Building a Corpus-based Historical Portuguese Dictionary: Challenges and Opportunities

Arnaldo Candido Junior* — Sandra Maria Aluísio*

** Center of Computational Linguistics (NILC)/ Department of Computer Sciences, University of São Paulo, Av. Trabalhador São-Carlense, 400, 13560-970 - São Carlos/SP, Brazil*

arnaldoc@ icmc. usp.br

sandra@ icmc.usp.br

ABSTRACT: Historical corpora are important resources for different areas. Philology, Human Language Technology, Literary Studies, History, and Lexicography are some that benefit from them. However, compiling historical corpora is different from compiling contemporary corpora. Corpus designers have to deal with several characteristics inherent in historical texts, such as: absence of a spelling standard, pervasive use of abbreviations plus their spelling variations, lack of space between words, irregular use of hyphenation, non-standard typographical symbols. This paper addresses the challenges posed in processing the corpus designed for the Historical Dictionary of Brazilian Portuguese (HDBP) project, which is composed of texts from the sixteenth through the beginning of the nineteenth century, and the solutions found to support the compilation of a Historical Portuguese dictionary based on this corpus.

RÉSUMÉ: Les corpus historiques sont des ressources importantes pour différents domaines: a Philologie, la Technologie du Langage Humain, les Études Littéraires, l'Histoire et la Lexicographie en tirent profit. Toutefois, la compilation des corpus historiques est différente de la compilation des corpus contemporains. Les concepteurs de corpus doivent faire face à des problèmes inhérents aux textes historiques, tels que: l'absence d'une norme orthographique, l'utilisation généralisée des abréviations en plus de leurs variantes orthographiques, le manque d'espace entre les mots, l'utilisation irrégulière des traits d'union, les symbols typographiques non standard. Ce document aborde les défis posés dans le traitement des corpus conçus po ur le Dictionnaire Historique du Portugais Brésilien (DHPB), qui est composé de textes du XVIe jusqu'au début du XIXe siècle, et les solutions trouvées pour appuyer la compilation d'um dictionnaire du portugais historique basé sur ce corpus.

KEY WORDS: historical corpora, corpora processing, historical dictionaries, Brazilian history

MOTS CLÉS: corpus historique, traitement de corpus, dictionnaires historiques, histoire du Brésil.

TAL. Volume 50 – n° 2/2009, pages 73 à 102

1. Introduction

The *Historical Dictionary of Brazilian Portuguese* project (HDBP) (Giusti *et al.*, 2007; Vale *et al.*, 2008; Candido Jr, Aluísio, 2008a; Candido Jr, Aluísio, 2008b), funded by the National Council for Scientific and Technological Development (CNPq), began in late 2006 and will run until 2010. The project's aim is to build a historical dictionary of Brazilian Portuguese covering the period from the sixteenth century through the beginning of the nineteenth century. It is based on a historical corpus that contains texts from the same period, compiled within the scope of the project. In the last two years of the project, efforts are being directed to the creation of the dictionary, a task involving only lexicographers and terminologists.

The HDBP project fills a gap in Brazilian history, following the example of several languages that are already supported by historical dictionaries or have historical dictionary projects under way. Historical European Portuguese has the *Deparc* (*Dicionário Etimológico do Português Arcaico* – Etymological Dictionary of Old Portuguese) (Machado Filho, 2005), whose aim is to create a historical dictionary for the period between the thirteenth and the sixteenth centuries, and the *Dictionary of Medieval Portuguese Verbs* (Xavier, 2008). The Oxford University Press continues to work on the *Historical Dictionary of American Slang* (Lighter, O'Connor and Ball, 1994). Its new version will probably contain more than 35,000 entries and is based on a corpus of more than 10,000 texts. The *Dictionary of the Scots Language* (*DSL*) (*Dictionary of the Scots Language*, 2008) is an online tool that comprises two historical dictionaries: the *Dictionary of the Older Scottish Tongue* (*DOST*) (from the twelfth to the seventeenth centuries) and the *Scottish National Dictionary* (*SND*) (from the eighteenth century to the 1970s). The *Historical Dictionary of Icelandic* (Pind *et al.*, 1993) spans the period from 1540 to the present. The *Nuevo Diccionario Histórico del Español* (*New Historical Dictionary of the Spanish Language*) (Ruiz and Martínez, 2008) is being developed by a team of 20 philologists. Besides these ongoing projects, some researchers emphasize the need for specialized dictionaries. Mahoney (1998), for instance, argues that it is necessary to create a diachronic and descriptive historical English dictionary of astronomy, since the ones available are synchronic, prescriptive, and encyclopedic, which makes them of little use for reading historical texts in the field. Mahoney supports the creation of a dedicated corpus-based dictionary of astronomy that includes all obsolete terms and changes of meaning, and also lists and defines concisely the astronomical lexicon from early English to the present day.

As is the case for the HDBP, many of the projects mentioned above are corpus-based. In some of them, corpora were adopted from the beginning, whereas in those started before the corpus processing technology was available, corpora were introduced at a later stage.

Initiatives for building historical corpora, mainly those that follow the principles of Corpus Linguistics (McEnery and Wilson, 2001), are particularly important, since such challenges are rare and make it possible to preserve the history of a country and its linguistic records, besides favoring the study of the evolution of a language in the period under investigation. As an example of projects to build historical corpora for the Portuguese language, we can mention the Tycho Brahe Project, the Portuguese Corpus, the Program for a History of the Portuguese Language (PROHPOR), and the Digital Corpus of Medieval Portuguese.

The Tycho Brahe Project[1] (Paixão de Sousa and Trippel, 2006), whose purpose is to model the relation between prosody and syntax from Classical to Modern European Portuguese, contains tagged and parsed texts written by Portuguese authors born between 1435 and 1845. Currently, this corpus has 52 texts (2,356,811 words), publicly available for research, by means of a two-stage system of linguistic annotation: morphological (applied to 26 texts) and syntactic (applied to three texts).

The Portuguese Corpus contains texts from both Brazilian and European Portuguese, and is publicly available[2] as well. Its texts were written between the fourteenth and the twentieth centuries. It has now 45 million words and includes texts from other corpora, such as the Tycho Brahe and the Brazilian Portuguese reference corpus of the project Lácio-Web (Aluísio *et al.*, 2004).

The corpus of the BIT-PROHPOR (*Banco Informatizado de Textos do Programa para a História da Língua Portuguesa* – Computerized Text Bank of the Program for a History of the Portuguese Language) is used in the Deparc project mentioned above. Both Deparc and BIT-PROHPOR are part of the project "Program for a History of the Portuguese Language"[3].

Researchers at the Universidade Nova de Lisboa have built the Corpus Informatizado do Português Medieval (CIPM – Computerized Corpus of Medieval Portuguese)[4], comprising Latin-Romance texts from the ninth to the twelfth centuries, and Portuguese texts from the twelfth to the sixteenth centuries, totaling some two million words. The *Dictionary of Medieval Portuguese Verbs*, mentioned previously, was based on this corpus.

However, of the four projects mentioned, only the latter two are dedicated to building historical dictionaries, and neither of them focuses on Brazilian Portuguese. The HDBP project will produce the first historical dictionary applied to the Brazilian variant, which began to differ from European Portuguese as early as the first centuries of our history. The HDBP fills a gap in Brazilian culture with a dictionary that describes the vocabulary of Brazilian Portuguese from the beginning of the country's history. Although some vocabulary had already been forged on this side of

1 http://www.tycho.iel.unicamp.br/~tycho/.
2 http://www.corpusdoportugues.org/.
3 http://www.prohpor.ufba.br/projetos.html.
4 http://cipm.fcsh.unl.pt.

the Atlantic, at that time the Brazilian variant still depended on European Portuguese. However, even at that early period, people faced a world materially and culturally different from what was known in Europe, and they needed to resort to European Portuguese to designate these previously unnamed referents of their new universe. Hundreds of native languages were then spoken in Brazil and had their own vocabulary for designating elements of the Brazilian fauna and flora, but these words did not belong to European Portuguese. Habits and institutions gradually began to form in this new society, as a result of the blend of new cultures. Inevitably, new words formed 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 living in this country, allows us to explore and unearth the vocabulary repertoire used from the sixteenth through the eighteenth centuries.

The corpus for the first three years of the HDBP project is completely compiled, and contains 2,458 texts annotated with basic Text Encoding Initiative (TEI) (Tei Consortium, 2006) and about 7.5 million simple forms, i.e., the total number of words in the corpus that are composed of letters that belong to a Historical Portuguese alphabet especially created to process the corpus with corpus processing tools. There are approximately 368,000 unique simple forms.

The texts selected for the corpus include letters written by Jesuit missionaries, documents of the *bandeirantes* (members of the exploratory expeditions that pushed Brazilian borders far into inland areas), reports of the *sertanistas* (explorers of Northeastern Brazil), documents of the Catholic Inquisition, inventories, and testaments, among others.

