Agostinho Miguel
Magalhães Salgueiro

agostinhosalgueiro@uc.pt
https://orcid.org/0000-0003-1077-9911
CELGA-ILTEC, University of Coimbra
Coimbra, Portugal

https://doi.org/10.4467/K7501.45/22.23.18075

Toponymy and Grammatical Gender: A Description From Portuguese

Abstract

One of the most common questions regarding the proper use of a toponym in Portuguese is related with the (i) obligation, (ii) possibility or (iii) interdiction to employ an article as a toponymic gender marker. Every Portuguese speaker acknowledges that the only possible position for an article attributing gender to a place name is to its left; also, it is well known that a mandatory or possible *gender article* in Portuguese is never a constituent of the place name it precedes. Nevertheless, language users still struggle to draw general rules that allow them to better understand the grammaticality of gender articles preceding toponyms. As one would expect, the less familiar a toponym is to a speaker, the harder it becomes for them to predict its gender value.

In Portuguese, toponyms derived from nouns (synchronically transparent place names, mainly) are the ones commonly labeled as more prone to be preceded by an article, but no extensive research has ever been done to evaluate if this assumption is, in any way, corroborated by user-based data or by data extracted from official toponymic resources. In this paper, considering data from the official resource for Portuguese toponymy, the *Vocabulário Toponímico*, a set of rules is drawn describing some mandatory, possible or unauthorized interactions between gender articles and toponyms.

Keywords

Portuguese toponymy, grammatical gender, toponymic gender-labeling, *Vocabulário Toponímico*

1. Introduction

Grammatical gender is a formal property of toponyms, and different linguistic systems can allow the association of either one or more values to it. Contrary to what can be observed in languages that only accept gender-neutral toponyms, like English, Portuguese has two possible gender values:

(1) **Gender value 1**: 'if a toponym is masculine or feminine'.
E.g.: *O Porto*[1] *é uma cidade linda. / *[ø] Porto é uma cidade linda.*
'Porto is a beautiful city.'

(2) **Gender value 0**: 'if a toponym is gender-neutral' (neither masculine nor feminine).
E.g.: *[ø] Brasília*[2] *é a capital do Brasil. / *A Brasília é a capital do Brasil.*
'Brasilia is Brazil's capital.'

In both oral and written utterances, the gender value of a Portuguese toponym can always be inferred. However, if a toponym is pronounced/written with no context, this property does not come to the surface, that is, without grammatical context, even native speakers can struggle to identify the gender of an endonym they are not familiar with. Thus, a **familiarity** trace is

[1] Portuguese city names are mainly gender-neutral. *Porto* is one of the few exceptions, being a homonym of the masculine noun *porto* 'harbor'. Interestingly, exonyms like the English *Oporto* disclose how a language that only has gender-neutral toponyms can process a definite article that marks gender in the departure language but is not part of the endonym, *per se*: in this case, English neglects the article's masculine trace by simply amalgamating the vowel *o-* in the beginning of the place name *Porto* and capitalizing it.

[2] *Brasília* derives from the country's name, *Brasil*, but both toponyms have different gender values: *Brasil* (masculine) is also the name of a type of wood; *Brasília* (gender-neutral) shares the morphological radical with *Brasil* but incorporates the suffix *-ia* (which is a productive one in the formation of Portuguese toponyms). The fact that *Brasília* derives from an endonym (and one with such importance – at country level) may trigger a gender-neutral value whose core function is to internally underline that the named geographic referent is not at the same hierarchical level as the country, avoiding possible ambiguity derived from a transparent morphological proximity. Furthermore, contrary to what the literature describes, the final *-a* has no prominent role in gender attribution – in this example, its addition does not trigger a (feminine) gender value 1.

key for the homogeneous use of unique gender values associated to each Portuguese toponym.

