The CPLP Corpus: A Pluricentric Corpus for the *Common Portuguese Spelling Dictionary* (*VOC*)

***Maarten Janssen*[1], *Tanara Zingano Kuhn*[1], *José Pedro Ferreira*[1], *Margarita Correia*[1,2]**
[1]*CELGA-ILTEC, Universidade de Coimbra,* [2]*FLUL, Universidade de Lisboa*
E-mail: maartenjanssen@uc.pt, tanarazingano@uc.pt, jpf@uc.pt, margaritacorreia@uc.pt

Abstract

The Pluricentric Corpus of the Portuguese Language (CPLP Corpus) aims to provide comparable corpora for the national varieties of the countries where Portuguese is an official language, making it possible to undertake corpus-based comparisons among the varieties of these countries. It is intended as a publicly available corpus for comparative linguistics and language resource development, but furthermore constitutes one of the pillars of the *Vocabulário Ortográfico Comum da Língua Portuguesa* (*VOC*), the official spelling dictionary for Portuguese. The headword list in *VOC* is partly derived from lexicographic tradition, which is to date based almost exclusively on the European and Brazilian varieties, and partly made up of words retrieved from the CPLP corpus, many of them included for the first time in official language resources for Portuguese. This double inclusion route aims at presenting an integral (i.e., non-contrastive) and increasingly balanced perspective on all the varieties. This paper describes the general design of the corpus, the challenges faced in its development, as well as the way it was used in the compilation of *VOC*.

Keywords: corpus, pluricentric languages, Portuguese, spelling dictionaries

1 Introduction

Portuguese is spoken by over 260 million people (Reto et al. 2016) in Africa, Asia, Europe and South America. It is an official language of Angola, Brazil, Cape Verde, Guinea-Bissau, Equatorial Guinea, Mozambique, Portugal, São Tomé and Príncipe, and Timor-Leste. Despite this geographical heterogeneity, attention has been given mostly to the varieties of Brazil and Portugal, meaning that language resources for the others are scarce (Branco et al 2012). One of the resources that are missing is a corpus of Portuguese in which these under-represented varieties are given the same status as the varieties from Brazil and Portugal. While some multi-varietal corpora have been built, most notably the Corpus Africa (Nascimento et al. 2008), which covers African varieties, their size and coverage are limited. The corpus described in this article, called the *Corpus Pluricêntrico da Língua Portuguesa* – CPLP Corpus ('Pluricentric Corpus of the Portuguese Language') aims to provide a corpus that represents all varieties equally, rather than one focused on Brazilian and European Portuguese only.

This paper consists of two main parts. The first introduces the CPLP Corpus, which is comprised of sub-corpora from different countries, describing the compilation process, from data extraction to the methods used for cleaning, tagging, and balancing the corpus. Moreover, we will give an overview of the particular challenges concerning the creation of a pluricentric corpus of this type. A general description of the characteristics of the corpus will also be provided.

The second part demonstrates how the CPLP Corpus is integrated into *VOC*, in which it plays a double role. On the one hand, the CPLP corpus constitutes one of the two pillars used to establish the headword lists of the official national spelling dictionaries, which are available as marked-up sub-selections of the entire dictionary. And on the other hand, the online interface of the dictionary uses the

corpus to indicate for each word in *VOC* whether it has a high, low, or medium frequency in a given country. Before moving on to those two main parts, the next section will provide some background of the way the corpus came to be, and the rationale behind it.

