Collecte de données orales pour la langue sereer

Floriane Brennus[1,1] *, Laurent Besacier*[1,2] *Sylvie Voisin*[1,3]
(1) Master IDL, Université Grenoble Alpes
(2) Laboratoire LIG, Université Grenoble Alpes
(3) Laboratoire DDL, Université Lumière Lyon 2
floriane.brennus@etu.univ-grenoble-alpes.fr,
laurent.besacier@univ-grenoble-alpes.fr,
sylvie.voisin@ish-lyon.cnrs.fr

RÉSUMÉ

La langue sereer est la troisième langue du Sénégal. Elle possède ses différentes variantes recensées au nombre de cinq. Nous nous proposons de collecter des données orales pour la variante la plus présente chez nos locuteurs. Nous utilisons pour cela l'application mobile LIG-Aikuma, qui permet de faire éliciter de la parole aux locuteurs, à partir de vidéos (nous utilisons pour cela le corpus *trajectoires*). Cet article vise à décrire la méthodologie employée pour cette collecte de données sur la langue sereer, puis à présenter un bilan des données orales collectées.

Abstract

Data collection for the sereer language

The Sereer language is the third language of Senegal. It has five different variants. We propose to collect oral data for the most common variant among our speakers. For this we use the mobile application LIG-Aikuma, which allows speakers to elicit speech from videos (we use a video corpus called *trajectoires*). The purpose of this paper is to describe the methodology used for this oral data collection in Sereer, and then to present an overview of the spoken data collected so far.

RÉSUMÉ EN LANGUE NATIONALE

Traduction en langue nationale du titre

[1]

MOTS-CLÉS : sereer, LIG-AIKUMA, collecte, enregistrements, transcription, documentation des langues

KEYWORDS : sereer , LIG-AIKUMA, collecting, recordings, transcription, language documentation

[1] ici sera prévu dans la version finale un résumé en wolof

1. Introduction

Cet article vise à décrire la méthodologie employée pour une collecte de données en langue sereer. Nous utilisons pour cela l'application mobile LIG-Aikuma, qui permet de faire éliciter de la parole aux locuteurs, à partir de vidéos (nous utilisons pour cela le corpus *trajectoires*). Un bilan des données orales collectées est aussi présenté dans cet article.

1. Description de la langue sereer

La langue sereer est la troisième langue du Sénégal. Cette langue possède différentes variantes, on en recense cinq, d'après Waly Faye 1979, Mc Laughlin 1992, et Renaudier 2012: le sereer Siin également appelé sereer Jaxaaw ou Singandum, une variante standard parlée au Centre-Ouest du Sénégal, dans la région de Sine-Saloum, le sereer a'ool, une variante parlée dans la région du Baol qui, du fait de sa proximité géographique avec Dakar (la capitale), subit une forte influence du wolof; le sereer Jegem ou sereer de la Petit-Côte, parlé à Mbour et à Joal, le sereer de Fadiouth/Palmarin (Fadiouth est une petite île rattachée à la ville de Joal), et enfin, le sereer Nyomiñka, parlée dans le Saloum, que Renaudier va requalifier de « parler Mar Lodj » car selon elle, le terme Nyomiñka permettait surtout de désigner une population sereer qui vit de la pêche (même si au cours de ses recherches, elle a découvert que la population sereer de l'île de Mar ne vivait plus de la pêche).

1.1. Inventaire de phonèmes

1.1.1. Les phonèmes vocaliques

Le système vocalique de la langue sereer est composé de cinq timbres vocaliques ([a], [i], [u], [o], [e]) qui s'opposent selon leur trait de longueur. Le système vocalique du sereer est donc un système à dix phonèmes.

