



HAL
open science

The First Annotated Corpus of Historical Basque

Ainara Estarrona, Izaskun Etxeberria, Ricardo Rikardo, R. Etxepare, Manuel Padilla-Moyano, Ander Soraluze

► **To cite this version:**

Ainara Estarrona, Izaskun Etxeberria, Ricardo Rikardo, R. Etxepare, Manuel Padilla-Moyano, Ander Soraluze. The First Annotated Corpus of Historical Basque. Digital Scholarship in the Humanities, 2021, 10.1093/lc/fqab066 . hal-03505658

HAL Id: hal-03505658

<https://hal.science/hal-03505658>

Submitted on 31 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The First Annotated Corpus of Historical Basque

Journal:	<i>Digital Scholarship in the Humanities</i>
Manuscript ID	DSH-2021-0030
Manuscript Type:	Full Paper
Date Submitted by the Author:	22-Feb-2021
Complete List of Authors:	Estarrona, Ainara; University of the Basque Country Faculty of Computer Science, Department of Computer Languages and Systems Etxeberria, Izaskun; University of the Basque Country Faculty of Computer Science Etxepare, Ricardo; IKER-CNRS Padilla-Moyano, Manuel; University of the Basque Country - Alava Campus Soraluze, Ander; University of the Basque Country Faculty of Computer Science
Keywords:	digital humanities, historical corpora, natural language processing (NLP), diachronic syntax, history of Basque

SCHOLARONE™
Manuscripts

The First Annotated Corpus of Historical Basque

[Here authors]

[And here affiliations]

This paper presents the elaboration of a morphosyntactically annotated diachronic corpus of Basque, and the first results obtained in the processing of historical varieties of this language with computational techniques. The corpus size is around one million words, expanding from the fifteenth to the mid-eighteenth century and encompassing the most significant written production in all historical dialects. Morphosyntactic tagging allows for systematic searches at different levels of complexity; additionally, a rich set of metadata enables searches based on socio-historical criteria too. This is not only the first tagged corpus of historical Basque but also a means to improve language processing tools by analyzing historical varieties more or less distant from the present-day standard language. Moreover, this project aims to set a model for further works in historical corpora of Basque and inform similar projects on other languages.

Keywords: digital humanities, historical corpora, natural language processing (NLP), diachronic syntax, history of Basque

1. Introduction

In other languages, there are different resources for lexical, morphological, or syntactic searches in old texts, e.g. the *Penn Parsed Corpora of Historical English* (Kroch and Taylor, 2000; Kroch *et al.*, 2004, 2016), the *Tycho Brahe Corpus* [historical corpus of Portuguese] (Galves *et al.*, 2017), the *Icelandic Parsed Historical Corpus* (Wallenberg *et al.*, 2011) or the *Parsed Old and Middle Irish Corpus* (Lash, 2014). All these corpora are provided with morphologic and syntactic annotation. However, no appropriate instruments of this type have ever been developed for Basque and the project we present in this paper, [PROJECT NAME], wants to tackle such a challenge.

The only searchable database for Basque historical linguistics today is the *Classic Basque Corpus*, located at the Institute of the Basque Language of the University of the Basque Country. For different reasons that will be explained in Section 4.1, the *Classic Basque Corpus* is a tool of very limited value for linguists. Its existence, with its usable format, is nonetheless crucial for the feasibility of our project, as it provides a first operational basis to work on.

1
2
3 36 The main goal of our project is the systematic diachronic study of several grammatical
4
5 37 features of Basque in the light of the current theoretical discoveries. The second stated goal is
6
7 38 to create a comprehensive annotated historical corpus of Basque that comprises both part-of-
8
9 39 speech and syntactic annotation as well as a rich set of metadata structure. This way, it will be
10
11 40 possible to search the annotated corpus by words, lemmas, grammatical categories, by
12
13 41 sequences of grammatical categories, and by specific structural configurations (such as
14
15 42 relatives, correlatives, finite subordination, and others) as well as by publication year, author
16
17 43 or place of publication, thanks to the mentioned metadata. Hence, the preparation of resources
18
19 44 for the syntactic exploration of the historical corpus in Basque will allow us to achieve the
20
21 45 first goal since the syntactically annotated historical corpus can show us how certain syntactic
22
23 46 structures have evolved over the centuries. The final output of our project will be a search
24
25 47 interface to browse the corpus.

26
27 48 This article focuses on the second goal of the project, that is, the construction of the
28
29 49 morphosyntactically annotated historical corpus of Basque. Having developed a lot of NLP-
30
31 50 based technology for the analysis of standard Basque, we now have to tackle the text
32
33 51 normalization, considering that once ancient or dialectal texts are normalized, the developed
34
35 52 NLP tools could be applied for the linguistic analysis of the corpora. We will explain the
36
37 53 implementation and evaluation of computational techniques for corpus normalization. Our
38
39 54 methodology is partially defined by previous research on the automatic normalization of
40
41 55 historical and dialectal Basque texts (Etxeberria *et al.*, 2014; Etxeberria *et al.*, 2016) — we
42
43 56 must stress, however, that the method was tested and evaluated in an experimental
44
45 57 application. Thus, a relevant contribution of this paper is testing the method in a real scenario,
46
47 58 considering periods and dialects never before analyzed through NLP tools.

48
49 59 After having explained the starting point for the project of an annotated historical
50
51 60 corpus, in Section 2 we summarize the state of research in historical Basque linguistics and
52
53 61 indicate the need for annotated tools for the study of the historical syntax of Basque. In
54
55 62 Section 3 we explain and describe the reference corpus. Section 4 proposes an overview of the
56
57 63 tasks involved in the normalization of the corpus, as well as the subsequent steps leading to
58
59 64 the creation of a searchable database. After methodological issues, in Section 5 we present the
60
61 65 results obtained so far applying computational methods and Natural Language Processing
62
63 66 tools to historical Basque varieties. Then we provide some details about the search interface
64
65 67 under construction in Section 6. Finally, in Section 7 we formulate the conclusions and point
66
67 68 out future orientations.
68
69 69

2. On Basque Historical Linguistics

Perhaps because of its peculiar status as a non-Indo-European language and as an isolate in the European linguistic domain, much of the literature on the Basque language has been concerned with issues of language ancestry and origin. This type of research has venerable predecessors in Basque studies, starting with the Basque-Iberian hypothesis proposed by Schuchardt (1907), and his later Afroasiatic hypothesis (1914), or the Basque-Caucasian one (Uhlenbeck, 1924; Lafon, 1951–1952).

The second half of the twentieth century witnessed the establishment of a rigorous approach to Basque historical linguistics, thanks to the ground-breaking work of Luis Michelena, which provided a clearer picture of the earlier stages of the language, and therefore a basis for any further comparison (for an overview see Trask, 1997). This systematic work was based on both a careful examination of the Basque textual corpus and the application of the comparative method to the rich internal variation of Basque. Michelena's work was crowned by important achievements: we have at present, for instance, a valuable representation of the Proto-Basque phonological system (Michelena, 1977 [1961]), and of the phonological changes that led to the present-day Basque system. Lafon's essential work must be mentioned too (1944, 1998), even though the vicissitudes of his time did not favor the consolidation of something like a school following his wake. Given the scientific context of the time, Michelena's and his associates' and later specialists' work has been largely oriented toward the establishment of earlier stages of Basque — where *earlier* should be understood as earlier than the grammar directly attested by the textual corpus starting in the fifteenth century. Afterward, some of the most important works have aimed to go beyond the stages of the Proto-Basque reconstructed by Michelena (Lakarra, 1995, 2005, 2006).

