The Gulf of Guinea Creole Corpora

Tjerk Hagemeijer*, Michel Généreux*’ **, Iris Hendrickx*’ *,**

Amália Mendes*, Abigail Tiny*, Armando Zamora****

*Centro de Linguística da Universidade de Lisboa - Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
** EURAC research - Viale Druso1, 39100 Bolzano, Italy
***Centre for Language Studies, Radboud University Nijmegen - P.O. Box 9103, NL-6500 HD Nijmegen, the Netherlands
****Universidad Nacional de Guinea Ecuatorial - Avenida Hassan II, S/N, 661 Malabo, Equatorial Guinea

t.hagemeijer@clul.ul.pt; michel.genereux@eurac.edu; i.hendrickx@let.ru.nl;
amalia.mendes@clul.ul.pt; abigail.tiny@hotmail.com; zamora.segorbe@gmail.com

Abstract

We present the process of building linguistic corpora of the Portuguese-related Gulf of Guinea creoles, a cluster of four historically related languages: Santome, Angolar, Principense and Fa d’Ambô. We faced the typical difficulties of languages lacking an official status, such as lack of standard spelling, language variation, lack of basic language instruments, and small data sets, which comprise data from the late 19th century to the present. In order to tackle these problems, the compiled written and transcribed spoken data collected during field work trips were adapted to a normalized spelling that was applied to the four languages. For the corpus compilation we followed corpus linguistics standards. We recorded meta data for each file and added morphosyntactic information based on a part-of-speech tag set that was designed to deal with the specificities of these languages. The corpora of three of the four creoles are already available and searchable via an online web interface.

Keywords: Gulf of Guinea creoles, corpus annotation and management, language documentation

1. Introduction

We present the process of building corpora of the Portuguese-related Gulf of Guinea creoles (GGCs), a cluster of four languages currently spoken on four islands in West-Africa: Santome (ST) and Angolar (ANG), spoken on São Tomé, Principense (PR) spoken on Príncipe, and Fa d’Ambô (FA) spoken on Annobón and Bioko.[1]

The islands of S. Tomé and Príncipe form the Democratic Republic of S. Tomé and Príncipe, where Portuguese is the official and nowadays predominant native language. Annobón and Bioko are part of Equatorial Guinea, where Spanish is the main official language, but the dominant languages are Pichi (English-based creole), Fang and Bubi (Bantu languages). The map in Figure 1 locates the Gulf of Guinea islands.

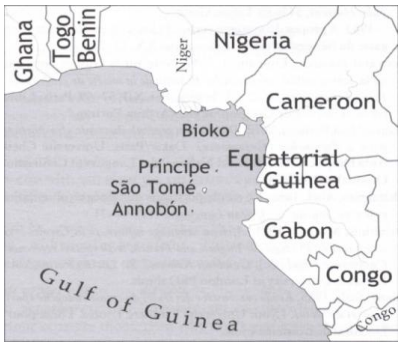

Figure 1. Map of the Gulf of Guinea.

[1] For the language names, we follow Michaelis *et al.* (2013).

Despite the limited mutual intelligibility between the four creole languages, it is usually assumed that they are genetically related, descending from a single creole proto-language that arose on São Tomé in the early 16th century as the result of language contact between Portuguese, the language that donated the large majority of their lexicon, and several African languages from the Benue-Congo family, a branch of the large Niger-Congo phylum (e.g. Ferraz, 1979; Hagemeijer, 2011).

Despite varying degrees of vitality, the GGCs should be considered endangered languages for the following reasons: i) they lack an official status, ii) their speech communities are small and have been gradually abandoning their language due to the presence of other, more widely spoken or more prestigious languages in a multilingual environment, and (iii) due to the absence of active language policies. Michaelis *et al.* (2013) provide the following number of speakers: ANG – 5,000; FA – 4,500~5,000; PR – less than 100; ST – 100,000. The situation of PR is particularly critical because language transmission between generations has been interrupted on a wide scale. The first language studies and written samples of these languages date back to the second half of the 19th century and, on a more positive note, the GGCs have been relatively well documented in academic work from the 1970s on. Apart from ST, the GGCs have hardly been used as a written medium by their speech communities.

