Dealing with dialectal variation in the construction of the Basque historical corpus

Ainara Estarrona[1], **Izaskun Etxeberria**[1], **Ricardo Etxepare**[2],
Manuel Padilla-Moyano[2], **Ander Soraluze**[1]
[1]HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
[2]CNRS - IKER (UMR 5478)
{ainara.estarrona}{izaskun.etxeberria}{ander.soraluze}@ehu.eus
{r.etxepare}{manuel.padilla}@iker.cnrs.fr

Abstract

This paper analyses the challenge of working with dialectal variation when semi-automatically normalising and analysing historical Basque texts. This work is part of a more general ongoing project for the construction of a morphosyntactically annotated historical corpus of Basque called *Basque in the Making (BIM): A Historical Look at a European Language Isolate,* whose main objective is the systematic and diachronic study of a number of grammatical features. This will be not only the first tagged corpus of historical Basque, but also a means to improve language processing tools by analysing historical Basque varieties more or less distant from present-day standard Basque.

1 Introduction

In many languages other than Basque, different historical corpora exist which are annotated morphologically and syntactically, and allow lexical, morphological or syntactic searches in historical texts, e.g. the *Penn Parsed Corpora of Historical English* (Kroch and Taylor, 2000; Kroch et al., 2004; Kroch et al., 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). However, no appropriate instruments of this type have ever been developed for Basque.

In this paper we present the work we are carrying out in the construction of a historical corpus of Basque considering the dialectal variation of historical texts. This work is part of a more general ongoing project called *Basque in the Making (BIM): A Historical Look at a European Language Isolate*[1], which has two main objectives: First, an exhaustive diachronic study of different grammatical features of the Basque language; second, the creation of a morphosyntactically annotated historical corpus that will enable comprehensive diachronic analysis.

The final output of our project will be a search interface to browse the corpus. This interface must be useful for analysing 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. Taking all this into account, BIM is an interdisciplinary project, where experts on Linguistics (IKER centre)[2] and Natural Language Processing (HiTZ centre)[3] work together.

This being the general scenario of the project, in this article we will focus on how we are dealing with dialectal variation in text normalisation. Together with that, we will also mention the first steps taken towards the adaptation of the morphosyntactic analyser of standard Basque developed by the Ixa group to be able to correctly analyse historical texts (by definition dialectal ones).

After having presented the overview of the project, in Section 2 we will make an exposition of the related work. In Section 3 we will present the corpus we are working with. The first steps taken in the

[1]http://ixa2.si.ehu.es/bim/en
[2]https://www.iker.cnrs.fr/?lang=fr
[3]http://hitz.eus/en

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 79–89
Barcelona, Spain (Online), December 13, 2020

adaptation of the morphosyntactic analyser will be presented in Section 4. Section 5 will be dedicated to the text normalisation process. In Section 6 we will outline and discuss the main experimental results and the normalisation strategy adopted in the light of the results. Finally, we will review the main conclusions and future work in Section 7.

2 Related work

Historical and dialectal texts present problems from an NLP point of view, since NLP tools developed for contemporary standard language often fail in handling the linguistic varieties encountered in such texts. A majority of NLP tools are designed to process newspaper texts written in contemporary language, but the characteristics of the standardised modern texts are not shared by historical texts: standard variants adhere to orthographic and grammatical norms which may be comparatively recent in the written corpus of the language. Therefore, the creation of a morphosyntactically annotated historical corpus using standard NLP tools needs a previous step such as normalisation of the texts.

Text normalisation has attracted a lot of interest over the past years, particularly normalisation of historical and dialectal texts, but also of informal texts such as those collected on Twitter (Alegria et al., 2014). Several techniques have been used for this task, but it can be said that nowadays machine-learning based techniques are the most popular ones, that is, systems that learn from examples of standard-variant pairs. First methods used for automatic normalisation of historical texts were rule-based methods. For instance, in the construction of the *Tycho Brahe Corpus of Historical Portuguese*, re-write rules were used to normalise historical texts (Hirohashi, 2005). These methods do not need hand-annotating training data, but they do need linguist experts to detect which rules to apply in the normalisation process, which is costly and not always successful.

