Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5047–5053
Marseille, 20-25 June 2022

The PALMA Corpora of African Varieties of Portuguese

Tjerk Hagemeijer, Amália Mendes, Rita Gonçalves, Catarina Cornejo, Raquel Madureira, Michel Généreux
University of Lisbon, School of Arts and Humanities, Centre of Linguistics
{t.hagemeijer, amaliamendes, ritamgg}@letras.ulisboa.pt, catarinacornejo@gmail.com
raquelmadureira@edu.ulisboa.pt, michel.genereux@claurendeau.qc.ca

Abstract

We present three new corpora of urban varieties of Portuguese spoken in Angola, Mozambique, and São Tomé and Príncipe, where Portuguese is increasingly being spoken as first and second language in different multilingual settings. Given the scarcity of linguistic resources available for the African varieties of Portuguese, these corpora provide new, contemporary data for the study of each variety and for comparative research on African, Brazilian and European varieties, hereby improving our understanding of processes of language variation and change in postcolonial societies. The corpora consist of transcribed spoken data, complemented by a rich set of metadata describing the setting of the audio recordings and sociolinguistic information about the speakers. They are annotated with POS and lemma information and made available on the CQPweb platform, which allows for sophisticated data searches. The corpora are already being used for comparative research on constructions in the domain of possession and location involving the argument structure of intransitive, monotransitive and ditransitive verbs that select Goals, Locatives, and Recipients.

Keywords: comparable spoken corpora, African varieties of Portuguese, language variation and change

1. Introduction

After the independences in the 1970s, the former Portuguese colonies in Africa adopted the standard of European Portuguese as their exclusive official language. Ever since, there has been an increasing body of work on the African varieties of Portuguese (AVPs), in particular on the Angolan and Mozambican varieties, a tendency which accompanies a more generalized interest in postcolonial varieties and language acquisition, variation, and change (e.g., Álvarez López et al., 2018; Gonçalves, 2010; Mesthrie and Bhatt, 2008).[1]

Despite this tendency, there are still many gaps in our knowledge of AVPs, since in-depth research is often concentrated on a relatively small number of linguistic domains or features (e.g., number agreement, clitics, argument structure) and mostly restricted to single varieties or subvarieties. Language contact with Bantu and creole languages is frequently considered to be the driving force behind the observed patterns. This type of conclusion, however, is often the consequence of a methodology which strongly focuses on the patterns diverging from European Portuguese, the exclusive official language adopted by the former Portuguese colonies in Africa, and hereby often overlooks the substantial group and individual variation that characterizes these varieties, including convergence with standard European Portuguese (cf. section 6 for some examples).[2]

While an important set of contemporary corpora for the Brazilian and European variety of Portuguese is available[3], few searchable resources are (publicly) available for the AVPs (cf. section 2), which is particularly striking if we consider that Angola and Mozambique are and will increasingly be demographic centres of the Portuguese language, with a current total of over 60 million inhabitants and an estimate of approximately 210 million inhabitants in 2100 (Müller de Oliveira, 2016), most of which are and will be Portuguese speakers.

The PALMA project, hosted by the Centre of Linguistics of the University of Lisbon (CLUL), seeks to answer these shortcomings by creating new, comparable, contemporary, spoken linguistic corpora of the urban varieties spoken in the capitals of Angola (ANG), Mozambique (MOZ), and São Tomé and Príncipe (STP) (Figure 1).

Figure 1: Map of the countries represented in the corpus.

The relevance of these corpora also stems from the fact that the urban varieties spoken in the countries' capitals, especially in Luanda and São Tomé, show a high degree of nativization of Portuguese, including substantial monolingualism.[4] This characteristic distinguishes

[1] The linguistic research on AVPs has been compiled and is frequently updated at the Cátedra de Português Língua Segunda e Estrangeira, hosted by the Universidade Eduardo Mondlane: https://catedraportugues.uem.mz/variedades-nao-europeias.

[2] For an overview of the issues laid out in this paragraph and linguistic features, see, for instance, Gonçalves (2010, 2013) and Hagemeijer (2016).