Simplex place names ending in *-a*(*s*) are usually assumed as being feminine. When they end in *-o*(*s*), they are generally interpreted as masculine. As for complex place names, the same rule tends to be applied, but with a regular focus on the place name's core unit, usually a proper noun, because the gender value of a toponymic constituent that is homonymous to a noun is always transparent. Even so, this inference cannot be taken as a flawless criterium for the proposal of a general rule that accounts for an accurate gender value associated to each Portuguese toponym. If it could, we would not find statements in Portuguese reference grammars like the one in Raposo et al. (2013) that it is a frequent reason for conversation which names of countries allow, require or do not admit the definite article, as well as whether or not there is any generalization that regulates this use (and if it exists, it has not yet been discovered) (p. 1018).

I propose that the only accurate way to access gender information associated to a toponym in Portuguese – to all toponyms, not only to country names – is by verifying if the name can be combined with a preceding article and, if it can, by clearly identifying if the article is masculine or feminine, sidelining the thematic index when one is available (often as a homograph form of the article that can precede it). By doing so, the grammatical **gender value 1** of Portuguese toponyms will follow from a gender feature located in a nonconventional nuclear unit: an article. For this reason, within the scope of this paper the term 'gender article' is adopted, underlining that, when it is acceptable, a gender article always **precedes** a place name – it is never a constituent of the simplex/complex name. When it cannot precede a place name, the name has **gender value 0**. In this regard, it is important to note that three country names, *Espanha* 'Spain', *França* 'France' and *Itália* 'Italy', now accept the previously interdicted feminine gender article (at least in the standard norm of European Portuguese). As so, each of these countries presently can have two homonymous endonyms, one that is feminine (**gender value 1**) and another one that is gender-neutral (**gender value 0**).[3] This duality can result

[3] The distinction and validation of two homonymous toponyms, based on different gender values, acquiesces with Villalva (2000), who wrote that the so-called uniform nouns, such as *artista* 'artist', are responsible for the occurrence of different agreement marks in syntactic constituents that specify, modify or predicate them, and thus proposed that they have two gender specifications, that is, that the gender contrast involves two lexical units (p. 227).

from a regularization through a process of synchronic analogy. That is, the ending *-a* in these frequently used country names (the same is true with the toponym *Inglaterra* 'England') influences the acceptability – and the resulting grammaticality – of a gender article (Salgueiro, 2016, p. 26), but this recent acceptability still does not outpace the gender-neutral homonym. These examples are also important to bear in mind, even in this theoretic frame, where they are understood as secondary, because they do show that toponyms' thematic indexes have some of the same gender-defining features as Portuguese nouns' thematic indexes.

Furthermore, this **gender article** assessment can be validated in all the language's national varieties, even if there can be some punctual output variation, namely between the Brazilian norm, on one hand, and the national norm of each other country that integrates the Community of Portuguese Language Countries (CPLP) (e.g.: [*o*] *Timor-Leste* / [*Ø*] *Timor-Leste*).

2. Types of constraints governing toponymic gender values

I propose that a gender article in Portuguese is always governed by one of three types of constraints. Hence, when a toponym is used in a given utterance, a (1) **mandatory**, (2) **possible** or (3) **interdicted** constraint is activated, verifying the grammaticality of a gender article.

(1) **Mandatory**: 'a gender article must be used in an utterance'.
E.g.: *Cheguei a**o** Porto / *Cheguei a[ø] Porto.*
'I have arrived at Porto'

(2) **Possible**: 'a gender article can be used in an utterance'[4].
E.g.: *Sou de/d**o** Pombal*; *Sou de/d**a**s Lajes das Flores.*
'I'm from Pombal'; 'I'm from Lajes das Flores'

[4] When the pair of place names with two possible values is illustrated, the most frequent gender value is presented first.