Character level Statistical Machine Translation techniques (CSMT) have been applied by Pettersson et al. (2013) by treating normalisation task as a translation problem. Then, in Pettersson et al. (2014) they evaluate and compare their approach with a memory-based filtering and a Levenshtein-based approach considering five languages. The SMT-based approach generally works best.

Using the same CSMT approach Scherrer and Erjavec (2016) develop a language-independent word normalisation method and test it on a task of modernising historical Slovene words. They perform two sets of experiments: supervised and unsupervised. In the first one, they use the lexicon of word pairs as training data to build a CSMT system. In the second one, they simulate a scenario in which word pairs are not available. They show that both methods produce significantly better results than the baselines.

In our previous works (Etxeberria et al., 2016; Etxeberria et al., 2019) a different approach is presented. The method learns to map phonological changes using a noisy channel model that combines weighted finite-state transducers (WFST) and language models. In Etxeberria et al. (2019) our approach is compared with the CSMT methods explained in Pettersson et al. (2014) and in Scherrer and Erjavec (2016) using same historical corpora. The results show that the WFST approach produces similar or better scores for the six languages tested.

Lately, neural network architectures have become popular for a variety of NLP tasks, and they have been also applied to text normalisation (Korchagina, 2017; Bollmann, 2018; Tang et al., 2018). The results obtained by deep learning are good. However, these methods usually train on a lot of manually labelled data. In Tang et al. (2018) eight different NMT models are applied to the spelling normalisation task in several languages (English, German, Hungarian, Icelandic, and Swedish). The authors carry out particular experimentation on Swedish by increasing the size of the training set and they conclude that the performance of NMT models is highly related to the size.

3 The corpus

The Basque historical corpus covers the most representative written production between the 15th and the 18th century. It is a time span within which all historical dialects of Basque are represented and it is also the temporal span that divides Archaic and Old Basque from Early Modern Basque[4]. Texts have been

[4]We are already working on another project called SAHCOBA and funded by Spanish Ministry of Science and Innovation (MICINN, RTI2018-098082-J-I00) which addresses both Early and Late Modern Basque.

selected on the basis of: 1) their representativeness; 2) the existence of reliable editions; and 3) their social context.

As far as corpus size is concerned, our goal is to get as much text as possible, because the wider the corpus, the more relevant the results of the research become. At the time being, we are creating a reference corpus of around one million words. Given the problems and limitations associated to the written past of languages, specially in a language like Basque, this range is considered acceptable for a historical corpus (Claridge, 2009).

3.1 The reference sub-corpus

Basque is an extremely fragmented language; a number of dialects and sub-dialects spread over an area of 10,000 km2 (see Figure 1). The dialectal split began in the early Middle Ages (Mitxelena, 1981), and during the last centuries, the linguistic distance between dialects has been increasing to the extent that today peripheral varieties are not mutually intelligible in oral speech by non-trained speakers.

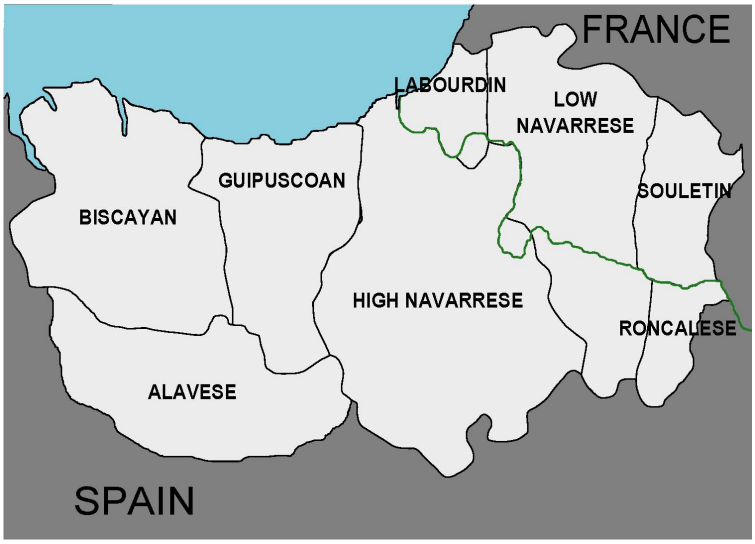

Figure 1: Historical Basque dialects. Alavese and Roncalese are extinct varieties. The green line represents the French-Spanish border.

