Automatic detection of spelling variation in historical corpus
An application to build a Brazilian Portuguese spelling variants dictionary

Rafael Giusti, Arnaldo Candido Jr, Marcelo Muniz, Lívia Cucatto, Sandra Aluísio
University of São Paulo, NILC, CP 668,13560-970, São Carlos/SP, Brazil
rg@grad.icmc.usp.br, arnaldoc@icmc.usp.br, marcelo.muniz@gmail.com, liviacucatto@yahoo.com.br, sandra@icmc.usp.br

Abstract

The *Historical Dictionary of Brazilian Portuguese* (HDBP), the first of its kind, is based on a corpus of Brazilian Portuguese (BP) texts from the sixteenth through the eighteenth centuries (and some texts from the beginning of the nineteenth century), being developed under the sponsorship of the Brazilian funding agency CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). It is a three-year project that started in 2006 to fill a gap in Brazilian culture with a dictionary describing the vocabulary of Brazilian Portuguese from the beginning of the country's history. The corpus totals more than 3,000 texts with approximately 7.5 million words. Our working corpus, i.e. the corpus already processed by the corpus processing system UNITEX (http://www-igm.univ-mlv.fr/~unitex/), is coded in Unicode (UTF-16) and totals 1,733 texts, 57.1 MB, and 4.9 million words. A difficulty in dealing with historical corpora to carry out lexicographic tasks is the identification of all spelling variants of a specific word, since spelling variation distorts frequency counts, a usual criterion to select dictionary entries. In our project, another challenge is to select all variants of a dictionary entry that are in the corpus to illustrate the absence of an orthographical system in the aforementioned centuries and to provide example sentences for them. This paper introduces both an approach based on transformation rules to cluster distinct spelling variations around a common form, which is not always the orthographic (or modern) form, and the choices made to build a dictionary of spelling variants of BP based on these clusters. Currently, we have forty-three rules manually developed, which generated 12,189 clusters of spelling variants, totalling 27,199 variants from our working corpus. After a careful analysis of these clusters, we adopted the DELA format to build our dictionary. The BP dictionary of spelling variants enables sophisticated searches in the historical corpus using UNITEX, giving support to build the main dictionary of the HDBP project. Moreover, the variants of a given word can be searched using an application named *Dicionário* we have developed to display dictionaries in DELA format. As we also use Philologic (http://philologic.uchicago.edu/index.php) to support the building of the HDPB, we carried out a comparative evaluation between our approach to cluster distinct spelling variants and AGREP (http://www.tgries.de/agrep/), which is used in Philologic to check for similar or alternative spellings.

1. Introduction

This research is part of the *Historical Dictionary of Brazilian Portuguese* (HDBP) project, in which we have built a Brazilian Portuguese corpus of texts from the sixteenth through the eighteenth centuries. Organizing such a historical dictionary required a comprehensive and time-consuming analysis of documents, published texts, and manuscripts produced by eyewitnesses to the early stages of Brazilian history. One important difficult to compile the corpus derived from the absence of a press in Colonial Brazil, which had a precarious communication system. Only after

1808, when fleeing from Napoleon's army, did the Portuguese monarchy transfer the government of the Portuguese empire to Brazil and improved communications. There are also peculiarities concerning language that had to be considered, such as biodiversity and the multifaceted cultural traditions from different regions of the country. To implement the project, we set up a network of researchers from various regions of Brazil and Portugal, including linguists and computer scientists from eleven universities. This team comprises eighteen PhD researchers, with complementary skills, and twenty-three graduate and undergraduate students.

This project will fill a gap in Brazilian culture with a dictionary describing the vocabulary of Brazilian Portuguese as of the beginning of the country's history. At that time, the Brazilian language was still dependent upon European Portuguese, even though some vocabulary was already forged on this side of the Atlantic. On the one hand, speakers of that early period faced a world materially and culturally different from what was known in Europe; therefore, they needed to designate referents of this new universe that were hitherto unnamed using words of the Portuguese linguistic system. The hundreds of native languages then spoken in Brazil had their own vocabulary for designating elements of the Brazilian fauna and flora, but these words did not exist in European Portuguese. On the other hand, habits and institutions gradually began to form in this new society with infusion of new cultures, resulting in new words that were different from those used in the Portuguese metropolis. A careful analysis of texts about Brazil written by Brazilians, or by Portuguese who were transferred to this country, allows us to explore and unearth the vocabulary repertoire used from the sixteenth through the eighteenth centuries. Once this task of finding and recording such data in dictionary format is completed, the results of this research will be made available to the Brazilian society and scholars of Brazilian studies.

To build the corpus we collected documents in public archives and libraries all over Brazil and also in Portugal. The corpus totals more than 3,000 texts with about 7.5 million words. Our working corpus, i.e. the corpus already processed to work with the corpus processing system UNITEX, totals 1,733 texts, 4.9 million words and 57.1 MB (we are using Unicode (UTF-16)). To process this large corpus, we have faced the typical problems one is likely to encounter when dealing with old documents, starting with text digitalization. J. Rydberg-Cox (2003) and R. Sanderson (2006) state that, in historical texts of Latin, Greek and English, to mention just a few languages, broken words at the end of lines are not always hyphenated; word breaks are not always used; common words and word-endings are abbreviated with non-standard typographical symbols; uncommon typographical symbols are pervasive also in non-abbreviated words; and spelling variation is common even within the same text. We encountered the same problems in the HDBP project.

Particularly, the non-existence of an orthographical system in the aforementioned centuries generated a Babel of graphic systems, used by the many different scribes or copyists who wrote those texts. This problem is also faced by large-scope digitization initiatives, such as Google Search Books[1], which are collecting vast quantities of searchable historical material, making the problem shift from scarcity to abundance, according to Rosenzweig[2]. This new scenario will motivate researchers to apply NLP

[1] http://books.google.com/
[2] http://chnm.gmu.edu/resources/essays/scarcity.php

(natural language processing) tools to historical data (Crane and Jones, 2006). However, some problems will appear in this endeavour. First of all, the large amount of spelling variants of a word makes it difficult to use successfully standard indexing techniques for Information Retrieval (Hauser *et al*., 2007; Ernst-Gerlach and Fuhr, 2006; Braun, 2002) and NLP tasks such as named entity extraction (Crane and Jones, 2006). Second, it is useless to apply corpus annotation tools trained on contemporary language data to historical texts, since they will not deal with spelling variants of the same word (Rayson *et al*., 2007).

