Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 59-66 recibido 31-03-2019 revisado 07-05-2019 aceptado 15-05-2019

Spelling Normalisation of Basque Historical Texts

Normalización de Textos Históricos Vascos

Ainara Estarrona1[1], Izaskun Etxeberria[2], Ander Soraluze[1], Manuel Padilla[3]
[1]Centre Nationale de la Recherche (CNRS)-IKER (UMR 5478)
[2]University of the Basque Country (UPV/EHU)-Ixa group
[3]University of Pau and Pays de l'Adour (UPPA)-IKER (UMR 5478)
{ainara.estarrona, izaskun.etxeberria, ander.soraluze, manuel.padilla}@ehu.eus

Abstract: This paper presents a computational method and its evaluation in a real scenario with the aim of normalising Basque historical texts in order to be analysed using standard Natural Language Processing tools (NLP). This normalisation work is part of a more general ongoing project 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 of the Basque language.
Keywords: Text normalisation, Digital Humanities, Finite-state technology, Historical Corpus, Diachronic Syntax

Resumen: En este artículo se presenta y evalua en un entorno real un método computacional con el objetivo de normalizar textos históricos vascos para que, una vez normalizados, puedan ser analizados con herramientas estándar de Procesamiento del Lenguaje Natural (PLN). Este trabajo de normalización forma parte de un proyecto en curso más general llamado *Basque in the Making (BIM): A Historical Look at a European Language Isolate*, cuyo objetivo principal es el estudio sistemático y diacrónico de ciertas características gramaticales de la lengua vasca.
Palabras clave: Normalización de textos, Humanidades Digitales, Tecnología de estado finito, Corpus Histórico, Sintaxis diacrónica

1 Introduction

In this paper we present a computational method and its evaluation in a real scenario with the aim of normalising Basque historical texts so that, once normalised, they can be analysed with standard NLP tools. 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].

The main objective of this project consists of a systematic diachronic study of a number of grammatical features of Basque as we know it today in order to know the evolution of certain grammatical structures and to compare them with the theoretical discoveries of the current diachronic synthesis and theoretical linguistic. Hence, previous research work in an academic environment is applied and tested on a real digital library where more dialects, periods and phenomena have been analysed and evaluated.

The second stated goal is to create a comprehensive annotated historical corpus of Basque that will comprise 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 aforementioned metadata. In a previous project carried out between the IKER centre[2] based in Bayonne (France) and the Ixa group[3] of the University of the Basque Country, a search interface was created to analyse the syntactic variability of the Basque language in different dialects (Uria and Etxepare, 2011). This interface is called *Basyque (Basque Syntactic Database)*[4] and we plan to use this experience now to create a search interface to analyse syntactic variability in a diachronic way. Thus, the preparation of resources for the syntactic exploration

[1]http://ixa2.si.ehu.es/bim/en

[2]http://www.iker.cnrs.fr/?lang=fr

[3]http://ixa.si.ehu.es/

[4]http://ixa2.si.ehu.es/atlas2/index.php?lang=en

ISSN 1135-5948. DOI 10.26342/2019-63-6

of the historical corpus in Basque is essential to perform systematic diachronic examinations on the main characteristics of Basque grammar. Taking all this account, *BIM* is an interdisciplinary project, where experts on Linguistics and Natural Language Processing take part.

As said before, in this article we will focus on the implementation and evaluation of computational techniques for corpus normalisation. We base our work on previous research carried out on the automatic normalisation of both historical and dialectal Basque texts (Etxeberria et al., 2014; Etxeberria et al., 2016).

Consequently, the methodology to be followed is already partially defined. It must be stressed that the method has been previously tested and evaluated always, however, in a experimental scenario. An important contribution of this paper is testing the method in a real application where periods and dialects that have never been analysed before are also considered, adding complexity to the normalisation task.

After having presented the project briefly, in Section 2 of the paper we will make an exposition of the related work. In Section 3 we will present the corpus with which we are going to work. Section 4 will be dedicated to the normalisation process including manual annotation and automatic normalisation. The main experimental results will be outlined and discussed in Section 5. Finally, we will review the main conclusions and preview future work in Section 6.

2 Related work

Several techniques have been used for the normalisation of historical texts and they can be roughly divided into three groups:

- Rule-based methods (hand-written phonological grammars) were the first approach, but they require a big amount of manual work.
- Unsupervised techniques: systems that work without supervision. Applying edit-distance (Levenshtein distance) or phonetic distance (by i.e., the *Soundex* algorithm) are popular solutions. Such approaches are often used as a baseline for testing new systems (Jurish, 2010).
- Machine-learning based techniques: systems that learn from examples of standard-variant pairs. These techniques are the most popular nowadays and our approach is within them.

