Bigorna – a toolkit for orthography migration challenges

José João Almeida[1], André Santos[1], Alberto Simões[2]

[1] Departamento de Informática, Universidade do Minho, Portugal
[2] Escola Superior de Estudos Industriais e de Gestão, Instituto Politécnico do Porto, Portugal
jj@di.uminho.pt, pg15973@alunos.uminho.pt, alberto.simoes@eu.ipp.pt

Abstract

Languages are born, evolve and, eventually, die. During this evolution their spelling rules (and sometimes the syntactic and semantic ones) change, putting old documents out of use. In Portugal, a pair of political agreements with Brazil forced relevant changes on the way the Portuguese language is written.

In this article we will detail these two Orthographic Agreements (one in the thirties and the other more recently, in the nineties), and the challenges present on the automatic migration of old documents spelling to their actual one.

We will reveal Bigorna, a toolkit for the classification of language variants, their comparison and the conversion of texts in different language versions. These tools will be explained together with examples of migration issues. As Birgorna relies on a set of conversion rules we will also discuss how to infer conversion rules from a set of documents (texts with different ages).

The document concludes with a brief evaluation on the conversion and classification tool results and their relevance in the current Portuguese language scenario.

1. Introduction

Languages evolve. This evolution can be natural and gradual, when speakers change habits during decades, or can be forced and drastic, when some kind of regulatory institution defines some rules (or heuristics) on how the language orthography will change.

The case of the recent evolution on the Portuguese language is the latest. This change was mainly a political issue to approximate the different countries that speak Portuguese. Details on the story of this evolution can be found on section 1.1..

In any of the two presented evolution cases, documents lose importance and relevance as soon as their orthography is outdated.

With the latest changes on the Portuguese language, and the amount of information that is typical of this era, it became necessary to develop automatic methods for language modernization.

This document discusses Bigorna, a toolkit developed to help on the update of documents orthography. It was developed to be as language independent as possible, relying in a set of rules learned by the comparison of old and actual corpora orthographic differences.

This section will present the main steps of Portuguese language evolution, the motivation for our work, what was the initial state before the project beginnings, and Bigorna main goals. It will be followed by a section on language resources construction and tools development. Section 3. presents a brief evaluation of the developed tools. It is followed by some conclusions and future work issues.

1.1. Portuguese Language Evolution

The Portuguese language evolved during the last century through successive orthographic agreements, especially between Portugal and Brazil. A full time-line of changes can be consulted in (Wikipedia:Reforms, 2010). Some of the described changes are just political, some other had real impact. In this article we are interested into the latest, that can be summarized as:

- **1931:** This was the first orthographic agreement between Portugal and Brazil trying to approximate both languages. This change abolished the silent *s* from words such as *sciência*, *scena*, *scéptico*. It also changed the way compound verbs were written, from *dir-se há* and *amar-te hei* to *dir-se-á* and *amar-te-ei*.

- **1945:** This new reform just hit the language in Portugal, cleaning a set of different accents, like removing completely the umlaut (trema, in Portuguese), and removing the circumflex accent in homograph words such as, *acêrto* and *acerto*, *cêrca* and *cerca*, *côr* and *cor*, *fôra* and *fora*, *dêsse* and *desse*, and so on.

- **1973:** While not a political agreement, Portugal followed Brazil and abolished accent marks in secondary stressed syllables (*tètralogia* and *tetralogia*).

- **1990/2009:** On December the 16th of 1990 the Portuguese Language Orthographic Agreement (PLOA) (da República, 1991) was approved in Lisbon by Angola, Cape Verde, Guinea-Bissau, Mozambique, Sao Tome and Principe, Brazil, Portugal and a delegation of galician observers. Its goal is to unify, as widely as possible, the general vocabulary of the Portuguese language.

