DOI: 10.30816/ICONN5/2019/67

A proposal for a modern multilingual gazetteer for the Portuguese-speaking countries

Agostinho Salgueiro
CELGA-ILTEC, University of Coimbra

José Pedro Ferreira
CELGA-ILTEC, University of Coimbra, Portugal

Margarita Correia
FLUL, University of Lisbon, Portugal

Abstract: A recently enforced orthographic agreement for Portuguese ended a century-long linguistic standardization divide within Portuguese-speaking countries. Over the past few years, for the first time ever, the development of shared language tools for all CPLP countries has been in progress. This includes the first standardized common toponymic resource, *Vocabulário Toponímico* (VT).
But what about other languages sharing their space with Portuguese? Official toponymy in the CPLP space is pervasively recorded only in Portuguese, and in Cabo Verde, Guinea-Bissau and Timor-Leste, for instance, the most widely spoken languages are all but absent from gazetteers and place-name signs, perhaps by virtue of those languages largely sharing their lexical base with Portuguese.
In this paper we address the conundrum of developing the first common normative toponymic resource for many Portuguese-speaking countries. We share arguments as to why it can be used as an adequate model for the toponymic codification of other languages spoken in the CPLP, as a common toponymic data platform and, from there, as a tool that can be used to strengthen linguistic diversity.
Keywords: toponymic standardization, multilingual gazetteer, *Vocabulário Toponímico*, Portuguese.

1. Introduction

In this paper we outline the history of the orthographic normalization of Portuguese, until the present common norm, the first applied in the context of the Community of Portuguese-speaking Countries (CPLP). We emphasize the fundamental role of having a valid single orthographic norm for the elaboration of multinational monolingual linguistic resources, as it alleviates the constrains commonly associated to the existence of political borders, and we use, as models to follow, the *Vocabulário Ortográfico Comum da Língua Portuguesa* (VOC) – the official vocabulary for Portuguese – and, in particular, the *Vocabulário Toponímico* (VT)[1], the official toponymic resource for Portuguese, contained in VOC.

Starting from a digitally implemented monolingual toponymic resource, we advocate extending VT to a shared-base multilingual tool. The VT was developed for Portuguese and for Portuguese users/learners, but it can be made available to communities outside of the CPLP, enabling the improvement of our own work with feedback from onomastician peers and simultaneously presenting an effective platform for the potential improvement of (i) older projects, (ii) ongoing research projects and (iii) related work still to be developed for other languages. As so, this contribution starts by proposing a common basis for normalized toponymic principles to deal with CPLP's national languages, leaving an open window for parallel systems containing toponymic data in other countries' languages to be included *a posteriori*.

Thus, in the current phase – the one that follows the making of the first versions of VOC and VT – we understand that, with respect to toponymy, the natural follow-up involves providing CPLP's languages other than Portuguese with a platform for harmonized toponymic data, according to current linguistic norms and with a structural basis already tested with Portuguese. It is with this mindset that we propose VT as a common platform for the public availability of standardized toponymy from CPLP's languages. The work already developed for Portuguese, its general public acceptance and the ease of including new language systems in the database are the pillars of our proposal. This work, as with VT's prerogative, focuses on universal, free and user-friendly access to normalized and standardized toponymic data.

2. Portuguese: 12th century, 1911, 1990, today

Although Portuguese has been written for over 800 years, language coding, under a spelling norm, is relatively recent. The first official orthographic norm (FOLP11) was applied in 1911 – and led to a split between Brazil and Portugal during the twentieth century, with advances towards and setbacks in the continuous goal of achieving a common orthography. The existence of two official orthographies to which different political spaces were bound had deep implications for the way the language unity was perceived.

After several agreements that never actually got to being applied, a proposal signed in 1990 (AOLP90) by all the then-independent Portuguese-speaking countries has been put in practice in a gradually larger number of countries since 2008. AOLP90 is virtually applicable in all Portuguese-speaking countries, namely Angola, Brazil, Cabo Verde, Guinea-Bissau, Equatorial Guinea, Portugal, Mozambique, Sao Tome and Principe and Timor-Leste[2]. As such, it is an essential tool for the unity in diversity of this pluricentric language.

[1] *Vocabulário Toponímico* is available at https://voc.cplp.org/index.php?action=toponyms.

[2] CPLP gathers over 260 million people, and this number is growing at a fast pace, especially due to high birth rates in the community's countries located in the Southern Hemisphere.

