Journal of Portuguese Linguistics

Mattos, A. P. B., & Oliveira, M. S. D. (2020). Kalunga in the lusophone context: A phylogenetic study. *Journal of Portuguese Linguistics,* 19: 2, pp. 1–24. DOI: https://doi.org/10.5334/jpl.224

**RESEARCH PAPER**

# Kalunga in the lusophone context: A phylogenetic study

Ana Paulla Braga Mattos[1] and Márcia Santos Duarte Oliveira[2]

[1] Aarhus University, DK

[2] Universidade de São Paulo, Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Bolsista de Produtividade de Pesquisa), BR

Corresponding author: Ana Paulla Braga Mattos (anapaullabm@gmail.com)

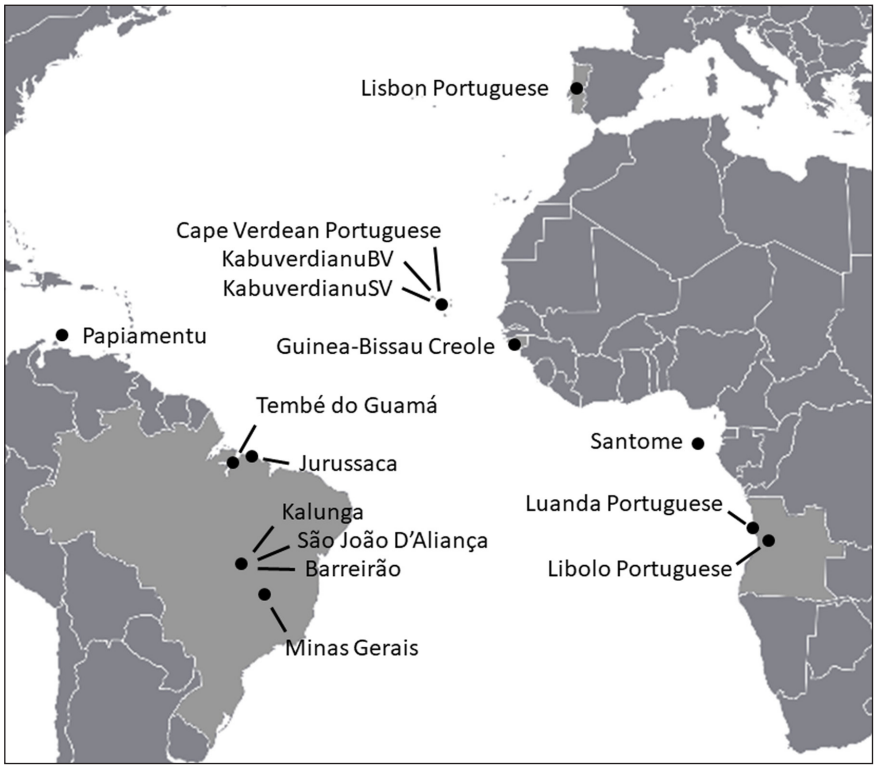
Kalunga is a variety of Afro-Portuguese spoken in a rural community located in the state of Goiás, Brazil. In this study, we compare Kalunga with other varieties of Portuguese spoken in Angola, Brazil, Cape Verde, and Portugal and Portuguese-based creoles from a contact linguistics perspective. We investigate typological similarities, differences, and possible connections between these varieties. The results support previous sociohistorical and linguistic studies that reveal significant differences between Kalunga and standardized varieties of Portuguese, and the typological distinction between creoles, more vernacular varieties, and more standard varieties.

**Keywords:** Kalunga; Portuguese-speaking world; Afro-Portuguese; creole languages; typology; phylogenetics

## 1 Introduction

This paper compares varieties of language in the *lusophone* context from a contact linguistics perspective, with focus on Kalunga, a variety of Portuguese spoken by an Afro-Brazilian community in the state of Goiás (Mattos 2019). By varieties of language in the *lusophone* context, we refer to varieties that have a connection with the Portuguese language, either a genealogically-based relationship or a contact-based one. In this paper we use the term 'variety' broadly, to refer to the set of Portuguese dialects and to the Portuguese-lexified creole languages. We investigate similarities and differences among eleven varieties of Portuguese spoken in Angola, Brazil, Cape Verde, and Portugal, which vary between more standard and more vernacular varieties, and five Portuguese-based creoles. These comprise i) six Brazilian Portuguese vernacular varieties – Kalunga, Barreirão and São João d'Aliança (Goiás), Minas Gerais (Minas Gerais), Jurussaca, and Tembé do Guamá (Pará); ii) Standardized Brazilian Portuguese; iii) two Angolan Portuguese vernacular varieties – varieties spoken in Libolo and Luanda; iv) Cape Verdean Portuguese vernacular; v) Portuguese from Lisbon and the surrounding areas; vi) five Portuguese-based creoles – Kabuverdianu (Sotavento – SV – and Barlavento – BV – varieties), Santome, Guinea-Bissau Creole and Papiamentu. These language varieties and their geographical locations are indicated on **Map 1**.[1]

[1] We do not indicate Standardized Brazilian Portuguese on Map 1, since it is not a variety spoken in a specific geographic region of Brazil. See our definition of Standardized Brazilian Portuguese in Section 2.




**Map 1:** Varieties of language in the Portuguese-speaking world.

Our data consist of feature value assignments based on speech data collected and analysed by specialists in the respective varieties. We use dialectal and typological features that address two sources. First, we have taken features discussed in previous descriptive studies of varieties of Portuguese, focusing on the linguistic phenomena identified in Kalunga (e.g. Mattos 2016; Mattos 2019; Mattos *in press*; Oliveira, Campos & Fernandes 2011; Figuereido & Oliveira 2013). Second, we have used more general typological features, based mainly on the shared features from the *Atlas of Pidgin and Creole Structures APiCS* (Michaelis *et al.* 2013) and the *World Atlas of Language Structures WALS* (Dryer & Haspelmath 2013). In order to conduct this comparative investigation, we used the *Splitstree4* software (Huson & Bryant 2006) to apply computational phylogenetic methods (e.g. Felsenstein 1985; Dunn 2015). A phylogenetic approach allowed us to visualize degrees of similarity and difference among the varieties of Portuguese.

In various disciplines, scholars have adopted the phylogenetic approaches of biology, making it possible to track evolutionary patterns and degrees of similarity among entities. Linguists have used phylogenesis to compare large amounts of data e.g. Dunn *et al.* 2005; McMahon & McMahon 2003; McMahon & McMahon 2006). In the field of language contact, researchers have compared a number of languages when investigating structural–grammatical differences between creoles and non-creoles, indicating a typological profile for creole languages (e.g. Bakker *et al.* 2011), and the classification of varieties related to a certain language group. Examples include Kortmann and Lunkenheimer's large-scale study (2013) of the English-speaking world, Perez *et al.*'s study (2017) of varieties of languages in the Spanish-speaking world, Sippola's work (2017) on the Iberian creoles, and Daval-Markussen's study (2017, 2019) of varieties of languages in the French-speaking world.

Comparative studies of varieties of language in the Portuguese-speaking world have been carried out for only a few varieties, and consider a limited amount of data (e.g. Holm 1992; Lipski 2008; Lucchesi, Baxter & Ribeiro 2009; Petter 2009; Figueiredo & Oliveira 2013; Teixeira & Araujo 2017). None of these used computational phylogenetic methods.

Therefore, this study is the first large-scale computational phylogenetic work to gather empirical data on a number of varieties of Portuguese and creole varieties from Africa, America and Europe, using both typological and dialectal features.

This study investigates the interrelations between varieties in the Portuguese-speaking world, based on empirical data. More specifically, we examine how Kalunga Portuguese, spoken by an Afro-Brazilian community, relates to a range of more standard and vernacular varieties of Portuguese, and to creole varieties with a Portuguese lexifier. It also highlights the relevance of data-driven studies in contact linguistics, and the importance of methodological consistency in comparative studies. All in all, it contributes to the classification and analysis of new language varieties, and to a better understanding of the outcomes of language contact in the Portuguese-speaking world, and more specifically, of the linguistically diverse situation in Brazil.

The paper is organized as follows: in Section 2 we discuss relevant concepts from the field of language contact, and we define the linguistic and sociohistorical background of the varieties studied. Section 3 presents the selection of features and coding used in our data set, and the phylogenetic method used to compare the data. Our analysis and results are discussed in Section 4, and Sections 5 and 6 present the discussion and conclusions.

## 2 Relevant concepts and background information

There is no consensus in the literature on contact linguistics about the definition and classification of terms such as 'vernacular variety', 'creole language', and 'standardized variety'. Without engaging in a deeper theoretical discussion on this matter, we present our definitions of the relevant terminology here, together with concise background information about the varieties compared in this study.

For convenience, we divide the varieties studied into three categories, based on their sociohistorical conditions:

(i) Portuguese-based creoles
(ii) Portuguese vernacular varieties
(iii) Standardized spoken Brazilian Portuguese

The first category, (i) Portuguese-based creoles, refers to a very heterogeneous group of languages called 'creoles'. In this study, we use the following approach to this term, inspired by Winford (2003) and Bartens (2013): creoles form a category of 'new creations of language' that emerge in specific contact situations. They usually have 'one lexifier language, i.e., they derive the bulk of their lexicon from one language, whereas the other levels of the language structure are a result of complex processes' (Bartens 2013: 65). In other words, the lexical and grammatical roots of Portuguese-based creoles usually derive from Portuguese, but the grammatical system is reorganized, and hence quite different (for a range of views on the characteristics and definitions of creole languages, see Bickerton 1981; Thomason & Kaufman 1988; McWhorter 1998, 2005; Mufwene 1996; Parkvall 2000; DeGraff 2005; Bakker 2008; Bakker *et al.* 2011, 2017). In our study, the sample of languages recognized by specialists as Portuguese-based creoles, that is, whose lexicons derive from Portuguese, comprises the Kabuverdianu Sotavento and Barlavento varieties (e.g. Quint 1998; Baptista *et al.* 2007), Santome Creole (e.g. Hagemeijer 2017), Guinea-Bissau Creole (e.g. Intumbo, Inverno & Holm 2012), and Papiamentu (Maurer 2013). There are debates in the literature about the classification of Papiamentu as a Portuguese-based creole or a Spanish-based creole. Nonetheless, recent studies (Jacobs 2012; Freitas 2016) show strong evidence of the Portuguese component of Papiamentu, which justifies the inclusion of this Caribbean language in our comparative study.

