Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2

Martin Reynaert, Iris Hendrickx and Rita Marquilhas

TiCC research group, Tilburg University, Tilburg, The Netherlands
Centro de Linguística da Universidade de Lisboa, Lisboa, Portugal

E-mail: reynaert@uvt.nl, iris@clul.ul.pt, rmarquilhas@fl.u.pt

Abstract

We present a comparison of two statistical tools for spelling normalization of historical Portuguese. The VARD2 tool has been originally developed for Early Modern English but has been successfully ported to the Portuguese language. The second tool TICCL was developed for English and Dutch. The VARD2 tool was explicitly developed for historical data, while TICCL aims to handle spelling and Optical Character Recognition variation in very large corpora of digitized 19th and 20th century text. Here we detail both tools, their methods, strengths and weaknesses and their performance on the task at hand.

1 Introduction

There are several reasons why spelling normalization of historical text is essential. Information extraction or retrieval tasks on a historical corpus cannot be handled by any standard search system. If the user would query for a particular modern key word in the corpus, such a system will not be able to retrieve all relevant matches for the query as the different spellings of a word will not be recognized and retrieved. Creating a corpus version that is normalized for spelling would alleviate this problem.

Furthermore, automatically adding linguistic information such as part-of-speech (POS) tags will be much easier on a normalized version of the historical text. Tools such as POS-taggers have been developed for contemporary text and these tools will make more errors when labeling unknown spelling variants [12]. [8] have shown that automatic normalization of the historical data leads to more accurate POS-labeling. In a previous study we have already shown that automatic normalization of the historical data leads to more accurate POS-labeling [8].

Another motivation for spelling normalization is that these old texts with many different spellings and archaic words are difficult to read for non-specialists. These

historical texts are a part of the country's cultural heritage that should be publicly available. A modernized version facilitates the accessibility of such historical texts for the general public.

Here we present a comparison between two different tools, VARD2 [2] and TICCL [16], for automatic spelling normalization of a corpus of historical Portuguese. VARD2 has already been compared against two other spelling checkers, the MS-Word spelling checker and Aspell on historical English text [13]. It was shown that VARD outperforms the other tools as it has a higher precision. TISC (Text-Induced Spelling Correction), the precursor of TICCL, has also been compared with Aspell, Ispell and the MS-Word spelling checker for contemporary Dutch and English text [15] and reached much higher precision than the other systems due to taking into account the full vocabulary of the texts to be corrected and due to making use of bigram information, thereby performing context-sensitive spelling correction on non-words. Bigram correction is now also implemented in TICCL.

In this investigation we apply both tools to Portuguese, a morphologically rich language substantially different from either English or Dutch, for which these tools were originally developed.

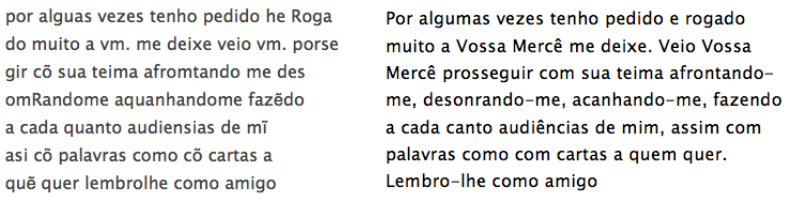
Moverover, the corpus used in these investigations is the Portuguese historical CARDS-FLY corpus that consists of digitally transcribed collections of personal letters and was developed for the historical study of Portuguese language and society. Obviously, handwritten personal letters contain more spelling variation than letters that were produced by professional scribes or clerks or than typeset and printed books that were published by printers.

In the next section we first give a brief account of the purpose and history of the CARDS-FLY corpus. In section 3 we discuss related work on spelling normalization for historical text. We present both tools in section 4 and present our experiments and results on a subset of the full corpus that was manually normalized for spelling in section 5. We discuss our findings in section 6.

