non-locative cases, such as the comitative (*-ekin*), benefactive (*–entzat*), instrumental (*-z*) and prolative (*-tzat*).

With regards to the nominal inflexion of the dialects, it is important to note that there are many differences from one dialect to another: some of these are differences in suffixes (dialects can use different suffixes), but some of them are differences in phonological rules (PhRs). Examples of differences in suffixes include e.g. in the associative case, the *-areki*(*n*) suffix of the central part of the Basque territory, the *-arekila*(*n*) allomorpheme of the eastern part, and the *–agaz* suffix of the western part. There are many differences of this kind. And there are also different PhRs in the dialects; these PhRs are not always known in every locality of one dialect, while some of them are scattered throughout the whole territory; there can be deletion rules, dissimilation, assimilation, addition, monophthongization, etc. To give another example, in the absolutive plural case of one word ending with the "o" vowel we have the following different options in the dialects: *-oak*, *-ook*, *-ok*, *-uak*, *and -uek.*

3 Data

Data was taken from the fifth volume of the EHHA, which refers to nominal morphology. From its 256 maps, it gathers all nominal inflexion cases; each case with words ending with all the vowel phonemes in Basque, in consonants and diphthongs, and cases in three forms according to the number (indefinite, singular and plural). The volume also gathers other questions about nominal morphology, such as the indefinite determinants. There are 188 questions about nominal inflexion cases. It encompasses the entire inventory of nominal inflexion cases in Basque (Euskaltzaindia, 1993). For this presentation we have selected the information gathered in 51 maps which belong to 51 questions: the information related to roots which finish with the *-o* vowel.

This is the first time that we have used these data for statistical analysis, and therefore, it is the first time that the 145 localities of the Basque language have been analyzed from a geolinguistic point of view. Because of that, this work is the first step towards more in-depth investigations in the future. Overall, the total sum of data that we analyze in this research is of 7,395 items (51 questions x 145 localities).

I will not discuss the reliability of the data here; everyone knows the value of linguistic atlases. In all fields of science which involve surveys there is some concern, not only in linguistics and dialectology. In the EHHA project, the survey methodology was treated as one of the most important tasks in the project; we created our own system for the survey time and for gathering data. Interviewers were required to take and transcribe all responses given by informants and, in addition to asking the direct questions, to make some proposals, asking them if they knew the word and whether or not they knew the grammatical form (Aurrekoetxea 2002). In the publication, all responses and all accepted proposals (as words used in the locality and by the informant) have been accepted as publishable and are therefore published (Euskaltzaindia 2010). There is one difference between the responses and the accepted proposals in the published material: while the responses are in bold, the accepted proposals are in italics and have an asterisk (*) in front of them. In this study, however, we have not kept this difference in mind and we take all of them to be responses. The reason for not identifying this difference is our conviction that both answers are part of the speaker's linguistic competence: the response is active competence and the assumed proposal is passive competence.

4 Responses and underlying representations

In this work we display two types of data: responses in orthographic transcription and underlying representations. The responses in orthographic transcription have been taken directly from the EHHA, but taking only the last vowel of the root and the nominal inflexion suffix of the case. For example, in the absolutive plural case in the first two localities, we have the data given below in (1):

(1)	Leioa	**astóak**	(ast-o	+	ak	-oak)
	Getxo	**ástok**	(ast-o	+	ak	-ok)
		'donkey'		+	abs. suf. pl. 'the'	

In the majority of the localities, the word used to gather data for this case was "asto" 'donkey'. What we did was to take only the last vowel of the root (*-o*) and the suffix of the case (*-ak* for the absolutive plural): in the locality of Leioa the part corresponding to the response is *-oak* and in Getxo, *-ok*. Therefore, the answer entered in the database was "oak" for the first locality *ok* (Leioa) and *ok* for the second (Getxo), and so on. We consider that in the *ok* answer there is a monophthongization or simplification of the vowels (*oak* > *ok*), taking into account the scientific literature on dialectal nominal inflexion and the PhR of the different varieties.

As far as the underlying representation of these responses is concerned, we propose a linguistic analysis of the data using the theoretical framework of traditional generative grammar, of SPE-style phonological rules. The framework has to be chosen by the researcher, according to the data and the objective of their study. From the different frameworks of structural dialectology, traditional generative dialectology and optimality theory (see Barbier, 2010 for a very short and comprehensible introduction to this subject), we have opted for generative grammar, because there are very few works about Basque variation using the OT framework (for the comparison of a phonological rules analysis and a constraints analysis of Basque, see Hualde 1997 and San Martin 1998). Taking into account the responses of the first question (the absolutive plural of the *-o* ending root), in all cases the suffix of the absolutive plural is the morpheme *-ak*; so, the underlying representation is *-oak* for all surface structures found in the dialects. And therefore, for all responses of this case the underlying representation will always be *oak*.

Here I use the term 'underlying representation' to group all of the surface structure, in this case the allomorphs, into just one form or word which is placed in the underlying representation of the language. Using underlying representations to measure the dissimilarities between localities means that we are working at a higher linguistic level. This level entails an abstraction from the phonetic data; we go up from the phonetic data to linguistic features, from speech to language, from a series of infinite variation of the data (phonetic variation) to a finite number of linguistic features of the linguistic system of each dialect or locality. This level corresponds to the phonological level. Therefore, the achieved result is the phonological distance.

There is a third possibility for measuring the dissimilarities: to use phonological rules. In the nominal inflexion case that we are using (absolutive plural case), and taking into account all of the responses gathered, we have the options listed below in (2):

(2) *-ok*, *-oak*, *-uak*, *-uk*, *-uek*, *-oog*, *-oag*, *-oek*

All of these forms can be inserted into a hierarchical structure as follows in Fig. 3. In this schematic structure, we can distinguish five phonological rules (symbolized by different letters: A, B, C, D and E) which are operating in the "-o + -ak" case: one dissimilation rule, two assimilation rules, one voiceless rule and one monopthongization rule.