The second category, (ii) Portuguese vernacular varieties, includes a diverse set of spoken vernacular forms from various geolinguistic areas. Different processes of acquiring the Portuguese language were at work in the formation of these vernacular varieties. Accordingly, the outcomes of these processes are varieties that diverge from one another to varying degrees when viewed from a linguistic perspective. These varieties are used in everyday communication. Previous studies have include the following terms in the category of Portuguese vernacular varieties: 'vernacular Brazilian Portuguese' (cf. Mello 1996); 'Afro-Brazilian Portuguese' (cf. Lucchesi, Baxter & Ribeiro 2009); 'Afro-Indigenous Portuguese' (cf. Oliveira *et al.* 2015); 'Indigenous Portuguese' (cf. Christino 2015).

In this second category, we have included urban and rural varieties of Brazilian Portuguese, and varieties of Portuguese spoken by communities that show a degree of isolation from mainstream Brazilian society, at a cultural and/or sociohistorical and/or geographical level. In addition to these Brazilian varieties, we include varieties of Portuguese spoken in Angola, Cape Verde, and Portugal. Subsequently, we present the geolinguistic areas of the Portuguese varieties examined in this study, and some aspects of their sociolinguistic.

(1) Brazil
- Afro-descendant communities

**Kalunga** is a remnant *quilombola*[2] community located in the northwest of Goiás state. The area of the community inhabits is located in three municipalities: Cavalcante, Teresina de Goiás, and Monte Alegre. Kalunga is a rural Afro-Brazilian community. It is one of 2958 recognized *quilombola* communities in Brazil, and it is recognized by the Brazilian government as the biggest (in terms of territory) Brazilian *remanescente quilombola* 'remnant maroon community'. Approximately 5000 people live in villages spread over an area of 2632 km$^2$, and the infrastructure conditions vary significantly from village to village. The data examined in this study is from the villages of Vão de Almas and Vão dos Moleques, considered to be the most isolated of the Kalunga areas, and difficult to access. There is no electricity in these two villages (Mattos 2019).

- Afro-indigenous communities

**Jurussaca** is located in the north of Pará state. It is one of the eight areas in the state with *quilombola* communities (NAEA – *Núcleo de Altos Estudos da Amazônia* 'Nucleus of Advanced Studies of Amazonia' 2005). The population of Jurussaca is composed of approximately 500 to 600 people. It is located ca. 25 km from Bragança city and 10 km from the very small town of Tracuateua, and the community is considered to be culturally isolated (Oliveira *et al.*2015: 153; Oliveira, Campos & Fernandes 2011: 131). In some ethnolinguistic studies, scholars have argued that the speech variety used in Jurussaca is Afro-Indigenous (Figueiredo & Oliveira 2013; Campos 2014; Oliveira *et al.* 2015).

**Tembé dos Guamá** is located in the Alto Rio Guamá Indigenous territory, in the northeast of Pará state. The indigenous Tembé language belongs to the Tembé macro-linguistic group. Their self-denomination is *tenetehara* 'people'. According to Machado and Eying (2018: 8), ca. 4168 people live in this territory, although only 2546 consider themselves indigenous. The territory comprises 33 villages: 17 in the Gurupi river region (south)

[2] According to the *Fundação Cultural Palmares*, the definition of *quilombo* is a community formed by "descendants of enslaved Africans that retained subsistence and religious cultural traditions through the centuries" (www.palmares.gov.br, our translation). For a historical account on the term, see e.g. Martiniano (1998, ch 1).

and 16 in the Guamá region (north). The Alto Rio Guamá area is currently undergoing a language shift from the Tembé language (a language of the Tupi-Guarani branch) to an Afro-Indigenous variety of Portuguese (p.c. Mara Jucá). The data analysed in this study is from the following villages in Tembé do Guamá: Sede, São Pedro, Frasqueira, Ita Putyr, Ituaçu e Piná'à.

- Rural and urban areas of Brazil

**Barreirão** refers to the variety of Portuguese spoken in the rural Barreirão village located in the Chapada dos Veadeiros National Park, in the state of Goiás. The village comprises ca. 20 families. The community is located ca. 45 km (dirt road) from São João D'Aliança city. There is electricity and running water in the community, and people have access to television (Mattos 2017, field notes).

**São João D'Aliança** is a small rural town located in the Chapada dos Veadeiros National Park, in the state of Goiás. The city has ca. 13,000 inhabitants. The city is located 160 km from Brasília, the capital of Brazil.

**Minas Gerais** refers mainly to the Portuguese spoken in the metropolitan area of Belo Horizonte, which is the capital of Minas Gerais state, and the sixth largest city in Brazil. The Belo Horizonte metropolitan area has approximately 6 million inhabitants. The Minas Gerais data examined in this study also includes varieties of Portuguese spoken in rural villages in the north and south of the Minas Gerais state (Mello 2012).

(2) Cape Verde

The **Cape Verdean Portuguese** data examined in this study was collected in the capital of the archipelago, on the island of Santiago (Sotavento islands), where there are inhabitants who originate from all the Cape Verdean islands. Kabuverdianu Creole is the dominant language in Cape Verde, and the islanders' first language. The majority of the population also speak Portuguese as L2, with various levels of proficiency. However, recent surveys have shown that Cape Verdeans use Portuguese in many social and communicative contexts (Lopes & Oliveira 2018; Alexandre 2018).

(3) Angola

**Libolo Portuguese** refers to the Portuguese spoken in the Libolo municipality, located south of the Kwanza River, in the Cuanza Sul Province. Libolo has ca. 87,244 inhabitants. The population is mostly from the Ambundu ethnic group, which speaks Kimbundu and Portuguese (both as L1 and L2). Libolo is a rural area, where different varieties of Kimbundu (e.g. Kissama and Kibala) meet. In the classification of Bantu languages, Libolo belongs to the H23 linguistic zone, which is transitioning to being a R10 zone (Lewis *et al.* 2015). According to Figueiredo and Oliveira (2013: 118–119), the Libolo area is also in contact with the Songos, a subgroup of Ovimbundu people, who speak Umbundu

**Luanda Portuguese** is spoken in Luanda, the capital of Angola. According to the 2014 census, there are 2,194,747 inhabitants who live in six urban districts in Luanda. The population comes from different Bantu ethnic groups, especially Ambundu (Kimbundu speakers). Nowadays, the main language spoken in Luanda is Portuguese. According to data from 2016, reported in *Ethnologue* (Simons and Fenning 2018), 40% of the population of Angola speak Portuguese as L1.

(4) Portugal

**Lisbon Portuguese** data relates to the Portuguese spoken in Lisbon and the surrounding urban areas. In Portugal, the variety of Portuguese spoken in the urban areas of Lisbon and Coimbra has privileged status. It is referred to as 'standard Portuguese' by Raposo (2013: 401–428). Despite this status, we decided to include it in this group of vernacular varieties.

The third category – iii) standardized spoken Brazilian Portuguese – refers to a variety of spoken Portuguese that resembles the 'standard form'. It may be identified with the variety used by mainstream TV news programmes (Massini-Cagliari 2004: 5). In this study, we suggest that standardized spoken Brazilian Portuguese refer to the variety represented by the language spoken in Brazilian TV news programmes, when the speakers present real-time news, do not use a teleprompter, and do not have time to rehearse. In these circumstances they produce spontaneous speech, rather than a [more] rehearsed speech (p.c. journalist Marcia Moretti).[3] Two national (not local) TV channels that target a large part of Brazilian society were the basis for data collection.

In short, the three language categories compared in this study represent a range of varieties used in the Portuguese-speaking world, with diverse profiles, in relation to standardization, social and geographical dimensions. In our comparison, we may be able to verify how much these three language categories converge and diverge, and how relevant this division into groups is.

## 3 Methods

### *3.1 Data set*

For this study, we have chosen sixteen Portuguese language varieties, listed linguistic features thereof, and contacted professional linguists considered experts on the selected varieties. The database includes as many varieties of Portuguese and Portuguese-based creoles as possible that meet the following criteria: experts have conducted fieldwork on the variety, they have collected spoken-language data, and they are willing to code our tables of features and share language examples. We also made sure that the experts followed similar guidelines with respect to their methods of data collection and organization. This means, for instance, that the data collected must relate to spontaneous speech, that the data represent situational variation and is representative of the variety in question, and that the quality of their recordings was good enough to permit some acoustic analyses. Thus, we worked with speech data collected and analysed according to similar parameters. Moreover, the individual experts coded typological and dialectal features that were carefully developed on the basis of descriptive studies and previous studies of contact languages. All this means that we have a unique database that presents relative homogeneity in terms of data collection and data organization, consequently, an internally comparable dataset.

As we requested in our questionnaire, each specialist with whom we worked coded two tables of features of the language variety, and was asked to provide illustrative examples. The two tables of features are

[3] Planned speech belongs to a type of written corpora, which is another language diasystem (Berruto 1993). Written and spoken corpora should not be used as one unified type of data in studies that compare similarities and differences between varieties (for a detailed discussion of this topic, e.g. Mello 2016).

1. **Typological Features Table:** This table presents an adapted version of the 48 typological features shared by APICS (Michaelis *et al.* 2013) and WALS (Dryer & Haspelmath 2013). The table comprises two lexical, two phonological, and forty-four morphosyntatic features of the world languages. The original version with the forty-eight features may be found at https://apics-online.info/wals. The adapted version is presented in Appendix 2.
2. **Dialectal Features Table:** This table consists of 57 dialectal features based on our descriptive studies and previous studies of contact varieties of Portuguese (Mattos 2016; Mattos, *in press*; Mattos 2019; Oliveira, Campos & Fernandes 2011; Figueiredo & Oliveira 2013). It comprises 12 phonetic and phonological features and 45 morphological and syntactic features. This selection of features was based on special properties observed mainly in Kalunga Portuguese (Goiás, Brazil) and/or other vernacular varieties of Portuguese, such as Jurussaca Portuguese (Pará, Brazil) and Libolo Portuguese (Angola) (for more on these varieties, see Section 2). The full *Tabela de traços de variedades de português e de línguas crioulas de base portuguesa* 'Table of features of Portuguese varieties and Portuguese-based creoles' (Mattos & Oliveira 2017), referred to in this study as the 'Dialectal Features Table', is presented in Appendix 1.