3. The first common toponymic resource for Portuguese

With a single orthographic norm, the possibility to develop transnationally shared language resources became a reality. Multilateral efforts arose from the recognition of the need for a common spelling dictionary, and its making was carried out within the context of the CPLP, including experts from all its countries, who formed autonomous national teams working in cooperation with a central coordinating team. VOC, in its first version, included more than 300 000 entries encompassing national varieties of the language, with explicit marking of existing variation between countries. For reference, it can be described as "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" (Ferreira *el al.* 2012: 1072).

At an early stage of the development of VOC, it was decided to proceed with the selection and systematic treatment of toponyms within the CPLP, recognizing the relevance such words have in the lexicon of any language, namely the fact that "[names] support structuring geographic space, since name and concept are mostly closely linked" (Jordan 2012: 125). A properly developed toponymic database allows users to clearly associate orthographic forms (the place names) with the places they designate (the geographic referents). Normalized and standardized toponymic databases are paramount for the functioning of any modern state and its institutions.

Bearing in mind the advantages associated with a tool that enables countless applications across State structures to serve populations, VT, a hierarchical system with relational subsets composed of toponymic normalized and standardized synchronic data, was born at the heart of VOC, with nearly 73 000 toponyms in its first version. This number was reached through the inclusion of (i) toponyms with administrative relevance in the participating CPLP countries, mostly obtained from databases of the national statistical institutes of each country; (ii) the names of countries and capitals of non-CPLP countries, obtained through data of the United Nations and from the collaboration with European Union institutions, specifically with experts involved in the drafting of the EC's Inter-institutional Style Guide for Portuguese.

For the present proposal, we do not make a comprehensive description of VT's features. Nonetheless, we share some information that is relevant to characterize our toponymic coding model. Users have access to several types of information in VT, as shown in Figure 1 (formal properties of toponyms), Figure 2 (type of entity and division level) and Figure 3 (geographic code). Every VT entry provides:

a) the formal properties of toponyms:

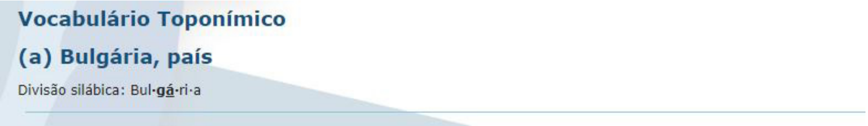
(i) word class;

(ii) grammatical gender, including non-marked[3] ((a), (o) or (Ø));

(iii) syllable division (a dot marks each syllable boundary);

[3] Portuguese nominal forms have one of two grammatical genders, masculine or feminine, which can be determined through their usage in context. Nevertheless, in the case of toponyms,

(iv) word stress (stressed syllables are underlined and written in bold).



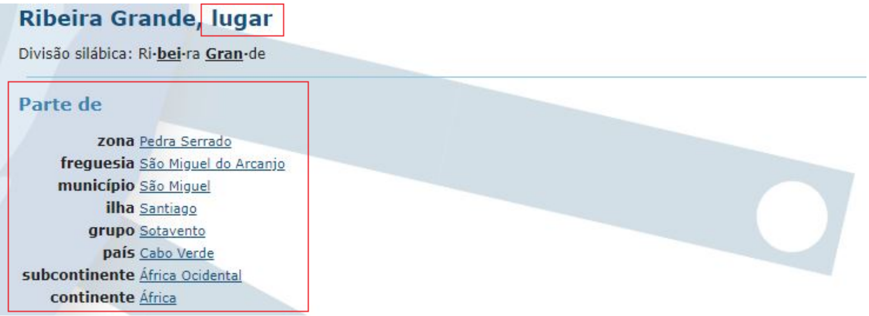
Vocabulário Toponímico

(a) Bulgária, país

Divisão silábica: Bul·**gá**·ri·a

Figure 1. Example of the layout of the formal properties of toponyms in the VT.

b) encyclopedic information:
(i) type of entity (these types range from *planet* to *locality*);
(ii) division level (position within the hierarchical system);
(iii) geographic code (available explicitly through the URL).



