

Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora

¹Oto Vale, ²Arnaldo Candido Jr.*, ²Marcelo Muniz, ¹Clarissa Bengtson, ¹Lívia Cucatto, ¹Gladis Almeida, ⁴Abner Batista, ⁴Maria C. Parreira, ³Maria Tereza Biderman, ²Sandra Aluísio

Federal University of São Carlos (UFSCar), University of São Paulo (USP), State University of São Paulo (UNESP)

¹ UFSCar - Department of Linguistics - Via Washington Luís, Km. 235 – C.P. 676 - CEP 13.565-905 - São Carlos, SP Brazil

² USP - Centre of Computational Linguistics (NILC)/ Department of Computer Sciences

C.P. 668 - CEP: 13560-970 - São Carlos, SP, Brazil

³ UNESP, FCLAR - Department of Linguistics - Rod. Araraquara - Jaú Km1, Bairro dos Machados, C.P. 174, CEP: 14.800-901, Araraquara, SP, Brazil

⁴ UNESP, IBILCE - Department of Linguistics - Rua Cristóvão Colombo, 2265, Bairro: Jardim Nazareth, CEP: 15054-000 São José do Rio Preto, SP, Brazil

E-mail: otovale@gmail.com, arnaldoc@icmc.usp.br, marcelo.muniz@gmail.com, clabengtson@hotmail.com, liviacucatto@yahoo.com.br, gladis.mba@gmail.com, abnerfortunato@gmail.com, mcparrreira2002@yahoo.com.br, mtbider@attglobal.net, sandra@icmc.usp.br

Abstract

Abbreviated forms offer a special challenge in a historical corpus, since they show graphic variations, besides being frequent and ambiguous. The purpose of this paper is to present the process of building a large dictionary of historical Portuguese abbreviations, whose entries include the abbreviation and its expansion, as well as morphosyntactic and semantic information (a predefined set of named entities – NEs). This process has been carried out in a hybrid fashion that uses linguistic resources (such as a printed dictionary and lists of abbreviations) and abbreviations extracted from the *Historical Dictionary of Brazilian Portuguese (HDBP)* corpus via finite-state automata and regular expressions. Besides being useful to disambiguate the abbreviations found in the *HDBP* corpus, this dictionary can be used in other projects and tasks, mainly NE recognition.

1. Introduction

The *Historical Dictionary of Brazilian Portuguese (HDBP)*, the first of its kind, is based on a corpus of Brazilian Portuguese texts from the sixteenth through the eighteenth centuries (including some texts from the beginning of the nineteenth century). The HDBP is a three-year project, which started in 2006, developed under the sponsorship of CNPq, Brazil. Organizing this historical dictionary has required an extensive, time-consuming analysis of documents, published texts and manuscripts produced by eyewitnesses in the early stages of Brazilian history. One important difficulty in compiling this corpus derived from the absence of press agencies in colonial Brazil, which had a precarious communication system. Only in 1808, after escaping from Napoleon's army, did the Portuguese monarchy transfer the government of the Portuguese empire to Brazil and improved communications. Moreover, peculiarities affecting language must be considered, such as biodiversity and multifaceted cultural traditions from different regions of the country. To implement the *HDBP* project, we created an integrated network of researchers from various regions of Brazil and Portugal, including linguists and computer scientists from 11 universities. Our team comprises 18 researchers holding a PhD, with complementary skills and expertise, and 23 graduate and undergraduate students.

This project fills a gap in Brazilian culture, for it is developing a dictionary that describes the vocabulary of Brazilian Portuguese in the beginning of the country's history. At that time, Brazilian language was still dependent on European Portuguese, even though some vocabulary was

already being coined on this side of the Atlantic. On the one hand, the speakers of those days faced a world materially and culturally different from what was known in Europe; therefore, they needed to designate referents of this new universe, which were hitherto unnamed, using words from 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, customs and institutions gradually began to form in this new society with the infusion of new cultures, resulting in new words, different from those used in the Portuguese metropolis.

To build the corpus, we collected documents in public archives and libraries all over Brazil and in Portugal. This corpus totals 2,458 texts; 287,570 sentences; 16,505,808 tokens (of which 368,850 are different from each other); 7,492,473 simple forms¹ (of which 368,529 are different from each other); and 82.2 MB. In a similar endeavor related to European Portuguese, researchers of the Universidade Nova de Lisboa have built the "Corpus Informatizado do Português Medieval"², comprising Latin-Romance texts from the ninth to the twelfth centuries, and Portuguese texts from the twelfth to the sixteenth centuries, totaling some 2 million words. Our corpus was built to be processed with corpus processing system UNITEX³ (Unicode – UTF-16)

* Scholarship CNPq, Brazil.

¹This is the total number of words in the corpus that are composed of letters belonging to Historical Portuguese alphabet.

² Digital Corpus of Medieval Portuguese: <http://cipm.fcsh.unl.pt>.