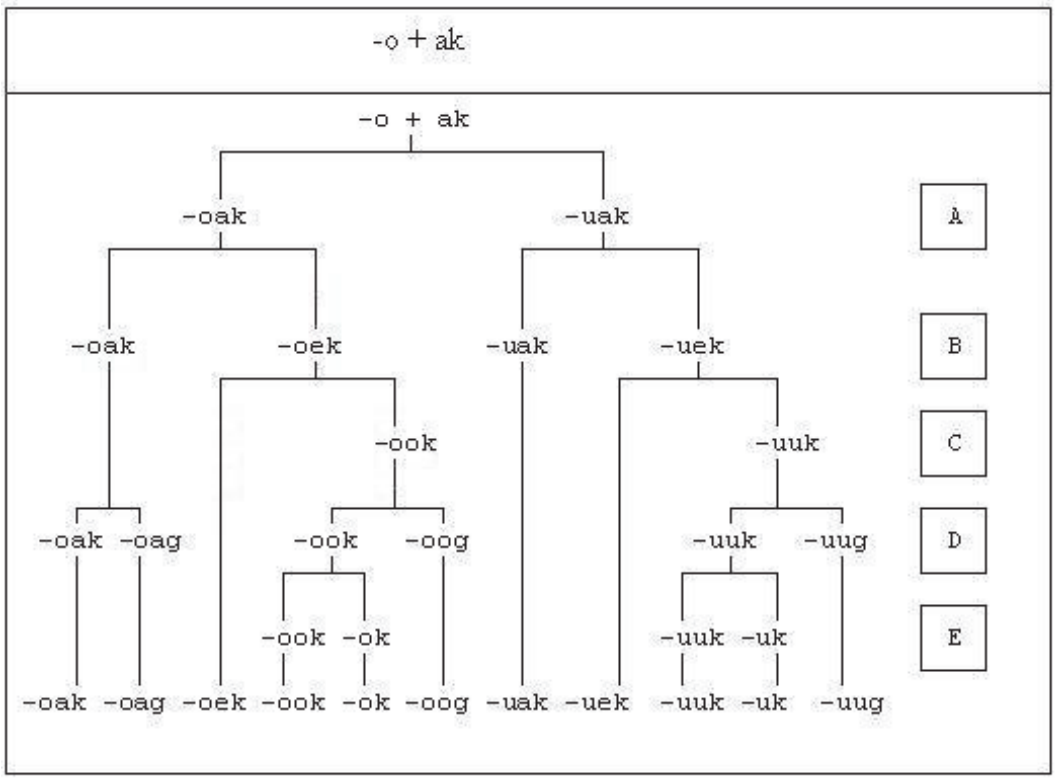


Fig. 3: Hierarchical structure of PhR of *-o+ak* case.

As one can see, in this tree we have five phonological rules:

a. on the A level a dissimilation rule, which changes from the underlying form *-oak* to the surface form *-uak*;
b. in B, an assimilation-1 rule, which changes the *-uak* > *-uek*; or *-oak* > *-oek*;
c. in C, an assimilation-2 rule, which changes from *oek* > *ook* and *-uek* > *-uuk*;
d. in D, a voiceless rule, which changes the final *-k* > *-g* (*-oak* > *-oag*, *-ook* > *-oog*, *-uuk* > *-uug*);
e. And in E, a monophthongization rule which changes two identical vowels into one: *-ook* > *-ok*, *-uuk* > *-uk.*

5 Analysis of the data

It is commonly accepted that linguistic atlases sometimes, and in some questions, show empty answers (henceforth EA), and in others, they show more than one response or multiple responses (MR). However, these features are not the consequence of one gathering methodology; at least this is not the case for the EA, because empty answers can be found in all linguistic atlases.

As far as the statistics and cartography of the data are concerned, we will use the online *Diatech* tool.[2] We have two reasons for proposing the use of this tool. These are to process the MR in a more accurate way, and to measure linguistic distances in two ways: phonetic distances, using Levenshtein distances (Heeringa 2004) and phonological distances, using Relative Index Value-RIV (Goebl 1978; 1981: 357–361; 1992: 436–438). To our knowledge, this tool is the only one which provides users with both measure indexes. The *Diatech* tool has been created with two main aims: to appropriately solve the problem of MRs and, in our personal effort, to assist in the task of making the research about linguistic variation carried out in different languages more portable (Aurrekoetxea et al. 2013).

If the use of appropriate statistical procedures is important, it is even more important to know what are we measuring, and how we do so. Do we measure the linguistic distance but taking only the surface structure of the dialects, or should we measure the underlying representation of the dialects? As has been shown on more than one occasion (Clua 2010), the linguistic distance is very different with linguistic analysis (using underlying representations) and without it (using phonetic representations). Both distances are interesting but they are different, and on some occasions very different. Therefore, before starting to analyze the quantification of the distance, we have to make some decisions, which only dialectologists can and must make.

If we assume that all languages and dialects are linguistic systems and if we want to measure the distance between them, we should measure the distance at the phonological level, which takes into account the linguistic system. To clarify the terminology: by phonetic feature or phonetic level I mean data gathered in the surveys without any linguistic analysis; these data can be written using phonetic or orthographical transcription. And by the phonological level I mean that a linguistic analysis is carried out rather than using raw data; such phonological features have been obtained using different theoretical points of view (structural, generative, optimality, etc.). They are not data, but linguistic features, an abstraction of the data gathered in the surveys.

Now we have more than one level for measuring linguistic distances between dialects: the phonetic and the phonological level. What is the difference between them? To measure the phonetic distance, researchers use gathered data without carrying out any linguistic analysis, while to measure the phonological distance the researcher does not measure the data but the linguistic feature. The difference between the two measurements can be great as has been shown many times (Aurrekotxea 1995; Clua 2010). Two linguistic distances will be calculated in this work: phonetic and phonological distances.

