Les langues créoles de São Tomé : transcrire pour écrire

Emmanuel Schang

CORAL (UPRES-EA 3850) – Université d'Orléans
Faculté LLSH
BP 46527, 45065 ORLEANS Cedex 2 (France)
emmanuelZschang@univ-orleansZfr

Mots-clés : Créoles portugais, golfe de Guinée, São Tomé et Príncipe, corpus oraux
Keywords: Creole languages, Gulf of Guinea Portuguese-based Creoles, São Tomé and Príncipe, spoken corpora

Résumé

Je présente une initiative visant à mettre à la disposition du public (scientifiques et Organisations Non Gouvernementales) des corpus oraux transcrits sur les créoles portugais du golfe de Guinée ainsi qu'une base de données lexicales des termes utilisés dans ces corpusZ Ce projet vise à contribuer à l'essor de l'écriture en forro et en angolarZ

Abstract

This paper presents an ongoing project for the Gulf of Guinea Portuguese-based Creole languagesZ It aims at delivering transcribed spokencorpora and a lexical databaseZ

1 Introduction

Je présente dans cet article une initiative de sauvegarde et de développement des langues créoles de São Tomé (République Démocratique de São Tomé et Principe) utilisant les nouvelles technologiesZ Encore peu décrites, ces langues sont également assez peu connuesZ Ce projet a pour objectif d'augmenter la documentation sur ces langues par le biais de corpus oraux transcrits et d'une base de données lexicalesZ Il prend place dans une initiative plus vaste, CreolData, visant à créer une base de données lexicales informatisée sur les créoles portugais d'Afrique (Schang & alii, *à paraître*), rassemblant d'autres créolistes et s'appuyant sur le dictionnaire de Jean-Louis Rougé (Rougé 2004)[1]Z

[1] Pour l'instant, ces projets n'ont pas fait l'objet de demandes de financement propres et reposent sur la libre contribution de chercheurs universitairesZ

Les rapports qu'entretiennent ces langues créoles avec le portugais et les langues africaines encore parlées dans certaines communautés seront également abordés car il est nécessaire de situer ces langues dans leur environnement sociolinguistique pour appréhender les difficultés dans le passage à l'écriture et à l'*informatisation* de ces languesZ

2 Les créoles portugais du golfe de Guinée

2.1 Quelques mots d'histoire

Il est impossible de décrire précisément en quelques lignes la situation linguistique d'un pays aussi complexe que São Tomé et PrincipeZ Cependant,on peut tracer quelques grands traits qui permettront au lecteur de se faire une idée de la situationZ

La République Démocratique de São Tomé et Príncipe (capitale : São Tomé) comprend deux îles : São Tomé et PríncipeZ Indépendantes depuis 1975, ces deux îles situées dans le golfe de Guinée couvrent une superficie d'à peine 1000 km^2 au total pour une population d'environ 175000 habitantsZ

L'île de São Tomé est, selon toute vraisemblance, inhabitée lorsque les navigateurs portugais la découvrent en 1471Z Elle ne sera peuplée de façon définitive qu'à partir de 1482, par des portugais relégués (*degredados*), des esclaves venant du royaume du Bénin puis de la région Congo-Angola, un millier d'enfants juifs convertis de force en 1492 et une poignée d'aventuriers commerçants venus de toute l'Europe (vZ Caldeira 1999)Z

Au succès de l'époque de la canne à sucre (16ème siècle) succèdera une période troublée de déclin économique jusqu'au 19ème siècle, qui verra les beaux jours des plantations de cacao et de café et l'apport d'une main d'oeuvre contractuelle (parfois forcée) originaire principalement du Cap Vert, de l'Angola et du MozambiqueZ

2.2 La situation actuelle

De ce passé, l'île hérite d'une situation sociolinguistique complexeZ En effet, à côté du portugais qui est la langue officielle de l'île on trouve deux langues créoles distinctes : le *forro* (ou lungwa santomé) et l'*angolar* (ou ngola)Z Le forro est le créole majoritaire parlé à la fois dans la capitale (São Tomé) et dans l'ensemble de l'île (ainsi que sur l'île de Principe, à côté du créole local, le lung'ie)Z L'angolar est le créoleparlé actuellement par les pêcheurs de l'îleZ Mais historiquement, il s'agit de la langue des esclaves fugitifs réfugiés dans les montagnes de l'île qui seront à l'origine de révoltes importantesZ Les Angolares, chassés des terres, seront contraints au 19ème siècle à devenir pêcheurs et à vivre sur les plages de l'îleZ