Compiling this historical dictionary was a comprehensive and time-consuming task of analyzing documents, printed material, and manuscripts produced by eyewitnesses to the early stages of Brazilian history. A significant difficulty derived from the absence of a press in colonial Brazil, which had a precarious communication system. Only after 1808 were communications improved, when the Portuguese monarchy fled from Napoleon's army and transferred the government of the Portuguese empire to Brazil. In addition, we had to consider some peculiarities concerning language: biodiversity and multifaceted cultural traditions. Therefore, to implement the project we decided to 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.

During the project design, we learned that, despite the many computer tools available to process corpora, only a few were able to fulfill the requirements for building historical Portuguese corpora as expected. Some of the problems we encountered are described in Section 2.1. Before deciding on Unitex (Paumier, 2006) and Philologic (University of Chicago, 2008) as the tools to use in the project, we made a comparison of free software for processing corpora, as shown in Section 3. We also detected the need to develop a tool to write entries with an interface

customized for the HDBP requirements, since existing tools are adequate to terminological and/or contemporary dictionaries, but not useful for historical texts.

Another prerequisite was to build glossaries (or computational lexicons) of abbreviations and spelling variants to support the creation of the historical dictionary. This issue demanded special attention, because abbreviations not correctly expanded can limit the effectiveness of information extraction and retrieval systems in digital libraries, hinder electronic index creation from a corpus, and reduce the capability of Natural Language Processing (NLP) tools, such as taggers, parsers, and named entity recognition (NER) systems that enrich corpora linguistically. Within the scope of the HDBP project, incorrect abbreviation expansion prevents the correct editing of dictionary entries. However, manual expansion of each and every abbreviation in a several-million-word corpus is time-consuming, expensive, and difficult – if not impossible, as is the case when noun abbreviations are ambiguous. This is why we had to tackle this problem in a different way, explained in Section 4.2.

As mentioned before, historical texts do not comply with a spelling standard and produced a large amount of spelling variants, making it difficult to use successfully the standard indexing techniques for information retrieval (Hauser *et al*., 2007; Ernst-Gerlach and Fuhr, 2006; Braun, 2002) and NLP tasks (Crane and Jones, 2006). Besides, it is useless to apply corpus annotation tools trained on contemporary language data to historical texts, since they will not deal with the spelling variants of a word (Rayson *et al*., 2005). Whenever a dictionary is being compiled, spelling variants hamper the search for agreement between words, limiting the number of possible examples. Our approach, explained in Section 4.2, is to apply a series of transformation rules to a list of single words extracted from a corpus to group different spellings around a common spelling.

Therefore, in this paper we summarize the work carried out to compile the HDBP historical corpus, as well as to build resources, methodologies, environment, and tools especially for the project. Some of these resources and tools are freely available[5], and they can be reused by other projects dealing with the Portuguese language or even adapted to projects dealing with other languages. This paper is organized as follows. Section 2 demonstrates the processes used to compile and pre-process the HDBP corpus. Section 3 presents a comparison between corpus processing tools. Section 4 describes the glossaries developed to support corpus access and dictionary creation. Section 5 introduces the system for writing entries. Section 6 details the computational environment for processing corpora employed in the HDBP project, which can also be used in similar projects.

5 http://www.nilc.icmc.usp.br/nilc/projects/hpc/.

2. Compiling the corpus

The HDBP was compiled from printed documents, manuscripts, and Portable Document Format (PDF) image files. Manuscripts were keyboarded, whereas original printed documents were processed by Optical Character Recognition (OCR). PDF files were converted into TIFF files before being scanned. All texts were coded in Unicode UTF-16, which allowed us to preserve symbols commonly found in Brazilian historical texts but already fallen into disuse, such as the symbol "long s" (ſ). Next, texts were submitted to semi-automatic cleaning and annotation. Cleaning consisted of removing from texts undesired parts such as headers, footers, and line numbers. Each text was then supplied with administrative metadata, such as author's name, page numbering, and document title, to be used with both of the corpus processors, Unitex (metadata are not taken into account either by frequency count or concordancer) and Philologic. We employed the TEI P4 lite tagset[6], including paragraph annotation. Figure 1 illustrates this process.

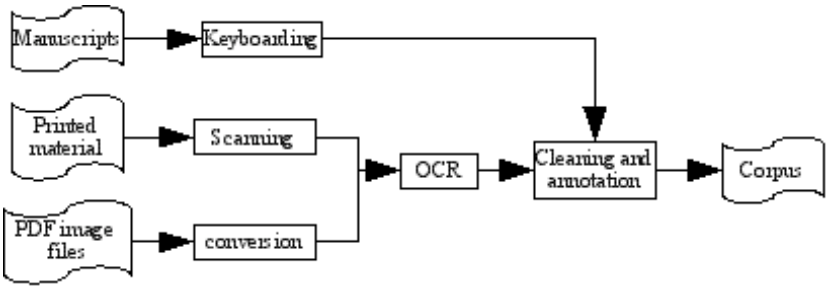


Figure 1. *Corpus compilation process*

The HDBP corpus will not be publicly available at first, since it is necessary to obtain authorization from publishing companies to circulate the texts (although there are some in the public domain, most of them belong to current editions under copyright law).

2.1. Compiling historical corpora: some issues

Rydberg-Cox (2003) and Sanderson (2006) list some issues concerning the compilation of historical corpora: words broken at the end of a line in historical Latin, Greek, and English texts, to mention just a few languages, since words are not always hyphenated; word-breaks that are not always used; abbreviated common words and word-endings, using non-standard typographical symbols; uncommon typographical symbols in non-abbreviated words; and spelling variation even within

6 http://www.tei-c.org/Guidelines/Customization/Lite/.

the same text. These issues arose in the HDBP project as well, and we describe them below.

There was no unified spelling system in the centuries covered by the project. At that time, scribes and copyists used a Babel of graphic symbols. Some features common to Portuguese texts prior to the eighteenth century, observed by (Menegatti, 2002), are double consonants, inconsistent use of diacritical marks, and vowel interchange. Given that it is important for lexicographical work to retrieve all occurrences of a lexia and that spelling variants occur even within the same text, lexicographical tasks become more difficult.

There are several tools to detect spelling variants automatically (Archer *et al.*, 2006; Hirohashi, 2004; Rayson, 2005). In (Giusti *et al.*, 2007), we proposed a method for using manual transformation rules to detect spelling variants automatically. This proposal is detailed in Section 4.1.

Another issue regarding the compilation of historical corpora is abbreviation. The scribes' habit of abbreviating words to make handwriting easier produced many thousands of abbreviations. Therefore, to understand texts correctly, it was necessary to expand them, a task that poses two main difficulties. The first refers to the use of modern knowledge sources, since gazetteers, encyclopedias, and heuristics currently in use do not deal directly with the characteristics of historical material, which describes people, places, and other entities that often do not appear in modern sources (Crane and Jones, 2006). The second, and perhaps the most important, is that even if we had adequate knowledge sources for expanding abbreviations, these are highly ambiguous with respect to meaning, which is critical for understanding correctly not only the abbreviations themselves but the whole text (Kerner *et al.*, 2004). Although there are techniques for expanding abbreviations automatically in contemporary languages (Terada *et al.*, 2004), there is not much research yet on treating abbreviations found in historical texts. An alternative is to use glossaries of abbreviations to support manual expansion while searching the corpus. We have chosen this approach – described in Section 4.2 – for the HDBP project, because it is faster to implement and less prone to errors.

Regarding the problem of uncommon typographical symbols, good character coding and adequate tags to denote them, such as the tag "<symbol>" from the TEI tagset, are useful for treating them. Unicode is particularly important for treating historical corpora, which are full of characters not allowed in the usual encoding patterns, such as the symbol "æ" (combination of "a" and "e") and the symbol "m̃" (as in "com̃ercio", commerce). In the HDBP project, we have chosen Unicode precisely because it can represent all symbols found in our historical texts.

Another feature that also makes searching the corpus difficult is word junction. In this case, the most appropriate solution is to split words. Junctions between prepositions and nouns are frequent, as in "acargo" (which, if split, becomes "a cargo" – under the responsibility of) and "depernambuco" ("de Pernambuco" – "from Pernambuco"), and there are several other examples, including articles

(“ocapitão”: “o capitão” – “the captain”), pronouns (“seusfilhos”: “seus filhos” – “their children”), proper names (“FranciscoCoelhoBitancur”: “Francisco Coelho Bitancur”), and even more complex cases involving different parts of speech (“seriamaisconveniente”: “seria mais conveniente” – “it would be more convenient”). For the HDBP project, we created a manually compiled glossary to explain junctions and support searches in the texts, using the TEI pattern for annotating junctions with the tag “<choice>” which makes it easier to replace occurrences of junctions in the corpus if such a version is desired.

