Iris Hendrickx, Rita Marquilhas

From old texts to modern spellings: an experiment in automatic normalisation

We aim to tackle the problem of spelling variations in a corpus of personal Portugese letters from the 16th to the 20th century. We investigated the extent to which the task of normalising Portuguese spelling can be accomplished automatically. We adapted VARD2 (Baron and Rayson, 2008), a statistical tool for normalising spelling, for use with the Portuguese language and studied its performance over four different time periods. Our results showed that VARD2 performed best on the older letters and worst on the most modern ones. In an extrinsic evaluation, we measured the usefulness of automatic normalisation for the linguistic task of automatic POS-tagging and showed that automatic normalisation of spelling helps improve the performance of the POS-tagger.

1 Introduction

Our goal was to reduce the problem of spelling variations in the Portuguese CARDS-FLY corpus of personal letters written over a period of approximately 500 years, from the 16th to the 20th century. As the letters were written by a diverse group of authors, some of whom were semi-illiterate, and most of the manuscripts predate the first standardisation of Portuguese spelling which only took place in 1911, they contain many spelling variations. We wanted to make the corpus available for further linguistic research and also make it accessible to a wider community.

As the corpus is being prepared for research into language change, given the exceptional, near-spoken status of the texts, it was advisable to preserve the original spellings in a paleographic edition appropriate for research into sound variations and change, and for discourse analysis focussing on the specific behaviour of the social agents. However, a standardised corpus is essential for lexical and grammatical research, since this is the only format that can be given the necessary mark-up, such as POS tagging, semantic tagging, and syntactic parsing, enabling it to be used as an empirical tool for testing theories of lexical and syntactic change.

On the other hand, the rarity of the documents in question also make them valuable to the lay public, since they represent part of the country's cultural heritage, reflecting the everyday lives of ordinary people, especially those who suffered hardship. The 16th to 19th century corpus comprises original epistolary texts produced by servants, children, wives, lovers, thieves, soldiers, artisans, priests, political campaigners and many other social agents who fell foul of the Inquisition and the civil courts, two institutions that habitually seized personal correspondence for use as evidence. The letters in the

20th century corpus come from personal collections compiled by the families of former soldiers, emigrants, and political prisoners. Editions intended for the lay public have the understandable obligation to provide readers with a clean text, free of distracting variations in spelling.

Our aim was therefore to add an additional orthographic component to the historical documents, which involved modernising the spelling. Modernising or normalising spelling is not an arbitrary step: if done manually it involves an enormous workload that can only be carried out by a specialist in historical linguistics and if done automatically, errors inevitably occur.

Here we investigate the extent to which the task of modernising Portuguese spelling can be accomplished automatically. We adapted VARD2 (Baron and Rayson, 2008) a well-studied statistical tool for normalising spelling, for use with the Portuguese language and studied its performance over four different time periods. The performance of this automatic normalisation tool was evaluated using intrinsic and extrinsic methods. Firstly, the automatically normalised text was compared with manually normalised text. Secondly, as an *evaluation of use*, the effect and usefulness of automatic spelling normalisation was evaluated in terms of the task of automatic grammatical tagging.

The performance of a POS-tagger used on a data set with the original, non-standardised spelling was compared with the effect of automatically normalised text and, as the upper bound, on manually normalised text. This research is similar to the work of Rayson et al. (2007) who investigated the usefulness of pre-processing with VARD2 for POS-tagging historical English and showed that normalisation does help to improve the performance of the POS-tagger.

The paper is structured as follows. In the next section we first discuss previous approaches to historical spelling normalisation. In Section 3 the experimental setup is explained, in particular the adaptation of the VARD2 tool to Portuguese in Section 3.1. Section 3.2 describes the historical corpus and provides additional background information on historical spelling changes in Portuguese. In Sections 3.3 and 5 the results of the experiments in normalisation and POS-tagging are presented and the conclusion appears in 6.

2 Related work

The problem of spelling variations in historical texts has been investigated from different perspectives and with different aims. Automatic text retrieval on historical data suffers severely from spelling variation and a common approach to this problem is not to modernise the full text collection, but to expand the search query to cover lexical variants (e.g. (Koolen et al., 2006; Hauser and Schulz, 2007; Ernst-Gerlach and Fuhr, 2007; Gotscharek et al., 2011)).