2 Corpus

The study of language history can be supported by rather 'popular' sources that exist by the thousands, unpublished and ignored, on the stacks of public archives of the western countries. These sources are namely private letters, in our case, confiscated by courts of law. They differ from literary texts and from institutional documents in that they were not written for posterity or public reading. They were not written to be preserved, but they were nevertheless so. They were meant to circulate in the private sphere but, again, they went public. The judges of different courts - either religious, or civil or military - used the letters as instrumental proof, so their style quality was irrelevant, and they would do even if poorly written. The only thing that mattered was their referential contents.

In Early Modern Portugal, two different courts, namely the Inquisition and the



por alguas vezes tenho pedido he Roga
do muito a vm. me deixe veio vm. porse
gir cõ sua teima afromtando me des
omRandome aquanhandome fazẽdo
a cada quanto audiensias de mĩ
asi cõ palavras como cõ cartas a
quẽ quer lembrolhe como amigo

Por algumas vezes tenho pedido e rogado
muito a Vossa Mercê me deixe. Veio Vossa
Mercê prosseguir com sua teima afrontando-
me, desonrando-me, acanhando-me, fazendo
a cada canto audiências de mim, assim com
palavras como com cartas a quem quer.
Lembro-lhe como amigo

Figure 1: Example of a manually transcribed letter from 1592 addressed to merchandiser Joào Nunes. English translation: *I have more than once asked Your Honour and begged Your Honour to leave me alone. But Your Honour has insisted on defying me, dishonouring me, lessening me, engaging in gossip about me at every corner, both by words spoken and by letters written to whoever you choose. I remind you, speaking as a friend...*

Royal Appeal Court (Casa da Suplicação) collected and filed in the courts proceedings many of these letters. In the CARDS Project (Cartas Desconhecidas), 2.000 of such documents were detected, contextualised, and transcribed by a team of linguists and historians of the Early Modern period. The project ran from 2007 to 2010. The role of the linguists was to decipher and publish the manuscripts with philological care in order to preserve their relevance as sources for the history of language variation and change. The role of the historians was to contextualise the letters' discourse as social events. Almost half of the documents came from early 19th century criminal lawsuits of the Royal Appeal Court, and the other half from Inquisition lawsuits of the 17th and 18th centuries (plus a small sample from the 16th). In a complementary way (10 per cent), aristocratic families' legacies were also searched for private letters. The whole set of transcriptions, accompanied by a context summary, was given a machine-readable format, which allowed for the assemblage of an online Portuguese historical corpus of the Early Modern Ages.

As a sequel to CARDS, the FLY project (Forgotten Letters, Years 1900-1974) was launched in 2010 by the same core team, now accompanied by Modern history experts, as well as sociology experts. The aim was to enlarge the former corpus with data from the 20th century. Since collecting personal papers from contemporary times is a delicate task, given the need to guarantee the protection of private data from the public scrutiny, the letters of the FLY project come mostly from donations by families willing to contribute to the preservation of the Portuguese collective memory having to do with wars (World War I and the 1961-1974 colonial war), emigration, political prison and exile. These were also contexts favourable for a high production of written correspondence with family and friends because in such circumstances strong emotions such as fear, longing and loneliness are bound to arise.

The CARDS-FLY corpus [6] is thus a linguistic resource prepared for the historical study of Portuguese language and society. Its strength lies in the broad social representativeness, being entirely composed of documents whose texts belong to the letter genre, the private domain, and the informal linguistic register.

The current version of the CARDS-FLY corpus contains 3,455 letters with 1,155,206 tokens involving 2.237 different authors and addressees.

