The Common Orthographic Vocabulary of the Portuguese Language a set of open lexical resources for a pluricentric language

José Pedro Ferreira[1], Maarten Janssen[2], Gladis Barcellos de Almeida[3], Margarita Correia[1], Gilvan Müller de Oliveira[4]

[1]Instituto de Linguística Teórica e Computacional – ILTEC, Lisbon,Portugal
[2]Institut Universitari de Lingüística Aplicada - IULA, Barcelona, Spain
[3]Núcleo Interinstitucional de Linguística Computacional - NILC, São Carlos, Brazil
[4]Instituto Internacional da Língua Portuguesa - IILP, Praia, Cape Verde
jpf@iltec.pt, maarten.janssen@upf.edu, gladis@ufscar.br, mcf@iltec.pt, iilpde@gmail.com

Abstract

This paper outlines the design principles and choices, as well as the ongoing development process of the Common Orthographic Vocabulary of the Portuguese Language (VOC), a large scale electronic lexical database which was adopted by the Community of Portuguese-Speaking Countries' (CPLP) Instituto Internacional da Língua Portuguesa to implement a spelling reform that is currently taking place. Given the different available resources and lexicographic traditions within the CPLP countries, a range of different solutions was adopted for different countries and integrated into a common development framework. Although the publication of lexicographic resources to implement spelling reforms has always been done for Portuguese, VOC represents a paradigm change, switching from idiosyncratic, closed source, paper-format official resources to standardized, open, free, web-accessible and reusable ones. We start by outlining the context that justifies the resource development and its requirements, then focusing on the description of the methodology, workflow and tools used, showing how a collaborative project in a common web-based platform and administration interface make the creation of such a long-sought and ambitious project possible.

Keywords: computational lexicography, lexical resources, corpus linguistics

1. Introduction

The spelling of Portuguese has changed several times over the past century, with the aim of coming up with a set of orthographic rules that are as close as possible in every Portuguese speaking country. These spelling reforms are usually implemented by means of a *vocabulary*, with the sense of word inventory, similar to what happens in other languages, such as for instance in Dutch with the *Woordenlijst Nederlandse Taal* (Jacobs, 1997).

In the context of an ongoing spelling reform which implements a common set of writing rules to all Portuguese speaking countries, there will be, for the first time, a Common Orthographic Vocabulary of the Portuguese Language (VOC), a resource containing information about the spelling and formal properties of words that is shared by every country in the Community of Portuguese Speaking Countries (CPLP). The work is being undertaken under the supervision of the International Institute for the Portuguese Language (IILP), CPLP's language bureau, with technical support from ILTEC and NILC, and the participation of teams from every CPLP country.

More than a mere word list with citation forms and corresponding POS tags, VOC aims to be both a reference for the orthography of Portuguese and a useful lexical resource for other purposes, in particular for scientific research and NLP. The project is currently being developed by teams in every CPLP country, the final release of its first version expected by mid-2014.

This paper contextualizes the project, outlines its aims, requirements and methodology, and presents the preliminary and expected results.

1.1. Ongoing spelling reform

In 1990 the countries where Portuguese is a state language agreed upon and signed a spelling reform. Despite the political consensus and the fact that it unifies the spelling rules used in the CPLP countries, the changes didn't start being implemented until the end of the last decade, from 2009 onwards[1]. A transitional period is currently under way, the new spelling being gradually implemented in different timelines for each country and sometimes for different sectors within each country (education, administration, media, etc.).

Once the reform implementation period is over, the spelling of Portuguese will be determined by a single legal document all over the CPLP, putting an end to a long period where different, national-level documents defined the orthographic rules. In 1911, the Portuguese government ended decades of fierce public dispute about the ideal character of the spelling of Portuguese by setting up an official orthography for the first time. However, it was a unilateral reform, which was not followed by Brazil, giving way to the existence of two national level legally binding spelling rule sets.

