Grammatical Annotation of Historical Portuguese: Generating a Corpus-based Diachronic Dictionary

Eckhard Bick[1], Marcos Zampieri[2,3]

[1]University of Southern Denmark
[2]Saarland University, Germany
[3]German Research Center for Artificial Intelligence (DFKI), Germany

Abstract. In this paper, we present an automatic system for the morphosyntactic annotation and lexicographical evaluation of historical Portuguese corpora. Using rule-based orthographical normalization, we were able to apply a standard parser (PALAVRAS) to historical data (Colonia corpus) and to achieve accurate annotation for both POS and syntax. By aligning original and standardized word forms, our method allows to create tailor-made standardization dictionaries for historical Portuguese with optional period or author frequencies.

Keywords: historical corpus, corpus annotation, dictionary

1 Introduction

Historical texts are notoriously difficult to treat with language technology tools. Problems include document handling (hand-written manuscripts, scanning, OCR), conservation of meta-data, and orthographical and standardization issues. This paper is concerned with the latter, and we will show how a modified parser for standard Portuguese can be used to annotate historical texts and to generate an on-the-fly dictionary of diachronic variation in Portuguese for a specific corpus, mapping spelling variation in a particular period, author or text collection. The target and evaluation data for our experiments come from the Colonia Corpus [16] whereas for the annotation pipeline we use the PALAVRAS parser [1].

Several large projects handling historical Portuguese are worth mentioning, among them the syntactically oriented Tycho Brahe Corpus [3, 5] and the lexicographical HDBP project [8] aiming at the construction of a historical dictionary of Brazilian Portuguese. A third one is the online 45M word Corpus do Português [4] which provides a diachronic cross section of both European and Brazilian Portuguese. Spelling variation was an important issue in both the Tycho Brahe and the HDBP projects. Though the Tycho Brahe project originally used tagger lexicon extensions, both projects ended up basing their variation handling on a rule-based regular expression methodology suggested by Hirohashi (2005) [7]. The HDBP version, called Siaconf, lumps variants around a common 'base form', but not necessarily the modern form, favoring precision (almost 100%) over recall

[8]. Hendrickx and Marquilhas (2011) [6] adapted a statistical spelling normalizer to Portuguese, recovering 61% of variations, 97% of which were normalized to the correct standard form. They also showed that spelling normalization improved subsequent POS tagging, raising accuracy about 2/3 of the distance between unmodified and manual gold standard input. Rocio et al. (2003) [12], assigned neural-network-learned and post-edited POS tags *before* morphological analysis, after hand-annotating 10,000 words per text *without* normalisation, then adding partial syntactic parses for the output, using 250 definite clause grammar rules developed for partial parsing of contemporary Portuguese. In our own approach, like HDBP, we adopt a rule-based normalization approach [2], but aiming at exclusively modern forms, both for lexicographical reasons, and to support tagging and parsing with standard tools *without* the need of hand-annotated data.

2 Motivation

A historical dictionary can take different forms, spanning from the purely philological aspect to automatically extracted corpus data and frequency lists. Thus, Silvestre and Villalva (2014) [15] aim at producing a historical root dictionary for Portuguese, based on lexical analysis, etymology and using other dictionaries, rather than corpora, as their source. By contrast, the HDBP dictionary is based on 10M words of corpus data, providing definitions and quotations for historical usage [9]. Spelling variation is not the primary focus of either, and the published HDBP lumps variants under modern-spelled entries (10,500). However, the HDBP group also provides an automatically extracted glossary of 76,000 spelling variants for 31,000 'common' forms, as well as a manually compiled list of 20,800 token fusions (junctions). While this glossary constitutes an extensive and valuable resource, there are number of gaps filled by our project:

1. The HDBP glossary uses only Brazilian sources, while Colonia is a cross-variant depository with a potentially broader focus.
2. Unlike our parser-based resource, the glossary does not resolve POS ambiguity, nor does it offer inflectional analysis.
3. At least in its current form, the glossary does not differentiate periods or authors, something our proposed live system is able to generate on the fly.
4. Modern and historical entries are mixed, and it is not possible to tell one from the other. Thus, consonant gemination is mostly regarded as a variant, listing *villa* under *vila*, but modern *tão* and *chamam*, for instance, are listed under the entry of *tam* and *xamam*.
5. Contributing to the last problem, the glossary strips acute and circumflex accents in its entries, creating ambiguity even in the modern, standardized form, e.g. *continua* ADJ (*contínua*) vs. V (*continùa/continûa/continúa*). And though ã and õ are maintained, a grapheme like -ão is not disambiguated with regard to -am, which it historically often denoted. Thus, the entry *matarão* may really mean *mataram*, which becomes clear from the entry *vierão* which is not ambiguous like *matarão*, but can only mean *vieram*.
6. The glossary contains fusions, not marked as such (e.g. *foime*), and does apparently not make use of the separate junction lexicon.

While some of these problems (4-6) could be addressed by reorganizing the data and aligning it with a modern lexicon, we believe that a live, automated system with a flexible source management, parser support, contextual disambiguation, and a clear variant2standardized entry structure can still contribute something to the field. We evaluate our method and results using the Colonia corpus, but our approach can easily be adapted for new or different data sets.

3 Corpus and Annotation

The Colonia corpus[1] is considered to be the largest historical Portuguese corpus to date. It contains Portuguese manuscripts, some of them available in other corpora (e.g. Tycho Brahe [5] and the GMHP corpus[2]), published from 1500 to 1936 divided into 5 sub-corpora per century. Texts are balanced in terms of variety, 48 European Portuguese texts and 52 Brazilian Portuguese texts.

Century	Texts	Tokens
16^{th}	13	399,245
17^{th}	18	709,646
18^{th}	14	425,624
19^{th}	38	2,490,771
20^{th}	17	1,132,696
Total	100	5,157,982

Table 1. Corpus Size by Century

Colonia has been used for various research purposes including temporal text classification [11, 17], diachronic morphology [10], and lexical semantics [13].

Grammatical annotation adds linguistic value to a corpus, complementing existing philological mark-up (source, date, author, comments) and allowing quantitative or qualitative linguistic research not easily undertaken on raw-text corpora. Since it is time consuming to annotate a corpus by hand, automatic annotation is often chosen as a quick means to allow statistical access to corpus data. Obviously, a historical corpus will present special difficulties in this respect, since the performance of a parser built for modern text may be impaired by non-standard spelling and unknown words. In addition, historical Portuguese is difficult to tokenize, because word fusion may follow prosodic rules and occur for many function and even content words not eligible for fusion in modern Portuguese. For our work, we tackled these issues by adding pre-processing modules and a lexical extension to the PALAVRAS parser [2]. Our annotation method involves the following steps (Fig 1):

1. A pre-processor handling tokenization issues such as Spanish-style clitics (fused without hyphen), preposition fusion and apostrophe fusion at vowel/vowel word borders.

[1] 1) Original version: `http://corporavm.uni-koeln.de/colonia`; 2) With our annotation and normalized lemmas: `http://corp.hum.sdu.dk/cqp.pt.html`

[2] `http://www.usp.br/gmhp/CorpI.html`

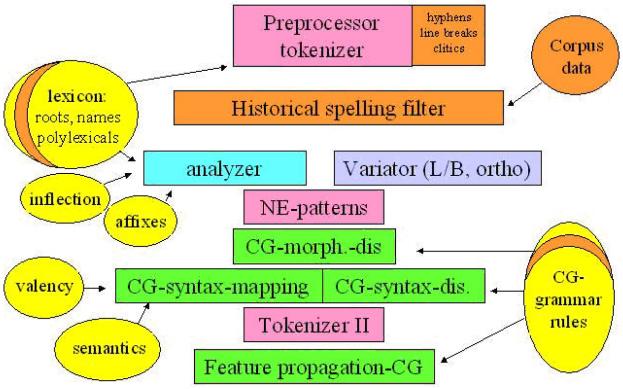

Fig. 1. System Components and Data Flow

2. A voting-based language filter blocking non-Portuguese segments from getting false-positive Portuguese analyses in the face of the orthographical 'relaxation' necessary for historical text.
3. A historical spelling filter, recognizing historical letter combinations and inflexion paradigms, and replacing words with their modern form where possible. Using a 2-level annotation, the original form is stored, while the standardised form is passed on to the parser. This module existed, but was extended with hundreds of patterns.
4. A fullform lexicon of modern word forms, built by generating all possible inflexion forms from the lemma base forms in the parser lexicon (1). This word list is used to validate candidates from (3) and for accent normalization.
5. An external dictionary and morphological analyzer, supplementing the parser's own morphological module. The module adds (historical and Tupi-Brazilian) readings to the (heuristic) ones for unknown words, allowing contextual Constraint Grammar (CG) rules to decide in cases of POS-ambiguity.