This forms the basis of the common understanding of *historical linguistics* in the Basque context (for a comprehensive overview, see Martínez-Areta, 2013). The evolution of the Basque language has been scarcely examined in those aspects which are directly observable in written records and make Basque typologically special in the European context. We know today, based on serious historical reconstruction, that Proto-Basque — corresponding approximately to the reconstructed Basque of the Roman period — showed typological properties that starkly diverge from present-day Basque. To begin with, Proto-Basque was very probably a VSO language, not an SOV language, one of the basic typological properties of present-day Basque (Gómez and Sainz, 1995). Ergativity, another outstanding property of Basque, may also be a relatively recent innovation, whose

1
2
3 104 morphological trace can be identified today in one of the locative particles (Lakarra, 2006).
4
5 105 The complex finite forms in Basque, with their long sequences of agreement and diathetic
6
7 106 affixes, a feature of the language that has fascinated European scholars for decades, can be
8
9 107 shown to be relatively recent: the emergence of number agreement marking in the auxiliary
10
11 108 can to some extent be tracked on the existing corpus, as can other instances of agreement,
12
13 109 such as dative agreement. The evolution of the determiner system, which evolved from the
14
15 110 demonstrative one (Manterola, 2015), can be directly examined, to a great extent, on textual
16
17 111 records, and the same goes for the category of morphological number, absent from earlier
18
19 112 stages of Basque. That is, many of the crucial steps leading to the special typological
20
21 113 properties of Basque can be accessed and examined in the light of the increasingly rich textual
22
23 114 evidence, and with the methods and analytic resources of modern diachronic linguistics. This
24
25 115 also implies carefully studying the contact-induced changes concerning the surrounding
26
27 116 romance languages, as well as considering the existing dialectal variation.

27
28 117 Nevertheless, no systematic study of this sort has ever been undertaken for Basque
29
30 118 historical syntax. The project [NAME] aims to address this challenge. It has two main goals:
31
32 119 i) to carry out a systematic diachronic study of some grammatical features of Basque, and
33
34 120 ii) the elaboration of a morphosyntactically annotated historical corpus of Basque. The corpus
35
36 121 will also have a rich metadata structure that will allow searches by dialect, time, or socio-
37
38 122 historical information. This will allow for complex and rich searches, e.g. by word, lemma,
39
40 123 part-of-speech, and also particular syntactic structures that we also annotate (relatives,
41
42 124 correlatives, etc.) and combinations of them. In this paper, we address this second goal in
43
44 125 particular. On the other hand, we can say that [NAME] is an interdisciplinary project,
45
46 126 involving experts in the field of historical linguistics and natural language processing
47
48 127 (henceforth NLP) and the collaboration between [NAME] and [NAME]centers.
49
50 128

46 129 **3. The Corpus**

50 131 *3.1 Design and Decisions*

51 132
52
53 133 We decided to establish a philologically-reliable corpus covering most of the textual
54
55 134 production between the fifteenth and mid-eighteenth centuries. This is, on the one hand, the
56
57 135 minimal span that includes regular attestations for all Basque dialects and, on the other hand,
58
59 136 it also divides Archaic and Old Basque from Early Modern Basque.
60

1
2
3 137 Usual issues conditioning the building of historical corpora are manual processes of
4
5 138 compilation, accessibility of materials, and availability of different types of texts (Claridge,
6
7 139 2009). In terms of representativeness, understood as the relationship between the corpus and
8
9 140 the body of the language, the ideal tends to the inclusion of ‘texts from as many different
10
11 141 categories of writing and speech as resources will allow’ (Hunston, 2009: 160). As far as
12
13 142 Basque is concerned, the corpus should be representative of all dialects with written tradition.

14
15 143 Regarding balance, which refers to the internal composition of the corpus, things are
16
17 144 even more complicated. The body of historical Basque is in many respects unbalanced.
18
19 145 Table 1 shows the figures for printed books in the four so-called historical *literary dialects* of
20
21 146 Basque. Until 1750, which roughly coincides with the closure point of our corpus, a
22
23 147 significant majority of the texts correspond to two dialects in the French side of the Basque-
24
25 148 speaking area: Labourdin and Souletin; moreover, nearly 90% of these published texts are
26
27 149 religious ones. Thus, if the goal of achieving balance among regional varieties might be
28
29 150 challenging for Germanic languages (cf. Claridge, 2009), then in the case of Basque it
30
31 151 becomes a chimera.

32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	Printings		First editions		Original works	
	1545–1749	1750–1879	1545–1749	1750–1879	1545–1749	1750–1879
Biscayan	2	74	2	24	1	13
Guipuscoan	8	187	7	69	3	44
Labourdin	53	206	25	53	12	22
Souletin	10	47	6	8	2	4

153 **Table 1**

154

155 Given the limitations that affect the body of historical Basque, in many cases, our criterion
156
157 has necessarily tended to the inclusion of most available materials, because, on the other hand,
158
159 size matters too. At the time being, we are creating a corpus of around one million words.
160
161 Considering the issues of all types associated with the written past of languages, especially in
162
163 cases like Basque, this size is considered acceptable for a historical corpus (cf. Claridge,
164
2009).

161 In a few words, on the one hand, our corpus is representative of all historical Basque
162
163 dialects, and it includes texts of as many genres as possible; on the other hand, balance has
164
165 been sacrificed for the sake of size. Indeed, ‘any attempt at corpus creation is therefore a
166
167 compromise between the hoped for and the achievable’ (Nelson, 2010: 60). Finally, let us

1
2
3 165 mention that our design has adopted the complete-text approach instead of selecting samples
4
5 166 of texts.

6
7 167

8 168 *3.2 Periodization and Texts*

9
10 169

11
12 170 In a commonly accepted periodization (see Gorrochategui *et al.*, 2018), the Basque textual
13
14 171 history is divided into four main periods: i) Archaic Basque, from the fifteenth century to
15
16 172 1600; ii) Old Classical Basque, from 1600 to 1745; iii) Early Modern Basque, from 1745 to
17
18 173 the end of the nineteenth century; and iv) Late Modern Basque, from the end of the nineteenth
19
20 174 century to 1968, year in which standard Basque was launched. Our project encompasses
21
22 175 Archaic and Old Classical periods; another project that addresses both Early and Late Modern
23
24 176 Basque is also underway [NAME]. The divide between Old Classical Basque and Early
25
26 177 Modern Basque, which defines the outer limit of our corpus, is mediated by the decisive
27
28 178 influence of Manuel Larramendi's works, in particular his grammar (*El Imposible Vencido*,
29
30 179 1729) and his *Diccionario Trilingüe* (1745).¹

31
32 180 Archaic Basque is represented by the four main texts from this period: i) Dechepare's
33
34 181 poems compilation, the first printed book in Basque (1545), written in Eastern Low Navarrese
35
36 182 dialect; ii) Lazarraga's manuscript, which includes a pastoral novel in Alavesse dialect
37
38 183 (ca. 1562); iii) Leizarraga's Calvinist translation of the New Testament into a sort of Northern
39
40 184 Basque *koine* (1571); and iv) *Refranes y Sentencias*, a compilation of sayings in Biscayan
41
42 185 dialect (1596). We also include most of the shorter texts edited by Michelena (1964) and
43
44 186 Sarasola (1983), as well as those compiled by Satrústegui (1989).

