MANUEL PADILLA-MOYANO

Le Dauphin project: a micro-corpus of correspondence in Lapurdian Basque of 1757

Abstract

As a result of a recent discovery, in this contribution the author presents the project of design of a micro-corpus of Basque correspondence of the 18th century. This little corpus reflects the language state of Lapurdian dialect in 1757 with the accuracy given only by spontaneous documents. Firstly, we shall explain the relevance of the finding of this documentation, and then we will proceed to a historical and linguistic contextualization. Then, we shall deal with the nature and contents of the letters, their authors and the linguistic richness of the corpus. Finally, *Le Dauphin* project will be placed in the context of other corpora, and we shall succinctly expound the works and kind of edition expected.

1. Introduction[1]

The aim of the *Le Dauphin* project is to compile and design a corpus of 18th century private correspondence. The letters will be

1 This introduction is a summarized adaptation of the dossier of the *Le Dauphin* project, recently submitted to the National Agency for Research (ANR) in France, in collaboration with our colleagues Aurélie Arcocha-Scarcia, Ricardo Etxepare, Xabier Lamikiz, Christophe Marquesuzaa and Jean-Philippe Talec.

transcribed, translated and annotated in order to facilitate subsequent linguistic, sociolinguistic, literary and historical treatments. The letters come from the ship *Le Dauphin*, which set sail from Bayonne for Louisbourg Île Royale (present-day Cap-Breton), in Canada, in April 1757. The vessel carried about two hundred letters, fifty of them written in Basque. They have recently been discovered in the National Archives of London (High Court Admiralty).

The relevance of this project is twofold. On the one hand, it collects documents which constitute a most illuminating contribution to the Basque written production of the 18th century. On the other hand, it develops a research methodology which takes a richly annotated corpus as a premise for the thorough examination of its content. The annotation will be executed by means of the TEI system, complying with the practices recommended by the *Très Grands Équipements* of the CNRS (*Centre National de la Recherche Scientifique*).

Among the benefits of this project, we firstly stress the digitalization, transcription, translation, annotation and filing of the correspondence through a computer platform. Secondly, these letters constitute a corpus that can be the object of research in the fields of history and dialectology of the Basque language, sociolinguistics, literature and history. We also intend to extend our research to further letters of similar nature from other ships whose documentation has been located in the National Archives of London.

From a linguistic point of view, these records present the great advantage of attesting a particular language state of the Lapurdian dialect –the basis of literary Basque. More specifically, the letters, dated 1757, represent a dozen of varieties in Labourd. As they are private documents produced with a communicative goal –mostly by humble people–, they reflect the dialect actually spoken in the mid-18th century more accurately than any other known text. Consequently, *Le Dauphin* corpus provides us with an original source of dialectal data that can be compared to other contemporary data.

From a sociolinguistic point of view, the corpus will make possible the study of alphabetisation processes in the context of the complex diglossic situation in the continental Basque Country in this period. It will allow to analyse the acquisition of writing skills, the

degree of grammatical codification of the Basque language, as well as the distribution of these phenomena in the Lapurdian society. As for the literary study, these letters offer a privileged access to the expression of the intimate and the daily in the practice of the epistolary genre in Basque, including aspects such as the discursive organization of the letter and the use of some stylistic formulae. The historical interest is also unquestionable. *Le Dauphin* is, therefore, a multidisciplinary project constructed from an exceptional corpus, not only in the framework of Basque studies, but also in the French and European contexts, with the potentiality of becoming the germ of future discoveries.

2. Contextualizing Le Dauphin corpus

2.1. The discovery of Le Dauphin

During the 17th and 18th centuries France and Great Britain rivalled each other for the control of the Atlantic Coast of Canada and its thriving fisheries. After the signing of the Treaty of Utrecht and the subsequent cession of Newfoundland (Terre-Neuve) and Acadia to the British Crown (1713), France colonised Île Royale (present-day Cap-Breton Island) and founded its capital, Louisbourg, which quickly became an important harbour. With the Seven Years' War (1756–1763) the situation in the French possessions of North America became complicated, since the very limited Atlantic emigration could not grant to the Kingdom of France the real control of these territories. In 1758, one year after the arrest of the ship *Le Dauphin*, Île Royale passed to British hands.

In 2007, during the course of his research on the transatlantic trade networks in 17th and 18th centuries, the historian Xabier Lamikiz discovered the Basque correspondence of the corsair ship *Le Dauphin* in the Archives of the High Court Admiralty in London. The letters, dated between February and April of 1757, were sent to

relatives and friends residing in Louisbourg Île Royale. They never arrived at their destination, since the English captured *Le Dauphin*. This documentation attests to the written use of the Basque language, which, as stated above, is used to convey the expression of intimate or daily topics in a spontaneous way. As will be pointed out (§ 2.2), this finding may be considered exceptional in the context of Bascology.

In the 18th century Basque seamen played an outstanding role in the Atlantic world. Sailors from Biscay and Guipuscoa, i.e. Basque coastal regions in Spain, enrolled in ships to Spanish America, whereas those from Labourd went to the overseas possessions of the Kingdom of France. Basque-origin toponyms in the coasts of Canada are evidence of the foregoing, as well as the existence of Basque pidgins in Iceland and Canada (Deen 1937/1991; Bakker et al. 1991). Following Lamikiz (2010: 66-67), there were three main reasons why Basque merchant ships tended to be manned by local mariners:

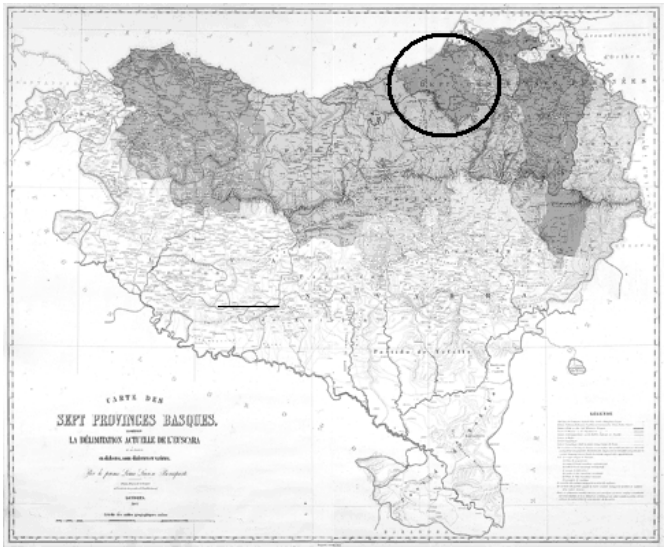
> First, Bilbao's merchant fleet was not large enough to exhaust the local supply of seamen [this is extensible to Lapurdian harbours such as Bayonne]. Secondly, while it was essential for captains to speak Spanish or French this did not apply to crews, most of whom spoke only Basque. […] Thirdly, their common geographical origin became an additional source of unity for the crew, and a tendency seen elsewhere in Europe.

2.2. Linguistic context

The written language of such letters is Lapurdian Basque (*lapurtera*), considered as the basis of the literary language. Despite the small size (see Map 1) and limited demographic weight of the region, Lapurdian was the first of the so-called *literary dialects* and it has outperformed other dialects in the literary production throughout the history. At this point, we should refer to a terminological difficulty, due to the specificity of Bascologist tradition. Although all Basque dialects have, to a certain extent, a written tradition, the distinction between *literary* and *non-literary* dialects is deep-rooted since Louis Lucien Bonaparte established a dividing line in 1863. On the one hand, we have Biscayan (*bizkaiera*), Guipuscoan (*gipuzkera*), Lapurdian and Souletin (*zuberera*), from the West to the East, all of them enjoying literary

consideration; on the other hand, there are the non-literary dialects: North and South High-Navarrese (*goinafarrera*) and Western and Oriental Low-Navarrese (*behenafarrera*). Nevertheless, present-day Basque dialectology has questioned such a distinction and Bonaparte's basis for the dialectal division itself.

The Basque textual history begun with the first printed book: Bernard Dechepare's *Linguae Vasconum primitiae* (Bordeaux 1545). According to Lakarra / Urgell (2008), from that time until 1750 (the symbolic beginning of modern *euskara*), the Basque written tradition comprises of around five hundred pages for the peninsular dialects. For the continental dialects, however, the corpus consists of around five thousands pages, most of which correspond to Lapurdian. Concerning the nature of the printed corpus of Basque, until 1900 –to establish an arbitrary limit– nearly 90% of the works are religious texts. Thus, Lapurdian is a relatively well-documented dialect, on which most of the written Basque tradition has been constructed, but the nature of the recordings is largely homogeneous.

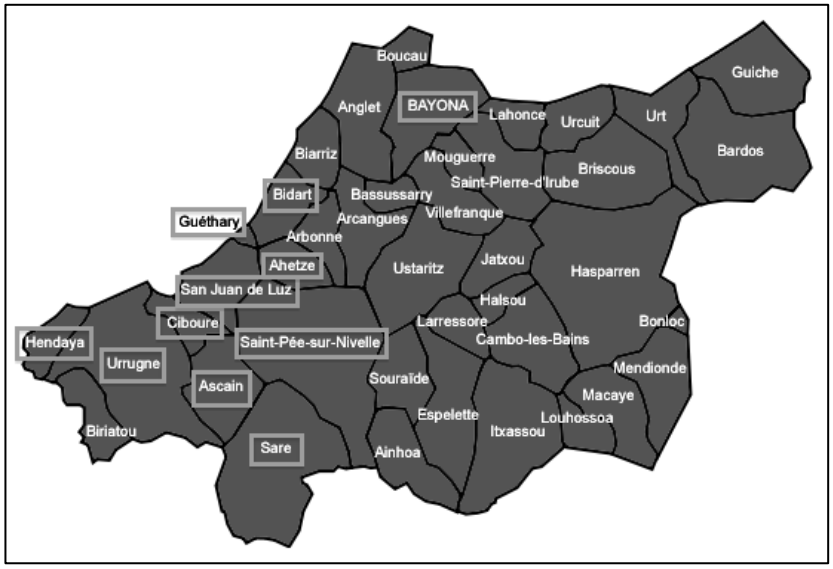


Map 1. *Carte des sept provinces basques montrant la délimitation actuelle de l'Euskara et sa division en dialectes, sous-dialectes et variétés* (London, 1863). The Lapurdian dialect is delimited by the circle.

It is precisely here that the importance of the letters from *Le Dauphin* lies. In languages such as English or Spanish the finding of several dozens of letters from the 18th century could be considered somewhat anodyne –at least from a linguistic point of view. In the case of Basque, however, it is a very different issue. In fact, these missives provide us with an authentic portrait of the language state of 18th century. Lapurdian. About ten of its local varieties are represented in the corpus (see Map 2). Moreover, the language of the letters differs greatly from the one found in religious texts: their spontaneous character –the humbler the authors were, the more *authentic* the language–, their communicative purpose and the subject matter raised make this little corpus an exceptional witness to 18th century Basque. In section 5.1. we will make reference to other correspondence-based corpora.

Point of origin	*Number of letters*
Sare / Sara	9
Guéthary / Getaria	8
Saint-Jean-de-Luz /	6
Ciboure / Ziburu	5
Saint-Pée / Senpere	5
Hendaye / Hendaia	4
Ascain / Azkaine	3
Urrugne / Urruña	2
Bidart / Bidarte	2
Ahetze / Ahetze	2
Bayonne / Baiona	1
Unknown origin	3
Total	*50*

Table 1. Origin of the correspondence.