The goal of GGC corpora is twofold: i) to carry out (comparative) linguistic research on the GGCs, for instance the reconstruction of lexical and grammatical features of the proto-language and ii) to function as a platform supporting language documentation, planning

and revitalization. Building the ST corpus, for instance, contributed significantly to the publication of the language's dictionary (Araújo & Hagemeijer, 2013).

In the next sections we will first present related corpora for other creole languages in section 2, followed by a more detailed description of the sources of the four Gulf of Guinea creoles in section 3. As these languages have an oral tradition and lacked a standard writing scheme, we needed to deal with spelling variation. In section 4 we describe how we tackled this problem. In section 5 the meta data schema is detailed. As we aimed to add linguistic information to the texts, we annotated all words in texts with part-of-speech (POS) information. For the small corpora this was done manually, but for ST we have more than 200K tokens, and here a part was labeled automatically using a POS-tagger trained on a manually labeled part. We describe this in more detail in section 6 and we conclude in section 7.

2. Review of creole corpora

Some eighty different creole languages are spoken around the world (Lewis 2009), but corpora have only been developed for a small subset. Corpora for Portuguese-related creoles are particularly scarce.

For the French-related creoles, a corpus of 200,000 words of Mauritian Creole is searchable online via a concordance interface as part of the website of the ALLEX project.[2] For Mauritian Creole, a diachronic corpus of 60 texts, of very different length and genre, written between 1721 and 1929, can be found online in a single webpage in html format, with author and date provided for each text (Baker & Fon Sing, 2007).[3] Another project, CREOLORAL, includes 3 hours of spontaneous spoken data of creoles from Martinique and Guadeloupe, translated to French, transcribed and annotated with phonetic and syntactic information in XML format.[4] The corpus is described as freely accessible by contacting the developers. Two Haitian Creole corpora have also been compiled and are freely downloadable: the Haitian Creole spoken and text data, at Carnegie Mellon University, and the Corpus of Northern Haitian Creole, at the Indiana University Creole Institute, which contains ten hours of interviews with 20 Haitians from the Cape Haitian region, with audio and transcriptions.[5, 6] The website of online journal Creolica makes available a corpus of 16 written texts, as well as short stories from Seychelles Creole and a corpus of Reunion Creole, most in pdf format, or in html.[7]

For the English-related creoles, we would like to mention the Corpus of Written British Creole (Sebba, Kedge & Dray, 1999), which counts around 12,000 words. This corpus is available for research purposes and consists of samples from different text genres, being manually annotated with tags that signal lexical, discourse, structure, and grammatical differences between Standard English and the creole. Two other corpora are available in book form: a corpus of Tok Pisin consisting of 1047 folktales that were translated to English (Slone, 2001) and the Corpus of Jamaican E-mail and other CMC (COJEC), a collection of emails and forum messages of about 40,000 words, written by Jamaican students (Hinrichs, 2006).

For Dutch-related creoles we refer to the Negerhollands database (den Besten *et al.*, 1996), a collection of historical texts from the U.S. Virgin Islands. The Surinam Creole Archive (SUCA) (van den Berg & Bruyn, 2008) contains a collection of early creoles in Suriname of about 550,000 words collected from heterogeneous sources.

The Dokumentation Bedrohter Sprachen (DoBeS Archive) contains a corpus of Sri Lanka Malay[8] and several creole corpora are available at the Endangered Languages Archive (ELAR), such as the corpus of Bastimentos Creole English[9] which includes a set of digitally recorded speech acts (audio and video) of the 600 speakers of the Creole community of Bastimentos Island, Panama. ELAR also includes a Portuguese-related creole, Malaccan Portuguese Creole, with video and audio recordings conducted at the Portuguese Settlement in 2011.[10] The audio and video files are paired with time-aligned orthographic transcriptions and English translations. Another resource for Portuguese-based creoles is the lexical database CreolData, under development at Université d'Orléans.