By selecting these two data sets, we were able to obtain a good balance with regard to the number and the types of features used in our comparison. The WALS/APiCS features may allow us to place the varieties we investigated in a typological framework, whereas the specifically chosen features listed in the dialect questionnaire reveal typological and historical connections among the varieties of Portuguese. We do not include semantic/pragmatic phenomena in the Dialectal Features Table, because they have not yet been studied in Portuguese varieties within Corpus Linguistics as much as morphosyntactic and phonetic phenomena. For instance, with regard to lexical features, our data is based on spontaneous speech collected by the individual researchers, therefore it does not share a list of terms that include accurate semantic criteria that would allow for a comparative analysis of dialects. A set of meanings and their forms would have been much more difficult to extract, especially considering the wide range of topics of the data involved.

All the dialectal features we studied have binary values (presence–absence). They comprise different linguistic phenomena, such as agreement, negation, word order, and simplification of complex onsets. Agreement is a morphosyntactic phenomenon, and we include features concerning gender and number-agreement variation in a noun phrase (NP), person and number-agreement variation in a verbal phrase (VP), and the presence or absence of agreement in both NP and VP. **Table 3** shows examples of four features related to the 'agreement phenomenon' set in the VP, extracted from the Dialectal Features Table. The table presents short descriptions of each phenomenon, followed by an example from a variety of Portuguese (often Kalunga) or Portuguese-based creole or a constructed example. In the text, but not in the original questionnaire, a comparison with 'standard Portuguese' is given, followed by what is found in the example.

Agreement markers in the VP have been studied in several varieties of Portuguese, especially Brazilian varieties (e.g. Monghilhott & Coelho 2002; Monte 2012; Rúbio 2012; Souza 2005; Lucchesi *et al.* 2009: ch. 14). Varieties of Portuguese mainly show variation in the verbal suffixes that mark person in the verb. Therefore, in the Dialectal Features Table, we made it possible to distinguish the suffix marker in the verb for first-person singular (Feature 13), first-person plural (Feature 14) and third-person plural (Feature 15).

**Table 3:** Features related to VP agreement, extracted and translated from Dialectal Features Table.

| Verbal Agreement | Yes | No |
|---|---|---|
| 13. Variation in the first-person singular marking on the verb[4] | | |
| (1) *Eu num **tem** filho* (Kalunga: Mattos 2016)<br>**1SG** no **have.3SG** son<br>'I don't have a son.'<br>*tenho > tem* 'have (1st sg.) > has' | | |
| (2) *Eu já **fez** o café* (Kalunga: Mattos 2016)<br>**1SG** already **make.PST.3SG** the coffee<br>'I already made the coffee'<br>*fiz > fez* 'made (1st sg.) > made (3rd sg.)' | | |
| 14. Variation in the first-person plural marking on the verb | | |
| (3) *Nós num **tem** filho* (Kalunga: Mattos 2016)<br>**1PL** no **have.PRES.3SG** son<br>'We don't have children'<br>*temos > tem* 'have (1st pl.) > have (3rd sg.)' | | |
| (4) *Nós já **fez** o café*<br>**1PL** already **do.PST.3SG** the coffee<br>'We already made the coffee'<br>*fizemos > fez* 'made (1st pl.) > made (3rd sg.)' | | |
| 15. Variation in the third-person plural marking on the verb | | |
| (5) *Eles não **vê** televisão* (constructed example)<br>**3PL** no **see.PRES.3SG** television<br>'They don't watch TV'<br>*veem > vê* 'see (3rd pl.) > see (3rd sg.)' | | |
| (6) *Eles já **fez** o café* (constructed example)<br>**3PL** already **make.PST.3SG** the coffee<br>'They already made the coffee'<br>*fizeram > fez* 'made (3rd pl.) > made (3rd sg.)' | | |
| 16. No person and number marking on the verb | | |
| (7) *pa-m bem* (Kabuverdianu: Quint 2005: 24)[5]<br>because-**1SG** come<br>'*porque eu venho*' 'because I come'<br>(*venho > bem*) | | |
| (8) *e ta kanta sábi* (Kabuverdianu: Quint 2009)<br>**3SG** **TMA** sing know<br>'*Ele sabe cantar*' 'He can sing'<br>(*sabe > sábi*) | | |

We also distinguish variation in use, from the absence of use of any suffix marker in the verb (Feature 16). In agreement with the literature, our data shows that the variation of suffixed markers in the verb is much less common in the case of first person singular than in third person singular and first person plural. Only the languages classified as creole languages appeared to be coded with 'yes' for the absence of suffixes in the verb to mark number and person (for verbal categories in creole languages, see Bakker, Post & der Voort 1994; Holm 1988: 148–171; Siegel 2007; Winford 2018).

'Negation' is also a morphosyntactic phenomenon that is present in the Dialectal Features Table. **Table 4** shows examples of features related to this subcategory.

In this set of features (20 to 23), we examine some constructions generally known as 'multiple negations', since previous experience indicated that the variety of negation

[4] Since the more and less standard varieties of BP show no differences between the second person singular and third person singular, and since the second person plural has the same form as the third person plural in many varieties of BP, we did not analyse the 2SG, 3SG and 2PL cases.

[5] The glossing and free translation into Portuguese in (7) and (8) are ours.

**Table 4:** Examples of features related to Negation, extracted and translated from the Dialectal Features Table.

| **Negation** | **Yes** | **No** |
|---|---|---|
| 20. Frequent use of multiple negation markers in the same sentence.[6]<br>(9) ***Não*** *vou lá* ***não.***<br>NEG go there NEG<br>'I don't go there'<br>(10) *Não veio ninguém.*<br>NEG came nobody<br>'Nobody came' | | |
| 21. Negative indefinite pronoun (*ninguém* 'nobody') in the subject position, followed by simple or double negation.<br>(11) ***Ninguém não*** *conseguia trabalho.*<br>Nobody NEG got job<br>'Nobody could get a job'<br>(12) *Nesse lugar aqui* ***ninguém num*** *tem futuro* ***não.***<br>In this place here nobody NEG have future NEG<br>'In this place here, nobody has a future'<br>(Kalunga: Mattos 2016) | | |
| 22. Combined use of ***nem*** 'nor' with ***não/num*** 'no' to mark negation.<br>(13) *Tinha ano que ele* ***nem num*** *vinha* (Kalunga: Mattos 2016)<br>had year that he nor NEG came<br>'There were years when he did not even appear'. | | |
| 23. Use of ***nunca*** 'never' disassociated from the quantifier value of the *nunca* 'never' (*jamais* 'not ever', *em tempo algum* 'at no time'). See example (15). Context to sentence with ***nunca* 'never'** in (14) (Kalunga: Mattos 2016):<br>(14) *A chuva destruiu a roça?*<br>the rain destroyed the harvest<br>'Did the rain destroy the harvest?'<br>(15) ***Num*** *distruo porque roça nós* ***nunca*** *tinha prantado né?*<br>NEG destroyed because harvest we never had planted PART<br>'It [the rain] didn't destroy [the harvest] because we hadn't planted the crops, right?' | | |

markers among the varieties of Portuguese is significant. Feature 20 relates to the frequent use of multiple negation markers in the same sentence, and it represents a more general and common case of multiple negation in varieties of Portuguese. Feature 21, the negative indefinite pronoun as a subject followed by a negative particle, is not as widely distributed as Feature 20, yet our data shows the presence of Feature 21 in many varieties of vernacular Portuguese and a few creole languages. This was rather unexpected, as in the literature, this feature was not reported for many varieties of BP. Feature 22 seems to be uncommon among the selected varieties, and its presence is reported only for Kalunga Portuguese, Minas Gerais Portuguese, Cape Verdean Portuguese, and Kabuverdiano. Feature 23 had not been reported in the literature for Portuguese varieties, but our data shows that this feature is present in seven language varieties, namely Kalunga, Barreirão, Minas Gerais, Tembé dos Guamá, Libolo, Cape Verdean Portuguese, and Kabuverdianu.

The number of dialectal features is not equally distributed among the various grammatical categories in our Dialectal Features Table (Appendix 1). There are more features in some grammatical categories than in others (e.g. four features relating to negation and agreement in VP, and two features relating to topic constructions). This difference

[6] By 'same sentence', we mean that the particle 'no' is part of the verbal negation, so we do not consider the particle 'no' when it is part of a different utterance.

occurs because we selected features that were expected to represent variation among the Portuguese language varieties, according to previous dialectological and descriptive studies. Therefore, in order to avoid bias in our results, we wanted to determine whether these features are really significant when mapping differences and similarities among the varieties. For this purpose, during the data analysis we experimented with various combinations of features. For instance, we randomly removed some features, we removed some selected categories of features, and we removed features that (apparently) overlap in the same grammatical category, to test whether we would obtain a different result or not. The results remained essentially the same.

The more general typological data used in this study relies on features selected for WALS and APICS. Based on their set of shared features, we adjusted these features by translating their descriptions into Portuguese. We provided relevant examples, in order to facilitate the contributing researchers' work. The features were coded using multi-state values. Because the dialectal features are binary, and the typological features are multi-valued, we do not combine dialectal and typological feature sets in the same analysis.

### *3.2 Phylogenetic approach*

When applying phylogeny, computational methods allow the quantitative comparison of a significant amount of data, in order to investigate the relationships between entities. Different algorithms may be chosen to yield the desired type of comparison. In linguistics, the number of studies using computational phylogeny has recently increased (e.g. Gray & Atkinson 2003; Nakhleh *et al.* 2005; Gray *et al.* 2009; Prokić and Nerbonne 2008; Sicoli and Holton 2014; Bakker *et al.* 2011; 2017). Computational phylogenetic techniques have been used mainly for evolutionary studies of language families, and dialect and language classification in historical linguistics.

In this study, we use SplitsTree4's Neighbour-Joining and Neighbour-Net algorithms (Huson & Bryant 2006) to calculate the number of similarities and differences between varieties. To do so, the algorithms use a distance-based method which uses a distance metric (Hamming distance[7]) to visually present the difference between two instances in the data set. In general terms, the distance between two pairs of language varieties in the graphic representation corresponds to the number of features these varieties do not share.

To build the graphic representation, first we have to encode the Dialectal and Typological features. This is done by answering 'yes' for the presence of a feature in a variety, and 'no' for the absence of a feature from a variety (for the Dialectal Features Table), or by choosing the value that best corresponds to the feature in the variety analysed (for the Typological Features Table). After encoding the features from both Dialectal and Typological Features Tables, the codes were converted into arbitrary numerical values. Missing data (where the participants did not know or did not have the information), were coded with a question mark ('?'). These scores are not included in the calculation. These encodings lead to matrices containing the language varieties and the scores for the features.