A côté de ces langues créoles locales, on entend toujours parler le créole capverdien (au travers de ses différents dialectes) mais aussi ce qu'on appelle les langues des Tongas (Rougé 1992) qui sont des vestiges de langues bantoues du Mozambique et d'Angola, et un portugais local (portugais des Tongas) parlé dans les plantationsZ

Dans ces conditions, le portugais, qui est la seule langue parmi celles citées précédemment à être véritablement enseignée à l'école, constitue la langue des élites et de l'émigration (puisqu'il existe une importante diaspora santoméenne à Lisbonne)Z

Mais dans un contexte d'indépendance, le forro incarne à la fois le sentiment national (la rue reprochera volontiers lors des élections à tel ou tel homme politique de ne pas parler le forro) et la tradition culturelleZ

3 Développement linguistique

Comme le soulignait JZ-LZ Rougé lors du Premier Colloque international sur les Langues Nationales qui s'est tenu à São Tomé en octobre 2001 (JZ-LZ Rougé 2001), il existe une différence importante entre écrire en créole et transcrire le créoleZ Si quelques tentatives de transcription ont vu le jour pour le forro et l'angolar, on peut dire qu'il n'existe pas de littérature écrite dans les créoles de São Tomé, pas plus d'ailleurs que d'ouvrages techniques ou scolairesZ L'usage du créole transcrit reste le fait de quelques initiatives personnelles (généralement de la poésie comme Paga Ngunu, etcZ)ZLa Radio Nationale et quelques Organisations non Gouvernementales ont fait parfois usage de slogans écrits en forro, sans toutefois chercher l'adoption de conventions orthographiques stablesZ Des initiatives liées à la Direction de la Culture tentent cependant de lancer un cours de créole à l'Ecole Polytechnique de São Tomé, sans grand impact pour l'instantZ

On peut dire sans déformer la réalité que le portugais est la seule langue écrite de São Tomé à l'heure actuelleZ

Faute de standardisation de la graphie, des formes concurrentes voient le jour pour noter certains phonèmes (faut-il noter [k] par *k* ou *c* dans [fika] "rester/être" ?)Z Bien que ceci soit un frein pour le développement de l'écriture en langue créole, il n'y a pas là un problème insoluble ou propre aux créoles (qu'on songe à des langues comme le luxembourgeois par exemple qui font face aux mêmes problèmes)Z Des solutions techniques existent qui ont été développées pour d'autres langues créoles (CreoleConvert http://hometownZaolZcom/mit2haiti/CrConvZhtm par exemple pour le créole haïtien)Z

Il existe bien sur Internet un groupe de discussion sur São Tomé qui cherche à promouvoir la culture santoméenne et le forro, mais l'impact de cette initiative retentit plus sur la diaspora que sur la population insulaire mêmeZ Comme ailleurs, l'utilisation du créole sur la Toile, par le biais des messages, s'apparente plus à du folklore (sympathique au demeurant) qu'à une volonté concrète de développement linguistiqueZ

Les corpus oraux que nous avons réalisés[2] sur le terrain montrent bien qu'en situation de conversation spontanée, le vocabulaire utilisé par d'authentiques créolophones est largement différent de celui utilisé par les internautesZ L'érudition et l'originalité sont revendiquées par les internautesZ L'exclusion de toute forme sensée être proche du portugais (perçu par certains comme l'adversaire linguistique du créole) est la règleZ On s'interrogera plus loin sur la pertinence de cette attitudeZ Ceci se retrouve par ailleurs dans toutes les mises en écriture des langues minoritaires (les autres langues créoles n'échappent pas à ce problème)Z

[2] Rapport de mission à São Tomé, mai 2004, J-L Rougé et EZ Schang (CORAL)Z

4 Genèse et philosophie du projet

Quelques explications sur la genèse de ce projet éclaireront probablement le lecteur sur la philosophie qui nous guide iciZ