[3] The PORTULAN CLARIN Infrastructure for the Portuguese language (https://portulanclarin.net) has a set of resources available in its repository. A listing of corpora for Portuguese is also available at Linguateca (https://www.linguateca.pt).

[4] The latest national censuses show that Portuguese is the most spoken language at home for 85% of the urban Angolan population (INE, 2016); 98,4% of the Santomean population speaks Portuguese (INE, 2013); 42% of the urban Mozambican

Portuguese from other former colonial and currently official languages in Africa, in particular French and English, which are typically L2.
Since the shift to L2 and L1 Portuguese is a 20th century phenomenon, which was accelerated by the independences and the massification of schooling in Portuguese, language change occurs at a faster pace than in stable L1 societies. Therefore, the three new corpora can be compared with older data sets and descriptions in order to assess ongoing linguistic changes and how and what features have been crystallizing.
The three corpora are currently used to support the specific research topics of the PALMA project, namely the syntax and semantics of possession and location in these AVPs, but are expected to function as an important tool to widen and deepen linguistic research on spoken AVPs, with potential for language planning.
In section 2 we will review other existing resources for AVPs and in section 3 we present the contents and metadata of the PALMA corpora. Section 4 is reserved for the discussion of the pre-processing and annotation of the corpus, while section 5 lays out how it can be explored in CQPweb. Section 6 contains the conclusions.

2. Review of corpora of African varieties of Portuguese

Another set of comparable corpora of African varieties was compiled by CLUL, covering five countries which have Portuguese as an official language – Angola, Mozambique, São Tomé and Príncipe, Cape Verde, and Guinea-Bissau. However, note that, in the latter two, Portuguese is typically a non-native language and, differently from the first three, does not fulfil the role of lingua franca.
The five corpora, which constitute the Corpus África, are around 640,000 words each and have the same percentage of spoken and written subparts (approx. 25,000 spoken words (4%) and approx. 615,000 written words (96%)). The written subcorpus is divided in newspapers (50%), literature (20%) and miscellaneous (26%) (Bacelar do Nascimento et al., 2006). The spoken subcorpus includes data from 1990 to 2006. The corpora are available through CQPweb.[5]
The Mozambique section of the Corpus África contains part of the materials collected and used for the large-scale project Panorama do Português Oral de Maputo (Stroud and Gonçalves, 1997; Gonçalves and Stroud, 1998, 2000), which constitutes a standard for research on AVPs.
In addition to the searchable corpora above, larger spoken data sets of AVPs can sometimes be found in individual research. Gonçalves (1991) contains an appendix with a spoken Mozambican corpus, collected in 1986-1987 for her doctoral dissertation, comprising 140,000 words based on 40 interviews with university students between 19 and 22 years old. This corpus was digitalized, annotated with POS and lemma in the scope of the project METANET4U[6], and is also available on the CQPweb platform.
Similarly, the appendix to Chavagne's (2005) doctoral dissertation contains a spoken interview-based corpus of urban Angolan Portuguese, comprising roughly 65,000 words. While other researchers have collected and used corpora for their work, we are not aware of other spoken, transcribed corpora of this type which have been made publicly available.

3. The PALMA corpora

The PALMA corpora encompass orthographic transcriptions of 205 semi-structured interviews, with a total of over 108 hours of audio recordings, corresponding to 1,097,702 tokens and 27,027 types, as shown in Table 1 below. Numbers in the table are counted after the tokenization of the corpus (see section 4) and include the speech of both the informants and the interviewers who are native speakers of these varieties. The interviews were audio recorded by nine different interviewers, of which five were speakers of the AVPs in question and four speakers of European Portuguese. The corpus has a total number of 1,356,614 tokens, which includes the tokens produced by non-native interviewers. The interviews, using a headset or handheld microphone, were carried out in different locations in or near the capitals of Angola, Mozambique, and São Tomé and Príncipe.

	inter-views	**hours**	**tokens (produced by native speakers)**	**years of recording**
ANG	58	34	393,745	2012, 2013, 2019
MOZ	70	42	380,958	2010, 2020
STP	77	32	322,999	2008, 2011, 2012
Total	**205**	**108**	**1,097,702**	

Table 1: Numbers for the three PALMA corpora.