1.1.2. Les phonèmes consonantiques

Le système consonantique du sereer comprend une trentaine de phonèmes, présentés dans la figure 1.

p	t	c	k	q
b	d	j	g	
ɓ	ɗ	ʃ		ʔ
f	r	s	h	x
w	l	y		
m	n	ɲ	ŋ	
mb	nd	nj	ng	

Figure 1: système consonantique du sereer

Sur la figure 1, on remarque la présence de consonnes implosives voisées: [ɓ], [ɗ], [ʄ]. Certaines variétés de sereer observent une opposition entre les consonnes implosives sonores et les consonnes implosives sourdes. On peut également constater la présence de la consonne prénasale uvulaire voisée [nG] dans certains dialectes du sereer.

1.2. La morphologie de la langue sereer_

La consonne initiale d'un lexème peut varier selon les motivations grammaticales, notamment selon la classe nominale (M. Renaudier, 2015). Les consonnes varient d'une manière spécifique. En effet, chaque consonne fait partie d'un ensemble de trois consonnes, et chacune de ces trois consonnes s'intègre dans un degré.

La figure 2 nous montre comment s'alternent ces consonnes en fonction de leur degré.

III	mb	nd	nj	ng	ŋg	mb	mb	nd	nj	ng	ɓ	ɗ	ʄ	l	y	m	n	ñ	ŋ
II	p	t	c	k	q	p	b	t	c	k	ɓ	ɗ	ʄ	l	y	m	n	ñ	ŋ
I	f	r	s	g	x	b^2	w	d	j	h	ɓ	ɗ	ʄ	l	y	m	n	ñ	ŋ

Figure 2: Alternances consonantiques du sereer

Ces alternances de la consonne initiale d'un lexème dépend des conditions suivantes:

-la classe du nom (dérivé ou non),

-l'accord en nombre du verbe avec le sujet,

-l'accord de l'adjectif avec le nom.

Ce système d'alternances consonantiques s'observe aussi bien sur les noms, que sur les verbes.

Ces alternances consonantiques sont dues à la classification mais également à la dérivation. Le choix du degré d'alternance dépendra ainsi du choix de la classe, mais également de l'opposition singulier/pluriel, ainsi que les diverses possibilités transcatégorielles. Comme nous l'expliquent K. Pozdniakov et G. Segerer (2006), ces mécanismes sont le plus souvent interdépendants. En effet, pour une dérivation quelconque, cela implique l'affectation à la base lexicale d'une marque de classe, ce qui entrainera inévitablement le choix d'un degré d'alternance. Cependant, il est possible que ce choix se trouve être en contradiction avec le degré d'alternance requis par l'opération de dérivation elle-même. Ce système s'avérant très complexe, nous ne le détaillerons pas ici.

2. Méthode de collecte

La collecte de données a été effectuée grâce à l'application LIG-AIKUMA[2]. Tout d'abord, AIKUMA est une application mobile développée à partir de 2013 par Hanke et Bird. Elle ne fonctionne que sur un système d'exploitation Android. Cette application possédait les fonctionnalités suivantes: enregistrement de parole spontanée (les locuteurs s'enregistrent eux même via des smartphones), le respeaking (qui permet d'effectuer un nouvel enregistrement à partir d'un enregistrement bruité ; les deux enregistrements étant alignés) et la traduction (la traduction effectuée est alignée avec l'enregistrement dans la langue source).

LIG-AIKUMA a apporté des fonctionnalités supplémentaires, telle que l'affichage automatique de formulaires de consentement, la constitution du profil du locuteur au début de chaque enregistrement (ce profil sera par la suite enregistré dans l'application), l'élicitation de parole à partir d'images, de textes ou de vidéos, la fonctionnalité respeaking a été repensée afin de pouvoir sectionner des segments de paroles à répéter, tout en restant aligné avec l'enregistrement de base. Enfin, l'ajout de la fonctionnalité vérification permet de corriger du texte correspondant à des transcriptions de signaux par exemple.