Other approaches attempt to modernise the spelling in the historical documents themselves. The VARD tools were developed for corpus linguistic research into Early Modern English. The original VARD tool consisted of a list of manually created mappings between historical variants and their modern versions. VARD2 (Baron and

Rayson, 2008) has an additional module that can search for variants and mappings. In the work of Rayson et al. (2005) the ability of VARD to detect spelling variants and suggest the correct modern spelling is compared with two commercial spelling correctors, MS-Word and Aspell, showing that VARD works better for historical texts since it detects fewer false positives. VARD2 will be discussed in more detail in Section 3.1.

Kestemont et al. (2010) describe an automatic approach to normalise the spelling in Middle Dutch Text from the 12th century. In this case however, they chose not to convert historical word forms to their modern counterparts, but to their modern lemma. They used machine learning to discover how to transform one spelling variant into another to resolve intra-lemma variation.

Several studies of spelling variation in Portuguese historical documents have been conducted and we were grateful to be able to re-use some of the resources already developed for historical Portuguese. We will briefly discuss this previous work and, in Section 3.1, explain how we used the available resources.

The Historical Dictionary of Brazilian Portuguese (HDBP) is constructed on the basis of a historical Portuguese corpus of 1,733 texts and approximately 5 million tokens. As there was no standard spelling at the time (16th to 19th century), it is not easy to create lexicographic entries on the basis of the corpus or produce reliable frequency counts. A rule-based method was therefore developed for the automatic detection of spelling variation in the corpus (Giusti et al., 2007). The HDBP researchers compared their automatic variant detection method with Agrep[1] using a small test set and showed that their method was more precise, whereas Agrep had much better recall. These experiments also led to a spelling variants dictionary containing approximately 30K clusters of variants.

Another resource available for Portuguese is the Digital Corpus of Medieval Portuguese (CIPM- Corpus Informatizado do Português Medieval)[2] which covers the 12th to the 16th century and counts around 2 million words. Rocio et al. (2003) describe how they annotated part of the CIPM with linguistic information such as POS-tags, morphological analysis and partial parse information. They did not proceed with modernisation but used automatic tools on the historical data as such, followed by a manual correction phase.

The Tycho Brahe Parsed Corpus of Historical Portuguese[3] is an electronic corpus of historical texts with prose from different text genres from the Middle Ages to the Late Modern era. The TBCHP contains 52 source texts but not all of them are annotated in the same way. Some of the texts maintain the original spelling variations, while in other texts, intended for part-of-speech and syntactic annotation, the spelling was standardised.

[1] Agrep: http://www.tgries.de/agrep/

[2] CIPM corpus: http://cipm.fcsh.unl.pt/

[3] TBCHP: http://www.tycho.iel.unicamp.br/∼tycho/corpus/en

3 Data and Methods

This section describes the adaptation of the VARD2 spelling standardisation tool for use with Portuguese, the corpus in question, which is a historical corpus of private letters written in Portuguese, and the experimental setup for normalisation and POS-tagging.

3.1 VARD2 for Portuguese

The aim of this study was to evaluate the performance and usefulness of the VARD2 tool for historical Portuguese. As mentioned in Section 2, VARD2 was developed for historical forms of English. It combines several resources to detect and replace spelling variants with a normalised form and, where possible, we adapted every module for use with the Portuguese language.

VARD2 uses a modern lexicon, a spelling variants dictionary list that matches variants against their modern counterparts, a list of letter replacement rules and a phonetic matching algorithm to predict normalised candidates for each variant detected, using an edit distance algorithm to determine the most likely candidate. The tool can be configured, since each module can be assigned a certain weight and can be individually configured in favour of recall or precision. When training the tools on a specific data set, new words and variants are added to the lists and the module weights are adapted accordingly. We replaced the modern frequency lexicon, spelling variants dictionary and letter replacement rules with Portuguese versions.