Kestemont, Daelemans, and Pauw (2010) carry out lemmatisation on a Middle Dutch literary corpus, presenting a language-independent system that can 'learn' intra-lemma spelling variation. This work employs a novel string distance metric to better detect spelling variants. The semi-supervised system attempts to re-rank candidates suggested by the classic Levenshtein distance, leading to substantial gains in lemmatisation accuracy.

Pettersson, Megyesi, and Tiedemann (2013) treat the normalisation task as a translation problem, using character-based SMT techniques (CSMT) in the spelling normalisation process. Then, in Pettersson, Megyesi, and Nivre (2014) they evaluate and compare their approach with a memory-based filtering and a Levenshtein-based approaches considering five languages. The SMT-based approach generally works best.

Using the same approach (CSMT) 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 work (Etxeberria, Alegria, and Uria, 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. During evaluation our approach is compared with the CSMT methods explained in Pettersson, Megyesi, and Nivre (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, and it can be considered within the state-of-the-art research.

Lately, some authors have begun applying NMT models in order to analyse whether they perform better than SMT models for the historical spelling normalisation task. Thus,

in (Bollmann and Søgaard, 2016) they explore the suitability of a deep bi-LSTM neural network architecture applied on a character level, and show that it outperforms the baselines (conditional random fields and Norma tool by Bollmann et al. (2012)).

In a similar way, eight different NMT models are applied in (Tang et al., 2018) to the spelling normalisation task in five languages, English, German, Hungarian, Icelandic, and Swedish. Using word accuracy as the evaluation metric, NMT models perform better than SMT models for four languages, but when CER is used, all the NMT models perform better than SMT models for all the languages. 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 this size.

In any case, for the moment we have ruled out neural methods due to the features mentioned of the BIM corpus: several historical periods, several dialects and short texts. Further research is necessary for a robust solution based on neural methods.

3 The corpus

The corpus covers the most representative written production between the 15th and the 18th century, a time span within which all historical dialects of Basque are represented. Texts have been selected on the basis of: 1) their representativeness; 2) the existence of reliable editions; and 3) their social context. To compile such a corpus, we have made use of some ancient texts that are available in the *Klasikoen Gordailua* website[5], which covers almost all the classic Basque texts over the last centuries. It is a corpus comprising more than 12 million words and therefore, it is therefore a solid base for the analysis of the development of the Basque dialects. The texts belong to different literary genres: narrative, poetry and verses. But we will focus on prose texts, considering that the syntactic structures we are interested in may be different from one genre to another. As the texts are classified according to time, language and literary genre, it will possible to take into account such parameters when searching for data. Besides, the corpus is transcribed and digitised in .pdf and .rtf formats, which makes it easy to access them for work.

However, as we know that some errors have occurred in the transcription process of such corpora, we consider the revision of the corpus is essential, and we have therefore carried out an important work of philological revision of the texts comparing them with their facsimiles. In this way, we have updated the spelling of the texts, while maintaining the phonological characteristics of each of them.

In terms of the corpus size, our goal is to analyse as many texts as possible, as a robust annotated corpus is necessary to achieve significant results. For the moment, we foresee starting with a 750.000 word reference corpus, expanding from the 15th century to the mid 18th century, the period encompassing Archaic and Old Basque. We have already finished the selection of those critical editions we will work on. When working with texts from different time periods and places, we are compiling a very rich corpus in respect to its quality. However, the diversity of dialects and varieties adds complexity to the normalisation task.

In this first step, relevant old works from different dialects and literary genres have been selected in order to be able to identify from the outset the challenges we will have to face during the process of labeling the entire corpus.

1. Joanes Leizarraga, *Jesus Krist Gure Jaunaren Testamentu Berria*, 1571.
2. Bernard Etxepare, *Linguae Vasconum Primitiae*, 1545.
3. *Refranes y Sentencias*, 1596 (anonymous).

The first work, Leizarraga, is a translation of the New Testament written basically in Lapurdian dialect. It consists of around 74,000 words and is undoubtedly the most extensive work documented in Basque until the 17th century.

The second work, Etxepare, is considered to be the first documented book in the Basque language. These are poems written in the Lower Navarrese dialect and consist of around 7,000 words.

Finally, the third work chosen, RS, is a compendium of sayings from the end of the 16th century written in an archaic Biscayan dialect and contains around 3,000 words.