More recently, several research projects dealing with English, German, French, and Portuguese, to mention just a few languages, have included the problem of spelling variation in historical corpus on their agendas (Rayson, Archer and Smith, 2005; Archer *et al*., 2006; O'Rourke *et al*., 1996; Hirohashi, 2005). One of them has developed a tool named VARD (VARiant Detector) to detect and normalise automatically variants of English language to their modern form (Rayson, Archer and Smith, 2005; Archer *et al*., 2006, Rayson *et al*., 2007). This solution includes a pre-processor for detecting historical spelling variants and inserting modern equivalents alongside them to avoid the need to retrain each annotation tool that is applied to the corpus. On the other hand, the part-of-speech (POS) tagger[3] developed to annotate the Tycho Brahe corpus[4] added historical variants to the POS tagger lexicon to deal with original (historic/ancient) spellings found in Portuguese texts written by authors born between 1435 and 1835. Later, in the scope of the Tycho Brahe project, researchers developed a methodology to normalise automatically the spelling variants of the corpus (Hirohashi, 2005), which served as a basis for this research. In section 2, we will review these and other approaches, such as the use of AGREP by Philologic to check for the similar or alternative spellings a search query has, and the RSNSR (Rule-Based Search in Text Data Bases with Non-standard Orthography) system (Archer *et al*., 2006), which focuses on finding and highlighting German spelling variation.

Therefore, the aim of this paper is twofold. The first is to introduce an approach based on transformation rules, i.e., letter and string replacement rules to cluster distinct spelling variations around a common form which is not always the orthographic (or modern) form. The second is to explain the format and the choices made to build a BP dictionary of spelling variants based on the generated clusters. Differently from the approaches mentioned above to normalise variants, the purpose of the system presented in this paper, named **Siaconf** (*Sistema de Apoio à Contagem de Freqüência em Corpus*/Support System for Frequency Counting in Corpus), is to support both the detection of spelling variants and the formulation of new transformation rules. Section 3 describes our approach to cluster the spelling variants of a word showing a) their frequency in the corpus, b) the different types of transformation rules, and c) the reports the system generates to give support to the task of evaluating the new transformation rules. In Section 4, we report an experiment to compare Siaconf to AGREP, which is used in Philologic. Section 5 explains the building of the Brazilian Portuguese dictionary of spelling variants. And finally, our conclusions and proposals for future work can be found in Section 6.

3 http://www.ime.usp.br/~tycho/relatorios/2000-2001/00_01.html
4 http://www.ime.usp.br/~tycho/

2. Related works

The VARD tool was developed to detect and normalise spelling variants to their modern equivalents in running text (Rayson, Archer and Smith, 2005; Archer *et al*., 2006, Rayson *et al*., 2007). It focuses on the English language and was trained on sixteenth to nineteenth-century texts. VARD does not make destructive changes in a corpus. Each normalisation is carried out adding up an XML tag that preserves the variant form found in the original corpus. This makes the use of automatic corpus manipulation tools easier, without destroying the historical features of a text. Its current version uses a combination of several linguistic resources manually developed and techniques derived from spelling checkers, and a scoring mechanism to select preferred candidates. The techniques and sources are SoundEx and several edit distance algorithms, a list of 45,805 variant forms and their modern equivalents, a small set of contextual rules in the form of word templates and POS tags, and letter replacement heuristics built with the help of a list of variants and equivalents to reduce the overhead to compile new lists of variants in new corpora. Tools with the same purpose as VARD depend much more on high precision than on high recall. Indeed, a performance comparison of VARD to MS-Word on the Lampeter corpus of Early Modern English Tracts[5] texts demonstrated that VARD is much more effective than MS-Word. VARD's accuracy can be attributed to its manually built regularization table (Rayson, Archer and Smith, 2005).

To normalise automatically a corpus, Hirohashi (2005) proposes a methodology that uses supervised machine learning techniques with indirect effectiveness evaluation. This evaluation is based on the variation of tag precision in a reference corpus. The normalising system is composed of three modules, one for generating transformation rules, one for training the normaliser, and the last one for applying the trained normaliser (Figure 1).

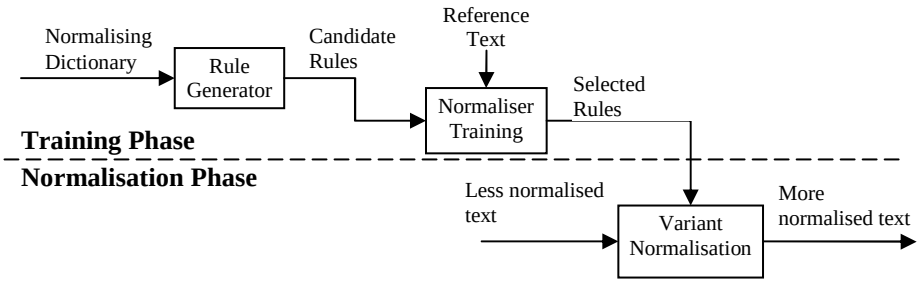

Figure 1. Training and Normalisation in Hirohashi's system (2005).

Hirohashi's methodology is based on transformation rules that normalise words replacing strings of symbols. The first module of the system is a generator of candidate replacement rules through the combination of word substrings from a previously built lexicon, in which word pairs consist of the modern form and its variant. This module generates a large amount of initial rules, but the following module removes most of them. The second module is the training one. It verifies the effectiveness of all candidate rules and selects a subset of such rules through the following process of indirect evaluation:

[5] http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM

1. The corpus is submitted to a process in which the replacement rule is applied to the text. This rule replaces character strings, but does not alter the order or the number of words in the text.
2. The corpus is processed by a POS tagger. Hirohashi used the Tycho Brahe tagger[6] on a text from the Tycho Brahe corpus[7].
3. A post-processor undoes the alterations carried out by the training module using the candidate rule, keeps the tagging generated by the Tycho Brahe tagger, and restores the original text.

The tagged corpus generated by these three steps is, then, compared with a manually tagged reference corpus. Since the Tycho Brahe tagger is spelling-sensitive, a better tagging efficiency means that the candidate rule is highly effective. On the other hand, a worse tagging efficiency means the candidate rule is little effective. Each candidate rule is considered individually. Their high or low effectiveness does not affect the evaluation of the following rules. The corpus used by the training module in step 1 is always the original corpus. The last module is the normalising module, which carries out the normalisation of text itself, applying the set of rules previously selected in the corpus. The order for applying the rules depends solely on the order in which the candidate rules were built based on the lexicon. This can pose a problem, since the final result of normalisation depends largely on the order of rule application, which is not trained by the system. The system evaluation with one of the texts from the Tycho Brahe corpus selected seventeen candidate rules, whereas, for example, the word pair[8] "elegância" "elegancia", which was in the Normalising Dictionary (see Figure 1), has alone generated eighty-four rules.

