MANUSCRIPTS AND MACHINES: THE AUTOMATIC REPLACEMENT OF SPELLING VARIANTS IN A PORTUGUESE HISTORICAL CORPUS

RITA MARQUILHAS AND IRIS HENDRICKX

Abstract *The CARDS-FLY project aims to collect and transcribe a diverse sample of historical personal letters from the 16th to 20th century in a digital format to create a linguistic resource for the historical study of the Portuguese language and society. The letters were written by people from all social layers of society and their historical, social and pragmatic contexts are documented in the digital format. Here we study one particular aspect of this collection, namely the spelling variation. Furthermore, on the basis of this analysis, we improved a statistical spelling normalisation tool that we aim to use to automatically normalise the spelling in the full collection of digitised letters.*

Keywords: historical linguistics, spelling variation, automatic normalization, Portuguese

I. INTRODUCTION[1]

Personal letters can have a twofold importance for historians. First, they play a supporting role as documents that contain first person testimonies (with all their flaws of accuracy, to be sure) ready for interpretation alongside all other available sources on whatever topic the historian is studying. Under this light, they often become 'a providential manna to feed biographies, the sketch of everyday life, the taste for intimacy and confidential matters'.[2] Secondly, historians can also find personal letters to be important for their own sake, if the context is that of a history of written culture.[3] Here they play the leading role,

International Journal of Humanities and Arts Computing 8.1 (2014): 65–80
DOI: 10.3366/ijhac.2014.0120

www.euppublishing.com/ijhac

given their status of social practices, 'traces of a complex reality that absorbs countless other practices and registers'.[4] For example, they enclose literate (and halfliterate) discourses on the practice of writing itself. Also, they are samples of intimate interactions, whose participants were conscious of the spatial-temporal discontinuity of their speech acts. They constituted either polite or impolite behaviour, either orderly, or disorderly conduct, depending on the observance of conventions valid for the historical communities in question.

For this second approach, nevertheless, rich collections of letters are mandatory because cultural interpretations have to be tested against a large quantity of data that represents the norm followed by social actors, and a thin quantity of exceptions that constituted possible marginal behaviours.

At the Linguistics Centre of the University of Lisbon (CLUL), such a large collection is being assembled, the CARDS-FLY corpus, in order to attend both the needs of cultural historians, and the needs of historical linguists. Historical linguistics is the study of language change through time, and original, non-literary sources are the most preferred data for the description and interpretation of such change. Spontaneous oral utterances would be the ideal data, but since their retrieval is impossible for language as spoken in past centuries, the personal letter discourse is the next best candidate. It offers the linguist the recording of a behaviour carried out by interactive speakers with a more informal attitude than the one adopted by writers of literary or institutional texts.

The CARDS corpus (*Cartas Desconhecidas – Unknown Letters)* is a collection of 2,000 personal Portuguese letters written between the 16th and the 19th century. The ones dating from 1500 to 1800 were mainly seized by a religious court (the Portuguese Inquisition) as instrumental proof to prosecute individuals accused of heretical beliefs. As for the 19th century ones, they were mainly seized by a Crown court (the *Casa da Suplicação*) as instrumental proof exhibited either by the prosecution or by the defence of individuals accused of anti-social or anti-political behaviour.

The project ran from 2007 to 2010, carried out by a mixed team of historians and linguists. 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. 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 Early Modern Ages.

In the sequence of 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 sociologists. 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 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 favourable contexts for a high production of written correspondence with family and friends because in such circumstances strong emotions such as fear, longing, and loneliness were bound to arise.

The CARDS-FLY *corpus* is thus a linguistic resource prepared for the historical study of Portuguese language and society. Its strength relies on the broad social representativeness, being entirely composed by documents whose texts belong to the letter genre, the personal domain, and the informal linguistic register.[5] The final goal is to have a total of 4,000 letters. By May 2013, the team had already transcribed a total of 3,809 letters involving 2,286 different participants (82 per cent men, 18 per cent women) and around 1,1 million words.