We show an example in Figure 2 of a digital transcription of a letter written in 1592, the version on the left side has the original spelling, apart from word boundaries normalization (except for enclitics), the right side was manually normalized for spelling[1]. This letter exemplifies the characteristics of this corpus of written historical letters: many letters do not contain punctuation marks, there are accents like the tilde that no longer have the same distribution in the current spelling, capitalization is used in a different way and does not signal sentence starts. Abbreviations such as *vm.* [P: Vossa Mercê E: Your Honour] are often used in these personal letters. Another difficulty is that there is not always a one-to-one mapping between words in the old and new spelling, since orthography was non-existing and creative spellings were far from rare, especially when writers were half-illiterate men and women.

3 Related work

As mentioned in the introduction, modernizing historical text aids information retrieval (IR) results. Another strategy is to adapt the search interface in such a way that it can cope with the spelling variation. This approach was taken by Gerlach et al. [4] who use a modern lexicon combined with transformation rules to expand the search query to capture also the spelling variants in the German historical text collection being searched. Other studies discuss several distance measures that augment the search query with fuzzy string matching [9] or acquire edit distance weights using unsupervised learning techniques [7].

A tool that was specifically developed to normalize the spelling of full texts, is the VARiant Detection (VARD) tool [13] that was developed for Early Modern English. We will discuss in detail its follow-up, the tool VARD2, in section 4.1.

Craig and Whipp [3] developed a method for automatic spelling normalization for early modern English. They combine list lookup of variants together with more sophisticated methods based on approaches taken in Word Sense disambiguation tasks to resolve ambiguous spelling variants that can be normalized to multiple modern forms.

The Historical Dictionary of Brazilian Portuguese (HDBP) is constructed on the basis of a historical Portuguese corpus of 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 to produce reliable frequency counts. Therefore [5] developed an automatic rule-based variant detection method and created a spelling variants dictionary containing approximately 30K clusters of variants (we refer to this list as the HDBP-variant list).

[1] Full description at: `http://alfclul.clul.ul.pt/cards-fly/index.php?page=infoLetter&carta=CARDS4006.xml`

The Corpus of Early English Correspondence (CEEC) [11] is a corpus similar to the CARDS-FLY corpus as both corpora consist of two collections of historical letters, although the CEEC, contrary to CARDS-FLY, is not based on previously unpublished material. A start is said to be made with a spelling normalization step using the VARD2 tool.

4 Tools

4.1 VARD2

The VARD2 tool [2] is a Java program with several options for the normalization of spelling variation in text. The tool offers an interactive mode in which the program suggests a list of candidates for each unknown word in the text and allows users to select the best choice in a manner similar to the Microsoft Word spelling correction module. The tool can also be used to automatically correct a full text and it can be trained and tuned by the user for a specific data set. In these experiments we opted for this last option.

VARD2 works as follows. Each word is checked against a modern lexicon. Words that are not present in the lexicon are potential spelling variants. Note that this limits VARD2's capacity to the detection of non-word errors. For each potential spelling variant, a list of candidate modern counterparts is generated using a variant list consisting of pairs of variants and their modern counterparts, a character rewrite rule list and a Soundex algorithm to find phonetically similar counterparts. These modules together determine the confidence weight that is assigned to each candidate modern equivalent. VARD2 has a confidence threshold that determines what weight is needed to actually replace the variant with the highest weighted modern equivalent that exceeds the minimum threshold. If no likely candidates are found, the variant is kept.

Hendrickx and Marquilhas previously adapted the VARD2 tool for the Portuguese language [8] and here we use their Portuguese version of the tool. They replaced the English versions of the modern lexicon, the variant list with pairs of variants and their modern counterparts and the rewrite rule list with Portuguese versions. As variant list they used the HDBP-variant list combined with a variant list extracted from training material. The rewrite rule list is based on the edit rules automatically generated by the DICER tool [1] which takes as input a list with spelling variants and modern equivalents and extracts character string transformation rules to capture the spelling variation. Only those rules that occurred 5 times or more were retained. The rules in this set that were too generic were manually edited to be made more specific. The final rewrite rule set of VARD2 consisted of 99 rules. VARD2 can be trained on a data set that is already manually normalized. When VARD2 is being trained, the program adds all normal words to the modern lexicon and adds all variants from the training material and their frequencies to the variant list. The different confidence weights for each replacement method are also adapted on the basis of the training data.