From the total of interviews recorded, a selection was made based on the informant's profile and the need to balance sociolinguistic criteria, as well as on the quality of the recordings, which includes the acoustics (many interviews were recorded in domestic and informal environments and/or outside) and, for instance, the talkativeness of the speakers. The length of the recordings with individual speakers is on average around 30 minutes. Each interview starts out with a questionnaire regarding the informant's sociolinguistic profile, including language use in the case of those who speak more languages in addition to Portuguese. The interviews are semi-structured: informants were mostly asked questions regarding their daily life, but without following any particular previously established script. The purpose of the interviews was explained beforehand to the informants, who gave the interviewers their written or spoken informed consent. From a sociolinguistic perspective, the corpora are fairly well balanced internally and among each other: within each corpus, the speakers are on average in their mid-thirties, quite evenly distributed over 4 age groups and 4 educational profiles, and gender proportions are balanced, which is shown in Table 2 below.

population most frequently speaks Portuguese at home (INE, 2019).

[5] http://gamma.clul.ul.pt/CQPweb/corpusafrica (CQPweb requires a registrated user account in order to log in.)

[6] http://metanet4u

In addition, a large majority of the speakers uses Portuguese as their L1 or primary language, although many of them also have some degree of proficiency in one or more languages from the Bantu group (Angola and especially Mozambique) or creoles (São Tomé and Príncipe). In this respect, the differences among the AVPs are related to the different sociolinguistic profiles of the three capitals where the recordings took place. With respect to the Angola corpus, it should be noted that most of the interviews (48/58) were originally recorded by Afonso Miguel for his doctoral research on Bantu loanwords in Portuguese spoken in Luanda (Miguel, 2019), and exhibit a more active role of the interviewer.

Subparts of the three corpora have been used for two other doctoral dissertations (Gonçalves, 2016; Nascimento, 2018), master theses (Cornejo, 2021; Gonçalves, R. 2010; Justino, 2011), and several research papers (e.g., Brandão, 2013; Pereira et al., 2018).

		ANG		MOZ		STP		TOTAL
		m	f	m	f	m	f	
Age	17-25	9	9	11	12	9	10	60
	26-35	6	6	11	11	10	11	55
	36-45	6	6	5	5	10	6	38
	46 >	9	7	10	5	11	10	52
Schooling	0-4 years	2	5	3	1	5	8	24
	5-9 years	12	6	11	7	11	12	59
	10-12 years	8	6	15	11	19	10	69
	higher education	8	11	8	14	5	7	53
								205

Table 2. Distribution of informants according to variety, gender, age, and schooling.

3.1 Corpus metadata

A set of detailed metadata describes the settings of the recording, the interviewer, and the sociolinguistic profile of the informant. The interviews were conducted by speakers of European Portuguese and by speakers of the three AVPs.

(i) Metadata describing the recording situation:
- Country
- Year of recording
- Place of recording
- Type of recording (interviews)

(ii) Metadata describing the interviewer(s):
- Code of the interviewer (in case there is more than one interviewer, this metadata field is duplicated)[7]
- Country of origin of the interviewer

(iii) Metadata describing the informant:
- Age
- Gender
- Level of schooling (including detailed information)
- First language
- Place of birth
- Place of residence
- Occupation
- Other languages spoken

Metadata are kept in a database and exported as a tab delimited file to be integrated in the CQPweb platform (cf. section 5).