---

[2] www.eudia.ehu.eus/diatech

5.1 *The choice of the linguistic distance*

Apart from the choice of the level of linguistic representation, we also have to choose the distance unit for counting the dissimilarities between data. There are different ways to count these dissimilarities and, therefore, there are different distance units to measure geolinguistic variation: the most widely used ones in dialectometry are the RIV, used mainly by the Salzburg school (Goebl 1978, etc.) and the Levenshtein distance (the so-called 'edit distance' and also 'string distance'), used in dialectology mainly by the Groningen school (Nerbonne & Heeringa 2010; Valls et al. 2012). There are others used in linguistics, including the Euclidean, Hamming distance (Séguy 1973), Manhattan distance, Canberra measure, binary or Minkowski measure (Lafkioui 2009), cophenetic distance, aggregate distance, and distributive distance or χ^2 (also called CHI-2 and KHI-2) (Philps 1985). Using different units, different distances are achieved; this is not catastrophic or disastrous. It is just like choosing the meter, or the mile to measure geographical distances. All of these are good units; the choice is the task of the researcher, according to the aim of his/her study. Here we will consider only the first two distance units: the Levenshtein distance to measure the phonetic distance and the RIV to measure the phonological distance.

5.2 *Phonetic distance*

The Levenshtein (Lv) distance has previously been successfully applied to measure the phonetic distance, mainly using orthographic responses, but it can also be applied to responses in phonetic transcription (Spruit, Heeringa and Nerbonne, 2008: 70–71). We consider that in the majority of the cases in which answers are built with the same or similar features (identical etymology or identical underlying representations, for example), this measurement can be applied. This distance uses strings to measure the distance between answers and gives the results in percentages. This distance is based on deletions, substitutions and additions which one word can undergo from one locality to another. The minimal difference undergone will be counted; the more changes in the word the greater the linguistic distance.

Although the Lv is a very good unit to measure words with an identical etymology, it is not recommended for words which have different etymologies; for example, it does not make sense to measure the phonetic distance between *apatx* and *azazkal* 'hoof' (EHH III: 233), as shown below:

apatx	*azazkal*	*difference*
a	a	0
p	z	1
a	a	0

tx	z	1
	k	1
	a	1
	l	1

Here, there is a difference of 5/7 (0.71). This is not a good measurement, because they are two different words and the two correspondences in the *a* vowel are fortuitous. So, the difference should be 1.00 and not 0.71.

Looking at another example, it does not make sense to talk about different phonetic measures between *zanaori*, *azenaio*, *pastenarre* and *karrota* 'carrot' (EHHA III: 117); they are very different words and the only distance that can be counted is 1.00.

And finally, there is no sense in comparing *kapatu*, *osatu* and *xikitatu* 'castrate' (EHHA III: 249) because all of them have the ending *-tu*, which is universally known to be one of the participles of morphemes of the Basque verb.

So, we recommend using the Lv when we have words with identical etymologies or when we have data with identical underlying representations or inputs. On the other hand, if all the gathered data have the same etymology, using categorical measurement units is not recommended for analyzing phonetic distance, because with categorical measurement a minimal pronunciation difference is counted as an entirely different word and is therefore counted in all cases as 100% difference.

When measuring words with different etymologies, it does not make sense to measure phonetic distance, because all the compared strings will be always or almost always different. Neither does it make sense to use the Lv, because fortuitous similarities in some strings can imply reduction or minimization of the distances.

In our work, in order to analyze the phonetic distance, the Lv is one of the most suitable measures. In effect, this distance works well for our data, because in the majority of the cases the underlying representation of the data of each question is the same. Because of this, we will apply this algorithm to measure phonetic distance. When dealing with this kind of data it is more worthwhile to use string measurements than categorical ones.

5.3 *Phonological distance*

As far as the phonological distance is concerned, we use underlying representtations as a basic feature. To measure this distance we choose a categorical or discrete measurement (the so-called 'nominal measurement'), and not a string measurement. Out of all of the options, we propose using the RIV. The categorical measurement fits better than other measurements to quantify the differences between dialects using the phonological level or underlying representation. At this level we are not talking about data, but about features;

abstract features which are part of the linguistic system. At this level there is no question as to whether two features have identical strings or not. There are different features which have no similarities. These underlying representations have been created by keeping in mind all the phonological rules that have been used in all the dialects to pass from the underlying representation to the surface structure, as has been shown previously. As a categorical or discrete measurement, the RIV measurement is not gradual and cannot be measured in percentages. All of the similarities are "0" (that is to say, two compared words or features are identical) and the dissimilarities are always counted as "1" (the compared words or features are different).

5.4 *Linguistic distance*

As we have said in section 4 (fig. 3), there is at least one more option for measuring the distance between the localities: to count the number of PhRs needed to pass from the underlying representation to one answer type; the more PhRs needed the greater the distance between them. If there is only one PhR needed to pass from the underlying representation to the response of one locality, the distance will be "1"; if we need two PhRs the distance will be "2", as follows:

- from *-o* + *ak* to *-oag*: distance 1 (one PhR needed to pass from *-oak* to *-oag*, the voiceless rule);
- from *-o* + *ak* to *-ook*: distance 2 (two PhRs needed to pass from *-oak* to *-ook*: the assimilation-1 and assimilation-2 rules);
- from *-o* + *ak* to *-ok*: distance 3 (three PhRs needed to pass from *-oak* to *-ok*: the assimilation-1, the assimilation-2 and the monophthongization rules);
- from *-o* + *ak* to *-uk*: distance 4 (four PhRs needed to pass from *-oak* to *-uk*: the dissimilation, the assimilation-1, the assimilation-2 and the monophthongization rules).

Nevertheless, we have no system for measuring these dissimilarities; as far as I know neither the Lv nor the RIV distance measures the dissimilarities in this approach. So, we will use only the first two methods in this paper.

6 Cartography of the data

The main work of the dialectologist, when analyzing regional variation, is to draw dialect areas and to analyze their limits. Both areas and limits have been studied using different methods and techniques for achieving more accurate approaches and for drawing more visible and understandable maps.