[5] http://klasikoak.armiarma.eus/

4 *Text normalisation*

Text normalisation is a necessary step in this project. Having developed a lot of NLP based technology for the analysis of standard Basque, we have now to tackle the text normalisation, considering that once ancient or dialectal texts are standardised, the developed NLP tools could be applied for the linguistic analysis of the corpora.

The process of text normalisation will be carried out in 2 phases. First, part of the corpus will be annotated manually, and then, the rest of the corpus will be normalised automatically.

4.1 Manual annotation

In previous work on the standardisation of historical Basque texts (Etxeberria, 2016), it has been demonstrated that manual annotation of 10% of the text is sufficient to obtain satisfactory results in automatic normalisation. Therefore, we have randomly collected and processed 10% of each of the 3 texts selected to proceed to the manual normalisation.

The processing of the text consists of 3 steps: tokenization, named entity recognition and lexical recognition. After these 3 steps, each word in the text will be marked with the corresponding labels (see Figure 1):

- ENT-Zuz: 'Correct entity', when the named entity recognition (Alegria et al., 2006) identifies a proper name (green colour).
- STD-Zuz: 'Correct standard', when the morphological analyser of standard Basque (Alegria et al., 1996) identifies the word or lemma (fuchsia colour).
- OOV: 'Out of vocabulary', when the morphological analyser of standard Basque does not identify the word or lemma (blue color).

The manual annotation is done using the *BRAT* tool[6] and the task is carried out by a single linguistic annotator with training in historical texts. Figure (1) shows the interface for the manual annotation.

The task of the annotator is to assign to each word of the text its corresponding standard form according to the dictionary of the

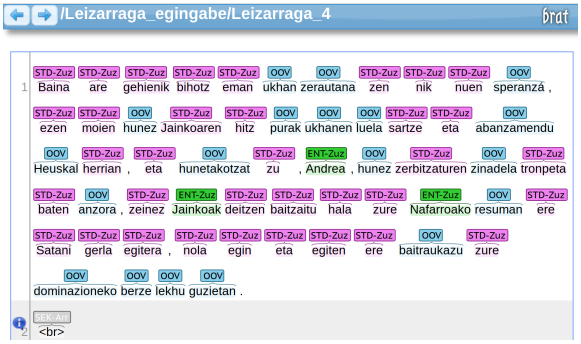

Figure 1: Interface for the manual annotation of the corpus in which a fragment of Leizarraga's work can be seen

Royal Academy of the Basque Language[7]. Thus, the goal of manual standardisation is to remove all OOV tags (in blue) based on a classification which is used to specify the type of variation. The classification of variants is represented using the following labels:

- OOV-Ald: 'Variant', when the annotator assigns a standard variant to the non-standard form (yellow colour).
- OOV-Zuz: 'Correct', when the morphological analyser does not recognise the word because it is archaic or exclusive of historical texts, but the annotator wants to point out that it is correct (fuchsia colour).
- STD-Ald: 'Variant', when an old word coincidentally looks like a modern Basque word ('false friends', yellow colour)[8].

In addition to this classification of variants, we also have the SEK-Ber label which means 'special section' (grey colour). This label is used to mark a paragraph where the annotator has found some interesting historical morphosyntactic phenomenon for subsequent morphonsyntactic analysis.

It should be stressed that in previous research work (Etxeberria et al., 2016) only Out-of-Vocabulary (OOV) words were revised and normalised. However, taking into

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

[7] https://www.euskaltzaindia.eus/hizkuntza-baliabideak

[8] We have also added a label to differentiate variants related to verbs or verbal auxiliaries (OOV-AldADIJ) because we believe it may be interesting to have these types of variants identified for future steps during the normalisation and morphosyntactic analysis process.

account that in this project we work for a real scenario, the annotator will have to check all the words one by one, also those that have the STD-Zuz label, since sometimes old words coincidentally look like modern Basque words ('false friends'), which adds an obvious complexity to the task.

One way to assess the complexity of the task in each work may be to measure the number of standard words and variants in each of them. Table 1 presents the total number of words of each of the works treated in this paper and described in Section 3, as well as the number of words that have the labels STD-Zuz, OOV and ENT-zuz in each of them (with their percentages). Finally, it shows the number of words manually annotated in each work.

	Leizarra	Etxepare	RS
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%
Manually	7,548	694	3,083

Table 1: Total number of words for each work, total number of words with STD-Zuz, OOV and ENT-Zuz labels and their percentages, and total number of words manually annotated

Thus, we can foresee that normalising *RS* (55% of standard forms) will be a much more difficult task than normalising Leizarraga's work (68%), as we will see in Section 5.