Finally, at the end of 2013 the Max Planck Institute, Leipzig, has published the *Atlas of pidgin and creole language structures online* (APiCS), a database that contains language samples of over 100 structural features of 76 pidgins and creoles.[11]

This review of existing corpora of creoles shows that the compilation varies enormously in terms of design, format and added information: some of the first collections are available as paper publications, or as digitalized document of paper texts, while more recent ones are planned as digital collections. Few creole corpora combine a digital format with a large diversity of genres and a systematic description of metadata, paired with a transcription of spoken data, although more projects of the kind are coming to life. In this respect, our corpora of the GGC are closely related to the methodology followed by corpora such as CREOLORAL or the Negerhollands database.

3. The corpora in numbers

The GGC creole corpora consist of a compilation of oral and written sources. Information related to the size of each of the four corpora can be found in Table 1.

[2] ALLEX project: http://www.edd.uio.no/allex/corpus/africanlang.html.
[3] http://concordancemmc.free.fr/Textes%20anciens.htm
[4] http://ircom.corpus-ir.fr/site/description_projet.php?projet=CREOLORAL
[5] http://www.speech.cs.cmu.edu/haitian/
[6] http://www.indiana.edu/%7Ecreole/
[7] http://www.creolica.net/spip.php?page=corpus
[8] http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI515582%23
[9] http://elar.soas.ac.uk/deposit/0171
[10] http://elar.soas.ac.uk/deposit/0123
[11] http://apics-online.info/

	tokens		types	sentences	texts
	spoken	*written*			
Angolar	10,406	-	992	825	16
Fa d'Ambô	54,197	9,226	3531	3461	29
Principense	14,605	1,303	1,173	1,301	19
Santome	103,651	113,294	6773	18,893	588

Table 1. The Gulf of Guinea creole corpora in numbers.

Table 1 shows that, except in the case of ST, there is a significant imbalance between spoken and written data, which relates to the absence of a written tradition in these languages. The spoken corpora comprise mainly transcriptions of recordings of folk tales and conversations that were recorded with native speakers of the languages. The spoken corpora of Angolar and Principense are based on transcriptions used in four written sources (Günther, 1973; Lorenzino, 1998; Maurer, 1995, 2009). The data in these publications were adapted from the orthographies used by these authors to the standardized writing system used for the corpora (see section 4 below). The texts in Günther (1973) and Lorenzino (1998) were adapted from phonetic transcriptions, whereas the texts in Maurer (1995, 2009) use an orthography that is largely identical to the one we use.

The spoken data of ST are the result of audio recordings of native speakers that were carried out in 1997 and 2001 on several locations in S. Tomé. The recordings were transcribed in 2002 and revised for the project. The spoken data of FA consists of transcriptions of recordings of native speakers from Annobón and Bioko made in 2012.

Following the transcriptions of the spoken published sources of Angolar and Principense, we deliberately left out typical spoken disfluencies like repetitions, hesitations and repair strategies. Relevant variation, however, was kept as much as possible.

The written material includes data from the oldest known written sources, which date back to the last quarter of the 19th century (e.g. Coelho, 1880-1886; Negreiros, 1895; Schuchardt, 1882) until the present and was collected from publications, private sources, such as pamphlets, obtained during field work trips, and from a blog written in ST. The variety of genres is limited and related predominantly to folklore, such as riddles, proverbs, lyrics of songs, and folk tales. The fact that these languages hardly reach into other domains (media, prose, and so forth) also reflects their endangered status.

4. Language standardization

In the absence of an official status, the GGCs also lacked an official orthography, which means that the used orthographies in non-academic work have been highly variable and generally quite inconsistent, ranging from etymological (Romance-based) orthographies to phonological writing systems, a problem that has been acknowledged for creole languages in general (e.g. Sebba, 1998).