³ <http://www-igm.univ-mlv.fr/~unitex/>

and with Philologic⁴ (Unicode – UTF-8), since the latter is web-based and includes several corpus-processing tools, as for example AGREP⁵, used to check for similar or alternative spellings in Philologic. To process this large corpus, we have faced the typical problems researchers are likely to encounter when dealing with old documents, starting with text digitalization. J. Rydberg-Cox (2003) and R. Sanderson (2006) state that, in historical Latin, Greek and English texts, to mention just a few languages, words broken at the end of a line 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 pervade non-abbreviated words; and spelling variation is common even within the same text. We encountered these same problems in the *HDBP* project. First of all, the non-existence of an orthographical system in the afore-mentioned centuries generated a Babel of graphic systems being used by many different scribes or copyists. Giusti et al. (2007) focus on this difficulty introducing both an approach based on transformation rules to cluster distinct spelling variations around a common form, which does not always correspond to the orthographic (or modern) form, and choices made to build a dictionary of spelling variants of Brazilian Portuguese based on such clusters. Another problem was scribes' habit of abbreviating words to facilitate handwriting – there are many thousands of such abbreviations. Therefore, for the correct understanding of texts, it was necessary to expand abbreviations, a task that presents two main difficulties. The first is related to the use of modern knowledge sources to perform expansion, since gazetteers, encyclopedias and heuristics currently in use do not address directly the needs of historical material describing people, places, and other entities that often do not appear in modern sources (Crane & Jones, 2006). The second, and perhaps most important, is that even if we had adequate knowledge sources for expanding abbreviations they are highly ambiguous with respect to meaning, which is critical for understanding correctly not only the abbreviations themselves but also the whole text (Kerner et al., 2004).

In general, if abbreviations are not expanded correctly they can limit the effectiveness of: i) information extraction and retrieval systems in digital libraries; ii) electronic index creation from a corpus; iii) Natural Language Processing (NLP) tools, such as taggers, parsers and named entity recognition (NER) systems to enrich corpora linguistically. Within the scope of the *HDBP* project, the failure of proper abbreviation expansion hinders the correct editing of dictionary entries. However, expanding each and every abbreviation manually in a several million-word corpus is a time-consuming, expensive and difficult – if not impossible – task, due to the inherent ambiguity of noun abbreviations, for example. In Section 4, we discuss our approaches to this problem. Automatic acronyms and abbreviations disambiguation have been given close attention in medical and biomedical domains, since text normalization is an important task for successful information retrieval and

extraction from texts in these areas (Pakhomov, 2002; Yu et al., 2002; Schwartz & Hearst, 2003; Dannélls, 2006). However, most of this automatic abbreviation disambiguation research has focused on modern scientific material, whereas historical corpora and digital libraries remain largely ignored (Rydberg-Cox, 2003). Moreover, NER systems have only begun to be implemented for digital libraries (Crane & Jones, 2006). Taking the above into consideration, the purpose of this paper is to present our ongoing work to build a large dictionary of abbreviations that contains the pair abbreviation and its expansion, together with morphosyntactic and semantic information. We are developing this process in a hybrid fashion, using linguistic resources (such as a digitalized printed dictionary of abbreviations and the authoritative lists of abbreviations that accompany the material that is being digitalized) and abbreviations extracted from the *HDBP* corpus via finite-state automata and regular expressions. Since expanding abbreviations is a costly process, these automata were created to recognize larger patterns of abbreviations that are NEs or the same pattern that has different types of NEs. For example, the dictionary semantic tags allowed us to identify new NEs, i.e., sequences of words that can be identified as personal names in some contexts, but that are categorized as place, river or organization names in other contexts. We have been working on an iterative process, which started with linguistic and common sense knowledge, to attribute initial NE categories to an abbreviation and later try to capture new NEs to update the respective dictionary entry. Both the new NE categories and the larger abbreviations gathered in this iterative process will be inserted in our dictionary of abbreviations, which will be useful for other projects and tasks, mainly named entity recognition and abbreviation disambiguation in the *HDBP* corpus. In the next section, we explain the details of the *HDBP* corpus and the graphic form of abbreviations found in it, as well as some historical corpus projects that addressed the same issue. In Section 3, we describe the process of building a dictionary of abbreviations. In Section 4, we consider possible applications of the dictionary in the *HDBP* project itself and in other scenarios: abbreviation lookup and expansion; search for spelling variations of abbreviations; and linguistic research. Section 5 contains our conclusions and final remarks.

2. The *HDBP* corpus and its abbreviations

The texts in the *HDBP* corpus were written by Brazilian authors and Portuguese authors who have lived in Brazil for a long time. Among the texts selected for our corpus, there are, for instance, letters of Jesuit missionaries, documents of the bandeirantes (members of the exploratory expeditions that pushed Brazilian borders far into inland areas), reports of sertanistas (explorers of Northeastern Brazil), and documents of the Inquisition. Table 1 shows more details about the composition of our corpus.