Map 2. Map of Labourd, nowadays integrated in the French Department of Pyrénées-Atlantiques. The points of origin of the letters carried by the ship Le Dauphin in 1757 are highlighted. Note that the Oriental area of the country is not Lapurdian-speaking, hence the representativeness of this dialectal corpus.

We must point out that the letters from the *Le Dauphin* corpus are not the only Basque correspondence known. Among other compilations, the 16th century spy missives found in the General Archive of Navarre (Floristán 1993; Satrústegui 1993) must be mentioned, as well as the cross-border official correspondence between the valleys of Soule (Basque Zuberoa) and Roncal, dated ca. 1616. Moreover, there are various sets of letters which respond to administrative needs of councils on both sides of the French-Spanish border. Even though they are private documents, they were written by persons from a high socio-cultural status. Unlike all of them, the correspondence of *Le Dauphin* constitutes a corpus which lends itself to dialectological, sociolinguistic and literary research: both because of the language register used and because of the multiplicity of authors and the geographical variety represented by them.

3. Correspondents' typology, epistolary uses and subject matters

When reading the letters from *Le Dauphin*, we realise that most of the authors of Basque letters are humble, hard-working people, who express their emotions, wishes and fears. The most obvious and conspicuous characteristic of the missives is the fact that they are written in Basque, which sheds light on the alphabetisation processes in the continental Basque Country. Indeed, the discovery of *Le Dauphin* correspondence comes to confirm Oyharçabal's thesis (2001), which postulates the existence of an alphabetisation system in Basque in the 17th and 18th centuries, gradually relegated to the lower strata of Basque society as the French language –obviously north of the Pyrenees– was entering the most cultivated circles. This Basque-speaking basic education took place in the *Petites Ecoles* (*Ororen eskolak*, lit. 'schools for everyone'), and was intimately close to the Catholic Church, which used it as a catechetical means –it must be kept in mind that the dioceses of the Kingdom of France took special care in fighting the Huguenot by means of indoctrination. In any case, such a schooling contrasts with the virtual absence of any kind of Basque-speaking alphabetisation south of the Pyrenees. That is exactly what we conclude from the comparison between Southern and Northern Basque seamen's letters. In Lamikiz' words (2010: 124):

> Despite the fact that most people in the Basque Country had Basque as their sole language, it was chiefly an oral culture and the use of written *euskara* was extremely rare [in the South of the Pyrenees]. That is why a note in Basque in the margin of one of [the Guipuscoan seaman] Luis de Echevarria's letters is such an important piece of evidence.

On the contrary, the correspondence from *Le Dauphin* shows an effective, habitual use of written Basque in the province of Labourd, only a few miles to the North. As pointed above, in the 18th century the sociolinguistic situation in the Basque provinces of France, and more specifically in Labourd, corresponds to an extremely complex diglossia. Unlike in the modern-day Labourd, the vast majority of the

population was Basque-speaking, many of them monolingual speakers. Furthermore, there was Gascon Occitan, a Romance neighbour of the northern dialects of Basque. For the non-monolingual Basque, the introduction of French was detrimental to the knowledge of Gascon Occitan. In addition, the most elevated strata knew also Latin, and Spanish was not ignored near the border.

Thus, in such a linguistic environment the correspondence from *Le Dauphin* proves the acquisition of writing skills in Basque in the heart of the Lapurdian society. We want to stress the proportion of women among the writers: wives, mothers or sisters sign thirty of the fifty letters, i.e. 60%. Although this percentage is obviously due to the fact that most of the addressees are men, we should not underestimate it: it is a manifestation of the high degree of alphabetisation among Lapurdian women, even among those who come from a low social status.

	Women	Men	Total	Women's share
Number of writers	30	19	49	61.2
Number of words	7,417	2,968	10,385	71.4

Table 2. Writers' gender.

Obviously, the virtual non-existence of a linguistic standard in this period –in spite of a fluctuating tradition– prompts every correspondent to resort to his local variety and to use a spelling system which reflects the absence of a unified standard; consequently, almost every letter presents different spelling rules. On the other hand, the existence and use of written uses and formulae, likely transmitted in those *Petites Ecoles*, is well demonstrated. The majority of the letters follows a clear structure including: i) the date; ii) an opening containing some standard expressions; iii) an account of all kind of events; iv) a more or less routinary farewell. These are some of the most utilized formulae, all of them subject to variation, transcribed in present-day spelling:

(1) *Hartzen dut libertate zuri bi lerroren eskribatzeko*
'I take the liberty of writing you a couple of lines'

(2) *Familia guzia kausitzen gara osasun perfeit batian eta desiratzen nuke zuria ere hala balitz*
'All the family is in perfect health, and I would wish that yours is so too'

(3) *Ez dut faltatu nahi izan presenteko okhasione hau zuri aditzerat eman gabe…*
'I did not want to lose the present occasion without communicating to you that…'

(4) *Ez dut zuri zer gaztiga baizik gelditzen naizela zure (aita/ama/espos…) fidela*
'I do not have anything else to tell you, except that I remain your faithful father/mother/husband…'

