The First Annotated Corpus of Historical Basque

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

OXFORD UNIVERSITY PRESS | Digital Scholarship in the Humanities

The First Annotated Corpus of Historical Basque

Journal:	*Digital Scholarship in the Humanities*
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

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.

The main goal of our project is the systematic diachronic study of several grammatical features of Basque in the light of the current theoretical discoveries. The second stated goal is to create a comprehensive annotated historical corpus of Basque that comprises both part-of-speech and syntactic annotation as well as a rich set of metadata structure. This way, it will be possible to search the annotated corpus by words, lemmas, grammatical categories, by sequences of grammatical categories, and by specific structural configurations (such as relatives, correlatives, finite subordination, and others) as well as by publication year, author or place of publication, thanks to the mentioned metadata. Hence, the preparation of resources for the syntactic exploration of the historical corpus in Basque will allow us to achieve the first goal since the syntactically annotated historical corpus can show us how certain syntactic structures have evolved over the centuries. The final output of our project will be a search interface to browse the corpus.

This article focuses on the second goal of the project, that is, the construction of the morphosyntactically annotated historical corpus of Basque. Having developed a lot of NLP-based technology for the analysis of standard Basque, we now have to tackle the text normalization, considering that once ancient or dialectal texts are normalized, the developed NLP tools could be applied for the linguistic analysis of the corpora. We will explain the implementation and evaluation of computational techniques for corpus normalization. Our methodology is partially defined by previous research on the automatic normalization of historical and dialectal Basque texts (Etxeberria *et al.*, 2014; Etxeberria *et al.*, 2016) — we must stress, however, that the method was tested and evaluated in an experimental application. Thus, a relevant contribution of this paper is testing the method in a real scenario, considering periods and dialects never before analyzed through NLP tools.

After having explained the starting point for the project of an annotated historical corpus, in Section 2 we summarize the state of research in historical Basque linguistics and indicate the need for annotated tools for the study of the historical syntax of Basque. In Section 3 we explain and describe the reference corpus. Section 4 proposes an overview of the tasks involved in the normalization of the corpus, as well as the subsequent steps leading to the creation of a searchable database. After methodological issues, in Section 5 we present the results obtained so far applying computational methods and Natural Language Processing tools to historical Basque varieties. Then we provide some details about the search interface under construction in Section 6. Finally, in Section 7 we formulate the conclusions and point out future orientations.

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

morphological trace can be identified today in one of the locative particles (Lakarra, 2006). The complex finite forms in Basque, with their long sequences of agreement and diathetic affixes, a feature of the language that has fascinated European scholars for decades, can be shown to be relatively recent: the emergence of number agreement marking in the auxiliary can to some extent be tracked on the existing corpus, as can other instances of agreement, such as dative agreement. The evolution of the determiner system, which evolved from the demonstrative one (Manterola, 2015), can be directly examined, to a great extent, on textual records, and the same goes for the category of morphological number, absent from earlier stages of Basque. That is, many of the crucial steps leading to the special typological properties of Basque can be accessed and examined in the light of the increasingly rich textual evidence, and with the methods and analytic resources of modern diachronic linguistics. This also implies carefully studying the contact-induced changes concerning the surrounding romance languages, as well as considering the existing dialectal variation.

Nevertheless, no systematic study of this sort has ever been undertaken for Basque historical syntax. The project [NAME] aims to address this challenge. It has two main goals: i) to carry out a systematic diachronic study of some grammatical features of Basque, and ii) the elaboration of a morphosyntactically annotated historical corpus of Basque. The corpus will also have a rich metadata structure that will allow searches by dialect, time, or socio-historical information. This will allow for complex and rich searches, e.g. by word, lemma, part-of-speech, and also particular syntactic structures that we also annotate (relatives, correlatives, etc.) and combinations of them. In this paper, we address this second goal in particular. On the other hand, we can say that [NAME] is an interdisciplinary project, involving experts in the field of historical linguistics and natural language processing (henceforth NLP) and the collaboration between [NAME] and [NAME]centers.