Since the emphasis is on word meaning, we have selected mainly published texts with minor editing. Examples of such editing are the separation of words that come together in the

⁴ <http://philologic.uchicago.edu/index.php>

⁵ <http://www.tgries.de/agrep/>

original text, the introduction of punctuation marks, paragraph mark-up to facilitate reading, and the insertion of letters and words in places where editors were sure (or almost) that such items were missing.

Data	XVI th	XVII th	XVIII th	XIX th
Texts (%)	6.24	26.39	59.78	7.59
Sentences (%)	6.30	18.32	64.34	11.04
Simple Forms (%)	7.60	20.18	62.57	9.65
Megabytes (%)	7.23	19.95	63.09	9.73

Table 1: Distribution of texts by century

This decision was made to avoid potential problems during corpus compilation; however, we still had to deal with the following issues: 1) guaranteeing consistent assignment of Unicode characters in the texts, since digitalization and OCR correction have been done by different groups geographically distant from one another; 2) treating the graphic variation that alters frequency counts in the corpus, thus causing difficulties for the selection of variants in dictionary entries; and 3) expanding the abbreviations that pervade the texts.

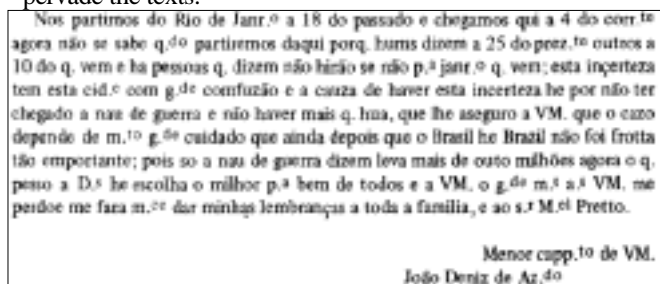


Figure 1: Excerpt from Cartas Remetidas para Lisboa em 12 de julho de 1715 In: Negócios coloniais, L. Lisanti (ed.)

There are several graphic forms for the abbreviations found in the *HDBP* corpus, some of which are shown in Figure 1:

a) abbreviations with a dot followed by a superscript piece of text, as in “Janr.o”/Janeiro (January) and “corre.te”/corrente (current), which appear in the first line;

b) abbreviations followed by a dot, as in “porq.”/porque (because) and in the three instances of “q.”/que (next/who/next, respectively) in the second and third lines. 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 processed computationally. The same symbol is used when the abbreviation does not possess a dot, but has a superscript chunk, as in “O sor Jesus xpo”/O Senhor Jesus Cristo (The Lord Jesus Christ), leading to 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 Gonçalves), “Jão”/João (proper name João), “Ido”/licenciado (licensed), “Ros”/Rodrigues (proper name Rodrigues), and “snr” or “snro”/senhor (sir).

An additional difficulty posed by abbreviations is that they hinder the correct segmentation of sentences in UNITEX (Friburger, 2002). In the historical corpus, this problem is

magnified by a large variation in the use of punctuation and capitalization, which also affects the recognition of named entities, since the corpus contains capitalized common words, as if they were proper names, and proper names in lower case. Table 2 illustrates problems related to abbreviations: ambiguity and variants. The first column shows 13 different expansions for the abbreviation “A”. The second column illustrates 13 different forms of abbreviating the name of the famous Brazilian city “Rio de Janeiro” (some of them in lower case), which makes 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

Table 2: Ambiguity and spelling variation in abbreviations

Most previous work on Brazilian Portuguese historical corpus expands abbreviations manually, such as the project “Para uma História do Português do Brasil”⁶ and “Projeto Programa para a História da Língua Portuguesa” (PROHPOR⁷). Also, the Tycho Brahe Project⁸ (Paixão de Sousa & Trippel, 2006), whose purpose is to model the relation between prosody and syntax in the process that led from Classical to Modern European Portuguese, contains tagged and parsed texts written by Portuguese authors born between the sixteenth and nineteenth centuries. These texts had their abbreviations expanded manually to facilitate tagging and parsing. Although this corpus is large for the task of syntactic analysis – it is currently composed of 46 texts and still growing – it remains manageable by manual markup made 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 abbreviations found in it will be expanded and annotated, based on well-accepted international standards in XML.

All projects mentioned above expand their abbreviations manually; however their development contexts differ from that of *HDBP*, which has only three years to develop both a large corpus and a dictionary. This is the reason why we had to approach the problem related to abbreviation expansion in

⁶ “For a History of Brazilian Portuguese”:

<http://www.lettras.ufrj.br/phpb-rj/>

⁷ “Project Program for a History of Portuguese Language”:

<http://www.prohpor.ufba.br/projetos.html>

⁸ <http://www.ime.usp.br/~tycho/>

a different way, detailed in Section 3.