2.2. Pre-processing the corpus

The tasks performed to pre-process the HDBP corpus were cleaning and annotating digital texts digitized as DOC files and converted into TXT with annotation. For this purpose, we developed the tool Protew (Candido Jr, 2008a). The TXT format allows for the generation of corpora in simplified XML, used to create corpora in the TEI format or in pure text format with cataloguing-in-publication information. To generate different corpus formats, we developed the tool Protej (Candido Jr, 2008a). Examples of tasks Protew and Protej can perform are converting the header into XML, removing hyphens automatically whenever possible, and treating line and paragraph numbering.

Figure 2 shows a percentage chart of corpus distribution by century. The values of the columns were normalized for visualization purposes (each color sums to 100%). There are few texts from the sixteenth century, because at that time not many Brazilians were literate, and besides, some of the documents have been lost due to the passage of time. This is a lesser problem for samples from the seventeenth century. The eighteenth century is represented by more texts. The nineteenth century is represented by few texts, in view of the fact that the corpus contains documents up to 1808 only.

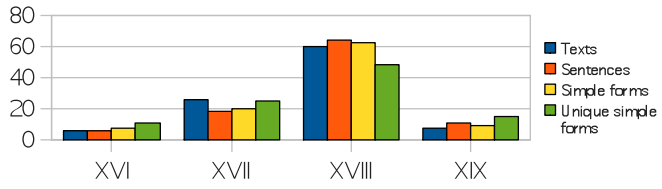


Figure 2. *Corpus distribution by century*

3. Accessing the corpus

Focusing on free software, we carried out a comparison among corpus processors to support our decision about which tools to use in the HDBP project.

Similar comparisons had already been made (Rayson, 2002; Schulze, 1994; Santos and Ranchhod, 2002; EAGLES 1995), but in general they were not focused on free software. Our comparison included five tools: GATE[7], Philologic, Corsis[8], Unitex, and Xaira[9].

These five corpus processors were evaluated as to software quality, using the six metrics defined in ISO 9126 (Eagles, 1995): functionality, reliability, usability, efficiency, maintainability, and portability. Table 1 shows some of the criteria adopted to analyze corpus processors. More details on Table 1 metrics and criteria can be found in (Candido Jr, 2008a).

Table 1. *Comparison of corpus processing tools*

Criterion	**GATE (build 2752)**	**Philologic 3.1**	**Unitex 2.0 beta**	**Corsis 0.1.3.2**	**Xaira 1.23**
Concordancer	yes	yes	Yes	yes	yes
frequency count	no	yes	Yes	yes	yes
glossary-oriented search	yes	yes	No	yes	yes
Annotation	yes (XCES)	yes (TEI-Lite)	partial (lexicon and local grammars)	partial (lexicon and local grammars)	yes (TEI or similar)
collocations or *n*-grams	yes	yes	no	no	yes
character coding	UTF-8	UTF-8	UTF-16	UTF, ISO, etc.	UTF-8/16
pre-processing time (in secs)	663	61.5	19.5	--- *	36.9
concordancer time (in secs)	212	1.5	8	13.5	0.7

* Corsis does not pre-process texts

The generation of concordances is an important resource when corpora are used to perform lexicographical tasks. In this prerequisite, Philologic, Unitex, and Xaira were good choices. GATE does not have a standard concordancer, but based on the resources it offers it is possible to create one. On the other hand, GATE has good resources that can be used in different kinds of research, such as parsing and corpus tagging. Corsis, in turn, has a user-friendly concordancer, but presented performance problems, since it does not index texts. Another of Corsis's problems is that it is still being developed, and consequently few resources are available. It is more useful for researchers who are looking for an alternative to WordSmith Tools[10].

With regard to the HDBP project, the most appropriate tools were Philologic, Unitex, and Xaira. Xaira was not used, in spite of having a great number of search resources, because we consider its interface difficult for beginners. For this reason, we chose Philologic and Unitex. We picked Philologic because of its user-friendly interface and because it centralizes data offered by web tools, besides allowing the survey of spelling variants by means of the edit distance algorithm AGREP (Approximate GREP), a fuzzy string searching program developed by Udi Manber and Sun Wu (1992). AGREP is used in Philologic similarity searchers to look for similar or alternative spellings for a query in a collection of texts. Unitex was chosen

7 http://gate.ac.uk/.
8 http://sourceforge.net/projects/corsis/.
9 http://www.oucs.ox.ac.uk/rts/xaira/.
10 http://www.lexically.net/wordsmith/.

because of its glossary processing tool, which simplifies searches for abbreviations and spelling variants.

4. Glossaries

We developed three glossaries in the project: (a) the glossary of abbreviations and their expansions, (b) the glossary of word junctions (manual, with 10,369 junctions), and (c) the glossary of spelling variants, to help in searching for concordances and frequency count. The glossary of abbreviations and the glossary of variants follow the DELAF formalism (Paumier, 2006) used by Unitex, and are detailed in Sections 4.1 and 4.2. These glossaries can also be accessed in the Procorph system, described in Section 6.

4.1. Spelling variants

Several researchers have dealt with the problem of spelling variants in historical corpora in projects about English, German, French, and Portuguese, to mention just a few languages (Rayson, Archer, and Smith, 2005; Archer *et al*., 2006; O'Rourke *et al*., 1996; Hirohashi, 2005).

Rayson, Archer, and Smith (2005) and Archer *et al*. (2006) describe a tool named VARD (VARiant Detector) for detecting and normalizing variants of the English language to its modern form automatically. VARD includes a pre-processor that detects historical spelling variants and inserts their modern equivalents in the system; consequently it does not have to retrain each and every annotation tool applied to the corpus. From a different point of view, the part-of-speech (POS) tagger[11] developed to annotate the Tycho Brahe corpus added historical variants to the POS tagger lexicon to manage original (historic/ancient) spellings found in Portuguese texts. Later, within the Tycho Brahe Project, researchers devised a methodology for normalizing spelling variants in the corpus automatically (Hirohashi, 2005).

Our research on spelling variant treatment is based on Hirohashi's (2005) methodology. We apply a series of transformation rules to a list of single words extracted from a corpus. Our aim is to group different spellings around a common spelling. Thus, the system that implements this approach can establish a relation between different spellings. It is expected that this relation will show spelling variations for any given word.

The system we developed was named Siaconf (*Sistema de Apoio à Contagem de Frequência em Corpus* - Support System for Frequency Count in Corpus). It

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

processes a corpus from an initial list of rules, built by diachronic linguists or by an expert who bases his work on diachronic linguistics, and provides three main types of detailed reports: (a) groupings/clusters including spelling variants of the same word, (b) information on the rules applied, and (c) a list of non-processed words. The grouping used in our research is different from the normalization approaches in Hirohashi (2005) and in the VARD tool. 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 does not exist in Brazilian Portuguese any more. Our aim is that groupings will reduce the impact of spelling variation on frequency count and that grouping contents will allow the study of spelling variation in the corpora compiled. For instance, using transformation rules, the following variants of "chão" were found: chaõ, xão, cham, chaão, and xam. Figure 3 shows four examples of clusters resulting from applying Siaconf to our corpus with frequencies for each spelling. The cluster "apelido" (nickname), for example, has 90 instances of actual words from the corpus.

mais (44,658)			indios (8341)		
	mais	(44,326)		indios	(5122)
	maes	(188)		índios	(2990)
	majs	(100)		jndios	(111)
	mays	(38)		yndios	(88)
	máis	(2)		imdios	(14)
	maís	(2)		hindios	(5)
	maïs	(1)		jmdyos	(5)
	mais	(1)		ymdios	(2)
				imdyos	(1)
				indíos	(1)
				jmdios	(1)
				ymdyos	(1)
apelido (90)			vila (5,218)		
	appellido	(48)		villa	(4,073)
	apelido	(30)		vila	(1,113)
	appelido	(7)		vyla	(13)
	apellido	(5)		vjlla	(9)
				vylla	(9)
				vjla	(1)

Figure 3. *Examples of spelling variation in "mais" (more), "Índio" (Indian/native inhabitant), "apelido" (nickname), and "vila" (village), in the report of groupings*

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 that will be processed by the rule. *C2* determines a substring in each W_i, which will be replaced by *S*. For example, the rule "(e[ao] e ei)" is applied as follows:

a) *C1* is tested against every form of the corpus and restricts the rule application

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

to those that contain the substring “ea” or the substring “eo”, for example: “aldea” (variant of small village).

b) *C2* determines the substring that will be replaced, for example: the letter “e” in “aldea”.

c) *S* determines the replacement string (“ei”), used to generate the new form, for example: “aldeia” (small village).

After applying different rules, several spellings G_i produce a new spelling *H*. Thus, it is possible to infer that spellings G_i are variants of the same word. For instance, the rules (ll, ll, l) and (y y i) can be applied to the spellings “vyla” and “villa”, respectively, resulting in the new spelling “vila”. Therefore they are highly likely to be variants of the same word. In addition, more than one rule can be applied to a given spelling, as shown in Table 2.