C'est au cours d'un travail à São Tomé en 1997 aux côtés des ONG dans des projets de microcrédit (Caixa de Poupança) qu'est née l'idée de favoriser l'émergence auprès des populations peu scolarisées (ici les femmes des pêcheurs angolares des plages du sud de l'île) de l'écriture en créoleZ Le projet de microcrédit n'a pas continué[3] mais l'idée est restéeZ

Le point de départ consistait à recueillir des enregistrements sur différents sujets de la vie quotidienneZ Ceux-ci devaient servir de base pour l'élaboration de documents pédagogiques et technique en créole, en réutilisant les termes employés par les informateurs lors des entretiensZ Il s'agissait de prendre le contre-pied d'une pratique courante consistant à chercher à traduire en créole des termes portugaisZ L'idée consistait donc à ne pas se servir de concepts exogènes (souvent issus de documents techniques portugais) mais d'employer les termes de la vie quotidienne autant que possibleZ Si un organisme voulait à relancer des projets de ce type, la base d'entretiens annotés que nous proposons lui serait disponible, évitant le coût d'une étude préliminaireZ

Mais à l'heure actuelle[4], les travaux visent essentiellement la communauté des linguistes en cherchant à sauvegarder des traces (orales) de ces langues créoles (le lung'Ie de Príncipe étant quasiment éteint, les autres créoles risquent de suivre cette voie)Z

Le projet est pionnier pour ce groupe de langues, mais modeste tant par les moyens mis en œuvre que par l'ambition avouéeZ Il s'agit pour l'essentiel de constituer des enregistrements transcrits et annotésZ Ceux-ci seront accessibles rapidement à la communauté scientifique et à qui le souhaitera pour une utilisation compatible avec la philosophie de ce projetZ

Ce projet concernant les créoles de São Tomé s'inscrit dans une initiative plus large visant à constituer une base de données lexicales sur les créoles portugais d'Afrique, CreolData (vZ Schang & alii *à paraître* et note 1)Z

Bien entendu, une telle approche va de paire avec le parti pris d'adopter autant que possible les logiciels libres et gratuitsZ

5 Démarche suivie

La démarche consiste à recueillir des enregistrements des créoles de São Tomé lors de missions de recherche sur le terrainZ Les entretiens enregistrés portent aussi bien sur les contes et la tradition orale que sur la vie quotidienneZ Les informateurs enregistrés sont des locuteurs de langue maternelle créole et utilisent le créole quotidiennementZ Les enregistrements récents sont faits directement au format numérique (fichiers Zwav, 16-bit, 44100hz)Z

---

[3] Pour des raisons qui n'ont pas de rapport avec la linguistiqueZ

[4] En raison d'un contexte politique peu favorable, pour de nombreuses raisons, tant nationales qu'internationales, qui ne permet pas de placer l'éducation en créole parmi les prioritésZ

Les entretiens sont transcrits ensuite par le ou les linguistes sous le contrôle d'un informateur à l'aide du logiciel Transcriber (Barras & alii 2001)Z