4.2 TICCL

Text-Induced Corpus Clean-up (TICCL) is described more fully in [16]. It is now a fully-fledged web application and service due to CLARIN-NL project TICCLops. The main lexical variant look-up mechanism in TICCL is based on anagram hashing. Informally, this works as follows: for all the words in the lexicon and in the corpus, a numerical representation called the anagram value is calculated by making the sum of the code page values of the individual characters of the word raised to power five. The numerical value obtained is used as the index key in a hash, the actual symbolic word(s) having this value are added as the hash value. Words consisting of the same bag of characters will have the same anagram value. This is why this is called anagram hashing. Given the anagram for a particular word (or set of anagrams), called the focus word, TICCL builds list A which contains the anagram values for all the individual characters in the bag of characters as well as the anagram values for all the possible combinations of any two characters in the bag of characters. TICCL also has list B, which has the same for all the characters in the alphabet. Given a focus word, all its near neighbours can now be found by exhaustively subtracting any value from list A from the focus word's anagram value and adding any value from list B. If the resulting anagram value is present in the anagram hash, a numerical near neighbour has been detected. Retrieving the symbolic values, i.e. the actual word or set of anagrams, the edit distance to the focus word for each needs to be checked. Correction Candidates or CCs are those instances that differ by less or equal the number of edits allowed by the Levenshtein distance (LD), the limit of which is set by a TICCL parameter.

The number of look-ups required per focus word depends on the size of the alphabet. In prior work, in order to reduce the search space, TICCL was used with a reduced alphabet. Its lexicons and the corpus it works on were thoroughly normalized by e.g. rewriting any character bearing a diacritical mark as a single digit and all punctuation marks as another digit, all numbers and digits present having been normalized into as single, different, digit. In this work, we retain all unicode points below a high unicode number. This has the benefit of not having to restore or retrieve the original diacritical word form for output purposes.

Further, we have worked only in what [16] calls **focus word mode**, the corpus not being very large. In character confusion mode, TICCL scales to the largest corpora.

New in the present work is that TICCL has been applied to Portuguese and has been equipped with both absolute correction [10] and bigram correction capabilities.

Converting TICCL to Portuguese involved little more than providing it with a Portuguese lexicon, which was the same one as used for VARD2. Derived from the lexicon is a **word confusion matrix** by applying TICCL's character confusion module. In its essence this matrix is a list of the anagram value references between each word in the lexicon and all the other words in the lexicon that are reachable within the confines of the particular Levenshtein or edit distance set. In the present

work we have limited ourselves to LD set to 2 edits.

This word confusion matrix in fact provides the list of all possible **confusables** (also known as real-word errors in spelling correction or 'false friends' [14]). The operative definition of confusables is therefore that they are those words that can be formed from any given focus word in the lexicon by applying at most the number of edits implied by the LD handled. Use of the word confusion matrix allows for preventing the system from returning a valid lexicon word for any given valid (because present in the lexicon) focus word. This in fact implies that in its current implementation, TICCL cannot perform real-word correction. We will return to this matter in the discussion of the results.