Table 2. *Grouping of “não” and “naõ” (variants of “not”) around spelling “nam”*

Words	**Rules applied**	**Spellings generated**
NAÕ	[óòõôõ] . o [^r][aã]o$ [aã]o am	"nao" "nam"
NÃO	[^r][aã]o$ [aã]o am	"nam"

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 spellings H_i are not orthographic, i.e., results from the process described are not necessarily normalized versions of a word. Currently, we are using 51 transformation rules. Our rules can be divided into six groups:

– Rules for spellings that have fallen into disuse. For example, replacement of “y” by “i”. “Y” and “i” sound the same in Portuguese. However, “y” has been replaced by “i” in all words, except for foreign words and proper names. Other rules are:

ee ee é	[áàäâ] . a	[ýỳÿŷ] . y	^ha ha a
ph ph f	[éèëê] . e	gu[ao] gu g	^he he e
pt pt t	[íìïî] . i	dh dh d	^hi hi i
th th t	[óòöô] . o	v$ v u	^ho ho o
ſ ſ s	[úùüû] . u	[^r][aã]o$ [aã]o am	^hu hu u
g[ei] g j			

– Rules for double consonants. For example, replacement of “ff” by “f”. Other rules are:

pp pp p	mm mm m	gg gg g	ll ll l
tt tt t	bb bb b	vv vv v	uu uu u
nn nn n	dd dd d	zz zz z	cc cc c

– Rules generated according to the orthographic norm. In the Portuguese orthographic norm, "m" and "n" sound the same when preceding consonants. However, "m" precedes only "b" and "p", whereas "n" precedes all other consonants. They are:

```
j[bcdfghklmnpqrstvwxz] j i
m[cdfghjklqrstwxz] m n
mn mn n
mpt mp n
mpt mpt nt
n[pb] n m
ct ct t
```

– Rules based on frequency, formulated to treat recurring patterns in spelling variations. For example, replacement of "chr" by "cr", as in Christo (Christ). Other rules are:

```
ch ch x
.acem$ c ss
aes$ aes ais
```

– Lexicalized rules: rules for specific words. For example: replacement of "o" by "u" in "Deos" (God).

– Automatic rules, based on Hirohashi's study (2005) of the automatic learning techniques on the Tycho Brahe corpus. It is not possible to use the same techniques on the HDBP, since the HDBP corpus does not have the same level of annotation as the one performed on the Tycho Brahe Project. An example is the replacement of "z" by "s" in the infix "zente", as in "presente" (gift/present). Other rules are:

```
ozo$ z s
serviss serviss service
preciz preciz precis
```

After applying these rules to our corpus, we identified 76,754 spelling variants in 31,069 word groupings. The report of non-processed words generated by Siaconf is useful for developing new rules. In this report, it is possible to find words with high frequency in the corpus that are not grouped by any rule.

A comparison between Siaconf and AGREP, used in Philologic, showed that Siaconf's precision is the highest possible (near 100%); however, AGREP performed better on recall. Siaconf's recall can be improved with the development of new rules. Both the Siaconf glossary and AGREP suggestions are available to HDPB researchers. Unlike transformation rules, there is no glossary for edit distance, since users can survey spelling variants on the fly as they access Philologic. There is also the possibility of creating a hybrid glossary with variants collected by transformation rules and variants collected by edit distance. We opted for not doing this, since the technique based on transformation rules prioritizes precision (which is useful for automatic tasks), whereas edit distance prioritizes recall (which is useful for manual tasks). We consider these techniques complementary. A comparison between the two strategies in terms of precision and

comparative recall (a measure employed in information retrieval systems) is shown in Table 3.

Table 3. *Comparing transformation rules and edit distance (Giusti et al., 2007)*

Strategy	True positives	False positives	Precision	Comparative recall
Transformation Rules (Siaconf)	36	0	100%	72%
Edit Distance (Philologic/AGREP)	41	196	21%	84%

Figure 4 shows DELAF entries that correspond to variants of "muito" (more/much).

```
muito,muito.N+VAR:ms/92.39%
muyto,muito.N+VAR:ms/7.16%
mujto,muito.N+VAR:ms/0.34%
muitto,muito.N+VAR:ms/0.08%
```

Figure 4. *Examples of entries conforming to the DELAF formalism*

Each entry is composed of variant, new spelling generated by Siaconf, word class, semantic attributes, information on inflection, and frequency of variant in the corpus. The whole process is automatic, so all entries are masculine singular (ms) nouns (N). A manual revision will be carried out later to insert grammatical and inflection data.

4.2. Abbreviations and their morphosyntactic and semantic information

In historical texts, the scribes' habit of abbreviating words to make handwriting easier has produced many thousands of different abbreviations. Hence to understand texts correctly, it is necessary to expand these abbreviated forms. Within the scope of the HDBP project, failing to expand abbreviations properly hinders the correct editing of dictionary entries. However, expanding each and every abbreviation manually in a several-million-word corpus is time-consuming, expensive, and difficult – if not impossible, due to the ambiguity inherent in noun abbreviations, for example.

Figure 5 illustrates the problems related to abbreviations: ambiguity and variants. The first column shows 13 different expansions for the abbreviation "A". The second illustrates 13 different forms of abbreviating the name of the Brazilian city "Rio de Janeiro" (some of them in lower case), which make them hard to memorize.

alteza (highness)	Rio de Jan.^{ro}
alvará (warrant)	Rio de Jan^{ro}
Amaro (proper name)	Rio de Janr.^{o}
Ana (proper name)	Rio de Jan.^{o}
anima (cheers up)	Rio de Jn^{ro}
ano (year)	Rio de janr^{o}
anos (years)	Rio de jan^{ro}
Antônio (proper name)	R ^{o} de jan^{o}
arroba (measure of weight, singular)	R ^{o} de Jan^{ro}
arrobas (measure of weight, plural)	R ^{o} de janer^{o}
Assembléia (assembly)	R ^{o} de Janr^{o}
assinado (signed)	R ^{o} de Jnr^{o}
Atual (current)	Rio de Janr^{o}

Figure 5. *Ambiguity and spelling variation in abbreviations (Vale et al., 2008)*

There are several graphic forms for the abbreviations found in the *HDBP* corpus:

a) abbreviations with a dot followed by superscript chunks of text, as in "Janr.^{o}"/Janeiro (January) and "corre.^{te}"/corrente (current);

b) abbreviations followed by a dot, as in "porq."/porque (because) and "q."/que (who).

To be consistent, we used the character "^" to denote superscript, thus generating the forms "Janr.^o" and "corre.^te" showed in (a) above, which can be automatically processed. The same symbol was used when the abbreviation did not have a dot but a superscript chunk, as in "O s^{or} Jesus xp^{o}"/"O Senhor Jesus Cristo" (The Lord Jesus Christ), producing the forms "s^or" and "xp^o". Other abbreviations display numerals, e.g., "8.bro"/"Outubro" (October), or other characters, e.g., "@" for the word "ano" (year). Some abbreviations only omit letters, as in "Glo"/"Gonçalo" (proper name "Gonçalo"), "Jão"/João (proper name "João"), "ldo"/"licenciado" (licensed), "Ros"/"Rodrigues" (proper name "Rodrigues"), and "snr" or "snro"/"senhor" (sir).

Most of the previous work on Brazilian Portuguese historical corpora expands abbreviations manually, as in "Para uma História do Português do Brasil"[13] ("For a History of Brazilian Portuguese") and "Projeto Programa para a História da Língua Portuguesa" (PROHPOR). Also, in the Tycho Brahe Project, abbreviations were expanded manually to make tagging and parsing easier. Although large for syntactic analysis, the Tycho Brahe corpus – currently composed of 52 texts and still growing – remains manageable by manual markup written with widely available standards in XML. The large-scale Germany-wide project Deutsch.Diachron.Digital (DDD) (Dipper *et al*., 2004) was set to build a diachronic corpus of German with texts from the ninth century (Old High German) to the present (Modern German) for linguistic, philological, and historical research. This is a long-term project – it is planned to run over seven years – and its large core corpus will reach 40 million words. The

13 http://www.letras.ufrj.br/phpb-rj/.

abbreviations found in it will be expanded and annotated, based on generally-accepted international standards in XML.

All the projects mentioned above expand abbreviations manually; however, their development contexts differ from that of HDBP, which has only four years to develop both a large corpus and a dictionary.

Automatic disambiguation of acronyms and abbreviations has deserved close attention in medical and biomedical domains, since text normalization is crucial for successful information retrieval and extraction in these areas (Pakhomov, 2002; Hong *et al*., 2002; Schwartz & Hearst, 2003; Dannélls, 2006). However, most of this automatic research has focused on modern scientific material, largely ignoring historical corpora and digital libraries (Rydberg-Cox, 2003). Taking this into consideration, we built a large dictionary of abbreviations containing pairs composed of abbreviations and their expansions, together with morphosyntactic and semantic information (a predefined set of named entities – NEs) (Vale *et al.*, 2008).