(3) **Interdicted**: 'a gender article cannot be used in an utterance'.
E.g.: *Nasci em* [ø] *Lisboa.* / **Nasci na Lisboa.*
'I was born in Lisbon'

This **type system** is useful not only to establish a clear distinction between existing types, but also to underline that each toponym is covered by only one of the existing types. It should be noted that if a geographical referent's name can have two different gender values, there are, in fact, two homonymous place names – in this case, two endonyms – to be considered (see (2), immediately above), under the premise that grammatical gender is a formal property of toponyms. Of course, if the article would be a *de facto* constituent of one of the two toponyms, this assumption would be much more intuitive, namely to non-native Portuguese speakers/learners.

Looking at the place name given above as an example, *Porto*, it is common to have English native speakers with Portuguese as an L2 produce ungrammatical utterances like (i) **Visitei Porto*; (ii) **Vim de Porto* or (iii) **Cheguei em Porto*, instead of (i') *Visitei o Porto* 'I visited Porto'; (ii') *Vim do Porto* 'I came from Porto' or (iii') *Cheguei ao Porto* 'I arrived at Porto'. This deviation happens because the only possible gender value in English, the neutral, is inadvertently assumed as also exclusive in Portuguese, and the speaker fails to verify the article constraint **mandatory** in the endonym *Porto* (in this case, [+masculine]).

Even to native speakers, the main difficulty that arises when dealing with types of constraints that govern the grammaticality of gender articles in Portuguese has to do with the linguistic output's frequent opacity. Cartography is one of the most common tools vehiculating the use of place names, and maps or georeferenced tools, such as Google Maps, don't provide the necessary linguistic context to language users, they just show the toponym by itself, stamping it on the correspondent geographic referent. Thus, some of the most vastly used tools that include toponyms do not convey any type of linguistic information that could cement a general rule for the understanding of gender values and types of constraints in Portuguese (other than the insufficient, already referred, thematic indexes of each toponym's lexical units).

3. Compositionally opaque names vs synchronically transparent names

In Portuguese, toponyms that derive from nouns (synchronically transparent place names, mainly) are ones commonly accepted as more prone to have a mandatory gender article, because their semantic motivation is easily accessible (or so it seems to a speaker that sees them as having a descriptive base). On the other hand, toponyms that are compositionally opaque to a contemporary language user – usually names described in Raposo et al. (2013) as having no descriptive meaning or being semantically arbitrary (the author's **canonical names**) – are expected to activate the **interdiction type** to gender articles.

In short, place names in whose lexical constituents it is (still) possible to identify a homonymous form of a noun, it being the nucleus of the NP, will, according to the literature, predictably accept the employment of a definite article (as expected, the masculine or feminine gender will accrue from the gender of the noun that originated it). Raposo et al. (2013) uses *Reino Unido* 'United Kingdom', as an example to illustrate that the employment of a definite article can be expected when toponyms are formed on the basis of a defined description whose core is a noun with which the article agrees in gender and number (p. 1019). But if it were this simple to acknowledge the acceptable type of gender article constraint, the mapping of the biunivocal relations between each compositionally transparent toponym and its gender value would not pose any difficulties to the implementation of automatic labeling processes focused on this formal property. Unfortunately, this is not the case for Portuguese. Raposo et al. (2013) even states that it is not clear if these toponyms [like *Reino Unido*] should be considered canonical or having a descriptive base, that is, if the semantic motivation behind them is already lost, or not, to the speakers[5] (p. 1019).

[5] The place names given in Raposo et al. (2013) as examples to account for this type of occurrence are always complex, but I will assume that what the authors call "other similar cases" also include simplex place names, i.e., toponyms with only one morphological radical.

4. The *Vocabulário Toponímico*

In the last decade, Portuguese, as a pluricentric language, has become a particularly interesting object within toponymic research. The commitment by the CPLP member-states to develop cooperation efforts for an adequate orthographic standardization of the language followed from the formal integration, in 2014, of the Common Orthographic Vocabulary of the Portuguese Language (VOC) in the CPLP heritage. Since then, VOC has been the official resource for Portuguese orthography. Not only is it official, but VOC also "represents a paradigm change, switching from idiosyncratic, closed source, paper-format official resources to standardized, open, free, web-accessible and reusable ones". It is also the first "free-access pluricentric 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 et al., 2012, p. 2), currently with more than 300,000 entries, from several national language varieties (see Ferreira et al., 2017).