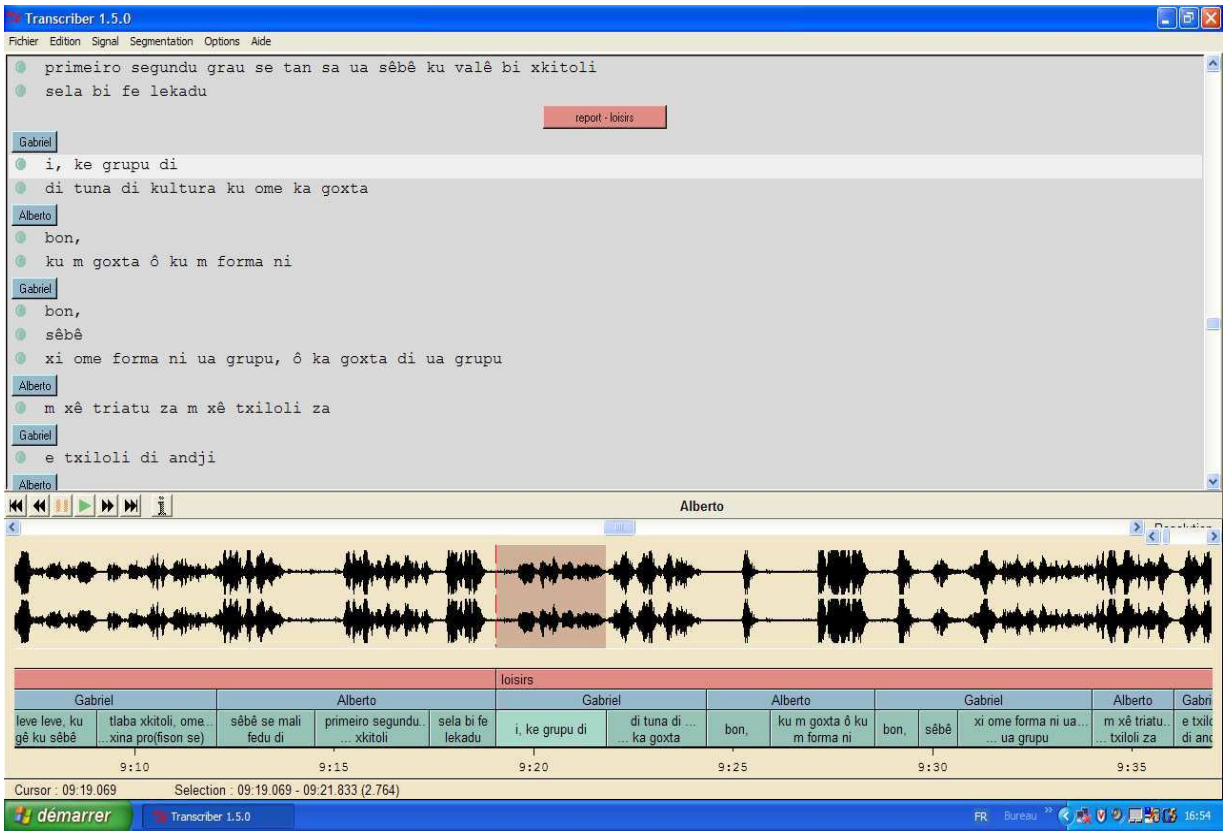


Figure 1 : transcription des entretiens avec Transcriber.

La graphie choisie pour transcrire les entretiens est celle qui est utilisée par JZ-LZ Rougé dans son dictionnaire (Rougé 2004)Z Aucune graphie officielle n'étant adoptée à l'heure actuelle à São Tomé (même si des propositions dans ce sens ont été faites au Colloque International sur les Langues Nationales de São Tomé et Príncipe), il paraît intéressant d'opter pour une graphie qui soit compatible avec les différents créoles portugais d'Afrique car certains termes sont communs aux différents créoles (par exemple *kume* "manger")Z Nous reviendrons plus tard sur ce pointZ