3.2 Transcription of recordings

The corpora are orthographically transcribed in standard textual format (a version of the CHAT format (Mac Whinney, 1994; Cresti and Moneglia, 2005)) with the annotation of speaker turns. The textual string is divided into utterances. The transcriptions keep a faithful description of what is said by the interviewer and by the informant. Morphological forms with a different grammatical use from standard European Portuguese are kept as long as these forms are registered in dictionaries, for instance, in case of constructions lacking number agreement: *as casa bonita* instead of *as casas bonitas* 'the beautiful houses'. Loanwords from the languages in contact with Portuguese were kept and, where necessary, adapted, since not all of them have (yet) been integrated in Portuguese dictionaries. The main non-linguistic and paralinguistic acoustic events in the speech flow are reported into transcripts. Overlapping speech, prosodic breaks, extralinguistic elements, incomprehensible words or sequences, and disfluencies such as filled pauses are all encoded in the transcription. Repetitions and unfinished words are marked in order to be singled out during the development of lexical resources for these three AVPs. Repetitions are frequently cases of filled pauses, when the speaker is planning the utterances ahead. Therefore, when counting the occurrences of a type it is useful to be able to eliminate such repetitions. Cases of abandoned structures are also labelled to enable future discourse studies. A list of the labels used to mark the properties of spoken language in the transcriptions is provided in Table 3.

Labels	Speech features
/	short prosodic break
//	end of a declarative or exclamative utterance
?	end of an interrogative utterance
+	word or utterance interrupted by another speaker or abandoned by the informant or repeated (e.g., estudei+ 'I studied')
&word+	incomplete word, interrupted by another speaker or abandoned by the informant
(word)	word unclear due to audio quality or phonetic phenomena
xxx	incomprehensible word
yyyy	sequence of incomprehensible words
hhh	extralinguistic element

[7] Eight interviews of the São Tomé corpus were conducted by two interviewers.

&ah, &eh, &hum	filled pauses
[token] [token]	utterances from two different speakers that overlap

Table 3: Labels used for transposing speech features in the transcriptions.

Example (1) provides an excerpt from the S. Tomé and Príncipe corpus and illustrates some of the codes used in the transcription:

(1) *porque a ideia foi nossa // isso foi através da+ porque éramos de+ éramos+ não tínhamos um mercado / vendíamos na rua / hhh e então a gente começámos a crescer / e outros colega nós éramos mais novos // começámos a crescer e dissemos que / de qualquer forma a gente não podia viver daquele jeito para todo o tempo //*
Translation: because the idea was ours // that was through the+ because we were from+ were+ we didn't have a market / we were selling on the street / hhh and then we started growing up / and other colleague we were younger // we started growing up and we said that / in any case we couldn't live like that forever //

Here, short prosodic breaks and the end of a declarative utterance are marked with / and //, respectively; *hhh* marks an extralinguistic element; a word followed by the sign + can correspond to abandoned content (the speaker rephrases their utterance immediately after uttering that word) or repeated content (the speaker repeats a fragment including that word or that word alone). Notice that the lack of number agreement in the Noun Phrase *outros*PL *colega*SG 'otherPL colleague' is kept in the transcription.
The transcriptions are anonymized by removing personal information of the participants. Speech turns are identified as ENT for the interviewer and INF for the informant, and each interviewer and informant receives an arbitrary numerical code.

4. Corpus Annotation

The corpus has been tokenized and tagged with POS and lemma information. For tokenization, we applied the LX-tokenizer, which splits punctuation marks from words and detects sentence boundaries (Branco and Silva, 2004). This tokenizer is developed specially for Portuguese and can handle typical Portuguese phenomena such as contracted word forms and clitics (including cases of mesoclisis, which are rare in in spoken Portuguese).
We tagged the corpus with a version of MBT (Daelemans et al., 1996), a memory-based tagger, trained on the written CINTIL corpus (644K tokens) that contains 80 POS-tag labels (multiword unit labels were left out) (Barreto et al., 2006). The tagger was used to tag the written subset of the Reference Corpus of Contemporary Portuguese (CRPC) and reached 95,5% accuracy (Généreux et al., 2012).
For the lemmatization, we use a version of MBLEM (Van den Bosch and Daelemans, 1999), adapted to Portuguese to lemmatize the CRPC with a reported accuracy of 96,7% (Généreux et al., 2012). MBLEM combines a dictionary lookup with a machine learning algorithm to tag words with their lemmas. As a dictionary list we used an in-house produced list of lemma and wordform-POS mappings. The dictionary list consists of 102,000 word forms mapped to 27,860 lemmas with a total of 120,768 wordform-lemma combinations. MBLEM uses the POS information to limit the set of possible lemmas for each word form.
The tagger was trained on written texts and on the European Portuguese variety and is applied to three corpora of spoken transcriptions of other varieties of Portuguese, which affects the quality of the annotation. The corpora contain a number of lexical items that are borrowed from the Bantu and creole languages that do not occur in Portuguese dictionaries, in particular nouns. The tagger was unable to recognize these forms, and as a result they are often annotated as proper names instead of common nouns.
As mentioned above, some interviewers are speakers of European Portuguese and not of one of the AVPs. For an adequate description of each AVP, in corpus queries it is important to be able to distinguish between word forms that were produced by native speakers, either interviewer or interviewee, and word forms produced by non-native speakers of that variety. With this goal in mind, besides POS and lemma, a third level of annotation "speaker" was added with two labels: "ns" for native speakers and "nns" for non-native speakers.
Finally, two structure attributes were added. A first level attribute is <text>. A second level of structure attribute defines speech turn boundaries and marks the beginning and the end of the interviewer's speech turn, and the beginning and the end of the informant's answer, <ent> and <inf> respectively. By combining these structural units with the metadata of the variety spoken by the participants, a third positional annotation was then applied to each token.