Throughout the last century, there were several proposals for *strong* reforms, which would have unified the spelling of almost every word, but they were all unsuccessful. In 1990, an agreement was finally reached on a weaker proposal, which eliminates purely orthographic variation but condones several cases of pre-existing variation with linguistic motivation.

[1] For a more detailed description of the reform process, see Ferreira, Lourinho & Correia, 2012.

1.2. Vocabularies in the main stage

Previous spelling reforms of Portuguese were implemented using officially-backed orthographic vocabularies with a large number of citation forms, sometimes along with POS and other information (Verdelho, 2002). Although the original legal text of this reform required the development of a common orthographic vocabulary within two years, this was not to be achieved in the following twenty, and, before the project this paper presents, the development of such a resource had never gone beyond initial planning.

Historically, this is of great relevance. There was a first spelling agreement between Portugal and Brazil, in 1931, but the official (national level) vocabularies interpreted the reform rules differently. This was a result, on the one hand, from the inaccuracies and technical malaises of the spelling agreements in general and from political reasons, and, on the other hand, from individual and uncoordinated efforts without a common and integrated methodology.

With the 1990 spelling reform, the same initially happened: Portugal and Brazil each developed official, national level vocabularies in the new orthography, presenting some different interpretations of the same rules. Furthermore, several commercial lexicographic companies published dictionaries in the new orthography, each again with differences in the application of the rules.

1.3. An unbalanced tradition

To this date, good lexicographic inventories have only been developed for Brazil and Portugal. These countries already had their own official, national level word lists, but they had never been fully integrated with each other. The implications of the sizeable number of spelling variants condoned in the reform's rules had thus never been fully assessed and treated lexicographically.

In the other six CPLP members, there are no national level linguistic resources, either corpora or lexicographic works, albeit the fact that in some cases there are clearly national standard emergence processes taking place.[2]

2. Project description

2.1. Aims

The main objective of VOC is to build a free-access lexical information database representing the contemporary lexicon of Portuguese as a whole, in a framework and set-up that is common to every CPLP country. The resulting resource should go beyond merely replicating the information contained in existing lexicographic works; it also aims to reflect current language usage in every country where Portuguese enjoys official status. In a number of countries, this is the first time that a lexicographic treatment of their national variety will take place. Given the fact that VOC implements an international agreement, the methodology presented in this paper and the decisions taken during the compilation and building process result from multi-lateral decision making structures.

The biggest – but not sole – purpose of VOC is to be the key resource in the application of the ongoing spelling reform, determining a definitive interpretation of the rules defined in the official text of the reform through its application to specific word forms. Apart from implementing the spelling changes, it will allow for the establishment of more objective criteria for the spelling of traditionally problematic orthographic contexts, such as those involving hyphenation rules and loan word adaptation, which the text of the reform did not always fully specify.

Beyond that, VOC is being built to put forward a free and accessible lexical resource which is useful for the general public, researchers and developers alike, taking into consideration both its future expansion and reusability by the community, and further language planning development efforts. To be able to contribute toward orthographic homogeneity, any new resource needs to take into account the role played today in setting up linguistic standards by NLP-related tools such as spell checkers, translation memories, and machine translation tools.

2.2. Requirements

Given the unbalanced and to a certain extent disjoint existing lexicographic traditions, VOC can't be a mere repository of the information already contained in existing lexicographic resources. Instead, in the process of cataloguing the existing lexicographic resources, updating them to the new spelling rules, and normalizing their contents, VOC has to furthermore reflect the current linguistic usage in every country.

On the other hand, while the reform calls for the resource to normalize the existing technical and scientific terminologies (a task beyond the scope of setting the orthographic forms of words), at this stage the work will focus on registering and linking whatever terminologically relevant entries are caught up during the compilation process, leaving the uphill task of further normalization to be multi-laterally evaluated after VOC's conclusion, in a gradual, thoughtful and informed way.