  The urge to bring closer all variants of Portuguese implies some changes on the structure and content of the language itself. On those content changes, the phonetic criteria (or pronunciation one) was chosen instead of the etymological one. This led to some cases of multiple spelling (whenever a word is pronounced differently according to the variant in use) like *facto* and *fato*, or *suntuoso* and *sumptuoso*. The lack of consensus about whether some words are to be updated or

not, and the absence of a common orthographic vocabulary (as foreseen in article 2 of the PLOA) gets this process even more complex.

While this latest political change was accepted and should be put in practice soon, real users of the language will continue to write as they learned for some time, leading to a double orthographic form. This means that the two forms (before-1990 and after-1990) are relevant at the moment.

1.2. Motivation

The motivation to this work is two fold: to update old documents in the public domain (with more than 70 years after author dead) to current used orthographic form, and the need to update current documents to the future[1] orthographic form, as defined by PLOA:

- With the actual tendency to public accessible repositories or libraries in the Internet, and as a special case Project Gutenberg[2], a lot of books that are currently under public domain are being scanned and transcribed. While it is crucial to keep the original orthographic form in the archive for linguistic studies (for example), their modernization is also relevant, namely for reading purposes.

  This leads to the necessity to develop a tool to perform the systematic migration of language from before 1930 to the currently used (before 1990/2009 reform).

- We are in the information era. A lot of different documents are currently available and there will be the need to update them to the new orthographic form.

  The PLOA implementation demands a revision and an update of the existing linguistic tools and the creation of new ones, and the complexity of these tasks increase due to the multiple spelling cases (see section 1.3.).

A great help to both processes would be the creation of a set of tools capable of inferring migration rules through the comparison of texts. Tools like those would be useful to develop applications that people could use to face the issues of PLOA. While this document will mainly focus PLOA implementation, the development tools also have a role in updating old documents spelling, or in future language updates.

1.3. Summarizing

The current scenario in the Portuguese language, as created by the PLOA:

- the spelling changes dictated by the agreement cannot be determined automatically, as they rely on phonetic criteria and may sometimes be ambiguous (different accents, for instance);
- therefore, it is important to create and maintain an *Orthographic Agreement's Knowledge Base* — `OAKB`, a table containing lemmas, changes and rules based on the currently existing lists, new examples found and the results of processing texts. This table will be updated and evolve (one might never know when it is complete);
- the main problem is on how to determine which words are candidates to integrate the `OAKB` (what are the changing words).

It was clear the need of a project dedicated to the migration process. It was Bigorna project birth.

1.4. Bigorna's design goals

In order to help in the migration process, it is necessary to have some resources and tools. The first set of requirements included:

- collect all the resources related to PLOA that are already available in the Internet;
- create a spell-checker dictionary for the new Portuguese language;
- build a tool to convert text from the pre-agreement Portuguese to the updated Portuguese language;
- build a classification tool, to identify the variant of Portuguese used in a given text;
- make the previous tools work with as few dependencies as possible.

We soon understood that the set of changing words was being updated frequently as PLOA is not defined by exhaustion (it does not list all words, just defines a few heuristics). This lead to the following decisions:

- center all the information about the word changes in an unique table (`OAKB`);
- build a system, that given that table, would generate:
  - the new spell-checker dictionary;
  - the language migration tool;
  - the Portuguese language variant classifier
- develop tools to help in the construction and manipulation of `OAKB`:
  - tools to find lemma changing candidates from texts;
  - a tool to extract lexical differences from pairs of texts;
- make the previous tools as generic and reusable as possible.

2. Bigorna Development

This section describes the Bigorna approach and its development details. We will first focus how resources were compiled into `OAKB` and then how the spell checker dictionary was updated. Follows a description about the created tools: conversion tools, the language classifier and the lexical comparison tools.

[1] We will use "current orthographic form" to refer to the orthographic form before 1990 agreement, and the "future form" for the orthographic form after 1990 agreement.

[2] `http://www.gutenberg.org/`

2.1. Compiling OAKB

To start the compilation of resources to integrate the knowledge base we started by searching for free resources related to the PLOA. We found several online dictionaries, spell-checkers, vocabularies and a growing number of publications about the subject.