5. The PALMA corpora on CQPweb

The corpora have been made available on CQPweb (Hardie, 2012), an online interface that allows users to query the corpora for concordances of word forms, sequences of words, and different layers of annotation, with additional functionalities such as collocations. While CQPweb is especially suited for written corpora and not prepared to handle some of the speech features that are labelled in the transcriptions, the decision to install the corpora on the same platform as other corpora of interest for the work on the AVPs relates to the possibility of enabling contrastive and diachronic analysis of data, annotated with the same tagset and tools, in particular CRPC, useful to compare word forms in the PALMA corpora with the European Portuguese variety, as well as the Corpus África, which includes data for ANG, MOZ and STP from the period 1990-2006.

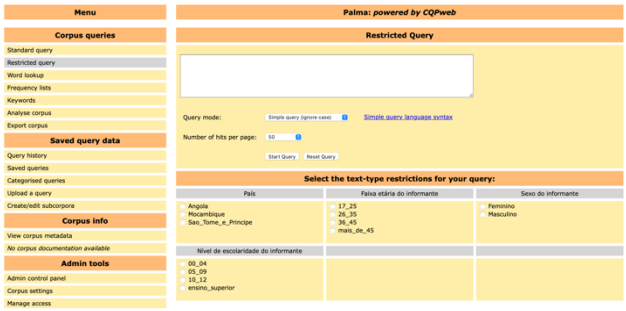

Figure 2: The restricted query options on the CQPweb for the PALMA corpora.

The CQPweb platform allows users to create frequency lists and to restrict the search query according to the following metadata regarding the informants: 'country' (Angola, Mozambique, São Tomé and Príncipe), 'age' (4 groups), 'gender' (male/female), and 'level of schooling' (four groups) (cf. Figure 2). The restricted queries, by making full use of the metadata of the three corpora, enable cross-linguistic comparison through sophisticated search patterns using the *simple query language syntax* (Hoffman et al., 2008) and the *CQP syntax* (Evert, 2009).

The annotated corpora were converted to .vrt files, with two levels of structural attributes: <text>, corresponding to each single corpus file, and turn, either interviewer <ent> or informant <inf>; and three positional attributes (POS, lemma and speaker (native or non-native). Queries using restrictions on metadata may be combined with POS categories and lemmas. Also, using the structural annotation of speech turns, it is possible to restrict the queries to the speech turn of the interviewers or the informant. Using the layer of annotation "speaker", it is possible to exclude word forms produced by non-native interviewers from the hits of the corpus query, in order to restrict the search results to data of the AVPs only.

