A Corpus of Santome

Tjerk Hagemeijer, Iris Hendrickx, Haldane Amaro, Abigail Tiny
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
t.hagemeijer@clul.ul.pt; iris@clul.ul.pt; amaro25@hotmail.com; abigail.tiny@hotmail.com

Abstract

We present the process of constructing a corpus of spoken and written material for Santome, a Portuguese-related creole language spoken on the island of S. Tomé in the Gulf of Guinea (Africa). Since the language lacks an official status, we faced the typical difficulties, such as language variation, lack of standard spelling, lack of basic language instruments, and only a limited data set. The corpus comprises data from the second half of the 19th century until the present. For the corpus compilation we followed corpus linguistics standards and used UTF-8 character encoding and XML to encode meta information. We discuss how we normalized all material to one spelling, how we dealt with cases of language variation, and what type of meta data is used. We also present a POS-tag set developed for the Santome language that will be used to annotate the data with linguistic information.

Keywords: Santome, Creole, S. Tomé and Príncipe, Corpus, Standardization, Annotation

1. Introduction

Santome (São-Tomense, Forro) is a Portuguese-related creole language spoken on the island of S. Tomé in the Gulf of Guinea (Africa). The language has no official status, but according to the 2001 census 72,4% of the population over 5 years old spoke Santome (as a first or second language), on a total population of 137,599 (RPGH – 2001, 2003). For Portuguese, the official and most widely spoken language on the island, this percentage was 98,9%, which shows that there is a high degree of bilingualism. Nevertheless, Santomeans, in particular younger generations, have been shifting towards Portuguese. The result is a gradual loss of diglossia and language attrition of the creole.

Together with Ngola (Angolar), Fa d'Ambô (Annobonense) and Lung'Ie (Principense), which have much smaller numbers of native speakers, Santome is the direct descendant of the proto-creole of the Gulf of Guinea that came into being in the 16th century, on the island of S. Tomé, as the result of language contact between Portuguese, the lexifier language, and several Benue-Congo languages, in particular Edo (Edoid) and Kikongo (Bantu) (Ferraz, 1979; Hagemeijer, 2011). The first concise language studies with samples of Santome date back to the second half of the 19th century (e.g. Schuchardt, 1882; Negreiros, 1895). Since the monograph of Ferraz (1979), the number of studies on this creole has increased significantly. Yet, it still lacks basic language instruments, such as a reference grammar and a dictionary.

The electronic Santome corpus is being built as part of the project *The origins and development of creole societies in the Gulf of Guinea: An interdisciplinary study*. (cf. section 8). Within this project, the corpus under construction will be used primarily for linguistic purposes, in particular data extraction and comparison with the other three Gulf of Guinea creoles mentioned above, in order to reconstruct properties of the proto-creole of the Gulf of Guinea. In addition, the corpus has potential for tasks related to language planning in S. Tomé and Príncipe, such as the development of dictionaries and text materials. To our best knowledge, no electronic corpora of the type presented here have yet been built for Portuguese-related creole languages. Despite a few exceptions mentioned in the next paragraph, this also applies to creole languages with other lexifiers. In part, this may be related to the fact that many creole languages are small minority languages lacking an official orthography or even a writing tradition altogether. Consequently, corpus building can be costly and labor intensive: it requires fieldwork trips for data collection, transcribing, revising, standardizing, and so on.

We would like to mention a few corpora we found for other creole languages. The Corpus of Written British Creole (Sebba, Kedge & Dray 1999) counts around 12,000 words of written English-based Caribbean Creole. The corpus consists of samples from different text genres and is manually annotated with tags that signal lexical, discourse, structure, and grammatical differences between Standard English and the creole. The corpus is available for research purposes. A corpus of 200.000 words of Mauritian creole, a French-based creole language, is available online and searchable via a concordance interface as part of the website of the ALLEX project[1]. There is also a corpus of Tok Pisin (English-related with a Melanesian substrate) consisting of 1047 folktales that were translated to English and published in book form (Slone, 2001). Furthermore, the new digital era offers new

1 ALLEX project
http://www.edd.uio.no/allex/corpus/africanlang.html.

possibilities to gather corpus data for certain creole languages. An example is the COJEC corpus of Jamaican creole, a collection of emails and forum messages of about 40,000 words, written by Jamaican students (Hinrichs, 2006)[2].