Finally, the time needed to manually annotate the whole corpus has been estimated and it can be seen that about 143 words per hour are labeled. Therefore, a person would need about 3 years to label the entire corpus manually and this is obviously why we plan to implement computational techniques to normalise the corpus. Once 10% of the text is manually normalised, the rest will be automatically normalised using statistical learning methods, which we will explain in detail in the following Section.

4.2 Automatic normalisation

All the word pairs in the learning set are taken into account for the learning process: some of them are labeled as *Zuz* (correct) and other as *Ald* (variant), but all of them are going to be used in 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 (correct or variant) but the second one contains only the word pairs labeled as variants.

We have reapplied the method previously cited in (Etxeberria et al., 2014). This method uses *Phonetisaurus*[9], a Weighted Finite State Transducer (WFST) driven phonology tool (Novak, Minematsu, and Hirose, 2012).

After collecting the word pairs (historical–standard) in the learning set into a dictionary, the application of the tool includes three major steps:

1. Sequence alignment. In order to align historical words and their equivalent standard words (not words and pronunciations), the results of the alignment process are joint grapheme/grapheme chunks that we use in the next step.
2. Model training. A joint n-gram language model is trained using the aligned data and then converted into a WFST (after some tuning, a joint 7-gram model was generated).
3. Decoding. The default decoder used in the WFST-based approach finds the best hypothesis for the input words, given the WFST obtained in the previous step. It is also possible to extract the k-best output hypotheses for each word.

Phonetisaurus tool has been used to learn the changes that occur within the word pairs 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. When multiple possibilities for a corresponding historical variant exist, some filtering becomes necessary and the quality of this filtering becomes very important to improve the results. Based on the tuning process carried out in our previous work (Etxeberria et al., 2016), the number of candidates is set to 5.

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

Regarding the filtering process of the generated candidates, the first filter is obvious: the transductions that do not correspond to any accepted standard word form are eliminated. For selecting standard words, a morphological analyzer of Basque is used (Alegria et al., 2009). From the remaining candidates, if there are any, the most probable transduction according to *Phonetisaurus*'s weight model is selected. If there is no standard candidate for a given input, we have tried different options: 1) giving the input word (historical) as output (standard); 2) the most probable transduction according to *Phonetisaurus*'s weight model is selected.

Another important option could be taking into account the memorised word pairs. There are exceptions, but a word is usually always standardised in the same form, therefore it could be a good idea to take advantage of the learned word pairs.

4.3 Evaluation

Once we have part of the text manually annotated and the method applied, we evaluate the quality of the automatic normalisation method we propose. We have carried out 10-fold cross-validation experiments on the manually annotated text.

Results are given based upon precision, recall and F-score IR measures.

5 Results and discussion

We have started with Leizarraga's work and we have used 10% of the text manually annotated for evaluation. Different experiments have been carried out based on the filtering process explained in Table 2 (Section 4.2). In the first one, we have evaluated the results when *Phonetisaurus* does not give an answer if it does not find a standard option among the first 5 options it offers (silence). In this case, the precision was very high (0.9630). Using information about previously seen examples (memory) recall rises 3 points and the F-score almost 2.

In any case, the goal for the project is that the system always offers an answer, therefore we have carried out two other experiments. When the system does not find a standard form among the first 5 options, it gives as output the same input (out=input) or it gives as an output the first of them (first WFST).

The last experiment consists of combining memory and the first proposal of the WFST (2nd. and 4th. options) obtaining the best recall and F-score (memory + first). This option has been chosen for the all the texts of the project and results obtained for the three texts can be seen in Table 3. P, R and F have the same value when always one option is proved, so, only one figure is shown for each one. We have added in the table the results obtained if we only consider the variants with the intention of really seeing how the system behaves in the most difficult cases and pointing out the difficulty of the task[10].

Filter	P	R	F
Silence	0.9630	0.8784	0.9187
Memory	**0.9695**	0.9070	0.9372
Out=Input	0.9245	0.9245	0.9245
First WFST	0.9368	0.9368	0.9368
Memory+First	0.9424	**0.9424**	**0.9424**

Table 2: Results for Leizarraga's work. When *Phonetisaurus* does not find a standard option, four approaches are tested: silence (no output), memory (output if it has been seen in the training set), echo of the input and first proposal of the normalizator.

After fixing the filter, the next work was that of Etxepare. Taking into account that the dialects of Leizarraga and Etxepare are close, we tried to use the same system trained with Leizarraga to normalise Etxepare and evaluated with the manually annotated sample. The results were not as expected with an F-Score of 0.7516 taking into account all the words, and only an F-score of 0.4617 for the variants. Using only the sample manually annotated the results improved (F-Score 0.5958 for the variants) but were still clearly insufficient, therefore we proposed 2 possible solutions: i) to annotate manually more text, and ii) to use for learning both Leizarraga and Etxepare. We decided to combine both sources and the results improved considerably as can be seen in Table 3 (2nd. column).