Dans ce projet, deux fonctionnalités ont été utilisées principalement: celle d'élicitation de parole à partir de vidéos, et celle de traduction. Les vidéos sont des vidéos fournies par Sylvie Voisin, sur le thème de la trajectoire. Ces vidéos ont déjà été utilisées dans le cadre du projet BULB (Breaking the Unwritten Language Barrier). Il y a au total, 76 vidéos, d'une durée de quelques secondes chacune. La première partie de l'enregistrement est consacrée à l'élicitation de parole. Le locuteur va donc visionner chaque vidéo et la décrire simultanément en langue sereer. Pour chaque vidéo correspond alors un fichier son en langue sereer. La seconde phase d'enregistrement consiste à traduire en français chaque fichier audio produit en langue sereer grâce à la fonctionnalité traduction. Les fichiers traduits seront alignés avec le fichier audio en langue sereer correspondant, et donc, avec la vidéo qui a suscité cette élicitation de parole.

Pour chaque enregistrement effectué avec un locuteur, nous avons tenu un journal de bord afin de rendre compte de leur déroulement.

---

2 https://lig-aikuma.imag.fr/

Les enregistrements ont eu lieu principalement sur Lyon. Pour trouver les locuteurs, nous avons tout d'abord cherché du côté de notre entourage (belle-famille sénégalaise, amis sénégalais). Est-ce qu'ils connaissaient d'autres personnes sénégalaises, plus spécifiquement des personnes sereer, dans la région? Pouvions-nous les contacter pour leur parler de notre projet, dans le but de les faire participer? Grâce à cette méthode, nous avons pu être en contact avec un locuteur, malheureusement trop loin géographiquement, mais il nous a mis en contact avec son oncle habitant sur Grenoble. Nous avons aussi contacté l'association des étudiants sénégalais de Lyon et de Grenoble. Celle de Grenoble n'avait malheureusement pas de contacts sereer. Cependant, le secrétaire général de celle de Lyon est sereer. Nous avons ainsi pu rencontrer trois locuteurs. Nous avons aussi cherché des locuteurs dans des restaurants sénégalais, et contacté d'autres associations de sénégalais ou de sereer. Certaines ont répondu en paraissant intéressées mais n'ont pas donné suite et d'autres n'ont à ce jour, pas encore donné suite. Aujourd'hui nous possèdons des enregistrements complets via la tablette et LIG-AIKUMA, pour 5 locuteurs.

2.1. Journal de bord pour les enregistrements complets effectués

2.1.1. Enregistrements avec trois locuteurs (de Popelguine [382], de Dakar [381], et de NGuénième [380]), au laboratoire DDL

Ces enregistrements ont pu avoir lieu grâce à l'association des étudiants sénégalais de Lyon contactés via un réseau social. Le paragraphe ci-dessous présente quelques notes extraites du journal de bord de la première auteur de cet article.
"

_382 : Nous avons effectué l'enregistrement seuls dans la salle. J'ai expliqué le déroulement de l'enregistrement. Puis l'enregistrement débute. Tout se passe bien. Il a juste une difficulté pour une vidéo, où le locuteur ne connaît pas le terme « escaliers » parce que d'après lui "il n'y a pas de mot pour ce que le peuple sereer n'a pas inventé. ». Je le remets en confiance et il termine l'enregistrement sans que j'ai à intervenir. Vient ensuite la deuxième passation, pour la phase traduction. Là aussi, tout se déroule comme prévu du point de vue du locuteur. Cependant, dans l'application, pour la session traduction, lors de certaines élicitations, l'application quittait automatiquement.

_381 : Les deux locuteurs suivant nous ont ensuite rejoint dans la salle. Chacun a rempli son questionnaire socio-linguistique ainsi que son formulaire de consentement. L'enregistrement du locuteur 381 s'est également bien passé. Au début, dans un souci de bien faire, il décrivait la vidéo dans une position de narrateur, avec un sereer plus « standardisé ». Ses amis, présents dans la salle, l'ont repris, et lui ont expliqué qu'il devait dire ce qu'il voyait sur la vidéo normalement, comme s'il parlait à quelqu'un, sans chercher à vouloir faire des phrases sophistiquées. L'enregistrement a donc repris et il a réussi à se mettre rapidement dedans. Ses amis étaient bienveillants et tenaient à ce qu'il parle avec un niveau plus élevé (dans le souci que je puisse exploiter au mieux mes données). La phase traduction s'est avérée plus compliquée, il ne cessait de répéter que c'était difficile de traduire du sereer au français et il demandait de l'aide à ses amis, ou cherchait leur approbation. Comme pour le locuteur 382, nous avons eu ce problème de bug de l'application pour quelques vidéos.