To contradict this tendency, in 2009 the Ministry of Education and Culture of S. Tomé and Príncipe invited a team to develop a writing system for the country's creole languages. The result was a phonology-oriented writing proposal – Alfabeto Unificado para as Línguas Nativas de São Tomé e Príncipe (ALUSTP, Pontífice *et al.*, 2009) – which was ratified in 2010, but has not been object of public discussion.[12] The main principle of this proposal is a one-to-one phoneme-grapheme correspondence, but it also deals with word boundaries, cases of contractions, compounds, reduplications, idiophones, etc. All the data of the GGC corpora were adapted to this spelling. The written texts were scanned with OCR software or copied manually and then adapted to the proposed standard in a text editor. The original spelling was not recorded in the manual transcription but can be consulted through the originals, which will be published online as digitalized pdf files, except for those texts that fall under copyright rules (monographs). The names of the pdf files match the file names in the searchable corpus.

The following text in ST, written by composer Gete Rita, illustrates an adaptation of an original text included in the corpus.

USSUA POR SANGAZUZA

O autor Géte Rita

Duentxi sá cama nfélumu
Nê nguê sêbê dê
Non sá ni dgêlu scá guadá
Pa djá nozadu dê
Cúma cu duentxi
Cá iógó ?

Bis

Còro

Kéga cá passa canuá
Ê candá cu xinta nauá
Xi mina tlabá zudá pédê
Sá ê scá guadá liqueza ledá
Bilá pagá massada kiá am
Bamu zunta coplá mindjan
Pá a piá xi duentxi cá iógó.

Bis

Fim

Figure 2. Original song text in ST by Gete Rita.

Adapted text
Dwentxi sa kama nfelumu
Nê ngê sêbê dê
Non sa ni djêlu xka gwada
Pa dja nozadu dê
Kuma ku dwentxi
Ka yogo?
Kega ka pasa kanwa
Ê ka nda ku xinta ni awa
Xi mina tlaba zuda pe dê
Sa ê xka gwada likêza leda
Bila paga masada kia an
Bamu zunta kopla mindjan
Pa a pya xi dwentxi ka yogo.

Table 2. Song text in ST by Gete Rita.

[12] For discussion of ALUSTP, see Araújo (2010).

Based on this sample text, some differences between the conventions in ALUSTP and the earlier original writings can be highlighted:

- The phoneme /k/ is systematically represented as grapheme <k>, whereas the original shows inconsistency (*cu, cá, kéga, copla*, etc.)*;*
- Morpheme boundaries often require readjustments (*candá > ka nda* lit. 'TAM walk'*; nauá > ni awa* lit. 'in water'*; pédê > pe dê* lit. 'father his');
- Significant reduction of accents. Santome exhibits a contrast between open-mid vowels ([ɛ], [ɔ]) and close-mid vowels ([e], [o]), which are frequently and non-systematically marked with acute and circumflex accents in the original version (*kéga, iógó, sêbê*, etc.). In the adapted version, we maintain the circumflex accent for close-mid vowels and use no accent for open-mid vowels (*kega, yogo, sêbê*). Accents on /a/ (<á>), for instance, are redundant because there is no contrasting pair (*sá>sa; scá>xka*, etc.)
- Treatment of nasal vowels underwent systematization (*am>an; mindjan>mindjan*)

Fa d'Ambô posed several challenges, in particular because this creole presents a number of morphophonogical features that show greater divergence from the other three GGCs. The following short poem was published in a traditional poetry bundle by Lêdjam (2008).

Original version	**Adapted version**
Ôxy kê já fó gêêza	*Ôxi k'en xha fô gêêza*
Ken já têndê fá tômbô Llave	*K'en xha tendê fa tombô Llave*
Ken bayà nã dadji	*K'en baya nan dadji*
Menfò tendê fé tômbô Llave	*M na fô tendê fe tombô Llave*
Pá me bá ba kü na dadjif	*Pa m na ba baya ku nan dadji f*
M'sajé tôsê	*M sa khe tôôsê*

Table 3. Poem in FA by N.M. Lêdjam (2008: 30).

Since FA, spoken in Equatorial Guinea, is not contemplated in ALUSTP, but still similar to the other creoles, we made a number of adaptations that reflect the specificity of this language, such as the use of <kh> to represent fricative velar /x/ (*já>xha; sajé>sa khe*), where the original <j> of course reflects the Spanish orthography. In the other GGCs, <j> was already representing postalveolar fricative /ʒ/. Several intricate morphophonological processes of this language were treated as well, such as the contraction involving pronouns and other functional material (e.g. *ken>k'en*, derived from *ku m* '…that I…'; *menfò>m na fô* lit. 'I not can'). Morpheme boundaries often had to be redefined (e.g. *menfò* above; *dadjif> dadji f* lit. 'age not').