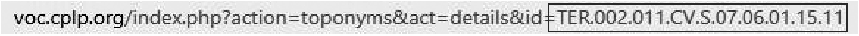
Ribeira Grande, lugar

Divisão silábica: Ri·**bei**·ra **Gran**·de

Parte de

zona	Pedra Serrado
freguesia	São Miguel do Arcanjo
município	São Miguel
ilha	Santiago
grupo	Sotavento
país	Cabo Verde
subcontinente	África Ocidental
continente	África

Figure 2. Type of entity and division level: *Ribeira Grande* as an example.



voc.cplp.org/index.php?action=toponyms&act=details&id=TER.002.011.CV.S.07.06.01.15.11

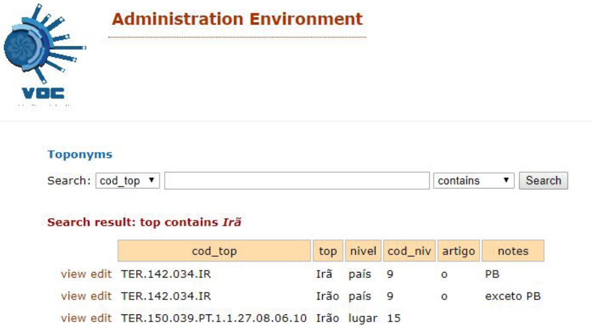
Figure 3. Geographic code: *Ribeira Grande* as an example.

As for the data codification model for the lexicographic micro-structure, each entry encompasses a mandatory set of information, which includes:
(i) a written form (the toponym itself);
(ii) a code adhering to *ISO 3166–1 alpha 2* and M49 (three digit) codification;
(iii) a unique identifier, or ID;
(iv) the toponym hierarchical level code;
(v) the name of the hierarchical level of the toponym;
(vi) gender data;
(vii) a reference to national usage restrictions.

The resource has its own administration environment, where the above set of information is easily accessible and editable. The tags *top*, *cod_top*, *cod_niv*, *nivel*

grammatical gender is not always marked. We assume that a toponym has non-marked gender when it does not accept the use of an article.

(level), *artigo* (article) and *notes* account for the data of each toponym[4]. Every entry contains a toponym-code pair, meaning that there can be two identical forms pertaining to different toponyms, and two toponyms for a single geographic entity. When one code is shared by two toponyms, the unique ID feature includes a) the pair *cod_top/top* or b) the pair *cod_top/artigo*. These cases also allow us to easily identify toponyms that refer to the same geographic entity but are used in different varieties of Portuguese. As for gender data (vi), tagging of entries for hierarchically lower levels is still to be developed. These levels include toponyms that are not so commonly used, so the access and selection of this type of information must be done not only with conventional sources, such as encyclopedias and *corpora*, but for the most part by painstakingly harvesting them from other existing resources, such as official Web pages of municipalities, or through time-consuming field work in specific parishes.



Toponyms

Search: cod_top ▼ | | contains ▼ | Search

Search result: top contains *Irã*

	cod_top	top	nivel	cod_niv	artigo	notes
view edit	TER.142.034.IR	Irã	país	9	o	PB
view edit	TER.142.034.IR	Irão	país	9	o	exceto PB
view edit	TER.150.039.PT.1.1.27.08.06.10	Irão	lugar	15		

Figure 4. Data codification model on the VT. *Irã-* as an example.

4. Portuguese as a reference for the development of national toponymic databases

We have seen above that the handling of different types of information associated with toponyms is an important part of VT, a resource that aims at being both as in-depth and as user-friendly as possible. In the process of compiling and compartmentalizing this information in the database it is essential to bear in mind the existence of internal variety and external variation in a pluricentric language as widely spoken as Portuguese. By *external variation*, we mean the variation that occurs between toponyms of two different varieties, each used in a different country (in many cases, this type of variation has the toponym used in the Brazilian variety on one side and the toponym accepted in the remaining CPLP countries on the other). We can take as an example the pair *Romênia/Roménia*, where the first form is used exclusively in the Brazilian variety and, simultaneously, the second one is only unacceptable in this variety of the language.

[4] In addition to the fields already available in the lexicographic micro-structure of the VT, a new one, labeled as *lang*, can be developed to identify each subsequent language to be included in the database.

This is due to the fact that the orthographic rules in the AOLP90 allow for some variation, e.g., in this instance, different diacritics to account for the height value of certain vowels in a stressed position. With regard to *internal variation*, we refer to cases where a geographic referent may have different designations in the same country, which in some instances stems from distinct languages inspiring the toponymic form. The pair *Inhaca/KaNyaka* serves to illustrate this reality: both forms are used in Mozambique, each reflecting a particular history and take on etymology. The first form, *Inhaca*, is the one used in Portuguese, and <c> is used because it is one of the two prototypical ways to represent [k] in the writing system of this language (the other one is <qu>). The choice between one of two possible toponyms, such as in the example *Inhaca/KaNyaka*, is mainly context dependent, i.e., it depends on the linguistic context.