3. The Corpus

3.1 Design and Decisions

We decided to establish a philologically-reliable corpus covering most of the textual production between the fifteenth and mid-eighteenth centuries. This is, on the one hand, the minimal span that includes regular attestations for all Basque dialects and, on the other hand, it also divides Archaic and Old Basque from Early Modern Basque.

Usual issues conditioning the building of historical corpora are manual processes of compilation, accessibility of materials, and availability of different types of texts (Claridge, 2009). In terms of representativeness, understood as the relationship between the corpus and the body of the language, the ideal tends to the inclusion of 'texts from as many different categories of writing and speech as resources will allow' (Hunston, 2009: 160). As far as Basque is concerned, the corpus should be representative of all dialects with written tradition.

Regarding balance, which refers to the internal composition of the corpus, things are even more complicated. The body of historical Basque is in many respects unbalanced. Table 1 shows the figures for printed books in the four so-called historical *literary dialects* of Basque. Until 1750, which roughly coincides with the closure point of our corpus, a significant majority of the texts correspond to two dialects in the French side of the Basque-speaking area: Labourdin and Souletin; moreover, nearly 90% of these published texts are religious ones. Thus, if the goal of achieving balance among regional varieties might be challenging for Germanic languages (cf. Claridge, 2009), then in the case of Basque it becomes a chimera.

	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

Table 1

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

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

mention that our design has adopted the complete-text approach instead of selecting samples of texts.

3.2 Periodization and Texts

In a commonly accepted periodization (see Gorrochategui *et al.*, 2018), the Basque textual history is divided into four main periods: i) Archaic Basque, from the fifteenth century to 1600; ii) Old Classical Basque, from 1600 to 1745; iii) Early Modern Basque, from 1745 to the end of the nineteenth century; and iv) Late Modern Basque, from the end of the nineteenth century to 1968, year in which standard Basque was launched. Our project encompasses Archaic and Old Classical periods; another project that addresses both Early and Late Modern Basque is also underway [NAME]. The divide between Old Classical Basque and Early Modern Basque, which defines the outer limit of our corpus, is mediated by the decisive influence of Manuel Larramendi's works, in particular his grammar (*El Imposible Vencido*, 1729) and his *Diccionario Trilingüe* (1745).[1]

Archaic Basque is represented by the four main texts from this period: i) Dechepare's poems compilation, the first printed book in Basque (1545), written in Eastern Low Navarrese dialect; ii) Lazarraga's manuscript, which includes a pastoral novel in Alavese dialect (ca. 1562); iii) Leizarraga's Calvinist translation of the New Testament into a sort of Northern Basque *koine* (1571); and iv) *Refranes y Sentencias*, a compilation of sayings in Biscayan dialect (1596). We also include most of the shorter texts edited by Michelena (1964) and Sarasola (1983), as well as those compiled by Satrústegui (1989).

As for Old Basque, our corpus includes the bulk of the written production of this period, which is considerably larger than that of the Archaic Basque — the reader can get an idea looking at the chronological list in the web *Klasikoen Gordailua* [= The Classics' Repository].[2] Additionally, the above-mentioned collections by Michelena, Sarasola, and Satrústegui include many texts of the Old Basque period. Despite its less-than-fully satisfactory philological quality, the inclusion of Satrústegui's compilation is justified because of the heterogeneous and rare dialectal materials it contains, some of them corresponding to High Navarrese, a dialect poorly represented in the central periods of the history of Basque.

4. Methodology and Tasks

To create a morphosyntactically tagged historical corpus of Basque, we propose three main steps. First, the corpus must be designed, compiled, and prepared. Then, the texts must be normalized, to make the automatic analysis by NLP tools possible — i.e. we need to convert the historical texts into standard Basque. After this normalization process, we will analyze and annotate the texts using the parsers developed by the [NAME] team. Below we shall detail each of these steps.