The matrices with numerical values may be converted by the SplitsTree4 program into relative numerical distances between the compared entities. Then, the Neighbour-Joining algorithm converts these distances into a graphic in the form of a tree. The more similar the entities are, the closer together they are in a tree. The numerical distances help to visualize the closeness and remoteness calculated between the compared varieties. Besides the visual distances between the entities, the Neighbour-Net algorithm also shows different possible ways where a split may be made. The web-like network represents the conflicting signals in the form of multiple places where splits may be made.

[7] In the Splits Manual (p. 10) (www.splitstree.org), it is referred to as 'Uncorrected P'.

An advantage of both algorithms in Splits Tree is that they provide an easy-reading general sense of how the different Portuguese varieties are related among themselves and in relation to Kalunga. Other cluster analysis algorithms, such as Principle Component Analysis, do not provide a well-displayed format of the groupings. A drawback of Splits Tree is that features responsible for the groupings are not informed. Therefore, in order to overcome this, we organized a heatmap of features based on the results provided by Splits Tree, to identify which features are more probably responsible to define the groupings. These features are discussed in Section 4.1 below.

Moreover, the foregoing algorithms have proven suitable for comparing language varieties when they descend from a common ancestor, as is the case with the varieties of Portuguese compared in this study (for references regarding the adequacy of distance methods, e.g. Dunn, 2015: 6; for references regarding the performance of these algorithms to compare dialects, e.g. Prokić, 2010: ch.3). As the process of creole language formation involves various languages, we cannot claim that they descend from one common ancestor. However, since the aim of this study is to verify the typological relationship between Kalunga and other varieties, and since all the creoles (the so-called Portuguese-lexifier creoles) included in this comparative study are in some way related to Portuguese, we consider this distance-based method suited to the type of comparison undertaken.

## 4 Language varieties compared

In this section, we report the results of the phylogenetic analyses. In 4.1 and 4.2, we present the results of the analysis of the Dialectal and Typological features, respectively. A detailed description of all features mentioned in this section, and the relevant examples, may be found in Appendix 1 (Dialectal features) and Appendix 2 (Typological features).

### *4.1 Dialectal features*

Our first test relates to dialectal similarities among the 16 varieties of Portuguese, including Portuguese creoles. In **Figure 1**, all 57 dialectal features are included for all 16 language varieties analysed in this study. The Neighbour-Net algorithm is used in

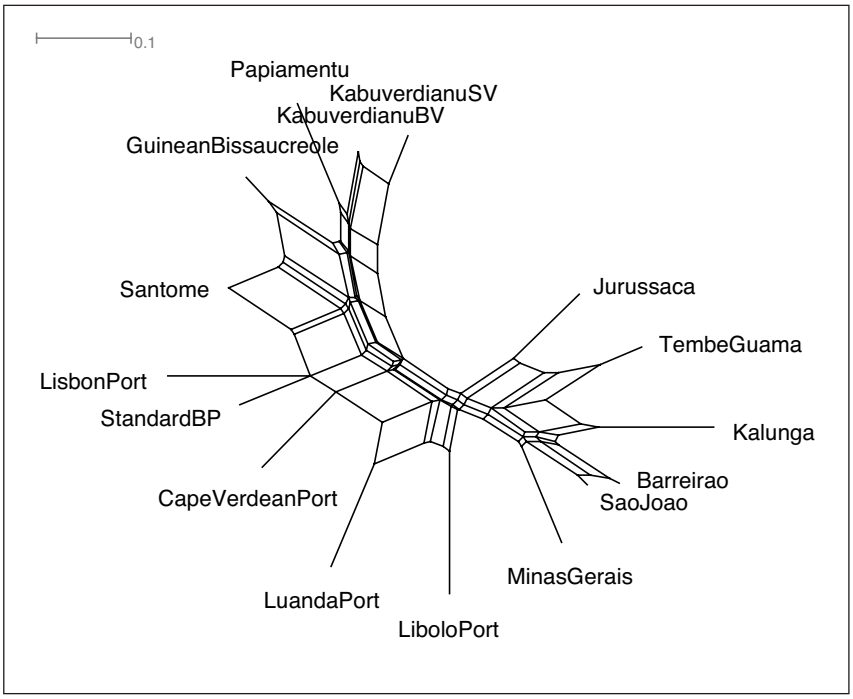



**Figure 1:** Split network of a Neighbour-Net of the 57 dialectal features of the 16 language varieties.

this analysis. The branches producing web-like boxes show connections and conflicting signals between the entities, where the splits occur. There are no clear-cut separations of groups in **Figure 1**, but there are clear patterns. The Brazilian vernacular varieties (Minas Gerais varieties, São João, Barreirão, Kalunga, Tembé do Guamá, and Jurussaca) appear to cluster at the bottom of the network.

At the top of **Figure 1**, the five creoles appear close together, and the Kabuverdianu creole varieties of Sotavento and Barlavento occur especially close together. Papiamentu is also close to the Kabuverdianu and Guinea-Bissau Creole, which is consistent with studies that show the historical relations between Papiamentu and the Upper Guinea creoles, especially between Papiamentu and Kabuverdianu (for the relationship between Papiamentu and Kabuverdianu, e.g. Jacobs 2012; Freitas 2016). Santome, a Gulf of Guinea creole, appears somewhat separate from the Upper Guinea creoles (Guinea-Bissau Creole and Kabuverdianu) and Papiamentu. The normalized distances representing proximity between Santome and the other creoles are: Guinea-Bissau Creole = 0.07, Kabuverdianu (Sotavento) = 0.26, Papiamentu = 0.41. However, the connection between Santome and the other creoles shown in the network is less certain, because a significant amount of data for this language is missing (over 50% of the features), in comparison with the other language varieties. We also ran a similar analysis (57 dialectal features) in which we removed Santome from the language set, to test the effect of this language in the outcome. The major difference was that the group of creole languages clustered more closely, and was more distant from the other groups. The result was otherwise similar to the tree in **Figure 2**.

Spoken Lisbon Portuguese appears relatively close to spoken Standardized Brazilian Portuguese (SBP) and Cape Verdean Portuguese (CVP), followed by Luanda Portuguese and then Libolo Portuguese. The position of SBP in the network shows that SBP tends to

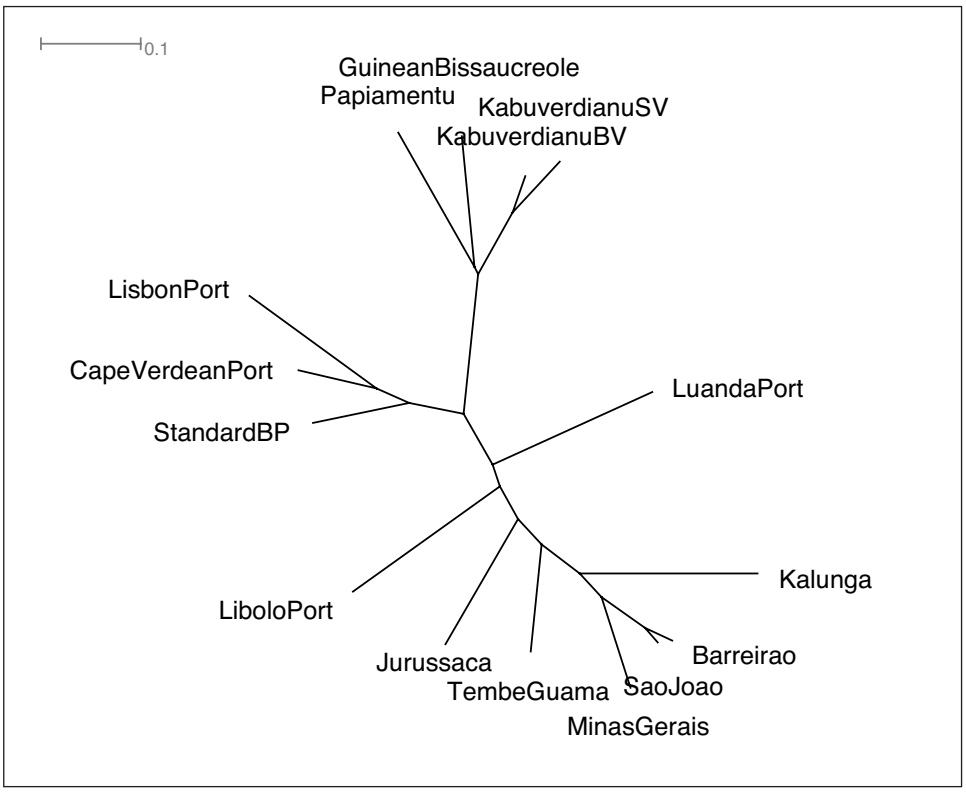



**Figure 2:** Split tree of a Neighbour-Joining Tree of 45 morphosyntatic dialectal features of 15 language varieties.

be closer to Lisbon Portuguese than to Brazilian vernacular Portuguese varieties, even though only spoken language data is considered in this study. The proximity of CVP to SBP and Lisbon Portuguese corresponds to the findings of recent studies that note the use of Portuguese mainly as an L2 variety in Cape Verde (a non-nativized variety), and to the use of European Portuguese as a referential variety for CVP speakers (Alexandre 2018). However, there is an increase in the use of Portuguese in a wider range of communicative contexts in Cape Verde, where only Kabuverdianu was used previously (Lopes & Oliveira 2018). In contrast with the situation in Cape Verde, Portuguese is spoken as L1 variety in Angola (for studies on varieties of Portuguese spoken in Angola, e.g. Inverno 2011; Figueiredo & Oliveira 2013). The Luanda Portuguese variety appears closer to the more standard varieties (Lisbon, SBP and CVP) than Libolo Portuguese does, which is expected, due to their sociolinguistic settings: Libolo Portuguese is a rural variety, and Luanda Portuguese is spoken in the capital of Angola.

At the bottom of the network in **Figure 1**, the six Brazilian vernacular varieties (BVP) are grouped together. At the right of the network, the Tembé do Guamá and Jurussaca varieties, from the north of Brazil, are close together. Kalunga has a long, independent branch, and appears between these two northern varieties and São João and Barreirão, which appear very close together. In terms of the sociolinguistic similarities between Kalunga and these other varieties, Kalunga, Tembé do Guamá, and Jurussaca are varieties spoken by African and Afro-Indigenous communities, and Kalunga, São João, and Barreirão are varieties spoken in neighbouring geographic areas, namely, the northeast of the province of Goiás. Minas Gerais varieties are also found in the BVP group, and, besides geographic proximity, it shares several sociohistorical aspects with the varieties in Goiás.

