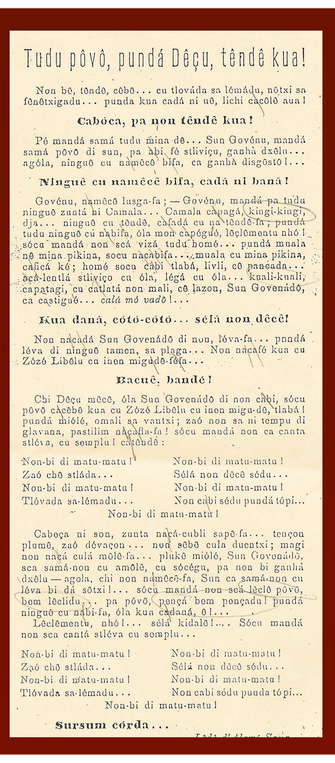
A Corpus of Santome

Tjerk Hagemeijer (t.hagemeijer@clul.ul.pt), Iris Hendrickx (iris@clul.ul.pt),
Haldane Amaro (amaro25@hotmail.com), Abigail Tiny (abigail.tiny@hotmail.com)

Centro de Linguística da Universidade de Lisboa (CLUL)



Tudu pôvô, pundá Dêçu, têndê kua!

Non bô, têndê côbô... cu tlovâda sa lómâdu, nôtxi sa fônôtxigadu... punda kua cadá ni uô, lichi cacôlê aua!

Cabôca, pa non têndê kua!

Pó mandá samá tudu mina dê... Sun Govénu, mandá samá pôvô di sun, pa bbi fô stliviçu, ganhá dxôlu... agóla, ninguê cu namécê bifa, ca ganhá disgôstó!...

Ninguê cu namécê bifa, cadá ni banâ!

Govénu, namécê lusga-fa; — Govénu, mandá pa tudu ninguê zuntá ni Camala... Camala capagá, kingi-kingi, dja... ninguê cu tendê, cadadá cu na tendê-fa; pundá tudu ninguê cu nabifa, óla mon capéguê, lóçlómentu nhó! sócu mandá non secá vizá tudu homé... pundá muala nê mina pikina, socu nacabifa..., muala cu mina pikina, cafécá kê; homé socu cabi tlabá, livli, cô pancadá... secá-lentlá stliviço cu óla, légá cu óla... knali-knali, capatagi, cu catlatá non mali, cô lazon, Sun Govenádo, ca castiguê... *calá mó vadô!*...

Kua daná, cóto-cóto... séla non dêçê!

Non nacadá Sun Govenádo di nou, lóva-fa... pundá lóva di ninguê tamen, sa plaga... Non nacaló kua cu Zózó Libólu cu ineu migadó-fóla...

Bacué, bandê!

Chi Dêçu mêcê, óla Sun Govenádo di non cabi, sócu pôvô cacabê kua cu Zózó Libólu cu ineu miga-dó, tlabá! pundá miólé, omali aa vantxi; zaó non aa ni tempu di glavana, pastilim nacalla-fa! sócu mandá non ca canta stléva, cu somplu! cadundê:

Non-bi di matu-matu! Non-bi di matu-matu!
Zaó chê stláda... Séla non dêcê sédu...
Non-bi di matu-matu! Non-bi di matu-matu!
Tlóvada sa-lémadu... Non cabi sédu pundá tópi...
Non-bi di matu-matu!

Cabeça ni sou, zunta naca-cubli sapô-fa... teuçon plumê, zaó dévaçon... nou sobô cala duentxi; magi non naçá culá molê-fa... plukê miólé, Sun Govenádo, sca samá-non cu amôlê, cu sócégu, pa non bi ganhá dxôlu — agola, chi non namécê-fa, Sun ca samá-non cu lóva bi dá sôtxi!... sócu mandá non sca lêclê pôvô, bem lêclidu... pa pôvô, pançá bem pançadu! pundá ninguê cu nabi-fa, óla kua cadaná, ê!...
Lêclêmentu, nhó!... séla kidalô!... Sócu mandá non sca cantá stléva cu somplu...

Non-bi di matu-matu! Non-bi di matu-matu!
Zaó chê stláda... Séla non dêcê sédu...
Non-bi di matu-matu! Non-bi di matu-matu!
Tlóvada sa-lémadu... Non cabi sédu punda tópi...
Non-bi di matu-matu!

Sursum córda...

1 – Introduction

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 São Tomé in the Gulf of Guinea (Africa). Santome is the result of language contact between Portuguese, the lexifier language, and several Benue-Congo languages, in particular Edo (Edoid) and Kikongo (Bantu) [2,3]. As 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.

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 Santome.

2 – Corpus

The corpus contains written and oral sources and has about 184K words. The number of available sources is limited. A large part of the corpus can be placed in the domain of folklore (folktales, riddles, song texts). The written part contains a few newspaper articles and books, pamphlets, cultural magazines and a blog. The spoken corpus comprises transcriptions of recordings made in 1997 and 2001 with twenty native speakers. These spoken recordings have been freely transcribed in the sense that we have tried to match written text as much as possible. Since we do not yet have copyrights for all the materials used in the corpus, at a first stage we plan to make the corpus available for concordances in an online interface, CQPweb [4], which allows users to search for concordances.

written subcorpus		
genre	**# files**	**# words**
mixed	10	22,652
music	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

3 – Language standardization

Despite the limited writing tradition in Santome, the orthographies have been highly variable. A word like [kwa] 'thing', for instance, has been written in the following ways: *cua, cuá, qua, quá, kua, kuá, kwa, kwá*. Therefore, we converted all orthography to ALUSTP,a phonology-oriented writing proposal developed in 2009 and ratified in 2010 by the Ministry of Education of Culture of S. Tomé and Príncipe [5]. The main principle of this proposal is a one-to-one phoneme-grapheme correspondence (see for instance the use of "k" [k] in the adapted version and its counterparts in the original orthography).