4.1 Compilation, Revision, Correction, and Spelling-Update

Our starting point is the *Basque Classics Corpus* of the Institute for the Basque Language of the University of the Basque Country,[3] with a size of 11.9 million words. It includes texts from the fifteenth to the early twentieth centuries pertaining to all the historical dialects of Basque and presents a rich variety of genres in both verse and prose (drama, translations, original literary narrative, religious prose, sayings, official and personal correspondence among others). Therefore, it offers a solid basis for analyzing the evolution of historical dialects. Unfortunately, this existing historical corpus, with all its obvious potential, cannot be directly used as it is for academic purposes:

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

b) Secondly, the corpus is highly unreliable, as it has been compiled on the basis of the voluntary contribution of scholars from different backgrounds and of unequal competence. Moreover, some of the transcriptions are based on outdated editions, and therefore reproduce errors that have been corrected in later ones. The transcriptions may also be incorrect due to typing errors, or to the wrong interpretation of the original forms. The use of this resource for linguistic exploitation, therefore, requires a thorough revision of the texts.

c) Thirdly, the most important part of the corpus, corresponding to the literary production, is transcribed in present-day standard orthography, a choice that keeps the original morphological forms, but results in the lack of relevant phonological information.

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

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

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

The main criterion for this philological work is the modernizing of the spelling — not to confuse with the adoption of present-day standard Basque orthography. The goal is to update the spelling of the texts so that NLP tools can work with them, always without distorting their phonological features. That is, while in the *Basque Classics Corpus* the spelling of the texts has been updated according to the rules of present-day standard Basque, in our corpus the modernizing of the spelling preserves the phonological shape of each text. For instance, Eastern Basque dialects have a set of aspirated plosive phonemes *ph*, *th*, *kh* that is not represented in the spelling system of standard Basque. As a result, most of the transcripts of Eastern Basque texts in the *Basque Classics Corpus* do not reflect aspirated plosives, although in original texts they are usually represented by special graphemes.

4.2 Normalization of the Corpus

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

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

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

4.2.1 Manual Normalization

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

- ENT-Zuz: "Correct entity" [from Basque *entitate zuzena*], when the recognizer of proper names (Alegria *et al.*, 2006) identifies a proper name, the word is labeled with the tag ENT-Zuz (green-colored).
- STD-Zuz: "Correct standard" [from Basque *estandar zuzena*], when the morphological analyzer (Alegria *et al.*, 1996) based on the *Lexical Database of Basque* [*Euskararen Datu Base Lexikala*, henceforth EDBL] (Aldezabal *et al.*, 2001) 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* tool[4] (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.