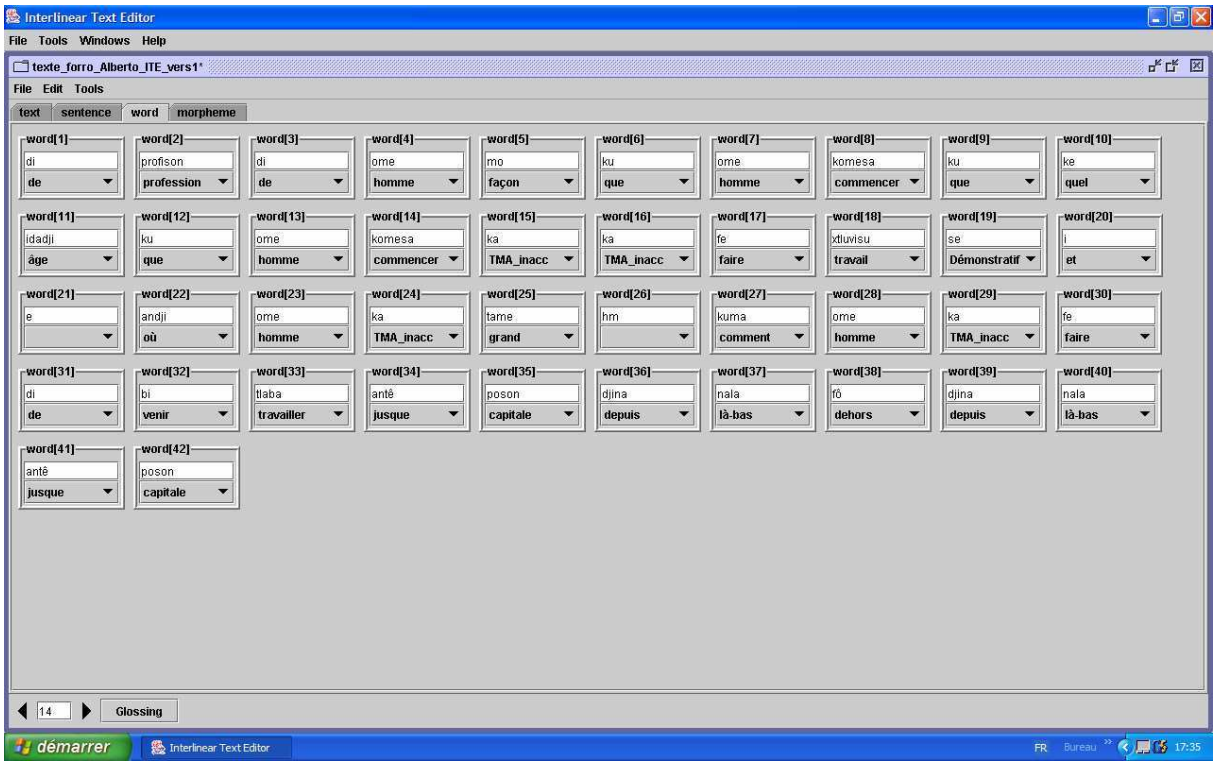
Transcriber permet par la suite de travailler sur des fichiers texte ou des fichiers XML contenant la transcriptionZ

```
<Section type="report" topic="to2" startTime="559.069" endTime="1040.730">
<Turn speaker="spk1" startTime="559.069" endTime="564.326">
<Sync time="559.069"/>
i, ke grupu di
<Sync time="561.833"/>
di tuna di kultura ku ome ka goxta
</Turn>
<Turn speaker="spk2" startTime="564.326" endTime="568.552">
<Sync time="564.326"/>
bon,
<Sync time="565.938"/>
ku m goxta ô ku m forma ni
</Turn>
<Turn speaker="spk1" startTime="568.552" endTime="574.0">
<Sync time="568.552"/>
bon,
<Sync time="569.644"/>
```

```
sêbê
<Sync time="570.635"/>
xi ome forma ni ua grupu, ô ka goxta di ua grupu
</Turn>
<Turn speaker="spk2" startTime="574.0" endTime="576.053">
<Sync time="574.0"/>
m xê triatu za m xê txiloli za
</Turn>
<Turn speaker="spk1" startTime="576.053" endTime="577.425">
<Sync time="576.053"/>
e txiloli di andji
</Turn>
<Turn speaker="spk2" startTime="577.425" endTime="578.897">
<Sync time="577.425"/>
txiloli Santana
</Turn>
<Turn speaker="spk1" startTime="578.897" endTime="581.28">
<Sync time="578.897"/>
sa ome sa blabu ni kwa se
</Turn>
```

Figure 2 : segment du fichier XML contenant la transcription de l'entretien

Ensuite, la transcription sous format texte sert d'entrée à l'élaboration d'une glose à l'aide de Interlinear Texte Editor (désormais ITE; Lowe & alii 2004 et http://michelZjacobsonZfreeZfr/ Les mots sont alors glosés[5] en français (figZ 3) et une proposition de traduction du texte est élaboréeZ ITE permet la réalisation d'une nomenclature (figZ 4) des termes utilisés dans les entretiensZ Il comprend également un concordancier très utile au linguisteZ Son principal avantage dans le cadre de ce projet est qu'il utilise XML et nous permet alors, grâce aux feuilles de style XLST, de pouvoir mettre en forme les données selon nos impératifs du momentZ



[5] La glose étant entendue comme une traduction littérale du mot créole en françaisZ

Figure 3 : la glose mot par mot des entretiens avec ITE.