The project RSNSR (Archer *et al*., 2006; Ernst-Gerlach and Fuhr, 2006) is an interdisciplinary attempt to support the conservation of cultural heritage, whose aim is to provide means to perform a reliable full-text search in documents written before the German unification of orthography in 1901. In this project, researchers developed a search engine to retrieve historical documents for experts and interested users who are not language experts. RSNSR uses a rule-based fuzzy search engine to retrieve text data independently of its orthographical realization. The rules adopted derive from several sources, such as statistical analyses, historical material, and linguistic principles. As the web-based system focuses on finding and highlighting historical spellings, its demand for recall is much more important than precision. Archer *et al*. (2006) describe the automatic process to generate the transformation rules in this system. The process is similar to Hirohashi's approach, since it is also based on a dictionary of word and variant. However, RSNSR uses triplets (contemporary word form, historic variant form, and frequency of spelling variant in the corpus). When the process for generating automatic transformation rules was compared with two other approaches (manual rules and variant graph) (Ernst-Gerlach and Fuhr, 2006), the automatic rules achieved a slightly inferior frequency, but a recall that is nineteen percent better than the variant graph.

[6] http://www.ime.usp.br/~hiro/normatizador.tar.gz
[7] http://www.ime.usp.br/~tycho/
[8] The word pairs in the dictionary have this form: word "x" with modern spelling followed by word "x" with old spelling.

Agrep (approximate grep) is a fuzzy string searching program, developed by Udi Manber and Sun Wu (1992). Agrep selects the best-suited algorithm[9] for a query from a variety of well-known fastest string searching algorithms, including Manber and Wu's bitap algorithm, based on Levenshtein distances. The others are mgrep, amonkey, mmonkey. Agrep is used in Philologic similarity searchers to check for similar or alternative spellings for a query, given a collection of texts. Figure 2 shows the results of a search with the word "giboia" (a kind of snake) in our working corpus that returned five matches: four of them are real variants; one is the plural form (giboias).

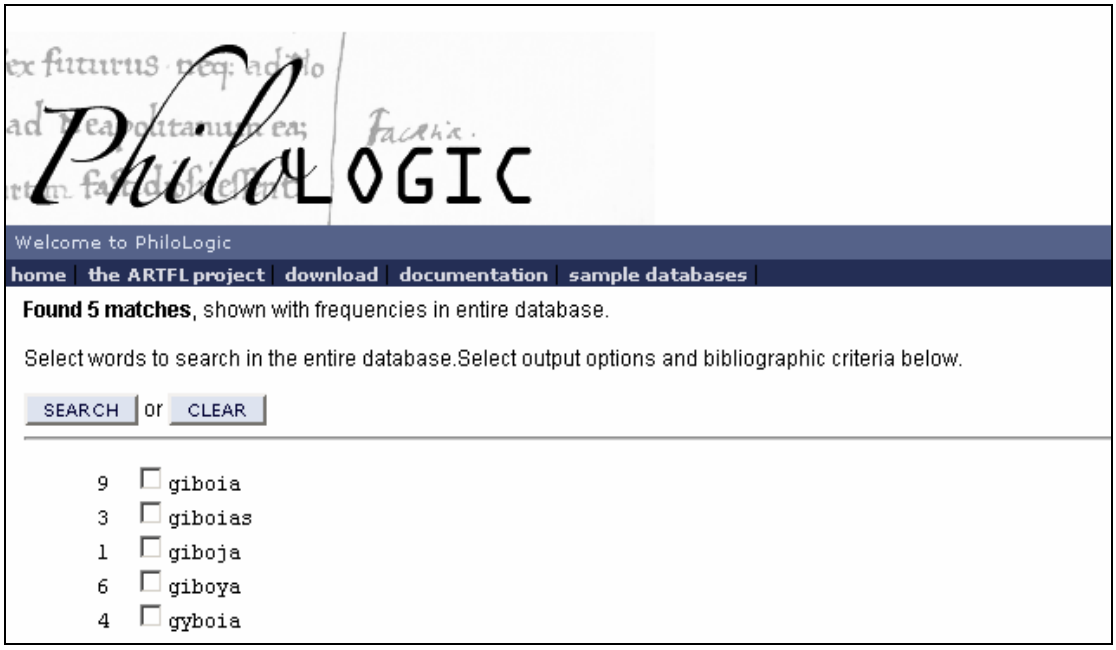

Figure 2. Search results using Philologic similarity searchers.

3. Our approach: a process to support the spelling variation detection

Our approach consists in applying a series of transformation rules, with the same format as those proposed in Hirohashi (2005), to a list of single words from a corpus aiming at grouping different spellings around a common spelling. Grouping spelling as described, the system that implements this approach is able to establish a relation between different spellings. It is expected that this relation shows spelling variation for any given word.

The system we developed is named Siaconf (Sistema de Apoio à Contagem de Freqüência em Corpus/Support System for Frequency Counting in Corpus) and was built in PERL[10]. It is freely available[11], although its documentation is only in Portuguese. This system processes a corpus from an initial list of rules, built by diachronic linguistics or by an expert who bases his work on diachronic linguistics, and makes available three main types of detailed reports:

a) groupings/clusters including spelling variants of the same word;
b) information on the rules applied; and

[9] http://www.tgries.de/agrep/#ALGORITHMS
[10] http://www.perl.com/
[11] http://moodle.icmc.usp.br/dhpb/siaconf.tar.gz

c) a list of non-processed words.

The grouping used in our research is different from the normalisation approaches mentioned in Section 2 ((Hirohashi, 2005) and the VARD tool), because we are not trying to find the orthographic equivalent of a variant that belongs to the corpus, although this happens in most of the cases. For instance, the words “chaõ” and “chaão” (variants of floor) are grouped around the spelling “xam”, which currently does not exist in Brazilian Portuguese (see Figure 3). On the contrary, our aim is that the groupings reduce the impact of spelling variation on the frequency count and that the content of groupings allow for a study of spelling variation in the compiled *corpora*.

```
chão,xam
chaõ,xam
xão,xam
cham,xam
chaão,xam
xam,xam
```

Figure 3. Example of cluster in Siaconf, which groups six words of the working corpus (first strings of entries above) with the word xam (second string of entries above).

