Actas do Colóquio Internacional São Tomé e Príncipe numa perspectiva interdisciplinar, diacrónica e sincrónica (2012), 597-609

Instituto Universitário de Lisboa (ISCTE-IUL), Centro de Estudos Africanos (CEA-IUL), ISBN: 978-989-732-089-7
Instituto de Investigação Científica Tropical (IICT), ISBN: 978-989-742-002-3

Abigail Tiny º Haldane Amaro º Iris Hendrickx º Tjerk Hagemeijer
Centro de Linguística (CLUL)
Universidade de Lisboa, Portugal
abigail.tiny@hotmail.com º amaro25@hotmail.com º iris@clul.ul.pt º t.hagemeijer@clul.ul.pt

O Forro: A Construção de um Corpus[1]

Este trabalho apresenta o processo de construção de um corpus de material oral e escrito do forro (santome), um crioulo de base lexical portuguesa falado na ilha de São Tomé. O corpus compreende dados da segunda metade do século XIX até ao presente. Abordamos as dificuldades típicas de línguas não oficiais que são predominantemente orais, tais como a normalização ortográfica e um conjunto de dados mais restrito. Para a compilação do corpus seguimos padrões linguísticos de corpora e para codificar os metadados utilizámos a codificação de carateres UTF-8 e XML. Definimos um conjunto de metadados e apresentamos as etiquetas desenvolvido para a anotação dos dados com informação linguística.

Palavras-chave: São Tomé, crioulo, forro (santome), corpus linguístico, normalização, anotação

This paper presents the process of building a corpus of spoken and written material of forro (santome), a Portuguese-related creole spoken on the island of S. Tomé (Gulf of Guinea, Africa). The corpus comprises data from the second half of the 19th century until the present. We address the usual difficulties related to non-official languages that are predominantly spoken, such as orthographic normalization and a relatively small data set. For the corpus compilation we followed corpus linguistics standards and used UTF-8 character encoding and XML to encode meta information. We also present a POS-tag set developed for forro that will be used to annotate the data with linguistic information.

Keywords: São Tomé, creole, forro (santome), linguistic corpus, normalization, annotation

[1] Este trabalho é financiado pela Fundação para a Ciência e Tecnologia (FCT) no âmbito do projeto "As origens e o desenvolvimento de sociedades crioulas no Golfo da Guiné: Um estudo interdisciplinar" (PTDC/CLE-LIN/111494/2009). Uma versão anterior e ligeiramente diferente deste trabalho foi publicada como Hagemeijer, Hendricxk, Amaro, Tiny (2012).

Introdução

O forro é uma língua crioula falada na ilha de São Tomé no Golfo da Guiné. De acordo com os censos de 2001, 72,4% da população acima dos 5 anos de idade falavam forro (como primeira ou segunda língua), numa população total de 137.599 (RPGH – 2001, 2003). Para o português, a língua oficial e a mais falada na ilha, essa percentagem era de 98,9%, o que pressupõe que há um grau elevado de bilinguismo. No entanto, convém realçar que o português tem vindo a dominar cada vez mais o panorama linguístico são-tomense, especialmente entre as camadas mais jovens, resultando numa perda gradual das línguas crioulas autóctones.

Tal como angolar, o fa d'ambo (crioulo de Annobón) e o lung'ie (crioulo do Príncipe), que apresentam um menor número de falantes, o forro descende do proto-crioulo do Golfo da Guiné que se formou no século XVI, na ilha de S. Tomé, em consequência do contacto linguístico entre o português, a língua lexificadora, e várias línguas benue-congo, em particular o edo (da família edóide, Nigéria) e o kikongo (da família banta, República Democrática do Congo e Angola) (e.g. Ferraz, 1979; Hagemeijer, 2011).

Os primeiros estudos do forro acompanhados de dados linguísticos datam da segunda metade do século XIX (e.g. Schuchardt, 1882; Negreiros, 1895), mas é especialmente desde a monografia de Ferraz (1979) que se tem vindo a verificar um aumento significativo do número de estudos sobre o forro. Apesar do crescente interesse académico nesta língua, ainda faltam instrumentos linguísticos de referência, tais como uma gramática e um dicionário.