From the vocabularies found, one stands out, a *list of changed words*, containing over 9 000 entries, published by Instituto de Linguística Teórica e Computacional (ILTEC, 2008). Another list, not as complete, was compiled by Projecto Quintus (Quintus, 2008).

Each entry on this list contained both the European and Brazilian Portuguese versions of a given word together with that word's updated form and a brief comment indicating the preferential new form for each Portuguese variant.

All lines where there was no change accordingly to PLOA were deleted and the remaining ones were used to compile our first OAKB version. This resulted in approximately 2 600 words.

After the reorganization process, the structure of OAKB is:

```
oakb = entry*

entry =
  pt_pt            : word
  pt_br            : word
  pt_oa            : word*
  preferencial_pt  : word
  preferencial_br  : word
  type     : Capit | Hyphen | Accent | Normal
```

The field *type (of change)* was used in order to make possible to define different ways of dealing with different situations (remotion of capitalization depends on context, remotion of accent marks depends not). Presently we only differentiate the hyphenation changes, but in the future the other types will be discriminated as well.

2.2. Updating the dictionary vocabulary

To prepare a spellchecker dictionary we decided to use jSpell, an open-source morphological analyzer (Almeida and Pinto, 1995; Simões and Almeida, 2002). jSpell has the advantage that its updates can be easily propagated (using *Chuveiro de Dicionários* (dos Santos Vilela, 2009)) to dictionaries for other spell-checkers engines (aspell, ispell, hunspell and myspell), some of which are used by well known open source applications like Firefox, Thunderbird or OpenOffice.

JSpell also has the ability to, when provided with a list of lemmas and their derivation rules, generate an exhaustive list of all derived words. The relevance of this feature is that the language updating can affect just the lemma and/or the generation rules, and a set of words will be automatically fixed. Also, the generated lists can be compared with the entries into OAKB easily.

The update of the dictionary was performed by searching the words from OAKB in jSpell's European Portuguese dictionary and replacing them by their updated version. When there was more than one possible form, we chose the version recommended for European Portuguese.

```
Function newdic(oakdb,dicjs)
  for ( x ∈ dom(oakb)
        ∧ oakb[x].type = normal
        ∧ x ≠ oakb[x].preferpt
        ∧ x ∈ dom(dicjs))
    {
      neww ← oakb[x].preferpt
      dicjs[neww] ← dicjs[x]
      delete dicjs[x]
    }
```

From the 2 600 words in OAKB, just 960 were related directly with a lemma in jSpell's dictionary. Note that from these 960 lemmas jSpell generates a total of 11 500 words.

2.3. Language Conversion Tools

Our second task was the development of a conversion tool, to migrate texts from the pre-agreement spelling to the new one.

Due to the multiple spelling cases which are, moreover, dependent on the variant of Portuguese being used, it was not possible to create a single conversion tool. Therefore, we decided to create two different tools: one capable of updating European Portuguese (pt2ptao) and one other capable of updating Brazilian Portuguese (br2brao). Note that these tools convert the original variant to the after-agreement variant.

Once again, we started with OAKB and jSpell as our base tools. Typical entries in the jSpell's dictionary include the lemma, its morphologic properties, and the derivation rules valid for that lemma:

```
acalentar/#vt/XYPLD/
coiote/#nm/p/
laico/#a/fidp/
zinco/#nm//
```

OAKB was used to extract a list of the changing lemmas. This list includes the European Portuguese lemmas and their updated form (lemma_pt/lemma_ptao), and from jSpell dictionary were extracted the derivation rules valid for that same Portuguese lemma (lemma_pt/rules).