Queries of word forms, POS, and lemmas can be performed by using the *simple query syntax*. Queries that make use of the structural units or the additional layer of annotation "speaker" require the queries to be made with the *CQP syntax*. The two options are available in the main window of the CQPweb. In Table 4 we provide some examples of possible queries in the 3 corpora using the *CQP syntax* and combining the different layers of annotation.

	query in CQP syntax	**results**
1	[word = "sempre" & inf]	word form *sempre* 'always' produced by the informants
2	[word = "sempre" & ent]	word form *sempre* produced by the interviewers
3	[word = "canto" & pos = "CN.*" & falante = "ns"]	word form *canto* with POS "common noun", produced by participants that are native speakers of the variety (either interviewers or informants)
4	[word = "paizinho" & falante = "nns"]	form of address "paizinho" 'daddy' produced by non-native speakers of the variety
5	[word = "sempre" & falante = "ns" & ent]	word form *sempre* produced by interviewers that are native speakers

Table 4: Examples of queries using different layers of annotation.

Queries 1 and 2 restrict the interviewer or informant, query 3 combines word form, POS tag, and native speaker, and query 5 deals with word form, native speaker and interviewer. Query 4 has no hits in the corpora, while the same query with the value "ns" provides 4 hits, which leads us to conclude that this specific form of address is only used by native speakers. If we add to query 4 the structural unit "& ent", we obtain the same 4 results, leading to the additional conclusion that the form of address is always used by interviewers that are native speakers of the variety. For each hit, detailed metadata may be accessed through the option "File info for text". In the case of query 5, file info indicates that the form of address is always produced by the same interviewer. Besides the corpus queries, the CQPweb interface offers powerful statistical analyses, which are useful to find collocations or keywords. Details on the annotation and the different search options for this corpus are laid out in the annotation manual (Mendes et al.,2022).

6. Conclusion

The available statistical data from the national censuses show that the increase in number of L1 and L2 speakers of Portuguese in the three African countries at stake constitutes an unprecedented case in the context of former colonial languages in Africa (Hagemeijer, 2016). In order to investigate patterns of variation in these increasingly stabilizing urban varieties of Portuguese, new contemporary data are required. This new set of three comparable corpora answers the shortage of linguistic resources for Angola, Mozambique, and São Tomé and Príncipe, providing input for linguistic studies of variation within and across these nativized/nativizing AVPs, and for comparisons with other varieties of Portuguese. The availability of detailed sociolinguistic information of the informants will further contribute to our understanding of the factors that drive variation and change, such as language contact, schooling, and age. The comparable corpora will thus allow us to perform quantitative, informant-based analyses to detect patterns of microvariation and to relate these patterns across structures and across AVPs.

The three online corpora have already been put to use for contrastive studies carried out within the PALMA project on constructions in the domain of possession and location. For instance, work by Gonçalves, Duarte and Hagemeijer (forthc.) shows that, apart from substantial convergence with standard Portuguese – which is often given much less attention in the literature but is also a relevant finding on its own –, each variety exhibits different non-standard solutions for the expression of Recipient (dative) arguments in ditransitive structures. In addition to the standard European Portuguese strategy, where the Recipient is introduced by preposition *a* 'to' (*Dei o livro ao João* 'I gave the book to João'), strategies with other prepositions, in particular *para* 'to, for' and *em* 'in' (*Dei o livro para/em o João* 'I gave the book to João') or without prepositions, i.e. double object constructions (*Dei o João o livro* 'I gave John the book'), are also attested. Similarly, in the domain of the Goal arguments of the verbs of directed motion *ir* 'to go' and *chegar* 'to arrive', Hagemeijer et al. (forthc.) detected substantial variation between the standard use of preposition *a* 'to' and non-canonical *para* 'to, for', *em* 'in, and Ø (*João foi {a/para/em/Ø/Lisboa}* 'João went to Lisbon').

These preliminary findings also show the autonomy of each variety with respect to other varieties of Portuguese, rendering void the erroneous idea of the existence of an overarching variety, in addition to European and Brazilian Portuguese, that has sometimes been called "African Portuguese". Moreover, individual speakers often use different patterns (between standard and non-standard), which shows that competition between features and grammars starts, in fact, at the speaker-level.

The constitution of new, comparable corpora is therefore an important achievement on its own, because they relate

to the past and the future. On the one hand, the new data can be compared to the relevant older data of corpora, theses, and publications, especially for the better documented case of Mozambique, giving an insight into the rate and type of change that affect newly emerging L1 varieties. On the other hand, the corpora will enable future research on other structures and have the potential to contribute to the fields of educational linguistics (e.g., teacher training and materials) and language planning. Although the reference variety in the former Portuguese colonies continues to be European Portuguese, gradual development of new norms, a process of *Ausbau* (Kloss, 1967), may occur, as shown by the case of Brazilian Portuguese, and further promote the status of Portuguese as a pluricentric language.