Cet outil apparaît aussi utile pour les travaux de recherche en linguistique que pour la mise en forme des données recueilliesZ S'il est conçu pourpermettre une transcription juxtalinéaire aisée (vZ le programme Archivage du LACITO : http://lacitoZvjfZcnrsZfr/archivage/), l'utilisation qui en est faite dans ce projet est autreZ En effet, le lexique récupéré dans les entretiens (figZ 4) sert d'entrée à la base de données lexicales que nous élaboronsZ Les items lexicaux sont repris à partir du fichier XML contenant le lexique et constituent le point de départ d'une entrée lexicale de CreolDataZ

```
<?xml version="1.0" encoding="UTF-8"?>
<lexique>
  <item nb="2">
    <transcription>mendu</transcription>
    <glose>peur</glose>
  </item>
  <item nb="1">
    <transcription>matu</transcription>
    <glose>forêt</glose>
  </item>
  <item nb="2">
    <transcription>mindjan</transcription>
    <glose>médicament</glose>
  </item>
```

Figure 4 : extrait du lexique obtenu avec ITE.

La glose est alors abandonnée au profit d'une définition (s'inspirant ou provenant pour l'essentiel de Rougé 2004)Z

Les données sont alors présentées sous une forme s'inspirant largement des propositions de norme en cours (Lexique pour le TAL et Lexical Markup Framework[6])Z

[6] Documents et discussions disponibles sur wwwZnormalangueZorg et wwwZtc37sc4Zorg