The digital encoding of the letters follows a set of guidelines prepared by the Flemish project DALF: Digital Archive of Letters in Flanders based on the TEI P4 Guidelines.[6] This encoding offers a machine-readable file format that allows for the philological care critical editions demand. The mark-up language is XML, and the labels contents are the ones fixed by DALF for letters idiosyncrasies and by TEI for primary sources.[7]

The letters manuscripts were transcribed in a conservative way and features such as unreadable parts, scratched-out parts or perforations in the letters are encoded explicitly in the XML mark-up. Also the spelling of the original document is maintained, as this is relevant for the history of language change, a prospect that is always compromised when spelling normalisations are practiced by editors. On the other hand, the lack of normalisation for spelling creates a problem when the letters are seen as a target for corpus linguistics operations: morphologic annotation, parsing, semantic annotation, concordancing, word lists, and keywords. Such level of processing demands for a *corpus* in standard spelling, a resource also invaluable for historians focusing on the discursive features that manifest themselves through keywords and semantic fields present in the corpus.[8]

As we intend to use the corpus for this purpose, we are in need of a normalised version. Manual spelling correction is a laborious and time-consuming effort and therefore we decided to explore the possibilities for automatic normalisation. We already did some exploratory experiments along that path.[9]

Here we first give a detailed analysis of how spelling varied and changed over time in our *corpus* based on a statistical analysis of a sample taken from the CARDS-FLY corpus. Next we present some practical results of automatic spelling normalization. We conclude with a discussion of the benefits and limits of using statistical methods for spelling normalization but we conclude that the benefits of the procedure are indeed remarkable ones.

2. SETTING A STANDARD FOR WRITTEN PORTUGUESE

In the history of Portugal the standard norm for written language came late in time, only in 1911, one year after the Republican instauration. The standard adoption had been persistently proposed since the 18th century, following foreign examples, but there was never a favourable occasion for the Royal Academy of Sciences of Lisbon (*Academia Real das Ciências de Lisboa*) to produce a written model, neither in the 1700s nor in the 1800s.[10]

When a Portuguese orthography could finally be decided, there were two possible paradigms that would serve as alternative models: the *shallow* orthography, such as the Spanish and the Italian, which preserved phoneme-grapheme correspondences, and the *deep* orthography, followed in the French and the English spelling standards. The *deep* paradigm, more etymological, is a type of spelling where morphology, rather than phonology, is recoverable by literate people.[11]

The authors of the 1911 Portuguese spelling reform decided openly for the *shallow* paradigm. They motivated their choice as a way of creating the proper instrument that would lead to a quick progress of literacy rates in Portuguese society:

> What are the bases for the Portuguese orthography that our Commission proposes?
>
> There was, from the beginning of the works, two systems that could be followed. One of them was the French orthography, which, more or less coherently, is being imitated in Portugal for some time now. The other system is the one of the Spanish and Italian orthographies, much simpler, more rational, logical and easy to learn, much more adapted to the natural and even literary evolution of those languages, which is also similar to the evolution of Portuguese. What radically differentiates the orthography of those two official languages [Spanish and Italian] is the modification of the Latin spelling of innumerable Romanised Greek words to other spellings, much more similar to the value of the letters of such words in modern times.
>
> In order to make the teaching of reading and writing an easier task, the Commission found that the time had come to banish once and for all from the Portuguese writing, as they were banished from the Spanish and the Italian for a long time, [. . .] the symbols *ph*, *th*, *rh*, and *y* [. . .].
>
> Translated from the Portuguese *Bases da Reforma de 1911*.[12]

The 1911 reform put an end to a long search for a Portuguese standard for spelling. But it raised a diplomatic misunderstanding between Portugal and Brazil, a problem that took a new period of 100 years to be solved. In 1990, all the Portuguese speaking countries signed an agreement on a decisive spelling reform. In 2011 that reform was finally adopted by the Portuguese education system.[13]

3. AUTOMATIC SPELLING NORMALISATION

3.1. Related Work

Here we first give some examples of recent related studies that handle spelling variation in historical corpora in general and then focus on studies for the Portuguese language.