O corpus em construção será utilizado principalmente para fins linguísticos, mais especificamente para a extração de dados e a comparação com os outros três crioulos do Golfo da Guiné acima mencionados, com o intuito de reconstruir propriedades do proto-crioulo do Golfo da Guiné. O corpus poderá igualmente contribuir para tarefas de planeamento linguístico, tais como o desenvolvimento de dicionários, gramáticas e materiais de texto.

Tanto quanto pudemos averiguar, ainda não foram compilados corpora eletrónicos desta natureza para crioulos de base lexical portuguesa e, salvo raras exceções, esta afirmação também se aplica a línguas crioulas de outras bases lexicais. Esta lacuna poderá estar relacionada com o facto de muitas línguas crioulas serem línguas minoritárias que não possuem uma ortografia oficial nem uma tradição escrita. A compilação de corpora nestas condições torna-se uma tarefa custosa e trabalhosa, pois requer trabalho de campo para a recolha de dados, transcrição, revisão, normalização, etc.

Os poucos corpora que encontrámos para outras línguas crioulas são os seguintes. O corpus escrito do crioulo britânico[1] (Sebba, Kedge & Dray, 1999) contém cerca de 12,000 palavras do crioulo jamaicano utilizado na Inglaterra. O corpus consiste em diferentes amostras de géneros textuais e está anotado manualmente com etiquetas relativas ao léxico, às partes do discurso, à estrutura e às diferenças gramaticais entre o inglês padrão e o crioulo. Além deste corpus, está também disponível um

[1] *The Corpus of Written British Creole* no original.

corpus de 200,000 palavras do maurício, o crioulo de base lexical francesa das ilhas Maurícias, que pode ser pesquisado através de uma interface de concordância como parte do website do projeto ALLEX[2]. Há ainda um corpus do tok pisin, um crioulo de base inglesa com um substrato melanésio, constituído por 1,047 contos que foram traduzidos para inglês e publicados em forma de livro (Slone, 2001). A era digital também oferece novas possibilidades para a recolha de dados. Um exemplo é o corpus COJEC do crioulo jamaicano, uma coleção de e-mails e mensagens de um fórum com cerca de 40,000 palavras, produzidas por alunos jamaicanos (Hinrichs, 2006)[3].

O corpus

O corpus do forro consiste numa compilação de fontes orais e escritas. A partir da segunda metade do século XIX, a língua começou a ser escrita em fontes publicadas e não publicadas. Mas não se tratando de uma língua oficial, a quantidade de materiais produzida foi sempre bastante limitada.

Os materiais escritos do século XIX consistem em alguns poemas de Francisco Stockler e fragmentos recolhidos em textos de Coelho (1880-1886), Schuchardt (1882) e Negreiros (1895). As fontes também incluem alguns artigos de jornal dos anos 1920 e um número reduzido de livros e revistas que foram publicados após a independência do país (1975). Os livros e revistas culturais frequentemente intercalam textos em forro e em português e apenas alguns dos textos em forro apresentam uma tradução portuguesa. As fontes não publicadas compreendem uma série de panfletos dos anos 40 ou 50 do século XX, obtidos a partir de fontes privadas, e muitos textos não identificados (na maioria letras de músicas) recolhidos no Arquivo Histórico de S. Tomé e Príncipe. À exceção de uma ou outra fonte que ainda não fomos capazes de localizar, acreditamos que a nossa recolha abrange uma quantidade muito significativa dos materiais que foram produzidos em forro desde finais do século XIX. A produção em forro situa-se, frequentemente, no domínio do folclore (contos populares, provérbios, adivinhas, etc.) e raramente em domínos de prestígio como o jornalismo ou a educação. Note-se, por fim, que uma grande parte dos textos foram produzidos por um pequeno número de autores: o autor dos artigos de jornal da década de 1920 é também o autor dos panfletos (Francisco Bonfim[4]), uma boa parte dos letras de canções foi escrita por um reduzido número de compositores (e.g. Gete Rita) e a maioria dos provérbios resultam de uma só publicação (Daio, 2002). Também encontramos um blogue escrito em forro, da autoria de Jykiti Wakongo (pseudónimo). Os blogues constituem um género textual interessante com um estilo de escrita informal e pode ser visto como um diário *online* que expressa as opiniões pessoais do autor do blogue. O subcorpus escrito tem atualmente 99,658 palavras.