2. Corpus

The Santome corpus consists of a compilation of oral and written sources. Since the second half of the 19th century, the language has been written in published and non-published sources, but as it is not an official language, not much material has been produced. The 19th century language samples consist of a few poems by Francisco Stockler (collected from different sources), and language fragments collected in Adolfo Coelho (1880-1886), Schuchardt (1882) and Negreiros (1895). The published sources further include a few newspaper articles from the 1920s and a small number of books and magazines written after the country's independence (1975). The books and cultural magazines typically intersperse Santome and Portuguese texts and some of the Santome texts come with a Portuguese translation. The unpublished sources comprise a number of pamphlets from the 1940s or early 1950s obtained from private sources and many unidentified texts (mostly song texts) collected in the Historical Archive of S. Tomé and Príncipe. Apart from a few known sources that we were unable to locate so far, we believe that we have gathered a significant amount of the existing written materials. Many texts that have been produced can be placed in the domain of folklore (folk tales, proverbs, riddles, etc.). The fact that the language does not reach into other functional spheres typically associated with prestige, such as journalism and education, is one of the reasons why the language is not thriving and may quickly become more severely endangered. Finally, it should be noted that the number of authors that produced the texts is relatively restricted. The author of the newspaper articles from the 1920s is also the one who wrote the pamphlets, many of the song texts were written by a small number of song writers and most of the proverbs were collected by a single author. We also encountered one source of the new media, a blog written in Santome. Blogs are an interesting text genre with an informal writing style and can be seen an online diary expressing the personal opinions of the blog's author. The written subcorpus has currently 99,658 words.

The spoken corpus comprises transcriptions of recordings of predominantly folk tales told by story tellers and conversations and songs that were recorded in 1997 and 2001 with native speakers of the language from different locations in S. Tomé. This subcorpus has currently 52 transcribed recordings of 20 different speakers who produced a total of 84,951 words. The spoken recordings have been freely transcribed in the sense that we have tried to match written text as much as possible. Many of the typical oral phenomena, like fragmented words, extra- linguistic sounds, hesitations, repetitions and linguistic repairs, were not transcribed because we aimed to keep the texts fluently. This type of free transcription can be seen as an additional normalization step of the spoken material. Additional details on the written and spoken subcorpus can be found in section 4.

Since we do not have copyrights for all the materials used in the corpus, we cannot make the corpus freely available at this point. We plan to remedy this problem by making the corpus available for concordances in an online interface, CQPweb (Hardie, forthc.)[3], that allows users to search for concordances of word forms, sequences of words and POS categories. The platform will also allow users to create frequency lists and to restrict the search query to specific text types.

3. Language standardization

Since Santome has no official status, the (Romance-based) orthographies have been highly variable and often quite inconsistent, ranging from etymological orthographies to phonological writing systems, a well known problem for creole languages in general (e.g. Sebba, 1996). A word like [kwa] 'thing', for instance, has been written in the following ways: *cua, cuá, qua, quá, kua, kuá, kwa, kwá*. Many texts show an unnecessary proliferation of accents and irregular morpheme separation of function words (aspect markers and negation markers, for instance) and lexical items. The explanation for the popularity of etymological orthographies, i.e Portuguese-oriented orthographies, can be assigned to the fact that probably over 90% of the creole's lexicon is drawn from Portuguese, the official language. However, Portuguese etymons underwent significant phonological changes when they were historically incorporated in the creole (Ferraz, 1979) and a considerable number of etymologies are unknown or traceable to the African source languages. This has led us to adopt ALUSTP, the 2009 phonology-oriented writing proposal that was ratified in 2010 by the Ministry of Education of Culture of S. Tomé and Príncipe (Pontífice *et al.*, 2009).

(Original)	**(Adapted)**
Inen piscadô nón	*Inen pixkadô non*
di tudu bóca plé	*di tudu boka ple*
di téla cé non glavi ximentxi	*di tela se non glavi ximentxi*
cá chê ni ké d'inen	*ka xê ni ke dinen*
cu amuelê cu buá vonté	*ku amwêlê ku bwa vonte*
chê bá nótxi	*xê ba nôtxi*
chê bá Tlachia	*xê ba Tlaxa*
basta p'inen bála blé d'omali	*baxta pa inen ba ala ble d'omali*
bá bucá vadô panhã cé	*ba buka vadô panha se*

Table 1. Excerpt from Quinta da Graça (1989) in the original and adapted version.