In order to build this dictionary of abbreviations, we employed lexicons together with corpus processing tools, particularly to expand a printed dictionary converted to digital form (Flexor, 1991) and to enrich it with information about the NE categories appearing in the HDBP corpus. Flexor (1991) is a large, alphabetically organized dictionary of abbreviations from the sixteenth through the nineteenth centuries. Despite its large number of abbreviations (see Table 4), most of them are not found in our corpus (only 16% are part of the HDBP corpus). We conducted an experiment to retrieve abbreviations from the HDBP corpus using three simple heuristics to estimate the amount of abbreviations that were not in the Flexor dictionary. We found 7,045 abbreviations with three simple heuristics (words with superscript; words with a dot between letters, and words ending with some consonants); only 35% of the total (2,473) were in the Flexor dictionary. However, it is still useful, since it permits abbreviation expansion.

Table 4. *Abbreviations from Flexor (1991) by century, showing % of forms found in the HDBP corpus*[14] *(Vale et al., 2008)*

Simple and multi-word abbreviations by century					
Types	Sixteenth	Seventeenth	Eighteenth	Nineteenth	Total
Flexor	2,050	4,091	14,376	9,939	21,869
Flexor (%)	9.37	18.70	65.74	45.45	139.26*
Intersection of Flexor and Corpus	754	1,323	2,447	1,710	3,529
Intersection of Flexor and Corpus (%)	21.37	37.49	69.34	48.46	176.65*

14 Observe that abbreviations can occur in more than one century.

Coverage (%)	16.13

Our dictionary of abbreviations differs from its counterparts developed in Unitex mainly in the use of a larger number of attributes. These are the most important additional attributes: ABREV to denote abbreviation; SEC16, SEC17, SEC18, and SEC19 to show the century to which lexical entries refer (information from Flexor 1991) – the century attribute appears only in some entries, since it was not always possible to identify the period in which the abbreviation was used; <ENT> to denote a named entity (NE); and the tag <INIT>, a collocation to extract certain types of NE. Each NE receives further attributes, according to the category it belongs to. These categories were established by a taxonomy proposed in the evaluation contest of systems for recognizing named entities in Portuguese (HAREM[15]), organized by Linguateca. We have employed the ten HAREM top categories in our dictionary of abbreviations: person, organization, artifact, location, thing, event, abstract entity, quantity, time, titles/man-made things. Figure 6 shows some lexical entries in DELAF formalism. In the first line of Figure 6, "Brg^es" is the form found in the corpus, "Borges" is the canonical form (lemma), "N" (noun) is the word-class tag for the entry, "ENT+PESSOA+ABREV+SEC19" are further attributes, and "ms" (masculine singular) is the morphosyntactic tagging. We also included the expanded form ("Borges"), which may differ from the canonical form in some cases.

Our dictionary has 18,499 simple abbreviations, with 8,030 classifications of ENT, INIT, ENT+INIT. The dictionary of abbreviations was designed to recognize large patterns of complete abbreviations. It also contains a specific tag to treat jobs/professions and titles/forms of address, such as "capitão" (captain), "frei" (friar), "promotor" (prosecutor), "Ilustríssimo" (Most Illustrious/Honorable), "Dom" (Don), "Majestade" (Majesty), "Senhor" (Sir), and family relations, such as "cunhada" (sister-in-law), "primo" (cousin).

```
Brg^es,Borges.N+ENT+PESSOA+ABREV+SEC19:ms/Borges
Brag.,Braga.N+ENT+PESSOA+LOCAL+ABREV+SEC18:ms/Braga
Br^ça,Braça.N+ENT+VALOR+ABREV+SEC19:fs/Braça
7^bro,setembro.N+ENT+TEMPO+ABREV:ms/setembro
B^eis,bacharel.N+INIT+TITULO+ABREV:mp/bacharéis
B.,beco.N+INIT+LOCAL+ABREV+SEC18:ms/beco
Bat^am,batalhão.N+INIT+ORGANIZAÇÃO+ABREV+SEC16:ms/batalhão
Bas^tos,bastardo.N+INIT+PARENTE+ABREV+SEC19:mp/bastardos
```

Figure 6. *Samples of dictionary entries (Vale et al., 2008)*

In linguistic research, it is very important to know whom the text is about and to whom it is directed. If we determine the authorities being addressed in a specific text, we can identify the words used in that specific register, given that a letter written to an ordinary person does not contain the same words and level of formality

15 http://www.linguateca.pt/HAREM/.

as one written to a monarch. This identification is possible because we used both NEs and other specific tags.

The morphosyntactic and semantic annotation of the abbreviations dictionary using information from Flexor (1991) is complete. However, its expansion with abbreviations extracted from the HDBP corpus via specific search patterns to extract new NEs of a given NE category has just started. Thus far, we have collected samples of proper names, hydronyms, and places other than bodies of water, totaling 228 entries. To perform this task, we are running the same process defined for REPENTINO[16], a repository of modern Portuguese NEs, except for the last stage, in which we adopted the NE taxonomy defined in HAREM: 1) choose a category for which you intend to search examples of entities; 2) decide which is the most appropriate strategy to search for examples: a) by tag <INIT>, such as in Rio S. Francisco; b) by context, such as in "localizado na XXX" (located at XXX), which strongly suggests that "XXX" is a place; or c) by discriminating suffixes (modern organizations' names include characteristic particles such as "Ltda."/Ltd. or "S.A"/Co.); 3) construct the respective pattern to be searched in a given corpus processor or to act as an independent program, and start the search; 4) validate manually the candidates you obtained, considering the intended category; 5) include positive candidates in the repository; 6) if necessary, create a new category or subcategory, thus expanding the taxonomic classification system. To support this process, we have developed an application for recognizing NEs. It retrieves NEs from the HDBP corpus automatically, via pattern search, and stores the NE, its manually inserted expansion, and one of the ten HAREM top categories plus one sample sentence in the web repository of abbreviated historical NEs[17] in Brazilian Portuguese.

As a consequence of the large number of abbreviations and spelling variations related to both abbreviated words and expanded words, this process had to be adapted to historical corpora. The prerequisite for accepting a new NE from the corpus was that at least one of the components should be in the abbreviated form. Capitalization was not a viable requirement, since proper names are not always capitalized in historical corpora. To illustrate the adaptations of this procedure for retrieving new NEs from a corpus, we discuss a case study about hydronyms – names of rivers, streams, creeks, and brooks found in the HDBP.

Flexor's dictionary contains 18 entries following the pattern Rio XXX/River XXX, but eight of them refer to the city of Rio de Janeiro and not to bodies of water (the other are: R^o da Ribr^a, R^o de Reg^o, R^o de S. Fran^co, R^o dos Alm^das, R^o G^de, R^o G^re, R^o Gdr^e, R^o Gr^de, R^o G^re e R^o P^do). As we did not find anything about Creek XXX (or its variants: brooks, streams, etc.), we began with ten entries. The preferred search strategies were: patterns formed by tag <INIT> and contexts "naveg*"/navigate (on), which include several conjugations of

16 http://poloclup.linguateca.pt/repentino/.
17 http://nilc.icmc.usp.br:8180/renahb/.

the verb "to navigate". However, words tagged as <INIT> would appear in their abbreviated or expanded form and, besides, we would have to deal with spelling variations and synonyms.

To treat spelling variants, we adopted two resources: the dictionary of spelling variants, created according to the Siaconf methodology proposed in Giusti *et al*. (2007), described in Section 4.1, and the Philologic resource for searching similar patterns, which uses AGREP. To treat synonyms for river, we drew on the Brazilian Portuguese Electronic Thesaurus TEP (Greghi *et al.*, 2002).

5. Entry writing

Thirty-one systems for creating dictionaries and supporting lexicographical and terminological tasks are described in Universität Leipzig (2008). Most of them focus on terminology, but some provide the tools for developing general language and multilingual dictionaries. Overall, these systems were written for English, which reduces their performance when they are used to create dictionaries for Portuguese. In Haddad (1999), the authors confirmed that off-the-shelf supporting systems for lexicographical and terminological tasks are little used in the Canadian translation industry. For the most part, specific software is developed for each project, which suggests that, in general, off-the-shelf tools are not widespread in lexicographical and terminological research. It is also important to observe that usually this type of commercial software is expensive.

It is desirable that these tools be capable of managing databases, since fast access to them increases the productivity of entry writers. An example of a system with this functionality is System Quirk (Ahmad, 1994), which is divided into modules and has a Browser/Refiner that manages terminological databases. The tool Corplex (Simonsen, 2005) focuses on the management of entries and offers resources to support corporate lexicographical tasks (developed in companies and organizations). Its searching device stands out among these resources. For Portuguese, there is Corpógrafo[18], which permits corpus creation and processing, terminology extraction, and terminological database management with semantic and ontological relations. At this time, the environment e-Termos[19] (Almeida, 2006) is being developed. It is a collaborative web platform for supporting the creation of terminological products.