The VARiant Detector (VARD) tools aimed to detect spelling variation in Early Modern English and were created for corpus linguistic research.[14] The first version of the tool was based on a list of manually created mappings between historical variants and their modern versions. The latest version combined several different modules such as a list of letter replacement rules, a phonetic matching algorithm and an edit distance search method to detect spelling variation. We discuss a Portuguese version of VARD in the next section. Craig and Whipp have also worked on a tool for automatic spelling variation detection for Early Modern English but in the perspective of authorship attribution.[15]

For the corpus of Early Modern German, a spelling variation detection tool is currently under development.[16]. For the Spanish diachronic corpus, a study of the effect of automatic spelling normalisation has been conducted.[17] They compared two different strategies, namely to first automatically normalise the data before using an NLP tool or to adapt the NLP tool itself to handle spelling variation. For their purpose of Parts of Speech tagging, they argued that tool adaption is better as the original spelling is kept.

As for Portuguese, most of the available studies concerning the spelling change along Early Modern and Modern times have a cultural historical perspective, which means that what they analyse is the discourses of contemporary *élite* writers, mostly grammar authors and dictionary authors. Such discourses were either bitter criticisms because of the lack of a spelling standard for the language, or concrete proposals for a solution to that void.[18]

As for quantitative corpus-based approaches of the same spelling change, they had to wait for the assemblage of large Portuguese historical *corpora* covering the Early Modern and Modern era, a work that is being mostly undertaken in Brazil.

The Tycho Brahe team, of Campinas University, was the first to present statistical measurements of the spelling change phenomenon in order to solve the processing problems it raised,[19] followed by the Historical Dictionary of Brazilian Portuguese team (*Dicionário Histórico do Português do Brasil*).[20] This dictionary is constructed on the basis of a historical Portuguese corpus (16th to 19th century) of approximately 5 million tokens. As they needed a normalised corpus to produce reliable frequency counts for the dictionary, they developed a rule-based method to automatically cluster spelling variants together. They clustered spelling variants around one common word form that is not always a modern word form, but the most central word form in the cluster of related variants leading to a spelling variants dictionary.[21]

A resource very similar to the CARDS-FLY *corpus* is the Shared Diachronic Corpus: Personal Brazilian Letters (*Corpus Compartilhado Diacrônico: cartas pessoais brasileiras*), which consists of a Brazilian collection of historical personal letters from the 18th to 20th century.[22] The aim is to provide the academic community with a resource for the sociolinguistic history of Rio de Janeiro's society along 300 years. The documents in this collection have also been normalised for spelling, but all normalisation was done manually, with the help of a friendly tool, namely E-Dictor, offered by the above-mentioned Tycho Brahe project.[23]

3.2. DICER

Similarly to the Brazilian experiments, our study also uses a statistical corpus-based approach to get a better insight in the Portuguese spelling variation over the 16th–20th century time span. Our major originality is that we deal with an ultra-varied corpus, entirely made up of text within original letter manuscripts, either written by common people, or by *élite* people in common moments of their lives.

We extracted a random sample from the CARDS-FLY corpus of 200 letters. These letters were manually normalised to the modern spelling by a linguist. Each word in the documents that was labelled as spelling variant was paired with its modern spelling counterpart. This sample was intended both for a manual inspection and analysis of the spelling variation present in the data, and for the development of an automatic tool for spelling normalisation. For the latter purpose, we split the sample in two parts. We used a hundred letters for training and tuning the automatic normalisation tool for this specific genre. The other hundred letters are used for evaluation of the tool as we can compare the manual normalisation against the automatic normalisation produced by the tool. We set apart the evaluation set and excluded it from any manual analysis. Tuning an automatic tool to the errors in the evaluation set would lead to a tool that performs very well on this one set but it might lead to an overly optimistic estimation of the true performance of the tool on other, unseen material.

DICER (Discovery and Investigation of Character Edit Rules) is a statistical tool that creates a list of edit rules on the basis of a corpus labelled with spelling variants and their modern counterparts.[24] The tool uses these pairs to detect which character(s) differ between the variant and the modern word, and it produces simple edit rules that capture the steps to rewrite the old word form to the modern form. The edit rules express what characters are being changed, what type of operation (deletion, insertion or substitution) is applied, and on which location of the word (start, second, middle, penultimate or end).

To rewrite a spelling variant to its modern form may need multiple different rewrite rules. For example, *apezare* is a variant in our historical data for the modern form *apesar* 'despite' and the transformation requires two edit rules: 'substitute < z > with < s > ', and 'delete < e > '. DICER creates a new rule for every edit that it encounters in the corpus and therefor gives a full statistical and systematic overview of the spelling changes that are present in the corpus.