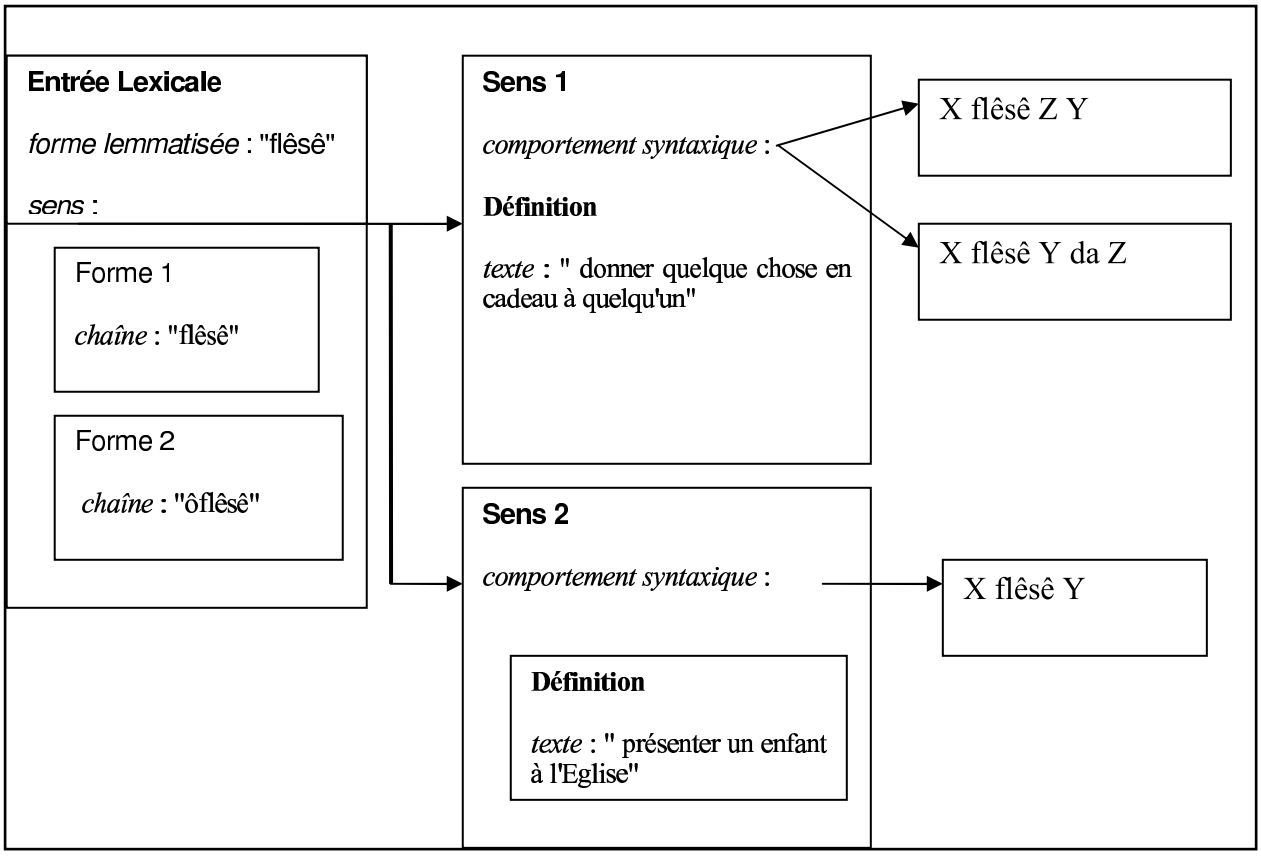


Figure 5 : schéma représentant l'entrée lexicale de flêsê "offrir".

La figure 5 illustre de façon schématique le traitement du mot *flêsê* "offrir"Z Il existe en forro deux variantes (au moins) *flêsê* et *oflêsê* qui ont deux sens distincts : 'présenter un enfant à l'Eglise' et 'offrir'Z Ce dernier sens autorise deux comportements syntaxiques (<X flêsê Z Y> et < X flêsê Y da Z>)Z

Pour l'instant, il n'y a pas de véritable traitement de l'interface syntaxe-sémantique car celle-ci pose un grand nombre de problèmes tant pratiques que théoriquesZ Faut-il utiliser la théorie classique des rôles thématiques (comme proposé dans Schang & alii *à paraître*) ou bien opter pour une autre théorie ? A l'heure actuelle, nous ne disposons pas d'éléments de réponse satisfaisantsZ D'un côté, la théorie des rôles thématiques pose problème : il n'est pas plus simple en forro qu'en français de savoir quels rôles attribuer pour les verbes *lire* ou *persuader*Z De l'autre, d'autres modèles plus complets existent mais ils sont lourds à mettre en œuvre (que ce soit la Théorie Sens-Texte ou le Lexique Génératif, vZ Bouillon & Busa 2001 pour une discussion)Z La morphologie (peu importante dans ces créoles en l'absence de flexion nominale et verbale ainsi que de procédés dérivationnels propres à la morphologie, vZ Schang (2000)) est également absente et fera l'objet de travaux à venirZ Le lexique étant issu des enregistrements recueillis, il se compose des atomes reconnus par la syntaxeZ Ce n'est probablement pas satisfaisant dans l'absolu, mais dans une première approche, il s'agit d'un choix de raisonZ

Quoi qu'il en soit, ces problèmes n'ont rien de spécifique aux créoles portugais et se posent à l'identique pour toutes les languesZ La différencetient ici essentiellement dans les moyens à disposition pour le traitement de ces problèmesZ Eneffet, le choix de coller au plus près à la proposition de norme Lexical Markup Framework relève ici plus de l'exercice de style que de

la nécessité économique : qui développera des logiciels de traduction pour une langue de moins de deux cent mille locuteurs ? C'est la volonté de partager les données qui prévaut ici et non pas les intérêts économiquesZ

6 Problèmes et développements futurs

Nous avons déjà vu dans les paragraphes précédents quelques problèmes liés à l'élaboration du projetZ J'insisterai ici sur la question de la variation (et de sa productivité), qui se pose dès lors qu'on cherche à décrire le créole ou à proposer des transcriptions du créoleZ

Certes, celle-ci n'est pas un phénomène propre aux langues créoles mais les créoles posent depuis toujours la question de l'identification des langues (vZ notamment Schang 2004 pour une discussion)Z Je ne parle pas ici de la coexistence de variantes d'un même mot (vZ figure 5 pour une solution à ce problème) mais de la difficulté de cerner les limites de ce qu'on appelle le créoleZ

Le problème que l'on rencontre tient à l'imbrication des systèmesZ Les locuteurs passent d'un système (devrais-je dire d'une langue ?) à l'autre, tant au plan lexical qu'au plan grammaticalZ C'est ce qu'illustre (1) où figure en italique un mot portugais fléchi au milieu d'une phrase en forroZ

(1) non na xka te *condições* (…)

/nous/négZ/Temps-Mode-Aspect/avoir/conditions/

nous n'avons pas la vie facile (...)