The following Portuguese resources were used to convert VARD2 to Portuguese. The Multifunctional Computational Lexicon of Contemporary Portuguese[4] contains 26,443 lemmas corresponding to 140,315 tokens. Only lemmas with a minimum lemma frequency of 6 were extracted from a sample from the contemporary Portuguese corpus CRPC (a corpus sample of 16,210,438 words) for inclusion in the lexicon. We filtered this lexicon to suit our purpose and removed all multi word expressions (for example "sem abrigo") and words with non-conventional spellings. The frequency counts for homonyms were reduced to one count for each particular word form. The word frequency list that we used has a total of 127,891 word forms.

We created a list of letter replacement rules based on the rule set described in detail by Giusti et al. (2007) and developed for historical Brazilian Portuguese. The rules encode accent changes, spelling changes, such as 'x' to 'ch', and letter combinations that are no longer used in modern Portuguese, for example 'th', 'ph', 'aes', or double consonants such as 'dd', 'ff' etc.

As mentioned in Section 2 a spelling variants dictionary[5] was created based on the Historical Corpus of Brazilian Portuguese. Giusti et al. (2007) have created a corpus-based tool to automatically generate and test rewrite rules that cluster spelling variants together. These groups are clustered around one common word form, the so-called head

[4] Lexicon is available for download at: http://www.clul.ul. pt/en/resources/88 -project-multifunctional- computational -lexicon- of-contemporary- portuguese-r

[5] BP spelling variants dictionary is available: http://www.nilc.icmc.usp.br/nilc/projects/hpc/

word of the cluster. As the original dictionary consists of clusters of spelling variants, and we needed a list of one-to-one mappings between variants and their modernized counterparts to integrate into the VARD2 tool, this variants dictionary had to be converted to meet our needs. As a logical choice, we initially mapped each variant in a cluster to the head word of the cluster. However, the head word is not always the modern or most frequent word form, although this is usually the case, and this implies that these automatic mappings sometimes lead to errors. Here is an example of a cluster from the spelling variants dictionary:

tambem	(12211)	
	tambem	(9002)
	também	(3160)
	tanbem	(47)
	ttambem	(1)
	ttanbem	(1)

The modern word form of this cluster is the accented version *também* (En: “also") and the cluster head *tambem* does not occur in the current modern lexicon. To prevent mappings between variants and non-modern word forms, every head word was checked to determine whether it occurred in our modern lexicon. If it did not, the most frequent word form in the cluster that did occur in the modern lexicon was selected. Closer inspection of the resulting variants list showed that this automatic mapping of variants and head words can still lead to some errors in cases where the head word occurs in the modern lexicon but is not the most obvious candidate, for example the “aviam -haviam" cluster. A manual correction phase of (at least the most frequent) variant clusters would certainly improve the variant list. We did not alter the phonetic matching algorithm of VARD2, but in future work we would like to evaluate this module for Portuguese.

3.2 The corpus

As mentioned in the Introduction, the CARDS-FLY corpus[6] was compiled from a rare collection of documents written by a variety of social agents living in difficult times. Later, disregarding the authors' intentions, the letters found their way into several archives instead of being destroyed, as might be expected in the case of everyday private papers. The manuscripts from the 1500-1800 period are personal letters that were, unusually, retained as part of religious legal proceedings, as evidence used by the Inquisition in heresy trials. Those from the 19th century were also used as legal evidence, this time in criminal cases heard by the Portuguese Royal Appeal Court (abolished in 1833) and civil cases that appeared before a regional court in the north-east. The 20th century letters date from 1901-1974 and consist mainly of manuscripts, together with some typed scripts, sent or received by soldiers who fought in World War I or in the Portuguese Colonial War, emigrants of Portuguese origin, and prisoners held by the political police. They were mostly kept in family archives and sometimes donated

[6] CARDS-FLY corpus: http://alfclul.clul.ul.pt/cards-fly/

to public documentation centres. A few others were archived by propaganda and censorship institutions. The whole collection is being processed electronically (involving transcription into XML-TEI file format) so that it can function both as a digital archive available to the general public and as a corpus intended for historical, linguistic and sociological research.