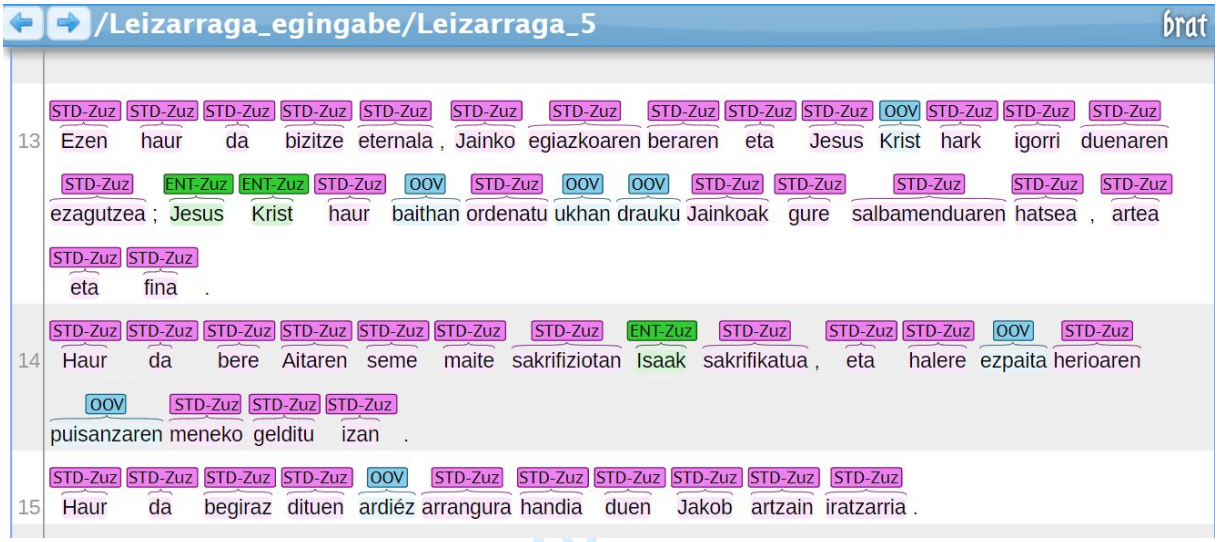

Figure 1

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:

i. 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].

ii. 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: *entrepresa* "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.

Among the decisions taken for normalization, we want to stress two of them: i) concerning noun morphology, to prioritize the standard forms of postposition cases: *-rano*, *-rekilako*, *-akgatik* → *-raino*, *-rekiko*, *-engatik*; ii) with regards to verbal morphology, to preserve the roots of auxiliary verbs not belonging to standard Basque. In this regard, we must mention that standard Basque has four auxiliary verbs: two transitive (**edun* "to have" and **ezan* [obscure semantics]) and two intransitive (*izan* "to be" and **edin* "to become"); nonetheless, the reality of historical dialects is much richer and more complex. Thus, our decision has implied the inclusion of the paradigms of five more auxiliary verbs into the morphological analyzer: *egin* "to do", **eradun* [causative of **edun*], **erazan* [causative of **ezan*], *eutsi* "to keep" and **iron* [obscure semantics]. Again, this task has involved exhaustive philological and linguistic work, since three of these five dialectal verbal paradigms have been standardized for the first time. By including them, NLP tools for Basque improve significantly their performance when processing historical texts — by definition dialectal ones.

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

- OOV-Ald: "Variant" [from Basque *aldaera*], when the human annotator assigns the standard form to the non-standard variant (yellow-colored).
- OOV-AldADIJ: "Variant – finite verbal form" [from Basque *aldaera – aditz jokatua*], as the previous label, but specifying that the word is a finite verbal form.
- OOV-Zuz: "Correct", when the morphological analyzer does not recognize the word, but it has an entry in the OEH, the annotator marks it as a correct form (fuchsia-colored).
- STD-Ald: "Variant", when the form of the text matches a current standard Basque word ("false friend", yellow-colored).
- STD-AldADIJ: "Variant – finite verbal form", as the previous label, but specifying that the word is a finite verbal form.

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

limited number but relatively high frequency of such cases, a specific list of items is being compiled, to manually tag and annotate them.

Moreover, OOV-AldADIJ and STD-AldADIJ labels are added prospectively. Given that both synthetic verbal forms and auxiliary verbs will be difficult to normalize, it is appropriate to identify them for further normalization tasks and morphosyntactic analysis. In addition to this classification of variants, we have a label called SEK-Ber (from Basque *sekzio berezia* "special section", gray-colored). The linguist charged with tagging uses this mark in the paragraph (s)he is working on when (s)he finds a morphosyntactic phenomenon of diachronic interest. The label SEK-Ber works as a kind of reminder during the subsequent automatic morphosyntactic analysis, so that we can easily identify which sentences or fragments might be more difficult for the analyzer, and thus we may be able to focus on them. We resort to this label mostly — but not only — for linguistic phenomena no longer occurring in the language, for instance, the prosecutive case (which historically merged with the ablative), certain verbal roots such as *-(g)idi*, the perfect past periphrases with auxiliaries **edin* / **ezan* in a the main clause (*mana zezan* "(s)he commanded", present-day Basque *manatu zuen* [same gloss]), old synthetic constructions of the subjunctive with the auxiliaries *izan* / **edun*, or prescriptive forms such as *egin albaiteza* "you will do [imperative meaning]".