_380 : Avant de débuter la passation de ce locuteur, j'ai demandé à ses amis s'ils voulaient rester dans la salle ou non. Je craignais qu'ils se lassent et perturbent l'enregistrement. Ils sont finalement restés et tout s'est très bien déroulé. J'ai de nouveau expliqué le déroulement de l'enregistrement, bien qu'il était présent à mes explications pour le locuteur 381, puis nous avons commencé. Tout s'est très bien passé, tout comme la phase de traduction, toujours avec ce bug cependant."

2.1.2. Enregistrement d'un locuteur de Moundé (383), au LIG, Grenoble :

Le paragraphe ci-dessous présente d'autres notes extraites du journal de bord de la première auteur de cet article pour les enregistrements du locuteur 383.
"Il s'agit de l'oncle d'un locuteur potentiel résident à Paris. Je ne pouvais l'interroger au vu de la distance. Il m'a donc confié les coordonnées de son oncle, avec lequel j'ai pu entrer en contact facilement et qui acceptera volontiers de participer.
Malheureusement, ce locuteur est arrivé avec un retard d'une trentaine de minutes. De plus, il avait une contrainte de temps, il devait être reparti pour 17h30. Ce qui a eu des conséquences sur les conditions d'enregistrement. Il était très pressé, pour chaque vidéo, il écourtait l'élicitation de parole, il arrêtait l'enregistrement sans avoir finit de parler. Je lui ai demandé plusieurs fois de faire attention, sans succès. Ensuite, la tablette a eu un problème quelques minutes après le début de l'enregistrement. Problème que j'ai dû régler en éteignant rallumant la tablette.(il était toujours aussi pressé). L'étape de traduction n'a donc pu avoir lieu. A la fin de l'enregistrement, conscient de l'échec de la session, il a proposé naturellement de renouveler l'enregistrement lorsqu'il serait plus disponible, chose que j'ai acceptée. Nous nous sommes donc vus une autre fois afin d'effectuer un nouvel enregistrement. Cette fois-ci, j'ai décidé de gérer la tablette moi même pendant la phase élicitation de parole afin de ne pas avoir les même problèmes que lors de notre précédent entretien . Il s'est avéré être un locuteur très disponible, nous nous sommes vus trois vendredi consécutifs afin qu'il m'aide dans mes débuts d'annotations de corpus."

2.1.3. Enregistrement d'une locutrice de Montélimar (384) au laboratoire DDL

Le paragraphe ci-dessous présente d'autres notes extraites du journal de bord de la première auteur de cet article pour le enregistrements du locuteur 384.
"Je suis rentrée en contact avec elle via des connaissances qui m'ont aidé à chercher des locuteurs sereer. Elle leur a confié être intéressée par le projet et c'est ainsi que j'ai obtenu ses coordonnées, et que j'ai pu prendre contact avec elle. Elle a tout de suite été très intéressée par le projet. L'enregistrement s'est très bien déroulé, malgré les soucis récurrents avec la tablette pendant la phase de traduction. Pendant l'enregistrement, je remarque qu'il s'agit d'une variété de sereer que je n'avais pas encore entendue. Il s'agit effectivement du sereer safène. (Ce n'est pas une variété de sereer, mais une autre langue canjin.) Je demande alors la traduction de mots que j'avais repéré en sereer sine, et je remarque que pour ces mots là, ils semblent très éloignés. Cette locutrice a également grandit dans un contexte autre que les autres locuteurs. En effet, elle a grandit en France, à Montélimar, en France. Cependant, ses parents lui parlaient dans leur langue maternelle: le sereer. Elle m'explique donc que son sereer, contrairement aux personnes vivant au Sénégal, ne peut être influencé par le wolof, puisqu'elle ne parle pas le wolof. "