At this point mention must be made to standard Basque, also known as unified Basque. Standard Basque is a literary variety constructed upon central dialects of the language. It was formulated in the 1960's, and today it has become the reference for speakers of all dialects, as well as the official form of the language. The Royal Academy of the Basque Language (1919) is the institution charged for the construction of unified Basque, and its decisions have normative character. The basis of standard Basque is formed by a spelling system, paradigms of noun- and verb morphology, syntactic rules, and an official dictionary. Historical dialects differ from standard Basque to different degrees; the most distant varieties are Biscayan in the West, and Souletin in the East.

Hence, in this corpus we are working with texts from different periods and dialects, and such diversity of materials adds complexity to the normalisation task. Consequently, we are carrying out this normalisation process in two phases. In the first phase we perform a manual normalisation of a sample of each text, and in the second phase, based on this manual work, we use computational techniques for the automatic normalisation of the rest of the text (see Section 5.2 for more details).

The manual work of normalising the whole corpus would take too much time. For this reason, we have created a sub-corpus which includes one or two texts representative of each dialect at each period. These works will be annotated semi-automatically and based on the results obtained in this sub-corpus the rest of the corpus will be annotated fully automatically. The works included in this sub-corpus are works that are relevant to the history of Basque and that, moreover, reflect the main characteristics of each historical dialect. This sub-corpus contains about 675,000 words, which is just over half of the total corpus. In

Table 1 we can see graphically each work included in the sub-corpus in its corresponding century and dialect[5]:

	G	B	Al	L	LN	HN	S
16th cent.		*RS*	Lazarraga / Bet	Lç	E		
17th cent.		Cap		Ax / Mat	Tt, *Onsa*	Ber, *Doc*	Bp
18th cent.	Lar, *ET*	Arz		He / Ch	AR	El	Mst

Table 1: G > Gipuscoan; B > Biscayan; Al > Alavese; L > Labourdin; LN > Low Navarrese; HN > High Navarrese, and S > Souletin.

Table 1 does not display the small 15th century texts compiled in two collections that will be mentioned in Section 5.1 (*TAV* and *Contr*). The gaps in the table mean that there are no works for that dialect and that century.

With regard to this paper in Section 6 we will present the results of normalisation for texts corresponding to the 16th and 17th centuries[6]

4 Adaptation of the morphosyntactic analyser

The morphosyntactic analyser was created for standard Basque, but dialectal texts show significant differences at all linguistic levels. First, there is an array of morpho-phonological phenomena, which confer a particular shape to every dialect; second, there are divergences in the inventory of case markers and other elements of noun morphology; third, extreme degree of variation in verbal morphology, which involves even different auxiliary verbs; last, but not least, specific lexical items.

Therefore, when we face the challenge of analysing historical and dialectal texts using natural language processing tools, we have two main tasks. On the one hand, to adapt the tools available for the standard so that they are capable of analysing historical texts, and on the other hand, to normalise or standardise historical texts so that the NLP tools are able to work with them. We believe it is necessary to undertake both tasks and in this section we will explain the preliminary work we have done for the adaptation of the morphosyntactic analyser of Basque.

The main reference for normalisation has been the dictionary of the Royal Academy of the Basque Language[7], but as far as morphology is concerned, we have followed two different criteria: i) concerning noun morphology, to prioritise the standard forms of case markers: *-rano* → *-raino* (terminative marking), *-rekilako* → *-rekiko* (comitative and relational suffix agglutination), *-akgatik* → *-engatik* (causal case); ii) with regards to verbal morphology, to preserve the roots of auxiliary verbs not belonging to standard Basque. And it is in this second point where we have had to carry out an adaptation work on the morphosyntactic analyser Eustagger (Alegria et al., 2002).