45
46 187 As for Old Basque, our corpus includes the bulk of the written production of this
47
48 188 period, which is considerably larger than that of the Archaic Basque — the reader can get an
49
50 189 idea looking at the chronological list in the web *Klasikoen Gordailua* [= The Classics'
51
52 190 Repository].² Additionally, the above-mentioned collections by Michelena, Sarasola, and
53
54 191 Satrústegui include many texts of the Old Basque period. Despite its less-than-fully
55
56 192 satisfactory philological quality, the inclusion of Satrústegui's compilation is justified because
57
58 193 of the heterogeneous and rare dialectal materials it contains, some of them corresponding to
59
60 194 High Navarrese, a dialect poorly represented in the central periods of the history of Basque.

195

56 196 **4. Methodology and Tasks**

57
58 197

1
2
3 198 To create a morphosyntactically tagged historical corpus of Basque, we propose three main
4
5 199 steps. First, the corpus must be designed, compiled, and prepared. Then, the texts must be
6
7 200 normalized, to make the automatic analysis by NLP tools possible — i.e. we need to convert
8
9 201 the historical texts into standard Basque. After this normalization process, we will analyze and
10
11 202 annotate the texts using the parsers developed by the [NAME] team. Below we shall detail
12
13 203 each of these steps.

14 204

15 205 *4.1 Compilation, Revision, Correction, and Spelling-Update*

16 206

17
18
19 207 Our starting point is the *Basque Classics Corpus* of the Institute for the Basque Language of
20
21 208 the University of the Basque Country,³ with a size of 11.9 million words. It includes texts
22
23 209 from the fifteenth to the early twentieth centuries pertaining to all the historical dialects of
24
25 210 Basque and presents a rich variety of genres in both verse and prose (drama, translations,
26
27 211 original literary narrative, religious prose, sayings, official and personal correspondence
28
29 212 among others). Therefore, it offers a solid basis for analyzing the evolution of historical
30
31 213 dialects. Unfortunately, this existing historical corpus, with all its obvious potential, cannot be
32
33 214 directly used as it is for academic purposes:

34 215 a) Firstly, this database only allows searching by word. Given the varied morphological
35
36 216 exponence that most of the grammatical words have in Basque, this corpus is of
37
38 217 limited value, and of zero value for scholars non-specialized in Basque.

39 218 b) Secondly, the corpus is highly unreliable, as it has been compiled on the basis of the
40
41 219 voluntary contribution of scholars from different backgrounds and of unequal
42
43 220 competence. Moreover, some of the transcriptions are based on outdated editions, and
44
45 221 therefore reproduce errors that have been corrected in later ones. The transcriptions
46
47 222 may also be incorrect due to typing errors, or to the wrong interpretation of the
48
49 223 original forms. The use of this resource for linguistic exploitation, therefore, requires a
50
51 224 thorough revision of the texts.

52 225 c) Thirdly, the most important part of the corpus, corresponding to the literary
53
54 226 production, is transcribed in present-day standard orthography, a choice that keeps the
55
56 227 original morphological forms, but results in the lack of relevant phonological
57
58 228 information.

59 229 As mentioned above (Section 3), the first task has been the definition of a reference corpus of
60
230 historical Basque. We have compiled the most significant written records from the fifteenth to
231 the eighteenth centuries, a period in which all the historical dialects of Basque are

1
2
3 232 represented. The texts have been selected according to three criteria: i) representativeness of
4 233 period and linguistic variety, ii) existence of reliable editions, and iii) sociolinguistic
5 234 relevance. When possible, we have prioritized prose texts rather than verse ones, as our
6 235 project focuses on syntactic structures.

7
8
9
10 236 Our reference corpus prioritizes those texts that have reliable critical editions. In this
11 237 regard, good critical editions exist for all texts of the sixteenth century, for most of the texts of
12 238 the seventeenth century, and to a lesser extent also for eighteenth-century texts. For those
13 239 works which lack a critical edition, we use the first edition as the reference text. As has been
14 240 explained, in the *Basque Classics Corpus* the transcriptions of the texts are based on editions
15 241 of different quality, and not all of them ensure the level of reliability that we need. For this
16 242 reason, a crucial phase of our project is the conscious revision of the texts. At this point we
17 243 compare the transcriptions with their facsimiles (and/or with reliable critical editions) and,
18 244 depending on the quality of each one, we opt for one of the following choices: i) to correct the
19 245 transcript, or ii) to undertake a new one. This philological task is highly time-consuming,
20 246 albeit necessary for the sake of basing our corpus on reliable versions of historical texts.

21
22
23
24
25
26
27
28
29 247 The main criterion for this philological work is the modernizing of the spelling — not
30 248 to confuse with the adoption of present-day standard Basque orthography. The goal is to
31 249 update the spelling of the texts so that NLP tools can work with them, always without
32 250 distorting their phonological features. That is, while in the *Basque Classics Corpus* the
33 251 spelling of the texts has been updated according to the rules of present-day standard Basque,
34 252 in our corpus the modernizing of the spelling preserves the phonological shape of each text.
35 253 For instance, Eastern Basque dialects have a set of aspirated plosive phonemes *ph*, *th*, *kh* that
36 254 is not represented in the spelling system of standard Basque. As a result, most of the
37 255 transcripts of Eastern Basque texts in the *Basque Classics Corpus* do not reflect aspirated
38 256 plosives, although in original texts they are usually represented by special graphemes.

39
40
41
42
43
44
45
46 257

47 48 258 *4.2 Normalization of the Corpus*

49
50 259

51 260 Most NLP tools have been conceived to process journalistic texts written in the present-day
52 261 language. Of course, the features of modern standardized languages and those of historical or
53 262 dialectal texts are by no means the same. To begin with, standard varieties i) constitute the
54 263 base for dictionaries and grammars, ii) are written in a standard spelling according to which
55 264 most texts are published, and iii) have a large amount of text available in an electronic format
56 265 that can be used to develop NLP tools. Conversely, historical and/or dialectal texts present the

1
2
3 266 opposite situation, so standard NLP tools cannot be directly used. For this reason, corpus
4 267 normalization is an essential step toward (automatic) morphosyntactic analysis of historical
5 268 and/or dialectal texts.

8 269 We have compiled a rich corpus of texts from different periods and varieties, and of
9 270 very high quality from a philological and linguistic point of view. At this point, let us
10 271 underline that the need to cope with extremely diversified grammatical materials adds
11 272 considerable complexity to the normalization process. Consequently, we are carrying out this
12 273 normalization process in two phases. In the first phase, we perform a manual normalization of
13 274 each text, and in the second phase, based on this manual work, we use computational
14 275 techniques for automatic text normalization. In other words, an automatic system will learn
15 276 from the manually normalized and labeled sample, and from there it will automatically
16 277 normalize everything else in the text.
17 278

25 279 *4.2.1 Manual Normalization*

27 280
28 281 A basis for the normalization of Basque historical texts can be found in the normalization
29 282 work previously developed. Etxeberria (2016) employed Axular's *Gero* (Labourdin dialect,
30 283 1643) and Mogel's *Peru Abarka* (Biscayan dialect, nineteenth-century) to perform
31 284 normalization experiments. One of her conclusions was that the manual normalization of 10%
32 285 of the text is enough to achieve acceptable results in automatic normalization. Accordingly,
33 286 we will use a randomly generated sample that contains 10% of each text in our corpus for
34 287 manual normalization. However, exceptions are depending on the size of the text; in fact, we
35 288 are tagging manually most of the texts of the Archaic Basque period, because they are too
36 289 short to use a 10% sample. Before starting to treat the text by hand, pre-processing is done in
37 290 three steps: i) tokenization; ii) named entity recognition; and iii) lexical recognition. Once
38 291 these three steps are completed, each word in the text will be assigned a label, as shown in
39 292 Figure 1).