Initially, we define a set of rules that are applied to the corpus. Based on the analysis of the three detailed reports, and especially of the *list of non-processed words* (item c), it is possible to devise new transformation rules. The detailed reports help the expert to understand the groupings generated by the rules, to check for mistakes made by the system, and also to find out the cases that are not covered by the rules adopted. The rule generation is iterative and, at the end of the process, it is possible to build a dictionary of spelling variants. Figure 4 illustrates this process.

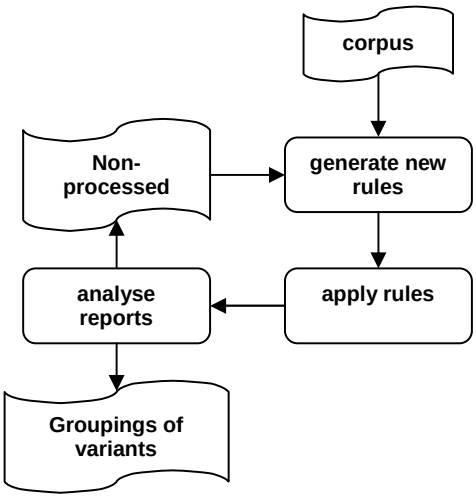

Figure 4. Iterative process for detecting spelling variants in a given historical corpus.

Figure 5 shows four examples of clusters resulting from the application of Siaconf to our working corpus. The report of groupings/clusters (item a) shows the word/spelling that groups actual examples from the corpus and the total frequency for the cluster. For instance, the cluster “apelido” (nickname) has 90 examples of actual words from the corpus, as discriminated in the table below, together with their individual

frequency. Observe that, in this case, “apelido” is also a word from the corpus and a word currently spoken in PB, differently from the example in Figure 3.

```
apelido  (90)                    nam  (37,100)
      appellido         (48)           não          (33,684)
      apelido           (30)           naõ          (2,652)
      appelido          (7)            nam          (439)
      apellido          (5)            nao          (325)

mais  (23053)                    vila  (5,218)
      mais         (22,918)            villa        (4,073)
      majs         (67)                vila         (1,113)
      maes         (38)                vyla         (13)
      mays         (30)                vjlla        (9)
                                       vylla        (9)
                                       vjla         (1)
```

Figure 5. Examples of spelling variations in “apelido” (nickname), “mais” (more), “não” (not), and “vila” (village), in the report of groupings.

In the following sections we present different parts of our research: in Section 3.1, the details of the building of the DHPB corpus; in Section 3.2, the format of transformation rules used in Siaconf; in Section 3.3, the six types of adopted rules to subclassify the forty-three initial rules; and in Section 3.4, the process to develop a new rule, using the system’s resources.

3.1 Working corpus

The current version of the DHPB corpus is composed of 1,733 texts, written by Brazilian authors or Portuguese authors who have lived in Brazil for a long time. The texts selected for our corpus include, for instance, Jesuit missionaries’ letters, documents of the *bandeirantes* (members of the exploratory expeditions which pushed the Brazilian borders far into inland areas), reports of *sertanistas* (explorers of Northeastern Brazil), and documents of the Inquisition. All documents were collected in their original versions or in digital format (PDF files composed of images). During the stage of corpus compilation, the texts are digitalized and pre-processed according to the flowchart shown in Figure 6.

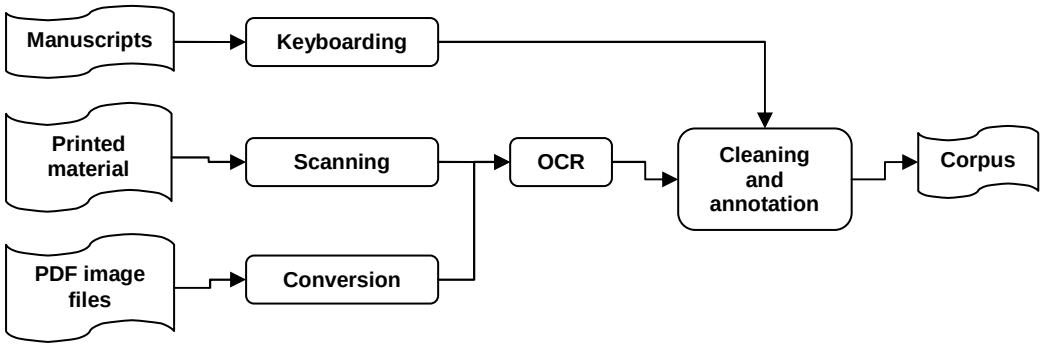

Figure 6. Stages of compilation of the DHPB corpus.

Manuscripts are manually keyboarded, whereas original printed documents are processed by OCR (Optical Character Recognition), and PDF files are converted into

TIFF files before OCR. Texts are coded in Unicode UTF-16, which makes possible to preserve symbols commonly found in Brazilian historical texts that fell into disuse over time, such as the symbol “long s” (ſ). Next, the texts are submitted to a semiautomatic cleaning and annotation process. The cleaning consists of removing from the texts undesired parts such as headers, footers, and line numbers. The annotation is made on text metadata, such as author’s name, page numbering, and document’s title. Then, the corpus is ready to be used in corpus processors such as UNITEX (Paumier, 2006) and Philologic. Table 1 shows details about the composition of our corpus.

Data	**Centuries**			
	16th	**17th**	**18th**	**19th**
Texts	11.16%	27.64%	52.06%	9.13%
Sentences (approx.)	28.99%	15.94%	43.17%	11.90%
Words	18.68%	20.67%	47.68%	12.98%

Table 1: Distribution of texts by century.

3.2. The format of transformation rules

The transformation rules adopted in our approach use regular expressions[12]. A transformation rule is a triplet (*C1 C2 S*), where *C1* and *C2* are regular expressions and *S* is a string. *C1* determines the rule’s coverage criterion, i.e., the forms W_i of the corpus will be processed by the rule. *C2* determines a substring in each *Wi*, which will be replaced by *S*. For example, the rule “(e[ao] e ei)” is applied as follows:

1. *C1* is tested against every form of the corpus and restricts the rule application those that contain the substring “ea” or the substring “eo”, for example, “aldea” (variant of small village).
2. *C2* determines the substring that will be replaced, for example, in the letter “e” in “aldea”.
3. *S* determines the replacement string (“ei”), used to generate the new form, for instance, “aldeia” (small village).