Below we show a detail of the DICER results summary, after the processing of the CARDS-FLY corpus sample of a hundred letters. The summary shows the operations involving word types (not tokens). The table captures the ten top edit rules on the modernisation of those types. We can see that the substitution of < z > by < s > , especially when the < z > letter appears in the middle or in the penultimate position, is the edit rule that has been applied most frequently, namely 193 times, as shown in the column labelled as 'Total' (see Table 1).

Since DICER finds all the edit rules involved in the modernisation process, it follows that a close examination of column 'Variant' *versus* column 'Standard', combined with the number of different word types that changed (column 'Total') will give us a good snapshot of the variation problems we have to face when dealing with the CARDS-FLY corpus.

The letters authors were either following old spelling traditions, later abandoned, or, in the case of half-illiterate authors, also struggling with the rationale of the general spelling usage of their time, either old or modern.

A computation of the spelling behaviour of those authors, as compared to modern Portuguese orthography, tells us that a total of 718 edit rules were needed in order to modernise the sample of 100 letters, and that these rules affected, one or more times, a sum of 3,450 different word types. When summing all operations of the 718 edit rules, we counted 4,225 different operations, which means that several of these word types had to be standardised step by step by multiple edit rules.

In order to have a manageable, humanly observable, sample of this large population of data, we only examined the rules that were applied at least three times, leaving aside the less frequent ones. The resulting sample had a large lexical representativeness (3,590 operations) but a feasible number of edit rules (only 171).

In the following two tables we show an interpretation of how the 171 top edit rules of the DICER tool could be distributed in terms of rule contents. The most frequent changes involved the spelling of phonological features (67 per cent), and, within these, the spelling of coronal fricatives was the most critical problem presented by our corpus variation (see Table 2 and Table 3).

Table 1. The DICER standardizing edit rules on the CARDS-FLY corpus (detail).

						Position				
#	ID	Operation	Variant	Standard	Total	Start	Second	Middle	Penultimate	End
1	8	Substitution	Z	S	193	0	4	132	46	11
2	20	Substitution	S	SS	164	2	20	89	53	0
3	149	Substitution	M	N	137	0	50	86	1	0
4	76	Insertion		-	123	0	1	117	5	0
5	40	Substitution	ÃO	AM	121	0	2	0	0	119
6	10	Substitution	S	C	118	23	10	78	7	0
7	45	Substitution	I	E	117	33	41	37	3	3
8	22	Substitution	I	Í	107	2	11	90	2	2
9	68	Substitution	E	I	106	23	35	40	8	0
10	6	Substitution	A	Á	92	5	8	36	9	34

Table 2. Causes for spelling variation in the CARDS-FLY corpus.

General cause	Specific cause	Word types to standardise
Phonology	coronal fricatives	860
Phonology	unstressed oral vowels written with < i > , < e > , < u > , < o >	456
Phonology	nasal vowels and diphthongs	426
Phonology	stressed oral vowels	408
Mixed	mixed	308
Graphic Tradition	abbreviations	267
Graphic Tradition	learned consonant groups, digraphs, and double consonants: < ct > , < pt > , < ph > , < pf > , < pp, < ff > , etc.	233
Syntax	enclisis: hyphenated verbal forms, with or without sandhi, followed by clitic pronoun *vs.* non hyphenated verbal forms	154
Graphic Tradition	etymological *vs.* non etymological initial < h >	136
Phonology	non standard phonology (dialectal variation)	132
Graphic Tradition	archaic letters: < y > *vs.* < i > , < u > *vs.* < v > , < i > *vs.* < j >)	95
Phonology	liquids /l, r, R/	63
Phonology	labialised velar stops /k^w, g^w/ *vs.* velar stop /k, g/[1]	52
TOTAL		3590

[1] We follow here Maria Helena Mateus and Ernesto d'Andrade, who present a case for the existence of segment /k^w/ in the phonology of Portuguese: M. H. Mateus and E. d' Andrade, *The Phonology of Portuguese* (Oxford, 2000).

The fact that the CARDS-FLY *corpus* is composed by original manuscripts, instead of printed texts, together with the large variety of their authors' social status, accounts for such a distribution of spelling variants. This means that much of the correspondence was written in a close-to-spoken manner, without the opportunity of being revised by a more literate copywriter.