- 49 293 ● ENT-Zuz: "Correct entity" [from Basque *entitate zuzena*], when the recognizer of
50 294 proper names (Alegria *et al.*, 2006) identifies a proper name, the word is labeled with
51 295 the tag ENT-Zuz (green-colored).
- 52 296 ● STD-Zuz: "Correct standard" [from Basque *estandar zuzena*], when the
53 297 morphological analyzer (Alegria *et al.*, 1996) based on the *Lexical Database of*
54 298 *Basque* [*Euskararen Datu Base Lexikala*, henceforth EDBL] (Aldezabal *et al.*, 2001)
55 299 identifies a word or a lemma (fuchsia-colored).

- OOV: “Out of vocabulary”, when the morphological analyzer does not identify the word or lemma (blue-colored).

Manual tagging will be performed with the *Brat* tool⁴ (see Fig. 1). Given the difficulty of the task, the annotator must be a linguist used to Basque historical texts. Figure 1 shows the interface for manual tagging.

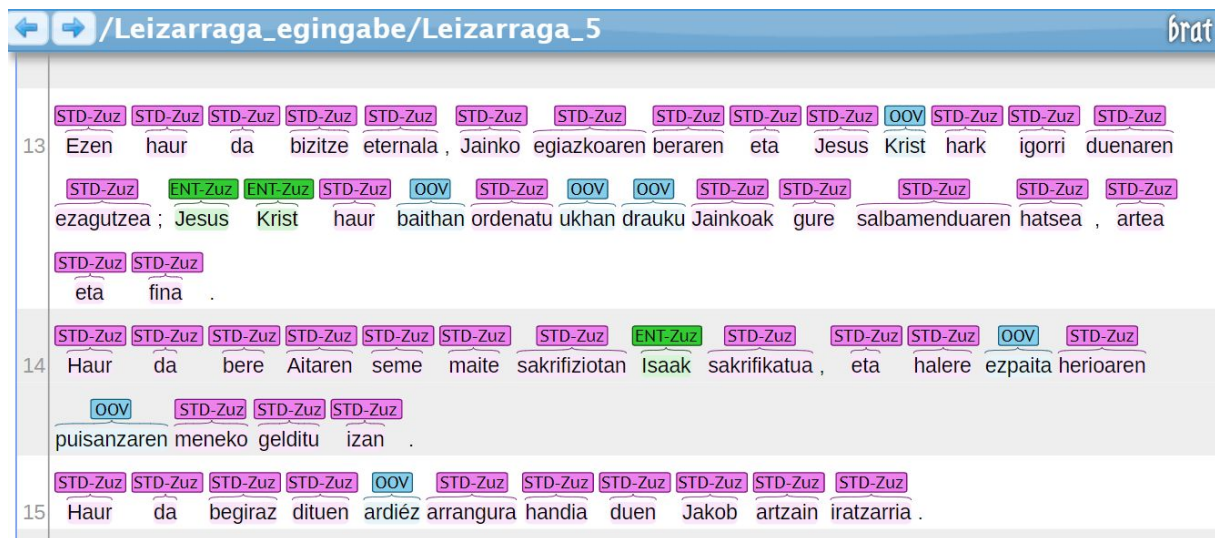


Figure 1

The mission of the annotator is to assign a standard form corresponding to each word in the text according to the dictionary of the Royal Academy of the Basque Language (*Euskaltzaindiaren Hiztegia*, henceforth EH).⁵ We follow two main criteria for manual normalization:

- When the words are variants of the same origin, we normalize the ones preferred by the EH: *andra* “lady”, *bertze* “(an)other”, *guzi* “all”, *saindu* “saint” → *andre*, *beste*, *guzti*, *santu* [same glosses].
- When the words do not have the same origin, or they are not variants of the same form, we follow the recommendations of the EH as if they were phonetic variants: *entrepesa* “company”, *xipi* “small” → *enpresa*, *txiki* [same glosses].

By doing so, the variants are linked, which means that if a corpus user in the future launches a search by lemma in the interface, the search engine will provide them with examples of all variants of that lemma. In the case the EH does not contain the word of a text in the corpus, our second main source is the *Basque General Dictionary*, henceforth OEH (Michelena and Sarasola, 1987–2005). Accordingly, we take the main variant considered as a standard form by the OEH — let us remember, the normalization we are performing just for the parser.

1
2
3 324 Among the decisions taken for normalization, we want to stress two of them:
4
5 325 i) concerning noun morphology, to prioritize the standard forms of postposition cases: *-rano*, *-*
6 326 *rekilako*, *-akgatik* → *-raino*, *-rekiko*, *-engatik*; ii) with regards to verbal morphology, to
7
8 327 preserve the roots of auxiliary verbs not belonging to standard Basque. In this regard, we must
9
10 328 mention that standard Basque has four auxiliary verbs: two transitive (**edun* “to have” and
11 329 **ezan* [obscure semantics]) and two intransitive (*izan* “to be” and **edin* “to become”);
12
13 330 nonetheless, the reality of historical dialects is much richer and more complex. Thus, our
14
15 331 decision has implied the inclusion of the paradigms of five more auxiliary verbs into the
16
17 332 morphological analyzer: *egin* “to do”, **eradun* [causative of **edun*], **erazan* [causative of
18 333 **ezan*], *eutsi* “to keep” and **iron* [obscure semantics]. Again, this task has involved
19
20 334 exhaustive philological and linguistic work, since three of these five dialectal verbal
21
22 335 paradigms have been standardized for the first time. By including them, NLP tools for Basque
23
24 336 improve significantly their performance when processing historical texts — by definition
25
26 337 dialectal ones.

27 338 The ultimate goal of this manual tagging is to remove every OOV tag (blue colored).
28
29 339 Variants are classified according to the following labels:

- 30
31 340 ● OOV-Ald: “Variant” [from Basque *aldaera*], when the human annotator assigns the
32
33 341 standard form to the non-standard variant (yellow-colored).
34
35 342 ● OOV-AldADIJ: “Variant – finite verbal form” [from Basque *aldaera – aditz jokatua*],
36
37 343 as the previous label, but specifying that the word is a finite verbal form.
38
39 344 ● OOV-Zuz: “Correct”, when the morphological analyzer does not recognize the word,
40
41 345 but it has an entry in the OEH, the annotator marks it as a correct form (fuchsia-
42
43 346 colored).
44
45 347 ● STD-Ald: “Variant”, when the form of the text matches a current standard Basque
46
47 348 word (“false friend”, yellow-colored).
48
49 349 ● STD-AldADIJ: “Variant – finite verbal form”, as the previous label, but specifying
50
51 350 that the word is a finite verbal form.

52
53 351 It should be stressed that in previous research work (Etxeberria *et al.*, 2016) only Out-of-
54
55 352 Vocabulary (OOV) tokens were revised and normalized. However, given the remarkable
56
57 353 grammatical diversity of historical and dialectal varieties with which we are working, the
58
59 354 annotator has to check all the words one by one, even those that have the STD-Zuz label,
60
355 since sometimes old words coincidentally look like modern Basque words, which adds an
356 obvious difficulty to the task (for example, the eastern Basque demonstrative *haur* “this”
357 happens to be homophonous to the standard and eastern Basque *haur* “child”). Due to the

1
2
3 358 limited number but relatively high frequency of such cases, a specific list of items is being
4
5 359 compiled, to manually tag and annotate them.