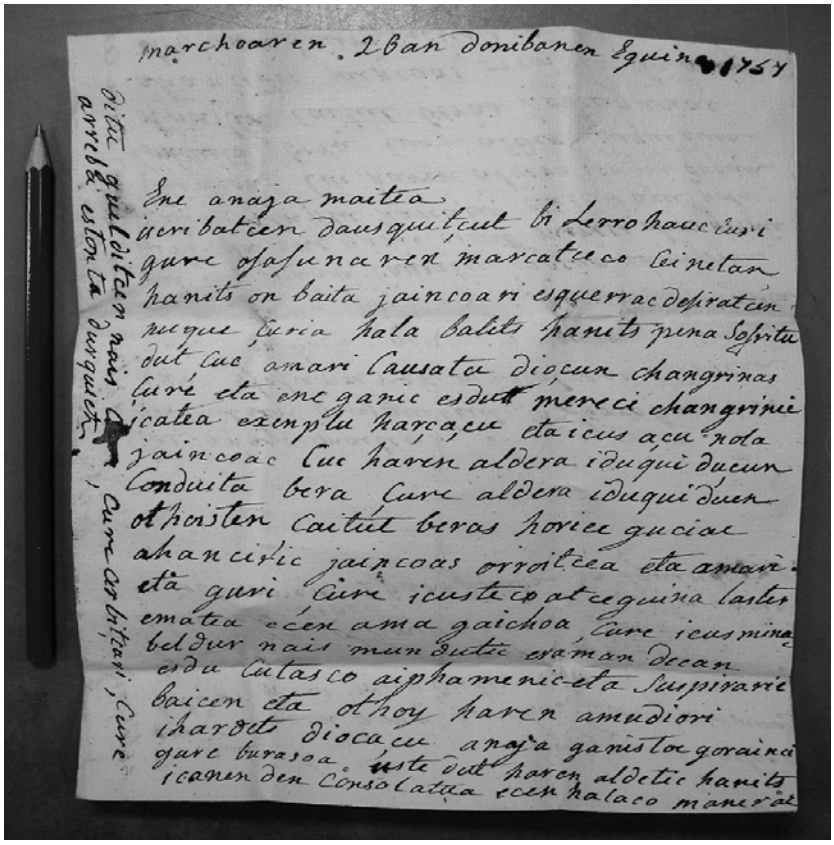


Figure 1. A letter dated the 26th of March of 1757 in Saint-Jean-de-Luz (Donibane-Lohitzune). Picture by Xabier Lamikiz.

Some of the letters show a certain degree of diastratic variation: several illiterate correspondents must have dictated their letters. Additionally, some of them openly admit some degree of ignorance, even though they largely manage to write what they wanted to express: *Ene haur maitea, enaquinan nola eman mila saspi ehunac* 'My dear daughter, I did not know how to put the nineteen seventy [= the year in numbers]'. In the opposite extreme, some missives must have been written by someone with a perfect command of the most elevated registers of Basque and familiarized with the reading of ascetic works, as the use of stylistic and rhetoric figures or the calligraphic *decorum* indicates.

Concerning the contents of the correspondence, the relations of the goods sent to or received from Louisbourg abound: usually mentioning canned food, drinks, clothes, shoes and the like. Economic affairs have a privileged place, too: the payment of debts and the bills of exchange (*letra chanjiac*) are very usual:

(5) *Monsieur Haranchipic othoisten çaitu erraiteas Samatchico semeari eztuela horren letraric batere içan; lettra chanja bat içan duela horren partes, bainan ezdela oraino pagatua, harçen baitu ungui goardatuco tuela, harceco esperanca oray baduela, ezdaquiela cer guerthatuco den. Monsieur Beauvassin Arracholacoac oray esperanca emaiten duela. Gauça bera erranen dioçu Martin Canderatzi, Chotilen semeari: ezduela Monsieur Haranchipic errecibitu oraino diruric, oray esparançan dagoela.*

'Monsieur Haranchipi begs you to say to the son of Samatchi [oikonym] that he has not received any letter from him; that he has received a bill of exchange from him, but it has not yet been paid, and he takes it in order to keep it carefully, now he has hopes of receiving the payment, but he does not know what will happen, and now Monsieur Beauvassin of La Rochelle gives him hope. You will say the same to Martin Canderatz, Chotil's son: that Monsieur Haranchipi has not yet received any money, and now he holds out the hope.' (Martin d'Etchart. Sare, 15th March 1757).

It is possible to find even a detailed report of accounts for a business left in Labourd or, as in the following sample, some kind of malicious information:

(6) *Erranen darotçut çure ontassuna guibelat hari della, çure ama andrea gai jçan gabes; eta asqueneco sasoina galdu du, harrias porrocaturic. Hortic jujatuco duçu çure ahalas lagundu behar duçulla çure ama andrea, baldin nahi baduçu çure ontasuna conserbatua jçan dadin, jauna ene iloba.*

'I shall tell you that your personal assets are decreasing, since your mother is not able [to administer it]; and she has lost the crop, ruined by the hailstone. Hence you will judge that you have to help Madam your mother by all means, if you want your goods to be kept, my dear nephew [Sir my nephew].' (Martin Larralde, Sieur de Bastidaguerre. Senpere, 16th March 1757).

Seven Years' War broken out, references to the tricky situation of the moment are present in almost every letter:

(7) *Hemen ez da guerla hotsic baicin, baitugu herrico soldadoac Bayonan, eta marinelak cortsuan dabilza eta cerbait eguiten dute, eta bertce cenbait Erregueren cerbitzuan dire.*

'Here there is nothing but war rumour; we have the soldiers of the village at Bayonne, and the sailors work as privateers and they do something, and some others serve the King'. (Marittipo de Subiet. Azkaine, 16th March 1757).

Most of the time, the correspondent fervently wishes the addressee to come back as soon as possible, even though the most cautious among them ask for patience until the situation changes, as this mother who begs her daughter to wait:

(8) *Badaquiçu orobat ume batec badubela obligacionea buraso baten obeditceco, beras hortaracots othoits eguiten darotçut neure ahal gucias guerla hunec dirabeino hor egoteas, eta guero, baquea eguiten den pontutic, etçherat erretiratceas; horra cer othoits dudan çuri eguiteco, eta desiratcen nuque entçuna banints ene othoisean.*

'Likewise, you know that a child has the obligation to obey a parent; therefore I beg you with all my might to stay there while the war persists and then, from the moment in which peace will be made, to come back home; here is the plea I make to you, and I would wish to be heard in my plea.' (Maria Dihitx. Senpere, 1st March 1757).