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

Finally, let us mention that we have estimated the time needed to normalize the whole corpus by hand. According to it, if the annotator tags 143 words per hour, (s)he would need about three years to treat the entire corpus. Since this is not feasible with the resources we have in this project, we chose to use computational techniques to normalize the corpus. In any case, it is not only a budgetary issue: making a virtue of necessity, the choice of automatic tagging leads to improvement in the automatic analyzer of the Basque language, as this tool is being trained in the historical and dialectal forms of Basque.

4.2.2 Automatic Normalization

To achieve the automatic normalization of our corpus we follow the normalization method presented by Etxeberria *et al.* (2019). This method is based on statistical learning and proposes a manually tagged sample for the learning process. All the pairs of words in the manual sample will be considered in the learning process; some of them will be tagged with

the label *Zuz* ("correct"), and the rest with the label *Ald* ("variant"), but we will use every pair for automatic learning. This method involves *Phonetisaurus*,[6] a grapheme-based phonological tool, guided by finite-state transducers (WFST) (Novak *et al.*, 2012, 2016). By introducing "variant/standard" and "standard/standard" pairs, the tool learns graph sequences. That is, we use *Phonetisaurus* to teach the changes that occur within the word pairs in the learning sample, and to create grapheme-to-grapheme models. After this training, the system should be able to propose standard forms for variants not previously analyzed.

After all this, we need to evaluate the quality of automatic normalization. We have applied the 10-fold cross-validation technique, dividing the manually tagged sample into ten parts. Of these ten parts, we used nine files for learning, and the remaining tenth to evaluate the results. We performed the same experiment ten times, changing the file used for evaluation, and we calculated the arithmetic average of these ten experiments to rate the normalization quality of each text (see Section 5).

4.3 Morphosyntactic Tagging

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

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

Finally, we review the morphosyntactic analysis performed by the automatic parser, with particular attention to paragraphs marked with the label SEK-Ber (see Section 4.2.1); if a structure is poorly analyzed by the tool, we correct it. In this respect, we need to formulate a number of new rules to properly parse morphosyntactic features that do not occur in present-day standard Basque. Thus, the parser will learn through manually corrected and formulated items to improve its performance. In other words, after completing this process we will have a morphosyntactic parser of historical Basque ready.

5. Learning through Normalization Tools: Results

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

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

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

	Leizarraga	Dechepare	*RS*
Only variants	87.68%	70.38%	66.83%
All words	94.24%	86.46%	80.66%

Table 2

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

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

text through what the automatic normalizer had learned from Leizarraga but, as expected, we obtained poor results, as Leizarraga's and Dechepare's linguistic varieties are quite different, particularly concerning non-standard forms: 75% for the whole of words, but only 46% for variants. Bearing this in mind, we decided to use only the manually normalized sample of Dechepare's text in the learning process. By doing so we improved the results, but they were not yet good enough to guarantee a successful normalization (59.58% for variants). To address this situation, we had two possible options: i) to tag more texts manually, or ii) to apply what was learned from Leizarraga's text processing to machine learning. Before starting tagging more by hand, we opted out for the second option, and the results improved significantly: 86.46% for all words and 70.38% for variants. The results of this last experiment suggest that in the future it might be a reasonable solution to have a normalization system for each dialectal area, instead of undertaking a normalization effort for each text.

The last text processed was *RS*, for which we got poorer results: 80.66% for all words and 66.83% for variants. The reason is that the very archaizing language considerably increases the difficulty of the normalization process. The obvious conclusion is that the more a text differs from the standard language, the more difficult its normalization becomes. Table 3 shows the numbers for standard words, variants, and entities in each text analyzed.

	Leizarraga	Dechepare	*RS*
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

Table 3

As shown in Table 3, the results obtained in the evaluation are completely consistent with the number of variants present in each text. In the case of *RS*, almost half of the words are not standard ones, and this complicates automatic normalization. The results obtained in *RS* point out that it is worthy to do a special effort in manual tagging when processing texts which differ from the standard language.