O corpus oral inclui transcrições de gravações de contos predominantemente populares narrados por contadores de histórias, bem como conversas e canções que foram gravadas em 1997 e 2001 com falantes nativos do forro provenientes de

[2] Projeto ALLEX: http://www.edd.uio.no/allex/corpus/africanlang.html
[3] Os e-mails de COJEC foram publicados num anexo do livro.
[4] Também conhecido como Faxiku Bêbêzawa, em forro.

diferentes locais de S. Tomé. Este subcorpus tem atualmente 52 gravações transcritas produzidas por 20 falantes diferentes, com um total de 84,951 palavras. Nas transcrições, muitos dos fenómenos da oralidade, tais como palavras fragmentadas, sons extralinguísticos, hesitações e repetições não foram anotados, visando manter os textos tão fluentes quanto possível. Na secção 4 encontram-se mais detalhes sobre o subcorpus escrito e oral.

Tendo em conta que ainda não obtivemos os direitos de autor de todos os materiais utilizados no corpus, não nos é possível, de momento, disponibilizar o corpus integral. Pretendemos, no entanto, disponibilizar o corpus para concordâncias numa plataforma online, o CQPweb (Hardie, no prelo), que permite aos utilizadores pesquisar concordâncias de palavras, sequências de palavras e categorias de palavras. A plataforma também permitirá aos utilizadores criar listas de frequências e restringir a pesquisa a tipos de texto específicos.

Padronização da língua

Tendo em conta que o forro não tem estatuto oficial, as ortografias, de base românica, têm sido variadas e frequentemente inconsistentes, situando-se num *continuum* que vai de ortografias de forte inspiração etimológica a ortografias de cariz fonético-fonológica. Este um problema bem conhecido para as línguas crioulas na sua generalidade (ex. Sebba, 1996). A título de exemplos, uma palavra como [kwa] 'coisa', por exemplo, tem sido escrita das seguintes formas: *cua, cuá, qua, quá, kua, kuá, kwa, kwá.* Há frequentemente uma proliferação desnecessária de acentos e uma irregular separação de palavras. No caso do forro, a explicação para a popularidade das ortografias etimológicas, i.e., ortografias baseadas no português, pode ser atribuída ao facto de uma larga maioria dos vocábulos que compõem esta língua serem provenientes do português, a língua oficial do país. Contudo, os étimos de origem portuguesa sofreram muitas vezes mudanças fonológicas drásticas aquando da sua incorporação histórica no crioulo (Ferraz, 1979). Há ainda um número considerável de etimologias desconhecidas ou relacionadas com o léxico de línguas africanas como o edo e o kikongo.

Estes fatores levaram-nos a adotar o Alfabeto Unificado para a Escrita das Línguas Nativas de S. Tomé e Príncipe (ALUSTP), uma proposta de escrita fonológica preparada em 2009 (Pontífice *et al.)* e homologada pelo Ministro da Educação e da Cultura de São Tomé e Príncipe em 2010[5]. O principal objetivo desta proposta é o de fazer corresponder a cada fonema (som) apenas um grafema (letra). A seguir, apresentamos um trecho de uma letra do conjunto Coimbra Nova, de 1970, seguido da versão adaptada à ortografia utilizada.

[5] O ALUSTP é discutido em Araújo (2010).

COIMBRA NOVA SAUDA O POVO

Coimbla fláam;
Tudo pôvô putuguêgi
Cu bi féça di Santomé Plôdôso
Sá punda homé migo d'homé
Só cá tònué c'opé ni féça.
Púnda chi nôtxi cubli
Migu cá ponté cámia chê stlàda

Diz Coimbra;
Todo o povo português
Que veio à festa de S.Tomé Poderoso
É porque quando um homém é amigo douti
Toca-lhe com pé na festa.
Porque caso anoiteça,
O amigo indica-lhe o caminho

2

Coimbla bila flaam:
Cúlo sá uê,
Plamám' sá tlaxi moli,moli;
Tudo nom cu bê plama zá,
Nom cá pidgi tádgi plô gimóla.
Punda póbli sá catxibo 'Dêço,
Lico cu démónó sá cóléga.
Tudo ninguê cu cá têdgêlo
Sá fédo chóchi cu démónó,
Punda óla cu nom pecadô cá sá lico,
Nom cá quêcê di sama Dêço.
Eia Ê!
Séla pá Sum di nom bi mundo
Bi sdluga tudo bem cú mali.