Since the spelling reform allows for some level of (not purely orthographic) variation in certain contexts, both within and between different linguistic varieties, it is very important to keep track of cases where a specific word form is only used in one or more countries, linking the varying forms and framing their usage. For this and other purposes, it is important that the provenance and sources for every word is tracked and added to each entry.

Given the high social and political profile of VOC, the methodology and processes pursued during the work should be made clear and be replicable, contrary to what is standard practice by private lexicographic publishers. While focusing on settling down the spelling rules, VOC should be seen as an opportunity to create other useful formal lexical information resources in a structured and integrated way.

[2] E.g., for Mozambique, see Gonçalves, 2010.

2.3. Design options

The methodology was set forth by a work group and politically validated by every country in the context of IILP, the CPLP bureau for language matters. The decision fell on an easy to update, collaborative, cloud-based platform, with an open, digital publication support as the result.

The lexical database is built using the OSLIN model (Janssen 2005), which provides a very good web-based administration interface allowing for the quick creation of large-scale lexical resources and providing access to well-tested and easy to use maintenance tools (Ferreira et al. 2008). The entire creation process of VOC is based on that set of web-based administrative tools, allowing for decentralized collaborative lexicographic development, ideal for a multi-national project such as this.

Given, on the one hand, the need to represent Portuguese as whole, and, on the other hand, the poor representation of a number of varieties in the existing lexicographic works, both lexicographically compiled works and corpus retrieved information are used, although under strict restrictions and close lexicographic supervision to ensure the quality of the end product.

The end user interface (ILTEC, 2006) provides users with the ability to search for word forms by literal strings, or more advanced searches including partial matches or pattern-based queries. The entry for each lemma contains the inflectional paradigm with the explicit spelling of each of the inflected forms, the orthographic syllable division for word breaking purposes, and functional links to other entries, mostly to morphologically related words. Each of these resources provides data that can be accessed through the entry for any given lemma, or via a dedicated lexical resource which makes the analysis of related sets of words easier to do. Other information will be added in the near future, most notably the phonetic transcription for several varieties of Portuguese (Ashby et al. 2008), information which is currently not available freely anywhere for Portuguese. Besides being made available through this online interface, the data will be shared in a range of standardized formats, namely those put forth by the Lexical Markup Framework (ISO/IS 24613:2008), to ease their usage for NLP development and other purposes.

3. Methodology

3.1. Workflow summary

National vocabularies to be integrated in VOC are built from a mix of existing lexicographic sources, corpus-extracted frequency lexica and two semi-automatic formal neologism tracking systems (Janssen, 2008; Martinez, 2012). Most of the weight is put on the existing lexicographic works, adapted to VOC's model and double-checked by means of corpus-extracted lexica. This ensures a good balance between reflecting tradition (as presented by the more prescriptive lexicographic works) and up-to-date description (as ensured by the more recent sources in the corpus and by the neologism trackers).

The vocabularies are previously adapted to the spelling reform using in-house built tools (Ferreira, Lourinho & Correia, 2012) and manual pattern-based verification, and inserted into the lexical management platform. There, a number of different types of formal data are automatically generated for each entry (inflection, syllable division, stress position, derivational relations) and individually evaluated and inserted by team members.

3.2. Existing sources

The bulk of the work of the VOC project consists of two interrelated tasks: firstly, to merge and entwine the existing national level official lexicographic resources already in place for Portugal and Brazil; and secondly, to retrieve and integrate data representative of the other national varieties from corpus-based word lists in general built specifically for this project.

The lexical representation model and the administration environment are the same as the ones used for the European Portuguese national vocabulary (Correia, 2010). Since different lexicographic works more often than not have different inclusion and lexical identity criteria, along with incompatible POS tag-sets – especially when comparing those published in paper with digital lexical resources –, every source inserted into the database has to first be adapted to the common format and criteria, in order to make it comparable.

