Un corpus oral transcrit de kriol

Jean-Louis Rougé[1] Emmanuel Schang[1] Ana Luís[2] Flora Badin[1] Eugène Tavares[3]
(1) LLL- Université d'Orléans
(2) CELGA-ILTEC- Universidade de Coimbra
(3) Université de Ziguinchor
jean-louis.rouge@univ-orleans.fr, emmanuel.schang@univ-orleans.fr, aluis@fl.uc.pt, flora.badin@univ-orleans.fr, etavares@univ-zig.sn

RÉSUMÉ

Nous présentons la constitution d'une ressource sur le kriol, langue créole à base lexicale portugaise, parlée en Guinée-Bissau. Il s'agit dans un premier temps de la mise à disposition de 25 heures d'enregistrements transcrits, glosés et traduits. Cet article propose une présentation des enjeux (§2) ainsi qu'une description de la méthode choisie pour le traitement des données (§3).

ABSTRACT

We present the development of a language resource for Kriol, a Portuguese-based Creole spoken in Guinea-Bissau, consisting of 25 hours of transcribed, glossed and translated recordings. In this article,we lay out the sociolinguistic context of Kriol (§2) and offer a description of the method chosen for the processing of the data (§3).

RÉSUMÉ EN LANGUE NATIONALE

Este artigo apresenta a criação de um recurso linguístico para o Kriol, um crioulo de base lexical portuguesa falado na Guiné-Bissau. Trata-se de um recurso constituído por 25 horas de gravações transcritas, glosadas e traduzidas. Começamos por apresentar o contexto sociolinguístico do Kriol (§2) para depois oferecer uma descrição do método escolhido para o processamento dos dados (§3).

MOTS-CLÉS : kriol, creole, ressource, corpus oral, transcription.

KEYWORDS: Kriol, Creole, resource, spoken corpus, transcription.

1 Introduction

Nous présentons ici un projet en phase initiale consistant à la fois en la mise à disposition de données recueillies depuis une dizaine d'années (faisant suite à des travaux entrepris depuis 1977) en Guinée-Bissau et à Ziguinchor et à la préparation d'un recueil de nouvelles données sur le kriol (créole à base lexicale portugaise principalement). Il s'agit donc d'un projet de constitution de ressources sur une langue d'Afrique, le kriol.
Dans un premier temps, nous expliquons les motivations qui nous conduisent à la constitution d'un tel projet, en donnant le contexte socio-historique dans lequel est parlé le kriol.
Nous présentons ensuite la méthodologie suivie pour l'exploitation des données et les perspectives.

2 Le kriol

2.1 Présentation socio-historique

Le kriol est un des créoles portugais d'Afrique. A la différence du créole cap-verdien et des créoles portugais du Golfe de Guinée (forro, angolar, lung'Ie et fa d'Ambu), il s'agit d'un créole continental, ce qui a des conséquences sur les contacts auxquels il est soumis.
Il est fort probable que jusque -au moins- au 18e siècle il était présent plus au nord en particulier sur la Petite Côte sénégalaise. Traditionnellement, on lui attribuait deux usages, d'une part celui de lingua franca du commerce côtier, d'autre part celui de marqueur identitaire fort des communautés dites luso-africaines, surtout regroupées à partir de la fin du 16e siècle dans et autour des "villes" (Cacheu, Geba, Farim, Ziguinchor, puis Bissau). A partir du début du 20e siècle et jusque dans les années 80, la situation s'est simplifiée. Schématiquement, on distingue :

— des variations régionales : kriol de Cacheu, de Bissau, de Geba, de Ziguinchor... En Guinée le kriol tel que parlé à Bissau prend progressivement pour des raisons politiques et démographiques l'ascendant. En Casamance, comme conséquence de l'indépendance du Sénégal et de la wolofisation de la région, le créole entame un déclin certain qui se traduit à la fois par une perte importante de locuteurs et la généralisation (au moins à Ziguinchor) de façons de parler où alternent français/wolof/kriol (Rougé, 2013) (Rougé, 2010) (Nunez, 2015).
— deux manières de parler le kriol (variation situationnelle) selon qu'on veut se faire comprendre du plus grand nombre ou de la communauté réduite (*kriol lebi* - "créole léger" v.s. *kriol fundu* - "créole profond").