With such a large scope, early in its development it became clear that VOC needed to include the language's toponymy. From 2013 forward, a dedicated toponymic database within VOC started to be developed: "Vocabulário Toponímico" (VT) was the first toponymic database ever developed for Portuguese with a pluricentric approach and common transnational standardization criteria,[6] and it still is the only one with such characteristics. As a hierarchical system of toponymic synchronic data (including relational subsets), VT presently accommodates more than 72,000 standardized toponyms, and maintains a structure that allows further sets of toponyms to be incorporated.

The formal properties of a toponym considered in VT are (i) word class; (ii) syllabification; (iii) word stress; and (iv) gender. Properties (i) to (iii) are already labelled in every entry; labeling of (iv) is currently on hold, but it has already been done for higher levels (for the names of geographic referents with higher national administrative relevance, as well as for all country names and country capitals' names). The understanding is that the microstructure of a system like VT, for languages with more than one gender value, must include explicit

[6] Free access at https://voc.cplp.org/index.php?action=toponyms

information regarding mandatory, possible or interdicted usage of an article with a toponym, facilitating gender value readings to all users.

5. Gender-labeled data from VT

In this paper, I share some gender-related regularities using toponymic data from Portugal and, based on them, propose a set of principles for the attribution of gender values to Portuguese (1) Districts/Islands and municipalities' names; (2) Toponyms with the generic *Rio* 'river' or *Ribeira* 'brook'; and (3) Names of countries and national capitals.

5.1. Districts/Islands and municipalities from Portugal

In its second highest hierarchical level – *distrito/ilha* 'district/island' – Portugal has 29 official geographical names, and they are mainly gender-neutral. The only exceptions are (i) [*o*] *Porto* (masculine) and [*a*] *Guarda* (feminine), both in continental Portugal; and (ii) names of all eleven islands from the two Portuguese archipelagos (nine in Azores and two in Madeira), all necessarily feminine, when they include the initial generic *Ilha* 'Island'.[7] Even so, four of these complex toponyms can still include a gender-neutral preposition: (1) *Ilha de São Jorge*, (2) *Ilha de São Miguel*, (3) *Ilha de Santa Maria* and (4) *Ilha de Porto Santo*.[8]

[7] The fact that these geographic referents' names can vary based on the inclusion/omission of the generic *Ilha* is probably justified by "the locality type of islands – being isolated in water" (Gammeltoft, 2018, p. 133). If that is the case, the explicit reference "to the island itself" will not only describe a physical property of the referent but may implicitly underline the political status of the two archipelagos as autonomous regions, with an analogy between water as a physical barrier and political autonomy.

[8] Focusing on the example (*Ilha de*) *Porto Santo*, it is interesting to notice that the feminine complex name accepts the gender-neutral preposition *de* immediately before the generic *Porto*, but the truncated toponym *Porto Santo* (still complex, but without the generic *Ilha*) is always masculine (mandatory type).

In the first three cases, below the generic *Ilha* we find regular hagiotoponymic elements (by rule, gender-neutral in Portuguese, according to VT), so the truncated toponyms that derive from those three examples will be gender-neutral. This is true even with the finding that anthroponymic elements or sacred designations, such as *Miguel* or *Santa*, respectively, do accept a **gender value 1** on their own (e.g., *São* (masculine) / *Santa* (feminine) 'saint'; Miguel (masculine) / Maria (feminine)).[9] In the fourth case, *Ilha de Porto Santo*, the specific element *Santo* 'sacred' is likely to also have some hagiotoponymic interpretation, reflected on a gender-neutral preposition, but must be read as a pseudo-hagionymic adjective, because contrary to what happens in toponyms like *Santa Maria*, it is not the head of a two element NP – also, it can never exist as a simplex toponym.