In this presentation we will use only cluster analysis. It is well known that there are many cluster types; nevertheless, we will only use hierarchical clustering, which is the most widely used one. This statistical procedure is also called 'automatic classification', 'numerical taxonomy', 'botryology' and 'typological analysis'. Out of all the possible clustering types, in this work we will use "agglomerative hierarchical clustering". I propose using this to identify the hierarchy of the analyzed localities. Before using the clustering statistical procedure, I did not know what the relationship between these localities was. I use it as a basic first step. Concerning the algorithms, although there are different options (including Ward, Complete-linkage, Single-linkage, average linkage, median, centroid, and UPGMA), we will use the *Ward* algorithm, which is an unbiased one.

The output of the hierarchical clustering is the dendrogram, in which the objects (here the localities) are placed in the x-axis, whereas the y-axis marks the distance at which the clusters merge. Although there are many methods to classify objects hierarchically, it is known that "these methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters" (http://en.wikipedia.org/wiki/Cluster_analysis). Moreover, when the aim of the researcher is to project the outputs onto the map, this step becomes essential. The partition of the dendrogram is absolutely necessary and it must be carried out by the researcher, according to their knowledge and the scientific literature.

The cartography of the data is produced using Voronöi polygons and by coloring the polygons of all localities included in each cluster with the same color, as is usual in dialectometrical cartography.

6.1 *Cartography of the data using phonetic distance*

First of all we display a cluster analysis showing the phonetic distance, using the Lv, as stated previously (fig. 4). Our cartographical analysis has been carried out with the following two points taken into consideration: the scientific literature about this kind of variation in Basque and the best cutting of the dendrogram. The dendrogram has been cut into 7 clusters (the localities belonging to each one are colored with a different colour). Beginning from the left-hand side, the first big cluster is made up of localities which are colored differently: the red color which includes 22 localities and the light red with 25 localities. The second main cluster is formed of light yellow with 14 localities, and the light blue cluster with 37 localities. And the third main cluster is formed of three clusters: the dark blue with 17 localities, the green with 13 localities and the dark orange, with 17 localities. To achieve these 7 clusters we have cut the dendrogram at the 250 level of the scale (0-2500); so it gathers 10% of the difference.

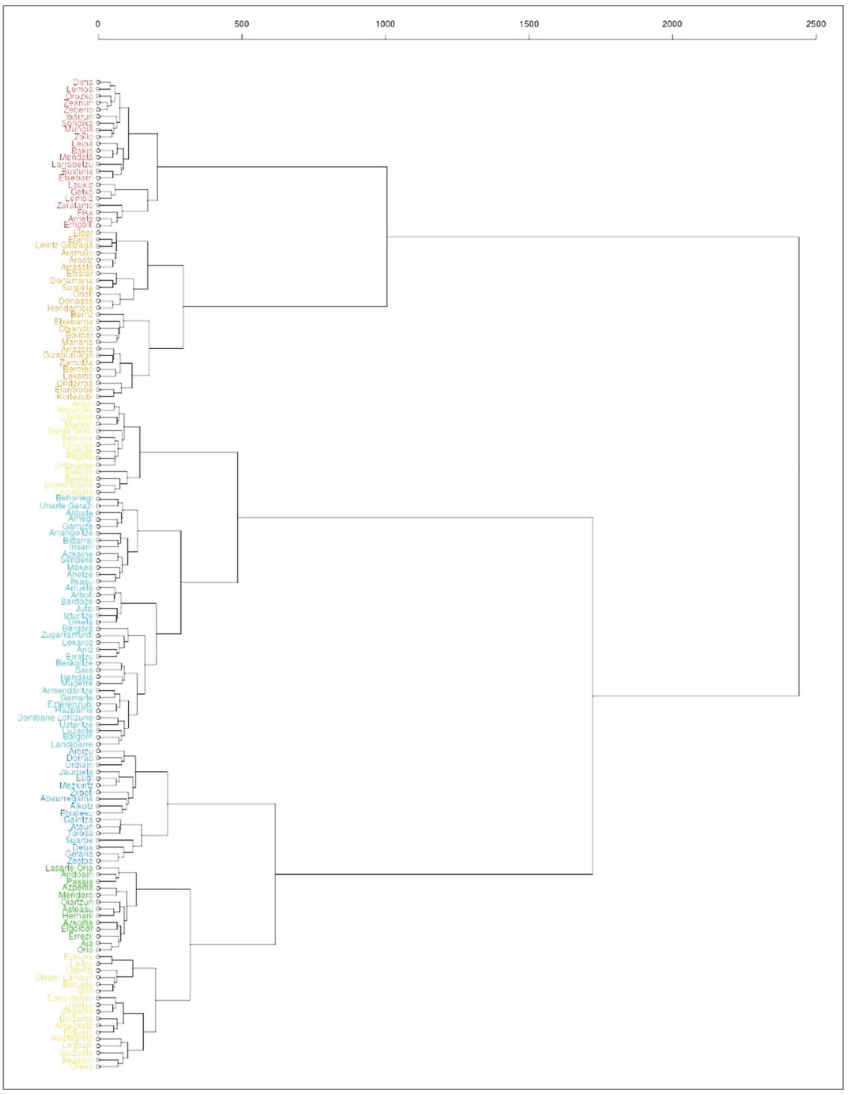
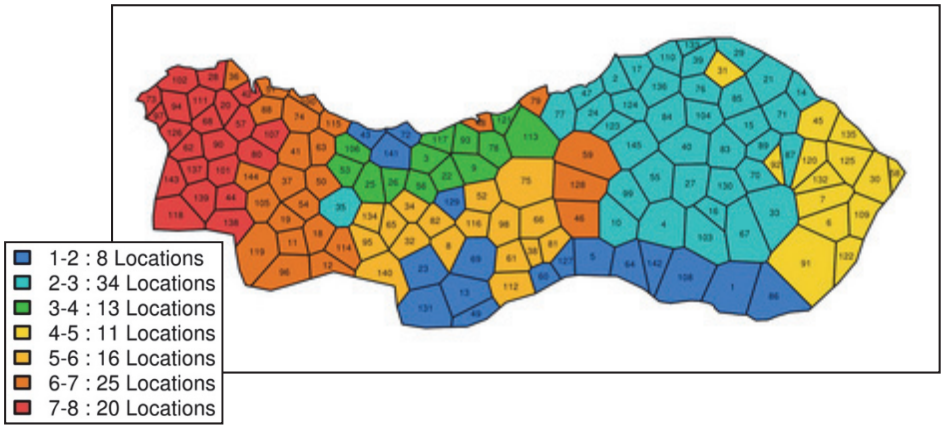