Dans les années 90, divers événements - politiques, démographiques, etc. - en Guinée-Bissau Casamance ont créé une situation nouvelle qui a eu des conséquences linguistiques, en particulier en ce qui concerne la vie du kriol. C'est dans ce nouveau cadre que se place le projet "corpus kriol". Les travaux descriptifs depuis Marques Barros (1908) des différentes variétés de ce kriol ne manquent pas ((Doneux et Rougé, 1988) (Kihm, 1994) (Scantamburlo, 1999) . Elles ne prennent pas en compte la situation linguistique complexe engendrée par cette nouvelle situation.

2.2 Le contexte actuel

2.2.1 En Guinée

La situation du kriol en Guinée Bissau est le fruit d'éléments externes qui se sont développés à partir de l'indépendance (1973). Le premier constat est d'ordre démographique ; la population de la Guinée Bissau a plus que doublé en trente ans (estimée autour de 750.000 habitants en 1980, 1.759 159 en 2010) Ce qui est dû d'une part à une forte natalité, mais aussi à l'arrivée importante de migrants qui n'est pas vraiment prise en compte dans les chiffres officiels. Il s'agit surtout de guinéens le plus souvent locuteurs du peul et de commerçants sénégalais dont la présence a changé considérablement le "paysage" : ainsi, sur le marché de Bandim, grand marché de Bissau, de nombreuses boutiques s'ornent de portraits d'Amadou Bamba, fondateur de la confrérie sénégalaise des Mourides, et on y entend parler wolof.
Avec l'indépendance (1973/1974), une administration dont le kriol est de fait la langue orale, s'est installée dans tout le pays, des fonctionnaires des villes ont pris leurs fonctions à l'intérieur du pays. En parallèle, un fort exode rural a vu les populations urbaines - et en particulier à Bissau, la capitale, dont la population, qui aurait doublée entre 1991 et 2007, est estimée à 400.000 habitants - prendre un poids considérable. Cependant il faut aussi prendre en compte qu'à d'autres moments des mouvements inverses ont pu se produire, comme pendant la guerre civile de 1998, où une grande partie de la population de la ville s'est réfugiée à l'intérieur du pays pour fuir combats et bombardements. Tous ces mouvements de population ont bien évidemment renforcé les contacts linguistiques.
La démocratisation et/ou la massification de l'enseignement dont la langue est le portugais a eu pour conséquences paradoxales, en raison du manque de formation des maîtres, le développement du kriol à l'intérieur du pays et celui des interférences kriol/portugais.
Le développement des radios puis de la télévision et de leur espace de diffusion a grandement contribué à la diffusion du kriol. Au niveau langagier la première conséquence de cet état de fait est l'évolution très nette du paysage linguistique en une trentaine d'années. Si l'on considère le tableau (Tableau 1) élaboré à partir des recensements de 1979 et de 2009, outre la progression du pulaar conséquence évidente des migrations depuis la Guinée Conakry, on constate le doublement de la population créolophone. Les recensements ne donnent pas de renseignements suffisamment fins pour faire le tri entre créole parlé comme première langue et comme deuxième langue, on peut quand même estimer que ces chiffres correspondent à deux grandes tendances : a. l'augmentation du nombre des enfants qui ont comme langue première de socialisation, voire comme langue unique, le kriol et b. le développement du kriol comme L2 dans des populations (peuls, mandingues, mais aussi balante) chez lesquelles il était peu pratiqué et, par voie de conséquence, l'apparition de nouveaux contacts linguistiques.