In general, these inconsistencies in the original texts do not only vary significantly within writings of the same author, as shown above, but also from author to author. Adapting all the different original orthographies represented a heavy workload, to which can be added that some of the original texts (the majority of which written on typewriters) were in bad state of conservation.

Instances of language variation were maintained as much as possible, in particular in the spoken corpora. For instance, there are cases of variation between postalveolar fricative /ʒ/ and affricate /d͡ʒ/, as in *dja~ja* 'day and *mindjan~minjan* 'remedy' or between alveolar fricative /s/ and postalveolar fricative /ʃ/, as in progressive aspect marker *ska~xka*. Note that we still lack studies on what factors motivate different types of variation in these creoles (e.g. linguistic context, geography, age, etc.). While variation is detectable in the spoken data and can be transcribed accordingly, the written corpus is of course less reliable, because the underlying phonetic realization of written forms is often not crystal-clear. This can be illustrated by the ST progressive aspect marker *ska~xka* mentioned above. This morpheme is generally written *scá* and it is impossible to know from the written data which spoken variant (/ska/ or /ʃka/) underlies this form. The maintenance of cases of variation was established in ALUSTP because of its usefulness in the discussion on additional standardization of these languages

Although the writing system used for the corpora generally diverges from those that have been used for these languages, a similar system has been used by some Santomean authors (e.g. Daio, 2002) and in academic publications. From our own experience with native speakers, readability of the languages is fully ensured by the standardized texts.

5. Meta data

The format of the corpora follows the general norms for corpus linguistics (e.g. Wynne, 2005) and uses UTF-8 character encoding and XML annotation for the meta data. We encoded the meta data of the corpora texts, like author and date, in a simple XML format that is compatible with the P5 guidelines of Text Encoding Initiative (TEI consortium, 2007), using the following XML meta data tags:

- language: one of the four GGCs
- type: spoken or written
- title: the title of the text (if any)
- author: the author of the text (if known)
- date: the date of publication or recording
- period: the periods in which the texts fall
- source: book, newspaper article, (cultural) magazine, pamphlets, unknown.
- genre: prose, conversations, poetry, proverbs, riddles, song texts, mixed, other
- age: the age of the recorded speaker (spoken corpora)
- place of recording: the place of recording (spoken corpora)
- notes: any type of additional information, such as the name of publisher and the place of publication.

In light of the predominantly folklore-related materials that were obtained, we did not follow text typology recommendations used for large corpora. While most tags are self-explaining, a short note is in place for 'genre'. The classification in genres relates to the amount of data that was available for each genre but without establishing a division that would be too fine-grained for the size of the data set. Different genres may be useful for different linguistic purposes: larger portions of text, such

as folk stories and conversations often reveal different linguistic properties than, for instance, proverbs and riddles. "Mixed" genre includes publications – in particular cultural magazines – with different types of texts that belong to one of the other genres. In these cases the main header receives the label "mixed", but we applied sub headers in line with the TEI guidelines[13] to tell apart genres in the text. This strategy was also adopted for other changes in the header data, for instance a change of authors within a collection of poetry[14].

6. POS annotation

A tag set of 35 POS-tags was prepared based on the data and on our knowledge of the languages. The tag set is based on the guidelines by Leech & Wilson (1996) and on the CINTIL tag set that was developed for the Portuguese CINTIL corpus (Barreto *et al*., 2006). Adaptation of the grammatical categories was crucial, because of the substantial typological differences between the GGCs and Portuguese. The POS-tag set below is an updated version of the one described in Hagemeijer *et al.*(2012) for ST, which was the first and largest corpus to be annotated, and has been successfully applied to the other GGCs, requiring no further adaptations.