Fig. 4: Cluster analysis using orthographic data and Lv

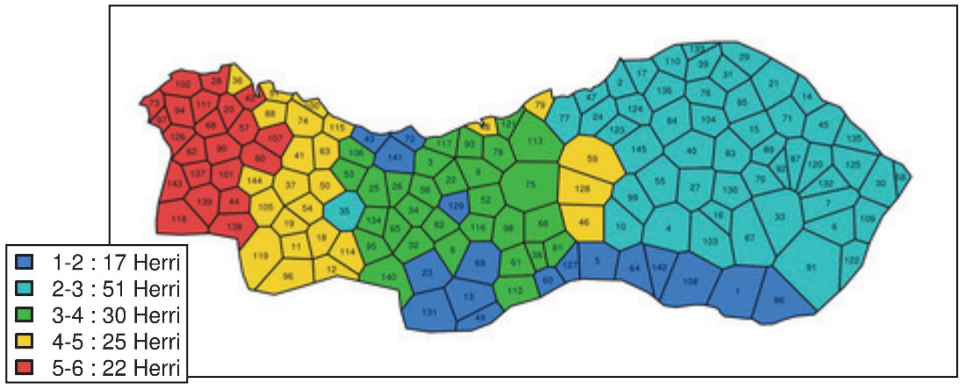
The reason that we have cut it into 7 clusters is because Basque dialects have been divided into 7 dialects since L. L. Bonaparte (1868). This division has been accepted with only slight changes and used widely since then until Zuazo's work, which, using both the isogloss system and following the way of dividing dialects in traditional dialectology, divides Basque dialects into 5 areas. Nevertheless, in this work we will consider only different partitions and their projection onto the maps. If we project the division of the dendrogram and the resulting 7 clusters onto the map (map 1) we show the distribution of each one in the geography of the Basque language.



Map 1: Cartographical representation of orthographic data using Lv-7 clusters

The map does not show a coherent dialectal structure: the majority of the areas show localities which are found outside their space; only the red and the green areas have all their localities grouped together. Light blue, dark blue, light orange, dark orange and yellow have one or more localities outside their areas.

Going up in the level of the scale in the dendrogram, and instead of taking the 250 level, if we take the 600 level we will get 5 clusters: two clusters have been grouped with others: the two oranges of map 1 have been put together (yellow color) in this map, and the blue and yellow of map 1 into blue (see map 2).



Map 2: Cartographical representation of orthographic data using Lv-5 clusters

This map shows two coherent areas (red and green areas) and three non-coherent areas (yellow, and dark and light blue). The coherent areas are in the corners of the map: on the far left and on the far right; nonetheless, we have to point out the light blue area, which is located in the eastern part, has one

locality which is found outside the main area. With regard to the non-coherent areas in the middle of the map, the odder areas are the dark blue one (which is mainly located in the southern part of the map and has localities in the northern and the central parts of the map) and the yellow one (which is mainly in the western area and has localities in the central part of the map which have historically belonged to a different dialect). Using phonetic data, it is not possible to obtain a coherent map, in which all areas include all their localities. Neither a 4 cluster partition of the dendrogram, nor a two cluster partition produces a coherent map, from the point of view of traditional dialectology.

6.2 *Cartography of the data using phonological distance*

Contrasting with the previous analysis, we display the second cluster analysis which shows the phonological distance between dialects, in this case with underlying representations and using the RIV distance. The hierarchical classification has been divided into 7, 5 and 3 clusters or dendrems. To visualize these clusters we have used different lines: the red line to divide the dendrogram into 7 clusters, the green line to divide it into 5 clusters and the blue line to divide it into 3 clusters.

The dendrogram shows a hierarchical classification of 145 localities. Cutting the dendrogram at the 500 level (red line) of the scale (0-8000) we obtain 7 clusters (see fig. 5). The localities belonging to each cluster are colored with different colors (see below). At this level only 6.25% of the variation is taken into account. Taking only this 6.25% of the variation, 7 groups (or clusters) have been formed from 100 localities. This means that there are not many differences or dissimilarities between the analyzed localities. The green line cuts at the 800 level of the scale, which is 10% of the variation; at this level 5 clusters are created. If we look at the figure and consider that the first two clusters create a new one a little above, and that the other three clusters are more or less at the same level, it would be better to cut at the 900 or the 1000 level and obtain only three clusters. Nevertheless, the three clusters cut is made at the 3300 level of the scale, which represents 41.25 % of the variation. Even though the distance between the red and green lines is small (very little linguistic distance), the distance between the green and blue line is great (bigger linguistic distance).