2 The emails of COJEC are published in the appendix of the book.

3 CQPweb: http://cqpweb.lancs.ac.uk/

The main principle of this proposal is a one-to-one phoneme-grapheme correspondence. We decided to standardize all material in the corpus to this spelling. The excerpt from a poem written by Quintas da Graça (1989) in Table 1 above illustrates the original writing system and the system used for the corpus.In the original text, [ʃ] is represented by the following graphemes: 's' (*piskadô*), 'x' (*ximentxi*) and 'ch' (*chê*). In the adapted version 'x' occurs in all the contexts. The sound [k] is represented by 'k' (*ké*) and 'c' (*bóca*) in the original text and becomes 'k' in the orthography we use. Santome exhibits a contrast between open-mid vowels ([ɛ], [ɔ]) and close-mid vowels ([e], [o]), which are respectively marked with acute and circumflex accents in the original version. In the adapted version, we maintain the circumflex accent for close-mid vowels and use no accent for open-mid vowels. In the case of vowel [a], accents are redundant altogether, because there is no contrasting pair. An example of morpheme separation follows from the form *p'inen* and *bála*, which become *pa inen* (lit. for they) and *ba ala* (lit. 'go there').

It follows from these examples that adapting all the different original orthographies represents a heavy workload. All the texts were scanned with OCR software or copied manually and then adapted to the proposed standard in a text editor. The original texts are all typewritten (the majority on typewriters) but sometimes in bad state of conservation. Instances of language variation (e.g. *djêlu* ~ *jêlu* 'money'; *idligu* ~ *igligu* 'smoke') were maintained as much as possible, in particular in the spoken corpus. With respect to variation, the written corpus is of course less reliable, because it is not always crystal-clear what variant underlies a given written form. By including variation, the corpus will also be useful to analyze quantitative and regional variation, which can then be used in language planning. The corpus and the variation found therein is also being used in a forthcoming Santome dictionary with over 4,000 lexical entries (Araújo & Hagemeijer, in preparation).

4. Meta data

The format of the corpus 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 decided to encode the meta data about the corpus texts like author and date in a simple XML format that is compatible with the P5 guidelines of Text Encoding Initiative (TEI consortium, 2007). Next, a brief explanation of the XML meta data tags is provided.

- language: In addition to Santome, the project will build a corpus for the other three Gulf of Guinea creoles (cf. introduction). Note, however, that the corpora of these three languages will be much smaller and mostly restricted to spoken data due to the absence of a writing tradition
- corpus: Spoken or written
- title: The title of the text (if any)
- author: The author of the text (if known)
- age: the age of the recorded speaker (spoken data)
- place of recording: geographical location of the recording (spoken data)
- date: The date of publication (if any), which can be exact or approximate. Unless we found evidence to the contrary, we assumed that publication dates are close to the date of writing.
- source: We use the following list of sources: book, newspaper article, (cultural) magazines, pamphlets, online, unknown.
- genre: We use the following list of genres for the written corpus: prose, poetry, proverbs, riddles, song texts, mixed, other. For the spoken corpus, there are three genres, namely prose (folk tales and other stories), music and conversations.
- notes: Tag reserved for any type of additional information, such as the name of publisher and the place of publication.

While some of the tags speak for themselves, a few notes are in place here, particularly with respect to the typology of genre. In light of the predominantly folklore-related materials that were obtained, we did not follow text typology recommendations used for large corpora. Since the main goal of the corpus within the project concerns linguistic analysis, the different genres can serve different purposes. Most importantly, prose should be set apart from the other genres. The narratives in the corpus including folk tales and (personal) stories, as well as the blog, are the best means for investigating specific linguistic topics that require larger portions of text (e.g. clause-linking or anaphoric relations). In proverbs, riddles or poetry, on the other hand, one might find archaic lexicon or structures that are less likely to be found in prose. Another criterion underlying the classification in genres relates to the amount of data that was available for each genre. A more fine-grained division would have lead to genres with smaller amounts of material. In Table 2 we present how the number of files and words is divided over the genres.

written subcorpus		
genre	**files**	**words**
mixed	**10**	**22.652**
music (song texts)	**169**	**21.081**
poetry	**11**	**4.442**
prose	**59**	**40.364**
proverbs	**3**	**9.081**
other	**4**	**1.936**
subtotal	**257**	**99.658**
spoken subcorpus		
conversation	7	20.945
prose	43	62.844
music	2	802
subtotal	**52**	**84.591**
TOTAL	**309**	**184.249**

Table 2. Distribution of files and words across the different genres in the Santome corpus

The high number of files in the category "music" derives from the fact that we are dealing with unpublished sources, often a song text on a sheet of paper. Many of the proverbs, on the other hand, were published in a single volume (Daio, 2002). Finally, the "mixed" genre includes publications – in particular cultural magazines – with different types of texts that belong to one of the other five genres. In these cases the main header receives the label "mixed", but we applied subheaders in line with the TEI guidelines[4] to distinguish between genres in the text, for instance … . This strategy was also adopted for other changes in the header data, for instance a change of authors within a collection of poetry. For the spoken part, we only have material of three different text genres. The largest part is made up of told stories (prose).