(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*

4 – Meta data

We encode meta data about the corpus texts like author and date in a simple XML format that is compatible with the P5 guidelines of the Text Encoding Initiative (TEI). We use the following fields:

- 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
- 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: Book, newspaper article, (cultural) magazines, pamphlets, online, unknown.
- genre: See table
- Notes:Additional information, such as publisher, place of publication.

5 – POS Annotation

Some examples of POS annotation.

(1) Sentence from the corpus:

Ola nansê ka xka nda ku migu
When 2PL TAM TAM walk with friend
CJ PRS TAM TAM V PREP CN
sela nansê toma kwidadu ê
must 2PL take care PRT
MOD PRS V CN PRT
'When you are hanging out with friends, you must be careful. '

(2) Reduplication (fully or partially):

tlêxi-tlêxi 'in groups of three' RED:NUM
tlê-tlêxi 'all three' ' RED:NUM

(3) Ideophones

kabêsa wôlôwôlô 'foolish person' (lit. head+id.) CN ID
sola potopoto 'cry intensely' (lit.cry+id.) V ID
vlêmê bababa 'intense red' (lit. red+id.) ADJ ID

Tag	Category	Examples (Santome)
ADJ	Adjective	glavi 'pretty', vlêmê, 'red' bon 'good'
ADV	Adverb	oze 'today', za 'already', yôxi 'yes'
ART	Article	ũa, inen.
CJ	Conjunction	i 'and', ô 'or', maji 'but', ola 'when'
CN	Common Noun	mosu 'boy', ope 'foot', ke 'house'
COMP	Complementizer	kuma 'that', ku 'who, which, that', xi 'if'
CONX	Sentence connector	so 'and then', soku 'and then'
DEM	Demonstrative	se 'this', sala 'that', ise 'this one'
DGT	Digit	0, 1, 42, 12345, 67890, …
EXC	Exclamative	Kê … 'How …'
FOC	Focus marker	so, soku.
FW	Foreign word	
ID	Ideophone	bababa, nwininwini, sũũũ, potopoto, ...
INDF	Indefinite	tudaxi 'everybody', nadaxi 'nothing'
INT	Interrogative	kuma 'how', andji, 'where'
ITJ	Interjection	kaki 'exclamation of surprise'
MOD	Modality Marker	sela 'must', milhon 'better…'
NEG	Negation marker	na, fa, fô, nantan, naxi.
ON	Onomatopoeia	plaplaplaplapla 'walking sound'
NUM	Numeral	ũa 'one', dôsu 'two', plumêlu 'first'
PM	Presentational marker	avia 'there was/were', ya 'here is/are'
PNM	Proper Name	Zon, Maya, Kolema, Txindadji
PNT	Punctuation Mark	., ?, ;,:,…,!
POSS	Possessive	mu(n) 'my', dinen 'their'
PP	Participle	fladu 'said', bixidu 'dressed'
PREP	Preposition	antê 'until', di 'of', djina 'since'
PRS	Personal	n 'I', ami 'I', bô 'you', non 'we'
PRT	Particle	an, en, ê, fa, fan, ô.
QNT	Quantifier	kada 'each', tudu 'all, every', yô 'many'
RED:xx	Reduplicated category	dôsu-dôsu 'in groups o f two'
REFL	Reflexive	mu 'myself', dê 'himself/herself'
RV	Residual value	abbreviations, acronyms, etc.
STT	Social Title	sun 'Mr.', san 'Misses', dôtôlô 'doctor'
TAM	Tense-Mood-Aspect marker	ka, ká, xka, ska, tava~ta.
V	Verb	mêsê 'want', pô 'may, can', sa 'be'

6 - Final remarks

Corpora for the three other Gulf of Guinea creoles are under way and, together with the Santome corpus, will be used for comparative research leading to the reconstruction of properties of the proto-Gulf of Guinea creole. We further expect this corpus to be an important tool in language maintenance and revitalization namely through the development of other language resources. Finally, a more wide-spread corpus-based approach to creole languages using large amounts of data will endow comparative research on creoles with tools that allow to assess claims regarding typological similarity between these languages.

References

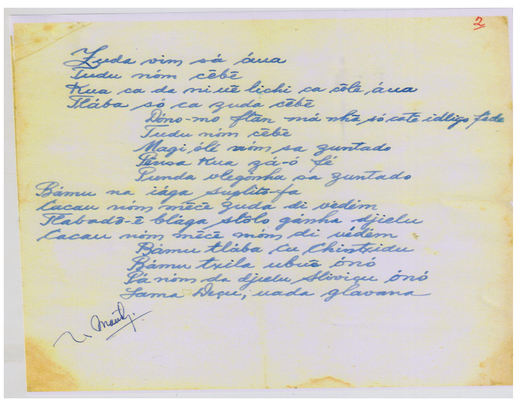
[1] RGPH – 2001. (2003). Características educacionais da população – Instituto Nacional de Estatística. S. Tomé e Príncipe.

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

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

[4] Hardie, A (forthcoming) "CQPweb - combining power, flexibility and usability in a corpus analysis tool".

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



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).

More information:

http://www.gulfofguineaceoles.com/

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Contact
Av. Professor Gama Pinto, 2, 1649-003 Lisboa, Portugal
Tel.: +351 21 790 79 00 | URL: www.clul.ul.pt

CLUL | Centro de Linguística da Universidade de Lisboa