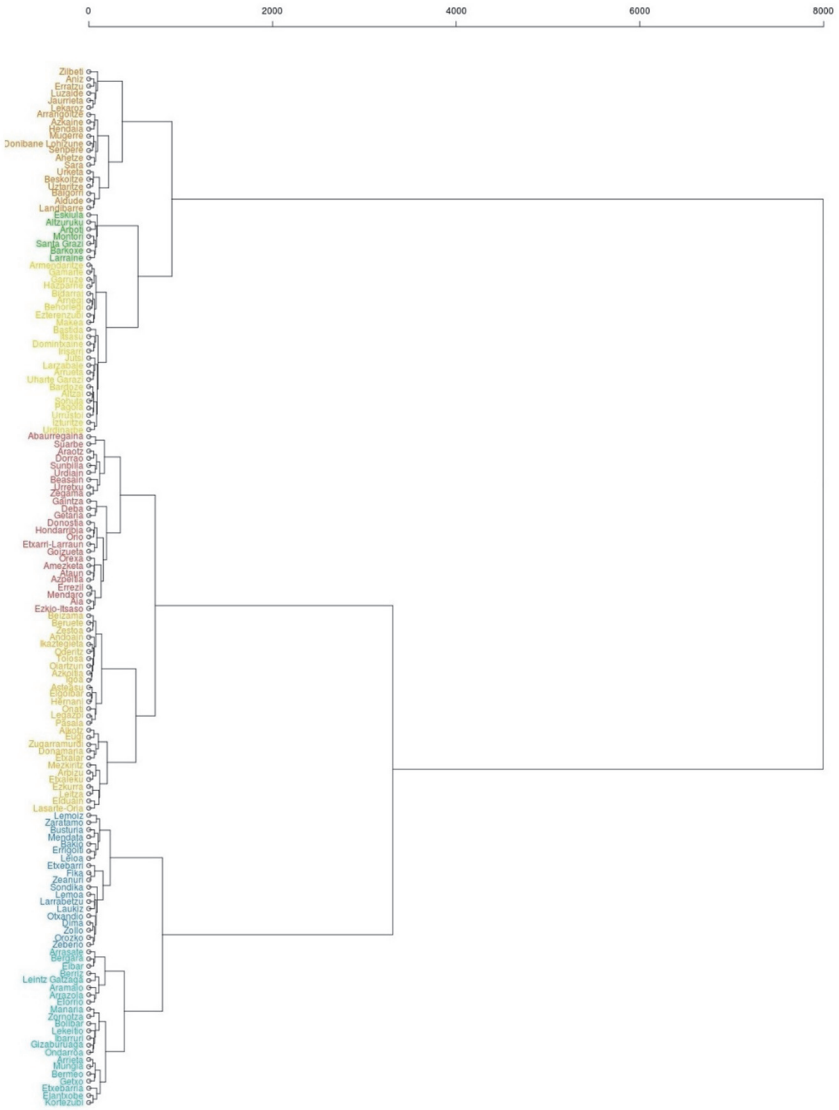
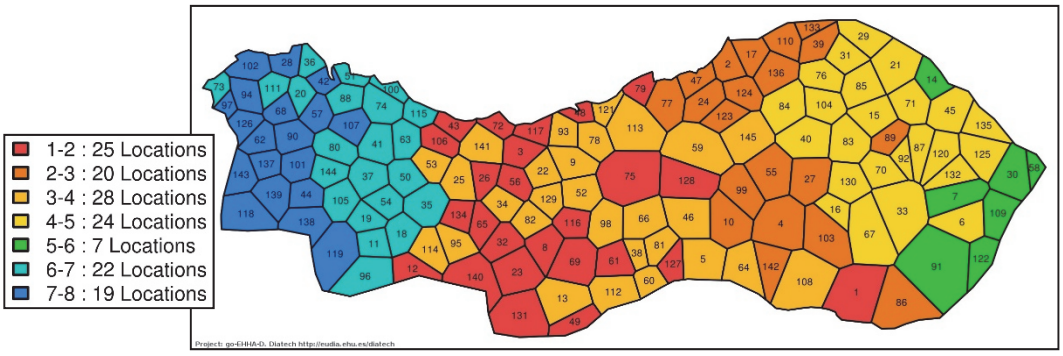


Fig. 5: Cluster analysis using underlying features and RIV distance-7 clusters

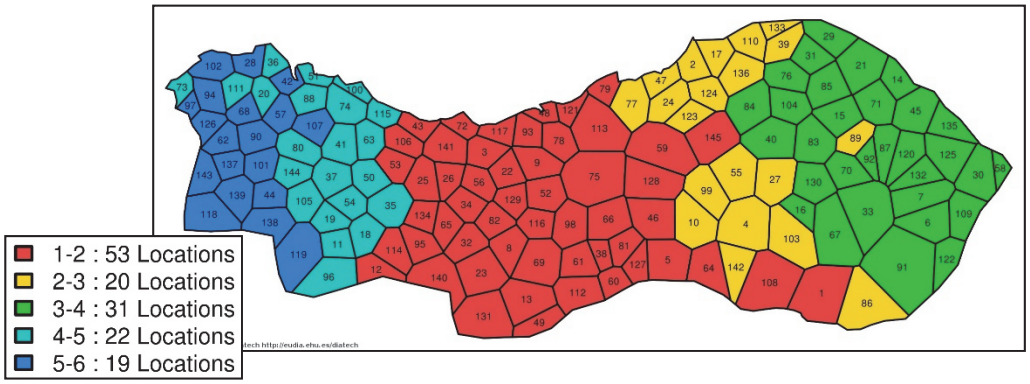
Projecting these cuts onto the maps, we can draw three maps. Map 3 has been projected by cutting the dendrogram at the 500 level of the scale (0-8000), which encompasses 6.25% of the distance from the beginning. The resulting map shows almost no consistent linguistic areas.