Portuguese districts' names are **gender value 0**, except for *Porto* and *Guarda*. Fundamentally, the Portuguese islands' names are **gender value 1**, but should be labeled as:

(i) **feminine** when they are complex and headed by *Ilha* 'island' (e.g.: *[a] Ilha do Corvo*).

(ii) **masculine** or **feminine** when they are simplex (with the exception of *Porto Santo*), depending on the gender of a homonymous noun, synchronically available in the general lexicon (e.g., *[as] Flores*).

(iii) exceptionally **gender value 0** when they are complex and headed by a hagionymic designation (instead of *Ilha*) – even if the designation itself is masculine (like *São*/*Santo*) or feminine (like *Santa*) (e.g., *[ø] Santa Maria*).

Even small variance like the one here presented already shows that different levels of analysis for the admeasurement of toponymic gender in Portuguese must be equated, depending on the complexity of each place name (taxonomy related properties, or number and weight of constituents, for instance, can trigger different gender values).

As for the names of Portuguese municipalities (*municípios*),[10] from 308 geographic referents, 252 have a **gender value 0** name (with an **interdicted** type of constraint), including several place names that are homonymous with

[9] These are the prototypical gender values of anthroponyms such as *Miguel* or *Maria*. Even so, it is possible to have a feminine *Miguel* (e.g., *Dulce Miguel*) or a masculine *Maria* (e.g., *Gonçalo Maria*) if a complex name includes them but is not headed by them.

[10] In Portugal, *município* 'municipality' is the administrative division immediately under *distrito* 'district' and above *freguesia* 'parish'.

nouns prototypically associated to hydronyms, such as *Lagoa* 'lagoon' or *Lagos* 'lake(s)'; 45 have an exclusive **gender value 1** name (with a **mandatory** type of constraint); and 7 can either have a **gender value 1** name or a **gender value 0** name (they have a **possible** type of constraint).[11] As an initial proposal, it is only acceptable to claim that Portuguese municipalities' names tend to be **gender value 0**.

5.2. Toponyms with the generic 'river' or 'brook' from Portugal

From the gender-related data available in VT, namely the one associated to toponyms with the generic element *Rio* 'river' (masculine) or *Ribeira* 'brook' (feminine), it is possible to draft solid general rules, based on the recurrent assignment of the same gender values (at least in European Portuguese).

Portuguese place names with the generic element *Rio* 'river' (deriving from a river name) are gender-neutral, and should be labeled as:

(i) **gender value 0** (with the constraint type **interdicted**) (e.g.: *Sou de Rio Torto* / **Sou do Rio Torto* 'I'm from Rio Torto').

Nonetheless, if a complex string like '*River x*' has this generic element omitted, the truncated place name tends[12] to be **gender value 1**, namely when it is synchronically transparent. As an example, the truncated toponym [*o*] *Pinhão* (synchronically homonym to 'pine nut'), a parish whose name derives from [*o*] *Rio Pinhão* (a river), allows a clear distinction between the referent with administrative relevance and the hydronym – two different entities with the same (masculine) specific element. Hypothetically, if the parish's name maintained the generic element *Rio* 'river' (*Rio Pinhão*), it would most certainly be gender-neutral, following the rule presented above. Thus, hydrotoponyms (here I present evidence coming from the ones with the generic element *Rio*) also show that gender types can have an important role when a clear distinction

[11] Interestingly, except for [*ø/o*] *Crato* and [*o/ø*] *Baião*, these are synchronically transparent names: [*as/ø*] *Lajes das Flores* 'flower slabs', [*ø/o*] *Pombal* 'pigeonry', [*Ø/o*] *Gavião* 'hawk', [*a/ø*] *Chamusca* (assuming a derivation from *chamuscar* 'to singe') and [*o/ø*] *Peso da Régua* (even if *peso* derives from the archaic form *penso* 'meal or the place where transportation animals would eat').