In **Figure 2**, the graphic representation compares only morphosyntactic features from the Dialectal Features Table (we removed the phonetic/phonological features) and 15 language varieties, since Santome was removed in this case, owing to insufficient data. Here we have used a neighbour-joining tree, which measures only linguistic distances, without considering possible conflicting signals. The shape of, and the positions of the varieties in **Figure 2** are similar to those of **Figure 1**. There is a more clearly visible separation of three language clusters in **Figure 2**: creoles, more standard varieties, and more vernacular varieties. The creoles are grouped together at the top of the tree. At the left of the tree, there are the more standard varieties (SBP, Lisbon Portuguese, and CVP). Then, the vernacular varieties appear with long independent branches, except for Barreirão and São João, which are close together. In this tree, the Angolan varieties cluster with the BVP varieties. Among the vernacular varieties, the Angolan varieties are the nearest to the standard-variety group. Some relevant features that contribute to this separation are Feature 21, 'Negative indefinite pronoun (*ninguém* 'nobody') in the subject position, followed by simple or double negation', which is not present in Lisbon Portuguese, SBP, and the Angolan Portuguese varieties (Luanda and Libolo), but encountered elsewhere, and Feature 34, 'use of preposition *a* 'to' in dative construction', which is present in Lisbon Portuguese and Luanda Portuguese only.

Analysing the data set in detail, we verify that there are not many features that clearly separate these groups of languages. The presence and absence of features are generally distributed among this set of language varieties. Only Feature 20, 'Frequent use of multiple negation markers in the same sentence' is shared by all the varieties, that is, it is irrelevant for the graphed results. Only six features (16, 37, 3, 29, 46, and 53) are present in, or absent from specific varieties. These features are discussed below.

Features 16 – 'no person and number marking on the verb morphology' – and 37 – 'no grammatical gender in adnominals, as, determiners, pronouns and adjectives' – are present

in the creole varieties only. 'No grammatical gender' and 'no inflectional morphology' are referred to in the literature as prototypical features of creole languages, when compared to their lexifiers (e.g. McWhorter 2005; Daval-Markussen and Bakker 2017). Some of the vernacular varieties, including Kalunga, present variation in marking grammatical gender, person, and number in the verb, as the examples in Features 13 and 36 show.

Features 3 – 'the use of affricative /ʧ/ in context of /t/ plus approximant /r/' as in */trabalha/* > [ʧa'baɱ] – and 29 – 'no preposition in genitive construction' as in *a luz (de) nós* 'lit. the light (of) we/our light' – are present in Kalunga only. Palatalization is a common phenomenon among the studied languages of the world, including BP. However, when the set of language varieties selected for this study is considered, palatalization with affricates in the phonological context described in Feature 3 appears to be a more specific phenomenon of Kalunga. This suggests that the phenomenon is an outcome of an internal language-development process. Concerning Feature 29, 'no preposition in genitive construction' in Kalunga, it is important to consider two things. Firstly, it is not common among the language varieties analysed to mark possession with the genitive form *de*, 'of' + 1st person pronoun. Only Kalunga, Libolo, Papiamentu, and Guinea-Bissau Creole present Feature 24, 'possession with the genitive form *de* 'of' + tonic personal pronoun, as in *a vaca* ***de nós***, 'lit. the cow of us/our cow' (see appendix 1 for a more detailed presentation of this feature). Secondly, in Kalunga there is variation with respect to the use and the non-use of a preposition in genitive constructions such as *a vaca (de) nós*. Since a variationist study was not conducted, it is difficult to track the process that might be involved in the phenomenon described for Feature 29.

The two features not present in Lisbon Portuguese only are Features 46 – *que* 'that' as the default relative pronoun', as in *a menina que o pai (dela) é angolano mora no Brasil* lit. 'the girl that (her) father is Angolan lives in Brazil/the girl whose father is Angolan lives in Brazil' – and Feature 53 – 'element in topic position – to the left of the sentence – without having been moved from any position of the sentence', as in ***Você,*** *você gosta disso?* 'you, you like it?'

With respect to Feature 46, *que* 'that' as the default relative pronoun', it is common among varieties with intense contact history to have one 'default' element to indicate grammatical functions, instead of a set of elements. A default locative preposition and a default relative pronoun illustrate the case. Lisbon Portuguese is the only variety in our data to include the use of a set of relative pronouns. Even SBP appears to have reduced the use of this set of relative pronouns. For instance, the use of *que* 'that', instead of *cujo* 'whose', as described in our data.[8]

The phenomenon described in Feature 53, 'element in topic position – to the left of the sentence – without having been moved from any position of the sentence', also seems to be an outcome of language contact. It is absent only in Lisbon Portuguese,[9] but it seems to be a common phenomenon in BP varieties and in creole languages in general. The sentences (i) ***Você, cê*** *gosta de cantar* 'you, you like to sing', and (ii) ***abo bu*** *gosta di kánta* 'you,

[8] For Feature 46, we refer to the use of relative pronouns in dependent relative constructions (DRC), and not to the relative pronouns in free relatives (FR). DRC are constructions that are semantically and syntactically related to an antecedent, as in *eu concordo com* ***as coisas que*** *você sugere*, 'I agree with **the things that** you suggest'. FR do not relate to an antecedent; the reference is incorporated in the relative pronoun, as in *eu concordo com* ***o que*** *você sugere*, 'I agree with **what** you suggest'.

[9] Feature 53 does not deal with topicalization (which involves the movement of an element to the left 'periphery of the sentence' without any resumptive element), nor with left dislocation (topicalization of an element that is resumed by another element). Both topicalization and left dislocation are widely mentioned in the literature on varieties of European Portuguese (cf. Mateus et al. 2003: 492–501; Raposo et al. 2013: 403–422). Instead, the phenomenon described in Feature 53 deals with an element placed to the left of the sentence that is semantically co-referent with the subject of the sentence. This subject is also realized as a weak pronoun (see, e.g. Kato 1999). In our reformulated Dialect Features Table (in preparation), Feature 53 is no longer treated as a Topic phenomenon.

you like to sing', in BP and in Kabuverdiano, respectively, illustrate this phenomenon. In these sentences, *você* and *abo* have discoursive properties, and they are semantically co-referent with the weak pronouns *cê* and *bu*, respectively. In Kabuverdiano, a clitic may be co-referent with its antecedent pronominal element to the left of sentence as in the sentence, *mi* ***m****-kánta* 'I, I sing' (Quint 2003: 218, our translation).

Apart from the six features (16 and 37 for creoles, 3 and 29 for Kalunga, and 46 and 53 for Lisbon Portuguese) just described, there are other features that may be predictors of language groups, due to differences in the experts' interpretations of the definition of each feature, and to the apparently less precise formulation of the features. For instance, Features 17, 18, and 19, which deal with specific properties of tense, mood, and aspect related to verbal morphology, are not present in either creole or standardized varieties of Portuguese. These features are present in vernacular varieties. Since creole languages usually have preverbal TMA particles, rather than inflectional verb morphology, Features 17, 18, and 19 are marked 'no' for creole varieties, meaning that they are not found in the creoles. The same features are coded as 'no' for standardized varieties, but for a different reason: the standardized varieties do have a more heterogeneous inflectional morphology than the vernacular varieties, and here, these features go in the opposite direction, that is, towards a more homogenous verb morphology. Therefore, the proximity of creole varieties and standardized varieties in the trees is partly due to the features that they do not share with vernacular varieties, rather than the features that these two groups share. In the same vein, the 12 phonetic and phonological features (Features 1 to 12) are present in vernacular varieties, but not in creoles and standardized varieties.

The feature groupings of the African and Afro-Indigenous varieties of Portuguese (Kalunga, Jurussaca, and Tembé do Guamá) suggest that these varieties have specific traits in common. However, they do not appear as a homogeneous group, probably because different linguistic processes were at work in the formation of these varieties, and/or they have developed in different ways. For instance, when comparing Kalunga, Tembé do Guamá, and SBP, we notice that both Kalunga and Tembé do Guamá show significant divergences from standardized varieties. However, Kalunga and Tembé do Guamá do not share the same differences with respect to SBP. When Kalunga and Tembé do Guamá do not share a set of features, we may consider two possibilities: i) Kalunga shares this set of features with SBP and the varieties of Goiás (São João and Barreirão), that is, Tembé is the particular case, or ii) Tembé do Guamá shares this set of features with SBP and Jurussaca (geographically close to Tembé), that is, Kalunga is the particular case.

**Figure 3** visualizes the distances among the Brazilian Portuguese varieties, based on a comparison of 45 morphosyntatic features. In this tree, Jurussaca and Tembé do Guamá cluster at the top, and Barreirão and São João, at the left. Minas Gerais is closer to Barreirão and São João than the other varieties. Kalunga and SBP have very long, independent branches. If we take the distances presented in **Figure 3**, and, based on those, place all the language varieties along a bi-dimensional BVP continuum, Kalunga would be the variety that is furthest removed from SBP. However, it is not clear where the Minas Gerais, Tembé, Jurussaca, Barreirão, and São João varieties would be placed along this bi-dimensional continuum.