	1979	2009
Kriol	44,3%	84,11%
Balante	24,5%	20,89%
Pulaar	20,3%	26,48%
Mandinka	10,1%	13,69%
Mandjaku	8,1%	7,72%
Pepel	7,2%	8,42%

TABLE 1 – Recensements de 1979 et 2009

Le poids pris par la capitale, par le développement des médias et de leur diffusion à l'intérieur du pays a provoqué la disparition presque totale des anciennes façons de parler régionales (créole de Cacheu, créole de Geba) au profit de celles de Bissau et en particulier des formes hautement lusitanisées.
Le recours au portugais a toujours été présent dans les productions en créole bissau-guinéen. Avec la lutte de libération, puis l'indépendance, cette influence a été renforcée en particulier à cause de l'utilisation du créole pour aborder des thèmes qu'il prenait rarement en charge (politique, administration, etc.) et du développement de la radio. Deux exemples illustrent bien ce que sont ces façons de parler qui ont pu être appelées kriol di radio. Dans le premier, il s'agit d'abord et avant tout d'alternances codiques nettement identifiables pour un créolophone (le créole est en italique, le portugais en gras)

(1) Kamarada ki na sukuta, no na pidi bo **atenção** *pa um komunikadu di* **Conselho da Revolução** *na* **voz** *di si* **responsavel maximo Comandante de Brigada, João Bernardo Vieira**, *Nino. Ma antes tudu no misti konta kamaradas kuma kil* **mensagem** *k'e obi ba na bos di Rafael Barboza i foi* **um lapso**. *I gosi no na pidi bo* **especial atenção** *pa sukuta palabra di* **chefe máximo di Conselho da Revolução, Comandante de brigada João Bernardo Vieira** (Radio Nacional de Guiné Bissau 15/11/1980)
Camarades qui écoutez, nous demandons votre attention pour un communiqué du Conseil de la Révolution par la voix de son plus haut responsable, commandant de brigade Joãoo Bernardo ieira, Nino, mais avant tout nous voulons dire aux camarades que le message qu'ils ont entendu par la voix de Rafael Barbosa était une erreur. Et maintenant nous vous demandons votre spéciale attention pour écouter les paroles du chef le plus haut du conseil de la révolution, commandant de Brigade João Bernardo Vieira.

Dans le second, cité par Kihm (1994) mais datant à peu près de la même époque, tous les mots peuvent être considérés comme kriol, mais la structure quant à elle est celle du passif portugais, avec présence d'une part d'un auxiliaire et d'autre part de l'expression de l'agent, structure inconnue du kriol "traditionnel".

(2) Kil asasinus yera ba komandadu pa un branku (RNGB in Kihm, 1994)
dem assassins être passé commandé par indef blanc
'ces assassins étaient commandé par un blanc'

Ce qui constitue une nouveauté ce n'est donc pas ces formes de créoles, ni même leur développement prévisible, mais le fait qu'elles se développent chez des locuteurs qui ne connaissent que ces variétés de kriol, tout en parlant par ailleurs différentes langues de la région. La majorité des nouveaux locuteurs sont exposés au kriol des médias et, plus généralement, aux productions influencées par le portugais et par ailleurs, ils apprennent souvent en même temps le kriol et le portugais, éprouvant des difficultés à différencier les deux langues.
Dans l'exemple qui suit, un néo-locuteur du kriol produit une phrase comprenant à la fois du portugais, du kriol traditionnel et du peul :

(3) I ka **explorado** *konno* i ka tene bon governu (Corpus 2014 - Marabout peul Bissau)
3sg neg exploité(Ptg) comme(Pl) 3sg neg avoir bon gouvernement(Ptg)
ce n'est pas exploité parce qu'il n'y a pas un bon gouvernement

Couto et Embalo (2010) présente la réalité linguistique en Guinée comme un continuum linéaire de variétés : portugais <-> portugais créolisé <->créole lusitanisé <-> kriol traditionnel <->kriol nativisé <-> langues natives

Compte tenu de ce que nous venons de présenter, cette description de la situation linguistique de la Guinée est totalement idéaliste. Sortir de ce genre de simplifications est entre autres un des objectifs du projet corpus kriol.

2.2.2 En Casamance

La cession de la Casamance à la France en 1886 avait eu deux effets sur le créole. Il s'était d'une part retrouvé isolé du monde lusophone et avait continué son évolution au contact du français et des langues parlées dans la région. D'autre part, la politique du colon français visant le déclassement de la bourgeoisie luso-africaine de Ziguinchor et son assimilation au reste de la population a produit une certaine unification du kriol casamançais. Après l'indépendance du Sénégal, le kriol a d'une part perdu son statut de langue véhiculaire au profit du wolof et par ailleurs il subit une forte influence du français qui se manifeste par de fortes doses d'alternance codique dans les discours et par des interférences grammaticales (Rougé, 2013). Dans l'exemple qui suit on notera, outre les alternances codiques nombreuses, l'utilisation de kuma avec le sens du français 'comme' en lieu et place du créole suma.