[12] See rule above for complex toponyms that are headed by a hagionymic designation (**gender value 0**).

is needed between related geographic referents with homonym names: in this case, between the river's name *Rio Pinhão* (**gender value 1**: masculine) and the hypothetical parish's name *Rio Pinhão* (**gender value 0**).

As for brooks, the Madeira and Azores islands are the only two regions where complex toponyms with the generic *Ribeira* 'brook' (feminine) always have **gender value 1** (e.g., the municipality [*a*] *Ribeira Grande*). Being relatively small, Portuguese islands have no rivers, only brooks. And unlike larger geographic referents, like the aforementioned rivers, brooks are most likely assumed by local community members as proprietary natural features, so each community will refer to the administrative referent using the more familiar/transparent toponym, the one that is homonym to the hydronym (the brook itself): a feminine name.

Fundamentally, Portuguese place names with the generic element *Ribeira* 'brook' (deriving from a brook name) are gender-neutral, and should be labeled as:

(i) **gender value 0** if their geographic referent is in continental Portugal (e.g., [*ø*] *Ribeira Branca* / [*ø*] *Ribeira de Pena*).

(ii) exceptionally **gender value 1** if their geographic referent is in the Madeira or Azores islands (e.g., [*a*] *Ribeira Quente* / [*a*] *Ribeira Grande*).

5.3. Names of countries and national capitals

Country names – and national capitals' names, almost to the same extent – are among the best-known toponyms in any given language. Hence, their increased visibility, placing them in a top position on a familiarity scale, mostly associated to geographic, historical or cultural proximity, is sufficient to avoid major doubts when it comes to a choice between two (or more) gender values. A high degree of familiarity helps to consolidate unique associations between each place name and its gender value (arising from clear constraint specificities), but sometimes makes it harder to decode the specific linguistic feature that triggers them.

Looking at the names of all the countries recognized by the United Nations, in the European Portuguese variety all but 33 of them are preceded by an article.[13] Thus, the **gender value 1** is prevalent in Portuguese country names.

[13] The 33 **gender value 0** country names in European Portuguese are: *Andorra*, *Angola*, *Antígua e Barbuda*, *Barbados*, *Cabo Verde*, *Cuba*, *Granada*, *Israel*, *Madagáscar*, *Malta*, *Marrocos*,

Curiously, from the nine CPLP countries, only three have a mandatory article, [*o*] *Brasil* 'Brazil', [*a*] *Guiné-Bissau* 'Guinea-Bissau' and [*a*] *Guiné Equatorial* 'Equatorial Guinea', and two of them used to have an **interdicted constraint** until recently (*Guiné* as a toponymic element, only started to require a mandatory article in the 20th century (see www.corpusdoportugues.org). Also, among country names there are only eight cases of variation between **gender value 0** and **gender value 1** relating to the same geographic referent. Three of those have already be explained above ([*ø*] *Espanha* / [*a*] *Espanha* 'Spain'; [*ø*] *França* / [*a*] *França* 'France' and [*ø*] *Itália* [*a*] *Itália* 'Italy'), another one is *Chipre* 'Cyprus', that can be named using the most frequent toponym [*o*] *Chipre* or the traditional form [*ø*] *Chipre*; and the last four cases ([*o*] *Belize* / [*ø*] *Belize*; [*o*] *Maláui* / [*ø*] *Maláui* 'Malawi'; [*o*] *Omã* / [*ø*] *Omã* 'Oman' and [*ø*] *Jibuti* / [o] *Jibuti* 'Djibouti') refer to countries that are geo-culturally distant to the majority of present-time Portuguese people, while having synchronically opaque Portuguese exonyms (furthermore, only *Omã*, ending in *-a* (even if it is nasal), could eventually induce a feminine gender reading, but the existing variants are [*ø*] *Omã* (gender-neutral) and [*o*] *Omã* (masculine)).