De façon générale, il est absolument arbitraire d'exclure les mots portugais du forro et de chercher à créer un créole "pur" sachant que les mots d'origine portugaise constituent près de 95 % du lexique du forroZ

Par ailleurs, s'il est aisé de repérer ce qui est typiquement angolar (défini en creux comme ce qui n'est ni forro ni portugais), on peut remarquer que l'angolar et le forro sont très proches, tant par leur vocabulaire que par leurs structures grammaticalesZ La vision organique des langues en tant qu'entités distinctes qui vivent et meurent est dans ce contexte mise à malZ Dans des enregistrements spontanés, il est très souvent difficile d'estimer s'il s'agit de portugais (avec un fort accent local) contenant des mots créoles ou bien d'une variété acrolectale (proche du portugais) du créoleZ Il enva de même entre les deux créoles forro et angolaresZ

La conception que je défends consiste à dire que les mots n'appartiennent pas à une langue mais que celle-ci les utiliseZ En effet, dire, parexemple, que le mot *taxi* est un mot français, portugais ou forro n'a guère de sensZ On trouvera dans les trois langues quelque chose qui, à la variante de prononciation près, est le mot *taxi*Z Doit-on noter qu'il s'agit d'un mot d'origine grecque ? Pourquoi pas, tant qu'il s'agit de dire qu'il s'agit d'une origine grecque et non pas d'un mot grecZ Ainsi, je prends le parti de ne pasexclure les mots d'origine portugaise des transcriptions et du lexique que je mets en placeZMais s'agit-il alors d'un lexique forro ou d'un lexique correspondant à un/des locuteurs(s) dans une situation de communication donnée ? C'est assurément la seconde réponse qui est la bonneZ Dans ce cas, je ne peux que proposer la description de situations de communication plutôt que la description d'une langueZ De ce point

de vue, on ne peut qu'être d'accord avec la devise de linguasphère (http://wwwZlinguasphereZorgZ): "dans la galaxie des langues, la voix de chaque personne est une étoile"Z

Remerciements

Je remercie mon collègue Jean-Louis Rougé pour sa collaboration à ce travail, ainsi que Laurent Romary et Gil Francopoulo pour leurs conseils ponctuelsZ

Références

BARRAS, CZ,GEOFFROIS, ÉZ,WU, ZZ,LIBERMAN, MZ(2001), Transcriber: development and use of a tool for assisting speech corpora productionZ*Speech Communication*, *33*(1-2): 5–22Z

BOUILLON, PZ,BUSA, FZ (2001)Z*The language of word meaning*Z Studies in NLP, Cambridge University PressZ

CALDEIRA, AZ (1999),*Mulheres, sexualidade e casamento em São Tomé e Príncipe.* Lisboa: CosmosZ

LOWE, JZ,JACOBSON, MZ,MICHAILOVSKY, BZ (2004), Interlinear Text Editor Demonstration and Projet Archivage Progress ReportZ Actes de :*4th E-MELD (Electronic Metastructure for Endangered Languages Data) workshop on language engineering: Linguistic Databases and Best Practice. Detroit. 15-18 juillet 2004*Z

ROUGE, JZ-LZ(1992), Les langues des TongaZ Actes de :*Colóquio sobre 'Crioulos de base lexical portuguesa', Universidade de Lisboa, junho de 1991, E. d'Andrade e A. Kihm (eds)*Z

ROUGÉ, JZ-LZ(2001), Escrever i/o transcrever as línguas de São Tomé e PríncipeZ Actes de : *Collóquio Internacional sobre as línguas nacionais. São Tomé, oct. 2001.*

ROUGE, JZ-LZ (2004),*Dictionnaire étymologique des créoles portugais d'Afrique*Z Paris : KarthalaZ

SCHANG, EZ (2000), L'émergence des créoles portugais du golfe de GuinéeZ Presses Universitaires du SeptentrionZ Thèse de Doctorat del'Université Nancy 2Z

SCHANG, EZ (2004), Identification automatique des langues :que faire des créoles ? Actes de : *Workshop MIDL 2004. Paris, nov. 2004. Presses de l'ENST.*

SCHANG, EZ,ROUGE, JLZ,ESHKOL, IZ,PETIT, MZ (à paraître en 2005), CreolData : une base de données lexicales informatisée sur les créoles portugaisZ*Les créoles, Revue Française de Linguistique Appliquée*, DZ Fattier (edZ)Z