Additionally, we plan to further enhance the corpora by creating text-audio alignment using the EXMARaLDA tool (Schmidt, 2012) and share the results on the TEITOK platform, which will be particularly useful for research in the domain of phonetics and phonology, two significantly understudied domains when it comes to the AVPs.

The PALMA corpora will be distributed through the PORTULAN CLARIN repository for academic use, with a CreativeCommons license, pending authorisation for user authentication.

7. Acknowledgements

This work was developed as part of the project "Possession and location: microvariation in African varieties of Portuguese" (PALMA), supported by the Fundação para a Ciência e a Tecnologia (FCT, Portugal) under the grant PTDC/LLT-LIN/29552/2017.

8. Bibliographical References

Álvarez López, L., Gonçalves, P. and Avelar, J. (eds.) (2018). *The Portuguese language continuum in Africa and Brazil*. Amsterdam: John Benjamins Publishing Company.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela, A., Bettencourt Gonçalves, J., Oliveira, S. M. and Santos, R. (2006). The African Varieties of Portuguese: Compiling Comparable Corpora and Analysing Data-derived Lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006),* pages 1791-1794, Genoa, Italy. ELRA.

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M.F., Nunes, F. and Silva, J. (2006). Open resources and tools for the shallow processing of portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy. ELRA.

Branco, A. and Silva, J. (2004). Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004*), pages 507-510. ELRA.

Brandão, S. (2013). Patterns of agreement within the Noun Phrase. *Journal of Portugese Linguistics*, 12(2): 51-100.

Chavagne, J.-P. (2005). *La langue portugaise d'Angola: étude des écarts par rapport à la norme européenne du portugais*. Doctoral dissertation. Lyon: Université Lumière.

Cornejo, C. (2021). *Forms of address in formal situations in Angolan Portuguese*. MA thesis. Ljubljana: Faculty of Arts of the University of Ljubljana.

Cresti, E. and Moneglia, M. (eds.) (2005). C-*ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins.

Daelemans, W., Zavrel, J., Van den Bosch, A., and Van der Sloot, K. (2010). *MBT: Memory-Based tagger*. Reference guide. ILK Technical Report Series 10-04.

Evert, S. (2009). *The CQP Query Language Tutorial*. https://cwb.sourceforge.io/temp/CQPTutorial.pdf

Généreux, M., Hendrickx, I. and Mendes, A. (2012). Introducing the reference corpus of contemporary portuguese on-line. In *Proceedings of LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 2237-2244. Istanbul, Turkey. European Language Resources Association (ELRA).

Gonçalves, P. (1991). *A construção de uma gramática de português em Moçambique: aspetos da estrutura argumental dos verbos.* Doctoral dissertation. Lisbon: Universidade de Lisboa.

Gonçalves, P. (2010). *A génese do português de Moçambique*. Lisboa: INCM.

Gonçalves, P. (2013). *O português em África*. In E. P. Raposo et al. (orgs.), *Gramática do português*, vol. 1, Lisbon: Fundação Calouste Gulbenkian.

Gonçalves, P. and Stroud, C. (orgs.) (1998). *Panorama do Português Oral de Maputo*, Vol. III. Maputo: INDE.

Gonçalves, P. and Stroud, C. (orgs.) (2000). *Panorama do português oral de Maputo*, Vol. IV. Maputo: INDE.

Gonçalves, R. (2010). *Propriedades de subcategorização verbal no português de São Tomé*. MA thesis. Lisbon: Universidade de Lisboa.

Gonçalves, R. (2016). *Construções ditransitivas no português de São Tomé.* Doctoral Dissertation. Lisbon: Universidade de Lisboa.

Gonçalves, R., Duarte, I. and Hagemeijer, T. (forthc.). Dative microvariation in African varieties of Portuguese. *Journal of Portuguese Linguistics*.