6. The Search Engine and the Interface

Computational linguists are building a search-engine-based web-application to consult the annotated data as well as information related to the project. The architecture and design of such application are based on different analytic tools. This includes part-of-speech tagging (Alegria *et al.*, 1996; Aduriz *et al.*, 2000; Alegria *et al.*, 2002), chunk analysis (Alegria *et al.*, 2008), and surface syntactic disambiguation (Arriola, 2015).

Thus, the final output of our project will be a search interface to browse the corpus. This interface must be useful for analyzing diachronic syntax. Therefore, it must be able to perform complex searches, both in terms of metadata (period, dialect, author, gender, etc.) and morphosyntactic characteristics (original form, lemma, part-of-speech, case ending, auxiliary verb root, time, aspect, mode, etc.), as well as any combination of them. A morphosyntactically annotated corpus will allow looking for structures that may be of interest from the perspective of Basque historical linguistics, for example:

- Postnominal relatives: search [noun + relative] structure.
- The so-called *archaic aorist*: search [verb + subjunctive past + main sentence], or [*EDIN lemma + past + main sentence] — **edin* is an intransitive auxiliary verb used in the expression of non-indicative values; thus, today its forms of the past express what conventionally is called *subjunctive*; however, in Archaic Basque such forms could be used to express perfective past in the main sentence.
- The special word order of negation: search [main verb + lemma EZ "not" + auxiliary verb] structure.
- Old periphrases of habituality: search [verbal noun + finite forms of the verbs *eroan* "carry"/ *eraman* "carry" / *joan* "go"] structure.

Finally, we want to stress another relevant feature of the search interface. When the user will browse the corpus and receive the list of results, (s)he will have the possibility to click on an example to display the page of the text in updated spelling, as well as the corresponding image of the facsimile, as shown in Figure 2.