It is then possible to say that tendentially, Portuguese country names have **gender value 1**. For now, my proposal is to label them as:

(i) **gender value 1** (e.g.: [*o*] *Uruguai*).
(ii) exceptionally **gender value 0** if they are one of the 33 names presented in this paper's footnote 13 (e.g.: [*ø*] *Angola*).
(iii) exceptionally with a **possible constraint** if they are *Spain, Italy* or *France* (e.g.: [*ø*] França / [*a*] França 'France'); or also *Belize*, *Chipre*, *Jibuti*, *Maláui* and *Omã* (*e.g.*: [*o*] *Belize* / [*ø*] *Belize*).

It is also important to note that, in speech, toponyms are not infrequently replaced by designations like *cidade* 'city', *país* 'country' or *vila* 'village', just to name a few possibilities, or associated with them in some way. When this occurs, the gender of the noun that replaces the toponym is naturally present in what appears to be a gender article, and subsequent inflections will grammatically agree with the noun, even if it is omitted. Thus, in these cases, masculine or feminine inflection in an article or in an adjective is not

Mianmar, *Moçambique*, *Montenegro*, *Nauru*, *Nicarágua*, *Níger*, *Palau*, *Porto Rico*, *Portugal*, *Quiribáti*, *Salvador*, *Santa Lúcia*, *São Cristóvão e Neves*, *São Martinho*, *São Tomé e Príncipe*, *São Vicente e Granadinas*, *Singapura*, *Timor-Leste*, *Tonga*, *Trindade e Tobago*, *Tuvalu* and *Vanuatu*.

to be associated with the toponym, even if contiguity exists, but with the noun – as stated in Salgueiro (2016). As so, the example given above referring to one of the few countries that accept two exonyms in European Portuguese, *Omã*, may deserve further discussion, as the masculine variant labeled in VT is most likely accepting an article that actually agrees with the omitted noun *país* 'country' (masculine). Of course, if this is the case, it should not be understood as a rule applied to country names that are masculine, nor to city names that are feminine, but as an exception to be considered when analyzing toponyms that are unfamiliar to a linguistic community. In any case, this proposal calls our attention to an additional layer of complexity in utterance processing, because in Portuguese it is possible to find utterances with, for instance:

(i) a gender-neutral toponym and a feminine (or masculine) adjective that agrees with the omitted noun:
E.g.: *Maputo é linda*.
'Maputo is beautiful.'
*[**A cidade*** [-MASC] *de]* [*ø*] *Maputo é linda**a*** [-MASC].
'[The city of] Maputo is beautiful.'

(ii) a feminine (or masculine) article that agrees with the omitted noun and a gender-neutral toponym:
E.g.: *A Díli dos anos 90 era bem diferente*.
'Dili from the 90s was so different.'
*A [**cidade*** [-MASC] *de]* [*ø*] *Díli dos anos 90 era bem diferente.*
'The [city of] Dili from the 90s was so different.'

As for the names of national capitals (all of which are cities, like the ones given above as examples), the **gender value 0** is almost transversal. Most of these names are synchronically opaque, *Brasília* being one of the exceptions, eventually because a **national capital** status can have higher weight in the attribution of a gender value than a **synchronic transparency** feature. Even if that is the case, [*a*] *Praia* (Cabo Verde's capital city), being homonymous to a very high frequency noun in Portuguese ('beach'), is still one of the eight national capitals that have a **gender value 1**. The remaining ones are four complex toponyms headed by the feminine generic *Cidade* 'City' and specified by the (**gender value 1**) country name: [*a*] *Cidade do Cuaite* 'Kuwait (City)', [*a*] *Cidade do México* 'Mexico City', [*a*] *Cidade do Mónaco* 'Monaco (City)' and [*a*] *Cidade do Panamá* 'Panama City'; and three simplex toponyms that are synchronically opaque: [*o*] *Cairo*, [*o*] *Luxemburgo* 'Luxembourg' and [*o*] *Vaticano*

'Vatican City' – *Luxemburgo* and *Vaticano* being homonyms of the (also **gender value 1**) country name.