3. Building a dictionary of abbreviations

In recent years, NLP researchers have focused on standardizing methods to construct linguistic resources, which led to the development of tools now accepted internationally. One of these construction standards, DELA (Dictionnaires électroniques du LADL), was developed at LADL (Laboratoire d'informatique documentaire et linguistique, University of Paris 7, France), jointly with the corpus-processing tool INTEX (Silberstein, 2000). DELA became the standard tool for developing electronic lexicons in the research network Relex⁹. These lexicons are used with INTEX, and now also with its open-source counterpart UNITEK (Paumier, 2006). This format allows for declaring simple and compound lexical entries, which can be associated with grammatical information and inflection rules. These dictionaries are linguistic resources specifically designed to perform automatic text processing operations. Types of DELA are DELAF, which comprises inflected simple words, DELAC and DELACF, for non-inflected and inflected compound words, respectively. The dictionaries of simple words (DELAS and DELAF) are lists of simple words associated with grammatical and inflectional information. The grammatical information is mainly morphological and corresponds to gender, number, degree, case, mood, tense, and person. However, with this format, it is possible to add syntactic and semantic information gradually (Ranchhod, 2001). DELAF lexical entries have the following general structure:

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

3.1 Customizing UNITEK

Processing lexicographical tasks in a corpus is easier when computational lexicons are available, and that was the reason why we adopted UNITEK in the HDBP project. UNITEK supports several languages, including Portuguese. Language-specific resources are grouped in packets referred to as idioms. When the UNITEK-PB (Muniz et al., 2005) was created, a lexicon for contemporary Brazilian Portuguese was incorporated into it. However, due to the peculiarities of historical texts, several changes had to be implemented and a new idiom was created, named “Português Histórico” (Historical Portuguese). These changes included characters that are no longer used in Portuguese, such as the long s (ſ) and the tilde (~) over consonants. Some diacritical marks differ from the ordinary diacritics currently used in Portuguese, because the former can be placed over consonants. For instance, an accent mark over “m̃” was common in Historic Portuguese. The introduction of such characters was made possible using Unicode when the text was being compiled.

3.2 The hybrid process to build a dictionary of abbreviations

3.2.1 Printed resources

In order to build our dictionary of abbreviations, we employed lexicons together with corpus processing tools,

especially to expand and enrich a digitalized printed dictionary (Flexor, 1991) 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. Although it has a large number of abbreviations (see Tables 3 and 4), most of them are not found in our corpus (only 16% appear in the HDBP corpus). We performed an experiment to recover abbreviations from the HDBP corpus using three simple heuristics, to estimate the amount of abbreviations in the corpus that is not present in the Flexor dictionary. We found out 7,045 abbreviations with the heuristics; only 35% of them (2,473) are in the Flexor dictionary. However, the Flexor dictionary is still worth using as it has abbreviations expansion. This dictionary is being revised to eliminate entries that could be considered spelling variants, as in the following example (pairs are composed of abbreviation and expansion): (Bês, bens); (Bêz, bens); (Bãda, banda), since the tilde was part of the writing system of historical Portuguese.

Simple and Multi-word Abbreviations by Century					
Types	XVIth	XVIIth	XVIIIth	XIXth	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
Coverage (%)	16.13				

Table 3: Abbreviations from Flexor (1991) by century, showing the % of forms found in the HDBP corpus¹⁰.

This hybrid approach to build a dictionary has already been successfully used to develop a dictionary of anthroponyms (Baptista, Batista and Mamede, 2006) and was adopted in the HDBP project as well. Besides, we have employed the authoritative lists of abbreviations found in the books we digitalized.

Simple and Multi-word Abbreviations by n-grams							
Types	1	2	3	4	5	6 or +	Total
Flexor	17,872	1,624	833	527	302	711	21,869
Flexor (%)	81.73	7.42	3.81	2.41	1.38	3.25	100.00
Intersection of Flexor and Corpus	3,237	234	33	18	5	2	3,529
Intersection of Flexor and Corpus (%)	91.75	6.60	0.94	0.51	0.14	0.06	100.00

Table 4: Abbreviations from Flexor (1991), by size

Thus far, we have digitalized and processed abbreviations from Flexor (1991) and some of the authoritative lists of

⁹ <http://ladl.univ-mlv.fr/Relex/introduction.html>

¹⁰ Note that abbreviations can happen in more than one century.

abbreviations to be used in the UNITEK system. Initially, the information we had to include in the entries, gathered from printed resources, was just the abbreviation, its expanded form, and the century in which the text had been written. However, considering information retrieval, we soon found out that the canonical form was also extremely important and should be in the dictionary, as it is required in the DELA format. Therefore, we added this information, and now a search for the canonical form capitão (captain), for instance, produces the following forms (nonexhaustive list): Capitão, capitam, Capitaõ, cappitão, Cappitam, capitães, Capitães, capitans and the abbreviated forms (nonexhaustive list):