2.2. Echecs rencontrés

Le paragraphe ci-dessous présente quelques notes extraites du journal de bord de la première auteur de cet article et concerne les problèmes rencontrés, les sessions d'enregistrmeent qui se sont mal passées.

2.2.1. Enregistrement avec une locutrice Sereer de Dakar

"Il s'agissait d'un cadre non formel, nous étions au domicile de la locutrice. L'enregistrement été délicat à mener. La personne est arrivée en retard. Dès son arrivée, elle me précise qu'elle ne sera disponible qu'une heure. Le temps me semble un peu court pour effectuer les enregistrements convenablement. Nous tentons malgré tout quitte à n'effectuer que la phase élicitation de parole. La phase de traduction aurait pu être effectuée auprès d'un autre locuteur, parlant la même variété de sereer. J'explique le déroulement de la passation. Commence donc l'enregistrement, je constate que la locutrice rencontre des difficultés, elle se repasse la vidéo plusieurs fois, peine à trouver ses mots. J'ai l'impression qu'elle fait une traduction du français vers le sereer. Elle fait plusieurs pauses me disant qu'il lui manque des mots, que si elle était en train de parler avec un autre locuteur sereer, elle dirait ces mots qui lui manquent en français directement. Nous avons ensuite été dérangée par son téléphone portable lui signalant un appel téléphonique auquel elle a répondu. Réduisant un peu plus le temps de passation réel. Ensuite elle me fait part de son malaise, qu'elle serait plus à l'aise à son retour du Sénégal dans trois semaines, car elle sera « replongée » dans sa langue maternelle pendant ces trois semaines, et qu'elle se serait de nouveau refamiliarisée avec le sereer. Elle m'avait déjà évoquée sa position particulière par rapport au sereer : c'est sa langue maternelle mais elle ne l'utilise que très peu. J'ai donc accepté de refixer un rendez-vous avec elle à son retour. Depuis son retour de Sénégal, j'ai eu beau la relancer, elle ne donnera pas suite à mes prises de contacts."

2.2.2. Enregistrement d'une locutrice de Pikine, à son domicile à Saint Priest :

"Enregistrement à son domicile, dans une résidence calme. Beaucoup de mal à se plonger dans la vidéo, elle posait beaucoup de questions dans la peur de « mal faire ». Par exemple: « Là je dis quoi ? Donc exemple, là je dis une femme lance un ballon à un homme ? . Est-ce que je dois décrire tout ce que je vois ? Si des mots ne me viennent pas ? Est-ce que je peux visionner toutes les vidéos d'abord avant de les traduire ? » Un début difficile et long qui a impacté le déroulement de l'enregistrement. Elle avait également une contrainte temporelle et devait s'absenter à une certaine heure. Elle m'a proposé de rester chez elle le temps de son rendez-vous, puis nous avons terminé la partie traduction. Pendant la phase traduction : elle envoyait des messages, elle changeait la conversation. Malheureusement, lors d'une mauvaise manipulation, j'ai perdu toutes les données relative à cette locutrice."