5. POS annotation

Once a few minor issues related to the uniformization of the data and the headers are settled, we plan to start the enrichment of the corpus with linguistic annotation, namely part-of-speech (POS) tagging. The following tag set has already been prepared based on a small subset of the data and on our knowledge of the language. It still needs testing on a larger data set. The tag set is based on the guidelines by Leech & Wilson (1996) and on the CINTIL tag set that was developed for Portuguese CINTIL corpus (Barreto *et al*., 2006). The adaptation of the grammatical categories was crucial, because Santome is typologically very different from Portuguese and shows greater resemblance to certain West-African languages, such as Edo, its main substrate language (Hagemeijer, 2011; Hagemeijer & Ogie, 2011) or languages from the Kwa cluster (e.g. Aboh, 2004).

Santome is a strongly isolating language without any inflectional morphology and only two productive derivational morphemes. Reduplication and compounding, however, are productive morphological strategies. For reduplicated categories we propose RED: followed by the label of the category that is being reduplicated. Numerals, for instance, can be fully or partially reduplicated (RED:NUM).

(1) *tlêxi-tlêxi* 'in groups of three'
(2) *tlê-tlêxi* 'all three'

In addition to more standard tags, we propose a number of tags that are highly language specific. Ideophones are a special word category consisting of modifiers with specific phonological properties that normally occur with a unique lexical item (nouns, verbs, adjectives).

(3) *kabêsa wôlôwôlô* 'foolish person' (lit. head+id.)
(4) *sola potopoto* 'cry intensely' (lit. cry+id.)
(5) *vlêmê bababa* 'intensely red' (lit. red+id)

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

Tag	Category	Examples
ADJ	Adjectives	*glavi* 'pretty', *vlêmê* 'red'
ADV	Adverbs	*oze* 'today', *yôxi* 'yes'
ART	Articles	*ũa* 'a(n)'
CJ	Conjunctions	*maji* 'but', *punda* 'because'
CN	Common Nouns	*mosu* 'boy', *ope* 'foot, leg'
COMP	Complementizers	*kuma* 'that'
DGT	Digits	*0, 1, 42, 12345, 67890*
DEM	Demonstratives	*se* 'this, that', *xi* 'that'
EXC	Exclamatives	*kê* 'what'
FOC	Focus markers	*so, soku*
FW	Foreign words	mostly Portuguese words
ID	Ideophones	*sũũũ* (*pya sũũũ* 'stare at', lit. look+ID)
INT	Interrogatives	*kuma* 'how', *andji* 'where'
ITJ	Interjection	*kaka!* (surprise)
MOD	Modality Markers	*sela* 'must'
NEG	Negation markers	*na, fa, fô*
NUM	Numerals	*dôsu* 'two', *tlêxi* 'three'
PP	Participles	*bixidu* 'dressed', vadu 'split'
PM	Presentational marker	*avia* 'there was'
PNM	Part of Name	*Zon* 'John'
PNT	Punctuation Marks	*., ?, (, ...*
POSS	Possessives	*mu* 'my', *bô* 'your'
PREP	Prepositions	*antê* 'until', *ku* 'with'
PREP:NOM	Nominal prepositions	*basu* 'under(neath)', *wê* 'in front of'
PRS	Personals	*n* 'I', *ê* 's/he, it'
PRT	Particles	*an* (interrogative particle)
QNT	Quantifiers	*kada* 'every', *tudu* 'all'
RED:xx	Reduplicated Categories	*kume-kume* 'keep eating' (RED:V)
REFL	Reflexives	*mu, bô, dê, non, ...*
RV	Residual Value	abbreviations, acronyms, etc.
SPV	Special Verbs	*loja* 'to encircle, around', *pê* 'to put, in'
STT	Social Titles	*sun* 'Mr.', *san* 'Mrs.'
TAM	Tense-Aspect-Mood markers	*ka, xka, tava, ta.*
V	Verbs	*fla* 'to speak', *mêsê* 'to want'

Table 3. POS-Tag set for the Santome corpus annotation.

The language further exhibits grammaticalized preverbal Tense-Mood-Aspect morphemes (TAM), a number of grammaticalized modality markers (MOD) that typically occur clause-initially and discourse-related particles (PRT). Some of these morphemes are illustrated in the following sentence.

(6) *Ola nansê* ***ka*** ***ska*** *nda ku amigu,*
when 2PL TAM TAM walk with friend

sela *nansê toma kwidadu* ***ê.***
MOD 2PL take care PRT
'When you are hanging out with friends, you must be careful.