2 Background

Portuguese spelling is determined in legally-binding, state-sanctioned documents, which means that in the countries where Portuguese is an official language, official documents written in Portuguese are obliged to follow the official spelling. Until recently, there were two official spelling norms for Portuguese: the *Formulário Ortográfico* ('Orthographic Guidelines') from 1943, which was followed in Brazil, and the *Acordo Ortográfico da Língua Portuguesa* ('Portuguese Language Orthographic Agreement') published in 1945, which was the norm in Portugal, Angola, Cape Verde, Guinea-Bissau, Mozambique, and São Tomé and Príncipe. In 1990, the Portuguese-speaking countries signed an orthographic agreement treaty (*Acordo Ortográfico da Língua Portuguesa - AOLP1990*, 'The Portuguese Language Orthographic Agreement'), with the objective of unifying the official spelling rules in different countries. The actual application of the spelling rules for Portuguese is traditionally made clear through *vocabulários* ('spelling dictionaries'), which up until AOLP1990 were typically national-level and were only published for Brazil and Portugal. In order to support the implementation of the AOLP1990, the countries that signed the treaty projected the creation of a common spelling dictionary for all countries, which is called the *Vocabulário Ortográfico Comum da Língua Portuguesa* (*VOC* – 'Common Spelling Dictionary of the Portuguese Language', Ferreira, Correia, & Almeida (Orgs.) 2017).

The agreement was only officially implemented in 2009, and the development of *VOC* started in the following year, under the supervision of the International Institute of the Portuguese Language (IILP), a multilateral bureau for language policy of the Community of Portuguese-Speaking Countries (CPLP – *Comunidade dos Países de Língua Portuguesa*). During a transitional period in which both the old and new spelling norms were simultaneously accepted, *VON – Vocabulário Ortográfico Nacional* ('national-level spelling dictionaries') were published in Brazil and Portugal following the new spelling rules of the AOLP1990. In the context of the development of *VOC*, these spelling dictionaries were made compatible and integrated into a single framework called OSLIN (Janssen 2005). At the same time, new spelling dictionaries were developed from scratch for the first time for the other countries (*VON* for Cape Verde, Mozambique, Timor-Leste and São Tomé and Príncipe are already available, with development still on-going for Angola and Guinea-Bissau). *VOC* is a free-access spelling dictionary that aims to represent the contemporary lexicon of Portuguese as a whole, in a framework and set-up that is common to all countries in CPLP (Ferreira et al. 2012).

The collection of lexical data for the development of *VOC* involved, among other methods, corpus-based acquisition of lexical entries. It was decided that headword lists of at least 30,000 entries per country would be extracted from lexicographic and para-lexicographic sources and from corpora provided by each country, intending to represent the way Portuguese is used in written contexts at a national level. An evaluation of existing corpora indicated that they were too small to attain this objective, and a decision was thus made to develop new corpora for the compilation of the national spelling dictionaries. The project to create those corpora set common design criteria, such as corpus size, textual genre types, and balance, to guarantee equal representativity among different national varieties of Portuguese (Almeida et al. 2013). However, due to a number of reasons, the compilation of some of these corpora was never finished: in some instances there were political reasons (e.g., Guinea-Bissau), while in others (e.g., Timor-Leste) it was the lack of available digital sources that prevented attainment of the initial objectives. The closest to completion were the corpora for Cape

Verde and Mozambique, which reached over 90% of corpus compilation goals, enough to fully enable their primary intended role in VOC.

Given the unquestionable importance of the existence of corpora for these currently under-studied and under-resourced linguistic varieties of Portuguese, and the fact that a great part of the compilation work and source acquisition had already been started, a decision was made to fully develop these corpora as an independent project, the *Corpus Pluricêntrico da Língua Portuguesa* (CPLP corpus). The CPLP corpus treats Portuguese as a pluricentric language in the sense of Clyne (1992:1), who defines such as language as one with several interacting centers, each providing a national variety with its own norms. Baxter (1992) affirms that Portuguese is a pluricentric language with two standards, the Brazilian and European varieties, and that "[t]he two standards differ from each other in phonology, morphology, syntax, lexicon, spelling and pragmatics" (Baxter 1992:35). Although in official discourse Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Príncipe, and Timor-Leste adopt the European variety as their standard language, it has been shown that some of these countries have emerging autonomous norms (Gonçalves 2010). Nevertheless, the varieties of those countries are not yet fully described and codified. Thus, one of the purposes of the CPLP corpus is to contribute to the study of these under-represented varieties of Portuguese, and serve as a base for the development of representative language resources for them.