For the sake of historical accuracy, the letters were divided into different time periods, taking into account the serialisations already proposed for the history of Portuguese language. Not all Portuguese historical linguists working on serialisations agree on the chronology for milestones in language change as Martins2002 explains. However, they do agree on the convenience of distinguishing between Old Portuguese, Classical Portuguese and Modern Portuguese, following the traditional classification in general history that distinguishes between the Middle Ages (from the end of Antiquity up until the Renaissance), the Early Modern Age (up until the liberal Revolutions), and the Contemporary Age.

Our corpus contains sources for the study of both Classical and Modern Portuguese and a dividing line therefore had to be drawn between the two in the early 19th century. However, a second milestone was needed since a great deal of debate surrounds the Classical Portuguese period with regard to innovations in European Portuguese, especially syntax, vis-à-vis Brazilian Portuguese. In the current state of the art, the beginning of the 18th century represents such an important milestone (Galves and de Sousa, 2005) and we therefore subdivided the Classical letters into those dating from 1500-1700 and those dating from 1701-1800. As for the Modern letters in our corpus from the period 1801-1974, on the one hand we had to take into account the fact that we were dealing with written texts that generally adopted non-standard spellings, but also the fact that the Republican decree of 1911 had instituted the first national spelling agreement (Castro et al., 1987). Prior to this, despite several debates, there was no standard way of spelling Portuguese and the discussion was very much divided between the 'Sonics' and the 'Etymologists'. The Sonics fought for phonographic spelling using diacritics (matéria versus materia) and the absence of learned consonantal clusters of Greek or Latin origin (catedral versus cathedral). The Etymologists advocated the reverse, which had a better established tradition in Portuguese writing.

In terms of the division of our corpus into time spans, we considered that the effects of the 1911 spelling reform would only be evident by 1931, when the children who had started grammar school in 1911 had already become adults using correspondence as a common interactive practice. The Modern letters in our corpus were therefore divided into two groups, namely 1801-1930 and 1931-1974.

The current version of the corpus contains 1,802 letters from which we randomly selected a subset of 200 letters for the experiments, respecting the frequency division of the different centuries. This subset was manually annotated by a linguist to be used as training and evaluation material. The texts were tokenised (punctuation was separated from words) and we converted any names in the text into the string 'NAME'. The names in the modern letters were already anonymous and, following this conversion, all the documents had the same form of representation for names. For the purposes of our experiments, this data set was split into 100 letters for training the VARD tool, and 100 for the evaluation set.

Table 1 presents the statistics for the evaluation set, showing the number of letters, tokens and the average number of spelling normalisations made by the human annotator. As might be expected, more corrections per letter were found in the oldest letters, which were also the shortest. In the modern letters only 4% of the tokens were normalised. The 18th century letters are remarkably long in comparison with the other letters. One possible explanation for this is that, on the one hand, the corpus contains more letters from the 18th than the 16th and the 17th centuries and so there is a greater likelihood of obtaining long letters. In addition, the lower-classes were gradually becoming literate (or semi-illiterate) during the 19th and 2oth century and would therefore have tended to restrict themselves to short letters dealing with urgent matters.

Table 1: Statistics for the evaluation set of 100 letters, divided into the four time periods. # Tok/file shows the average number of tokens per letter, '#Norm/file' the average number of manual spelling corrections per letter and '% Norm/tok' is the percentage of all tokens that is normalised.

Period	Files	Tokens	#Tok/file	# Norm/file	% Norm toks
1500-1700	10	2262	226.2	56.9	25.2
1701-1800	28	13913	496.9	120.8	24.3
1801-1930	43	14343,	333.6	60.7	18.1
1931- 1974	19	6817	358.8	16.1	4.2

Even though the letters from the final period 1931-1974 basically use modern spelling, an average of 16 spelling changes per letter can still be observed. The type of spelling changes here are mostly due to hypercorrections associated with the inner logic of Portuguese 'sonic' orthography, which contains many inconsistencies in grapheme -phoneme correlations.