Another domain that requires special attention is adpositions. Like many West-African languages (e.g. Kwa languages), the prepositional function in Santome can be expressed by prepositions (7), nouns (8) and (defective) verbs (9), for which we will respectively use the following tags: PREP, PREP:NOM and SPV. The tag SPV makes it possible to distinguish between the use of certain verbs as main verbs (V) and verbs in the second position in serial verb constructions.

(7) *Ê xê* ***ni*** *ke.*
'S/he leave PREP.LOC house
'S/he left home.'
(8) *Ê sa* ***wê*** *ke*
3SG be eye house
''S/he's in front of the house
(9) *Ê saya kanwa* ***pê*** *ple.*
3SG pull canoe put beach
'S/he pulled the canoe on the beach.'

For the POS annotation we will use an automatic POS-tagger trained on a manually annotated sample both to support the annotation process and to tag the rest of the corpus automatically. For the annotation, the tagger can speed up the POS tagging as manual verification is faster than annotating every word manually. We aim to have a sample of 50K words of the corpus manually verified and the other part will be tagged automatically.

6. Final remarks

We presented the construction and annotation of a corpus of Santome. The process of resource creation for a creole language like Santome has to deal with problematic issues like lexical language variety both in spoken and written material, a small body of written material with spelling variance, lack of standardized resources, such as a standard spelling, dictionary or grammatical reference. We aimed to address these issues by 1) collecting all written material that we could find into one uniformly encoded corpus 2) adding meta data information 3) standardizing the spelling of the written material to one systematic spelling 4) transcribing spoken material in the same spelling format 5) development of a POS-tagset for Santome. We expect the corpus to be a useful resource to establish the degree of relatedness between the four Gulf of Guinea creoles and a tool in language maintenance and revitalization, partly through the development of other language resources. Despite the fact that creole languages constitute different genetic units and not a single language family, it is often highlighted that they share certain linguistic (typological) properties. Therefore we believe that a more widespread corpus-based approach to these languages will endow comparative research on creoles with tools that allow for investigating these claims based on larger amounts of data.

7. Acknowledgements

The Santome corpus is 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) and the FCT program Ciência 2007/2008 (Iris Hendrickx).

8. References

Aboh, E. (2004). *The morphosyntax of complement-head sequences: Clause structure and word order patterns in Kwa*. Oxford: Oxford University Press.

Aráujo, G. & Hagemeijer, T. (in preparation). *Dicionário santome-português / português-santome*. São Paulo: Hedra.

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 the 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy.

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

Daio, O. (2002). *Semplu*. S. Tomé: Edições Gesmédia.

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

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. & Ogie, O. (2011). Edo influence on Santome: evidence from verb serialization and beyond. In Claire Lefebvre (ed.), *Creoles, their substrates, and language typology*. Amsterdam, Philadelphia: John Benjamins, pp. 37-60

Hardie, A (forthcoming) "CQPweb - combining power, flexibility and usability in a corpus analysis tool". Online available: http://www.lancs.ac.uk/staff/hardiea/cqpweb-paper.pdf

Hinrichs, L. (2006). *Codeswitching on the Web: English and Jamaican Creole in e-mail communication*. (Pragmatics and Beyond New Series 147). Amsterdam: John Benjamins.

Geoffrey Leech and Andrew Wilson (1996). EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora. Technical Report. Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-TCWG-MAC/R.

Negreiros, A. (1895). *Historia Ethnographica da ilha de S. Tomé*. Lisbon.

Pontífice, J. *et al.* (2009). *Alfabeto Unificada para as Línguas Nativas de S. Tomé e Príncipe (ALUSTP)*. São Tomé.

Quintas da Graça, A. (1989). *Paga Ngunu*. S. Tomé: Empresa de Artes Gráficas.

RGPH – 2001. (2003). *Características educacionais da população – Instituto Nacional de Estatística*. S. Tomé e Príncipe.

Schuchardt, H. (1882). Ueber das Negerportugiesische von S. Thomé. *Sitzungsberichte Wien* 101. 889-917.

Sebba, M. (1996). Informal orthographies, informal ideologies spelling and code switching in British Creole. Cadernos de Linguagem e Sociedade, Vol. 2, No 1.

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: Feb 27, 2012)

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) , Masalai Press, Oakland, California.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. www.tei-c.org/Guidelines/P5/ (Date of access: Feb 25, 2012).

Wynne, M. (2005). *Developing Linguistic Corpora*: a Guide to Good Practice. Oxford: Oxbow Books.