Cap ^{acens}	Cap ^{ams}	Cap ^{ans}	Cap ^{ens}	Cap ^{im}
Cap ^{es}	Cap ^{ms}	Cap ^{ns}	Cap ^s	Capão
Cap ^{tens}	Cap ^{tes}	Capa ^{ens}	Capitt ^{es}	Capp.
Capm ^s	Capn ^{es}	Capn ^s	Capns	Capp ^{ão}
Capp ^{acens}	Capp ^{es}	Capp ^{tes}	Capt ^{es}	Capp ^{im}

Our dictionary of abbreviations differs from its counterparts developed in UNITEK, mainly in the use of a larger number of attributes. The most important attributes that have been added are: ABREV, used to denote abbreviation; SEC16, SEC17, SEC18, and SEC19 to indicate the century to which the lexical entry refers (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>, which is a collocation to extract certain types of NE. Each NE receives additional 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 (HAREM¹¹), organized by Linguatca. Among the ten HAREM categories, we have employed nine of them except OBRA (titles, man-made things). Figure 2 shows some lexical entries in DELA format. In the first line of the Figure 2, Brg^{es} is the form found in the corpus, Borges is the canonical form (lemma), N (noun) is the part-of-speech tag for the entry, ENT+PESSOA+ABREV+SEC19 are additional 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.

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/bacharel 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:ms/bastardos

Figure 2: Entry samples from the dictionary

We have already processed letters A, B, C and some of the authoritative lists of abbreviations. From the 3051 simple abbreviations under letter A, 814 are named entities (<ENT>)

and 548 have the tag <INIT>. 1789 were simple abbreviations. There are also 430 multi-word abbreviations in letter A. From the 488 simple abbreviations under letter B, 260 are named entities (<ENT>) and 138 have the tag <INIT>. Only 107 were common abbreviations. Some entries classified as <ENT> are <INIT> as well, such as “Barb^{ro}” (barber), a family name and a pattern used to introduce this profession. There are also 45 multi-word abbreviations in letter B, such as “Bn^s Ay^s” (Buenos Aires) and “Brigad^{ro} Insp^{or}” (Brigadeiro Inspetor/Inspector Brigadier). As for letter C, from the 2187 simple abbreviations, 364 are named entities and 853 have the tag <INIT>. There are also 510 multi-word abbreviations in letter C. All the multi-word abbreviations will be annotated later.

3.2.2 Generic patterns to extract different categories of NEs for an abbreviation

The use of heuristics is efficient for extending lexicons of NEs, such as in the search for words (or n-grams) that begin with a capital letter that is not in the beginning of a sentence or in the search for words followed by titles and forms of address. Thus, heuristic rules allow for the identification of named entities. However, the identification of some abbreviated NEs, such as “V. M.” (Vossa Mercê/archaic Portuguese for “you”), is difficult, because the dot that follows V makes the NE look like the beginning of a sentence (“. M.”), and therefore impossible to be retrieved using the heuristic rule mentioned above¹².

An experiment¹³ was performed to investigate NEs in the historical corpus, in order to extend and enrich the dictionary of abbreviations. This experiment was carried out with the dictionary of abbreviations described in Table 3. The three lists of abbreviations for letters A, B, and C and some short lists of abbreviations were first tagged with HAREM categories. The NEs received the tag <ENT>, whereas all entries received the tag <ABREV>. The tag <INIT> was created to designate abbreviated collocations found at the left of certain types of NEs, thus yielding three subcategories that were not present in HAREM, viz. <TITULO> (for jobs/professions and titles/positions), <PARENTE> (for family relations), and <TRATAMENTO> (for forms of address, since they are very pervasive in Flexor’s dictionary). Besides, all ten HAREM categories were used to subcategorize <INIT>.

First, using the dictionary of abbreviated forms, we performed a search in the corpus for tag <ABREV> (rule 1), which resulted in 1,795,519 occurrences. Several of them were not abbreviated forms, but stopwords with similar formats. In addition, several occurrences were actually orthographic variants that looked like abbreviations, such as bom/bõ. This prompted us to re-examine the list of abbreviations to remove non-abbreviated variants of stopwords and abbreviated forms of stopwords (we call this the pre-processing phase). For instance, the prepositions “por” and “para” were abbreviated as “p.”. However, this was also the abbreviation for padre (priest). Re-examining the list led us to create the rule 2 for searching forms in a UNITEK graph or using the following regular expression to locate retrieve abbreviated forms or the form “p.” preceded by determiners: <ABREV>+((o+ao+do+ho).p\.)

¹² Note that sentence breaking was not performed in the corpus preprocessing phase.

¹³ In this experiment, we have used UNITEK version 2.0 and set UNITEK to find the longest matches in its searches.