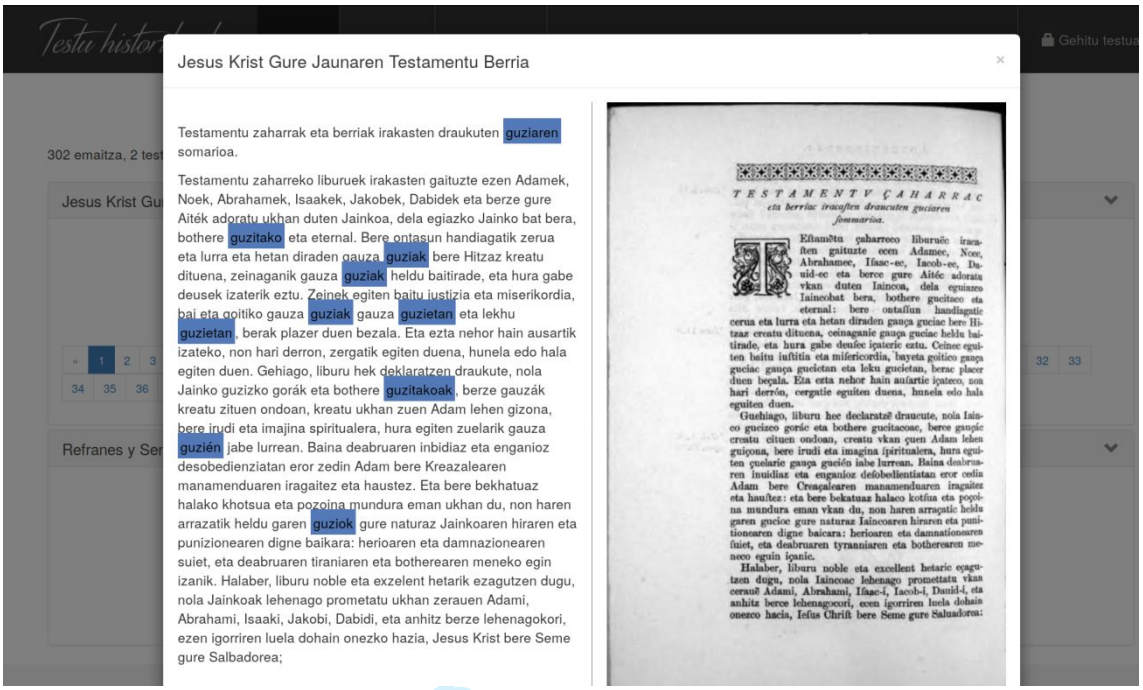

Figure 2

7. Conclusion

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

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

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

In the nearest future, the major challenge is the adaptation of automatic morphosyntactic analysis to historical Basque texts. What is more, we collect and consider theoretical linguists' proposals and suggestions to progress in the implementation of functionalities of the search engine and the design of the interface. In the longer term, our goal is the enhancement of the corpus, extending its time scope until the twentieth century — indeed a project has been recently funded to undertake the task.

To sum up, the corpus we are constructing aims to remove a major hindrance for the study of Basque diachronic syntax, which is the absence of such resources for this language. In this sense, the tool we are developing will be not only an important contribution to Basque Studies, but it also pursues the improvement of the tools for language processing of Basque and lays the foundations for future projects. Ultimately, some lessons learned from the work with highly diversified grammatical materials may inform projects on other languages that have similar aims.

Notes

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_hiztegianbilatu&view=frontpage&Itemid=410&lang=eu>

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

Funding

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

Acknowledgments

Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements | Acknowledgements

Captions

Table 1 Books printed in Basque between 1545 and 1879. The 1545-1749 columns roughly correspond to the span of our corpus (adapted from Sarasola, 1976)

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

Table 3 Size of each text, number, and proportion of the words labeled with the STD-Zuz, OOV, and ENT-Zuz tags

Figure 1 Interface for manual normalization of the corpus showing a fragment of Leizarraga's New Testament

Figure 2 The interface displays a page of Leizarraga's New Testament in both modernized spelling (left) and facsimile (right)

References

Aduriz, I., Agirre, E., Aldezabal, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K., Sarasola, K. and Urkia, M. (2000). A Word-Level Morphosyntactic Analyzer for Basque. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhauer (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens: European Language Resources Association (ELRA).

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G. and Lersundi, M. (2001). EDBL: A General Lexical Basis for the Automatic Processing of Basque. In S. Bird, P. Buneman and M. Liberman (Eds.), *IRCS Workshop on linguistic databases.* Philadelphia: University of Pennsylvania. Institute for Research in Cognitive Science.

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

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

Alegria, I., Arrieta, B., Carreras, X., Díaz de Ilarraza, A. and Uria, L. (2008). Chunk and Clause Identification for Basque by Filtering and Ranking with Perceptrons. *Procesamiento del Lenguaje Natural, 41*, 5–12.

Alegria, I., Artola, X., Sarasola, K. and Urkia, M. (1996). Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, *11*(4), 193–203.

Arriola, J. M. (2015). Different Issues in the Design and Implementation of a Rule Based Grammar for the Surface Syntactic Disambiguation of Basque. In E. Bick and K. Hagen (Eds.), *Proceedings of the Workshop on Constraint Grammar - methods, tools and applications; at NODALIDA 2015* (pp. 1–9). Vilnius: Institute of the Lithuanian Language.

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

Etxeberria, I. (2016). *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz* [= Normalization of linguistic variants using phonological and morphological inferences]. (Unpublished doctoral dissertation). University of the Basque Country, San Sebastián.

Etxeberria, I., Alegria, I. and Uria, L. (2019). Weighted finite-state transducers for normalization of historical texts. *Natural Language Engineering, 25*(2), 307–321.

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

Galves, Ch., Andrade, A. and Faria, P. (2017). *Tycho Brahe Parsed Corpus of Historical Portuguese*.
http://www.tycho.iel.unicamp.br/~tycho/corpus/ (accessed 29 January 2021).