2.3. Bilan et leçons après ces enregistrements

La recherche de locuteurs sereer en métropole française s'est avérée compliquée. Il serait évidemment plus facile de trouver des locuteurs sereer au Sénégal afin de découvrir au mieux la langue sereer mais également la culture qui s'y rattache. Cependant, il n'est pas impossible de trouver des locuteurs sereer en France. Pour ce faire, nous nous sommes rapprochés de la population sénégalaise de la région Rhône Alpes. La diaspora sénégalaise est très présente en France. Il ne faut pas hésiter à contacter les associations, à aller dans des lieux de rencontres de personnes sénégalaises comme des restaurants par exemple. L'attitude à adopter est également importante, il faut savoir montrer un intérêt pour leur langue, leur culture. Ils sauront percevoir notre motivation, et se montrer également motivé pour nous aider par la suite. Les personnes face à nous pourront également connaître d'autres personnes sereer avec lesquelles nous pourrions être amenés à être en contact par la suite, pour des enregistrements. Il ne faut pas hésiter à demander s'ils connaissent d'autres personnes sereer, afin d'élargir notre champ de recherches.
La présentation de nos recherches est important dans la prise de contact, et l'attrait des locuteurs éventuels. Cependant, il faut veiller à ne pas en dire trop. Ne pas évoquer ce qu'on recherche exactement, pour ne pas orienter les entretiens. En effet, dans les soucis de bien faire, les locuteurs pourraient insister sur des formes orales qu'ils n'utiliseraient pas forcément dans des discours spontanés. Il faut donc expliquer les recherches sans les détailler totalement. Lors de l'enregistrement, il faudra opérer de la même manière pour l'explication des consignes. Il faut évidemment expliquer le déroulement de la session d'enregistrement au locuteur, afin qu'il sache réellement à quoi s'attendre. Ne pas chercher à lui mentir, bien lui expliquer qu'il y a 76 vidéos (même si le nombre peut sembler impressionnant), et préciser que les vidéos sont très courtes, afin d'éviter de susciter une certaine lassitude du locuteur. La première partie des enregistrements consiste à éliciter de la parole grâce à l'application LIG-AIKUMA. Il est préférable d'accompagner les explications de manipulations sur la tablette, puisque le locuteur sera généralement le détenteur de la tablette (sauf cas exceptionnel, où pour garantir la qualité de l'enregistrement, nous avons dû manipuler nous même la tablette à la place du locuteur).
Le cadre des enregistrements est également très important. Par peur de déranger les locuteurs, on pourrait vouloir proposer d'effectuer les enregistrements chez le locuteur. Cependant, nous ne pensons pas qu'il s'agisse de la solution idéale. Nous ne connaissons pas le domicile du locuteur, ni son environnement, cela pourrait compromettre les enregistrements d'une part. D'autre part, la formalité du cadre est importante. Le cadre non formel peut décrédibiliser le statut de chercheur. Selon notre propre expérience, en France, nous avons pu remarqué que le cadre formel d'une salle dans un laboratoire apportait une certaine légitimité à nos recherches, et une meilleure considération. Nous avons pu voir une nette différence entre les enregistrements effectués à domicile, et ceux effectués en laboratoire.

Le déroulement de nos enregistrements peut s'avérer assez laborieux pour le locuteur. Les 76 vidéos pour l'élicitation de parole sont classées selon un ordre aléatoire afin d'éviter une certaine habituation. De plus, des vidéos de distraction sont insérées afin de permettre au locuteur de ne pas comprendre le coeur de la recherche. À la fin de cette phase d'élicitation, nous avons mis en place une phase de traduction en français pour chaque fichier enregistré en sereer. Effectuer les deux phases en une session d'enregistrement peut s'avérer assez long. La durée sera alors d'une heure trente environ, ce qui demande un réel investissement de temps pour le locuteur, mais également un effort cognitif.
Avec le recul, la phase de traduction située directement après la phase d'élicitation peut avoir ses limites. Le locuteur sereer peut faire appel à ses souvenirs de la vidéo pour fournir une pseudo traduction la plus fidèle à la vidéo correspondante, et ne pas effectuer une traduction des propos réellement entendus en sereer. On pourrait résoudre ce problème en effectuant l'enregistrement de la traduction avec un nouveau locuteur, de même variété de sereer, qui lui n'aurait pas vu les vidéos.