Standard Basque has four auxiliary verbs: two transitives (**edun*[8] "to have" and **ezan* [obscure semantics][9]) and two intransitives (*izan* "to be" and **edin* "to become"). Nevertheless, the historical dialects were much richer and more complex. Our goal has been to make our tools capable of collecting and correctly analysing all that complexity and that is why we have included the paradigms of five more auxiliary verbs in the morphosyntactic analyser. These auxiliary verbs are the following: *egin* "to do", **eradun* [causative of **edun*], **erazan* [causative of **ezan*], *eutsi* "to keep" and **iron* [obscure semantics][10]. This task has involved exhaustive philological and linguistic work, since three of these five

[5]For the abbreviations of the works we refer to the *Basque General Dictionary* of the Royal Academy of the Basque Language (*Orotariko Euskal Hiztegia)*: https://www.euskaltzaindia.eus/components/com$_o ehberria/pdf/02 - erreferentzia - bibliografikoak.pdf$

[6]Due to time constraints we cannot present the results of Tt, *Onsa* as we have not yet finished the manual annotation of this work.

[7]https://www.euskaltzaindia.eus/en/

[8]We use the symbol '*' to refer to reconstructed or unattested forms.

[9]Reconstructed form, its semantics is not clear, although some etymological connection with the verbal root *za "to be" can be postulated.

[10]Reconstructed verb. The forms of **iron* are restricted to the expression of modal meanings with transitive verbs.

dialectal verbal paradigms (**eradun, *erazan* and **iron*) have been standardised for the first time. By including them, on the one hand, NLP tools for Basque improve significantly their performance when processing historical and dialectal texts, and on the other hand, we are able to collect this dialectal variation to enable linguists to search for these auxiliary verbs that no longer exist in standard Basque. For this process the standard transducer used by Eustagger is extended with new lexical entries and phonological rules obtained after the philological and linguistic work. It should be mentioned that this extended version of the analyser is only used to analyse historical texts, while the previous version is the one used for standard Basque, i.e. we now have two analysers, one for standard and one for historical texts.

5 Methodology for text normalisation

As we said before text normalisation is a necessary step in this project. Once ancient and dialectal texts are normalised, the NLP tools developed for standard Basque could be applied for the linguistic analysis of corpora.

The normalisation method used is the WFST-based method presented in Etxeberria et al. (2016) and Etxeberria et al. (2019) and used also in Estarrona et al. (2019). This approach uses *Phonetisaurus*, a WFST-driven phonology tool (Novak et al., 2012; Novak et al., 2016), based on OpenFST (Allauzen et al., 2007), which learns mapping of phonological changes, using a noisy channel model. It is the strongest method found in previous work with dialect normalisation in Basque (Etxeberria et al., 2014) and when it has been tested with other languages in order to be compared with the CSMT methods proposed in Pettersson et al. (2014) and Scherrer and Erjavec (2016) it has obtained similar results (Etxeberria et al., 2019). For the moment we have ruled out neural methods due to the features mentioned of our corpus: several historical periods, several dialects and short texts. Further research is necessary for a robust solution based on neural methods.

The process of text normalisation will be carried out in two phases. First, part of the text will be manually annotated, and then, the rest of the text will be automatically normalised.

The purpose of this article is not to explain the normalisation process that has already been explained in detail in Etxeberria et al. (2019) and Estarrona et al. (2019); therefore here we will only give a brief summary of each of the normalisation phases.

5.1 Manual normalisation

Previous work on normalisation of Basque historical texts (Etxeberria, 2016) has shown that manual normalisation of 10% of the text is sufficient to obtain satisfactory results in automatic normalisation. Thus, we have randomly collected 10% of each of the texts. However, we have also found that it is necessary for the work to be of a minimum length for the results to be acceptable, so we have decided that works with less than 6,000 tokens will be normalised manually in their entirety. Following this criterion, we have manually normalised two works of the 16th century: i) the anonymous compendium of sayings *Refranes y Sentencias* (3,078 tokens) written in an archaic Biscayan dialect, and ii) the Christian doctrine of Betolaza (1,050 tokens) written in the Old Alavese dialect. In addition to these two works, three collections which include small texts from the 15th, 16th and 17th centuries written in different dialects have been manually labelled: *Textos Arcaicos Vascos* (TAV) by Michelena, *Contribución al estudio de Textos Arcaicos Vascos* (Contr) by Sarasola and *Euskal Testu Zaharrak* (ETZ) by Satrustegi.