¹¹ http://poloxldb.linguatca.pt/harem.php?l=classificacao_v3_sem

With this regular expression, all abbreviated forms in the dictionary can be retrieved, plus the form “p.” preceded by determiners, thus decreasing the number of abbreviated forms. On applying the pre-processing cited above and the rule 2, the number of abbreviations dropped to 804,939. Before applying a search using tags, we tested the hypothesis that a significant number of abbreviated forms were either an NE or were in the vicinity of an NE. This test was carried out with the rule 3 depicted by:

```
((<ABREV>+((o+ao+do+ho).p\.).)(<MOT>+<MOT><MOT>+<MOT><MOT><MOT>+<MOT><MOT><MOT><MOT>))
```

applied to a search for abbreviated forms containing one to four elements. The number of retrieved abbreviations was 469,640. Therefore, further strategies are necessary to identify NEs, since we observed in our corpus that abbreviations tend to be close to each other. We carried out another search using tags, in which we replaced the tag <ABREV> by <INIT> (rule 4):

```
((<INIT>+((o+ao+do+ho).p\.).)(<MOT>+<MOT><MOT>+<MOT><MOT><MOT>+<MOT><MOT><MOT><MOT>))
```

With the rule 4, the number of occurrences dropped to 22,196. More than 50% were the abbreviated form “S.”, which stands for “Saint”, and abbreviated forms of address such as “S. M.” (Sua Majestade/His or Her Majesty), “S. A.” (Sua Alteza/His or Her Highness). The names of saints, however, were commonly other types of NE, not associated with PERSON. In fact, they were abbreviations for names of places (fazenda/farm, arrayal/hamlet, mosteiro/monastery, aldeia/village, bairro/district, villa/village etc.), rivers (Corgo/Brook, rio/river), organizations (mosteiro/monastery, fortaleza/fortress). We can check this information looking for such words at the left of the abbreviated form in the excerpt shown in Figure 3. This analysis of abbreviations productivity is useful for identifying and contextualizing new NEs. The use of these new attributes allows for sophisticated searches in the *HDBP* corpus and in other historical corpora. This is important, because we intend to make this resource available for research under request, since we cannot make it public due to copyright issues. It will be possible, for instance, to search for all NEs from the eighteenth century or for all NEs related to persons in their abbreviated forms.

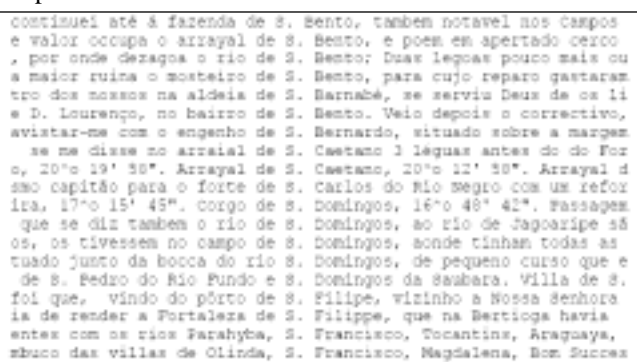


Figure 3: Concordances retrieved from sentences that have the abbreviated form “S.”

3.2.3. Specific patterns to extract new NEs of a given category

The use of generic search patterns in corpus processing tools shown in Section 3.2.2 is easily carried out by linguists and helps to get insight of the lexical patterning of historical

corpora. Moreover, it is more useful to retrieve ambiguous classes of NEs, as illustrated in Figure 4 for the name of saints, since all of them are under letter S. However, we believe it is not efficient to enlarge the dictionary with new and different entries of a given category and its subspecifications. To perform this focused task, we are applying the same process defined for REPENTINO¹⁴, a repository of NEs from modern Portuguese. This process is run in six steps, but the last one was not applied, for 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 the 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 have in their names 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 conduct the search; 4) validate manually the obtained candidates, considering the intended category; 5) include positive candidates in the repository; 6) if necessary, create a new category/subcategory, thus expanding the taxonomic classification system.

However, this process had to be adapted to historical corpora, because they have a large number of abbreviations and spelling variations related to both abbreviated words and expanded words. The requirement to accept a new NE from the corpus was that at least one of the components should be in the abbreviated form. We could not adopt the requirement of capitalization, since in historical corpora proper names are not always capitalized. 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 (1991) contains 18 entries with the pattern Rio XXX/River XXX, but eight of them refer to the city of Rio de Janeiro (the rest 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}); there is nothing about Creek XXX (or its variants, brooks, streams), so we began with ten entries. The chosen search strategies were: pattern formed by tag <INIT> and contexts “naveg*/navigate (on), that includes the several conjugations of the verb to navigate. However, words tagged as <INIT> could appear in their abbreviated or expanded form and, besides, we would have to deal with spelling variations and synonyms (see Table 5). To deal with spelling variations, we adopted two resources: the *HDBP* dictionary of spelling variants, created according to the SIACONF methodology proposed in Giusti et al. (2007)¹⁵, which employs 43 transformation rules to cluster variants under one orthographic form, and the Philologic resource of searching for similar patterns, which uses AGREP. The *HDBP* dictionary of spelling variants has 18,082 clusters, totaling 41,710 variants. In spite of producing false-positives, AGREP helps to complete variants resulting from

¹⁴ <http://poloclup.linguatca.pt/repentino/>

¹⁵ Available at <http://moodle.icmc.usp.br/dhpb/siaconf.tar.gz>

SIACONF. To deal with synonyms of river, we used the Brazilian Portuguese Electronic Thesaurus TEP (Gregghi et al., 2002).