(4) I kel chapelle ali ki purmedu grisya di Sigicor
- Mais i chapelle o be grisya ?
- I chapelle pabia kel tenpu **kuma** I ka ten ba gintis… kriston ka cu ban
- Mais le magasin kuma k'e panga-l ? C'est avec des voûtes, si bu oja-l, on dirait grisya ou bien i chapelle k'e panga. C'est avec des voûtes, on dirait une église ou bien c'est une chapelle ? (Corpus 2011 Carvalho. Zigguinchor)
C'est cette chapelle, ici, qui est la première église de Ziguinchor/Mais c'est une chapelle ou une église ?/ C'est une chapelle parce que à cette époque comme il n'y avait pas de gens… Les chrétiens n'étaient pas nombreux/ Mais le magasin, comment l'ont-ils construit ? C'est avec des voûtes, on dirait une église ou bien c'est une chapelle ?

Depuis une vingtaine d'année, la situation sociopolitique ainsi que des aménagements infra structuraux ont permis le renforcement des relations entre Casamançais et Bissau-Guinéens. En 1998 la guerre civile en Guinée-Bissau a conduit de nombreux guinéens à franchir la frontière. Certains ont même fixé leur résidence non seulement à Ziguinchor mais aussi dans toute la Casamance. La réfection de la route entre Ziguinchor et Bissau et la construction de deux ponts pour remplacer les bacs a permis de vivifier le commerce entre les deux villes et de recréer des réseaux familiaux et sociaux. L'ouverture d'une université à Ziguinchor a aussi attiré de

nombreux jeunes Guinéens. Ainsi, s'est formée en Casamance une communauté pour laquelle la norme est l'utilisation d'un parler bilingue portugais-créole avec tout ce que cela comporte d'alternances et d'interférences. On remarquera que ces façons de parler sont aussi celles des émissions des radios guinéennes reçues toujours plus en Casamance. Outre la présence massive de populations pratiquant le créole "comme en Guinée", on constate l'émergence de nouvelles façons de parler le créole, caractéristiques de locuteurs de moins de cinquante ans, où se mêlent kriol traditionnel casamançais, français, parfois wolof, et aussi kriol guinéen moderne. L'influence du portugais est particulièrement remarquable pour des personnes ne parlant pas et n'ayant pas appris cette langue. La formation d'un mot comme *voyagem* - pour 'voyage' en kriol *byas* (formé à partir du français voyage et du préfixe portugais -agem) est emblématique de ces façons de parler dans certaines communautés casamançaises.
Ainsi, se côtoient en Casamance à la fois le créole traditionnel (particulièrement bien conservé dans les villages alentours de Ziguinchor : Adéans, Sindone...), des formes francisées de ce même créole, le kriol tel que parlé en Guinée Bissau, et les formes "hybrides" qui viennent d'être évoquées.

Pour résumer, de nouvelles "dynamiques créoles" ont vu le jour ces dernières année, des deux côtés de la frontière, ce qui pose à la fois la question de la redynamisation démographique de cette langue et celle du changement linguistique. L'ensemble linguistique kriol de Guinée et de Casamance se présente donc comme une multitude de façons de parler au contact à la fois de langues africaines, du portugais moderne (et dans le cas de la Casamance du français).
C'est dans ce contexte, que naît le projet "corpus kriol", dans le cadre d'une coopération entre des chercheurs de Coimbra , d'Orléans et de Ziguinchor.