The previous step to manual normalisation is to pre-process the text. This pre-processing consists of three phases: tokenization, named-entity recognition and lexical recognition. After this pre-processing, all the words that have not been recognised and that are therefore likely to have to be normalised will be marked in the text with an OOV ('out of vocabulary') tag. The task of the linguistic annotator is to normalise all those words marked as OOV, that is, to assign to each of them its corresponding standard based on the dictionary of the Royal Academy of the Basque Language, and also to check the rest of the words to identify possible false friends.

The manual normalisation is done using the BRAT tool (Stenetorp et al., 2012)[11] and the task is carried out by a single linguistic annotator with training in historical texts. The time needed to manually

[11] http://brat.nlplab.org/index.html

annotate the whole corpus has been estimated and it can be seen that about 143 words per hour are labelled. Therefore, a person would need about 4 years to label the entire corpus manually, and this is obviously why we plan to implement computational techniques to normalise the corpus.

5.2 Automatic normalisation

For the automatic normalisation process we have reapplied the method previously cited in Etxeberria et al. (2019). This method uses *Phonetisaurus*[12], a Weighted Finite State Transducer (WFST) driven phonology tool (Novak et al., 2012; Novak et al., 2016).

After collecting the word pairs (variant-standard) tagged in the manual normalisation process into a dictionary, the application of the tool requires three steps: a) sequence alignment: in this step the historical and standard words are aligned, obtaining a joint grapheme/grapheme chunks used in the next step; b) model training: using the aligned data obtained in the previous step, a model is trained and converted to a WFST; c) decoding: given the WFST obtained in the previous step, the decoder finds the best hypothesis for the input words.

We use the *Phonetisaurus* tool to learn the changes that occur within the word pairs (variant-standard) in the learning set, which by itself produces a grapheme-to-grapheme system. Once this model is trained and converted to a WFST format, it can be used to generate correspondences between previously unseen words and modern standard forms.

6 Results and discussion

In our previous work (Estarrona et al., 2019), the phonological induction inspired method we have chosen has been evaluated on three works of the 16th century selected from the sub-corpus: *Refranes y Sentencias*, Leizarraga's translation of the New Testament and Etxepare's *Linguae Vasconum Primitiae*.

The method achieved an accuracy value between 80% and 95% in the experiments and the results suggested, on the one hand, that automatic normalisation should be done taking into account for the learning process texts from the same dialect or nearby dialects and, on the other hand, that the more text we have for the learning process the better the results we get.

The main objective of the set of experiments presented below is to validate these hypotheses in order to define the normalisation strategy to be followed in the texts that will not be manually tagged.

6.1 Experiments and results

In order to evaluate the quality of the automatic normalisation method used we have performed 10-fold cross-validation experiments on the manually annotated text. The text selected for manual annotation has been divided into 10 files, therefore getting the 10 sets for learning and testing has been straightforward.

We have carried out the same experiments for all the works included in the 16th and 17th centuries[13]. To compare the improvements of our method, first we have calculated the baseline values using two approaches: 1) by giving the input word as output; 2) by looking at the memorised word pairs (variant-standard) and if there is an input word form, providing its normalised version. In case the word form is not memorised, we give as output the input word.

After the baseline has been set, in the first experiment we have evaluated the normalisation considering only the evaluated work. The second experiment has consisted in using for the learning process texts from the same dialectal region of the evaluated work, and for that purpose we have defined two major dialectal groups, the Eastern dialects (Labourdin, Low Navarrese and High Navarrese) on the one hand, and the Western dialects (Biscayan and Alavese) on the other (see Figure 1). Finally, in the third experiment, we have used for learning all the texts manually normalised so far.

All the word pairs in the learning set are taken into account for the learning process. However, we have prepared two different lists of word pairs for the test set: the first one contains all the word pairs in the test set including those which do not need to be normalised, and the second one contains only the

[12] https://github.com/AdolfVonKleist/Phonetisaurus

[13] Except *Refranes y Sentencias* which was manually tagged in its entirety and Tt, *Onsa* which we have not labelled yet.

word pairs labelled as variants, those which need to be normalised. Nevertheless, we have to point out that the real scenario we will always work on is that of considering all the words in the text.

Results are given based upon accuracy, i.e., the percentage of words that have been normalised correctly, and are presented in Table 2.