6 360 Moreover, OOV-AldADIJ and STD-AldADIJ labels are added prospectively. Given
7
8 361 that both synthetic verbal forms and auxiliary verbs will be difficult to normalize, it is
9
10 362 appropriate to identify them for further normalization tasks and morphosyntactic analysis. In
11
12 363 addition to this classification of variants, we have a label called SEK-Ber (from Basque *sekzio*
13
14 364 *berezia* “special section”, gray-colored). The linguist charged with tagging uses this mark in
15
16 365 the paragraph (s)he is working on when (s)he finds a morphosyntactic phenomenon of
17
18 366 diachronic interest. The label SEK-Ber works as a kind of reminder during the subsequent
19
20 367 automatic morphosyntactic analysis, so that we can easily identify which sentences or
21
22 368 fragments might be more difficult for the analyzer, and thus we may be able to focus on them.
23
24 369 We resort to this label mostly — but not only — for linguistic phenomena no longer occurring
25
26 370 in the language, for instance, the prosecutive case (which historically merged with the
27
28 371 ablative), certain verbal roots such as *-(g)idi*, the perfect past periphrases with auxiliaries
29
30 372 **edin / *ezan* in a the main clause (*mana zezan* “(s)he commanded”, present-day Basque
31
32 373 *manatu zuen* [same gloss]), old synthetic constructions of the subjunctive with the auxiliaries
33
34 374 *izan / *edun*, or prescriptive forms such as *egin albaiteza* “you will do [imperative meaning]”.

35
36 375 This prominent issue of the forms that vanished from present-day Basque and require
37
38 376 specific annotation must be anticipated in the annotation manual. Other tags may also be
39
40 377 employed, e.g. Proper Noun, Spelling Error, Improper Segmentation, and others.

41
42 378 Finally, let us mention that we have estimated the time needed to normalize the whole
43
44 379 corpus by hand. According to it, if the annotator tags 143 words per hour, (s)he would need
45
46 380 about three years to treat the entire corpus. Since this is not feasible with the resources we
47
48 381 have in this project, we chose to use computational techniques to normalize the corpus. In any
49
50 382 case, it is not only a budgetary issue: making a virtue of necessity, the choice of automatic
51
52 383 tagging leads to improvement in the automatic analyzer of the Basque language, as this tool is
53
54 384 being trained in the historical and dialectal forms of Basque.

55 385 56 386 *4.2.2 Automatic Normalization*

57 387
58 388 To achieve the automatic normalization of our corpus we follow the normalization method
59
60 389 presented by Etxeberria *et al.* (2019). This method is based on statistical learning and
61
62 390 proposes a manually tagged sample for the learning process. All the pairs of words in the
63
64 391 manual sample will be considered in the learning process; some of them will be tagged with

1
2
3 392 the label *Zuz* (“correct”), and the rest with the label *Ald* (“variant”), but we will use every pair
4
5 393 for automatic learning. This method involves *Phonetisaurus*,⁶ a grapheme-based phonological
6
7 394 tool, guided by finite-state transducers (WFST) (Novak *et al.*, 2012, 2016). By introducing
8
9 395 “variant/standard” and “standard/standard” pairs, the tool learns graph sequences. That is, we
10
11 396 use *Phonetisaurus* to teach the changes that occur within the word pairs in the learning
12
13 397 sample, and to create grapheme-to-grapheme models. After this training, the system should be
14
15 398 able to propose standard forms for variants not previously analyzed.

15 399 After all this, we need to evaluate the quality of automatic normalization. We have
16
17 400 applied the 10-fold cross-validation technique, dividing the manually tagged sample into ten
18
19 401 parts. Of these ten parts, we used nine files for learning, and the remaining tenth to evaluate
20
21 402 the results. We performed the same experiment ten times, changing the file used for
22
23 403 evaluation, and we calculated the arithmetic average of these ten experiments to rate the
24
25 404 normalization quality of each text (see Section 5).

25 405

27 406 *4.3 Morphosyntactic Tagging*

29 407

30 408 Once the normalization of the corpus is completed, the NLP analyzers are able to parse the
31
32 409 texts. In this project, we tag the corpus employing *Eustagger*, a morphosyntactic analyzer
33
34 410 developed by the Ixa team (Ezeiza *et al.*, 1998). This parser processes the text in three steps:
35
36 411 i) tokenization, ii) lemmatization, and iii) parsing — both segmentation and assignment of
37
38 412 syntactic functions.

39 413 Firstly, the parser divides the text into words (tokens). Secondly, it identifies the
40
41 414 lemma of each word in the text and, thirdly, it proposes a morphosyntactic analysis. On the
42
43 415 one hand, the parser segments each word into morphemes and assigns them a value; on the
44
45 416 other hand, it marks syntactic functions too (subject, direct object, indirect object, and so on).

46 417 Finally, we review the morphosyntactic analysis performed by the automatic parser,
47
48 418 with particular attention to paragraphs marked with the label SEK-Ber (see Section 4.2.1); if a
49
50 419 structure is poorly analyzed by the tool, we correct it. In this respect, we need to formulate a
51
52 420 number of new rules to properly parse morphosyntactic features that do not occur in present-
53
54 421 day standard Basque. Thus, the parser will learn through manually corrected and formulated
55
56 422 items to improve its performance. In other words, after completing this process we will have a
57
58 423 morphosyntactic parser of historical Basque ready.

58 424

60 425 **5. Learning through Normalization Tools: Results**

426

427 In this section, we will present the results obtained in the application of Etxeberria's (2016)
 428 normalization method to some texts of our historical corpus of Basque. More specifically, we
 429 have carried out experiments on three of the main texts of the Archaic Basque period (until
 430 ca. 1600):

- 431 • Dechepare's *Linguae Vasconum Primitiae*, the first book printed in Basque
 432 (Bordeaux, 1545), a collection of poems written in Eastern Low Navarrese.
- 433 • Joanes Leizarraga's *Iesus Krist Gure Iaunaren Testamentu Berrria* [= *New Testament*
 434 *of our Lord Jesus-Christ*] (La Rochelle, 1571), Calvinist inspired translation written in
 435 a sort of koine of eastern dialects.
- 436 • *Refranes y Sentencias (RS)* (Pamplona, 1596), a compilation of sayings that reflect an
 437 archaizing variety located in or near Bilbao (Biscay).

438 As explained in Section 4.2.2, we have randomly selected 10% of each text to normalize it
 439 manually. In any case, we have decided to tag the text of *RS* fully by hand, because of the
 440 very special characteristics of this work, made up of sayings formulated in a highly archaizing
 441 language. However, although the whole text of *RS* has been manually normalized, we found it
 442 interesting to evaluate the result that our normalization method would have in a text of this
 443 type, so we also have applied the 10-fold cross-validation evaluation method to this work.
 444 Table 2 displays the rate of correct analysis achieved by automatic normalization in each text.
 445 On the one hand, we have worked out the success rate considering the whole of words in the
 446 text, as it will be in real situations; on the other hand, we have also calculated the number of
 447 right analyses performed by automatic normalization considering only non-standard words.

448

	Leizarraga	Dechepare	<i>RS</i>
Only variants	87.68%	70.38%	66.83%
All words	94.24%	86.46%	80.66%

449 **Table 2**

450

451 The first work through which we evaluated the automatic normalization method was
 452 Leizarraga's *New Testament*, and we found out that the results were very good, reaching a
 453 success rate of 94% for the whole of words in the text.