3 The CPLP Corpus

The pluricentric nature of the CPLP Corpus is twofold. On the hand, the CPLP Corpus aims to encompass comparable sub-corpora that are representative of the written Portuguese variety in each of the countries of the CPLP. On the other hand, while all tasks related to computational processing, homogeneity of formats, and balancing are the responsibility of a core team, the sources and data for each country are vetted by national-level teams who in the future should manage their own representative corpora. This means multiple centers build and manage the CPLP Corpus.

Furthermore, the CPLP corpus is balanced. On the one hand, country-specific sub-corpora have the same approximate size; on the other, each sub-corpus has the same internal make up, containing comparable representativity of the same textual genres. The balancing over the different genres is given in Table 1.

Table 1: Distribution by genre of the *VON* corpora.

Journalistic	25%	Texts taken mostly from newspapers, but also including magazine texts, taken from online versions where available, or provided by local teams from written printed sources.
Parliamentary	25%	Transcriptions of the parliamentary meetings, partly representing the local formal spoken variety of the language.
Literary	20%	Consisting of prose and provided by local publishers or directly by the authors.
Academic	25%	Academic texts written at the universities of the various countries, typically available through open repositories, and focusing as much as possible on common specific domains (Health, Education, Sea and Environment, Agriculture and Energy, Law).
Other	5%	Privately-produced texts taken from websites within each country, typically blogs.

The differences in available data for different countries is huge, meaning that in order to create a balanced corpus with equal sizes for all countries the varieties with less resources create a bottleneck for

the entire corpus. To make sure that a sizable corpus could nevertheless be provided for the larger varieties, several sub-corpora were defined. The sub-corpora sizes were planned in terms of thresholds, the less-represented countries being the bottleneck for the minimum threshold planned for all sub-corpora (Guinea-Bissau, São Tomé and Príncipe, Timor-Leste): a corpus counting material for all the countries, but with a modest scale of only three million tokens per variety. A second corpus of 30 million tokens per country is the second threshold, for those countries for which this is an attainable goal: Portugal, Brazil, Mozambique, Angola, and Cape Verde. For those benefiting from a greater wealth of data, which is to say Brazil and Portugal, a further threshold was planned for data acquisition in tandem.

An additional challenge that we face derives from the fact that most African newspapers written in Portuguese publish a substantial amount of material from news agencies based elsewhere, especially in Portugal, as extensive analyses have shown. This makes the creation of a clean pluricentric corpus for Portuguese, and likely for any pluricentric language, problematic: there is no easy and reliable way to verify through the data itself whether a text published in, for example, Mozambique, is indeed representative of the variety of Mozambique, or rather copied from or written by authors from different varieties. Moreover, the fact that, contrary to Brazil, Portuguese orthography was already the same before AOLP1990 in Portugal and African countries, means that discards using orthographic variation as a cue for variety distinction is a method that can reliably be used to tell apart Brazilian and European Portuguese texts (see Kuhn et al. 2017). Therefore, in order to keep the sub-corpora as representative of each variety as possible, which is to say, to reduce the number of texts not actually belonging to the variety in question minimum, a decision was made that the CPLP corpus for the African varieties would reject a large number of texts acquired from potentially problematic sources, in some instances giving preference to offline and less easy to process sources, which more reliably represent the variety of the country they were published in.

The CPLP Corpus will be made publicly available for consultation through TEITOK (Janssen 2016), a web-based platform for visualizing, searching, and editing corpora with both rich textual markup and linguistic annotation. A sub-selection of the corpus will be provided through the same platform for download, containing a balanced selection of texts that can be freely distributed.

4 *VOC* Integration

The main initial motivation behind the creation of the CPLP Corpus was the need for a reference corpus for *VOC*, which could provide source material for those countries with no lexicographic resources, along with frequency data for each of its constituting *VON*. To see how the CPLP Corpus was used for this purpose, it is important to highlight some of the structural decisions behind the design of the lexicon, as well as their motivations.