Searching patterns (63)	Sources (5)	Right Occurrences (112)
rio	river	79
arroio, córrego, corrente, regato, regueira, regueiro, ribeirão, ribeiro, riacho, rio, veia, veio	synonyms	13
arroyo, corrego, corego, corgo regueyro, ribeiraõ, ribeyrão, ribeyraõ, rybeirão, rybeyrão, rebeirão, rebeyrão, ribeirão, ribeiro, ribeyro, rybeiro, ribejo, rjbeyro, rybeyro, riaxo, ryo, rjo, rio, veyra, veyo	spelling variants	7
c [^] te, cor, cor [^] e, cor [^] te, corr [^] e, corr [^] te, cort [^] e, crr [^] e, curr [^] te, r [^] bro, r [^] o, r [^] ro, reb [^] o, rib [^] ro, rib [^] o, rib [^] ro, riber [^] o, ribr, ribr [^] o, ryb [^] o, ryb [^] ro, rybr [^] o, r [^] bro, r, r [^] o	abbreviations from Flexor (1991)	11
naveg*	context	2

Table 5: Searching patterns for hydronyms

The manual validation is the slowest step (we checked 27,808 occurrences in 160 minutes – 1,100 checkings per hour), but easier in concordancers, since the pattern formed by abbreviations stands out, which facilitates checking. As a result of this case study we have now 122 abbreviations under category LOCAL, specifically rivers and words related to watercourses, displaying their morphology in this semantic group in the *HDBP* corpus. Some examples are: Ribeyrão de N. Sr.^a do Carmo/Ribeirão de Nossa Senhora do Carmo; Corgo de S. Gonçalves/Córrego de São Gonçalves; rib.^o do Tombadouro, ribeirão do Tombadouro; coRego Ant.^o da Silua, Córrego Antonio da Silva; Rio M.^{el} Alves, Rio Marechal Alves; R^o doce, Rio doce.

4. Applications of the dictionary of abbreviations

An example of use of the *HDBP* dictionary of abbreviations is the application of UNITEX together with the software Dicionário¹⁶ (Muniz et al., 2005) to assist lexicographers in manually identifying possible canonical forms for an abbreviation or for expanding an abbreviation. Using the concordancer shown in Figure 4, a linguist may find examples of abbreviations in the corpus, but may not be aware of the possible expanded forms for a given abbreviation. On using the software Dicionário together with the concordancer, the abbreviations can be quickly identified and associated with their possible canonical forms and

categories. In the context of information retrieval for historical documents, it may be useful to gather, for example, texts reporting certain facts that happened in a certain place. If the index is the expanded abbreviation, we can easily find all abbreviations for a word, such as Bahia (a Brazilian city in the northeastern coast of Brazil). Next, we can locate passages in the corpus related to those abbreviations.



Figure 4: Search for pattern “b” in UNITEX; the program Dicionário helps the manual disambiguation of the abbreviation “b” (in the top right corner). All subcategorization of NEs will be included later.

Our dictionary of abbreviations was designed to recognize large patterns of complete abbreviations. It also includes a specific tag for dealing with jobs/professions and titles and 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). In linguistic research, it is very important to know whom the text is talking about or whom it is talking to. If we can determine the authorities that are being addressed in a specific text, we can identify the words used in that specific level of formality, given that a letter written to an ordinary person does not contain the same words and level of formality as one written to a monarch, and this is possible because we used NEs and other specific tags.

5. Conclusions and Future Work

To sum up, the *Historical Dictionary of Brazilian Portuguese* is not only a pioneer project, but also a fundamental tool for recapturing and registering the country’s early history through its vocabulary. The compilation of a corpus of historical texts is therefore a crucial step to achieve such aim, since it allows researchers to retrieve the lexicon of a given period. The lexical, morphological, syntactic, and typographic characteristics identified in these texts have been the object of study of various members of our team, which includes philologists, linguists and computer scientists. Among the peculiarities of historical texts, the abbreviated forms pose a special challenge. In addition to their high frequency and ambiguity, a researcher is also faced with the fact that, as far as historical

¹⁶ The software Dicionário is a Java application that handles any dictionary compacted in the DELA format, and allows searching for inflected words.