Also in the current implementation, **bigram correction** is applied. In terms of the informal discussion of how TICCL works above it is easy to see that given the anagram values for all combinations of two consecutive words in the corpus, the corpus bigrams, the same mechanism can be applied to retrieve bigram CCs. This is because the numerical distance between the anagram values for e.g. the English words 'cat' and 'rat' will be the same for the likely bigrams 'the cat' and 'the rat' or 'white cat' and 'white rat', i.e. the numerical anagram value distance for 'c' to 'r'. Bigram correction is here applied only to short words. In prior work, a lower word length threshold was always applied. The word length threshold in our unigram mode experiments here was set to six characters. For very short words the lexical neighbourhood is very dense, substituting just one or two characters leads to very many other short words. We here try to overcome this problem for short words by looking in the corpus for variants for the bigrams they occur in. In doing so, TICCL handles short word bigrams as if they were just ordinary unigrams, the only difference being that now the space character is also at play. In its essence, by searching in only the bigrams containing a particular short focus word, the possible search space is effectively and efficiently reduced by the contexts it shares with potential near neighbours. Having retrieved variant bigrams, the overlapping, exactly matching left- or right bigram part is then discarded and the remaining pair of unigram variants is further handled in exactly the same way as the longer word pair variants which would have been retrieved. This approach has its limitations and this too will be further dealt with in the results section.

The **absolute correction** strategy was defined by [10], who called it 'limited but very cost-effective'. We equipped TICCL with absolute correction capabilities based on the collection of lexical variants as present in the training set. Put in a misspelling dictionary, when one of the known historical variants is encountered, it is simply replaced by the contemporary form.

5 Experiments

For the evaluation experiments of TICCL and VARD2 we use a subpart of 200 letters from the CARDS-FLY corpus. These letters were manually normalized by one linguist but difficult cases were discussed with a second expert. This data set

was split in 100 letters for training and tuning the tools, and 100 letters were set apart as a true test set. The test set contains 37,372 tokens of which 6,978 (19%) are spelling variants that need to be detected and normalized by the tools. We measure the performance of the tools and compute accuracy, recall, precision and their harmonic mean, F-score, on the spelling variants.

In our experiments with TICCL on the training set we learned that absolute correction using all the pairs in the variant list was highly detrimental to precision. In the end we settled on a subdivision where words that are only in the corpus and not in the lexicon were allowed to be absolutely corrected. Words that are in both were evaluated on whether they were ambiguous or not, in the sense that they have more than a single possible resolution in the variant list. Those that were ambiguous were not let to be absolutely corrected, with the one exception of the pair 'q-que'; the others were let be handled by TICCL's proper correction mechanism. The ones that were not ambiguous were not corrected, with the exception of three pairs ('hum-um', 'porem-porém'[2] and 'exmo-excelentíssimo'), which were absolutely corrected. Absolute correction, when applied, was given precedence over whatever TICCL had retrieved in all cases. For ambiguous cases we retained the most frequent variant in the variant list only, which results in a single CC for all instances of absolute correction, enhancing ranking.

VARD2 was trained on the training material of 100 letters and this tuned each of the modules for this particular data set. We used VARD2 in its "batch" mode in which each detected spelling variant is automatically replaced by its best-first ranked word form.

In table 2 we show the best-first ranked performance of TICCL and VARD2 on the test set. The results reported for TICCL are those obtained when TICCL had access only to the variant list obtained from the training set. TICCL2 reports results obtained when TICCL had access both to the list obtained from the training set, as well as to the HDBP-variant list. VARD2 was trained on both.

Tool	acc	prec	recall	f-score
VARD2	94.65	96.99	73.63	83.71
TICCL	93.25	94.27	67.96	78.98
TICCL2	93.50	94.38	69.33	79.94

Figure 2: Best-first ranked results on the test set of 100 letters

5.1 Test results analysis

It should be noted that the results reported are necessarily precision and recall scores on tokens, not on word types because of the ambiguity of part of the original tokens which may have to be resolved to different contemporary word forms.

[2] Actually, 'porem' is ambiguous in Modern Portuguese but not in this corpus.