So, given this context, from a codification perspective and as users of a pluricentric language, how can we address in VT other languages sharing their usage space with Portuguese? While Portuguese is overwhelmingly dominant in Brazil and Portugal, that is not the case in most regions of some other CPLP countries. Despite this, official toponymy is pervasively established only in Portuguese. This is especially striking in Cabo Verde, Guinea-Bissau and Timor-Leste, where the most widely spoken languages are all but absent from gazetteers and name signs, perhaps by virtue of them largely sharing their lexical base with Portuguese and due to historical reasons.

The fact that this phenomenon occurs is a reflection of what commonly happens between languages that are phylogenetically related but, while used in tandem, have a different political status: resorting to the State language poses some clear advantages for users from a practical point of view without compromising intelligibility, even if some of these advantages are hampered by homograph expressions, homophone pairs or simply by structurally close lexical items having different types of semantic information associated with them. In the case of toponyms, semantic questions such as these do not arise in the same way, but it is still extremely relevant to be able to know when similar names, sometimes homophones, pertain to different languages. This task can be achieved easily by adding a tag to all entries of VT, for instance in a field identified as *lang*, as suggested above, one that identifies the current nomenclature as belonging to Portuguese. The same can then be made to each additional set of toponyms pertaining to a different language. This way, two toponyms that identify the same geographic referent and which belong to exactly the same hierarchical slot are not only associated with different languages but also have a unique ID, an unambiguous language identifier.

Hundreds of languages other than Portuguese are spoken in the CPLP countries as a whole. It would be unrealistic to aim to codify the toponymy of all of them within any single project, even more so in the short or medium term. It would make sense to use the existing platforms to start working on the languages that are closest to Portuguese and, simultaneously, the ones for which there are existing resources, or which enjoy thriving usage nationally. Cabo Verde Creoul, Guinea-Bissau Creoule and Tetum, from Timor-Leste, are obvious aims to start the enlargement and enrichment of VT, because they do largely share their lexical base with Portuguese and because there is a pressing need for normative national resources in these languages, and not only for the resources being developed for the three Portuguese varieties enjoying official status in each of these countries. Such a task would require steadfast diplomacy efforts and international partnerships, namely between the International Portuguese Language Institute (IILP), research institutions in specific countries and other national institutions involved in toponymy, such as Mozambique's INGEMO, the Brazilian IBGE or the Portuguese INE. From there, the platform provided by the VT can lead to an implementation process that suits all the participating countries' linguistic needs and, at the same time, ease the publication and dissemination of the data to each national population, providing a basis to protect multilingualism across the CPLP.

5. One toponymic database, a means for the protection of linguistic and cultural diversity

The fact that there is a fully implemented official toponymic resource for Portuguese can be an opportunity for countries interested in documenting and codifying national languages besides Portuguese with a view to protect them and foster their development. The use of a model such as VT's does not necessarily cannibalize other languages and can in fact be a powerful tool towards improving their status. It provides a platform with unique potential for consolidation at the national level, for cross-border visibility and for international recognition resulting from it[5]. One of the great advantages of the VT as a common platform lies in its continuous development, that is, the fact that it can always incorporate parallel language systems without distorting any of the ones already developed. In fact, the more diasystems are incorporated into the VT, the greater is the potential of the resource (i) for comparative studies, (ii) for the update of language policies in multilingual geographic areas and notably (iii) for the implementation of bi/multilingual place-name signs.

Like Jordan (2012: 124), we see language as an important part of group identity, and bi/multilingual place-name signs are one of the closest physical manifestations of the multicultural and multilingual richness available to populations: they reflect the political openness to toponymic coding based on the respect for diversity. It is often through such pieces of stone, wood or metal that language users ingrain the official status of place names, the places they go to, where they live. Place-name signs that account for more than one language contribute, in the long term, to the acceptance of cultural differences among people living in the same neighborhood[6]. The existence of bi/multilingual toponymic signs can, moreover, be used as a tool to alleviate conflicts with different genesis. By doing so, we can easily agree that the

[5] This model would also be adequate for any other pluricentric language, including English, which, considering its status, would be most relevant. However, the absence of a common orthography for English greatly hinders such purpose, at least at a transnational level.