A dedicated management tool in the OSLIN platform allows for the easy treatment of both lexicographic and corpus-extracted frequency lexica. In order to process a lexical resource, the tool first creates a translation table and then eases the task of adapting the entries in that resource to the lexical identity criteria and POS tag-set used in VOC. Once this has been done, all the *citation-form +POS-tag* pairs in that resource can be compared to the ones already in the VOC database. The comparison process adds an explicit link between the entries in the lexical resource and their corresponding VOC entries for those words that are already in the database; conversely, it tags the entries in the resource that are not yet in the VOC database as missing words. The tool then provides the option to reliably add all those missing words that match the selection criteria set out for VOC. The explicit link between the VOC database entries and each of the lexical resources in which they occur provides a solid layer backbone for the validation and cross-verification for each of the words in VOC.

This process is employed systematically to compare sources with each other, selecting the entries that are to be inserted into the database. Every new entry needs to be registered in at least one reference work *and* in an auxiliary source before being individually assessed and inserted by a team's lexicographer. In the case of Brazil, the reference work is the vocabulary of the country's Academy of Letters, which was compared with two computational lexica compiled from large tagged corpora, a spell-checker base lexicon by NILC (Nunes et al, 1996) and Corpus Brasileiro (Berber Sardinha, 2009).

3.3. New sources

The representation and comparability of the language varieties in use in Portugal and Brazil is ensured by a large body of existing lexicographic resources and corpora. The building process of the official Portuguese resource (VOP) followed criteria similar to the process described above for Brazilian Portuguese. It was based on two well-established lexicographic works by the Lisbon Academy of Sciences, and on one by the biggest private publishing company (Porto Editora), verified and complemented with information retrieved from a vast amount of available high-quality corpora, mostly those built by *Linguateca* (Santos, 2011).

For the national varieties of other CPLP countries, though, there are often no suitable resources, although there are clear indicators that national standards have been rising in several of them, particularly in Angola and Mozambique (Gonçalves, 2000). To overcome this lack of representation, each country has a team compiling corpora that are representative of their national variety from accredited local sources. To assure comparability between those varieties, common criteria were established for the creation of the corpus, building on genres and sources that are sufficiently available for all CPLP countries, and setting minimum size requirements for each one of those genres. This secures not only the representation of every country in the vocabulary, but also the creation of much sought-after comparable corpora for these currently under-studied linguistic varieties.

After previous computational processing, these corpora are treated using an OSLIN corpus management tool, which takes on lemmas, along with their inflectional paradigm forms and frequency index, and compares them with the entries already in the lexical database, similarly to what is done when adding sources based on lexicographic works. Since the first version of these resources is relatively small when compared to the large existing lexicographic traditions already incorporated into the VOC database, it is to be expected that only a small amount of national variety-specific lexicon will not already be in the database. This means that the national variety corpora need only to be linked automatically, making it easier for the team to focus on the cases that need the most attention, such as loan words from other languages spoken in those countries, many of which will be registered lexicographically for the first time.

3.4. Work flow

Every time a word is already registered in one of these resources, there is no new entry inserted into the database: instead, an explicit link between the word-form in the corpus and the lemma in VOC is created, similar to the links made for lexicographic resources. This eases the distinction between common and country-specific lexicon, since these links provide information for each word in VOC about its frequency of use in the various national varieties. However, subsequent cross-checking steps always have to be performed.

The remaining words – the inclusion candidates resulting from the differential of each of these resources with the current state of the database – are checked individually and verified manually by teams of lexicographers in the different CPLP countries using the same administrative lexical management processes and tools, namely for determining the corresponding inflectional paradigm of each lemma (Janssen, 2011). Every lexicographer goes through small sets of candidates for insertion at a time, sequentially treating lists of words that share the same formal properties (e.g. class, gender, affix, and theme vowel). This provides for the homogeneous treatment of similar cases and for human resource specialization.