The above results also reveal the most important stumbling block in the Portuguese modern spelling system when the researcher wants to modernise historical written matter. That stumbling block is the lack of correspondent letters for the distribution of voiced and voiceless coronal fricatives.

Table 3. Summary of spelling variation in the CARDS-FLY corpus.

General cause for variation	Frequency of word types to standardise	Rate of word types to standardise
Phonology	2397	66,7%
Graphic Tradition	731	20,4%
Mixed	308	8,6%
Syntax	154	4,3%
Totals	3590	100%

In the Middle Ages, Southern Portuguese dialects were already experiencing *seseo* (the merge of the dental-alveolar affricates /ts, dz/ and the dental-alveolar fricatives /s, z/).[25] Today only the archaic variety of the North-Eastern area keeps a distinction between four segments, articulating different fricatives in the middle of *passo* 'step', *paço* 'palace', *coser* 'sew', and *cozer* 'bake, steam'. Also, but later, from the 17th century on, the voiceless palatal affricate (traditionally written < ch >) merged with the voiceless palatal fricative (traditionally written < x >) in Southern and Central dialects, so that the phonological difference between words like *chá* 'tea', and *xá* 'shah' was lost.[26] All affricates disappeared in the innovative dialects, but since their traditional spelling was always kept by learned writers, including the ones that established the 20th century Portuguese orthography, it became a major source of variation in texts by poor writers along the centuries.

Nevertheless, if we split our data into chronological segments, it is clear that the major problem for 20th century uneducated letter writers is not the spelling of coronal fricatives. That problem is specific of earlier writers, especially the ones of the 18th and the 19th century. The major problem with standardizing the spellings of 20th century poor writers resides in the system of stressed vowels, which they normally write without the phonographic diacritics prescribed by the standard rules.

The other two more important sets of rules applied by the DICER tool have to do with the spelling of unstressed vowels and the spelling of nasal vowels and diphthongs, two phonological categories that are insufficiently mirrored by the Portuguese standard spelling. Neither the Spanish nor the Italian language, the overt examples that guided the creators of the Portuguese standard spelling in 1911, compare to Portuguese in what concerns the phonology of unstressed vowels and nasal vowels and diphthongs. So here the Portuguese spelling system became more etymological, less shallow, a feature that triggers several problems when it comes to standardizing historical data with many spelling variations.

3.3. VARD2

As a next step in our study we used the edit rules automatically generated by DICER to further improve the VARD2 tool for automatic spelling normalisation of historical Portuguese.[27] We already experimented with the tool VARD2 in a previous study, and here we show how DICER can contribute to a better performance.

VARD2 was initially developed for Early Modern English but we converted it to Portuguese. The system uses a modern lexicon to detect possible spelling variants in a historical input text. Words that do not occur in the modern lexicon are marked as possible candidates. The system checks for each candidate if it occurs in a variant dictionary, which lists frequent spelling variants and normalised equivalents. If the variant is listed, it is recognised as a true spelling variant and is replaced automatically by its modern equivalent. Otherwise, both rules based on phonological information and character rewrite rules are used to generate possible modern equivalents for the variant and associated confidence weights. One of the parameters of VARD2 is a confidence threshold that determines what weight is needed to replace the variant with the highest weighted modern equivalent that exceeds the minimum threshold. If no likely candidates are found, the variant is kept.

To convert VARD2 to the Portuguese language we replaced the English modules by Portuguese ones.[28] As modern lexicon we used the Multifunctional Computational Lexicon of Contemporary Portuguese.[29] We had created the variant list of spelling variants and their modern equivalents on the basis of an existing spelling variants dictionary extracted from the Historical Corpus of Brazilian Portuguese mentioned above.[30] We made several small improvements to the Portuguese modules in VARD2. When inspecting the modern lexicon, we noticed that even though it was extracted from a contemporary dictionary it still contained several archaic word forms. We attempted to filter out these word forms on the basis of a list of archaic word forms from the Houaiss dictionary.[31] We also used the list of spelling variants from the training sample of a hundred letters to filter the lexicon by deleting the variants and adding the modern word forms.

Furthermore, a manual check of the most frequent items in the spelling variant list was needed as we had already noticed that some variants were not mapped to a modern word form but to another, more frequent archaic word form. For example, in our previous experiments the variant list contained the archaic form *fforão* '(they) were/went' matched with equivalent *forão* instead of the correct modern counter part *foram*.