[6] "The identity of every person is composed of multiple layers […]. Group identity is composed of language, religion and all the other cultural elements shaping a social group" (Jordan 2012: 124).

advantages of being proactive in regard to the implementation of multilingual public place-name signs should not put up much debate.

If the VT structure is to be used as a base for a common toponymic tool with parallel linguistic systems that share geospatial referents, the homogeneous standardization criteria adopted in the VT should be seen as the basis of an effective effort to protect linguistic diversity and for the maintenance of multicultural features. This work starts from stakeholders in each community, namely political representatives and official institutions, e.g., through the objective of having multilingual place-name signs. In parallel, the *modus operandi* behind such an endeavor would ideally preclude a narrow collaboration between experts working with each language and community on the same platform and a technical team working towards integrating the data and ensuring adequacy and homogeneity.

A shared toponymic platform that is both standardized and multilingual can present several advantages and cover multiple use-cases, of which a few stand out. Firstly, it can be a powerful tool for the consolidation and protection of languages that are not spoken widely and that are more prone to be engulfed by more well-resourced and more widely spoken majority/official languages. A shared toponymic platform ensures that such less widely spoken languages are represented at the highest level, putting them on a par with the dominant languages they would share a platform with. Secondly, being represented in such a high-status resource could aid the reinforcement of identity roles based on the availability of normalized data not only for local languages, but also for more widely spoken ones sharing the same platform. Thirdly, such platform would also enhance international visibility of the less widely spoken languages included in it, as researchers and users in general can have simplified access to data pertaining to those languages when they use the resource with a view to obtaining data for the majority language.

Lastly, taking the CPLP as an example, such a resource can add cohesion within a larger community, bringing closer different communities and cultures from distinct countries. The replication of this model for other languages outside of the CPLP area would enable simplified access to parallel data, easing comparative studies and mass-scale linguistic contact studies, providing otherwise inaccessible data potentially interesting for fields such as history and anthropology.

References

Almeida, G.M., J.P. Ferreira, M. Correia, and G.M. Oliveira. 2013. Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. *Revista Estudos Linguísticos* 42 (1): 204–215.

Correia, M. 2015. Topónimos e ortografia – os topónimos no espaço público. *Congress about the National Languages of Angola.* Benguela: Piaget Benguela.

Correia, M., J.P. Ferreira, and G.M. Almeida. 2020. A gestão da ortografia da língua portuguesa: do desencontro ao Vocabulário Ortográfico Comum da Língua Portuguesa. *Estudis Romànics* 42: 277–286. DOI: 10.2436/20.2500.01.297

Ferreira, J.P., M. Janssen, G.M. Almeida, M. Correia, and G.M. Oliveira. 2012. The Common Orthographic Vocabulary of the Portuguese Language: A Set of Open Lexical Resources for a Pluricentric Language. In *Proceedings of the Conference on Language Resources and Evaluation (LREC): Vol. 2012 Proceedings*, 1071–1075. Istanbul: n.p.

Jordan, P. 2012. Place Names as Ingredients of Space-Related Identity. *Names and Identities (Oslo Studies in Language)* 4 (2): 117–131.

Jordan, P. 2019. The Endonym/Exonym Divide from a Cultural-Geographical Perspective. *Language and Society* 10: 5–21.

Kerfoot, H. and E.M. Närhi. 2006. *Manual for the National Standardization of Geographical Names.* New York: United Nations Publication.

Ormeling, F. 1993. Exonyms in Cartography. *UNGEGN Training Course in Toponymy for Southern Africa.* University of Pretoria: Pretoria.

Radding, L. and J. Western. 2010. What's in a Name? Linguistics, Geography and Toponyms. *The Geographical Review* 100 (3): 394–412.

Salgueiro, A.M. 2016. Topónimos no espaço da CPLP: o vocabulário toponímico. MSc diss. ISCTE-IUL, Lisbon. https://repositorio.iscte-iul.pt/handle/10071/12495 (accessed in September 2019).

Tichelaar, T. 2002. *Toponymy and Language*. Frankfurt: United Nations Publication.

Woodman, P. 2012. Endonyms, Exonyms and Language Boundaries. In *The Great Toponymic Divide. Reflections on the Definition and Usage of Endonyms and Exonyms*, P. Woodman (ed.), 75–78. Warsaw: Główny Urząd Geodezji i Kartografii.