By expanding each lemma_pt and each lemma_ptao in parallel (using the corresponding jSpell rules) we obtained a new table where each entry is structured as follows:

```
word_pt/word_ptao
```

As we did not have the derivation rules for Brazilian Portuguese words (because jSpell's Portuguese dictionary is European Portuguese specific), we decided to use the rules present on OAKB to derivate the corresponding Brazilian Portuguese words, and their updated form. Then, by a process similar to the above, we obtained a table with the following structure:

```
word_br/word_brao
```

The conversion tool was developed in Perl and we tried to minimize its dependencies to make it easier to install and use. For the very same reason, we decided to include the word table itself in the script file, resulting in a self-contained single file.

The tool performs the conversion by sweeping a given text looking for `word_pt` words and replacing them by the matching `word_ptao` (`word_br` and `word_brao` in the Brazilian version). We included some additional options which protect XML tags and LaTeX commands within the text.
Follows some examples of the tools interaction.

```
$ pt2ptao
A adopção do acordo implica a actualização
de ferramentas.
                ↓
A adoção do acordo implica a atualização
de ferramentas.

$ br2brao
Ele fez um vôo rasante sobre a aréia.
                ↓
Ele fez um voo rasante sobre a areia.
```

The different character encoding systems were the source of several problems and the target of a special care during this project. Both versions of the conversion tool, as well as the other programs developed by Bigorna, accept both ISO-8859-1 and UTF-8 encoded text.

2.4. Variant Classifier

A relevant goal on Bigorna is the identification of the Portuguese variant present on some text. This feature is crucial so we can use one of the tools presented above automatically without human intervention.
Another tool, named `whichPT`, was developed to identify whether the European or the Brazilian Portuguese variant is present on a document.
The tool incorporates a list of (just) European Portuguese words and another of (just) Brazilian Portuguese ones, compiled with a subset of the conversion tools lists.

```
Function calc_whichpt_lsts(dicpt,dicbr,oakb)
  for ( x ∈ dom(oakb)
     ∧ oakb[x].type = (normal or accent)
     ∧ oakb[x].pt_pt ≠ oakb[x].pt_br
     ∧ oakb[x].pt_pt ∈ dom(dicpt)
     ∧ oakb[x].pt_br ∈ dom(dicbr))
   {
    wpt ← oakb[x].pt_pt
    wbr ← oakb[x].pt_br
    justpt ← justpt ∪
         {x ∈ deriv(wpt,dicpt)| x ∉ dicbr}
    justbr ← justbr ∪
         {x ∈ deriv(wbr,dicbr)| x ∉ dicpt}
   }
```

The classification is performed by counting the number of words from the text that have a match in each of the lists; at the end, the result is the variant with the highest number of matches.
This implementation allows to easily expand the script to other variants (for example, ancient Portuguese). It is also possible to perform the classification ignoring the words inside XML tags or LaTeX commands, to print at the end the exact number of matches in each variant, or to output all the words matching each variant.

This tool was tested with two variants from the book *Amor de Perdição*, from the Portuguese author *Camilo Castelo-Branco*, one written in European Portuguese and the other in Brazilian Portuguese. The results were the expected.

```
$ whichPT AmorPerd.ptPT AmorPerd.ptBR
AmorPerd.ptPT     pt
AmorPerd.ptBR     br
```

2.5. Lexical Comparison Tools

The lists of words compiled above were based on previous existing lists. But there are other situations where there are no available lists of words. Although there are no lists, there are documents with different orthographic versions.
With that in mind, we built a tool to help on comparing two versions of a text and detect (linguistic) differences between two versions of a text with different spelling.
This script, named `lexdiff`, receives two files and compares them with the help of the Unix commands `egrep` and `diff`.
The result of `lexdiff` applied to previously mentioned versions of *Amor de Perdição* is:

```
$ lexdiff -ac AmPerd.ptBR AmPerd.ptPT
   32 acadêmico => académico
   14 idéia     => ideia
   12 redargüiu => redarguiu
   7 gênio      => génio
   4 refletiu   => reflectiu
    ...
```