Diz ainda a Coimbra:
Escuridão está à frente,
Claridade vem depois;
Todos nós que já vimos a claridade,
Pedimos tarde por esmola.
Porque pobre é escravo de Deus,
Rico e demónio sã colégas.
Toda gente que tem dinheiro,
Faz-se sócio do demónio,
Porque quando somos ricos,
Esquecemo-nos de invocar Deus
Ah!
É preciso que o Nosso Senhor venha
Julgar bens e males.

Texto adaptado:
Coimbla[6] fla an:
Tudu pôvô putugêji
Ku bi fesa di Santome Plôdôzu
Sa punda ome migu d'ome
So ka tono ê ku ope ni fesa
Punda xi nôtxi kubli
Migu ka pont'e kamya xê xtlada

Coimbla bila fla an:
Kulu sa wê
Plaman sa tlaxi moli-moli
Tudu non ku bê plama za
Non ka pidji tadji plô jimola

[6] Mantivemos o "C" (em vez de "k") em Coimbla por se tratar do nome do conjunto.

Punda pobli sa katxibu Dêsu
Liku ku demono sa kolega
Tudu ngê ku ka tê djêlu
Sa fedu xoxi ku demono
Punda ola ku non pekadô ka sa liku
Non ka kêsê di sama Dêsu
Êya ê!
Sela pa Sun di non bi mundu
Bi xdluga tudu ben ku mali

Apresentamos, de seguida, alguns exemplos de padronização do texto original (i a v).

i. O som [ʃ] (como no português 'pei**x**e') é, neste texto, representado pelos seguintes grafemas: 'ch' (*chi, chê, chóchi*), 'x' (*tlaxi, catxibo*) e 's' (*stlàda*). Na versão adaptada, todos este contextos apresentam o grafema 'x' (*xi, xê, xoxi tlaxi, katxibu, xtlada*).

ii. O som [s], que é representado na versão original pelos seguintes grafemas 'c' (*quêcê*), 'ç' (*feça, Dêço*) e 's' (*séla, sama, sá*), torna-se 's' na ortografia que usamos (*kêsê, fesa, Dêsu, sela, sama, sa*).

iii. O som [u] é representado pelos seguintes grafemas: 'u' (*putuguêgi, punda, cubli*) e 'o' (*cúlo, Dêço, lico, mundo*). Uniformizámos estas vogais como 'u' (*putugêji, punda, kubli, kulu, Dêsu, liku, mundu*).

iv. O forro apresenta um contraste entre as vogais médias abertas [ɛ] e [ɔ] (como no português 'm**e**stre' e m**o**ça), e as vogais médias fechadas [e] e [o], (como no português 'v**e**r' e 'b**o**lo'). Na versão original, as palavras são grafadas com acento agudo e circunflexo, como nos exemplos *féça* e *ch*ê e *p*ôv*ô* e *p*ó*bli*. Uma vez que esta marcação é redundante, optamos por marcar as vogais médias fechadas com circunflexo, ao passo que as vogais médias abertas não recebem acentuação (*fes*a, *x*ê, *p*ôv*ô*, *p*o*bli*). No caso da vogal 'a', a utilização de acentos é redundante por não haver um par contrastante como no caso das vogais médias.

v. A separação de palavras pode ser observada nos seguintes casos: *fláam > fla an* (dizer + partícula *an*), *tònué > tono e* (tocar + clítico).