When comparing Kalunga with the other language varieties in Goiás (Barreirão and São João), we find 13 features that Kalunga does not share with the other two varieties. Of these 13 features, Kalunga shares only one feature with SBP, that is, Kalunga has 12 features from different grammatical categories that are neither shared with the varieties of Goiás nor with SBP. This shows that the rural variety of Barreirão is closer to SBP than Kalunga is, and, as expected, Kalunga shares more features with Barreirão than São João D'Aliança. This proves that there are various linguistic degrees of divergences between

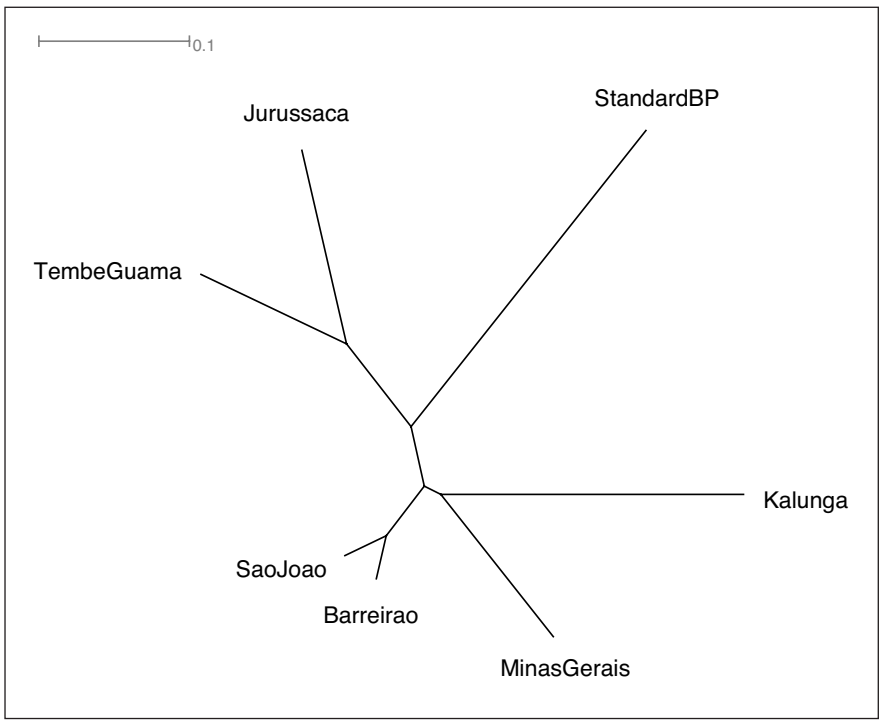



**Figure 3:** Split tree of a Neighbour-Joining Tree based on 45 morphosyntatic features of seven Brazilian Portuguese varieties.

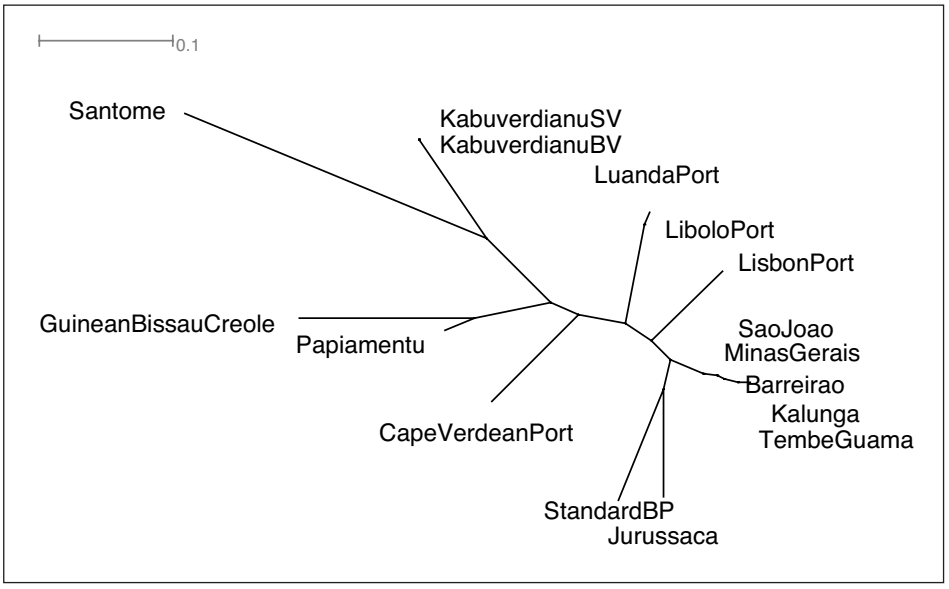



**Figure 4:** Split tree of a Neighbour-Joining Tree based on 48 Typological features, adapted from APICS/WALS of 16 language varieties.

Kalunga (Afro-Portuguese), Barreirão (rural Portuguese), and São João (Portuguese from a rural town). Relative to SBP, Kalunga is the most distant, followed by Barreirão, and then São João.

### *4.2 Typological features*

The last tree (**Figure 4**) represents the relationships among the 16 language varieties studied here, when the Typological Feature Table is used for comparison. Since this represents more general typological features adapted from the 48 shared features from WALS/APICS, the differences among the groups are less prominent. For instance, there

are no significant differences between the vernacular Brazilian varieties of Minas Gerais, São João, Barreirão, Tembé do Guamá, and Kalunga, which appear together at the right side of the network. The same is true of the Angolan varieties, Libolo and Luanda, which appear on the same branch of the tree. The creoles all appear on the left side of the network, but they do not form a clear single close cluster. In particular, Guinea-Bissau Creole and Santome stand apart from Kabuverdianu and Papiamentu. A factor that might explain the longer branch for Santome is that the features for this language were the only ones taken from the WALS/APICS atlases. We did not have an expert scoring the typological features for Santome, unlike the cases of all the other varieties. This different methodological approach may have affected the result.

It is not surprising that these general (typological) features present less marked differences, as all are varieties of Portuguese and Portuguese-lexifier creoles, and the selected features were designed to cover the typological variety of the languages of the world, whereas the dialectal features were selected to cover variation in varieties of Portuguese.

## 5 Considerations of similarities and differences in the lusophone context

Comparative studies based on linguistic corpora that include a significant amount of data that allow us to verify the historical and typological relationships among entities are highly relevant to language contact studies. The use of computational models such as those used for this study helps us to visualize the relationships among language varieties, as they map similarities and differences among these entities. However, the results presented in the previous section, although interesting and significant, should not be analysed without considering the sociolinguistic contexts in which the languages are spoken (c.f. Perez et al 2017). Moreover, the methods used in comparative studies must be considered when the results are discussed. Methodological concerns include the features selected for the study, methodological coherence in terms of data collection, the way researchers understand and score the features, and the way a language variety is defined, in order to compare it to another variety. For instance, we notice that the typological features from APICS/WALS, in the format they are presented in this study, as we see in the tree in **Figure 4**, cannot be utilized to map detailed similarities and differences among all 16 language varieties selected, as many of them end up as being close to identical in structure, as is the case of the five vernacular varieties at the right side of **Figure 4**.

Our language feature selection is certainly one of the greatest factors to have an effect on the results.[10] For instance, in our Dialectal Feature Table, we consider some morphosyntactic and phonological aspects of the Portuguese varieties in general, and some features for which Kalunga, in particular, stands apart from the other varieties. Considering our specific data set, previous linguistic and sociolinguistic studies of the varieties of Portuguese analysed, and with Kalunga at the centre of our comparison, our results, based on the Dialectal Feature Table, indicate that:

- there are three major clusters: creoles, standardized varieties, and vernacular varieties;
- Kalunga shares more features with the BVP varieties than with creoles and standardized varieties; generally, Kalunga shares more features with creoles than standardized varieties;
- Kalunga shares specific features with the language varieties spoken in the same geographical area of Goiás state, namely, Barreirão and São João; Kalunga shares more features with Barreirão (rural community) than with São João (country-side town);

[10] Refer to Section 3.1 to see some actions, in order to shorten the effects of some specific features of the results.

- Kalunga is also close to Tembé do Guamá, which is an Afro-Indigenous Portuguese variety;
- Kalunga differs from Lisbon Portuguese and SBP more than the other vernacular varieties do;
- SBP is closer to Lisbon Portuguese than it is to any Brazilian vernacular variety, which may suggest that the spoken data from Lisbon is indeed close to standardized Portuguese; also, this suggests that (spoken) BP still follows the (spoken) European Portuguese language as standard;
- Cape Verdean Portuguese is close to SBP and Lisbon Portuguese;
- The BVP varieties cluster;
- The Minas Gerais variety is close to the Goiás varieties;
- There is no special relationship between Kalunga and the analysed varieties of Portuguese spoken in Angola and Cape Verde, in Africa.

One aspect that may have had an effect on the analyses is the way in which the researchers interpreted the linguistic features, and how they analysed the linguistic phenomena. Feature values that are essentially similar may be analysed and classified differently by researchers. Therefore, similar features may have been scored differently according to the researcher's understanding. Moreover, the understanding of what is a categorical feature in the variety may vary from researcher to researcher. The potential impact of these observations of the outcomes has been reduced by our asking the contributors to provide examples or direct access to their data. After analysing the answers, it is sometimes possible to verify which features may be controversial, or may have led to different understandings among the researchers. In a few cases, we may observe a discrepancy between an example and the proposed score. In this study, we addressed that by double-checking such answers with the researcher.

As mentioned above, in this study we have considered only spoken language data, and we have attempted to use data that have been collected in compatible ways. However, with respect to methodology and for future works, we recommend that varieties of Portuguese, or areas where each variety of Portuguese is spoken, be more clearly delimited, as these play a role in how a language variety or a sociolinguistic area is defined. A more precise definition of varieties or areas may allow a more accurate comparative analysis, as we compare micro-areas such as those of Afro-Indigenous communities (Jurussaca and Tembé do Guamá), an Afro-Brazilian community (Kalunga), a rural community (Barreirão) with macro-areas, such as the Minas Gerais state in Brazil, and Lisbon and surrounding areas, in Portugal.

Extending the database by adding dialectal features and language varieties would allow a more fine-grained analysis of the relationship between linguistic varieties in the lusophone world, therefore we would be able to reach more informed conclusions in relation to the different categories of languages. For instance, it would be interesting to add to the comparison varieties of Portuguese spoken in Asia, more varieties of Portuguese spoken in Africa and Portugal, more Afro-communities in Brazil, such as in Bahia (where it is argued that there was a strong contact influence from African languages) and in the Rio Grande do Sul (where, apparently, the Afro-Brazilian communities were not as isolated from the rest of society as the other Afro-communities in Brazil). It would also be interesting to investigate varieties of Portuguese spoken close to the country's borders, as a significant amount of language contact exists there.

## 6 Concluding remarks

In this study, we analyse the relationship between varieties of spoken Portuguese and Portuguese-based creoles, based on morphosyntactic and phonetic/phonological features. The results show differences between the varieties of Portuguese and the creoles, and

reveal clusters of three major groups: creoles, more standard varieties, and more vernacular varieties. In general, these groupings are consistent with the group-language distinction made in Section 2 (i.e., Portuguese-based creoles, Portuguese vernacular varieties, Standardized spoken Brazilian Portuguese). The only exceptions are Lisbon Portuguese and Cape Verdean Portuguese, both of which appear in the results with SBP, suggesting that they could be categorized as standard varieties, instead of vernacular varieties. The results corroborate Raposo's (2013) categorization of spoken Lisbon Portuguese as a standard variety.

Kalunga appears close to the vernacular varieties of Portuguese, especially BVP varieties, and substantially distant from standardized varieties. In general, Kalunga is typologically closer to creoles than to standardized varieties, possibly owing to features that are outcomes of the specific contact situation from which Kalunga emerged (as discussed in Section 4.1). Also, Kalunga has specific features (also discussed in Section 4.1) that are not shared by other varieties, possibly features that are independent developments owing to the degree of Kalunga isolation.