For each changed word `lexdiff` outputs the number of times that change ocurred on the given texts.
The script is also able to calculate narrower differences. This option allows us to detect changing patterns (sequences of characters that get usually updated), independently of the word. Below is the result of this narrower approach to the same pair of texts:

```
$ lexdiff -m -ac AmPerd.ptBR AmPerd.ptPT
    36 et  => ect
    18 déi => dei
    17 güi => gui
    8  at  => act
    7  eç  => ecç
     ...
```

The option shown above creates a confusion matrix (CM) which relates each word with all the possible words it can be. Follows an extract of the CM created in the previous example:

```
et   => {  ect  => 36 },
déi  => {  dei  => 18 },
güi  => {  gui  => 17 },
at   => {  aat  => 1,
           apt  => 1,
           act  => 8  },
eç   => {  eaç  => 2,
           ecç  => 7,
           epç  => 2  },
```

There are many uses for `lexdiff`:

- it can be used to search for words that could be inserted into OAKB and, consequently, in the updated dictionary, the converters and the classifier, or to see how similar two given texts are;

- it can even be used as a spell-checker, although its main application will be the generation of tools like the ones previously described in this article.

At the moment we are developing a lexpatch tool that, given a changes file produced by lexdiff can update texts according to those changes.
The combined action of both of these tools will allow the automatic generation of converters after a learning process made from manual adaptations of texts.

3. Evaluation

In order to evaluate the tools previously described, we used different strategies and resources.
In this section we will discuss the process used and the results obtained in the evaluation of pt2ptao, br2brao and whichPT.
Given that lexdiff is not dependent of specific language resources, its evaluation deals with a different problem: if the algorithm is correct, if its results are useful, and if it can be used for other relevant tasks. While lexdiff was used in the tests descried here, we will not deal with its evaluation in this document.

3.1. Evaluating the Conversion Tools

In order to evaluate both versions of the conversion tool we decided to test other similar tools in order to make a common evaluation and compare the results obtained using the different systems.
The systems compared to evaluate the European Portuguese version (pt2ptao) were Priberam[3], Porto Editora[4] (PE) and Portal da Língua Portuguesa[5] (PdLP) conversion tools.
Results of the ortographic updating of "*Amor de Perdição*" are summarized in Table 1 and Table 2.
Given the lack of work power to select manually the changing words from the list of words present on the book, we compiled the set of words correctly changed in the four systems and used this set to compute the recall measure[6].

NrWords	*Amor de Perdição*	Changed
Total	48 011	81
Different	8 717	61

Table 1: Comparison of the number of words in *Amor de Perdição* and the number of changed words.

[3] http://www.priberam.pt/

[4] http://portoeditora.pt/

[5] http://www.portaldalinguaportuguesa.org/
Due to timing reasons, Portal da Lingua Portuguesa results were gathered using the online version of their conversion tool, which may have affected their results.

[6] The real number of words to be changed should be higher (probably not all the words were detected by the four systems), what would make the recall value lower.

NrWords	pt2ptao	Priberam	PE	PdLP
Total	69	80	78	48
Different	49	60	58	33
Precision	1.0	1.0	1.0	1.0
Recall	0.80	0.98	0.95	0.54

Table 2: Comparison of the number of changed words using the four different conversion systems.

After some analysis of the words not updated by pt2ptao we noted that it is not yet dealing with the task of lowercasing the months and days of the week.
We tried a similar approach to evaluate the Brazilian Portuguese version of this tool (br2brao). The systems chosen for the comparison were Priberam (they also have a conversion tool for Brazilian Portuguese) and Interney[7] conversion tools.
We tested these systems with a Brazilian Portuguese version of "*Amor de Perdição*", but we could not interpret the results; this task must be performed by someone more experienced with Brazilian Portuguese, as it requires a native-level knowledge of the language which none of the authors possesses.
During this process of evaluation we found some complex aspects that need a careful approach:

- One may not want to convert words which are part of composite proper names (names of streets, family names, etc);

- The remotion of capitalization should not be applied to words in the beginning of sentences;

- The PLOA document is not clear on how/when to apply the hyphenation rules.