3.3 Experiments in normalisation

After discussing the VARD2 tool used for standardisation and the corpus in detail, we now describe exactly how this tool was applied to the data and explore how well it performed with the Portuguese corpus. Our aim was also to investigate whether it was more practical to have one spelling modernisation tool for the complete Early Modern-Contemporary period, or separate tools for shorter periods and whether the advantage of having specialised tools for each period outweighed the disadvantage of less data, given that each specific tool would be trained using a smaller set.

To evaluate the performance of the tool, we compute accuracy, recall, precision and F-score for the words (excluding punctuation marks) in the test data. True positives (TP) refer to cases in which there was a spelling variant in the text and the modern variant was correctly predicted by the tool. False positives (FP) involve cases in which the tool erroneously predicted a spelling variant, and false negatives (FN) are the spelling variants that were not detected by the tool. True negatives (TN) are the remaining words correctly predicted as ‘not a spelling variant’. We compute accuracy, recall (R), precision(P) and the harmonic mean (F-score) between recall and precision (van Rijsbergen, 1979) as follows:

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{1}$$

$$P = TP/(TP + FN), R = TP/(TP + FP), F - score = 2 * P * R/(P + R) \tag{2}$$

As a first step the VARD2 tool was configured for the data set. VARD2 has two parameters that need to be set: the first establishes the weighting given either to recall or to precision, and the second is the replacement threshold which decides whether a potential variant should be replaced with the equivalent modern candidate. We set the first parameter to assign equal weight to recall and precision. To determine the value for the second parameter, we ran a series of experiments with different thresholds. We divided the training set in 80 letters for training and 20 as a development set. We tested the following settings: 1, 5, 10, 20,.. 90 for this threshold. The best F-score, 65.5%, for the development set was obtained with parameter 1. All the parameters tested between 5 and 40 obtained a score of 64% and a gradual decrease in performance was observed when the parameters were increased to values above 50. The best performance was obtained with the threshold set to 1 and this setting was therefore used in all further experiments.

When examining the errors made by VARD2 in the development set, one specific error stood out: the letter q is an abbreviation and is almost always standardised to que. As q itself is listed as a valid word in the modern lexicon, it was never detected as potential spelling variant. Since this q occurs very frequently and many errors were due to this mismatch, a rule was added to the tool to normalise each q to *que*.

In order to investigate whether it was better to have specialised tools trained separately for each time period, or one tool trained using the full training set for the whole period, we trained five versions of VARD2, one for each individual period and one for the whole period. In the first set we applied the VARD2 trained on all 100 training letters to the full test set of 100 letters, and then to the four individual test sets whose properties are listed in Table1.

4 Results of normalisation

The results from these experiments can be seen in Table 2. The second row shows the evaluation when training on 100 letters and testing on 100 letters. The VARD2 tool has a much higher precision than recall. The tool detected around 2/3 of the spelling variations (61% recall), and, if a variant was found, in 97% of the cases it was correctly changed to its modern counterpart. Row 3-6 of the table focuses on the performance of VARD2 in each specific time period. In the latest time period, 1931-1974, high accuracy and a remarkably low recall and F-score can be observed. In calculating accuracy, true negatives are counted, whilst the other measurements focus solely on the spelling variants, which comprise a much smaller set for this particular period. As shown in Table 1, these letters had the fewest spelling normalisations, amounting to only 16 changes per document on average, whereas the letters from the other periods had an average of at least 50 per document.

Contrary to expectations, the highest precision and F-scores were found in the oldest letters. Better results would have been expected for the period 1801- 1930, since we had the largest amount of training and testing material for this period. One explanation may be found in the nature of the data set, since this was a time when the lower-classes were becoming semi-literate and it therefore contains a set of letters produced by people who would never have put pen to paper in earlier times and who produced many creative misspellings that are extremely difficult to predict.