Tendo em conta a grande variação nas ortografias utilizadas, verifica-se que adaptação de todos os textos representa um trabalho meticuloso. Todos os textos escritos originais que compõem o corpus foram passados pelo scanner com o software OCR ou copiados manualmente e posteriormente adaptados de acordo com o ALUSTP num editor de texto. Muitos textos originais foram datilografados em máquinas de escrever e alguns escritos à mão. Frequentemente há emendas feitas no original que dificultaram a interpretação e alguns textos encontravam-se em mau estado de conservação, isto é, quase ilegíveis.

Estruturas que denotam variação linguística (ex: *djêlu/jêlu* 'dinheiro'; *idligu/igligu* 'fumo') foram mantidas sempre que possível, particularmente no corpus oral. Quanto à variação, o corpus escrito é claramente menos confiável, pois, nem sempre é

transparente qual a variação subjacente a uma determinada forma escrita. Por incluir a variação, o corpus torna-se também útil para análise quantitativa e possivelmente para determinar se há variação regional.

O corpus, e a variação nele encontrada, também estão a ser usados na preparação de um dicionário do forro com mais de 4,000 entradas lexicais (Araújo & Hagemeijer, em preparação).

Metadados

O formato do corpus segue as normas gerais para o corpus linguístico (e.g. Wynne, 2005) e utilizou-se a codificação de carateres UTF-8 e a anotação XML para os metadados. Decidiu-se codificar os metadados por autor e data num formato XML simples compatível com as diretrizes P5 do Text Encoding Initiative (TEI consortium, 2007).

Composição dos metadados:

língua: Para além do forro, o projeto irá construir um corpus para os outros três crioulos do Golfo da Guiné. Contudo, os corpora destas três línguas serão mais reduzidos e essencialmente restritos a dados orais devido à falta de uma tradição escrita.

- corpus: Oral ou escrito.
- título: Titulo do texto (se houver).
- autor: O autor do texto ou desconhecido.
- idade: Idade do falante (dados orais).
- local: Localização geográfica onde foram efetuadas as gravações (dados orais).
- data: Data da publicação (se houver), que pode ser exata ou aproximada. Não havendo evidências que contrariem, assume-se que as datas de publicação são aproximadas as datas em que foram escritos os textos.
- fonte: As fontes consideradas são as seguintes: livro, revista, artigo de jornal, panfleto, blogue e desconhecido.
- género: Os géneros considerados para o corpus escrito são: prosa, poesia, provérbios, adivinhas, música, misto e outros. Para o corpus oral, consideraram-se três géneros: prosa (para contos tradicionais e outras histórias), música e conversas.
- notas: Etiqueta reservada para informação adicional, como por exemplo o nome e o local de publicação.

No corpus, as narrativas ocupam um papel central, estando nelas incluídas os contos tradicionais, histórias (pessoais) e os textos retirados do blogue. Um dos critérios subjacentes à classificação do género está relacionado com a quantidade de dados disponível para cada género. Uma divisão mais pormenorizada iria conduzir a géneros com pequenas quantidades de material. No Quadro 1 está representada a distribuição de ficheiros e palavras por género.

Subcorpus escrito		
Género	**Ficheiros**	**palavras**
Misto	**10**	**22.652**
música (letras de músicas)	**169**	**21.081**
Poesia	**11**	**4.442**
Prosa	**59**	**40.364**
Provérbios	**3**	**9.081**
Outros	**4**	**1.936**
Subtotal	**257**	**99.658**
Subcorpus oral		
Conversação	7	20.945
Prosa	43	62.844
Música	2	802
Subtotal	**52**	**84.591**
TOTAL	**309**	**184.249**

Quadro 1 - Distribuição de ficheiros e palavras pelos géneros no corpus do forro.

O elevado número de ficheiros na categoria "música" deve-se ao facto de estarmos perante material não publicado, frequentemente letras encontradas em folhas de papel. Uma grande parte dos provérbios, por outro lado, foi publicada num único volume (Daio, 2002). Quanto ao género "misto", inclui publicações – revistas em particular - com diferentes tipos de textos pertencentes a um dos restantes cinco géneros. Nestes casos, no cabeçalho principal vigora a etiqueta "misto", porém, cada texto encontra-se subdividido para que se possa distinguir os diferentes géneros. Esta divisão é feita da seguinte forma: ... . Esta estratégia foi igualmente adotada para outras possíveis mudanças nos dados do cabeçalho, nomeadamente em casos de mudança de autor numa compilação de poesias. Para a parte oral, o material disponível enquadra-se em apenas três tipos de género pois grande parte é composta por histórias contadas (prosa).