After being manually checked, the data are batch-inserted in small subgroups into the VOC lexical database, with all the corresponding information mentioned before. A number of verification tools are regularly used to cross-check entries already in the database for derivational and formal consistency, where the check automatically suggests potential missing links and eases the process of integrating the sources of different provenance into a common resource. Links between variants are of particular importance to this, complementing the source attestation information. For that task in particular, a dedicated tool is being purposefully developed.

4. Partial and expected results

The project is divided into two parallel phases. The first, ongoing until July 2012, aims at setting up the methodology and tools, compiling the national corpora and integrating most of the European and Brazilian resources. The second phase, to be completed by July 2014, is to see the integration of the new national corpora data and further consolidation between those data and the pre-existing ones. The project aims at reaching a high-quality cross-variant common lexicon with around 300 000 entries in total.

5. References

Ashby, S., Ferreira, J. P., Barbosa, S. (2009). Introducing LUPo: An Accent Independent Pronunciation Lexicon for Portuguese. *Proceedings of Phonetics and Phonology in Iberia 2009*, Las Palmas de Gran Canaria, Spain, 17-18 June 2009.

Berber Sardinha, T., Moreira Filho, J. L., Alambert, E. (2009). The Brazilian Corpus: A one-billion word online resource. In Mahlberg, M., González-Díaz, V., Smith, C., *Proceedings of the Fifth Corpus Linguistics Conference, CL2009*, University of Liverpool, UK, 20-23 July 2009.

Correia, M. (coord.) (2010). *Vocabulário Ortográfico do Português*, http://www.portaldalinguaportuguesa.org/vop.html.

Ferreira, J. P., Barbosa, S., Janssen, M. (2008). MorDebe Admin: a lexicon management system. In *Proceedings of the XIII EURALEX International Congress*. Barcelona: IULA, UPF.

Ferreira, J. P., Lourinho, A., Correia, M. (2012). Lince, an end user tool for the implementation of the spelling reform. In Caseli, H. M., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), *Computational Processing of the Portuguese Language, PROPOR'2012, LNCS*

7243. Springer.

Gonçalves, P. (2000). Para uma aproximação língua-literatura em português de Angola e Moçambique. In *Via Atlântica, 4*.

Gonçalves, P. (2010). *A Génese do Português de Moçambique*. Lisboa: INCM.

ILTEC (2006). *Portal da Língua Portuguesa*, http://www.portaldalinguaportuguesa.org.

Jacobs, D. (1997). Alliance and Betrayal in the Dutch Orthography Debate. *Language Problems & Language Planning*, 21(2), pp. 103-118(16).

Janssen,M. (2005). Open Source Lexical Information Network. Paper presented at *Third International Workshop on Generative Approaches to the Lexicon*. (available at http://maarten.janssenweb.net/Papers/GL2005-mjanssen.pdf, retrieved May 4, 2012).

Janssen, M. (2008). NeoTrack – Un analyseur de néologismes en ligne. In *Actes du 1er Congrès International de Néologie des langues romanes(Cinéo 2008).*Barcelone, Espagne, 07-10 mai 2008.

Janssen, Maarten (2011). Computer-Aided Inflection for Lexicography Controlled Lexica. In Kosem, I., Kosem, K., *Electronic lexicography in the 21st century: New applications for new users*. Ljubljana: Trojina.

Martinez, W. (2012). *SIA, a corpus building and neologism tracking system*. Unpublished tool.

Nunes, M.G.V., Vieira, F.M.C., Zavaglia, C., Sossolote, C.R.C. (1996). *A Construção de um Léxico para a Língua Portuguesa do Brasil para Suporte à Correção Automática de Textos*. ICMSC-USP nr. 42.

Santos, D. (2011). Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa* 3 (2).

Verdelho, T. (2002). Dicionários portugueses, breve história. In Nunes, J. H., Petter, M. (orgs.), *História do saber lexical e constituição de um léxico brasileiro*. São Paulo: Humanitas.