Likewise, we can read the account of somewhat sordid events:

(9) *Segur da aditu içan tucula hemengo berri tristiac: ehaile gaisçoa preso harturic Jondone Laurendy egunian, 1755co urthean, Parisat eramadute, Sabat dorreco çena norbaitec hilçea dela medio, eta by hilabete badu jendec daraçatela bidean heldu dela libro bere etcherat, baignan oraino esta agueri gaisçoa, eta Ostaleriaco premua urthe berian Jondoni Jauni inguruan presso harturic Toulonnen da condenaturic galleretarat seculacots, adisquideac ongui mellaturic ez urcatçeco. […] Eta itxasoan beçala leihorrean ere fortunac arribatçen dire: Betrieneco jaun gaisçoa malobran hari dela lur peça bat gainera eroriric lehertu jçan da joan den udan, eta oren baten buruco hil confesaturic.*

'I am sure you have heard about the sad news from here: the unfortunate weaver, captured in Saint Laurent's day of year 1755, has been taken to Paris because of the murder of the late owner of Sabat Tower, and for two months people have said that he is coming back home free, but the poor weaver has not yet appeared. And the first-born son of Ostaleria, captured in the same year around Saint John's day, is in Toulon, sentenced to galley for life, his friends having interceded to avoid the hanging. […] And on land as well as on sea some incidents can happen: the last summer, when the unhappy sir of Betrienea was doing community works, a landslide fell over him and crushed him, and he died in an hour, after confession.' (David Borrotra. Ahetze, 27th March 1757).

Finally, correspondents often relate their difficulties; in the following sample a widow writes:

(10) *Chagriñac eta persequçioneac içatu tut famillarecin, non hauçitan hari bainais. Egia da presentean aphur bat osasuna piçhca bat badudala: suqhar langitac qitatu nau; ene edaria da tiçana, arno batere gabe.*

'I have suffered sorrows and persecutions from my family, in such a way that I am in litigation. It is true that at the present time I am feeling healthier: the slow fever has left me; herb tea is my only drink, no wine at all.' (Cathalin Marie Berrogain. Bayonne, 15th March 1757).

4. Linguistic interest

As a general consideration, familiar missives, because of the dialogue that they establish between writer and addressee, create favourable conditions for the colloquial uses of the language, especially when an unskilful correspondent writes the letter. Thus, our documentation attests many of the features denoting a weak command of the written code. In fact, poorly-read writers often show characteristics such as: i) traces of the pronunciation in the spelling; ii) hesitating morpheme-limits and/or agglutination of words; iii) approximate orthography (Martineau / Tailleur 2012: 294). Therefore, the attested language undoubtedly reflects the spoken Lapurdian more accurately than any other kind of writing does.

In addition, *Le Dauphin* corpus spells out a remarkable diatopic variation within the Lapurdian dialect itself: about ten local varieties covering most of the Lapurdian-speaking area. The fact that the correspondence involves 50 letters written by people from 10 places confers an extraordinary relevance upon the finding: the representativeness of the corpus is in direct proportion both to the number of writers and to the plurality of geographical origins. Obviously, a discovery of 50 letters written by the same person –even written by 50 people from the same town– could not represent a dialect in the way that the correspondence from *Le Dauphin* does. On this matter, we take into consideration the comparison between different kinds of archaeological findings, which illustrates the application of the information theory to the localization of texts of unknown origin (Reenen et al. 2009: 21–24).

The nature of the texts favours the apparition of some phonological changes rarely attested in the classic writers of this time, because such authors are subject to a writing tradition –even if it is a feeble one-; hence their somewhat conservative choices. As a result, while reading the letters from *Le Dauphin* we get the impression of a modern phonological appearance, with the occurrence of some phenomena such as:

- Anti-hiatus tendencies. First, the insertion of *b*, formulated as [*ua, ue → uba, ube*]: *trabailuan → trabailuban* ‘working’ (lit. *in work*), *datatua → dadatuba* ‘dated’, *zinduela → zindubela* ‘that you had (it)’. Second, the insertion of *j* after *i*, running with any other vowel: *berriac → berrijac* ‘news’, *abiatu → abijatu* ‘to start, to depart’, *guztiek → guztijek* ‘all (everyone) [ERG]’, *diozu → dijozu* ‘you do it to him/her’, *bi(h)otz → bijotz* ‘heart’, *batistio → batistijo* ‘baptism’ (unattested variant) or *amudiuaren → amudijuaren* ‘of the love’.
- The emergent fall of intervocalic fricative consonants, which starts with *d*: *baditugu → baitugu* ‘we have them’, *comoditate → comoitate* ‘occasion’; and continues with *g*: *gastiga → gastia* ‘to advise, to inform’, *nagusi → nausi* ‘boss’.
- Verbal contractions: *gastiga iezadazu → gastiadazu* ‘give me advice’ (lit. *advise me*), *errezebitu ditut → errezebitu tut* ‘I have received them’.
- Some astonishing archaisms such as the occurrence of *egin* ‘to do’ as an auxiliary verb outside the western dialects of the language or the use of subjunctive conditional clauses, both combined in the following example: *ahal badagizu* ‘if you can [SUBJ.]’.

As is well known, Basque verbal morphology exhibits an exuberant degree of variation, and this correspondence shows unexpected forms in the Lapurdian of 1757 and, occasionally, even unattested forms. Thus, particular mention must be made to a long missive entirely written in *noka*, one of the three allocutive moods of the language — apart from its three habitual arguments (absolutive, dative and ergative), the Basque verb is able to incorporate also a reference to the addressee, which implies both gender differentiation and register choices. In accordance with pragmatic rules, the *noka* mood is a familiar form of address between women; for obvious reasons, such discourses are extremely rare in the historical corpus of Basque. Furthermore, the allocutive moods are undergoing a significant process of recession in present-day *euskara*, even more pronounced in the case of *noka*.