After applying the different rules, several spellings G_i result in a new spelling *H*. Thus, it is possible to infer that the spellings G_i are variants of the same word. For instance, the rules (ll, ll, l) and (y y i) can be applied, respectively, to the spellings “vyla” and “villa”, resulting in a new spelling “vila”. Therefore, they have a great probability of being variants of the same word. In addition, more than one rule can be applied to a given spelling, as shown in Figure 7.

Words	Rules applied	Spellings generated
CHAÕ	ch ch x aõ aõ ão [^r][aã]o$ [aã]o am	"xaõ" "xão" "xam"
CHAÃO	ch ch x aã aã ã [^r][aã]o$ [aã]o am	"xão" "xaão" "xam"

Figure 7. Grouping of CHAÕ and CHAÃO (variants of floor) around spelling XAM.

[12] http://www.regular-expressions.info/

During this process, all rules are applied against all single forms in the corpus, generating a set of new spellings H_i. Each new spelling represents a grouping of spelling variations. It is worth mentioning that the spellings H_i are not orthographic, i.e., the results from the process described are not necessarily the normalised versions of a word.

3.3. Rules and groupings used

Currently, we are using a total of forty-three transformation rules, described in the following sections. After applying them in our working corpus, we identified 27,199 spelling variants, in a total of 12,189 word groupings.

3.3.1. Rules for spellings that fell into disuse

Brazilian Portuguese have abandoned some letters and digraphs over time, as the letter "y", which was replaced by "i". Four rules come within this scope. They are:

1. Replacement of "y" by "i" in every context.
2. Replacement of "ph" by "f" in every context.
3. Replacement of the grave accent (`) by the acute accent (´) over vowel "o" (ò → ó).
4. Replacement of "th" by "t".

3.3.2. Rules for double consonants

Menegatti (2002) examines the occurrence of double Latin occlusive and fricative consonants, often used to represent the stressed syllable by authors who did not adopt the traditional diacritic marks for this purpose. Such double consonants can be removed and replaced by a single instance of the same letter. The rule below illustrates how to manage double consonants:

- ff → f in every context.

Similarly, we developed rules for cases of pp, tt, cc, bb, dd, gg, vv, uu, and zz. In addition, based on the analysis of the double consonants in the corpus, we set rules for dealing with mm, nn, and ll, totalling thirteen rules to be applied to doubles consonants.

3.3.3. Rules generated according to the orthographic norm

It is impossible to apply automatically several of the modern orthographic norms without knowing the "misspelled" word. Many orthographic norms depend, for instance, on the stress of a word – which cannot be inferred without understanding the semantics of the spelling under study. Even so, many of them provide invaluable rules, such as that which establishes the use of "m" or "n" before consonants. Six rules come within this context:

1. Replacement of "m" by "n" when followed by consonants other than "p" or "b".
2. Replacement of "n" by "m" when followed by "p" or "b".

3. Replacement of “aã” (spelling that indicates a nasal sound) by “ã”.
4. Replacement of “aõ” by “ão”.
5. Replacement of the grave accent over “a” (à) by the acute (á), except when it is at the beginning of a word.
6. Replacement of suffix “aes” (used in some historical texts to indicate plural) by “ais”.

3.3.4. Rules based on frequency

Some rules were developed with the sole purpose of grouping spellings, with no intention of transforming them in modern spellings. All rules shown in this subsection derive from Menegatti (2002) and were validated in our working corpus.

1. Replace “chr” by “cr” in every context.
2. Replace “ch” by “x” in every context. In Portuguese, the digraph “ch” and the consonant “x” have the same sound. Although “ch” is still used, there are few words or no word in our corpora that are differentiated by these symbols (e.g.: “chá” (tea) and “xá”).
3. Replace “ee” by “é” in every context.
4. Replace “he” by “é” in every context.
5. Remove the mute “p” from the consonant cluster “pt”. Although this may cause undesired groupings (such as “apto” and “ato”), the analysis of reports has been showing that it is a beneficial rule.
6. Replace the consonant cluster “mpt” by “nt” (for example, “redemptor” and “redentor” (redeemer).
7. Remove the mute “c” in the consonant cluster “ct”.
8. Replace “v” by “u” when it is the last letter of a word (e.g.: “rev” → “reu”).
9. Replace the consonant “j” by the vowel “i” when preceded by another consonant.
10. Mark with an accent the first “i” in the suffixes “issimo”, “issima”, “issimos”, and “issimas”.
11. Replace the consonant cluster “mn” by consonant “n”.
12. Add the tilde to nasalize the suffix “oens”.
13. Replace “z” by “s” in the suffix “ozo”.
14. Replace the nasal suffixes “ao” and “ão” by “am” (“tão”, “são”), except when preceded by “r”, which can mean verb inflection (“saberão”, “sairão”). This rule aims at “denormalising” normalised forms to group more common or diversified non-normalised forms, such as “saõ", “saão”, “são” and “sam”.

3.3.5 Lexicalised rules

These are rules developed for specific words, which are not grouped, by any general rule, in spite of being very frequent in the corpus. The only lexicalised rule used was:

- Replace “o” by “u” in the suffix “deos” (“deus” (Christ) and “judeus” (Jewish)).

3.3.6 Automatic rules

In Hirohashi (2005), transformation rules are generated automatically from the Tycho Brahe corpus. Some of these rules are reused in our research, since they proved very efficient in the grouping task.

1. Mark with an accent the "a" in suffix "agio" (for example, "sufrágio" (suffrage)).
2. Replace "z" by "s" in the infix "preciz" (for example, "precisa", "precisando" (conjugated form of need)).
3. Replace "ss" by "ç" in the infix "serviss" (for example, "serviço" (service)).
4. Replace "z" by "s" in the infix "zente" (for example, "presente" (gift/present)).
5. Replace "c" by "ss" in the suffix "acem" (for example, "tirassem" (conjugated form of take)).

3.4 Development of a new rule

The report of non-processed words generated by Siaconf is useful to develop new rules. In this report, it is possible to find words with a high frequency in the corpus that are not grouped by any rule. Based on the analysis of this report, it is possible to find words such as "hum" ("um") (an/a) and "huma" ("uma") (an/a), and formulate rules to treat these cases, as follows:

- Remove "h" from prefix "hum". This rule can be written as "hum hum um".