Tag	**Category**	**ST examples**
ADJ	Adjectives	*glavi* 'pretty', *vlêmê* 'red'
ADV	Adverbs	*oze* 'today', *yôxi* 'yes'
ART	Articles	*ũa* 'a(n)', *inen* 'the'
CN	Common Nouns	*mosu* 'boy', *ope* 'foot, leg'
COMP	Complementizers	*kuma, ku, pa* 'that'
CONX	Conjunctions	*maji* 'but', *punda* 'because'
DEM	Demonstratives	*se* 'this, that', *xi* 'that'
DGT	Digits	*0, 1, 42, 12345, 67890*
FOC	Focus markers	*so, soku*
FW	Foreign words	mostly Portuguese and Spanish vocabulary
ID	Ideophones	*liku sonosono* 'very rich' (lit. rich+ID)
INDF	Indefinites	*nadaxi* 'nothing'
INT	Interrogatives	*kuma* 'how', *andji* 'where'
IPS	Incomprehensible sequences	
ITJ	Interjection	*kaka!* (surprise)
ME	tag for *me* 'even, -self, etc. in ST and corresponding forms in the other GGCs – *Zon me* 'even Zon'	
MOD	Modality Markers	*sela* 'must'
NEG	Negation markers	*na, naxi, nantan,fa, fô*
NUM	Numerals	*dôsu* 'two', *tlêxi* 'three'
ON	Onomatopees	
OTLO	tag for *otlo* '(an)other' in ST and corresponding forms in the other GGCs – *ôtlô ngê* 'another person'	
PM	Presentational marker	*ya* 'there … is'
PNM	Part of Name	*Zon* 'John', *Maya* 'Maria'
PNT	Punctuation Marks	*., ?, (, ...*
POSS	Possessives	*mu* 'my', *dê* 'his, her, its'
PREP	Prepositions	*antê* 'until', *ku* 'with', *di* 'of'
PRS	Personals	*n* 'I', *ê* 's/he, it', *non* 'we'
PRT	Discourse Particles	*an, ê, en, fan, ô*
QNT	Quantifiers	*kada* 'every', *tudu* 'all'
RED:xx	Reduplicated Categories	*dôsu-dôsu* 'in groups of two' (RED:NUM, lit. 'two-two')
REFL	Reflexives	*mu, bô, dê, non, ...*
RV	Residual Value	abbreviations, acronyms, etc.
STT	Social Titles	*sun* 'Mr.', *san* 'Mrs.'
TAM	Tense-Aspect-Mood markers	*ka, xka, tava*
V	Verbs	*fla* 'to speak', *mêsê* 'to want'

Table 4. POS tag set for the GGC corpora.

The GGCs lack inflectional morphology but exhibit, for instance, preverbal tense-mood-aspect markers, productive reduplication of many word categories, idiophones and clause-final discourse particles. The following examples illustrate a few features of the GGCs.

(1) *Am na thaka be wa ê ,*
PRS NEG TAM V NEG PRT PNT
punda n fô mionga welewele si e .
CONX PRS V CN ADV DEM DEM PNT
'I'm not going, because I've just come back from the sea.'
(Angolar, Maurer 1995: 191)

(2) *Ũa mosu se pletu lululu ku kaza*
ART CN DEM ADJ ID COMP V
ku mina di men mu .
PREP CN PREP CN POSS PNT
'A very black boy who married to my mother's daughter.'
(Santome)

(3) *Kêtê-kêtê ki n tê , n tolo*
RED:CN COMP PRS V PNT PRS V
da ningê tudu .
PREP CN QNT PNT
'What little I have, I have shared with all of you.'
(Principense, Maurer 2009: 185)

Due to the small size of the ANG and PR corpora, they have been fully manually annotated and revised. Only a small subpart (5,500 tokens) of the FA corpus has been annotated at this point. For the much larger corpus of ST, a data set of approximately 17,000 tokens was manually annotated. This training set was used to train and evaluate the automatic POS-tagging software. We

13 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html #DSDI

14 Note that such XML file with mixed genres was split in separate files in the CQPweb interface to accommodate search on genre type.

examined the performance of two different software systems that are developed for POS-tagging, namely the Memory-Based Tagger (MBT) by Daelemans *et al*. (2010) and the SVMTool by Giménez & Màrquez (2004). Both software programs are off-the-shelf tagger-generators and can be trained and tuned on the task by giving it training material. To tune the parameters of the programs for this particular data set, we split the data in 10 parts and performed tenfold cross-validation experiments (90% train, 10% test) computing the average accuracy over ten folds. We experimented with different settings to find optimal one. After experimenting different parameters, the SVMTool yielded a slightly higher accuracy (87,6%) and was used to tag the remainder of the ST corpus. We suspect that the data set size is too small to gain much from tuning as it will easily lead to overestimations.