Table 2: Precision, recall and F-score for VARD2 trained on 100 training letters in the evaluation set of 100 letters, divided over the four time periods.

time period	Acc	R	P	F-score
total	91.92	61.10	97.21	75.03
1500 -1700	91.75	68.72	98.74	81.04
1701 -1800	90.77	65.95	97.72	78.75
1801- 1930	91.06	56.0	96.81	70.96
1930 - 1974	96.46	35.23	87.5	50.24

In the second round of experiments we trained VARD2 on the time-specific subsets of the training set and tested on the same test subsets. The results are shown in Table 3. Again it can be observed that the best F-score performance occurs in the older letters and the worst in the most modern ones. When the results in Table 2 are compared with the results in Table 3, a slight improvement can be seen in the two oldest time periods and a decrease in the more modern periods, both in terms of accuracy and F-score. Recall is mainly affected by the change in training set, showing an increase for the oldest data. For the two more modern data sets precision slightly increases at the cost of the recall.

Table 3: Precision, recall and F-score for specialised VARD2 tools, trained and tested separately for the four time periods.

Period	Acc	R	P	F-score
1500-1700	92.52	71.88	98.55	83.13
1701-1800	91.81	70.24	97.49	81.65
1801-1930	90.61	53.45	97.08	68.94
1931-1974	96.32	31.88	87.96	46.80

In order to investigate the spelling variants that could not be corrected automatically, we analysed the most frequent errors in the final round of experiments for each of the different time periods. The most frequent error was the failure to recognise the traditional spelling of 'um' with 'h-' (43 times for the 1701-1800 period and 52 times for 1801-1930). VARD2 does not recognise this as a spelling variant, since hum is listed in the modern lexicon. For all periods we clearly observe that many spelling variants originate from the fact that writers find it difficult to master the diacritics. Furthermore, the older groups of letters have a large number of archaisms (e.g 'inda' and 'cousa' in the top 10 errors for the period 1500-1700) that are no longer used in current spelling, but are erroneously part of the VARD2 modern lexicon list and are therefore not recognised as spelling variants. The older letters also have a large percentage of abbreviations (e.g. v., va. and etcra.) which are difficult to recognise automatically, unlike the modern ones. Grammar teachers condemn the use of abbreviations, which they label bad style, and this 'lesson' seems to have been learned by the 20$^{\text{th}}$ century authors writing between 1931 and 1974.

Major trends involving confusion between different spellings were observed within the different periods. For the period 1500-1700, difficulty in mastering the etymological use of s/c/ss for the single sound [s] was evident, and , for the period 1701-1800, the etymological use of z/s for the single sound [z], whilst in the period 1801-1930 the phonetic spelling of 'i' for 'e' frequently occurs.

5 POS-tagging

Our other goal was to quantify the effect of text normalisation on the application of NLP tools such as a POS-tagger. We trained one POS-tagger on normalised texts from the Portuguese Tycho Brahe corpus and tested it both on non-normalised text and normalised text. The Tycho Brahe corpus contains 19 normalised and POS-tagged texts with a total of approximately 40K sentences and 891K tokens. The POS-tag set contains 280 different tags that express specific information such as gender and number.

We created an automatic POS-tagger by training MBT (Daelemans et al., 2007) on the 19 texts from Tycho Brahe. MBT is a memory-based machine learning system specifically developed to handle sequence labelling such as POS-tagging. When assigning POS tags to words, the previous labelings can be very informative in terms of the current decision: for example if the previous word is labeled as determiner, the current word is likely to be an adjective or noun. MBT takes its previous decisions into account when labelling words.

We tested the POS-tagger on three versions of our corpus of letters: the original unnormalised text, the text automatically standardised using VARD2 (trained on 100 letters), and on the gold standard of manual annotation. The results of these experiments are shown in Table 4. The first column shows the POS-tagging accuracy for all tokens (including punctuation marks) in the 100 letters from the test set.

The major source of errors made by a POS-tagger is unknown words that the tagger has not encountered previously in the training set. In the case of known words, the tagger assumes that it knows (on the basis of the training set) which labels are applicable for a certain word and chooses one label from this small sub set. For example the word *via* can be a verb or noun and the POS-tagger only needs to choose between these two. However, for an unknown word, the POS-tagger needs to consider all 280 possible tags. Therefore the accuracy rate for unknown words is lower than for known words as demonstrated in the last two columns of Table 4.

Table 4: POS-tagger accuracy for the evaluation set of 100 letters, based on the original non-normalised text, text automatically normalised by VARD2, and the gold standard created by manual annotation.