Anotação POS

Após a uniformização dos dados iniciou-se a fase de enriquecimento do corpus através de anotação morfossintática. A escolha das etiquetas no Quadro 2 em baixo baseou-se no nosso conhecimento prévio da língua e numa amostragem de dados para o qual nos guiámos pelo conjunto de etiquetas em Leech & Wilson (1996) e nas etiquetas CINTIL desenvolvidas para o corpus Português CINTIL (Barreto *et al.,* 2006).

Etiqueta	Categoria	Exemplos
ADJ	Adjetivos	*glavi* 'bonito', *vlêmê* 'vermelho'
ADV	Advérbios	*oze* 'hoje', *yôxi* 'sim'
ART	Artigos	*ũa* 'um(a)', *inen* 'os, as'
CJ	Conjunções	*maji* 'mas', *punda* 'porque', *ku* 'e'
CN	Nomes Comuns	*mosu* 'rapaz', *ope* 'pé, perna'
COMP	Complementadores	*kuma* 'que'
CONX	Conectores	*zao* 'então', *êlê manda* 'por isso'
DGT	Dígitos	*0, 1, 42, 12345, 67890*
DEM	Demonstrativos	*se* 'este(a), aquele(a)', *xi* 'aquele(a)'
FOC	Marcadores de Foco	*so, soku* 'é que'
FW	Estrangeirismos	maioritariamente palavras em Português
ID	Ideofones	*sũũũ* (*pya sũũũ* 'olhar fixamente', lit. olhar+ID)
INDF	Indefinidos	*tudaxi, nadaxi, nyuku~niku*
INT	Interrogativos	*kuma* 'como', *andji* 'onde'
ITJ	Interjeições	*kaka!* (surpreso)
MOD	Marcadores de Modalidade	*sela* 'ter de'
NEG	Marcadores de Negação	*na, fa, fô*
NUM	Numerais	*dôsu* 'dois', *tlêxi* 'três'
ON	Onomatopeias	plaplaplaplapla
PP	Particípios	*bixidu* 'vestido',*vadu* 'rachado'
PM	Marcadores de Apresentação	*avia* 'havia'
PNM	Nomes Próprios	*Zon* 'João'
PNT	Marcas de Pontuação	*., ?, (, ...*
POSS	Possessivos	*mu* 'meu', *bô* 'teu'
PREP	Preposições	*antê* 'até', *ku* 'com'

PRS	Pronomes Pessoais	*n* 'eu', *ê* 'ele, ela'
PRT	Partículas	*an* (partícula interrogativa)
QNT	Quantificadores	*kada* 'cada', *tudu* 'todos(as)'
RED:xx	Categorias Reduplicadas	*kume-kume* 'comer continuamente' (RED:V)
REFL	Pronomes Reflexivos	*mu, bô, dê, non, ...*
RV	Valor Residual	abreviaturas, acrónimos, etc.
STT	Títulos Sociais	*sun* 'Sr.', *san* 'Sra.'
TAM	Marcadores de Tempo-Modo-Aspeto	*ka, xka, tava, ta.*
V	Verbos	*fla* 'falar', *mêsê* 'querer'

Quadro 2 - Etiquetas para a anotação morfossintática do corpus do forro.

A adaptação das categorias gramaticais foi crucial, uma vez que o forro é uma língua tipologicamente muito distinta do português que apresenta uma maior proximidade de certas línguas do oeste de África, tal como o edo, a sua principal língua de substrato (Hagemeijer, 2011; Hagemeijer & Ogie, 2011).

Contrariamente ao português, que é uma língua flexional, o forro é uma língua isolante, i.e. sem flexão morfológica, que tem apenas dois morfemas derivacionais produtivos. A reduplicação e a composição são, contudo, estratégias morfológicas produtivas.