This study finds interesting similarities and differences among varieties of Portuguese. This paper presents the first quantitative phylogenetic analysis of the relationships among a wide range of varieties of Portuguese and Portuguese Creoles, based on spoken data. This analysis helps us to understand the relationship between Kalunga and the Portuguese varieties and Portuguese-based creoles.

## Additional Files

The additional files for this article can be found as follows:

- **Appendix 1:** Dialectal Features Table. DOI: https://doi.org/10.5334/jpl.224.s1
- **Appendix 2:** Typological Features Table. DOI: https://doi.org/10.5334/jpl.224.s2

## Acknowledgements

We would like to thank all the participating researchers for their immeasurable contributions to this paper. Without them, this study would not have been possible: Carlos Figueiredo (UM), Ednalvo Campos (UEPA), Eduardo Ferreira dos Santos (UNILAB), Fernanda Ziober (VU), Francisco João Lopes (Instituto Camões), Gabriel Antunes de Araújo (USP), Heliana Mello (UFMG), Manuele Bandeira de Menezes (UNILAB), Mara Jucá (UEPA), Márcia Moretti (TV journalist), Maria de Lurdes Zanoli (USP), Nélia Alexandre (ULisboa), Shirley Freitas (UNILAB), Silvana Silva de Farias Araujo (UEFS), Tommaso Raso (UFMG).

## Competing Interests

The authors have no competing interests to declare.

## References

**Alexandre, N.** (2018). Aquisição do português L2 em Cabo Verde: alguns aspectos morfossintáticos do contato [Acquisition of Portuguese L2 in Cape Verde: some morphosyntactic aspects of the contact]. In M. S. D. Oliveira & G. A. Araujo (Eds.), *O português na África atlântica* (pp. 139–164). São Paulo: HUMANITAS/FAPESP.

**Bakker, P.** (2008). The development of tense, mood and aspect in creole languages and the typology of affix order. In F. Josephson & I. Söhrman (Eds.), *Interdependence of Diachronic and Synchronic Analyses* (pp. 43–59). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/slcs.103.04bak

**Bakker, P., Borschenius, F., Levisen, C.,** & **Sippola, E.** (Eds.). (2017). *Creole Studies: Phylogenetic Approaches*. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/z.211

**Bakker, P., Daval-Markussen, A., Parkvall, M.,** & **Plag, I.** (2011). Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, *26*(1), 5–42. DOI: https://doi.org/10.1075/jpcl.26.1.02bak

**Bakker, P., Post, M.,** & **van der Voort, H.** (1994). TMA Particles and Auxiliaries. In J. Arends, P. Muysken & N. S. H. Smith (Eds.), *Pidgins and Creoles. An Introduction* (pp. 247–258). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/cll.15.27bak

**Bakker, P., Sippola, E.,** & **Borchsenius, F.** (2017). Methods – on the use of networks in the study of language contact. In P. Bakker, F. Borchsenius, C. Levisen & E. Sippola (Eds.), *Creole Studies: Phylogenetic Approaches* (pp. 59–78). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/z.211.04bak

**Baptista, M., Mello, H.,** & **Suzuki, M.** (2007). The syntax of Cape Verdean Creole and Guinea-Bissau Creole. In J. Holm & P. Patrick (Eds.), *Comparative Creole Syntax* (pp. 53–82). London: Battlebridge.

**Bartens, A.** (2013). Creole Languages. In P. Bakker & Y. Matras (Eds.), *Contact Languages: a comprehensive guide* (pp. 65–158). Berlin: De Gruyter Mouton.

**Berruto, G.** (1993). Le varietà del repertorio [The varieties of the repertoire]. In A. A. Sobrero (Ed.), *Introduzione all'italiano contemporaneo: la variazione e gli usi* (vol. 2, pp. 3–36). Roma: Laterza.

**Bickerton, D.** (1981). *Roots of Language*. Ann Arbor MI: Karoma.

**Campos, E. A.** (2014). A sintaxe pronominal na variedade afro-indígena de Jurussaca: uma contribuição para o quadro da pronominalização do português falado no Brasil [The pronominal syntax in the Afro-Indigenous variety of Jurussaca: a contribution to the pronominalization system of Portuguese spoken in Brazil]. PhD Dissertation, University of São Paulo.

**Christino, B. P.** (2015). Definite articles in Huni-Kuin Portuguese. In S. Gorovitz & I. Mozillo (Eds.), *Language Contact: Mobility, Borders and Urbanization* (vol. 1, pp. 17–31). Cambridge: Cambridge Scholars Publishing.

**Daval-Markussen, A.** (2017). The typology and classification of French-based creoles – A global perspective. In P. Bakker, *et al.* (Eds.), *Creole Studies: Phylogenetic Approaches* (pp. 175–191). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/z.211.08dav

**Daval-Markussen, A.** (2019). Reconstructing Creole. PhD Dissertation, Aarhus University.

**Daval-Markussen, A.,** & **Bakker, P.** (2017). Typology of creole languages. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *The Cambridge handbook of linguistic typology* (pp. 254–286). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316135716.009

**DeGraff, M.** (2005). Linguists' most dangerous myth. The fallacy of creole exceptionalism. *Language in Society*, *34*(4), 533–591. DOI: https://doi.org/10.1017/S0047404505050207

**Dryer, M. S.,** & **Haspelmath, M.** (Eds.). (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info

**Dunn, M.** (2015). Language phylogenies. In C. Bowern & B. Evans (Eds.), *The Routledge handbook of historical linguistics* (pp. 190–211). New York: Routledge.

**Dunn, M., Terrill, A., Reesink, G., Foley, R. A.,** & **Levinson, S. C.** (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, *309*(5743), 2072–2075. DOI: https://doi.org/10.1126/science.1114615

**Felsenstein, J.** (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15. DOI: https://doi.org/10.1086/284325

**Figueiredo, C. F. G.,** & **Oliveira, M. S. D.** (2013). Português do Libolo, Angola, e português afro-indígena de Jurussaca, Brasil: cotejando os sistemas de pronominalização [Libolo Portuguese and Jurussaca Afro-Indigenous Portuguese: comparing the pronominalization systems]. *Papia*, *23*(2), 105–186.

**Freitas, S.** (2016). A origem do Papiamentu: evidências para uma convergência de hipóteses [The origin of Papiamentu: evidence for a convergency of hypotheses]. *Papia*, *26*(2), 121–235.

**Gray, R. D.,** & **Atkinson, Q. D.** (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435. DOI: https://doi.org/10.1038/nature02029

**Gray, R. D., Drummond, A. J.,** & **Greenhill, S. J.** (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, *323*(5913), 479–483. DOI: https://doi.org/10.1126/science.1166858

**Hagemeijer, T.,** & **Rocha, J.** (2017). Creole languages and genes: The case of São Tomé and Príncipe. *Faits de Langues*, *49*, 167–182. DOI: https://doi.org/10.1163/19589514-04901011

**Holm, J.** (1988). *Pidgins and Creoles, Vol. I: Theory and Structure*. Cambridge: CUP.

**Holm, J.** (1992). Popular Brazilian Portuguese: a semi-creole. In E. d'Andrade & A. Kihm (Orgs.), *Actas do Colóquio sobre Crioulos de Base Lexical Portuguesa* (pp. 37–66). Lisboa: Colibri.

**Huson, D. H.,** & **Bryant, D.** (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*, 254–267. DOI: https://doi.org/10.1093/molbev/msj030

**Intumbo, I., Holm, J.,** & **Inverno, L.** (2012). Guinea-Bissau Kriyol. In S. Michaelis, P. Maurer, M. Haspelmath & M. Huber (Eds.), *Atlas of Pidgin and Creole Language Structures* (vol. 2). Oxford: Oxford University Press.

**Inverno, L.** (2011). *Contact-induced restructuring of Portuguese morpho-syntax in interior Angola*. PhD Thesis, University of Coimbra.

**Jacobs, B.** (2012). *Origins of a Creole. The History of Papiamentu and its African Ties*. Berlin: Mouton. DOI: https://doi.org/10.1515/9781614511076

**Kato, M. A.** (1999). Strong and weak pronominals in the null subject parameter. *Probus*, *11*(1), 1–38. DOI: https://doi.org/10.1515/prbs.1999.11.1.1

**Kortmann, B.,** & **Lunkenheimer, K.** (2013). *The Mouton World Altlas of Variation in English*. De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110280128

**Lewis, M. P., Gary, F. S.,** & **Charles, D. F.** (2015). Ethnologue: languages of the world. 18ª ed. Dallas, Texas: SIL. Available online: http://www.ethnologue.com

**Lipski, J.** (2008). Angola e Brasil: vínculos linguísticos Afro-Lusitanos [Angola and Brazil: Afro-Portuguese linguistic links]. *Veredas: revista da Associação Internacional de Lusitanistas*, *9*, 83–98.

**Lopes, F. J.,** & **Oliveira, M. S. D.** (2018). Estudos sobre o português falado em Cabo Verde: o 'estado da arte' [Studies on Portuguese spoken in Cape Verde: the 'state of the art']. In M. S. D. Oliveira & G. A. Araujo (Eds.), *O português na África atlântica* (pp. 101–138). São Paulo: HUMANITAS/FAPESP.

**Lucchesi, D., Baxter, A. N.,** & **Silva, J. A. A.** (2009). A concordância verbal [Verbal agreement]. In D. Lucchesi, A. N. Baxter & I. Ribeiro (Eds.), *O português afro-brasileiro* (pp. 331–372). EDUFBA. DOI: https://doi.org/10.7476/9788523208752

**Lucchesi, D., Baxter, A. N., Silva, J. A. A.,** & **Figueiredo, C.** (2009). O português afro-brasileiro: as comunidades analisadas [The Afro-Brazilian Portuguese: the analysed communities]. In D. Lucchesi, A. N. Baxter & I. Ribeiro (Eds.), *O português afro-brasileiro* (pp. 75–100). Salvador: EDUFBA. DOI: https://doi.org/10.7476/9788523208752

**Machado, M.,** & **Vanessa, E.** (Org.). (2018). Plano de Gestão Terra Indígena Alto Rio Guamá [Management Plan of the Alto Rio Guamá Indigenous Land]. Brasília: ECAM, 2018. 106 p.; il. Avalilable at http://ecam.org.br/wp-content/uploads/2018/04/pgta-tiarg-web3.pdf

**Martiniano, J. S.** (1998). *Quilombos do Brasil Central: séculos XVIII e XIX (1719–1888). Introdução ao estudo da escravidão*. [Quilombos of Central Brazil: 18th and 19th centuries (1719–1888]. Introduction to the study of slavery). MA Thesis. Universidade Federal de Goiás.