Tool	acc	precision	recall	f-score
TICCL-bi-rank3	94.11	94.62	72.57	82.14
TICCL-bi-rank5	94.35	94.71	73.89	83.01
TICCL-bi-rank10	94.55	94.78	74.92	83.69
TICCL-bi-rank20	94.66	94.82	75.52	84.08
TICCL-uni-rank20	94.42	95.03	73.99	83.20
VARD2-notraining	90.58	93.79	53.05	67.77
TICCL-bi-rank20-noabsolut	89.18	92.03	46.02	61.35

Figure 3: Results on the test set of 100 letters measuring TICCL's 3, 5, 10 and 20 first-best ranking with bigram correction and with absolute correction. Also shown is the effect of TICCL not performing bigram correction. Finally, the effects of VARD2 and TICCL not having been trained/using absolute correction with the variation list(s)

The accuracy of the original corpus before correction is 81.33%. This means that less than 20% of the original texts need to be normalized. VARD2 manages to improve texts by 13.32%. TICCL manages an improvement of 11.52% and TICCL2 reaches 12.17%.

VARD2 returns only best-first ranked results. These are compared with TICCL's best-first ranked results in Table 2. VARD2, being specially trained on manually edited rules specific for the task, is the clear winner. TICCL has not received any special training, but has had bigram correction at its disposal and has been equipped with new code for dealing with the absolute correction.

Results reported in the upper half of Table 3 show clearly that if TICCL's ranking mechanism might be improved, it can potentially best VARD2. These are results obtained when TICCL's absolute correction had access only to the variant list obtained from the training set. Also, in these experiments, TICCL lacked the benefit of a background corpus of contemporary Portuguese bigrams.

In the lower half of Table 3 we show the results of a few ablation tests. First we give the result of running TICCL in unigram mode only, then we show what VARD2 and TICCL manage to accomplish without having the benefit of the information in the variant list(s).

In unigram mode only TICCL is in fact a bit more precise. But it necessarily loses recall: it has not itself retrieved any variants for words shorter than six characters other than the ones it has been able to resolve through the absolute correction. This clearly shows the improvement due to the bigram correction.

Further in the same table we also show performance results obtained when the systems have not had the benefit of the domain specific variant list(s) either for training or for the purpose of absolute correction. Both benefit a great deal, but TICCL clearly most. As the CARDS-FLY corpus consists of data from a very specific textual genre with typical characteristics, we observe that training and tuning

the spelling checkers explicitly on this genre, leads to a substantial performance gain.

6 Discussion

The results of the comparison of TICCL and VARD2 shed valuable light on the problem of historical spelling normalization.

First and foremost, the results show that there is an upper limit to what can be achieved with what is essentially non-context sensitive correction on historical data. Probably neither of the systems in these tests have reached this upper bound, but both nevertheless get close to it.

The situation is that TICCL's absolute correction and VARD2's equivalent, a special purpose rule, work splendidly for cases such as 'q' which unambiguously should be normalized to 'que'. However, the instances of the single character 'v' in the historical letters are variously resolvable to 'via ' (on just 1 occasion), 'vossa' (42 times), 'vossas' (17 times) and 'vosso' (just once). The corpus contains many more similar instances. This implies full-fledged context-sensitive correction, which neither system can currently provide.

It would be legitimate to wonder if the task undertaken here should not be divided over two separate tasks, to be handled possibly by different systems and evaluated separately. The results show that about one quarter of the test instances cannot be solved by either of the systems. This implies that either the systems need to be equipped with mechanisms that do allow them to be solved, or that other systems or approaches should be sought and applied.

As it is, the systems are measured in part on test instances they were not designed to be able to handle. This in part obscures their capabilities of handling what they were meant to be able to handle.

TICCL will probably never be set to handle spelling variation exceeding LD 4. Even applying LD 3 would have an adverse effect on its precision. It could nevertheless, much in the way VARD2 is, be taught on the basis of the variant list(s) to look for specific higher edit distance variants. The variant list has, e.g. the pair 'exmo', an abbreviation, which should be expanded to 'excelentíssimo'. This is currently handled by the absolute correction, but the test set might as well have the pairs 'exma-excelentíssima', as well as the plural forms. By teaching TICCL to look for the anagram value for 'elentíssi', this desirable generalization would be achieved. This we will implement in future work.