Type	Total	Unknown	Known
# tokens	37,335	5,869	31,466
Original	76.86	42.34	87.06
VARD2	83.41	47.57	90.1
Gold	86.578	49.11	91.94

6 Conclusions

We have presented an approach to standardising the spelling of historical Portuguese and demonstrated how to adapt the statistical VARD2 normalisation tool for the Portuguese language by re-using several Portuguese resources currently available. Having split the data set into 4 time periods, it was observed that VARD2 performs best on the older letters and worst on the most modern ones. We also investigated whether it was more useful to have specialised normalisation tools for each time period, or whether the tool benefits more from one large training set covering the whole time period 1500 to 1974. The results show that for the Classical period the advantage of a specialised tool outweighs the smaller amount of data. Conversely, for the Modern period a tool trained using a larger, diverse data set works better. In terms of extrinsic evaluation, we measured the usefulness of automatic normalisation in terms of the more complex linguistic task of automatic POS-tagging and showed that automatic normalisation of spelling helps improve the performance of the POS-tagger. In all periods, the letter writers can be seen to struggle with two problems: i) how to master etymological spellings without knowing Latin, Greek, or Old Portuguese, ii) how to master phonographic spellings if they never obey purely phonetic facts, given that phonological (segmental and suprasegmental), morphological and lexical information always influences apparently phonographic principles to some extent. Nevertheless, the effectiveness of the 20th century Portuguese spelling reform can be clearly observed in our corpus, as well as its 'Sonic' profile: many etymological principles have clearly been abandoned. This trend was recently reinforced when all the Portuguese speaking countries in the world adopted a new more phonographic reform, celebrated in a 1990 treaty and implemented in Portuguese public education in 2011.

References

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Castro, I., Duarte, I., and Leiria, I. (1987). *A Demanda da Ortografia Portuguesa*. Edições João Sá da Costa, Lisbon, Portugal.

Daelemans, W., Zavrel, J., Van den Bosch, A., and Van der Sloot, K. (2007). MBT: Memory-Based Tagger, version 3.1, Reference Guide. Technical report, ILK Series 07-08.

Ernst-Gerlach, A. and Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the ACM/IEEE-CS conference on Digital libraries*, pages 333–341.

Galves, C. and de Sousa, M. C. P. (2005). *Romance Languages and Linguistic Theory 2003*, chapter Clitic Placement and the Position of Subjects in the History of European Portuguese, pages 97–113. Current Issues in Linguistic Theory 270. John Benjamins, Philadelphia.

Giusti, R., Candido, A., Muniz, M., Cucatto, L., and Aluísio, S. (2007). Automatic detection of spelling variation in historical corpus: An application to build a Brazilian Portuguese spelling variants dictionary. In *Proceedings of the Corpus Linguistics Conference*.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K., and Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *International Journal on Document Analysis and Recognition*, 14:159–171.

Hauser, A. W. and Schulz, K. U. (2007). Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pages 1–6, Borovets, Bulgaria.

Kestemont, M., Daelemans, W., and De Pauw, G. (2010). Weigh your words - Memory-Based Lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25:287–301.

Koolen, M., Adriaans, F., Kamps, J., and de Rijke, M. (2006). A cross-language approach to historic document retrieval. In *Advances in Information Retrieval: Proceedings of ECIR 2006*, volume 3936 of *LNCS*, pages 407–419. Springer Verlag, Heidelberg.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora. In *Proceedings of the Corpus Linguistics Conference (CL2007)*, University of Birmingham, UK.

Rayson, P., Archer, D., and Smith, N. (2005). VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In *Proceedings from the Corpus Linguistics Conference Series*, volume 1, Birmingham (UK).

Rocio, V., Alves, M. A., Lopes, G. P., Xavier, M. F., and Vicente, G. (2003). *Automated creation of a Medieval Portuguese partial Treebank*, volume 20 of *Language and Speech Series*, pages 211–230. Kluwer, anne abeillé edition.

van Rijsbergen, C. (1979). *Information Retrieval*. Buttersworth, London.