454 The second work analyzed was Dechepare's book, in which we conducted several
 455 experiments. First, in order to economize manual work, we tried to normalize Dechepare's

1
2
3 456 text through what the automatic normalizer had learned from Leizarraga but, as expected, we
4
5 457 obtained poor results, as Leizarraga's and Dechepare's linguistic varieties are quite different,
6
7 458 particularly concerning non-standard forms: 75% for the whole of words, but only 46% for
8
9 459 variants. Bearing this in mind, we decided to use only the manually normalized sample of
10
11 460 Dechepare's text in the learning process. By doing so we improved the results, but they were
12
13 461 not yet good enough to guarantee a successful normalization (59.58% for variants). To
14
15 462 address this situation, we had two possible options: i) to tag more texts manually, or ii) to
16
17 463 apply what was learned from Leizarraga's text processing to machine learning. Before starting
18
19 464 tagging more by hand, we opted out for the second option, and the results improved
20
21 465 significantly: 86.46% for all words and 70.38% for variants. The results of this last
22
23 466 experiment suggest that in the future it might be a reasonable solution to have a normalization
24
25 467 system for each dialectal area, instead of undertaking a normalization effort for each text.

26
27 468 The last text processed was *RS*, for which we got poorer results: 80.66% for all words
28
29 469 and 66.83% for variants. The reason is that the very archaizing language considerably
30
31 470 increases the difficulty of the normalization process. The obvious conclusion is that the more
32
33 471 a text differs from the standard language, the more difficult its normalization becomes. Table
34
35 472 3 shows the numbers for standard words, variants, and entities in each text analyzed.
36
37 473

	Leizarraga	Dechepare	<i>RS</i>
Number of words	73,610	6,860	3,083
STD-Zuz	50,321	4,416	1,709
%	68.36	64.37	55.43
OOV	20,924	2,403	1,352
%	28.42	35.03	43.85
ENT-Zuz	2,365	41	22
%	3.21	0.60	0.71
Tagged by hand	7,548	694	3,083

38
39
40
41
42
43
44
45 474 **Table 3**

46
47 475
48
49 476 As shown in Table 3, the results obtained in the evaluation are completely consistent with the
50
51 477 number of variants present in each text. In the case of *RS*, almost half of the words are not
52
53 478 standard ones, and this complicates automatic normalization. The results obtained in *RS* point
54
55 479 out that it is worthy to do a special effort in manual tagging when processing texts which
56
57 480 differ from the standard language.

58 481

59 482 **6. The Search Engine and the Interface**

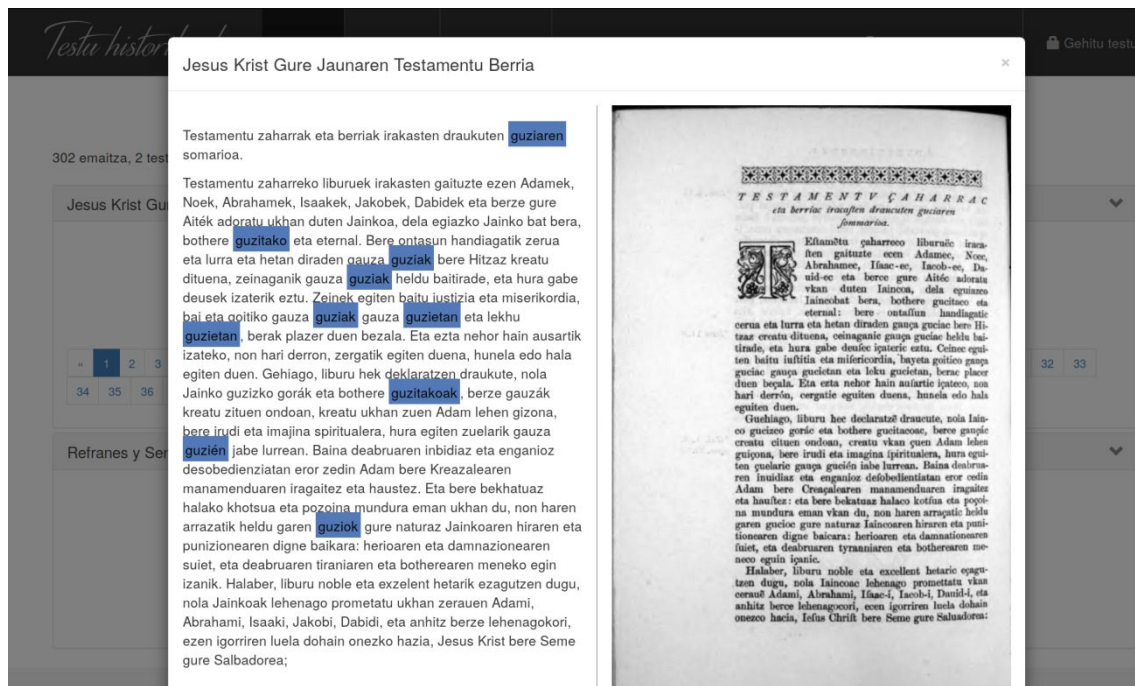
1
2
3 483
4
5 484 Computational linguists are building a search-engine-based web-application to consult the
6
7 485 annotated data as well as information related to the project. The architecture and design of
8
9 486 such application are based on different analytic tools. This includes part-of-speech tagging
10
11 487 (Alegria *et al.*, 1996; Aduriz *et al.*, 2000; Alegria *et al.*, 2002), chunk analysis (Alegria *et al.*,
12
13 488 2008), and surface syntactic disambiguation (Arriola, 2015).

14 489 Thus, the final output of our project will be a search interface to browse the corpus.
15
16 490 This interface must be useful for analyzing diachronic syntax. Therefore, it must be able to
17
18 491 perform complex searches, both in terms of metadata (period, dialect, author, gender, etc.) and
19
20 492 morphosyntactic characteristics (original form, lemma, part-of-speech, case ending, auxiliary
21
22 493 verb root, time, aspect, mode, etc.), as well as any combination of them. A
23
24 494 morphosyntactically annotated corpus will allow looking for structures that may be of interest
25
26 495 from the perspective of Basque historical linguistics, for example:

- 27 496 ● Postnominal relatives: search [noun + relative] structure.
- 28 497 ● The so-called *archaic aorist*: search [verb + subjunctive past + main sentence], or
29 498 [*EDIN lemma + past + main sentence] — **edin* is an intransitive auxiliary verb used
30
31 499 in the expression of non-indicative values; thus, today its forms of the past express
32
33 500 what conventionally is called *subjunctive*; however, in Archaic Basque such forms
34
35 501 could be used to express perfective past in the main sentence.
- 36 502 ● The special word order of negation: search [main verb + lemma EZ “not” + auxiliary
37
38 503 verb] structure.
- 39 504 ● Old periphrases of habituality: search [verbal noun + finite forms of the verbs *eroan*
40
41 505 “carry”/ *eraman* “carry” / *joan* “go”] structure.

42
43 506 Finally, we want to stress another relevant feature of the search interface. When the user will
44
45 507 browse the corpus and receive the list of results, (s)he will have the possibility to click on an
46
47 508 example to display the page of the text in updated spelling, as well as the corresponding
48
49 509 image of the facsimile, as shown in Figure 2.

50 510
51
52
53
54
55
56
57
58
59
60



511
512 **Figure 2**

513
514 **7. Conclusion**

515
516 We have presented the creation of the first morphosyntactically annotated historical corpus of
517 Basque. To that end, an interdisciplinary team has been formed by theoretical and
518 computational linguists and computer scientists. As undertaking the entire preparation of the
519 corpus by hand is not feasible, computational techniques are needed.