Therefore, a conversion tool needs to be able to identify phrases and entities in the text to perform accurate conversions.
This process of comparative evaluation, which was made possible with the cooperation of the teams responsible by the other conversion tools, has already lead to some improvements in some of these tools.

3.2. Evaluating the Language Classifier

In order to evaluate the classifier (whichPT) we used a set of 100 books in different Portuguese dialects (pt-pt, pt-br).
The classifier gave the correct answer in all the cases. One of the books was difficult to evaluate as it had a mixure of both dialects.
We then decided to stop the classification after obtaining a certain number of language clues (20). With this limitation we obtained 2 ambigous books.
This evaluation was also relevant as, after analyzing the detected clues, a type was found on our knowledge base. Correcting it the classified answered correctly for all books.

[7] http://www.interney.net/

4. Conclusions

We are now replicating the updates on jSpell's dictionary to other spell-checkers, using *Chuveiro de Dicionários*. The resulting dictionaries will be included as the official dictionary for Firefox, Thunderbird and OpenOffice in the near future.

Tools construction was only possible after the construction of OAKB. For this purpose the resources already available on the Internet were crucial. We would not have resources for the hand construction of this list.

With a proper knowledge base the identification of variant is simple and comparable to common language identification tasks.

The conversion tool requires, as already explained, some more linguistic knowledge that is not yet present on our tools, like the mentioned entity recognition. Nevertheless, current approach is already performing comparatively well.

The migration tools are being tested and should be implemented as web-services and released as stand-alone tools.

The lexical difference tool is already being used for other projects like Dicionário-Aberto (Simões and Farinha, 2010), where a definitions dictionary from 1913 was transcribed and is being migrated to the actual orthography.

While the full orthographic actualization is almost impossible to perform in an automatic way (mostly given the abundance of different words present on a dictionary), current experiments show that it is possible to update 89.39% of the words automatically.

5. Future work

The converters will be tested and improved with the help of `lexdiff` and `lexpatch`.

The `lexdiff` script and subsequent work could be used to improve the tools built in the previous stages of this project (for example, new converters can be developed merely by the analysis of handmade transcripts).

6. Acknowledgments

We would like to thank both the Priberam and the Porto Editora teams, who applied their commercial tools on our test cases for evaluation purposes.

7. References

J.J. Almeida and Ulisses Pinto. 1995. Jspell – um módulo para análise léxica genérica de linguagem natural. In *Actas do X Encontro da Associação Portuguesa de Linguística*, pages 1–15, Évora 1994.

Diário da República. 1991. Acordo ortográfico da língua portuguesa, 1990. Technical Report 193, série I-A, 23 de Agosto. `http://www.portaldalinguaportuguesa.org/index.php?action=acordo&version=1990`.

Rui Miguel Rodrigues dos Santos Vilela. 2009. Geração de dicionários para correcção ortográfica do português. Master's thesis, Escola de Engenharia, Universidade do Minho, Braga, Outubro.

ILTEC. 2008. Vocabulário de mudança. Inclui listas de palavras que mudam, Portal da Língua Porguesa. `http://www.portaldalinguaportuguesa.org/index.php?action=novoacordo`.

Projecto Quintus. 2008. Lista total (?) de palavras da norma culta lusoafricana que são modificadas pelo acordo ortográfico de 1990. Inclui lista de palavras que mudam, Projecto Quintus.

Alberto M. Simões and J.J. Almeida. 2002. Jspell.pm – um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas do XVII Encontro da Associação Portuguesa de Linguística*, pages 485–495, Lisboa 2001.

Alberto Simões and Rita Farinha. 2010. Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, (16), April. forthcomming.

Wikipedia:Reforms. 2010. Reforms of portuguese orthography. Wikipedia article, `http://en.wikipedia.org/wiki/Reforms_of_Portuguese_orthography`. [Online; accessed 10-March-2010].