When a rule is included in the system, it is convenient to check which words it will affect so as to ensure its precision. The report of applied rules allows us to check which words are affected by a given rule (Figure 8).

y y i		
	Daly ->	Dali (From there)
	Despoyes ->	Despoies (After)
	Houtrosy ->	Houtrosi (Also, Likewise)
	Muyto ->	Muito (Many, very)
	Outrosy ->	Outrosi (Also, Likewise)
	Pydimos ->	Pidimos (Conjugated form of verb ask)
	Prymeyramente ->	Primeiramente (Primarily)
	Primeyramente ->	Primeiramente (Primarily)
	foy ->	Foi (Conjugated form of verb go)
	ygreja ->	Igreja (Church)

Figure 8. Excerpt from the report of the application of rule y y i.

4. Evaluation

The evaluation of Siaconf by Agrep used in Philologic was not comprehensive. The performance measurements have been performed in two parts. In the first part, twenty-three words were chosen randomly in Siaconf's report of groupings/clusters, one for each letter of the Portuguese alphabet (except for X) plus K. These words

were applied to Philologic and the result was analysed in both systems to check whether they were real variants or not. We used the comparative recall, a measure employed in systems of Information Retrieval. Table 2 shows precision values and the comparative recall for this experiment. As it was expected, Siaconf's precision is the highest possible (one-hundred percent); however, concerning recall, Agrep's performance is better. Siaconf's recall can be improved with the development of new rules. We also intend to use both sources of groupings to build a dictionary of spelling variants, using words from the DHPB.

Technique	True positive	False positive	Precision	Comparative recall
Transformation rules (Siaconf)	36	0	100%	72%
Edition distance (AGREP)	41	196	20.92%	84%

Table 2. Precision and comparative recall in a comparative evaluation of performance of AGREP and Siaconf.

The second part of the performance measurements was based on five Siaconf clusters, which have the greatest frequency counts. By inspection on the Siaconf reports, we already knew that the most frequent words were very short ones. Therefore, we would like to evaluate the effect they have on AGREP, used in Philologic. Table 3 shows, in the first five rows, the five clusters that have the highest frequency counts. As we have customized the similarity search in Philologic to work with words with length greater than two, two words of those clusters (words "e" (and) and "em" (in)) had to be changed by others that are presented in the last two rows of Table 3. The last column of Table 3 shows the normalised words of the chosen clusters, used in Philologic similarity search.

Clusters	Frequency	Normalised words, used in Philologic similarity search
e	223549	
que	183917	que (that)
em	58147	
com	54617	com (with)
nam	44901	não (not)
mais	30394	mais (more)
seu	14941	seu (your)

Table 3. Siaconf clusters which have the greatest frequency counts

Table 4 shows the Precision and Comparative Recall results in the comparative evaluation of performance of AGREP and Siaconf for a few short words. In AGREP, these words present more false positives and more false negatives than those results of evaluation 1, causing AGREP precision to fall down and AGREP comparative recall to rise. In Siaconf, these short words cause the comparative recall value to fall down, since AGREP brings more true positives than Siaconf. Checking the results that

Philologic returns is a very tiring task for linguists, since the number of false positives is very high. However, AGREP has shown to be very useful to detect joint words and OCR errors.

Technique	True Positives	False Positives	Precision	Comparative Recall
Transformation rules (Siaconf)	7	2	77.77%	23.33%
Edition distance (AGREP)	27	217	11.06%	90%

Table 4. Precision and Comparative Recall results for the five words of the last column of Table 3

5. Building a Brazilian Portuguese dictionary of spelling variants

In recent years, NLP researchers were focusing their studies on standardization during the construction of linguistic resources. Such studies led to the achievement of the international standards and tools we use today. One of these standards, DELA (Dictionnaires électroniques du LADL), was developed at LADL (Laboratoire d'informatique documentaire et linguistique, University of Paris 7, France), using the corpus-processing tool INTEX (Silberztein, 2000). DELA became the standard for electronic lexicons in the research network Relex[13]. These lexicons were used with INTEX, and now are used with its open-source counterpart, UNITEX. This format allows the declaring of simple and compound lexical entries, which can be associated with grammatical information and inflection rules. These dictionaries are linguistic resources specifically designed for automatic text processing operations.

Variations of DELA include DELAF – which comprises inflected single words –, DELAC and DELACF – for non-inflected and inflected compound words, respectively. The dictionaries of single words (DELAS and DELAF) are simple lists of words associated with grammatical and inflection information. The grammatical information is mainly morphological and corresponds to gender, number, degree, case, mood, tense, and person. However, the format allows for a gradual inclusion of syntactic and semantic information (Ranchhod, 2001). The lexical entries in DELAF have the following general structure:

(*Inflected word*),(*canonical form*).(*part of speech*)[+(subcategory)]:*morphological behaviour*

5.1 Customising UNITEX to deal with the Historical Portuguese Corpus

Processing a corpus for lexicographical tasks is made easier if computational lexicons are available, that was the reason why we adopted UNITEX in the HDPB project. UNITEX supports several languages, including Portuguese. Language-specific resources are grouped in packets referred to as idioms. A lexicon for contemporary Brazilian Portuguese was included in the construction of UNITEX-PB (Muniz *at al.*, 2005).

[13] http://infolingu.univ-mlv.fr/Relex/Relex.html

However, due to the peculiarities of historical texts, several changes had to be made and a new idiom was created, named “Historical Portuguese (Brazil)”. Among the changes made, we included characters that are no longer used in Portuguese, such as the long s (ſ) and the tilde (~). Diacritic marks differ from the common ones in that they may be used with consonants. For instance, an accentuated “m̃” was common in Historic Portuguese. Using Unicode when the text was being compiled made possible to include such characters. Table 5 shows the generic symbols and diacritic marks used.

Symbol	Description	Unicode	Sample
^	combining circumflex accent	0302	quarŷ (*)
~	combining tilde	0303	comm̃andante (commander)
¯	combining macron	0304	cacaō (cocoa beans)
¨	combining dieresis	0308	muÿ (much, many)
̉	combining hook above	0309	sȯmente (only)
°	Combining ring above	030A	Å (abbreviation of Afonso)
̓	Combining comma above	0313	tinhaȯ
Æ	Latin capital letter AE	00C6	Æthyopia (**)
æ	Latin small letter ae	00E6	grati (**)
Œ	Latin small ligature oe	0153	Cœteris (**)
§	section sign	00A7	§ (denotes paragraph mark)
∃	turned capital f	2132	
ſ	Latin small letter long s	017f	Deſcobrio (find)
ƒ	Latin small letter f with hook	0192	
∃	Latin small letter turned e	01DD	
ɐ	Latin small letter turned a	0250	

(*) Indian name (**) Latin name

Table 5: Characters found in historical texts.