VARD2 uses a set of rewrite rules to generate the modern word form candidates. In our first approach we manually constructed such a list of rewrite rules based on our own intuitions and on the rule set described by Giusti et al.

Table 4. VARD2 scores on the development set with different thresholds for the rule set.

Threshold	Accuracy	Recall	Precision	F-score
5	93.0	74.3	98.5	84.7
10	93.0	739	98.5	84.5
25	92.8	73.0	98.6	83.9
50	92.2	70.6	98.7	82.3

Here we intend to investigate to what extent the automatically generated rewrite rules by the DICER tool can help improve the performance of VARD2. Our analysis and interpretation of the generated rule set presented above showed that the DICER was able to produce edit rules that capture a broad and diverse set of spelling changes.

As DICER generates a large rule list and some of the rules are based on evidence of only one occurrence, we decided to search for an optimal minimum frequency threshold for the rule set.[32] To get an indication for a suitable cut-off point, we ran experiments on the training set to see the effect of using rules that occurred at least 5, 10, 25 and 50 times. The higher the cut-off threshold, the smaller the rule set would be. The rule set with cut-off threshold 5 has 99 rules while a cut off of 50 only leaves 14 rules. We split the training sample in a part of 80 letters for training and 20 letters as a development set to determine the optimal rule set. We ran experiments with the different thresholds on the development set.

To evaluate the performance of the tool, we compute accuracy, recall, precision and F-score for the words (excluding punctuation marks) in the held out evaluation data. Recall expresses the number of cases in which there was a spelling variant in the text and the modern variant was correctly predicted by the tool, divided by the total number of predictions (errors because the tool predicted too many cases). Precision on the other hand focuses on the number of correct predictions divided by the number of true spelling variants in the data (errors because the tool missed some cases).

In table 4 we show the effect of varying the threshold on the development set. We do not observe huge differences between the different thresholds, but as the threshold of 5 had a slightly higher score, we decided to use this cut-off threshold for the experiments on the test set.

As we aim to study the effect of DICER edit rules on the VARD2 system, we made a comparison between the DICER edit rules, and the set of rules that we had manually created for our previous experiments. The manual rule set contains 62 different rules while the DICER rule set with threshold 5 contains 99 rules. When we compare the two rule sets, we notice only a few overlaps in rules. Both sets contain the rules to remove the double consonants < ll > , < nn > , < tt > ,

Table 5. A comparison on the test set of two versions of the VARD2 tool one with the DICER rule set and one with handcrafted rules.

Rule set	Accuracy	Recall	Precision	F-score
handcrafted	92.7	64.9	98.4	78.3
DICER	94.2	73.4	97.0	83.6

the substitution of < y > with < i > and some accent changes. The manual rule set contains many specific rules that cover multiple character strings such as 'substitute < zente > with < sente > at the End position'. The DICER tool however has more general rules that do capture the same event, for example the rule < z > - < s > is a generalisation of the 'substitute < zente > ' rule.

In the table 5 we show the results of the comparison VARD2 with the handcrafted rule set against a version of VARD2 trained with the DICER rule set with threshold 5 on the held out test sample of a hundred letters. Overall, we observe that VARD2 has a very high precision. The automatically generated rule set leads to a higher performance of 84 per cent F-score and 94 per cent accuracy. As shown in the table, the automatically generated rule set leads to a higher overall performance due to an increase of the recall. The DICER rule set enables the VARD2 tool to create a larger list of possible modern candidates thereby reducing the number of missed variants. For example, the variant *lansar* was not corrected by VARD2 trained with the handcrafted rule set, but it was correctly changed to *lançar* 'to launch' by the version trained with the DICER rule set as it included the edit 'substitute < s > with < ç > '.