As regards the lexicon, the domestic topics of the correspondence are reflected in a quantity of everyday-life words, many of which show a very low frequency in old texts. The corpus *Le Dauphin* contains terms concerning weave names or products from the fisheries, such as *ziriku* (a kind of silk), *firlango* 'hessian' (for a sack), *maripolisa* (certain type of jacket of which the first attestation in *euskara* comes thus one century earlier (cf. fr. *pelisse*), or *harbi pastea* 'paste of fish spawn'. Likewise, the letters provide us with some *hapax legomena* such as *aitaroxi* 'grand father' or *amaroxi* 'grand mother' (historical Lapurdian *aitatxi* and *amatxi*, respectively), which come to increase the richness of Basque vocabulary related to family.

Finally, the letters, because of their nature, offer a large amount of onomastic data. In addition to the expected anthroponyms, there are hypocorisms and nicknames such as *Mutxil* 'Guy', *Mothela* 'The Stammerer' (lit. *The Slow*), *Xotila* 'The Astute' or *Saindua* 'The Saint' (probably in an ironical way). We must also emphasize the importance of oikonyms, such as *Naussianea* ← *Nagusi* 'Boss' (house)' or *Patinenea* ← *Matin* 'Martin's' or *Mutchurdinenea* 'The Old Maid's'. With reference to other place-names, among the inevitable toponyms of Quebec there is a plethora of variants for Louisbourg or Niganiche, and we can find exonyms for places like Lisbon, once mentioned as a Spanish city: *Espanijarat, Lisibona eraten dioten portu batera* 'to Spain, to a harbour called Lisibona'; Pleymouth, which appears in the inessive case: *Anguellesec harturic Plemuan da* '(he) is in Pleymouth, captured by the English'; or the Island of Guernsey, also in the inessive: *hartubac dire Garnesein* '(they) are captured in Guernsey'.

5. Towards the design of a micro-corpus

5.1. Le Dauphin *among other corpora*

In view of the fact that all the letters are dated between February and April of 1757, *Le Dauphin* is a synchronic corpus since, as we have previously mentioned, it captures an image of the language state of the Lapurdian dialect in the mid-18th century. This said, it is evident that the letters also offer precious data for diachronic researches. With regard to the size, the Basque correspondence from *Le Dauphin* contains about 10,385 words; hence we prefer the term *micro-corpus* rather than *corpus*. Its length is significantly small compared to other Basque corpora already compiled, among which the following are the most relevant:

- Corpus Arakatzailea [The Browser Corpus], which contains most of the literary works of historical Basque, from 1545 to the 20th century. With material from 495 books, altogether 11.9 million words, its platform allows for searches by genre, author, period or dialect.
- XX. Mendeko Euskararen Corpus Estatistikoa [Statistic Corpus of the 20th Century Basque]. Its 4,600,000 words constitute a representative sample of the Basque written during the last century.
- Erreferentziazko prosa [Prose of Reference]. Over 25 million words of modern prose from 2000 to 2007, both from books and press. It is a closed corpus complemented by the recent *Dinamic Corpus of Reference*, which covers the period 2007–2011.
- The Eroski – Consumer corpus, which feeds on the magazine for the consumers of the Basque chain of supermarkets Eroski, thus implementing a language policy of equality between Spanish, Catalan, Galician and Basque −i.e. all the official languages in Spain except the Aranese (Occitan on the Aran Valley); consequently, the magazine is published in such

languages. The Basque version of this corpus, with 2,365,000 words, reflects the modern standard or *euskara batua* −unified Basque− in its written register.

- The Goenkale Corpus, named after the homonymous series that the Basque Television, ETB1, broadcasts non-stop since 1994. This corpus collects 2,418 episodes, with more than 800,000 dialogues and 11 million words. Its interest lies exactly in such mass of dialogues, which represents the spoken register of the modern standard.
- Lexikoaren Behatokiaren Corpusa [The Corpus of the Observatory for the Lexic]. This project is under way, and envisages the compilation of a large, balanced, lemmatized and annotated corpus fed by different media. At the end of 2011 it had over 18 million words.

Apart from its dimensions, *Le Dauphin* project responds to more specific needs since, as pointed above, our corpus is limited to epistolary texts written in a single dialect at a given moment. In a sense, *Le Dauphin* corpus could be compared to Bourciez' compilation (Aurrekoetxea / Videgain 2004), which collects 150 versions of the translation of the *Parable of the Prodigal Son* into the local varieties of 150 municipalities in the Basque Country in France, made in 1895. In spite of the fact that they are not spontaneous texts, such translations provide us with detailed dialectological information of northern Basque and, as in the case of our correspondence, they afford us the image of a language state at a given moment.

In any case, *Le Dauphin* project will benefit from the study of other correspondence-based compilations, such as the CEEC (Corpus of Early English Correspondence), with five daughter corpora among which the CEECE (Corpus of Early English Correspondence Extension) covers the eighteenth century. The methodological choices of such projects, led by T. Nevalainen and H. Raumolin-Brunberg, will be a reference for our work. Maybe in a way more similar to our project, we could mention the Scotch-Irish emigrant letters (Montgomery 1995), which have offered valuable information to diachronic research. Other interesting examples for us are the works on the evolution of the Acadian dialect of French based on familiar

correspondence (Martineau / Tailleur 2010), as well as most of the papers included in Dossena / Del Lungo Camiccioti (2012).

	Le Dauphin	*CEEC**	*CEECE**
words	10,385	2,597,795	2,219,422
collections	1	96	77
letters	50	5,961	4,923
writers	49	778	308
time span	1757	1410–1681	1653–1800

Table 3. Le Dauphin, the CEEC and the CEECE in numbers. *Data from Nevalainen & Raumolin-Brunberg (http://www.helsinki.fi/varieng/domains/CEEC.html).