Para as categorias reduplicadas usa-se RED: seguido da etiqueta da categoria que está a ser reduplicada. Os numerais cardinais, por exemplo, podem ser total ou parcialmente reduplicados (RED:NUM).

(1) *tlêxi-tlêxi* 'em grupos de três'

(2) *tlê-tlêxi* 'todos os três'

Para além das etiquetas padrão, foram criadas etiquetas específicas para esta língua. Os ideofones constituem uma categoria de palavras especial que consiste em modificadores com propriedades fonológicas específicas que ocorrem habitualmente com apenas um ou alguns itens lexais (nomes, verbos e adjetivos).

(3) *kabêsa wôlôwôlô* 'pessoa tonta' (lit. cabeça+ideofone)

(4) *sola potopoto* 'chorar intensamente' (lit. chorar+ideofone)

(5) *seku klakata* 'sequíssimo' (lit. seco+ideofone)

De seguida, apresentamos a anotação um trecho da letra de Coimbra Nova (cf. secção 3). A linha a. representa a frase em forro, a linha b. a glosa (o material funcional não traduzível tem uma etiqueta específica) e a linha c. a aplicação das etiquetas de anotação morfossintática.

a. Coimbla fla an:
b. Coimbra dizer PRT
c. PNM V PRT

a. Tudu pôvô putugêji
b. Todo povo português
c. QNT CN ADJ

a. Ku bi fesa di Santome Plôdôzu
b. Que vir festa de S. Tomé poderoso
c. COMP V CN PREP PNM ADJ

a. Sa punda ome migu d' ome
b. Ser porque homem amigo de homem
c. V CJ CN CN PREP CN

a. So ka tono ê ku ope ni fesa
b. É que TAM tocar ele/a com pé em festa
c. FOC TAM V PRS PREP CN PREP CN

a. Punda xi nôtxi kubli
b. Porque se noite cobrir
c. CJ CJ CN V

a. Migu ka pont' e kamya xê xtlada
b. Amigo TAM apontar -lhe caminho sair estrada
c. CN TAM V PRS CN V CN

Para a anotação morfossintática irá ser usado um etiquetador que é treinado a partir de uma amostra anotada manualmente e que permite a etiquetagem automática do resto do corpus.

Considerações Finais

Neste artigo apresentámos, de forma sucinta, a construção e a anotação do corpus do forro. No processo de criação de recursos para uma língua crioula como o forro há múltiplas questões por resolver, tais como a variação lexical, tanto do material escrito como do oral, a dimensão limitada do corpus com variação ortográfica, a falta de recursos padronizados como dicionários ou gramáticas de referência. Abordámos estas questões da seguinte forma: 1) recolhendo o máximo de material escrito para incluir num corpus uniformemente codificado; 2) adicionando metadados com informação; 3) utilizando uma ortografia sistemática para padronizar a grafia do material escrito; 4) transcrevendo o material oral no mesmo formato; 5) desenvolvendo um conjunto de etiquetas para a anotação morfossintática do forro.

Uma vez concluídos o corpus do forro e os corpora dos outros três crioulos do Golfo da Guiné, prevê-se que sejam recursos importantes na investigação sobre a relação linguística entre estas quatro línguas e ferramentas que contribuam para a manutenção e a revitalização destas línguas, designadamente através do desenvolvimento de outros recursos linguísticos.

Referências bibliográficas

Araújo, Gabriel (2010). Relações entre as fonologias das línguas crioulas de STP e a 'proposta ortográfica' ALUSTP. In 7º Congresso Ibérico de Estudos Africanos, 9, Lisboa - 50 anos das independências africanas: desafios para a modernidade : actas [Em linha]. Lisboa: CEA, 2010. Disponível em: http://hdl.handle.net/10071/2349

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

Barreto F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F. P., Nunes, F. e Silva, J. (2006). Open resources and tools for the shallow processing of Portuguese. Em *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, Tjerk; Hendrickx, Iris; Amaro, Haldane; Tiny, Abigail (2012). A Corpus of Santome. In *Proceedings of the SALTMIL-AfLaT workshop*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

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 (no prelo) "CQPweb - combining power, flexibility and usability in a corpus analysis tool". Disponível online: 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 ((Data de accesso: 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.