VOC is a reference resource for the implementation of the spelling rules defined by AOLP1990. Previous attempts at an orthographic agreement, namely in 1931, failed at the implementation level: different, national-level wordlists had divergent interpretations of a common legal text or explicitly introduced unilateral changes to the text itself. The contention points are historically rooted in incompatible views on the ideal character of the orthography: conserving traditionalized features or more transparently conveying phonemic information; forcing unique forms for all countries or allowing for country-level, phonemic-induced variation.

While retaining some non-phonemic features, AOLP1990 recognizes the fact that the spelling of Portuguese has a phonemic base, with country-level pronunciation motivating country-specific variants in some contexts. The most visible of these differences is that, much like French, the Portuguese spelling encodes the quality of vowels with diacritics, which motivates variation, since the vowel quality

can be different in Portugal and Brazil in a large class of contexts. Along with this, consonant clusters such as *ct* are written as they are pronounced, and the pronunciation again differs per country. As a result, AOLP1990 defines a number of cases in which the recommended spelling differs per country, as in the case of the vowel quality, where in Portugal 'anonymous' is written as *anónimo*, whereas in Brazil it is written as *anônimo*. For consonant clusters, in Brazil you write 'fact' as *fato*, whereas in Portugal you write *facto*. Yet, in purely legal terms, all spellings that are legally correct in one country are legally correct in all countries, allowing for common legal documents. So, in AOLP1990 there is a fundamental difference between a recommendable spelling (which is specific to each country), and a legally acceptable spelling (which is common to every country). As such, the online interface of *VOC* can display the entire accepted lexicon, showing all words for all countries, but by default will display the *VON* for the country from which is it being consulted.

Since these country-specific spellings have to reflect the pronunciation and, to some degree, the orthographic tradition within a country, it would not only be legally incorrect, but also impractical to have a central team define the spelling in each country. Therefore, a guiding principle in *VOC* is that each country is responsible for its own *VON*, both in terms of which words should form part of the core vocabulary for that country, and in the definition of what are the acceptable spellings of those words in the case of spelling variation.

For the development of *VON* a local lexicographic team was established in each country, responsible for the final validation of the contents of *VON* and for the definition of the sources to be taken into account for the constitution of the nationally-representative headword lists. Those data comprise the complete set of words already included in *VOC*, and all the words included in the (known) lexicographic tradition for that country. On top of this, the frequency of each of the words in the entire database in the corpus for that country, which is to say the frequency of each word in the sub-corpus of the CPLP corpus for that country, was also considered.

The corpus frequencies were provided as a guiding principle, and each local team was free to tailor the headword list independently of the frequency of each word in the corpus. For lexical selection, this means each *VON* is corpus-driven, but ultimately involves some degree of traditional lexicographic handpicking. And for the case of spelling variation, each local team had to decide whether only one or several of the officially allowed variants was correct for their *VON*. For some of those cases, corpus frequencies were of no use, since many of the words in question were among those having their spelling changed with the orthographic agreement (e.g. in every country apart from Brazil, *facto* would take that form regardless of its pronunciation, <ct> being orthographically opaque). For all those words used in the local variants (often loan words from languages spoken within the country, but not yet lexicographically registered), the corpus was the only legitimizing force, enabling the representation in official sources, for the first time, of a large number of perfectly valid and frequently used words in several countries.

Apart from being a guiding principle, the CPLP Corpus plays a second role in the online interface of *VOC* when displaying a specific *VON*. In order to give an indication of how common a word is, the system displays whether the frequency of the word is high, medium, or low in the corpus, where high means it is amongst the 10% most frequent words, medium means it is amongst the top 40%, and low means anything below that. These data are presented in the interface directly from the sub-corpus of the CPLP Corpus corresponding to the selected *VON*.