documents are concerned, there are no standard graphic forms, and abbreviations reflect this inconsistency, displaying a large number of variations. Taking this fact into account and to make a lexicographer's task feasible, special attention was given to abbreviations. An electronic dictionary of abbreviated forms is being built based on printed resources, using the DELA format, which allows us to categorize each new entry morphosyntactically, semantically and pragmatically. New NE categories of abbreviations were found using semantically categorized abbreviations, UNITEX graphs and regular expressions to examine the vicinity of abbreviated forms. Since the process to expand abbreviations demands considerable expertise, these automata and regular expressions were created to recognize only larger patterns of abbreviations that are NEs, spelling variations and synonyms of NEs or the same pattern that has different types of NEs, given that context will provide meaning. With regard to enlarging the dictionary of abbreviations, we focused on a specific NE category (places), subcategorizing it further (hydronyms). This experiment provided us with evaluation data with regard to time spent and productiveness rate of the semi-automatic approach we decided to adopt to guarantee high accuracy for the classification process. We concluded that this approach is worth pursuing once we need to guarantee a high precision classification. In the future, we intend to make this corpus and the dictionary of abbreviations available for those studies on history to which correct NE classification is crucial and mainly as a resource for NE recognition systems.

6. Acknowledgments

The authors are grateful to CNPq (Brazil) for supporting this research.

7. References

- Baptista, Jorge; Fernando Batista; Nuno Mamede. 2006. Building a Dictionary of Anthroponyms. In Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, R.Vieira et al. (eds.) Proceedings. Lecture Notes in Computer Science 3960, Berlin: Springer, pp. 21--30.
- Crane, Gregory; Alison Jones. 2006. The challenge of Virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 31--40.
- Dannélls, Dana. Automatic Acronym Recognition. In EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.
- Dipper, Stefanie; Lukas Faulstich; Ulf Leser; Anke Ludeling. 2004. Challenges in modelling a richly annotated diachronic corpus of German. In Proceedings of the Workshop on XML-based Richly Annotated Corpora.
- Flexor, Maria Helena M. O. 1991. Abreviaturas - Manuscritos dos Séculos XVI Ao XIX. 2nd ed. São Paulo: UNESP. 468 p.
- Friburger, Nathalie. 2002. Reconnaissance automatique de noms propres: Application à la classification automatique de textes journalistiques. Thèse (doctorat). Université de Tours. Tours.
- Giusti, R.; Candido Jr, A.; Muniz, M.; Cucatto, L.; A. Aluísio, S. 2007. "Automatic detection of spelling variation in historical corpus: An application to build a Brazilian Portuguese spelling variants dictionary". In Proceedings of the Corpus Linguistics 2007 Conference, Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.).
- Greggi, J. G.; Martins, R. T.; Nunes, M. G. V. Diadorim: a Lexical database for Brazilian Portuguese In Proceedings of the International Conference on Language Resources and Evaluation LREC 2002, Manuel G. Rodríguez and Carmem P. S. Araujo (Eds.), v. IV, pp. 1346--1350.
- Kerner, Yaakov HaCohen; Ariel Kass; Ariel Peretz. 2004. 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. 3230. Berlin: Springer, pp. 58--69.
- Muniz, Marcelo C.M.; Maria das Graças V. Nunes; Eric Laporte. 2005. UNITEX-PB, a set of flexible language resources for Brazilian. In III Workshop em Tecnologia da Informação e da Linguagem Humana, pp. 2059--2068.
- Paixão de Sousa, Maria Clara; Thorsten Trippel. 2006. Metadata and XML standards at work: a corpus repository of historical Portuguese texts. In Proceedings of V International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.
- Pakhomov, Serguei. 2002. Semi-supervised Maximum Entropy-based Approach to Acronym and Abbreviation Normalization in Medical Texts. In Medical Texts Proceedings of ACL 2002. Philadelphia.
- Paumier, S. 2006. Manuel d'utilisation du logiciel UNITEX. IGM, Université Marne-la-Vallée. Available at <http://www-igm.univ-mlv.fr/~unitex/ManuelUnitex.pdf>
- Ranchhod, Elisabete M. 2001. O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais. In Tratamento das Línguas por Computador. Uma Introdução à Lingüística Computacional e suas Aplicações, E. Ranchhod (ed.), Lisbon: Caminho, pp. 13--47.
- Rydborg-Cox, Jeffrey A. 2003. Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries. In Proceedings of JCDL, 03, pp. 372--373.
- Sanderson, Robert. 2006. "Historical Text Mining", Historical "Text Mining" and "Historical Text" Mining: Challenges and Opportunities. Talk presented at the Historical Text Mining Workshop, July 2006, Lancaster University, UK.
- Schwartz, Ariel M.; Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical texts. In Proceedings of the Pacific Symposium on Biocomputing (PSB) 2003.
- Silberstein, Max. 2000. Intex: a FST toolbox. Theoretical Computer Science, 231, pp. 33--46.
- Yu, Hong; George Hripcsak; Carol Friedman. 2002. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc. May--June, 9(3), pp. 262--272.