Map 3: the representation of 7 areas using phonological data, RIV distance and the *Ward* algorithm

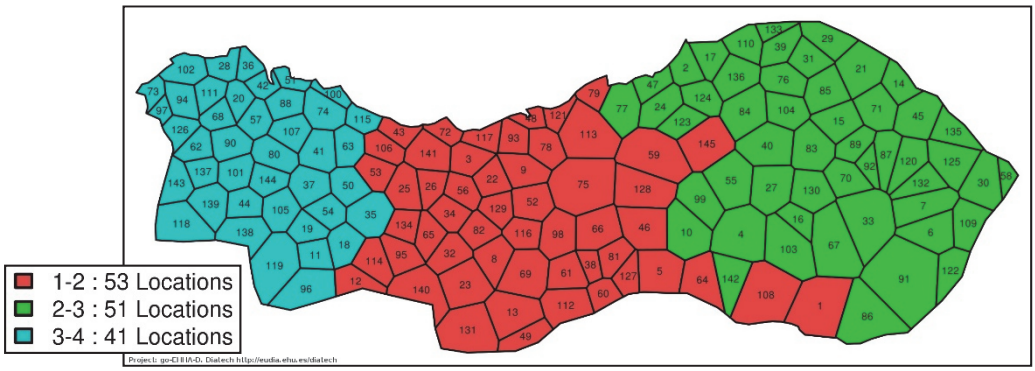
This map shows no linguistic coherence in any of the areas; not one of the colored areas has a coherent space; neither in the western part of the territory, nor in the central or the eastern parts. It is true that the blue color is found only in the western part, but the dark blue area has localities which belong to the light blue colored area; the red color is spread only in the central part, but this area is scattered with light orange-colored localities. The green color has eastern features that are located in the extreme eastern part. So, from the point of view of geolinguistic variation, we have to reject this map, because it does not fulfill the minimum requirement in dialectology, at least in traditional dialectology.

Nevertheless, departing from linguistic features but not from the geography, we can question whether it is possible to give the name of dialect areas to a sum of localities with the same linguistic features but without any geographical coherence. It is obvious that in traditional dialectology this question was unthinkable. But now we have different options to depart from (from geography or from linguistic features), more linguistic data, and powerful tools. Therefore, dialectology can go further than before.



Map 4: the representation of 5 areas using phonological data, RIV distance and the *Ward* algorithm

Map 4 (which is drawn by cutting the dendrogram at the 800 level, 10% of the differences) shows more consistent linguistic areas, especially in the green and red areas. The red area has been created with the red and the light orange areas of the dendrogram (fig. 5). It takes up the central part of the territory and encompasses 53 localities; it is the greatest and the most united area. The green area, in the eastern part, is also coherent, but has one locality in the middle of the area which does not belong to the group. The yellow group is divided into two areas and in addition has two other localities outside of its areas. In the western part, we can see the dark blue and the light blue areas: while the dark blue area has a coherent geographical structure, the light blue one has localities located inside the dark blue area.



Map 5: the representation of 3 areas using phonological data, RIV distance and the Ward algorithm

Map 5 has been drawn by projecting the result of cutting the dendrogram at the 3300 level of the scale. At this level of the scale, 42% of the distance has been taken into account. It shows three very consistent areas: the western area, the central area and the eastern area. Comparing the last two maps, we can see that the central area remains identical, no changes have been undergone, but the lateral areas of map 4 have been merged into two (the dark and light blue into blue, and the green and yellow into green).

Looking again at the dendrogram, the blue and red cutting lines (which have been used to draw maps 3 and 4) are very close to the beginning, or zero point, and include only 6.25% and 10% of the projected distance. So it seems that they cannot be used to distinguish dialects. On the other hand, the green line picks up more or less 50% of the differences or the distance. According to the scientific literature about Basque dialects, and looking at both the dendrogram and the maps, we claim that this partition can be taken as a good partition and the resulting map as a good dialect map.

Now we have the option to compare the phonetic and phonological analyses (figs. 4 and 5) and maps (maps 1 and 2 for the phonetic distance, and maps 3 and 4 for the phonological distance): maps 1 and 3 are divided into 7 areas and maps 2 and 4 into 5 areas. As you can see, the phonetic differences (fig. 4) are greater than the phonological differences (fig. 5). This is a consequence of both the different data and the different distance units that have been used. As far as the maps are concerned, the 7-group division (maps 1 and 3) shows many differences between the phonetic and phonological maps; in fact, the maps are very different. The 5-group division (maps 2 and 4) has differences in all groups. We can conclude that coming from the same starting point but using different linguistic analyses and distance units, we obtain different cartographical representations. Therefore, when analyzing linguistic distances it is very important to be explicit about which kind of data we are using (phonetic or phonological) and which distance unit we are using.

7 Discussion

In previous sections, we have seen that phonetic distance and phonological distances have to be measured differently (phonetic distance as a string distance and phonological distance as a categorical or discrete distance) and that both distances are very different. Following this, we have proposed measuring phonetic distance using the Lv algorithm and phonological distance with the RIV distance unit. It is, to my knowledge, the first time that using different linguistic distance units has been proposed to measure the phonetic representation and the underlying representation. Moreover, the only tool which supplies both measurements is the *Diatech* tool; this is why I have used this tool.

On the other hand, it can be considered pertinent to question what the most appropriate linguistic measurement is. If we consider that language is a system made up of different features, all of which have their specific position, and that dialect constitutes a linguistic system belonging to the language system, we must conclude that the most appropriate measurement is the one which takes into account all the linguistic features of the system. From this point of view, the phonological measurement fits best with linguistic theory. Despite this, I think that this standpoint is questionable and that the phonetic measurement can be defended as the most suitable one from the point of view of speech.

The use of the phonological measurement implies that no gathered data will be used to measure the distance between or among localities or varieties. Instead of this, linguistic features will be used. The verification of the existence of each feature in the system of each locality is not the subject of this contribution.

Accepting that the phonological measurement fits better than the phonetic one entails the choice of the linguistic theoretical framework to analyze the gathered data, as has been previously proposed. In any case, researchers must know that when choosing between different frameworks, the linguistic distances can be different.

8 Conclusions and future challenges

We have provided new data from the Basque dialects using data from the EHHA project, and carrying out what is until now the largest geolinguistic analysis of Basque nominal morphology, using data gathered from 145 localities.

We have presented the hierarchical classification or clustering analysis of the Basque dialects using two data types (phonetic data and phonological data) and two linguistic distances (the Lv algorithm and the RIV measurement), and showing the different structure of both data types: the phonetic representation and the underlying representation. I have used the *Diatech* tool, the only tool which provides both (the Lv and RIV) measurements.