5 Conclusion

The CPLP corpus is the first pluricentric corpus for Portuguese of a substantial size, large enough to guide the development of pluricentric lexicographic material. It grew out of a corpus designed for the

creation of the official *VOC* spelling dictionary for Portuguese in all countries of the CPLP, which from 2010 onwards provided one of the fundamental resources for the creation of the *VON* for each country. The CPLP corpus finishes what that initial corpus set out to establish, but could not fulfil at the time, which is to be a pluricentric reference corpus that can be used to give corpus-driven usage information on words in the various countries, hence implicitly providing information about whether words are used internationally or specific to a given variety. In time, the CPLP corpus will replace the initial *VOC* corpus as the basis for the frequency information in *VOC*. Moreover, it is also intended to be a basis for future pluricentric lexicographic resources for Portuguese.

The CPLP corpus can be used not only for much-needed lexicographic work, but also for broader linguistic research, such as comparative studies of the different varieties of Portuguese, thus making it a highly relevant resource for the study of Portuguese as a pluricentric language and highly valuable for future lexicographic work, given that there is no single existing dictionary for varieties other than those of Brazil and Portugal. Future work includes further integration of the CPLP Corpus into *VOC*, by extending its role to directly present a selection of example phrases in the corresponding sub-corpus for each word in a given *VON* that is registered in the corpus.

References

Almeida, G. B., Ferreira, J.P., Correia, M. & Oliveira, G.V. (2013). Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. In *Estudos Linguísticos*, 42(1), pp. 204-215.

Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., … Bacelar, F. (2012). *The Portuguese Language in the Digital Era/A Língua Portuguesa na Era Digital.* White Paper Series. Springer.

Baxter, A. (1992). Portuguese as a pluricentric language. In M. Clyne (ed.) *Pluricentric languages: Differing norms in different nations*. Berlin, New York: Mouton de Gruyter, pp. 11-43.

Clyne, M. (1992). Pluricentric languages – introduction. In M. Clyne (ed.) *Pluricentric languages: Differing norms in different nations.* Berlin, New York: Mouton de Gruyter, pp. 1-9.

Ferreira, J.P., Correia, M. & Almeida, G. B. (orgs.) (2017). *Vocabulário Ortográfico Comum da Língua Portuguesa*. Praia: Instituto Internacional da Língua Portuguesa / Comunidade dos Países de Língua Portuguesa.

Ferreira, J.P, Janssen, M., Almeida, G. B., Correia, M. & Oliveira, G.M. (2012). The Common Orthographic Vocabulary of the Portuguese Language: A set of Open Lexical Resources for a Pluricentric Language. In *Conference on Language Resources and Evaluation (LREC)*, Istanbul, pp. 1071-1075.

Gonçalves, P. (2010). *A Génese do Português de Moçambique*. Lisbon: Imprensa Nacional/Casa da Moeda.

Nascimento, M. F. B., Pereira, L. A. S., Bettencourt, J., Estrela, A., Oliveira, S., & Santos, R. (2008). *Corpus* África: as cinco variedades africanas do português. In S. Frota, A. L. Santos (eds) *Textos Seleccionados. XXIII Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa: APL, pp. 373–384.

Kuhn, T. Z., Janssen, M., Ferreira, J.P., Kosem, I. & Correia, M. (2017). Dealing with multiple orthographic standards within a single corpus: the case of Portuguese in the CoPEP corpus. In *Actes des 9èmes Journées Internationales de la Linguistique de corpus*, Grenoble, pp. 52-54.

Janssen, M. (2005). Open Source Lexical Information Network. In P. Bouillon, K. Kanzaki (eds.) *Proceedings of the Third International Workshop on Generative Approaches to the Lexicon, May 19-21 2005*. Genebra: École de Traduction et d'Interprétation – Université de Genéve, pp. 79-106.

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In *Proceedings of LREC 2016*. Portorož, Slovenia, pp.4037-4043.

Reto, L., Machado, F.L & Esperança, J.P. (2016). *Novo Atlas da Língua Portuguesa*. Lisboa: Imprensa Nacional-Casa da Moeda.