7. The corpora on CQPweb

The corpora of ST, ANG and PR have been made available for concordances in CQPweb (Hardie, 2012)[15], an online interface that allows users to search for concordances of word forms, sequences of words and POS categories. The platform also allows users to create frequency lists and to restrict the search query to specific text types. Users can restrict their search for genre (ST, PR and ANG), period (ST and PR), recording place (ST), source (ST) and type (ST and PR). CQPweb also offers the possibility to make sophisticated search patterns using the Simple query language syntax (Hoffman *et al*., 2008). Furthermore, the CQPweb interface offers powerful statistical analysis, which is very helpful to find collocations or keywords: given the size of these corpora, this should be particularly well-suited in the case of ST. Failing license agreements for our corpora to give the public full access, the CQPweb platform should offer the user a wide range of options to explore the intricacies of the GGC.

8. Concluding remarks

We presented the process of building a unique resource based on written and spoken data of the four GGCs, languages spoken by small island communities in West Africa. It was shown how we dealt with language standardization and annotation issues. The corpora were made searchable on CQPweb and are intended for both research purposes and tasks related to language planning. The documentation of the corpora with digitalized written texts in pdf format constitute an added value, in particular for speakers of these languages with little access to these often rare materials.

9. Acknowledgements

The Gulf of Guinea corpora are funded by the Portuguese Foundation of Science and Technology (FCT) as part of the project *The origins and development of creole societies in the Gulf of Guinea: An interdisciplinary study* (PTDC/CLE-LIN/111494/2009). This work would not have been possible without all the native speakers of these languages who kindly participated in the fieldwork: *Dêsu paga nansê da non* (ST). Finally, we would like to acknowledge the precious comments made by three anonymous reviewers.

[15]GGC are online available at:

http://alfclul.clul.ul.pt/CQPweb/Angolar/

http://alfclul.clul.ul.pt/CQPweb/Santome/

http://alfclul.clul.ul.pt/CQPweb/Principense/

10. References

Araújo, G. (2010). Relações entre as fonologias das línguas crioulas de STP e a 'proposta ortográfica' ALUSTP. In: *7º Congresso Ibérico de Estudos Africanos, 9: 50 anos das independências africanas: desafios para a modernidade*. Lisboa: CEA. https://repositorio.iscte-iul.pt/handle/10071/2349?mode=full

Aráujo, G. & Hagemeijer, T. (2013). *Dicionário livre do santome-português*. São Paulo: Hedra.

Baker, P. and Fon Sing, G. (2007) *The making of Mauritian Creole. Analyses diachroniques à partir des textes anciens*. Westminster Creolistics series, 9. Battlebridge, London, UK.

Barreto F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F. P., Nunes, F. and Silva, J. (2006). Open resources and tools for the shallow processing of Portuguese. In *Proceedings of LREC* 2006. Genoa, Italy, 1438-1443.

van den Berg, M., & Bruyn, A. (2008). The early Surinamese creoles in the Suriname Creole Archive (SUCA). *Linguistics in the Netherlands*, *25*(1), 25-36.

den Besten, J. B., van Rossem, C., & van der Voort, H. G. A. (1996). Linguistic annotation of the creole texts. *Die creol taal. 250 years of Negerhollands texts*. Amsterdam: Amsterdam University Press, pp. 49-280.

Coelho, A. (1880-1886). Os dialectos românicos ou neo-latinos na África, Ásia e América. In Morais Barbosa, Jorge (ed.) 1967, *Crioulos*. Lisboa: Academia Internacional de Cultura Portuguesa.