PALAVRAS' annotation scheme uses the following fields: (1) Word - (2) lemma [...] - (3) secondary tags (sub class, valency or semantics) - (4) part of speech (PoS), (5) inflexion, (6) syntactic function (@...) and (7) numbered dependency relations. For orthographically standardized historical words, (1) is the original word form, while the lemma (2) will indicate the modern lexeme. A special <OALT:...>tag in field (3) is used for normalized versions of the word form (1).

```
Esta [este] <dem> DET F S @>N  #1->2
povoaçam [povoação] <OALT:povoação> <Lciv> N F S @SUBJ>  #2->3
he [ser] <OALT:é> V PR 3S IND VFIN @FS-STA  #3->0
uma [um] <arti> DET F S @>N  #4->5
Villa [vila] <OALT:Vila> <Lciv> N F S @<SC  #5->3
mui [muito] <OALT:muito> <quant> ADV @>A  #6->7
fermosa [fermoso] <ORTO:formoso> ADJ F S @N<  #7->5
```

Because the added historical-orthographical information is contained in angle-bracketed tags, this annotation scheme is fully compatible with all PALAVRAS post-processing tools, allowing easy conversion into constituent tree format, MALT xml, TIGER xml, CoNLL format, PENN and UD treebank formats etc. However, in order to handle ambiguity and avoid false positives, normalisation patterns should only be applied for out-of-lexicon words, and multiple filtering options must be constrained by a modern lexicon. For this purpose we used a list of about half a million distinct word forms inflexionally constructed from PALAVRAS' lemma lexicon, as well a modern spell-checking list. Accent-carrying forms were checked both with and without accents, to allow for the fact that historical Portuguese was often more explicit in marking a vowel as closed or open, respectively (*cêdo, gastára, afóra*). The fullform list was also used to handle word fusion (junctions). For this task, unknown forms were systematically stripped of 'particle candidates' (prepositions, adverbs, pronouns/clitics), checking the remaining word stem against the modern word list. The following are the orthographical topics that were treated by a pattern-matching pre-processor:

```
geminated and triple consonants: (attenção, accumula, soffra, affligir)
word fusion: heide, hade -> hei de, há de
"Greek" spelling: ph->f, th->t, y -> i (mathematica, authores, systema)
nasals: em[dt]->en (bemdito), om[df]->on (comforme), aon->ão (christaons)
    chaotic -ão/am and -ões: áo, ào, âo, aõ, aò, àm, ao, ôes, óes, oens
extra hiatus-h: sahiu, incoherente, comprehender
z/s-dimorphism: isa -> iza, [aeu]z -> s, [óú]s$ -> z
s/c-ellision: sci -> ci, cqu -> qu: descifrada, sciência
lack of tonic accents: aniversario, malicia, razoavel, providencia, fariamos
superfluous accets: dóe, pessôa, enfrê
fluctuating accents:  nòs, serà, judaïsmo
```

To evaluate the effectiveness of orthographical filtering on the performance of the PALAVRAS parser, we did a small, inspection-based evaluation of 1 random sample per century, comparing the modified PALAVRAS with the original Treetagger [14] annotation as a baseline (provided on the Colonia website and produced without additional modules). Since the older texts require more orthographical intervention, the 20th century figures can also be used as a kind of baseline for PALAVRAS itself. Percentages in the table are for non-punctuation tokens (words), and the unknown/heuristic lemma count is without proper nouns.[3]

The table data indicates that the modified PALAVRAS outperforms the baseline for all centuries, and that the expected performance decrease for older texts is buffered by orthographical filtering. For both parsers a correlation between lexical coverage and tagging accuracy can be observed. A notable exception of the age-accuracy correlation is the 17th century text, which on inspection proved very modern in its orthography, probably due to the fact of it being a newspaper