This suggests that in the future, instead of having a system for each work, we could 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 dialect or group of dialects.

[10]The results decrease when evaluating only the variants, but it is necessary to bear in mind that the real scenario is the one in which the system normalises all the text, variants and standard forms included.

The last work was *RS*. This is a special work because it reflects a variety of language very far from the standard and very archaic. The first results using the manually annotated sample of 10% were clearly insufficient with an F-score less than 0.50. The work has about 3,000 words, therefore we thought it was convenient to make an effort and manually annotate it in its entirely. The results obtained using cross-validation can be seen in Table 3 (3rd. column).

	Leizar.	Etxep.	RS
Test-variants	0.8768	0.7038	0.6683
Test-all	0.9424	0.8646	0.8066

Table 3: Accuracy for *Leizarraga*, *Etxepare* and *RS* taking into account all words or only variants

The results clearly show the difficulty of the task in the case of works so distant from the standard variety of the language. As we have shown in Table 1 (Section 4.1), Leizarraga's work is the one that has more standard words (68%) and in which we obtain better results, followed by Etxepare's work (64%) and, finally, the worst results are for *RS*, which is the work farthest from the standard variety, with only 55% of standard words.

Results for *RS* are worse than expected in spite of annotating all the text. This is an important finding for the project and has led us to decide that sometimes the whole text should be manually annotated (luckily only for short texts).

6 Conclusions and future work

We have presented and evaluated in a real scenario a normalisation method for Basque historical texts for several historical periods and several dialects within the framework of a more general project of creating a syntactically annotated historical corpus of the Basque language.

Our phonological induction inspired method has been evaluated on three works of the 16th century selected from the corpus. The method achieves F-scores between 80% and 95% in the task. We believe our method is a good solution using a limited amount of supervision to achieve acceptable results without significant manual annotation effort.

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. Nevertheless, Moeller et al. (2019) 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.

It is important to underline that the experiments have been carried out on a real historical digital library, overcoming the limits of some academic repositories. Combining historical periods, dialects and scarcity of data, we have concluded that the results can be, in some cases, worse than expected (*RS* corpus).

In addition to the quantitative evaluation, we intend to carry out an error analysis in order to detect the weak points of our system and be able to search for strategies aimed at improving the results. A preliminary error analysis in the Leizarraga's work have been carried out concluding that an important part of the errors correspond to the system of auxiliary verbs. Something that we already foresaw, since the system of auxiliary verbs in historical Basque and its different dialects is especially complicated.

After normalising the corpus we will proceed to the automatic morphosyntactic analysis. This automatic analysis will be revised manually to detect errors and to proceed to the labeling 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.

Acknowledgments

We are indebted to Josef Novak for his useful help concerning Phonetisaurus, and to Jordi Porta, Eva Petterson, Yves Scherrer and Tomaž Erjavec for providing their corpora and for addressing our inquiries about their experiments. Grateful thanks to the three anonymous referees for helping us improving the paper. The research leading

to these results was carried out as part of the *BIM* project (Agence Nationale de la Recherche, France) and the *BERBAOLA* project (Basque Government funding, Elkartek KK-2017/00043).

References

Alegria, I., O. Arregi, N. Ezeiza, and I. Fernández. 2006. Lessons from the development of a named entity recognizer for basque. *Procesamiento del lenguaje natural*, 36.

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

Alegria, I., I. Etxeberria, M. Hulden, and M. Maritxalar. 2009. Porting basque morphological grammars to foma, an open-source tool. In *Finite-State Methods and Natural Language Processing*. Springer, pages 105–113.

Bollmann, M., S. Dipper, J. Krasselt, and F. Petran. 2012. Manual and semi-automatic normalization of historical spelling-case studies from early new high german. In *KONVENS*, pages 342–350.

Bollmann, M. and A. Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. *arXiv preprint arXiv:1610.07844*.

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

Etxeberria, I., 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.

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

Etxeberria, I., 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.

Jurish, B. 2010. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77.

Kestemont, M., W. Daelemans, and G. D. Pauw. 2010. Weigh your words—memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301.

Moeller, S., 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.

Novak, J. R., 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.

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

Pettersson, E., 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.

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

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

Uria, L. and R. Etxepare. 2011. Basyque: Aplicación para el estudio de la variación sintáctica. *Linguamática*, 3(1):35–44.