**Massini-Cagliari, G.** (2004). Language policy in Brazil: Monolingualism and linguistic prejudice. *Language Policy*, *3*(1), 3–23. DOI: https://doi.org/10.1023/B:LPOL.0000017723.72533.fd

**Mateus, M., Mira, H.,** et al. (2003). *Gramática da Língua Portuguesa*. [Portuguese Grammar]. 7 ed (pp. 489–502). Lisboa: Caminho.

**Mattos, A. P. B.** (2016). *The speech variety of Kalunga: an Afro-Brazilian community in Goiás, Brazil*. MA Thesis, Aarhus University.

**Mattos, A. P. B.** (2017). *Fieldwork notes on Barreirão rural community*. Manuscript.

**Mattos, A. P. B.** (2019). Kalunga: An Afro-Brazilian Portuguese Variety. Ph.D. Thesis, Aarhus University.

**Mattos, A. P. B.** (*in press*). The Afro-Brazilian community Kalunga: linguistic and sociohistorical perspectives. In E. Sippola & D. Perez (Eds.), *Postcolonial varieties in the Americas*. Berlin: De Gruyter Mouton.

**Mattos, A. P. B.,** & **Oliveira, M. S. D.** (2017). *A filogênese e o contexto lusófano: Kalunga, outras 'variedades' e línguas crioulas* [Phylogenesis and the Lusophone context: Kalunga, other 'varieties' and creole languages]. Paper presented at VII GELIC. Federal University Santa Catarina. Manuscript.

**Maurer, P.** (2013). Papiamentu. In S. Michaelis, P. Maurer, M. Haspelmath & M. Huber (Eds.), *The survey of pidgin and creole languages. Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford: Oxford University Press.

**McMahon, A.,** & **McMahon, R.** (2003). Finding families: Quantitative methods in language classification. *Transactions of the Philological Society*, *101*(1), 7–55. DOI: https://doi.org/10.1111/1467-968X.00108

**McMahon, A.,** & **McMahon, R.** (2006). Why linguists don't do dates: Evidence from Indo-European and Australian languages. In P. Forster & C. Renfrew (Eds.), *Phylogenetic Methods and the Prehistory of Languages* (pp. 153–160). Cambridge: McDonald Institute for Archaeological Research.

**McWhorter, J. H.** (1998). Identifying the creole prototype: Vindicating a typological class. *Language*, 788–818. DOI: https://doi.org/10.2307/417003

**McWhorter, J. H.** (2005). *Defining Creole*. Oxford: Oxford University Press.

**Mello, H.** (2012). Os corpora orais e o C-ORAL-BRASIL [The oral corpora and the C-oral Brazil]. In T. Raso & H. Mello (Eds.), *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal* (vol. 1, pp. 31–54). Belo Horizonte: Editora UFMG.

**Mello, H.** (2016). *Estudo empírico da fala baseado em corpus* [Empirical study of corpus-based speech]. Habilitation Thesis. Federal University of Minas Gerais.

**Mello, H. R.** (1996). *The genesis and the development of the Vernacular Brazilian Portuguese*. PhD Thesis, University of New York.

**Michaelis, S.,** et al. (2013). *Atlas of Pidgins and Creole Language Structures*. Oxford: Oxford University Press.

**Monghilhott, I.,** & **Coelho, I.** (2002). Um estudo da concordância verbal de terceira pessoa em Florianópolis [A study of the third person verbal agreement in Florianópolis]. In P.

Vandresen (Ed.), *Variação e mudança no português falado na Região Sul* (pp. 189–216). Pelotas: EDUCAT.

**Monte, A.** (2012). *Concordância verbal e variação: um estudo descritivo-comparativo do português brasileiro e do português europeu* [Verbal agreement and variation: a descriptive-comparative study of the Brazilian and European Portuguese]. PhD Thesis. São Paulo State University.

**Mufwene, S.** (1996). The Founder Principle in creole genesis. *Diachronica, 13*, 83–134. DOI: https://doi.org/10.1075/dia.13.1.05muf

**NAEA – Núcleo de Altos Estudos da Amazônia 'Nucleus of Advanced Studies of Amazonia'.** (2005). Quilombos do Pará, cd-rom. Belém: NAEA/UFPA & Programas Raízes.

**Nakhleh, L., Warnow, T., Ringe, D.,** & **Evans, S. N.** (2005). A comparison of phylogenetic reconstruction methods on an IE dataset. *Transactions of the Philological Society, 3*(2), 171–192. DOI: https://doi.org/10.1111/j.1467-968X.2005.00149.x

**Oliveira, M. S. D.,** et al. (2015). O Conceito de Português Afro-Indígena e a Comunidade de Jurussaca [The concept of the Afro-Indigenous Portuguese and the Jurussaca community]. In J. O. Avelar & L. A. López (Eds.), *Dinâmicas afro-latinas – língua(s) e história(s)* (pp. 149–178). Frankfurt am Main: Peter Lang.

**Oliveira, M. S. D., Campos, E. A.,** & **Fernandes, J. T. V.** (2011). Repensando a escola em Jurussaca a partir da "norma dos pronomes pessoais da comunidade" [Rethinking the school in Jurussaca from the 'norms of the pernonal pronouns of the community]. In A. S. A. Cunha (Ed.), *Entendendo quilombos, descontruindo mitos – a educação formal e a realidade quilombola no Brasil* (vol. 1, pp. 129–144). Guimarães, MA: SETAGRAF.

**Parkvall, M.** (2000). *Out of Africa*. London: Battlebridge.

**Perez, D. M., Sessarego, S.,** & **Sipolla, E.** (2017). Afro-Hispanic varieties in comparison – new light from phylogeny. In P. Bakker, F. Borchsenius, C. Levisen & E. Sippola (Eds.), *Creole Studies: Phylogenetic Approaches* (pp. 269–292). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/z.211.12per

**Petter, M.** (2009). O continuum Afro-Brasileiro do Português [The Afro-Brazilian continuum of Portuguese]. In C. Galves, H. Charlotte, R. Helder & F. Rosa (Eds.), *África-Brasil: Caminhos da Língua Portuguesa* (pp. 265–284). Campinas: Editora UNICAMP.

**Prokić, J.** (2010). *Families and Resemblances*. Groningen: Groningen Dissertations in Linguistics.

**Prokić, J.,** & **Nerbonne, J.** (2008). Recognizing groups among dialects. In J. Nerbonne, C. Gooskens, S. Kürschner & R. van Bezooijen (Eds.), *International Journal of Humanities and Arts Computing – Special Issue on Language Variation, 2*, 153–172. DOI: https://doi.org/10.3366/E1753854809000366

**Quint, N.** (1998). *Grammaire de la langue cap verdiene* [Grammar of the Cape Verdean language]. Paris – France: L'Harmattion.

**Quint, N.** (2003). *Parlons Capverdien – langue et culture* [Let's Speak Capeverdean: language and culture]. Paris: L'Harmattan.

**Quint, N.** (2005). *Le Créole Capverdien de poche* [The Cape Verdean pocket book]. Chennevières-sur-Marne: Assimil.

**Quint, N.** (2009). *O caboverdiano de bolso* [The Cape Verdean pocket book]. Chennevières-sur-Marne: Assimil.

**Raposo, E. B.,** et al. (Org.) (2013). *Gramática do português* [Portuguese Grammar]. Lisboa: Fundação Calouste Gulbenkian.

**Rúbio, C. F.** (2012). *Padrões de concordância e de alternância pronominal no português brasileiro e europeu: estudo sociolinguístico comparativo* [Patterns of pronominal agreement

and alternation in Brazilian and European Portuguese: a comparative sociolinguistic study]. PhD Thesis. University of São Paulo.

**Sicoli, M. A.,** & **Holton G.** (2014). Linguistic Phylogenies Support Back-Migration from Beringia to Asia. *Plos One*, *9*(3): e9172. DOI: https://doi.org/10.1371/journal.pone.0091722

**Siegel, J.** (2007). Recent evidence against the Language Bioprogram Hypothesis: the pivotal case of Hawai'i Creole. *Stud. Lang*, *31*, 51–88. DOI: https://doi.org/10.1075/sl.31.1.03sie

**Simons, G. F.,** & **Fennig, C. D.** (Eds.). (2018). *Ethnologue: Languages of the World, Twenty-first edition*. Dallas, Texas: SIL International. http://www.ethnologue.com (retrieved on October 2, 2018).

**Sippola, E.** (2017). Similarities and differences among Iberian creoles. In P. Bakker, F. Borchsenius, L. Carsten & E. Sippola (Eds.), *Creole studies: Phylogenetic approaches* (pp. 241–268). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/z.211.11sip

**Souza, P. D. S.** (2005). *A variação na concordância verbal na primeira fase do período arcaico da língua portuguesa: séculos XIII–XIV* [Variation in the verbal agreement in the first phase of the archaic period of the Portuguese language: 13th–14th centuries]. MA Thesis. Federal University of Bahia.

**Teixeira, E. P.,** & **Araujo, S. S. F.** (Eds.). (2017). *Diálogos entre Brasil e Angola – o português d'aquém e d'além-mar* [Dialogues between Brazil and Angola – the Portuguese below and beyond the sea]. Feira de Santana, BA: UEFS Editora.

**Thomason, S. G.,** & **Kaufman, T.** (1988). *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

**Winford, D.** (2003). *An introduction to contact linguistics*. Oxford: Blackwell.

**Winford, D.** (2018). Creole Tense–Mood–Aspect Systems. *Annual Review of Linguistics*, *4*, 193–212. DOI: https://doi.org/10.1146/annurev-linguistics-011516-034054

**How to cite this article:** Mattos, A. P. B., & Oliveira, M. S. D. (2020). Kalunga in the lusophone context: A phylogenetic study. *Journal of Portuguese Linguistics,* 19: 2, pp. 1–24. DOI: https://doi.org/10.5334/jpl.224

**Submitted:** 01 May 2019 **Accepted:** 21 December 2019 **Published:** 23 March 2020

**Copyright:** © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

]u[ *Journal of Portuguese Linguistics* is a peer-reviewed open access journal published by Ubiquity Press. **OPEN ACCESS**