3 Description du projet

Ce projet vise plusieurs objectifs complémentaires. Pour les linguistes, il s'agit de constituer une ressource de documents authentiques (enregistrements d'oral spontané) qui permette une analyse de la situation (socio-)linguistique des deux côtés de la frontière et aussi une analyse du changement linguistique à partir de productions authentiques attestées. Pour cela, il s'agit de rendre disponible les enregistrements glosés (Leipzig rules ou Universal POS tags `http://universaldependencies.org/u/pos`) et leur traduction en portugais et en français (corpus parallèle).
Du côté de l'outillage (versant plus informatique que linguistique), il s'agit de rendre possible la consultation de ressources complexes car comprennant de nombreuses interférences entre langues (corpus plurilingue et non multilingue). Des métadonnées précises OLAC Dublin Core enrichies de données locuteurs) ainsi qu'un repérage efficace (même sommaire) des alternances codiques est nécessaire. Cela nécessite (pour l'instant) un fort traitement manuel (transcription, annotation, traduction).

3.1 Enregistrements

Nous avons 47 enregistrements (effectués par J.-L. Rougé) d'une durée allant de 10min à plus d'une heure, pour un total de 25 heures environ. Ces enregistrements ont été effectués avec un enregistreur numérique Marantz sur plusieurs terrains : en Casamance, à Bissau et à Orléans (2

enregistrements). Comme il s'agit d'enregistrements de discours spontané, ils ont tous les défauts des enregistrements de terrain : bruits de fond, interruptions, etc.
Les enregistrements ayant vocation a être librement accessibles sur la plateforme CRDO-COCOON, l'équipe sélectionne les enregistrements diffusables sans restrictions de ceux qui resteront en accès réservé (certains enregistrements contiennent des opinions ou des informations qui ne doivent pas être rendues publiques).

3.2 Méthode

Le projet étant multi-sites (Orléans, Coimbra et Ziguinchor), la compatibilité et l'homogéneité des pratiques entre les équipes est un point essentiel. Cela repose sur la rédaction de guides (guidelines) précis mais aussi sur l'utilisation d'outils communs.

Pour l'instant, les transcriptions sont effectuées à l'université de Coimbra par un locuteur natif du kriol, sur un financement du CELGA-ILTEC. Le logiciel Transcriber (Barras *et al.*, 2001) est utilisé pour la transcription. Un guide d'annotation a été rédigé conjointement par les équipes du CELGA-ILTEC et du LLL. Il est probable que des transcriptions soient faites en Casamance par la suite. Pour l'instant, il est convenu de passer de Transcriber à un outil tel que ITE (`http://michel.jacobson.free.fr/ITE/`) pour les gloses juxtalinéaires. Cela demande quelques manipulations sur les fichiers, ce qui ne favorise pas l'autonomie des linguistes qui ne disposent pas de savoir-faire en informatique. Nous souhaitons donc favoriser l'usage de chaînes de traitement des données au sein d'une même plateforme si c'est possible, afin de diminuer les traitements sur les fichiers. Le passage de Transcriber à trjs (`https://github.com/christopheparisse/trjs`) est en test. Ce logiciel est conçu comme un outil utilisant la plupart des caractéristiques de Transcriber et ELAN.
Il est à noter que la segmentation du signal sonore se fait en fonction des groupes de souffle et non en fonction des 'phrases'. Ce choix a été fait pour des raisons d'efficacité et de rendement dans les transcriptions mais il pose bien entendu des problèmes (solubles à notre avis) dans les traitements ultérieurs (parsing notamment).
L'utilisation des balises de commentaire pour les annotations des événements (bruits, prononciations particulières, etc.) se fait conformément au guide d'utilisation de Transcriber mais cela pose problème également pour les exports (comme cela a été les cas pour l'annotation du corpus ANCOR par exemple (Muzerelle *et al.*, 2014)).
Par ailleurs, nous comptons sur l'intégration d'un outil de glose (similaire à ITE `http://michel.jacobson.free.fr/ITE/`) sur la plate-forme TXM (Heiden *et al.*, 2010) pour la tâche de glose du corpus.