[3] TreeTagger does not distinguish between common and proper nouns, but for the 'unknown' count, names were removed by inspection.

text (Manuel de Galhegos: 'Gazeta de Lisboa'), and as such subject to conscious standardization and proof-reading.[4]

Century	Words	Treetagger unknown (- PROP)	PALAVRAS heuristic lemma (- PROP)	Treetagger accuracy (POS)	PALAVRAS accuracy modified (POS)	PALAVRAS accuracy modified (synt. function)
16th	473	15.2	0.4	80.1	96.6	91.1
17th	432	0.7	0.0	96.5	98.8	94.4
18th	477	21.8	0.6	81.1	97.7	91.6
19th	372	1.3	0.8	95.2	98.1	93.3
20th	446	0.2	0.0	97.3	99.6	96.0

Table 2. System Performance per Century

Syntactic function assignment profited from orthographical filtering only indirectly, and historical syntactic variation (e.g. VS, VOS and OVS word order) were not addressed directly, leading to a moderate decrease in performance compared to modern text.

4 Generating a Diachronic Dictionary

From the automatically annotated Colonia corpus we extracted all wordforms that had undergone orthographical normalization (Table 3).

Century	Words	Orthographically Non-standard	Fused	Fused (relative)
16th	528K	4.11%	0.93%	22.63%
17th	577K	2.09%	0.25%	12.31%
18th	456K	2.88%	0.25%	8.68%
19th	2,459K	0.30%	0.08%	28.23%
20th	857K	0.17%	0.04%	23.52%

Table 3. Frequency of non-standard forms across centuries

This happened either by in-toto filtering or by inflexion-based lookup in the add-on lexicon. The frequency of such forms decreased, as one would expect, over the centuries. Token fusion followed this trend, but was lowest in the 18th century in relative terms (i.e. out of all orthographical changes).[5] Another trend across time is the decreasing use of Latin and Spanish. Our language identification module identified foreign chunks and excluded them from the analysis (Table 4). As can be seen, Latin and Spanish had a certain presence in Portuguese writing in the

[4] At the time of writing it was not clear if this text had been subject to philological editing in its current form, which might explain its fairly modern orthography.

[5] Parts of fused tokens were counted individually in the statistics, the token count is therefore higher than it would be counting the original text tokens as-is.

first 3 centuries of the Colonia period, enough to disturb lexicographical work if no language-filtering was carried out.[6]

Century	All Foreign	Latin	Spanish	Italian	French
16th	0.78%	0.49%	0.26%	0.01%	-
17th	0.78%	0.51%	0.24%	-	0.01%
18th	0.23%	0.17%	0.05%	-	-
19th	0.03%	0.02%	-	-	0.01%
20th	0.03%	0.01%	0.01%	-	0.01%

Table 4. Distribution of non-Portuguese text across centuries

Together, the orthographically non-standard words constitute a dictionary of historical Portuguese spelling with 10,400 wordform types, representing 52,000 corpus tokens. The dictionary contains 862 non-standard word fusion types (e.g. *ess'outra, fui-lh'eu, estabeleceremse*), representing around 5,000 tokens.

```
capitaens <OALT:capitães> (14; - 17th:10 18th:4 - -)
capitaes <OALT:capitães> (1; 16th:1 - - - -)
capitaina <ORTO:capitânia (5; 16th:5 - - - -)
capitam <OALT:capitão> (5; 16th:4 - 18th:1 - -)
capitan <OALT:capitão> (1; 16th:1 - - - -)
capitanîas <OALT:capitanias> (1; - 17th:1 - - -)
capitaõ <OALT:capitão> (3; - 17th:1 18th:2 - -)
```

5 Conclusions and Outlook

In this paper, we have shown that with an orthographical standardization module, a tagger/parser for modern Portuguese (PALAVRAS) can achieve reasonable performance across a wide range of historical texts, outperforming an unaltered statistical tagging baseline (Treetagger) by a large margin. Standardization was most important for the 16th-18th century, although some individual 17th century texts in our corpus already showed signs of standardization. Syntactically motivated grammar adaptations were not part of the current project, but are likely to further enhance performance, so future work should focus on this area.