From the data on VT, one can conclude that in European Portuguese there is no variation regarding the usage of gender articles with national capitals' names. Even at a CPLP scale, *Luxemburgo* may be the single example with two national variants – the gender-neutral one occurring only in Brazilian Portuguese.

Portuguese names of national capitals are almost always gender-neutral and they are to be labeled as:

(i) **gender value 0** (e.g.: [*Ø*] *Varsóvia* 'Warsaw').
(ii) exceptionally **gender value 1** if they are headed by the generic *Cidade* 'City' (e.g.: [*a*] *Cidade do México* 'Mexico City').
(iii) exceptionally **gender value 1** if they are *Praia*, *Cairo*, *Luxemburgo* or *Vaticano*.

6. Final remarks

Admittedly, the labeling of a toponymic formal property such as grammatical gender is a complex task when dealing with languages that accept more than one gender value. More so when one aims to extend the application of such a linguistic feature to the entire toponymy of a pluricentric language. This complexity partially justifies the fact that such work has never been done for Portuguese in a comprehensive way; another justification comes from the realization that until recently, simply put, no standardized toponymic (digital) tool existed for Portuguese, even less one with a pluricentric approach, and an official status, as recommended in UNGEGN (2006, p. 73). As such, it is assumed that an official pluricentric-based resource, like VT, plays a key role in the admeasurement and shared systematization of toponymic gender in languages with more than one possible gender value. That type of comprehensive development needs, however, to be supported by national authorities as a long-term project.

Some of the observable recurrences that can already be measured were shared in this text, following the initial work that we carried out in VT to account for gender values based on constraint types – presently, only

higher-level toponyms have this feature labeled, that is, the ones with administrative relevance, such as those naming countries and country capitals, or Portuguese municipalities and districts.

This paper aims to unravel standard rules for semiautomatic toponymic gender-labeling in Portuguese, by (i) clarifying the language's possible gender values, (ii) establishing a classification system by types of constraints governing toponymic gender values, (iii) describing a small sample of **gender triggers**, such as *Ilha* 'island', *Rio* 'river' or *Ribeira* 'brook', and (iv) identifying regularities in country names and country capitals' names. It focuses on the European Portuguese variant, as a starting point for multi-varietal exhaustive grammatical gender-labeling. Moreover, the grammatical gender-related information provided in this work can potentially be analogous in other languages.

References

Ferreira, J. P., Correia, M., & Almeida, G. de B. (Eds.). (2017). *Vocabulário Ortográfico Comum da Língua Portuguesa*. Praia: International Portuguese Language Institute (IILP)/ Community of Portuguese Language Countries (CPLP).

Ferreira, J. P., Janssen, M., Almeida, G. de B., Correia, M., & Müller de Oliveira, G. (2012). The Common Orthographic Vocabulary of the Portuguese Language: a set of open lexical resources for a pluricentric language. In N. Calzoari et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1071–1075). Istanbul: European Language Resources Association. https://aclanthology.org/L12-1616/

Gammeltoft, P. (2018). Island names. In C. Hough (Ed.), *The Oxford Handbook of Names and Naming* (pp. 125–134). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199656431.013.49

Raposo, E. P., Nascimento, M. B., Mota, M. A., Segura, L., & Mendes, A. (2013). *Gramática do Português* (Vol. 1). Coimbra: Fundação Calouste Gulbenkian.

Salgueiro, A. (2016). *Topónimos no espaço da CPLP: o Vocabulário Toponímico* [Master's thesis, ISCTE, Instituto Universitário de Lisboa]. Repositório do ISCTE-IUL. http://hdl.handle.net/10071/12495

United Nations Group of Experts on Geographical Names (UNGEGN). (2006). *Manual for the National Standardization of Geographical Names*. New York: United Nations. https://digitallibrary.un.org/record/570893

Villalva, A. (2000). *Estruturas morfológicas: Unidades e hierarquias nas palavras do português*. Braga: Fundação Calouste Gulbenkian.