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

Daio, O. (2002). *Semplu*. S. Tomé: Edições Gesmédia.Ferraz, L. (1979). *The creole of São Tomé*. Johannesburg: Witwatersrand University Press.

Ferraz, L. 1979. *The creole of São Tomé*. Johannesburg: Witwatersrand University Press.

Giménez, J. and Marquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal, pp. 43-46.

Günther, W. (1973). *Das Portugiesische Kreolisch der ilha do Príncipe*. Marburg an der Lahn.

Hagemeijer, T. (2011). The Gulf of Guinea creoles: genetic and typological relations. *Journal of Pidgin and Creole Languages*, 26:1, pp. 111-154.

Hagemeijer, T., Hendrickx, I., Haldane, A., Tiny, A. (2012). A Corpus of Santome. In *Proceedings of the SALTMIL-AfLaT workshop*. Istanbul, Turkey, 2012. European Language Resources Association (ELRA), pp. 61-66.

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

Hinrichs, L. (2006). *Codeswitching on the web: English*

and Jamaican Creole in e-mail communication. Pragmatics and Beyond New Series Vol. 147. Amsterdam: John Benjamins Publishing.
Hoffmann, S., Evert, S., Smith, N., Lee, D., & Berglund-Prytz, Y. (2008). *Corpus linguistics with BNCweb-a practical guide* (Vol. 6). Peter Lang.
Lêdjam, N.-M. (2008). *Cancionero oral annobonés*. Barcelona: Ceiba.
Leech, G. and A. Wilson (1996). EAGLES. *Expert Advisory Group on Language Engineering Standards. Recommendations for the Morphosyntactic Annotation of Corpora.* EAGLES Document EAG-TCWG-MAC/R. Version March 1996
Lewis, M. P. (2009). *Ethnologue: Languages of the world, 16th edition.* Dallas: SIL International.
Lorenzino, G. 1998. *The Angolar creole Portuguese of São Tomé: Its grammar and sociolinguistic history*. City University of New York: Ph.D. Dissertation.
Maurer, P. (1995). *L'Angolar: un créole afro-portugaise parlé à São Tomé*. Hamburgo: Helmut Buske Verlag.
Maurer, P. (2009). *Principense – Grammar, texts, and vocabulary of the Afro-Portuguese creole of the island of Príncipe*. Londres: Battlebridge Publications.
Michaelis, S, Maurer, P., Haspelmath, M. & Huber, M. (2013). *The survey of pidgin and creole languages, Vol. I:* Portuguese-based, Spanish-based and French-based Languages. Oxford: Oxford University Press.
Negreiros, A. (1895). *Historia Ethnographica da ilha de S. Tomé*. Lisboa: José Bastos.
Pontífice, J. *et al.* (2009). *Alfabeto Unificada para as Línguas Nativas de S. Tomé e Príncipe (ALUSTP)*. São Tomé.
Schuchardt, H. (1882). Kreolische Studien I. Über das Negerportugiesische von S. Thomé. *Sitzungsberichte der kaiserlichen Akademie des Wissenschaften zu Wien*. 101, II, pp. 889-917.
Sebba, M. (1998). Phonology meets ideology: the meaning of orthographic practices in British Creole. *Language Problems and Language Planning* 22(1), pp. 19-47.
Sebba, M., Kedge, S.; Dray, S. (1999). The corpus of written British Creole: A user's guide. http://www.ling.lancs.ac.uk/staff/mark/cwbc/cwbcman.htm (Date of access: 2014-12-3)
Slone, T.H. (2001). *One thousand one Papua New Guinean nights: Folktales from Wantok Newspapers*: Volume 1, Tales from 1972-1985 and Volume 2, Tales from 1986-1997 (Papua New Guinea Folklore Series). Oakland, CA: Masalai Press.
TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.6. [2014-01-20]. http://www.tei-c.org/Guidelines/P5/ (Date of access: 2014-12-3) .
Wynne, M. (2005). *Developing Linguistic Corpora*: a Guide to Good Practice. Oxford: Oxbow Books.