In general, the limitation of VARD2 to only detect non-word errors causes a major part of the errors. To give an example, the noun *circunstancia* was not detected as a spelling variant because it is listed in the modern lexicon where it represents a conjugation of the verb *circunstanciar* 'to state in detail'. However, the modern equivalent of the noun has an accent: *circunstância* 'circumstance'. The information about the grammatical function of a word in the sentence is not available and therefor the system cannot detect this variant. In other cases VARD2 will chose the most likely and closest modern variant, and this may not be the best option in a given context. Like the form *frea* that can either be an abbreviation of *freguesia* 'parish' or a variant of *fria* 'cold'. A context-sensitive tool is needed to solve this type of problems but this is a line of future research as there are currently not many context-sensitive spelling normalisation tools available, certainly not for historical texts.[33]

4. CONCLUSIONS

We have presented an analysis of the main types of spelling variation that we encountered in CARDS-FLY corpus, a corpus of Portuguese historical personal letters that lacks standardisation because it corresponds to extremely varying sources, which were transcribed in a semi-palaeographic way. The systematic account of all spelling changes in the corpus sample, as generated by the DICER tool, shows the mixed nature of Portuguese modern orthography, not so much shallow as their inventors wanted it to be. This mixed nature of the modern standard clashes both with etymological spellings within the corpus, and with phonological ones.

As spelling variation can be a hindrance for certain types of research and for automatic search in the corpus, we presented a series of experimental results with the VARD2 statistical normalisation tool. This tool can automatically normalise variants with an F-score of 84 per cent and a precision of 97 per cent. A high precision means that when VARD2 makes a correction, this is in general correct. The errors that it makes are caused by missing a spelling variant. This score is more than sufficient to be useful for automatic correction of the corpus as it is preferable to have a conservative tool making only those corrections that it is certain about.

We have shown that a systematic statistical analysis of spelling variation is a powerful way to both consolidate known changes in the spelling conventions and to discover new insights in the way people wrote in earlier times.

We also showed that both diachronic linguists and historians wanting to subject historical Portuguese sources to processing operations can have them modernised by an automatic way. They do not have to wait long years, nor to exhaust large human resources, in the operation of manually modernising the variant spellings of such texts, even if they were written by the poor-writer type of author. Additionally, the same procedure can always be adapted to new languages, since the tools we worked with were originally designed for English historical texts.

END NOTES

[1] Acknowledgements: This research is funded by the Portuguese Foundation of Science and Technology (FCT), under the project FLY (PTDC/CLE-LIN/098393/2008), and the FCT program Ciência 2007/2008.

[2] Translated from C. Dauphin, ‘Pour une histoire de la correspondance familiale’, *Romantisme* 90, (1995), 89–99. Cited here at 89.

[3] A. Petrucci, *Public lettering: script, power, and culture* (Chicago, 1993).

[4] Translated from Dauphin, ‘Pour une histoire de la correspondance familiale’, 89.

[5] D. Y. W. Lee, ‘Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle’, *Language Learning & Technology* 5, 3 (2001), 37–72. Cited here at 46 and 50.

[6] DALF, *Guidelines for the description and encoding of Modern correspondence material*, Version 1.0, 2003, http://ctb.kantl.be/project/dalf/.

[7] TEI, *Text Encoding Initiative*, P5 guidelines, http://www.tei-c.org/index.xml, last accessed 24 May 2013.

[8] Recent examples are D. Archer and J. Culpeper, ‘Identifying key sociophilological usage in plays and trial proceedings (1640–1760): An empirical approach via corpus annotation’, *Journal of Historical Pragmatics* 10, 2 (2009), 286–309, and D. Z. Mohd, G. Knowles and Ch. K. Fatt, ‘Nationhood and Malaysian identity: a corpus-based approach’, *Text & Talk – An Interdisciplinary Journal of Language, Discourse & Communication Studies* 30, 3 (2010), 267–287.

[9] I. Hendrickx and R. Marquilhas, ‘From old texts to modern spellings: an experiment in automatic normalisation’, *Journal for Language Technology and Computational Linguistics* 26, 2 (2011), 65–76.

[10] M. F. Gonçalves, *As ideias ortográficas em Portugal: de Madureira Feijó a Gonçalves Viana (1734–1911)* (Lisboa, 2003), 779–786.

[11] F. Coulmas, *The Blackwell encyclopedia of writing systems* (Oxford & Cambridge, Mass., 1996), 380.

[12] Reprinted by I. Castro, I. Duarte and I. Leiria, eds, *A demanda da ortografia portuguesa* (Lisboa, 1987), 152.