To create the *Historical Dictionary of Brazilian Portuguese*, we developed the tool Procorph (Candido Jr, 2008b). At the start of this project, entries were being written in MS Word, a poor solution for the task, because a writer does not have access to the entries being written by the others (entries are not distributed to prevent synchronization problems) and entry formatting is not automatic. Another problem is related to variations in entry form and contents, making it difficult to standardize

18 http://poloclup.linguateca.pt/ferramentas/gc/.
19 http://www.etermos.ufscar.br/index.php.

them. The name Procorph relates to corpus processing, since this is one of the tasks for which it was designed, as well as dictionary writing. To the best of our knowledge, this work is the only one in the area dedicated to supporting the construction of historical dictionaries in Portuguese. Procorph is used via the web, which makes access simpler for the project team, besides allowing entry standardization. The advantages of creating a web system are centralized data storage and sharing of entries among writers.

Some difficulties faced by lexicographers during entry writing motivated us to develop this tool. These were the main difficulties we faced: formatting problems, absence of a system to simplify references to sample texts, absence of a system for centralizing entries written by different lexicographers simultaneously. With historical dictionaries, extra difficulties arise, such as searching for spelling variants in entries and managing the dating in sample sentences. Besides simplifying the tasks performed by lexicographers, this system can also be used by the general public. In Correia (2008), the author emphasizes the consensus in the field of computational terminology and lexicography with respect to the fact that machine-usable dictionaries are much more efficient than their printed counterparts.

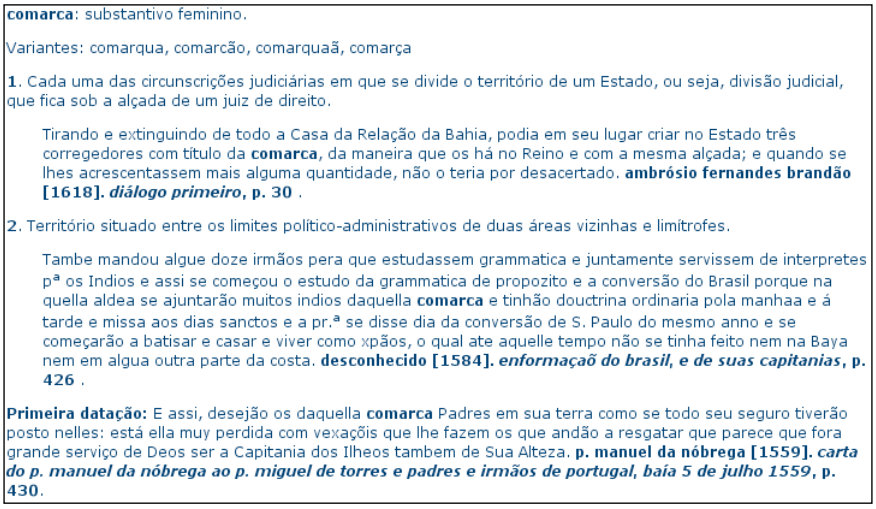
One of the objectives of developing this tool was making it capable of treating historical databases in general, so that it could be used in other projects to build historical dictionaries with minor adaptations. It is also possible to modify this tool to create systems focused on contemporary language dictionaries, since Procorph is free software, available under GPL[20] (General Public License). The program and its source code are publicly available[21], at no additional cost, and modifications are freely permitted. This system has a web interface, developed in PHP (PHP Hypertext PreProcessor), using the database MySQL The use of Javascript provided a more dynamic and simpler interface for editing entries.

The two main system screens provide entry searching and editing. Information stored in the database for each entry includes part of speech, gender and inflection, different meanings/definitions (or acceptations), related entries, observations, and sub-entries. Each definition is followed by an example sentence (an excerpt of a text from the corpus in which the entry is an example of the definition under consideration), as well as a reference to the text from which the definition was retrieved. The reference comprises the page on which the excerpt appears and the text code. Using this code, it is possible to obtain the title, the year of publication, and the author's name, generating references in a format similar to ABNT's (Associação Brasileira de Normas Técnicas – Brazilian Association of Technical Standards). Other system screens are the screen for listing texts used to collect sample sentences, the screen for searching spelling variants, and that for controlling users (only for users with administrative privileges).

20 http://www.gnu.org/licenses/gpl.txt.
21 http://www.nilc.icmc.usp.br/nilc/projects/hpc/.

In addition to information common to contemporary dictionaries, this system also allows the insertion of spelling variants. Abbreviations can be used together with spelling variants if the entry writer so wishes. Another specific resource for historical dictionaries is the control of the earliest date of an entry, which provides useful information for estimating the approximate period when the word started to be used in Portuguese. Each entry can be followed by sub-entries, which are complete entries (with the same attributes as the main entries) associated with a main entry and usually consisting of complex lexias. For instance, for the entry "mulher" (woman), sub-entries include "mulher do reino" (woman from Portugal), "mulher ama" (wet-nurse), "mulher moça" (maiden), and "mulher da terra" (native woman from Brazil). Figure 7 shows the entry "comarca" (district), created with Procorph.

Entries are stored in Unicode, which – as previously discussed – is capable of representing all symbols found in historical texts. However, it is not possible to keyboard some of these symbols using Brazilian keyboards. A feasible solution is to use programs such as the Character Mapping tool, available on Microsoft Windows. However, this is not convenient, due to the difficulty of locating the desired characters. Procorph's answer involves the use of character sets to denote those Unicode symbols difficult to keyboard. The advantage is the ease of converting character sets into their respective symbols automatically. Table 5 shows character strings and their respective symbols.



comarca: substantivo feminino.

Variantes: comarqua, comarcão, comarquaã, comarça

1. Cada uma das circunscrições judiciárias em que se divide o território de um Estado, ou seja, divisão judicial, que fica sob a alçada de um juiz de direito.

Tirando e extinguindo de todo a Casa da Relação da Bahia, podia em seu lugar criar no Estado três corregedores com título da **comarca**, da maneira que os há no Reino e com a mesma alçada; e quando se lhes acrescentassem mais alguma quantidade, não o teria por desacertado. **ambrósio fernandes brandão [1618]. *diálogo primeiro*, p. 30** .

2. Território situado entre os limites político-administrativos de duas áreas vizinhas e limítrofes.

Tambe mandou algue doze irmãos pera que estudassem grammatica e juntamente servissem de interpretes pª os Indios e assi se começou o estudo da grammatica de propozito e a conversão do Brasil porque na quella aldea se ajuntarão muitos indios daquella **comarca** e tinhão douctrina ordinaria pola manhaa e á tarde e missa aos dias sanctos e a pr.ª se disse dia da conversão de S. Paulo do mesmo anno e se começarão a batisar e casar e viver como xpãos, o qual ate aquelle tempo não se tinha feito nem na Baya nem em algua outra parte da costa. **desconhecido [1584]. *enformaçaõ do brasil, e de suas capitanias*, p. 426** .

Primeira datação: E assi, desejão os daquella **comarca** Padres em sua terra como se todo seu seguro tiverão posto nelles: está ella muy perdida com vexaçõis que lhe fazem os que andão a resgatar que parece que fora grande serviço de Deos ser a Capitania dos Ilheos tambem de Sua Alteza. **p. manuel da nóbrega [1559]. *carta do p. manuel da nóbrega ao p. miguel de torres e padres e irmãos de portugal, baía 5 de julho 1559*, p. 430**.

Figure 7. *Example of entry created with Procorph*

The different spelling variants of a certain entry are gathered while entries are being written. This is useful for selecting the most relevant example sentences in the dictionary and informing dictionary users about the different spellings they will find when looking up historical texts. The number of variants can be large (especially in

the sixteenth century). An example is the entry “prejuízo” (loss), which has ten known variants (prejuizo, preiuizo, preioizo, preijuiso, preyuizo, preyoizo, prejoizo, prejuiso, perjuizo, prejuiſo). However, it is difficult to perform a manual selection of spelling variants in the corpus. To alleviate this problem, Procorph has a glossary of spelling variants found automatically. Since the construction of the glossary was automatic and can have errors, the variants are not inserted automatically during entry writing. They need to be analyzed by writers beforehand. Moreover, the process of gathering variants automatically is not capable of detecting all possible variants for a certain entry.

Table 5. *Conversion of strings into Unicode*

Original	**Converted**
grati{ae}	gratiæ
{f}eito	ƒeito
c{oe}teris	cœteris
dis{s}cur{s}o	diſcurſo
{F-inv}ixit	ⅎixit
passad{a-inv}	passadɐ
Quar{circ}y	quarŷ
co{til}mércio	comércio
Caca{macron}o	cacaō
mu{trema}y	Mui
s{gancho}omente	sómente
tinha{virgule}o	tinhaȯ
{anel}Afonso	Åfonso
Quae{agudo}s	quaeś
apanh{breve}e	apanhĕ

A glossary of variants and junctions was included in the tool as well, as seen in Figure 8.

Figure 8. *Variants, abbreviations and junctions of “vila” (village)*

During the HDBP project, entry formatting was carried out by writers. However, this is a time-consuming task that impacts productivity negatively. Procorph solved this problem by formatting entries automatically. Another of its benefits is the low cost of applying formatting changes to all entries simultaneously, as well as the possibility of generating different versions of a dictionary. It is possible, for instance, to modify the system so that it generates an unabridged version and an abridged version in which sample sentences are removed for space reasons. Additionally, this tool provides automatic entry conversion into Microsoft Word, which has been widely used in the HDBP project.