5.2 Entry samples from the dictionary of spelling variants in DELA format

Figure 9 shows DELA entries that correspond to variants of “apelido” (nickname).

```
Appellidos,apelidos.N+VAR:ms/50.0%
apelidos,apelidos.N+VAR:ms/36.36%
appelidos,apelidos.N+VAR:ms/9.09%
apellidos,apelidos.N+VAR:ms/4.54%
```

Figure 9: Examples of entries in DELA format.

In each entry, we have: variant, new spelling generated by Siaconf, class of word, its semantic attributes, information on inflection, and frequency of variant in the corpus. As the whole process is automatic, all entries are masculine-singular (MS) nouns (N). A manual revision will be carried out later to insert grammatical and inflection data.

A possible change is to insert the lemmatised form of the spelling in the proposed structure. Searches based on the lemmatised form are particularly useful for verbs in

Portuguese, since they have a great number of inflections. The lemmatised form can be inserted in the place of the spelling generated by Siaconf:

appellidos,apelido.N+VAR:ms/50.0%

An alternative is to insert the normalised form as a semantic attribute:

appellidos,apelidos.N+VAR+apelido:ms/50.0%

The first strategy is faithful to the semantics of the DELA format. As for the second strategy, it also preserves the form generated by Siaconf.

The DELA dictionary of spelling variants is converted to the binary format used by UNITEX. This conversion optimises the time spent searching the corpus and the dictionary. The entries in the dictionary of variants can be looked up with the help of a tool named *Dicionário*, developed by one of the authors of this article. Figure 10 shows the search for variants of the word “apelido”. It is important to observe that the variant and the new spelling are inverted when compared to Figure 9, because of the indexing used in UNITEX.

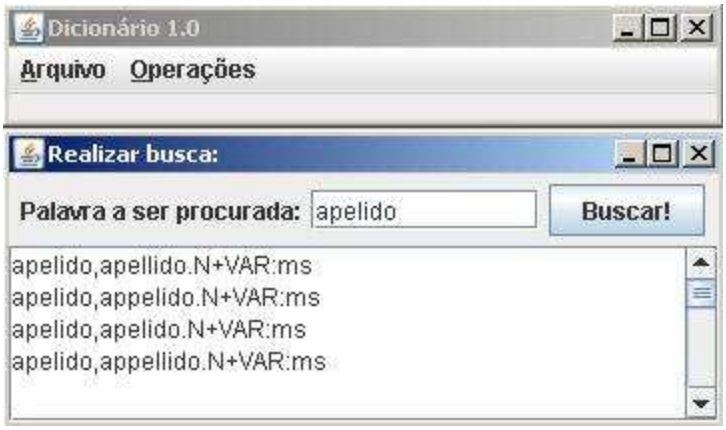

Figure 10: Search for variants using *Dicionário*.

We can see that, in Figure 10, there is no information about the frequency of each variant in the corpus. This is due to the fact that the frequency of variants is inserted as comments in the DELA format, and comments are removed when DELA is converted to the binary format. Possible solutions for this problem are to discretise frequencies and convert them to DELA attributes. For instance, it is possible to define the values “very rare”, “little frequent”, “frequent” for the intervals 0-9 percent, 11-49 percent, 50-100 percent, respectively. An example of entry with discretised frequency is:

appellidos,apelidos.N+VAR+Freqüente:ms

Discretisation is necessary because of the form UNITEX employs to search for semantic attributes. The spelling dictionary can also be used together with UNITEX for searching the corpus. Figure 11 shows the result of the search for the expression “<apelido.VAR>”, which returns all variants of “apelido”.

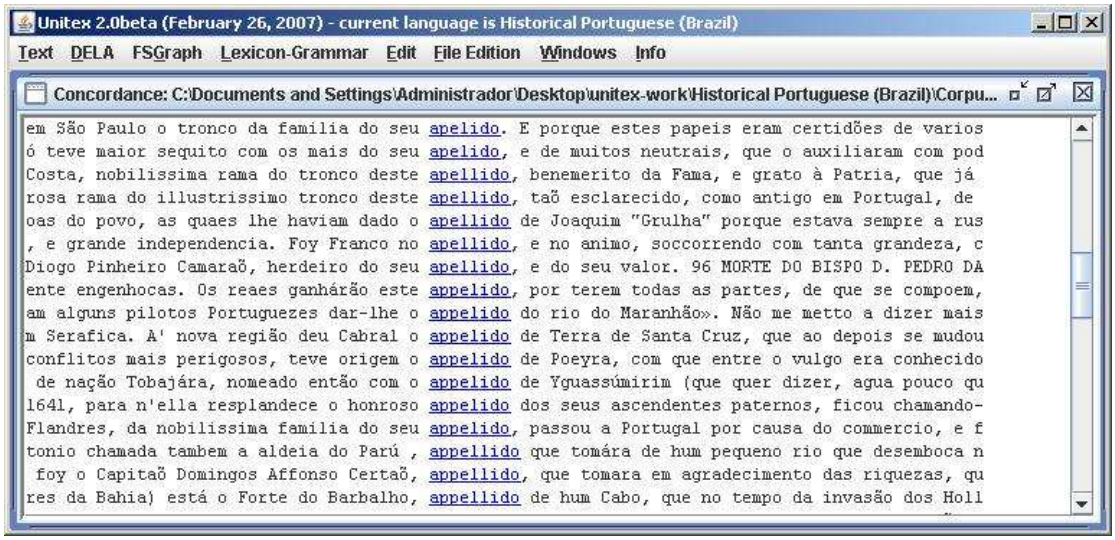

Figure 11: Search in the corpus with the aid of the dictionary of spelling variants.

The variants dictionary built in this investigation is freely available[14] to be used in other research on texts in historical Portuguese.