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

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

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

Kroch, A., Santorini, B. and Delfs, L. (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3. http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3 (accessed 29 January 2021).

Kroch, A., Santorini, B. and Diertani, A. (2016). *The Penn Parsed Corpus of Modern British English (PPCMBE2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1. http://www.ling.upenn.edu/ppche-release-2016/PPCMBE2-RELEASE-1 (accessed 29 January 2021).

Kroch, A. and Taylor, A. (2000). *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4. http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4 (accessed 29 January 2021).

Lafon, R. (1944). *Le système du verbe basque au XVI^e siècle*. Bordeaux: Éditions Delmas.

Lafon, R. (1951–1952). Concordances morphologiques entre le basque et les langues caucasiques. *Word 7,* 227-244 and *8,* 80–94.

Lakarra, J. A. (1995). Reconstructing the root in Pre-Proto-Basque. In J. I. Hualde, J. A. Lakarra and R. L. Trask (Eds.), *Towards a History of Basque Language* (pp. 189–206). Amsterdam: John Benjamins.

Lakarra, J. A. (2005). Prolegómenos a la reconstrucción de segundo grado y análisis del cambio tipológico en (proto) vasco. *Palaeohispanica, 5,* 407–459.

Lakarra, J. A. (2006). Protovasco, munda y otros: reconstrucción interna y tipología holística diacrónica. *Oihenart: cuadernos de lengua y literatura, 21,* 229–322.

Lash, E. (2014). *The Parsed Old and Middle Irish Corpus (POMIC)*. Version 0.1.

https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/ (accessed 29 January 2021).

Manterola, J. (2015). *Towards a History of Basque Morphology: Articles and demonstratives.* (Unpublished doctoral dissertation). University of the Basque Country, Vitoria-Gasteiz.

Martínez-Arena, M. (Ed.) (2013). *Basque and Proto-Basque. Language-Internal and Typological Approaches to Linguistic Reconstruction.* Frankfurt: Peter Lang.

Michelena, L. (1977 [1961]). *Fonética histórica vasca.* San Sebastián: Council of Guipuscoa.

Michelena, L. (1964). *Textos Arcaicos Vascos*. Madrid: Minotauro.

Michelena, L. and Sarasola, I. (Eds.) (1987–2005). *Orotariko Euskal Hiztegia* [= Basque General Dictionary]. Bilbao: Royal Academy of the Basque Language. https://www.euskaltzaindia.eus/index.php?option=com_oeh&view=frontpage&Itemid=413&lang=eu (accessed 29 January 2021).

Nelson, M. (2010). Building a written corpus What are the basics? In Anne O'Keeffe and Michael MacCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge.

Novak, J. R., Minematsu, N. and Hirose, K. (2012). WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In I. Alegria and M. Hulden (Eds.), *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing* (pp. 45–49). San Sebastián: Association for Computational Linguistics.

Novak, J. R., Minematsu, N. and Hirose, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, *22*(6), 907–938.

Sarasola, I. (1976). *Historia Social de la Literatura Vasca.* Madrid: Akal.

Sarasola, I. (1983). *Contribución al estudio y edición de textos antiguos vascos*. San Sebastián: Council of Guipuscoa and University of the Basque Country.

Satrústegui, J. M. (1987). *Euskal Testu Zaharrak (I)* [= Old Basque texts]. Pamplona: Royal Academy of the Basque Language.

Schuchardt, H. (1907). *The Iberische Deklination.* Vienna: Holder.

Schuchardt, H. (1914). Baskisch und Hamitisch. *Revista Internacional de Estudios Vascos, 8*(1), 76.

Trask, R. L. (1997). *The History of Basque.* London/New York: Routledge.

Uhlenbeck, C. C. (1924). De la possibilité d'une parenté entre me basque et les langues caucasiques. *Revista Internacional de Estudios Vascos, 15,* 565–588.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F. and Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC)*. Version 0.9. http://www.linguist.is/icelandic_treebank (accessed 29 January 2021).