4 Perspectives

Nous avons décrit jusqu'ici ce que nous commençons à faire dans ce projet, en partant des possibilités offertes grâce aux compétences cumulées dans les équipes qui constituent ce projet. Cependant, nous pouvons distinguer deux objectifs :

- sauvegarder les 25 heures d'enregistrements, les traiter et les mettre à disposition,
- lancer de nouvelles études sur le kriol en utilisant les techniques de recueil développées dans le cadre d'autres projets, tels que AIKUMA (Bird *et al.*, 2014) et BULB (Adda *et al.*,

2016). Ces projets nous paraissent particulièrement intéressants mais posent des questions sur leur utilisation sur nos données. Parmi ces questions figure la fidélité entre la répétition et la source. En effet, le recours à des locuteurs natifs pour effectuer des transcriptions (ou des répétitions) dans la situation complexe de contact de variétés, demande une attentionaux particularités sociolinguistiques. L'expérience du terrain nous a montré que dans ce type de situation le transcripteur (ou le répétiteur) a souvent une tendance prononcée à réinterpréter dans sa variété propre ou dans la norme à laquelle il se réfère le texte entendu. Par exemple, un Bissau-Guinéen répétera dans la variété guinéenne un enregistrement recueilli auprès de locuteurs casamançais et ainsi /pode/ deviendra /pudi/ '(il) peut', /meste/ deviendra /misti/ 'vouloir', etc. Cela demande à être pris en compte.

Par ailleurs, nous envisageons l'exploitation des données sous l'angle du contact par l'utilisation d'un schéma d'annotation tirant profit de l'expérience du projet CLAPOTY (Vaillant et Léglise, 2014). Nous retenons en particulier l'idée de ne pas assigner une langue donnée à un passage (v. les exemples donnés plus haut) mais de laisser la possiblité que le passage appartienne à plusieurs langues ou constitue une unité 'flottante' entre plusieurs langues. En effet, le repérage des phénomènes de contact par un jeu d'annotation intégrant la dimension plurilingue (et non multilingue) du corpus est certainement le défi majeur à relever afin de faire progresser les études sur les créoles (créolistique).

Remerciements

Ce projet bénéficie de financements de la part du CELGA-ILTEC, du LLL et du GDRI SEEPiCLa.

Références

ADDA, G., STÜKER, S., ADDA-DECKER, M., AMBOUROUE, O., BESACIER, L., BLACHON, D., BONNEAU-MAYNARD, H., GODARD, P., HAMLAOUI, F., IDIATOV, D. *et al.* (2016). Breaking the unwritten language barrier : The bulb project. *Procedia Computer Science*, 81:8–14.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

BIRD, S., HANKE, F. R., ADAMS, O. et LEE, H. (2014). Aikuma : A mobile app for collaborative language documentation. *In Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5.

DONEUX, J. L. et ROUGÉ, J.-L. (1988). *En apprenant le créole à Bissau ou Ziguinchor*. Editions L'Harmattan.

HEIDEN, S., MAGUÉ, J.-P. et PINCEMIN, B. (2010). Txm : Une plateforme logicielle open-source pour la textométrie-conception et développement. *In 10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

KIHM, A. (1994). *Kriyol syntax : The Portuguese-based creole language of Guinea-Bissau*, volume 14. John Benjamins Publishing.

MUZERELLE, J., LEFEUVRE, A., SCHANG, E., ANTOINE, J.-Y., PELLETIER, A., MAUREL, D., ESHKOL, I. et VILLANEAU, J. (2014). Ancor_centre, a large free spoken french coreference corpus :

description of the resource and reliability measures. *In LREC'2014, 9th Language Resources and Evaluation Conference.*, pages MUZERELLE14–150.

NUNEZ, J. J. F. (2015). *L'alternance entre créole afro-portugais de Casamance, français et wolof au Sénégal : une contribution trilingue à l'étude du contact de langues*. Thèse de doctorat, Université Sorbonne Paris Cité.

ROUGÉ, J.-L. (2010). Parler créole à ziguinchor au xxième siècle. *Sciences & Techniques du Langage-CLAD*, 7:75–87.

ROUGÉ, J.-L. (2013). Créole de casamance. émergence de nouvelles variétés. *Travaux-Cercle linguistique d'Aix-en-Provence*, 24:201–212.

SCANTAMBURLO, L. (1999). Dicionário guineense-português. *Lisboa : Universidade Nova de Lisboa*, 1.

VAILLANT, P. et LÉGLISE, I. (2014). À la croisée des langues. annotation et fouille de corpus plurilingues. *Revue des Nouvelles Technologies de l'Information*, pages 81–100.