520 Concerning methodology, after the design and compilation of the corpus, the first task
521 has been a philological validation of the texts. Then we have normalized the corpus so that the
522 automatic tools be able to analyze the texts. We normalize by hand a part of each text, and
523 based on these manually normalized parts and applying machine learning techniques, we
524 automatically normalized the remaining parts of the texts. Once the text is normalized, the
525 *Eustagger* parser is ready to tag the texts. Finally, a search interface to browse all the
526 morphosyntactic information stored in the corpus is under development.

527 The normalization of the text is crucial to guarantee the success of the project, and we
528 have shown that the results obtained in three texts of Archaic Basque validate the
529 methodology proposed. However, the more a text differs from the standard language, the
530 more its normalization becomes difficult. Consequently, to improve both the performance of
531 automatic NLP tools and the quality of the search engine, we decided to make a special effort
532 in manual work.

1
2
3 533 In the nearest future, the major challenge is the adaptation of automatic
4
5 534 morphosyntactic analysis to historical Basque texts. What is more, we collect and consider
6
7 535 theoretical linguists' proposals and suggestions to progress in the implementation of
8
9 536 functionalities of the search engine and the design of the interface. In the longer term, our
10
11 537 goal is the enhancement of the corpus, extending its time scope until the twentieth century —
12
13 538 indeed a project has been recently funded to undertake the task.

13 539 To sum up, the corpus we are constructing aims to remove a major hindrance for the
14
15 540 study of Basque diachronic syntax, which is the absence of such resources for this language.
16
17 541 In this sense, the tool we are developing will be not only an important contribution to Basque
18
19 542 Studies, but it also pursues the improvement of the tools for language processing of Basque
20
21 543 and lays the foundations for future projects. Ultimately, some lessons learned from the work
22
23 544 with highly diversified grammatical materials may inform projects on other languages that
24
25 545 have similar aims.

1
2
3 547 **Notes**

4 548

5 549
6 550
7 551
8 552
9 553
10 554
11 555
12 556
13 557
14 558
15 559
16 560
17 561
18 562
19 563
20 564
21 565
22 566
23 567
24 568

1. Manuel Larramendi provided the first truly influential orientations about the way Basque should be written, which involved not only orthography but most importantly, also word order and paradigmatic choices. Those recommendations were extremely successful during the second half of the eighteenth century and the nineteenth century, especially in Southern dialects (i.e. South to the French-Spanish border), and their influence may be perceived in highly sensitive domains of linguistic structure, such as word order. In this sense, the written language before Larramendi may be considered to be closer to the spoken language than the subsequent textual tradition.

2. <<https://klasikoak.armiarma.eus/krono.htm>>

3. <<https://www.ehu.eus/en/web/eins/euskal-klasikoen-corpusa>>

4. <<https://brat.nlplab.org/index.html>>

5. <https://www.euskaltzaindia.eus/index.php?option=com_hiztegiabiltatu&view=frontpage&Itemid=410&lang=eu>

6. <<https://github.com/AdolfVonKleist/Phonetisaurus>>

1
2
3 570 **Funding**

4
5 571 This work was supported by the *Agence nationale de la recherche* of France [ANR-17-CE27-
6 572 0011-BIM]; and the Ministry of Science, Innovation, and Universities of Spain [RTI2018-
7
8 573 098082-J-I00] to X. X.
9

10 574

11
12 575 **Acknowledgments**

13 576 Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements |

14
15 577 Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements |

16
17 578 Acknowledgements | Acknowledgements | Acknowledgements
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

1
2
3 580 **Captions**

4
5 581
6 582 Table 1 Books printed in Basque between 1545 and 1879. The 1545-1749 columns roughly
7
8 583 correspond to the span of our corpus (adapted from Sarasola, 1976)

9
10 584 Table 2 Success rate of automatic normalization in three Archaic Basque texts

11
12 585 Table 3 Size of each text, number, and proportion of the words labeled with the STD-Zuz,
13
14 586 OOV, and ENT-Zuz tags

15 587 Figure 1 Interface for manual normalization of the corpus showing a fragment of Leizarraga's
16
17 588 New Testament

18
19 589 Figure 2 The interface displays a page of Leizarraga's New Testament in both modernized
20
21 590 spelling (left) and facsimile (right)

1
2
3 592 **References**
4

5 593
6 594 **Aduriz, I., Agirre, E., Aldezabal, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K.,**
7
8 595 **Sarasola, K. and Urkia, M.** (2000). A Word-Level Morphosyntactic Analyzer for Basque. In
9
10 596 M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhauer (Eds.),
11
12 597 *Proceedings of the Second International Conference on Language Resources and Evaluation.*
13
14 598 Athens: European Language Resources Association (ELRA).

15 599
16
17 600 **Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G. and Lersundi,**
18
19 601 **M.** (2001). EDBL: A General Lexical Basis for the Automatic Processing of Basque. In S.
20
21 602 Bird, P. Buneman and M. Liberman (Eds.), *IRCS Workshop on linguistic databases.*
22
23 603 Philadelphia: University of Pennsylvania. Institute for Research in Cognitive Science.

24 604
25 605 **Alegria, I., Aranzabe, M. J., Ezeiza, N., Ezeiza, A., and Urizar, R.** (2002). Robustness and
26
27 606 customisation in an analyser/lemmatiser for Basque. In Federica Busa (Ed.), *LREC-2002*
28
29 607 *Customizing knowledge in NLP applications workshop* (pp. 1–6). Las Palmas de Gran
30
31 608 Canaria: European Language Resources Association (ELRA).

32 609
33
34 610 **Alegria, I., Arregi, O., Ezeiza, N. and Fernández, I.** (2006). Lessons from the Development
35
36 611 of a Named Entity Recognizer for Basque. *Procesamiento del lenguaje natural*, 36, 25–37.

37 612
38
39 613 **Alegria, I., Arrieta, B., Carreras, X., Díaz de Ilarraza, A. and Uria, L.** (2008). Chunk and
40
41 614 Clause Identification for Basque by Filtering and Ranking with Perceptrons. *Procesamiento*
42
43 615 *del Lenguaje Natural*, 41, 5–12.

44 616
45
46 617 **Alegria, I., Artola, X., Sarasola, K. and Urkia, M.** (1996). Automatic morphological
47
48 618 analysis of Basque. *Literary and Linguistic Computing*, 11(4), 193–203.

49 619
50
51 620 **Arriola, J. M.** (2015). Different Issues in the Design and Implementation of a Rule Based
52
53 621 Grammar for the Surface Syntactic Disambiguation of Basque. In E. Bick and K. Hagen
54
55 622 (Eds.), *Proceedings of the Workshop on Constraint Grammar - methods, tools and*
56
57 623 *applications; at NODALIDA 2015* (pp. 1–9). Vilnius: Institute of the Lithuanian Language.

58 624
59
60

1
2
3 625 **Claridge, C.** (2009). Historical corpora. In A. Lüdeling and M. Kytö (Eds.), *Corpus*
4 626 *linguistics. An International Handbook* (pp. 242–259). Berlin: Mouton de Gruyter.

5
6 627
7
8 628 **Etxeberria, I.** (2016). *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta*
9 629 *morfologikoa erabiliz* [= Normalization of linguistic variants using phonological and
10 630 morphological inferences]. (Unpublished doctoral dissertation). University of the Basque
11 631 Country, San Sebastián.

12
13 632
14
15 633 **Etxeberria, I., Alegria, I. and Uria, L.** (2019). Weighted finite-state transducers for
16 634 normalization of historical texts. *Natural Language Engineering*, 25(2), 307–321.