Hagemeijer, T. (2016). O português em contacto em África. In M. Martins and E. Carrilho (eds.), *Manual de linguística portuguesa*. Berlin: Mouton de Gruyter, pp. 43-67.

Hagemeijer, T., Leal, A., Madureira, R. and Cordeiro, J. (forthc.). Goal arguments of *ir* 'to go' and *chegar 'to arrive*' in three African varieties of Portuguese. *Journal of Portuguese Linguistics*.

Hardie, A. (2012). CQPweb-combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3): 380-409.

Hoffmann, S., Evert, S., Smith, N., Lee, D., and Berglund-Prytz, Y. (2008). *Corpus linguistics with BNCweb-a practical guide* (Vol. 6). Peter Lang.

INE (Instituto Nacional de Estatística de São Tomé e Príncipe) (2013). *IV Recenseamento geral da população e da habitação 2012. Características educacionais da população*.

INE (Instituto Nacional de Estatística de Angola) (2016). *Resultados definitivos do recenseamento geral da população e da habitação de Angola 2014.*

INE (Instituto Nacional de Estatística de Moçambique) (2019). *IV Recenseamento geral da população e habitação 2017.*

Justino, V. (2011). *A distribuição e a expressão gramatical do futuro do conjuntivo no português de Moçambique*. MA thesis. Lisbon: Universidade de Lisboa.

Kloss, H. (1967). Abstand languages and Ausbau languages. *Anthropological linguistics* 9 : 29-41.

Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R. Version March 1996.

McWhinney, B. (1994). *The CHILDES Project. Tools for Analyzing Talk*. Hilsdale, N.J.: Lawrence Erlbaum Associates.

Mendes, A., Hendrickx, I., Généreux, M., Hagemeijer, T. (2022). *Manual for the PALMA corpora on the CQPweb interface, version 1.0*. Lisboa: Centro de Linguística da Universidade de Lisboa.

Mesthrie, R. and Bhatt, R. (2008). *World Englishes. The study of new linguistic varieties*. Cambridge: Cambridge University Press.

Miguel, A. (2019). *Integração morfológica e fonológica de empréstimos lexicais bantos no português oral de Luanda.* Doctoral Dissertation. Lisbon: Universidade de Lisboa.

Müller de Oliveira, G. (2016). The system of national standards and the demolinguistic evolution of Portuguese. In R. Muhr (ed.), *Pluricentric languages and non-dominant varieties worldwide*, vol. II, Peter Lang Verlag, pp. 35-48.

Nascimento, F. (2018). *O sistema vocálico do português de São Tomé e o comportamento das vogais médias em contexto pretônico*. Doctoral dissertation. Rio de Janeiro: Universidade Federal do Rio de Janeiro.

Pereira, R., Hagemeijer, T. and Freitas, M. J. (2018). Consoantes róticas e variação no português de São Tomé. *Revista da Associação Portuguesa de Linguística 4* : 206-224.

Schmidt, T. (2012). EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 236-240. Istanbul, Turkey. ELRA

Stroud, C. and Gonçalves, P. (orgs.) (1997). *Panorama do português oral de Maputo*, Vols. 1 & 2. Maputo: INDE.

Van den Bosch and Daelemans, W. (1999). Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL'99, pages 285-292.

9. Language Resource References

Gonçalves, R., Hagemeijer, T., Alcântara, C., Cornejo, C., Madureira, R., Généreux, M. and Mendes, A. (2021). *PALMA Corpus São Tomé e Príncipe*. Lisboa: Centro de Linguística da Universidade de Lisboa.

Hagemeijer, T., Madureira, R., Cornejo, C., Justino, V., Campos, M., Gonçalves, R., Généreux, M. and Mendes, A. (2021). *PALMA Corpus Moçambique*. Lisboa: Centro de Linguística da Universidade de Lisboa.

Miguel, A., Cornejo, C., Madureira, R., Silva, D., Hagemeijer, T., Gonçalves, R., Généreux, M. and Mendes, A. (2021). *PALMA Corpus Angola*. Lisboa: Centro de Linguística da Universidade de Lisboa.