An important result from the new annotation of the Colonia corpus is a method for automatically producing tailor-made spelling dictionaries of historical Portuguese. The resulting dictionary for Colonia itself contains almost 10,000 entries with century frequency information. We hope that both the method and the resource will be useful not only for linguistic-lexicographical purposes, but also as a language-technology resource, making it possible to reduce the out-of-vocabulary problem encountered by statistical taggers when used on historical text. A problematic aspect of the fullform substitution strategy for unknown

[6] Note that the figures constitute a lower bound. In order to achieve a precision close to 100%, only chunks with at least 4 (clear Latin 3) non-name words were treated, so individual loan words or mini-quotes are not included.

words are false negatives, where a word matches an existing modern form, but still should have been changed (e.g. *noticia* V? vs. *notícia* N), and ambiguous cases like *estillo*, where the substitution *estilho V?* (Spanish ll/lh) was allowed to preclude the correct *estilo* N (gemination variant). Frequency ranking might help, but only to a certain degree, and an alternative strategy - as yet untried - would be to pass both readings on to the CG grammar module for contextual resolution, based on the differences in POS or inflection.

References

1. Bick, E.: PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. In: Working with Portuguese corpora. pp. 279–302 (2014)
2. Bick, E., Módolo, M.: Letters and Editorials: A Grammatically Annotated Corpus of 19th Century Brazilian Portuguese. In: Proceedings of the 2nd Freiburg Workshop on Romance Corpus Linguistics. pp. 271–280 (2005)
3. Britto, H., Finger, M., Galves, C.: Computational and Linguistic Aspects of the Tycho Brahe Parsed Corpus of Historical Portuguese. In: Romance Corpus Linguistics: Corpora and Spoken Language. pp. 137–146 (2002)
4. Davies, M.: Creating and Using the Corpus do Português and the Frequency Dictionary of Portuguese. In: Working with Portuguese corpora. pp. 89–110 (2014)
5. Galves, C., Faria, P.: Tycho Brahe Parsed Corpus of Historical Portuguese (2010), `http://www.tycho.iel.unicamp.br/tycho/corpus/en/index.html`
6. Hendrickx, I., Marquilhas, R.: From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation. JLCL 26(2), 65–76 (2011)
7. Hirohashi, A.: Aprendizado de Regras de Substituição para Normatização de Textos Históricos (2005)
8. Junior, A.C., Aluísio, S.M.: Building a Corpus-based Historical Portuguese Dictionary: Challenges and Opportunities. TAL 50(2), 73–102 (2009)
9. Murakawa, C.d.A.A.: A Construção de um Dicionário Histórico: o Caso do Dicionário Histórico do Português do Brasil-séculos XVI, XVII e XVIII. Estudos de Lingüística Galega 6, 199–216 (2014)
10. Nevins, A., Rodrigues, C., Tang, K.: The Rise and Fall of the L-shaped Morphome: Diachronic and Experimental Studies. Probus 27(1), 101–155 (2015)
11. Niculae, V., Zampieri, M., Dinu, L.P., Ciobanu, A.M.: Temporal Text Ranking and Automatic Dating of Texts. In: Proceedings of EACL. pp. 17–21 (2014)
12. Rocio, V., Alves, M.A., Lopes, J.G., Xavier, M.F., Vicente, G.: Automated Creation of a Medieval Portuguese Partial Treebank. In: Treebanks, pp. 211–227. Springer (2003)
13. Santos, D., Mota, C.: A Admiração à Luz dos Corpos. Oslo Studies in Language 7(1), 57–77 (2015)
14. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. pp. 44–49 (1994)
15. Silvestre, J.P., Villalva, A.: A Morphological Historical Root Dictionary for Portuguese pp. 967–971 (2014)
16. Zampieri, M., Becker, M.: Colonia: Corpus of Historical Portuguese. ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research pp. 77–84 (2013)
17. Zampieri, M., Malmasi, S., Dras, M.: Modeling Language Change in Historical Corpora: The Case of Portuguese. In: Proceedings of LREC. pp. 4098–4104 (2016)