As entries are available on the web, it is necessary to control users' access to the database and rights to write entries. In Procorph, there are four levels of access: user, writer, reviser, and administrator. Users can only navigate the database and look up entries, texts, and variants. Writers have permission to create entries and modify their own entries. Revisers have unrestricted access to the database and can alter entries written by anyone. Administrators have the same privileges as revisers and can also control the users registered in the database.

This system has not been implemented in the HDBP project yet, although it has been introduced to entry writers at a project meeting and lately tested by four of the 21 writers. According to their evaluation, the system had an excellent performance. However, we consider that a test including all lexicographers has higher potential for showing its limitations and suggesting improvements.

6. Computational environment for processing historical corpora

Our model for processing historical corpora was conceived based on experience acquired during the HDBP project. We focused on lexicographical activities, but this model suits several purposes in processing historical corpora in Portuguese. The environment is composed of modules that provide access to different corpus processing tools. The advantage of using modules is in how simple it is to add new resources to the environment, to replace inadequately functioning modules, and to customize modules for other corpus projects. Modules can be grouped in two architectures: corpus processing and glossary building.

The architecture for compiling corpora and building glossaries is composed of six modules, described as follows. The **cleaning and annotating module** removes undesired metadata from the text and annotates useful metadata. In a lexicographical task, examples of undesired metadata are footnotes and line numbers (useless information for users). Some structures, such as page numbers, chapter titles, and section subtitles, must be kept with appropriate annotation, since they provide useful information about texts. In the HDBP project, Protew and Protej perform the cleaning and annotating tasks. These tools were described in Section 2.2. After cleaning texts, digitizing (or keyboarding) errors may be found in the corpus. The **error detection module** analyzes patterns and is able to find automatically the most

frequent types of error. For instance, searching for words containing "1" and "0" can reveal digitizing mistakes, since the presence of these numerals in words is usually associated with failures in the recognition of characters "I", "L" or "O". The tool Siaconf searches for unknown symbols in the corpus to detect OCR errors. The **abbreviation extraction module** can be used to build glossaries of abbreviations from simple heuristics. Abbreviations can also be obtained from dictionaries of abbreviations such as Flexor (1991). In the HDBP corpus, the tool Protej pre-processes abbreviations and converts them into DELAF formalism, used in Unitex. Metadata are extracted by the **metadata extraction module,** and then can be included in different corpus versions. As each corpus processor provides different annotation patterns at different structural and linguistic levels, it is necessary to convert annotated texts into different formats. For this purpose, we developed the **version generation module**. Conversion between patterns can be made by means of the transformation language XSLT or by programs developed specifically for format conversion. A simple case of conversion is removing all XML structure to be used in tools that do not permit annotation. Format conversion increases corpus reusability. Finally, the **spelling variant extraction module** generates a spelling variant glossary based on edit distance techniques, phonetic analysis, and/or transformation rules. In the HDBP corpus, this task is performed by Siaconf, a tool described in Section 4.1.

The architecture for corpus access is based on the web environment and has the advantage of centralizing data storage, a feature typical of client-server systems. With an integrated environment, it is possible to guarantee that all researchers are working with the same database. Centralized data prevent inconsistencies in the database, since all researchers will have access to the most recently updated version of corpora and glossaries. Centralized data also minimize the cost of equipment necessary to process corpora. Workstations with modest configurations are satisfactory, given that servers perform most of the processing. Besides, many users are familiar with web interfaces, allowing for an environment of fast learning.

Basically, the architecture for corpus access provides modules grouped in three categories: corpus access, entry writing, and glossary access. In the HDBP project, Philologic and Unitex provide access to the corpus. The **modules to access glossaries** provide searches in abbreviation and spelling variant glossaries. For lexicographical searches for verbs (such as those in the HDBP project), a contemporary glossary can be used as a filter, allowing the identification of spelling variants and the detection of words that have fallen into disuse. This task is being performed with the help of Unitex. The **modules for writing entries** are the most specific in this architecture, since they apply only to lexicographical (or terminological) research. These modules allow users to insert entries in the database, as well as definitions, sample sentences, and references to the corpus. This task is being performed with the help of Procorph. A specific dictionary entry model has been defined for the HDBP project.

This architecture is divided into two parts: client and server. The client part is composed of modules implemented by scripts (programs executed on a web browser) that present and format data (lexias, concordances, and entries). The server part is composed of modules developed from heterogeneous technology. Modules for accessing corpora are connected to modules to access glossaries. Thus, it is possible to expand users' searches. For instance, users can search for all spelling variants and all abbreviations of a word. Likewise, modules for writing entries can use the services provided by modules for corpus access, simplifying the processes of finding sample sentences and reference to corpora.

7. Conclusions

In this paper, we have described the work of compiling a historical corpus to support the building of the HDBP, its challenges, and the solutions adopted for using this corpus for lexicographical tasks.

This work was motivated by issues in the treatment of historical corpora that arose during the HDBP project. Four tasks were identified: (a) compiling a historical corpus of Brazilian Portuguese, (b) building glossaries to support lexicographical tasks, (c) accessing the corpus, and (d) writing entries. This contribution encompasses methodology, resources (glossaries and the corpus as a whole), tools developed to process resources, and a tool developed to write dictionary entries, the main objective of the HDBP project. To the best of our knowledge, this work brings about innovations of a different nature and is the only one to offer a computational environment comprising all functionalities described in this paper to treat historical corpora. An additional contribution is the comparison among corpus processors, which can be useful for corpus linguistics researchers by informing their choice of tools. These contributions are freely available for use in other projects (except for material under copyright law).

Our glossaries are ready for reuse in historical corpus projects dealing with the Portuguese language. For other languages, the methodology employed to create the glossary of spelling variants can be adapted to extract patterns in a relatively direct way. The simple heuristics for extracting abbreviations automatically from the corpus can also be reused, in spite of a limitation in the methodology that does not allow for the automatic expansion of abbreviations. The methodologies for corpus compilation and entry writing became dependent on the tools employed. These tools, in turn, have functionalities specific to the needs of the HDBP project, as in the case of the information adopted in the TEI header. However, functions for dealing with page headers and footers and other cleaning operations on digitized texts can be reused on corpora in any language. We consider that other projects can reuse parts of the functionalities provided by our tools, with no need to develop new software. They are open source tools, and can even be adapted to the needs of each different project.

We have not been able to validate our methodology on other projects yet, because the low number of projects with characteristics similar to HDBP dealing with Portuguese imposes serious restrictions on validating the proposed methodology. Besides the HDBP, there are no other projects for creating dictionaries in Brazilian Portuguese as yet. To build historical corpora, only two (PROHPOR and Tycho Brahe) of the four projects cited in Section 1 are still being developed, and could benefit from our research. We are setting up a partnership with a project for creating the USP Brasiliana Digital Library[22] aiming at using our glossary of spelling variants, since our university (University of São Paulo – USP) is the guardian of one of the largest *brasilianas* in Brazil. We believe that in the future there will be more projects for building historical Portuguese corpora and dictionaries, which will make the validation of our methodology easier.

Concerning the evaluation of the creation of the glossary of spelling variants, we used traditional metrics in the field of Information Retrieval (precision and comparative recall). To evaluate the coverage of the dictionary of abbreviations that was digitized (Flexor, 1991), we developed an intersection of this dictionary and the corpus (only 16% are part of the HDBP corpus). In addition, we carried out an intersection of abbreviations extracted automatically from the corpus and the dictionary of abbreviations to estimate the amount of abbreviations not in the Flexor dictionary. We found 7,045 abbreviations with the heuristics; only 35% of the total (2,473) were in the Flexor dictionary, which is still useful, since it permits expanding abbreviations.

We observed that building the corpus and the dictionary is a huge task that demands effort and integration from many researchers. The work presented here was possible thanks to the help of the many participants of the HDBP project. Future work will include improving the proposed environment to identify more spelling variants; gathering more abbreviations; carrying out tests with users applying HCI (human-computer interaction) techniques; and extracting corpus metadata automatically through machine learning techniques.

In sum, the Historical Dictionary of Brazilian Portuguese is not only a pioneering project, but also a fundamental tool for recapturing and registering the country's early history through its vocabulary. Compiling a corpus of historical texts has been crucial for achieving this goal, allowing researchers to retrieve the lexicon of a given period. The lexical, morphological, syntactic, and typographic information gathered in these texts is being investigated by several members of our team, composed of philologists, linguists, and computer scientists. Historical texts have their peculiarities, and among them abbreviations pose a special challenge for researchers, since they are highly frequent and ambiguous. Researchers have also to face the fact that in historical texts there are no standard graphic forms, and abbreviations reflect this inconsistency, displaying a large number of variations.