	Baseline1	Baseline2	Test-all	Test-variants
Leizarraga	63.08	82.13	**94.24**	**87.68**
Leizarraga + 1	-	-	94.10	87.87
Leizarraga + 3	-	-	92.67	88.00
Axular	72.89	87.51	**96.75**	**90.87**
Axular + 1	-	-	96.57	90.53
Axular + 3	-	-	94.35	88.74
Materra	72.43	82.32	93.11	77.86
Materra + 1	-	-	**96.56**	**89.42**
Materra + 3	-	-	94.42	89.17
Etxepare	56.45	72.54	73.88	59.58
Etxepare + 1	-	-	87.63	73.69
Etxepare + 3	-	-	**90.90**	**84.01**
Beriain	61.43	80.43	91.69	81.31
Beriain + 1	-	-	**92.73**	**82.88**
Beriain + 3	-	-	91.71	85.11
Belapeire	44.09	68.43	**87.82**	**81.40**
Belapeire + 1	-	-	87.69	80.07
Belapeire + 3	-	-	86.28	80.50
Lazarraga	47.57	64.30	77.71	64.00
Lazarraga + 2	-	-	81.13	70.00
Lazarraga + 3	-	-	**84.13**	**75.06**
Kapanaga	41.76	61.94	80.08	70.87
Kapanaga + 2	-	-	**84.50**	**78.35**
Kapanaga + 3	-	-	84.21	77.31

Table 2: 1= Eastern dialect works (Labourdin, Low Navarrese, High Navarrese); 2= Western dialect works (Biscayan and Alavese); 3= All the works.

6.2 Discussion

As we can see in Table 2 our normalisation method improves in all cases the two baselines we have defined for the task. In the case of Etxepare the improvement between Baseline2 and our method is only slightly more than one point. This may be due to the fact that this work is very short (6,842 tokens) and therefore we do not have enough volume of manually annotated text for the learning process. This would be confirmed by the fact that in this case the best results are obtained by including all the texts we have in the learning process (90.90% and 84.01%).

As we have said in Section 6.1 before starting the experiments we had raised two main hypotheses: i) normalisation should be done taking into account for the learning process the nearby dialects, and ii) the more text for learning, the better the results.

The main conclusion we have drawn from this set of experiments is that the performance of normalisation improves substantially when treating texts of the same dialect or dialectal group for learning purposes which validates our first hypothesis. Nevertheless, in the case of Leizarraga and Axular, the results do not improve (rather they worsen somewhat) by including more texts in the learning process, perhaps because, being two long texts (77,780 and 102,000 tokens respectively) and both of them close to standard Basque as shown by the Baseline1, the results obtained in the first experiment leave little room for improvement. In addition, the fact that they are two works very close to standard Basque means that

including variants of other dialects has a negative influence on the results. Materra's work also reflects a variety very close to the standard, however it is a short work (6,750 tokens) and probably for this reason, the results of the second experiment benefit from increasing the volume of text in the learning process.

This leads us to the second major hypothesis of this normalisation work: the more volume of text we have for learning, the better the results. We have validated this hypothesis in the case of short texts as Etxepare's which achieves the best results in the third experiment when we use all the texts we have for the learning process. However, we have also seen in the rest of the works that including more text in the learning process without taking into account the dialectal variety does not improve the results. In the case of Lazarraga the results also improve with experiment 3. We have two possible explanations for this. On the one hand, it is not a very extensive work (12,500 tokens) and it benefits from the increase of text in the learning process. On the other hand, it is a work that reflects a rather archaic variety of the Alavese dialect and in experiment 3 we have incorporated all the works manually annotated up to now including the two collections that involve the more archaic texts: *TAV* and *Contr*. In addition, this work is written in a more eastern variety of the Alavese dialect and this would make the texts that do not belong to the Western dialect group help in the results.

We can therefore conclude that the second hypothesis alone would not be sufficient to achieve the best results, i.e. including more text in the learning process without taking into account dialectal variety, does not ensure improved results.

Consequently, instead of having a system for each work or a single general system for all works, we will implement a system for each dialect or dialects that are linguistically proximate, including for learning what is learned in all the works that belong to that group of dialects.

In the case of the Souletin dialect, as we can see in Table 2, the best results for Belapeire's work, written in this variety, are obtained by using the text itself. It is true that the same occurs in both Leizarraga and Axular, although the main difference between these works and Belapeire's, apart from the size, is that the first two are very close to the standard, while the third differs considerably as shown by Baseline1.