And finally, we have proposed carrying out a linguistic analysis of data using different theoretical frameworks, and gathered data or accidental features to use linguistic systemic features to analyze geolinguistic distances between localities. We know that by using different theoretical frameworks the obtained distances can be different. In the near future, dialectology will have to take into consideration these two linguistic levels in the analysis of linguistic distances.

References

Alvarez Enparantza, José Luis "Txillardegi" & Gotzon Aurrekoetxea. 1987. *Euskal dialektologiaren hastapenak*. Bilbao: UEU. www.inguma.org.

Aurrekoetxea, Gotzon. 1995. *Bizkaieraren egituraketa geolinguistikoa* [*The geolinguistic structure of the Biscay diale ct*]. Bilbao: UPV/EHU.

Aurrekoetxea, Gotzon. 2002. Algunas consideraciones sobre la contrapregunta en las encuestas lingüísticas. In *Mélanges offerts à Jean-Louis Fossat*. Ed. L. Rabassa. Université de Toulouse II-Le Mirail, CerCLid 11/2, 57–65. http://artxiker.ccsd.cnrs.fr/docs/00/07/14/00/PDF/Contrapregunta_encuestas_ling.pdf

Aurrekoetxea, Gotzon, Karmele Fernandez-Aguirre, Jesus Rubio, Borja Ruiz & Jon Sanchez. 2013. "DiaTech": A new tool for dialectology. *Literary and Linguistic Computing*, 28(1): 23–30. DOI: 10.1093/llc/fqs049.

Bonaparte, Louis Lucien. 1868. *Carte des sept provinces basques, montrant la délimitation actuelle de l'euscara*. Londres. Standford's Geographical Establishment.

Clua, Esteve. 2010. Relevancia del análisis lingüístico en el tratamiento cuantitativo de la variación dialectal. In *Tools for linguistic variation*, 151–166. Ed. Gotzon Aurrekoetxea & J. L. Ormaetxea. ds.). Bilbao: UPV/EHU.

Diatech. http://www.eudia.ehu.es/diatech .

Euskaltzaindia. 1993. *Euskal Gramatika Laburra: Perpaus Bakuna*. Bilbao: Euskaltzaindia.

Euskaltzaindia. 2010–2013, *Euskararen Herri Hizkeren Atlasa I–V* [Linguistic Atlas of the Basque Language I–V]. Bilbao: Euskaltzaindia. http://www.euskaltzaindia.net/.

Goebl, Hans. 1978. Analyse dialectométrique de quelques points de l'AIS (italien standard valdotain provençal alpin turinois milanais). In *Lingue e dialetti nell'arco alpino occidentale. Atti del Convegno Internazionale di Torino* (*1976*), 282–294 (and maps). Eds. G. E. Clivio & G. Gasca Queiraza. Turin.

Goebl, Hans. 1981. Eléments d'analyse dialectométrique (avec application à l'AIS). *Revue de linguistique Romane* 45: 349–420.

Goebl, Hans. 1983. Parquet polygonal et treillis triangulaire. Les deux versants de la dialectométrie interponctuelle. *Revue de linguistique romane* 47: 353–412.

Goebl, Hans. 1992. Problèmes et méthodes de la dialectométrie actuelle (avec application à l'AIS). In *Nazioarteko dialektologia biltzarra. Agiriak*, 429–475. Eds. Gotzon Aurrekoetxea & Charles Videgain. Bilbao: Euskaltzaindia.

Hualde, José Ignacio. 1997. Rules vs. Constraints: Palatalization in Biscayan Basque and Related Phenomena. In *Issues in the Phonology and Morphology of the Major Iberian Languages*, 79–99. Eds. F. Martínez-Gil & A. Morales-Front. Georgetown University Press: Washington.

Hualde, José Ignacio & Jon Ortiz de Urbina, J. 2003. *A Grammar of Basque*. Berlin: Mouton de Gruyter.

Lafkioui, Mena. 2009. Analyses dialectométriques du lexique berbère du Rif. *Studien zur Berberologie/Etudes Berbères*, 4: 133–150.

Laka, Itziar. 1994 *A brief grammar of Euskara*, *the Basque language*. http://www.ei.ehu.es/p056-12532/eu/contenidos/informacion/euskara_inst_lexiko_gramatika/eu_lex_gram/adjuntos/Laka2.pdf (acc. 9 October 2014).

Martínez Areta, Mikel. 2013. Basque dialects. In *Basque and Proto-Basque*, *Mikroglottika. Minority language Studies* 5: 31–87.

Nerbonne, John & Wilbert Heeringa. 2010. Measuring dialect differences. In Theories and Methods, vol. *Within series Language and Space*, 550–567. Eds. Jürgen Erich Schmidt & Peter Auer. Berlin: Mouton de Gruyter.

Philps, Denis. 1985. *Atlas dialectométrique des Pyrénées centrales.* Thèse de Doctorat d'Etat. 2 vols. Toulouse (unpublished).

San Martin, Itziar. 1998. An OT Account of the Formation of Definite Forms in the Vizcayan Basque Dialect of Markina. *University of Maryland Working Papers in Linguistics* 7: 63–80.

Séguy, Jean. 1973. *Atlas linguistique de la Gascogne*, *Vol 6 + volume annexe*, + *matrices dialectométriques* (*phonétique diachronique*, *phonologie*, *morpho-syntaxe*, *morphologie verbale*, *lexique*). Paris: CNRS

Spruit, Marco René, John Nerbonne, Wilbert Heeringa. 2009. Associations among linguistic levels. *Lingua* 119, Issue 11, 1624–1642.

Valls, Esteve, John Nerbonne, Jelena Prokić, Martin Wieling, Esteve Clua, Maria-Rosa Lloret. 2012. Applying the Levenshtein Distance to Catalan dialects: A brief comparison of two dialectometric approaches. *Verba. Anuario Galego de Filoloxía* 39: 35–61. http://www.usc.es/gl/servizos/publicacions/revistas/verba/.

Gotzon Aurrekoetxea • Linguistic and Basque Studies Department
University of the Basque Country • SPAIN • gotzonaurre@gmail.com