[13] Presidência do Conselho de Ministros, ‘Resolução do Conselho de Ministros n.º 8/2011’, *Diário da República*, 1.ª Série, n.º 17, January 25, 2011.

[14] 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’, *Proceedings of the corpus linguistics conference* (Birmingham, 2005).

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

[16] S. Scheible, R. J. Whitt, M. Durrell and P. Bennett, ‘For the A Gold Standard Corpus of Early Modern German’, *Proceedings of the 5th linguistic annotation workshop* (Portland, Oregon, 2011), 124-128.

[17] C. Sánchez-Marco, G. Boleda, J. M. Fontana and J. Domingo, ‘Annotation and representation of a diachronic corpus of Spanish’, *Proceedings of the seventh conference on international language resources and evaluation* (Malta, 2010), 2713–2718.

[18] Gonçalves, *As ideias ortográficas em Portugal*; M. L. C. Buescu, *Gramáticos portugueses do século XVI* (Lisboa, 1978); R. Marquilhas, ‘O acento, o hífen e as consoantes mudas nas Ortografias antigas portuguesas’, in I. Castro, I. Duarte, and I. Leiria, eds., *A demanda da ortografia portuguesa* (Lisboa, 1987), 103–116; M. H. Paiva, ‘Variação e evolução da palavra gráfica: o testemunho dos textos metalinguísticos do século XVI’, in *Actas do XII encontro nacional da Associação Portuguesa de Linguística*, 2 (Coimbra, 1997), 233–252.

[19] T. A. Menegatti, *Regras lingüísticas para o tratamento computacional da variação de grafia e abreviaturas do corpus Tycho Brahe* (Campinas, 2002).

[20] R. Giusti, et al., ‘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 CL2007* (Birmingham, 2007).

[21] BP spelling variants dictionary is available at: http://www.nilc.icmc.usp.br/nilc/projects/hpc/, last accessed 24 May 2013.

[22] The *Corpus Compartilhado Diacrônico* was created by the Laboratório de História do Português Brasileiro from the Universidade Federal do Rio de Janeiro in Brazil. More information can be found at http://www.letras.ufrj.br/laborhistorico/, last accessed 24 May 2013.

[23] M. C. Paixão de Sousa, F. N. Kepler and P. P. F. Faria, ‘E-Dictor: novas perspectivas na codificação e edição de corpora de textos históricos’, in *Caminhos da linguística de corpus* (Campinas, 2010).

[24] DICER is described in chapter 4 of the following thesis: A. Baron, ‘Dealing with spelling variation in Early Modern English texts, PhD dissertation’ (Lancaster University, 2011).

[25] L. F. Lindley Cintra, ‘Observations sur l’orthographe et la langue de quelques textes non littéraires galicien-portugais de la seconde moitié du XIIIe siècle’, *Revue de Linguistique Romane* 27 (1963), 59–77.

[26] P. Teyssier, *História da língua portuguesa* (Lisboa, 1982); I. Castro, *Introdução à História do Português* (Lisboa, 2006).

[27] A. Baron and P. Rayson, ‘VARD 2: A tool for dealing with spelling variation in historical corpora’, in *Proceedings of the postgraduate conference in corpus linguistics* (Birmingham, UK, 2008).

[28] For a detailed description of the Portuguese modules in our version of the VARD2 tool, we refer to the following paper: Hendrickx and Marquilhas, ‘From old texts to modern spellings’, sec 4.

[29] This Lexicon is available for download at: *Multifunctional Computational Lexicon of Contemporary Portuguese*, 2010, http://www.clul.ul.pt/en/resources/88-project-multifunctional-computational-lexicon-of-contemporary-portuguese-r.

[30] Giusti, et al., ‘Automatic detection of spelling variation in historical corpus’, sec 9.

[31] A. Houaiss, et al., *Dicionário Houaiss da língua portuguesa* (Rio de Janeiro, 2001). We wish to thank Mauro Villar for kindly granting us access to the digital form of the Houaiss dictionary’s archaic lexicon.

[32] The Dicer rules were manually converted to the VARD format and some rules were adapted as very general rules such ‘insert e anywhere’ slow down and ultimately crash the VARD program as they generate too many possibilities. To elevate this problem, such general rules were converted to more specific rules.

[33] Baron, ‘Dealing with spelling variation in Early Modern English texts’, sec 6.4, and sec 7.