These results are coherent with the marginal character of this language variety; indeed, Souletin differs to a great extent from the rest of the Basque dialects, because it has countless specific distinctive features at all linguistic levels: phonology, morphology, syntax and lexicon.

In order to validate this fact we wanted to test the influence of including Souletin in the learning process to normalise other works of the group of eastern texts. The two works linguistically closest to Souletin are those of Etxepare and Beriain, written in Low and High Navarrese respectively, as they are the most eastern varieties outside Souletin. For this reason, we think that if Souletin could be of help in normalisation, it would be in these two works.

Nevertheless, the results shown in Table 3 did not confirm our hypothesis, since the inclusion of Souletin in the learning process did not improve the results obtained for the works of Etxepare and Beriain which curiously remain at exactly the same value. This reinforces the thesis of historical dialectology that characterises Souletin as a marginal dialect with respect to the other varieties of Basque.

	Test-all	Test-variants
Etxepare + 1	87.63	73.69
Etxepare + 1 + Bp	87.63	73.69
Beriain + 1	92.73	82.88
Beriain + 1 + Bp	92.73	82.88

Table 3: 1= Eastern dialect works (Labourdin, Low Navarrese, High Navarrese); Bp= Belapeire.

In view of the results obtained in the experiments, we have decided for our normalisation strategy to treat Souletin separately from the other eastern dialects.

7 Conclusions and future work

In this paper we have presented the results obtained in the normalisation of Basque historical texts using a Weighted Finite State Transducer (WFST) driven phonology tool and taking into account the dialectal

variation when designing the experiments. Our initial hypotheses were two: i) by using texts from linguistically close dialects in the learning process, the results would improve, and ii) the results would improve as more text is fed. The results presented confirm the first hypothesis, since, in general, they are substantially improved by including in the learning process texts from the same dialectal group. However, the second hypothesis is not validated by the results, since in most cases by including all the texts available in the learning process, the results are worse than those obtained with texts from nearby dialects. This is with the exception of Etxepare and Lazarraga, but in both there are reasons that would explain these results and which have been discussed in Section 6.2. These results have led us to conclude that the best normalisation strategy for texts that will be normalised in a fully automatic way is to create a normaliser for each dialectal group. In the same way, the results obtained with the work written in Souletin have confirmed that given the marginal character of this dialect, the normalisation must be done separately from the rest of the eastern dialects.

We would like to underline that the results obtained reinforce the theories of historical dialectology in a way that has not been explored so far, that is, using computational methods. We would like to continue exploring this line of research in future work and in collaboration with dialectologists.

On the other hand, we have presented the preliminary work carried out to adapt the morphosyntactic analyser for standard Basque in order to correctly recognise and analyse the historical and dialectal variation of auxiliary verbs. Thanks to this work we have included five new paradigms in the lexicon used by the analyser. We have to emphasise that from the philological and linguistic point of view this has been a significant work, since three of these five dialectal verbal paradigms have been standardised for the first time.

The next steps in this project are the following: i) the semi-automatic normalisation of the 18th century works included in the reference sub-corpus; ii) the automatic normalisation of the rest of the corpus, and finally, iii) the morphosyntactic analysis of the corpus. For the normalisation we have already defined the strategy to follow, and for the automatic morphosyntactic analysis of the corpus we will use the Eustagger tool. A sample of this automatic analysis will be revised manually to detect errors and to proceed to the annotation of interesting morphosyntactic phenomena from the point of view of diachronic syntax. This syntactically annotated corpus will facilitate the systematic study of a number of grammatical features of the Basque in a diachronic way by means of a search interface, on which we are already working, and it will be the only tool of these characteristics existing for the Basque language. The annotated corpus and the search interface will be public and freely available to the research community.

Finally, we think it might be interesting to test the performance of different methods of normalisation based on neural networks. We are aware of the limitations of our corpus, in terms of its extension, in order to achieve adequate results using methods based on neural networks. For this reason, we have decided to work on this line of research in the elaboration of the corpus that will include the next stages in the history of the Basque language, Early and Late Modern Basque. We are already working on the choice of the works that will make up this new corpus. With this new project, we plan to extend the historical corpus to 12 million words, which would allow us to work with more guarantees on methods based on neural networks. In any case, there are authors (Moeller et al., 2019) who have tried different strategies to deal with these types of limitations and we believe that it would be an interesting line of investigation in the case of the Basque language.