5.2. A micro-corpus susceptible of extension

After the finding of the documentation from the ship *Le Dauphin* by Xabier Lamikiz in 2007, we hold out the hope that further discoveries of similar Basque correspondence will take place. This is a very reasonable expectation, given that Lapurdian sailors and fishermen have historically played an extraordinary role in the transatlantic context, and more specifically in the coasts of Canada. Furthermore, such a hope is reinforced by the fact that, among the letters effectively carried by *Le Dauphin*, we have found a missive dated in Saint-Pierre de Martinique, which would have been sent to Labourd. Someone in the High Court Admiralty classed this letter in error together with the documentation from *Le Dauphin*.

At the moment our corpus, defined as the Basque correspondence of 1757 carried by *Le Dauphin*, is a closed set of texts, but future discoveries could change the configuration of our project. In the case of an eventual extension of the corpus, we refer, *mutatis mutandi*, to Laitinen 2002. Since a large extension of the period attested may not be expected, however, such an extended corpus of Lapurdian transatlantic correspondence will remain, essentially, a synchronic one. To a certain extent, future discoveries

could widen the degree of diatopic variation, since some Lapurdian varieties are absent from *Le Dauphin* correspondence, and we lack linguistic witnesses of the Lapurdian people residing in Canada. If this were not the case, our corpus would become a precious instrument for elucidating the controversial topic of the evolution of Basque varieties in America.

5.3. Digital edition of Le Dauphin *corpus*

Le Dauphin project has a double partnership: on the one hand, the Center for Research on Basque Language and Texts (IKER UMR 5478), headed by the CNRS and the Universities of Bordeaux 3 and Pau; on the other hand, the Computer Laboratory of the University of Pau (LIUPPA), which incorporates thirty researchers belonging to two teams: T2I (Treatment and adaptation of spatial, temporal and thematic images) and MOVIES (MOdelling, VIsualisation, Execution and Simulation). Their cooperation is essential in a project aiming at the compilation and design of a reliable, well-annotated corpus, which needs the collaboration between researchers both from Computing and from Social and Human Sciences.

Since our main aim is the dissemination of a corpus not yet designed, the editing work becomes very important. In that regard, different versions are envisaged, one of them closer to a diplomatic edition and the other one conceived of in a critical and interpretative way. Additionally, a third version of the texts in standardized spelling is also expected. All over the process, philological criteria will govern our work, in order to achieve the highest degree of faithfulness to the text.

In addition to making possible the immediate access to the digitalized original documents, a computer platform will facilitate the insertion of the critical apparatus and the display of various informations on the text through XML (*eXtensible Markup Language*) tagging and annotation. Among such XML metalanguages we opt for the TEI system (*Text Encoding Initiative*), which offers a range of possibilities adapted to any editorial need in philological work, from

the material description to the eventual insertion of variants. In sum, we propose an enriched digital edition able to satisfy any expectation.

5.4. Tagging and annotation by means of the TEI system

Following habitual conventions in the annotation of manuscripts compilations, in the correspondence of *Le Dauphin* every letter becomes a corpus unit; each unit will be assigned a TEI header containing a complete description of established parameters. TEI headers have a hierarchical, tree-like structure which can be divided in four elements susceptible of grouping a variable number of items: i) <fileDesc>, a description of the electronic document; ii) <encodingDesc> for an account of the coding rules; <profileDesc>, which describes the text, regardless of its material form; and iv) <revisionDesc>, a history of the eventual changes in the electronic document.

Some of the features subject to annotation are: description of the manuscript, author, addressee, date, brief summary of the document, editor, description of the project, researchers, responsible institutions, editing criteria or textual typology. Tagging guidelines start from the segmentation of the linguistic units (*vid*. *TEI,* 15.1): characters (*c*), morphemes (*m*), words (*w*), phrases (*ph*) and clauses (*cl*). It is possible to assign a linguistic interpretation (15.4), as well as to annotate the linguistic features to be examined, whether they are phonological, morphological or syntactic items. In addition, we also consider the annotation of some socio-linguistic variables, such as the writers' gender, origin or socio-economic status. Likewise, the luxuriant variety of spelling will need a specific treatment throughout the annotation process, in order to facilitate eventual researches in orthographic systems.

In the framework of Basque Studies the annotation of manuscripts through the TEI system is not abundant. The *Bonaparte Archive* (University of Deusto) is the most relevant project. If we quote its authors, "the most complete –if not the only complete– collection of Basque dialectology. Every dialect, sub-dialect and even varieties of the Basque language are reflected in these manuscripts"

(Pagola, n.d.). The analyzer of the *Bonaparte Archive* allows for a semi-automatic morpho-syntactic analysis of the language. Furthermore, we could mention Fernández / Gómez 2009, even if their object of study is the evolution of Spanish in Biscay.

6. Conclusion

We have presented the project of compilation and design of a micro-corpus of correspondence in old Lapurdian Basque, based on a recent textual discovery. The documentation will benefit from a multidisciplinary research team led by the Center for Research on Basque Language and Texts (CNRS UMR 5478 IKER). Concerning methodological choices, we shall take into consideration the background of other correspondence-based corpora, such as the *Corpus of Early English Correspondence* (CEEC), even if their volume is very different. We bear in mind the tagging of both grammatical features and sociolinguistic variables. Since both the experiences on manuscript annotation and the design of micro-corpora do not abound in Basque Studies, the partners of the project *Le Dauphin* seek to adopt a rigorous research methodology in the field of Corpus Linguistics amenable to becoming a point of reference for similar cases not only for bascologists, but also for the compilation and design of corpora in other languages.

Acknowledgements

I would like to thank my colleagues Aurélie Arcocha-Scarcia (University of Bordeaux 3), Ricardo Etxepare (CNRS), Xabier Lamikiz (University of the Basque Country), Bernard Oyharçabal (CNRS), Jean-Philippe Talec (CNRS) and, in a special way, Charles

Videgain (University of Pau). I also thank Manuel Padilla-Cruz (University of Seville) for his linguistic advice and commentaries to this paper. Needless to say, all errors and omissions are mine.