The results in Table 3 over the various ranks show that TICCL potentially reaches the same or an even higher level of performance as VARD2 currently does. Our conversion of TICCL to Portuguese so far has been inadequate in that it cannot deal with the higher degree of morphological variation in Portuguese compared to Dutch and English which it had so far been applied to. Also, in these experiments it turned out that the ranking mechanism described in [16] did not deliver the results hoped for. The performance results reported in Table 3 were obtained by

leaving out the frequency information from the final ranking of CCs retrieved. In the absence of a large, contemporary background corpus, the frequencies observed in the historical test corpus were too sparse or totally unavailable for contemporary word forms and their bigram combinations. Best results, as reported, in these tests were obtained by the combination of ranking on frequency of character confusion observed and LD, only.

A fruitful path for future work is to study the strengths of both systems and to see how these might be combined. The DICER tool provides a wealth of statistics on the variation present in the training set. Certainly TICCL would benefit from direct use of these statistics in its ranking of CCs. Also, we should study how to address more properly the problem of morphological variability in TICCL's ranking. In conclusion, some generalization from the information available to TICCL in the absolute correction list should be of benefit. Certainly for higher LD variation which is highly present in these historical texts due to the high incidence of abbreviations, providing TICCL with the common character confusions above the LD limit it is set to work with, would allow it to emulate VARD2 in that it would then also be trained to explicitly identify and retrieve these variants.

7 Acknowledgements

TICCL has further been developed by Martin Reynaert in the Dutch NWO project Political Mashup. Iris Hendrickx was funded by FCT Doctoral program Ciência 2008. Rita Marquilhas is supported by the projects FLY (PTDC/CLE-LIN/098393/ 2008) and Post Scriptum (ERC, Adv Grant 2011, GA 295562).

References

[1] A. Baron. *Dealing with spelling variation in Early Modern English texts*. PhD thesis, University of Lancaster, Lancaster, UK, 2011.

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

[3] H. Craig and R. Whipp. Old spellings, new methods: automated procedures for indeterminate linguistic data. *Literary and Linguistic Computing*, 25(1):37–52, April 2010.

[4] A. Ernst-Gerlach and N. Fuhr. 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, 2007.

[5] R. Giusti, A. Candido, M. Muniz, L. Cucatto, and S. Aluísio. 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*, 2007.

[6] M. Gomes, A. Guilherme, L. Tavares, and R. Marquilhas. Projects CARDS and FLY: two multidisciplinary projects within linguistics. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[7] A. Hauser and K. Schulz. 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, 2007.

[8] I. Hendrickx and R. Marquilhas. From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation. *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(2):65–76, 2011.

[9] S. Kempken, W. Luther, and T. Pilz. Comparison of distance measures for historical spelling variants. In *Artificial Intelligence in Theory and Practice*, volume 217, pages 295–304. Springer Boston, 2006.

[10] J.J. Pollock and A. Zamora. Automatic spelling correction in scientific and scholarly text. *Commun. ACM*, 27(4):358–368, 1984.

[11] H. Raumolin-Brunberg and T. Nevalainen. Historical sociolinguistics: The corpus of Early English Correspondence. *Creating and Digitizing Language Corpora*, 2: Diachronic Databases:148–171, 2007.

[12] P. Rayson, D. Archer, A. Baron, J. Culpeper, and N. Smith. 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, 2007.

[13] P. Rayson, D. Archer, and N. Smith. 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), 2005.

[14] U. Reffle, A. Gotscharek, C. Ringlstetter, and K. Schulz. Successfully detecting and correcting false friends using channel profiles. *IJDAR*, 12(3):165–174, 2009.

[15] M. Reynaert. *Text-Induced Spelling Correction*. PhD thesis, Tilburg Universtity, 2005.

[16] M. Reynaert. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187, 2010. 10.1007/s10032-010-0133-5.