6. Conclusions and future work

In this paper, we described a methodology to detect spelling variants in historical texts written in Portuguese. Using this methodology, we built a dictionary of spelling variants, which is freely available, and a system for detecting spelling variations automatically. Our dictionary of spelling variants was evaluated for lexicographers, who reported some cases of variants not covered by the transformation rules, as was expected, given that Siaconf has a high precision, but a not so high recall. However, developing more transformation rules and building a dictionary that includes the results from AGREP can solve this problem. As a matter of fact, we are already formulating new rules. Based on the analysis of the corpus and the reports, it was possible to develop new phonetic rules. They were not tested against the words in the corpus yet, but we expect them to increase recall in the detection of spelling variations, with little interference in precision. These rules are: a) replacement of “ſ” (long s) by “s”; b) replacement of “g” followed by “e” and “i”; c) removal of accents (for example, “á”, “à”, and “é”); d) removal of “u” in infixes “gua” and “guo”; e) removal of “h” preceded by “d”; f) removal of mute “h” at the beginning of words. Although this last rule is contrary to the current orthographic norm for Portuguese, the evaluation of the reports shows that it produces good groupings, such as “helle”, “hele”, “elle”, and “ele” (he). We consider the approach that makes use of transformation rules as an efficient way to detect spelling variations in historical corpora, since, with just forty-three rules, we detected almost 30,000 variants in a corpus of 4.9 million words, and the number of mistakes made by the system was minimal. In addition, we have observed that, in a large number of cases, the system generates normalised versions of the words.

[14] http://moodle.icmc.usp.br/dhpb/spelling-variants.gz

References

Archer, D., A. Ernst-Gerlach, S. Kempken, T. Pilz and P. Rayson (2006) The identification of spelling variants in English and German historical texts: manual or automatic? In E. Vanhoutte et al. (eds.) Proceedings abstracts of Digital Humanities 2006, 3–5. Paris: Sorbonne.

Braun, L. (2002) Information retrieval from Dutch historical corpora. Master's thesis. Maastricht, Nederlands: Maastricht University, 2002. Available on-line at http://www.nici.ru.nl/~idak/publications/papers/scripties/scriptiebraun.pdf (accessed: 22 june 2007).

Crane, G. and A. Jones (2006) The challenge of Virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection, In G. Marchionini et al. (ed.) Proceedings of 6th ACM/IEEE-CS joint conference on Digital libraries, pp. 31-40. Chapel Hill, USA: ACM Press.

Ernst-Gerlach, A. and N. Fuhr (2006) Generating Search Term Variants for Text Collections with Historic Spellings, M. Lalmas et al. (ed.) Proceedings of 28th European Conference on Information Retrieval Research (ECIR 2006), pp. 49-60. London: Springer-Verlag.

Ernst-Gerlach, A. and T. Pilz (2006) Search methods for documents in non-standard spelling. Talk presented at Historical Text Mining Workshop, Lancaster University, UK. Available on-line at http://ucrel.lancs.ac.uk/events/htm06/PilzErnstGerlachHTM06.pdf (accessed 22 june 2007).

Hauser, A., M. Heller, E. Leiss, K. U. Schulz and C. Wanzeck (2007) Information Access to Historical Documents from the Early New High German Period, In C. Knoblock et al. (eds.) Proceedings of IJCAI-07 Workshop on Analytics for Noisy Unstructured Text Data (AND-07), pp. 147-154. Hyderabad, India. Available on-line at http://research.ihost.com/and2007/cd/Proceedings_files/p147.pdf (accessed: 22 june 2007).

Hirohashi, A. (2005) Aprendizado de regras de substituição para normatização de textos históricos. Master's thesis. IME: Universidade de São Paulo, Brasil. (In Portuguese)

Koolen, M., F. Adriaans, J. Kamps and M. Rijke (2006) A Cross-Language Approach to Historic Document Retrieval, In Lalmas et al. (ed.) Proceedings of 28th European Conference on Information Retrieval Research (ECIR 2006), pp. 407-419. London: Springer-Verlag.

Menegatti, T. A. (2002) "Regras Lingüísticas para Tratamento Computacional da Variação de Grafia e Abreviaturas do Corpus Tycho Brahe". Technical Report. Campinas, Brasil: Unicamp. (In Portuguese)

Muniz, M. C. M., M. G. V. Nunes and E. Laporte (2005) UNITEX-PB: a set of flexible language resources for Brazilian, In Rove Chishman (Ed.) Proceedings of III Workshop em Tecnologia da Informação e da Linguagem Humana (TIL), pp. 2059-2068. São Leopoldo, Brazil: Universidade do Vale do Rio dos Sinos.

O'Rourke, A. J., A. M. Robertson, P. Willett, P. Eley and P. Simons, (1996) Word variant identification in Old French. Information Research 2, no. 4.

Paumier, S. (2006) Manuel d'utilisation du logiciel UNITEX. IGM, Université Marne-la-Vallée. Available on-line http://www-igm.univ-mlv.fr/~unitex/ManuelUnitex.pdf (accessed 22 june 2007).

Ranchhod, E. M. (2001) O uso de dicionários e de autômatos finitos na representação lexical das línguas naturais, In E. Ranchhod (ed.) Proceedings of Tratamento das Línguas por Computador: uma Introdução à Linguística Computacional e suas Aplicações, pp. 13-47. Lisbon: Caminho.

Rayson, P., D. Archer and N. Smith (2005) VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historic corpora, In S. Hunston et al. (eds.) Proceedings of Corpus Linguistics 2005, vol. 1, no. 1. Birmingham: Birmingham University.

Rayson, P., D. Archer, A. Baron and N. Smith (2006) Tagging historical corpora: the problem of spelling variation, In L. Burnard et al. (eds.) Proceedings of Digital Historical Corpora - Architecture, Annotation, and Retrieval, no. 6491. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI).

Rydberg-Cox, J. A. (2003) Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries, In G. Henry et al. (eds.) Joint Conference on Digital Libraries (JCDL 2003), 372-373. Houston, USA: ACM Press.

Sanderson, R. (2006) "Historical Text Mining", Historical "Text Mining" and "Historical Text" Mining: Challenges and Opportunities, Talk presented at Historical Text Mining Workshop, Lancaster University, UK. Available on-line at http://ucrel.lancs.ac.uk/events/htm06/RobSandersonHTM06.pdf (accessed 22 june 2007).

Silberztein, M. (2000) Intex: a FST toolbox. Theoretical Computer Science 231, 33–46.

Sousa, M. C. P. and T. Trippel (2006) Building a historical corpus for Classical Portuguese: some technological aspects, In P. Baroni et al. (eds.) Proceedings of V International Conference on Language Resources and Evaluation (LREC 2006), pp. 1831-1836. Genova: ELRA.

Wu, S. and U. Manber (1992) Fast Text Searching Allowing Errors. Communications of the ACM 35, no. 10, pp. 83-91. New York, USA: ACM Press.