17 635
18 636 **Ezeiza, N., Alegria, I., Arriola, J. M., Urizar, R. and Aduriz, I.** (1998). Combining
19 637 stochastic and rule-based methods for disambiguation in agglutinative languages. In C. Boitet
20 638 and P. Whitelock (Eds.), *Proceedings of the 36th Annual Meeting of the Association for*
21 639 *Computational Linguistics and 17th International Conference on Computational Linguistics,*
22 640 *Volume 1* (pp. 380–384). Montreal: Association for Computational Linguistics (ACL).

23 641
24 642 **Galves, Ch., Andrade, A. and Faria, P.** (2017). *Tycho Brahe Parsed Corpus of Historical*
25 643 *Portuguese*.
26 644 <http://www.tycho.iel.unicamp.br/~tycho/corpus/> (accessed 29 January 2021).

27 645
28 646 **Gómez, R. and Sainz, K.** (1995). On the origin of the finite forms of the Basque verb. In J. I.
29 647 Hualde, J. A. Lakarra and R. L. Trask (Eds.), *Towards a History of Basque Language* (pp.
30 648 235–274). Amsterdam: John Benjamins.

31 649
32 650 **Gorrochategui, J.; Igartua, I. and Lakarra, J. A.** (Eds.) (2018). *Historia de la lengua*
33 651 *vasca*. Vitoria-Gasteiz: Basque Government.

34 652
35 653 **Hunston, S.** (2009). Collection strategies and design decisions. In A. Lüdeling and M. Kytö
36 654 (Eds.), *Corpus linguistics. An International Handbook* (pp. 154–167). Berlin: Mouton de
37 655 Gruyter.

38 656
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 657 **Kroch, A., Santorini, B. and Delfs, L.** (2004). *The Penn-Helsinki Parsed Corpus of Early*
4 658 *Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-
5 659 ROM, first edition, release 3.
6
7
8 660 <http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3> (accessed 29 January
9 661 2021).
10
11 662
12
13 663 **Kroch, A., Santorini, B. and Diertani, A.** (2016). *The Penn Parsed Corpus of Modern*
14 664 *British English (PPCMBE2)*. Department of Linguistics, University of Pennsylvania. CD-
15 665 ROM, second edition, release 1.
16
17 666 <http://www.ling.upenn.edu/ppche-release-2016/PPCMBE2-RELEASE-1> (accessed 29
18 667 January 2021).
19
20 668
21
22 669 **Kroch, A. and Taylor, A.** (2000). *The Penn-Helsinki Parsed Corpus of Middle English*
23 670 *(PPCME2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second
24 671 edition, release 4.
25
26 672 <http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4> (accessed 29 January
27 673 2021).
28
29 674
30
31 675 **Lafon, R.** (1944). *Le système du verbe basque au XVI^e siècle*. Bordeaux: Éditions Delmas.
32
33 676
34
35 677 **Lafon, R.** (1951–1952). Concordances morphologiques entre le basque et les langues
36 678 caucasiques. *Word* 7, 227–244 and 8, 80–94.
37
38 679
39
40 680 **Lakarra, J. A.** (1995). Reconstructing the root in Pre-Proto-Basque. In J. I. Hualde, J. A.
41 681 Lakarra and R. L. Trask (Eds.), *Towards a History of Basque Language* (pp. 189–206).
42 682 Amsterdam: John Benjamins.
43
44 683
45
46 684 **Lakarra, J. A.** (2005). Prolegómenos a la reconstrucción de segundo grado y análisis del
47 685 cambio tipológico en (proto) vasco. *Palaeohispanica*, 5, 407–459.
48
49 686
50
51 687 **Lakarra, J. A.** (2006). Protovasco, munda y otros: reconstrucción interna y tipología holística
52 688 diacrónica. *Oihenart: cuadernos de lengua y literatura*, 21, 229–322.
53
54 689
55
56 690 **Lash, E.** (2014). *The Parsed Old and Middle Irish Corpus (POMIC)*. Version 0.1.

- 1
2
3 691 [https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-](https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/)
4 pomic/ (accessed 29 January 2021).
5
6 693
7
8 694 **Manterola, J.** (2015). *Towards a History of Basque Morphology: Articles and*
9 *demonstratives*. (Unpublished doctoral dissertation). University of the Basque Country,
10 Vitoria-Gasteiz.
11
12 696
13 697
14
15 698 **Martínez-Arena, M.** (Ed.) (2013). *Basque and Proto-Basque. Language-Internal and*
16 *Typological Approaches to Linguistic Reconstruction*. Frankfurt: Peter Lang.
17
18 700
19
20 701 **Michelena, L.** (1977 [1961]). *Fonética histórica vasca*. San Sebastián: Council of Guipuscoa.
21
22 702
23
24 703 **Michelena, L.** (1964). *Textos Arcaicos Vascos*. Madrid: Minotauro.
25
26 704
27 705 **Michelena, L. and Sarasola, I.** (Eds.) (1987–2005). *Orotariko Euskal Hiztegia* [= Basque
28 General Dictionary]. Bilbao: Royal Academy of the Basque Language.
29
30 706 [https://www.euskaltzaindia.eus/index.php?option=com_oeh&view=frontpage&Itemid=413&l](https://www.euskaltzaindia.eus/index.php?option=com_oeh&view=frontpage&Itemid=413&lang=eu)
31 ang=eu (accessed 29 January 2021).
32
33 708
34 709
35
36 710 **Nelson, M.** (2010). Building a written corpus What are the basics? In Anne O’Keeffe and
37 Michael MacCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London/New
38 York: Routledge.
39
40 712
41 713
42
43 714 **Novak, J. R., Minematsu, N. and Hirose, K.** (2012). WFST-based grapheme-to-phoneme
44 conversion: Open source tools for alignment, model-building and decoding. In I. Alegria and
45 M. Hulden (Eds.), *Proceedings of the 10th International Workshop on Finite State Methods*
46 *and Natural Language Processing* (pp. 45–49). San Sebastián: Association for Computational
47 Linguistics.
48
49 718
50 719
51
52 720 **Novak, J. R., Minematsu, N. and Hirose, K.** (2016). Phonetisaurus: Exploring grapheme-to-
53 phoneme conversion with joint n-gram models in the WFST framework. *Natural Language*
54 *Engineering*, 22(6), 907–938.
55
56 722
57 723
58
59 724 **Sarasola, I.** (1976). *Historia Social de la Literatura Vasca*. Madrid: Akal.

- 1
2
3 725
4
5 726 **Sarasola, I.** (1983). *Contribución al estudio y edición de textos antiguos vascos*. San
6 727 Sebastián: Council of Guipuscoa and University of the Basque Country.
7
8 728
9
10 729 **Satrústegui, J. M.** (1987). *Euskal Testu Zaharrak (I)* [= Old Basque texts]. Pamplona: Royal
11 730 Academy of the Basque Language.
12
13 731
14
15 732 **Schuchardt, H.** (1907). *The Iberische Deklination*. Vienna: Holder.
16
17 733
18
19 734 **Schuchardt, H.** (1914). Baskisch und Hamitisch. *Revista Internacional de Estudios Vascos*,
20 735 8(1), 76.
21
22 736 Trask, R. L. (1997). *The History of Basque*. London/New York: Routledge.
23
24 737
25
26 738 **Uhlenbeck, C. C.** (1924). De la possibilité d'une parenté entre me basque et les langues
27 739 caucasiques. *Revista Internacional de Estudios Vascos*, 15, 565–588.
28
29 740
30
31 741 **Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F. and Rögnvaldsson, E.** (2011).
32 742 *Icelandic Parsed Historical Corpus (IcePaHC)*. Version 0.9.
33
34 743 http://www.linguist.is/icelandic_treebank (accessed 29 January 2021).
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60