22 http://www.brasiliana.usp.br/brasiliana_antiga/.

8. Acknowledgments

The authors thank CNPq (Brazil) for funding this research. Also, the authors would like to thank all the HDBP project members for their valuable work on corpus compilation and dictionary creation.

References

Ahmad, K. Language engineering and the processing of specialist terminology. In *The Language Engineering Convention/Journées du Genie Linguistique*. Paris, France: European Network in Language and Speech (ELSNET), 1994.

Almeida, G. M. B., Oliveira, L. H. M., Aluísio, S. M. A terminologia na era da informática. *Ciência e Cultura* (SBPC), v. 58, p.42–45, 2006.

Aluísio, S., Pinheiro, G. M., Manfrim, A. M. P, Oliveira, L. H. M. de, L. C. Genoves Jr., Tagnin, S. E. O. The Lácio-Web: Corpora and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In *Proceedings of LREC 2004*, Lisbon, Portugal, p. 1779-1782, 2004.

Archer, D., Ernst-Gerlach A., Kempken S., Pilz T., Rayson P. The identification of spelling variants in English and German historical texts: manual or automatic. In *Digital Humanities*, 2006, Paris: Sorbonne, p. 3-5, 2006.

Braun, L. Information retrieval from Dutch historical corpora (Master's thesis). Maastricht, Netherlands: Maastricht University, 2002.

Candido Jr, A. Criação de um ambiente para o processamento de corpus de Português Histórico. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, 131 p., 2008. Available at: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-21052008-103237/ (accessed: 20 February 2009).

Candido Jr, A., Aluísio, S. M. Um Ambiente Computacional para o Processamento de corpus de Português Histórico. In *IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence (WTDIA 2008) & VI Best MSc Dissertation/PhD Thesis Contest (CTDIA 2008),* Bahia, Brazil, p. 1-10, 2008.

Candido Jr, A., Aluísio, S. M.. Procorph: um Sistema de Apoio à Criação de Dicionários Históricos. In *VI Workshop em Tecnologias da Informação e da Linguagem Humana (TIL 2008)*, Vilha Velha, v. 1. p. 1-6, 2008.

Correia, M. Terminologia e lexicografia computacional. Available at: http://www.realiter.net/spip.php?article787#nb1 (accessed: 20 February 2009).

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

Dannélls, D. Automatic Acronym Recognition. In *EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics.* p. 167–170, 2006.

Dictionary of the Scots Language. Dictionary of the Scots Language, 2009. Available at: http://www.dsl.ac.uk/ (accessed: 20 February 2009).

Dipper, S., Faulstich, L., Leser, U., Ludeling, A. Challenges in modelling a richly annotated diachronic corpus of German. In *Workshop on XML-based Richly Annotated Corpora*. p. 21-29, 2004.

EAGLES DOCUMENT EAG-EWG-PR.2. Evaluation of Natural Language Processing Systems. 1995. Available at: http://www.issco.unige.ch/en/research/projects/ewg95/ (Accessed: 22 February 2009).

Ernst-Gerlach, A., Pilz. T. Search methods for documents in non-standard spelling. Talk presented at Historical Text Mining Workshop, Lancaster University, UK, 2006. Available at http://ucrel.lancs.ac.uk/events/htm06/PilzErnstGerlachHTM06.pdf (accessed: 20 February 2009).

Flexor, M. H. O. *Abreviaturas: Manuscritos dos séculos XVI ao XIX*. 2. UNESP: Brazil, 468 p., 1991.

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

Greghi, J. G., Martins, R. T., Nunes, M. G. V. Diadorim: a Lexical database for Brazilian Portuguese. In *International Conference on Language Resources and Evaluation LREC 2002*, v. IV, p. 1346–1350, 2002.

Haddad, R. Survey of the Canadian Translation Industry. Moncton, Canada: Canadian Translation Industry Sectoral Committee (Technical report), 1999.

Hirohashi, A. S. Aprendizado de regras de substituição para normatização de textos históricos. Master Thesis – Instituto de Matemática e Estatística, USP, São Paulo, Brazil, 2004.

Hong, Y., Hripcsak, G., Friedman, C. Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc*. May–June, 9(3), p. 262–272, 2002.

Kerner, Yaakov HaCohen, Ariel Kass, Ariel Peretz. Baseline Methods for Automatic Disambiguation of Abbreviations in Jewish Law Documents. In *EsTAL: International Conference on Advances in Natural Language Processing* N. 4, Alicante. Lecture Notes in Computer Science. Berlin: Springer, p. 58–69, 2004.

Lighter, J. E., O'Connor, J., Ball, J.. *Historical Dictionary of American Slang*. Random House. 1994.

Machado Filho, A. V. L. Projeto Deparc (Dicionário Etimológico do português arcaico). In *IV Congresso Internacional da Abralin*, Brasília, 2005.

Mahoney, T. J. Historical Astrolexicography and Old Publications. *ASP Conference Series*, Vol. 153, 1998.

McEnery, Tony, Wilson, Andrew. *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press, 2001.

Menegatti, T. A. Regras Lingüísticas para Tratamento Computacional da Variação de Grafia e Abreviaturas do Corpus Tycho Brahe (Technical report). Campinas, Brazil, 2002..

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

Paixão de Sousa, M. C., Trippel, T. 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)*, p. 1831-1836. Genoa: ELRA, 2006.

Pakhomov, S. Semi-supervised Maximum Entropy-based Approach to Acronym and Abbreviation Normalization in Medical Texts. In *Medical Texts Proceedings of ACL 2002.* p. 160–167, 2002.

Paumier, S. Manuel d'utilisation du logiciel Unitex. IGM, Université Marne-la-Vallée, 2006. Available on-line http://www-igm.univ-mlv.fr/~unitex/ManuelUnitex.pdf (accessed 20 February 2009).

Pind, J., Bjarnadóttir, K., Jónsson, J, H,, Kvaran, G., Magnússon, F., Svavarsdóttir, A. Using a Computer Corpus to Supplement a Citation Collection for a Historical Dictionary. *International Journal of Lexicography.* Oxford Press, 1993.

Hauser, A., Heller, M., Leiss, E., Schulz, K. U.,. Wanzeck, C. 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)*, p. 147-154. Hyderabad, India, 2007.

Rayson, P. E. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD Thesis, Lancaster University, September 2002.

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

Ruiz, J. R C., Martínez, M. G. Recopilación y estructuración del vocabulario de especialidad en el Nuevo Diccionario Histórico del Español (RAE). *XIII Congreso Internacional Euralex*, 2008.

Rydberg-Cox, J. A. Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries. In *Joint Conference on Digital Libraries.* Houston, USA: IEEE Press. v. 3, p. 372–373, 2003.

Sanderson, R. "Historical Text Mining," Historical "Text Mining" and "Historical Text" Mining: Challenges and Opportunities. Talk presented at Historical Text Mining Workshop, Lancaster University, UK, 2006. Available at http://ucrel.lancs.ac.uk/events/htm06/RobSandersonHTM06.pdf (accessed 20 February 2009).

Santos, D., Ranchhod, E. Ambientes de processamento de corpora em português: comparação entre dois sistemas. In *PROPOR'99*, Evora, 2002.

Schulze, B. M. *et al.* Comparative State-of-the-Art Survey and Assessment Study of General Interest Corpus-oriented Tools. Institut fur maschinelle Sprachverarbeitung, Universitat Stuttgart, 1994.

Schwartz, A. M., Hearst, M.. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Pacific Symposium on Biocomputing (PSB)*, p. 45–462, 2003.

Simonsen, Henrik K. CorpLex: Blueprints of a Corporate Dictionary and Editing System. In *Studies in Contrastive Linguistics*, Santiago de Compostela, Universidade de Santiago de Compostela, p. 453-460, September 2005.

Tei Consortium. *The TEI Guidelines*. Text Encoding Initiative Consortium, 2006. Available at http://www.tei-c.org/Guidelines2/ (accessed 20 February 2009).

Terada, A., Tokunaga, T., Tanaka, H. Automatic expansion of abbreviations by using context and character information. *Inf. Process. Management*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, no. 1, p. 31–45, 2004. Universität Leipzig. Terminology Management. 2008. Available at: http://www.uni-leipzig.de/~xlatio/software/soft-termiman.htm (accessed 20 February 2009).

University of Chicago. Philologic User Manual, 2008. Available at: http://philologic.uchicago.edu/manual.php (accessed 20 February 2009).

Vale, O. A., Candido Jr, A., Muniz, M. C. M., Bengtson, C. G., Cucatto, L. A., Almeida, G. M. B., Batista, A., Parreira, M. C., Biderman, M. T., Aluísio, S. M. Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. In *LATECH* 2008. Paris: ELRA, v. 1. p. 1-10, 2008.

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

Xavier, M. F. O Percurso Diacrónico dos Modais e Semimodais em Português e em Inglês e as suas Gramáticas. In *I Simpósio Mundial de Estudos de Língua Portuguesa.* São Paulo, 2008.