Acknowledgements

The research leading to these results was carried out as part of the *Basque in the Making (BIM): A Historical Look at a European Language Isolate* project (ANR-17-CE27-0011 - BIM, Agence Nationale de la Recherche, France) and the *Syntactically Annotated Historical Corpus in Basque* (SAHCOBA, RTI2018-098082-J-I00) project (Ministry of Science and Innovation (MICINN), Spain).

References

I. Alegria, M.J. Aranzabe, N. Ezeiza, A. Ezeiza, and R. Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6.

I. Alegria, N. Aranberri, P. Comas, V. Fresno, P. Gamallo, L. Padró, I. San Vicente, J. Turmo, and A. Zubiaga. 2014. Tweetnorm_es: an annotated corpus for spanish microtext normalization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2274–2278. European Language Resources Association (ELRA).

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer.

M. Bollmann. 2018. *Normalization of historical texts with neural network models*. Ph.D. thesis, Bochum, Ruhr-Universität Bochum.

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

A. Estarrona, I. Etxeberria, A. Soraluze, and M. Padilla-Moyano. 2019. Spelling Normalisation of Basque Historical Texts. *Procesamiento del Lenguaje Natural*, 63:59–66.

I. Etxeberria, I. Alegria, M. Hulden, and L. Uria. 2014. Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural*, 52:13–20.

I. Etxeberria, I. Alegria, L. Uria, and M. Hulden. 2016. Evaluating the Noisy Channel Model for the Normalization of Historical Texts: Basque, Spanish and Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1064–1069.

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

I. Etxeberria. 2016. *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz*. Ph.D. thesis, Universidad del Pais Vasco / Euskal Herriko Unibertsiatea.

Ch. Galves, A. Andrade, and P. Faria. 2017. Tycho Brahe Parsed Corpus of Historical Portuguese. url: http://www.tycho.iel.unicamp.br/ tycho/corpus/texts/psd.zip.

A. Hirohashi. 2005. *Aprendizado de regras de substituição para normatização de textos históricos*. Ph.D. thesis.

N. Korchagina. 2017. Normalizing Medieval German Texts: from rules to deep learning. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 12–17.

A. Kroch and A. Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). url: http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4.

A. Kroch, B. Santorini, and L. Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). url: http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3.

A. Kroch, B. Santorini, and A. Diertani. 2016. The Penn Parsed Corpus of Modern British English (PPCMBE2). url: http://www.ling.upenn.edu/ppche-release-2016/PPCMBE2-RELEASE-1.

E. Lash. 2014. The Parsed Old and Middle Irish Corpus (POMIC). url: https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irishcorpus-pomic/.

K. Mitxelena. 1981. Lengua común y dialectos vascos. *International Journal of Basque Linguistics and Philology*, 15:291–313.

S. Moeller, G. Kazeminejad, A. Cowell, and M. Hulden. 2019. Improving Low-Resource Morphological Learning with Intermediate Forms from Finite State Transducers. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 81–86.

J.R. Novak, N. Minematsu, and K. Hirose. 2012. WFST-Based Grapheme-to-Phoneme Conversion: Open Source tools for Alignment, Model-Building and Decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49.

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

E. Pettersson, B. Megyesi, and J. Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, volume 18, pages 54–69.

E. Pettersson, B. Megyesi, and J. Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. *Proceedings of LaTeCH*, pages 32–41.

Y. Scherrer and T. Erjavec. 2016. Modernising historical Slovene words. *Natural Language Engineering*, 22(6):881–905.

P. Stenetorp, S. Pyysalo, G. Topić, T. TOhta, S. Ananiadou, and J. Tsujii. 2012. Brat: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

G. Tang, F. Cap, E. Pettersson, and J. Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. *arXiv preprint arXiv:1806.05210*.

J.C. Wallenberg, A.K. Ingason, E.F. Sigurosson, and E. Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). url: http://www.linguist.is/icelandic$_treebank$.