References

Aurrekoetxea, Gotzon / Videgain, Charles 2004. *Bourciez-en "Recueil des idiomes de la région Gasconne" bildumako euskal testuak*. <http://artxiker.ccsd.cnrs.fr/docs/00/08/05/48/PDF/Haur_prodigoa_testuak.pdf>

Bakker, Peter / Bilbao, Gidor / Deen, Nicolaas / Hualde, José I. 1991. *Basque pidgins in Iceland and Canada*. San Sebastian: University of the Basque Country.

Deen, Nicolaas 1937/1991. Glossaria duo vasco-islandica. *International Journal of Basque Linguistics and Philology*. 25/2, 321-426.

Dossena, Marina / Del Lungo Camiciotti, Gabriella (eds) 2012. *Letter Writing in Late Modern Europe*. Amsterdam: John Benjamins.

Elspaß, Stephan 2012. The use of private letters and diaries in sociolinguistic investigation. In Hernández-Campoy, Juan M. / Conde-Silvestre, Juan C. (eds) *The Handbook of Historical Sociolinguistics*. Oxford: Wiley-Blackwell, 156-169.

Fernández, Patricia / Gómez, Sara 2009a. La edición enriquecida y en paralelo para el estudio del patrimonio documental vizcaíno. *Oihenart: cuadernos de lengua y literatura*. 24, 87-98.

Fernández, Patricia / Gómez, Sara 2009b. Un ejemplo de reutilización del patrimonio documental vizcaíno: el análisis lingüístico por medio de la marcación en XML. *Oihenart: Cuadernos de lengua y literatura*. 24, 99-119.

Floristán, José M. 1993. Conflictos fronterizos, espionaje y vascuence a finales del siglo XVI. 20 documentos inéditos. *Fontes Linguae Vasconum: Studia et documenta*. 63, 177-219.

IKER UMR 5478 (n.d.). *Correspondance en basque du bateau Le Dauphin* (report presented to the ANR) <http://www.iker.cnrs.fr/-correspondance-le-dauphin-1757-.html?lang=fr>

Isasi, Carmen 2004. La edición en el siglo XXI: Nuevos retos universitarios. In Jacob, Inés (ed.) *Capacidades humanizadoras de las TIC*. Bilbao: University of Deusto, 381-388.

Lakarra, Joseba / Urgell, Blanca 2008. *Euskararen historia* [Unpublished manuscript]. Vitoria-Gasteiz: University of the Basque Country.

Laitinen, Mikko 2002. Extending the Corpus of Early English Correspondence to the 18th Century. *Helsinki English Studies*. 2.<http://blogs.helsinki.fi/hes-eng/files/2011/03/HES_Vol2_Laitinen.pdf>

Lamikiz, Xabier 2008. Basque Ship Captains as Mariners and Traders in the Eighteenth Century. *International Journal of Maritime History*. 20/2, 81-109.

Lamikiz, Xabier 2010. *Trade and Trust in the Eighteenth-Century Atlantic World: Spanish Merchants and Their Overseas Networks*. Woodbridge: Royal Historical Society / Boydell Press.

Leech, Geoffrey 1993. Corpus Annotation Schemes. *Literary and Linguistic Computing*. 8, 275-282.

Martineau, France / Tailleur, Sandrine 2010. Correspondance familiale acadienne au tournant du XX^{e} siècle: Fenêtre sur l'évolution d'un dialecte. In Neveu, Franck *et al.* (eds) *Congrès Mondial de Linguistique Française*. Paris: Institute de Linguistique Française, 291-303.

Montgomery, Michael. 1995. The Linguistic Value of Ulster Emigrant Letters. *Ulster Folklife*. 41, 26-41.

Nevalainen, Terttu / Raumolin-Brunberg, Helena (n.d.). *Corpora of Early English Correspondence*. <http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html>

Nurmi, Arja (ed.) 1998. *Manual for the Corpus of Early English Correspondence Sampler CEECS*. Helsinki: Department of English of the University of Helsinki. [file text]. <http://icame.uib.no/ceecs/index.htm>

Oyharçabal, Bernard 2001. Statut et évolution des lettres basques durant les XVII^{ème} et XVIII^{ème} siècles. *Lapurdum*. 6, 219-287.

Pagola, Rosa M. et al. 2006. Edición digital para el análisis lingüístico automático del corpus Bonaparte. In Villayandre, Milka (ed.) *Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística*. León: Universidad de León, 1429-1441.

Pagola, Rosa M. *et al.* (n.d.). *Bonaparte Archive* <http://bonaparte.deusto.es/includes/index.php?&id_menu=1&id=29&id_idioma=eu&id_idioma=en>

Raumolin-Brunberg, Helena / Nevalainen, Terttu 2007. Historical Sociolinguistics: The Corpus of Early English Correspondence. In Beal, Joan C. / Corrigan, Karen P. / Moisl, Herman L. (eds) *Creating and Digitizing Language Corpora: Diachronic Databases*. Basingstoke: Palgrave Macmillan, 148-171.

Reenen, Pieter van / Rem, Margit / Wattel, Evert 2009. The Localization of Medieval Texts of Unknown Provenance. In Dossena, Marina / Lass, Roger (eds) *Studies in English and European Historical Dialectology*. Bern: Peter Lang, 19-66.

Satrústegui, Juan M. 1993. Relectura de los textos vascos de espionaje del siglo XVI. *Fontes Linguae Vasconum: Studia et documenta*. 64, 443-475.

TEI. 2007. *TEI: P5 Guidelines*. <http://www.tei-c.org/Guidelines/P5/>

Uitti, Karl D. et al. (n.d.). *The Princeton Charrette Project.* <http://www.princeton.edu/~lancelot/ss/>