3. Données collectées

3.1. Quantité de données collectées

Nous avons créé un dossier des fichiers audios pour chaque locuteur. Dans ce dossier se trouve les fichiers audio .wav ainsi que les fichiers .json correspondant. Chaque fichier son enregistré est automatiquement aligné avec la vidéo correspondante. De plus, toute traduction effectuée est également alignée avec le fichier d'élicitation de parole source. Tout se situe dans le nom de chaque fichier. Nous avons tout d'abord renommé tout nos fichiers afin de faciliter la reconnaissance des locuteurs, tout en préservant leur anonymat. Nous avons donc choisi de les renommer par « 38 » le département d'appartenance de notre faculté puis d'ajouter une numérotation. Mes locuteurs sont ainsi identifiés comme suit: 380, 381, 382, 383, 384. Nous avons créé un script afin d'extraire des statistiques quantitatives sur nos données.

La figure 3 montre les résultats obtenus :

Locuteurs	380	381	382	383	384	Totaux
Nombre de fichiers au total	128	131	140	159	151	709
Nombre de fichiers en français	52	56	66	83	76	333
Durée totale en français	4min 21s	5min 26s	5min 7s	6min 2s	3min 34s	24min 30s
Nombre de fichiers en sereer	76	75	74	76	75	376
Durée totale en sereer	5min 55s	4min 43s	4min 54s	8min 4s	4min 24s	28min 1s

Figure 3: Données enregistrées en sereer et en français selon chaque locuteur

Nous pouvons donc voir que le nombre total de fichiers en sereer est de 376 contre 333 en français. Cela s'explique par les complications que nous avons eu lors des enregistrements en phase traduction. En résumé, nous avons:
-pour le locuteur 380: 128 fichiers au total, pour une répartition de 76 fichiers soit 5min55sec en sereer, et 52 fichiers soit 4min21sec en français.
-pour le locuteur 381: 131 fichiers au total, pour une répartition de 75 fichiers soit 4min43sec en sereer, et 56 fichiers soit 5min26sec en français.
-pour le locuteur 382: 140 fichiers au total, pour une répartition de 74 fichiers soit 4min54sec en sereer, et 66 fichiers soit 5min07sec en français.
-pour le locuteur 383: 159 fichiers au total, pour une répartition de 76 fichiers soit 8min04sec en sereer, et 83 fichiers soit 6min02sec en français.
-pour le locuteur 384: 151 fichiers au total, pour une répartition de 75 fichiers soit 4min24sec en sereer, et 76 fichiers soit 3min34sec en français.

La durée totale de nos fichiers audios, en sereer et en français est de 52min et 51sec. La différence de nombre de fichiers en français par rapport au nombre de fichiers en sereer peut s'expliquer par les fichiers audios qui ne pouvaient être lus par la tablette lors de la phase traduction (lorsqu'il y a moins de fichiers en français qu'en sereer), mais également par le fait que lors de la phase traduction, certains locuteurs voulaient améliorer la traduction qui venait d'être effectuée (lorsque les fichiers traduits sont plus importants que les fichiers sereer). Lors de la phase d'annotation, nous avons donc conservé le fichier audio traduit le plus récent, qui correspond, pour le locuteur, à la meilleure traduction de ses propos.

3.2 Travaux en cours concernant la transcription du français

L'application LIG-AIKUMA permet un alignement automatique entre les fichiers sons produits en sereer, et leur traduction correspondante. Grâce à cela, la traduction en français pour chaque fichier audio sereer a ainsi pu être transcrite pour chaque fichier bénéficiant de sa traduction. Nous avons donc effectué ces transcriptions grâce aux logiciels Elan_CorpA[3] et Praat[4]. Ces transcriptions ont dû être effectuées manuellement, car un système de traitement automatique n'aurait pas pu traiter la parole de personnes francophones non natives.

La figure 4 nous montre un exemple de transcription sous Praat pour un locuteur sereer. La première tier correspond à la vidéo correspondant à cette élicitation de parole, la troisième tier correspond à la transcription de la traduction.

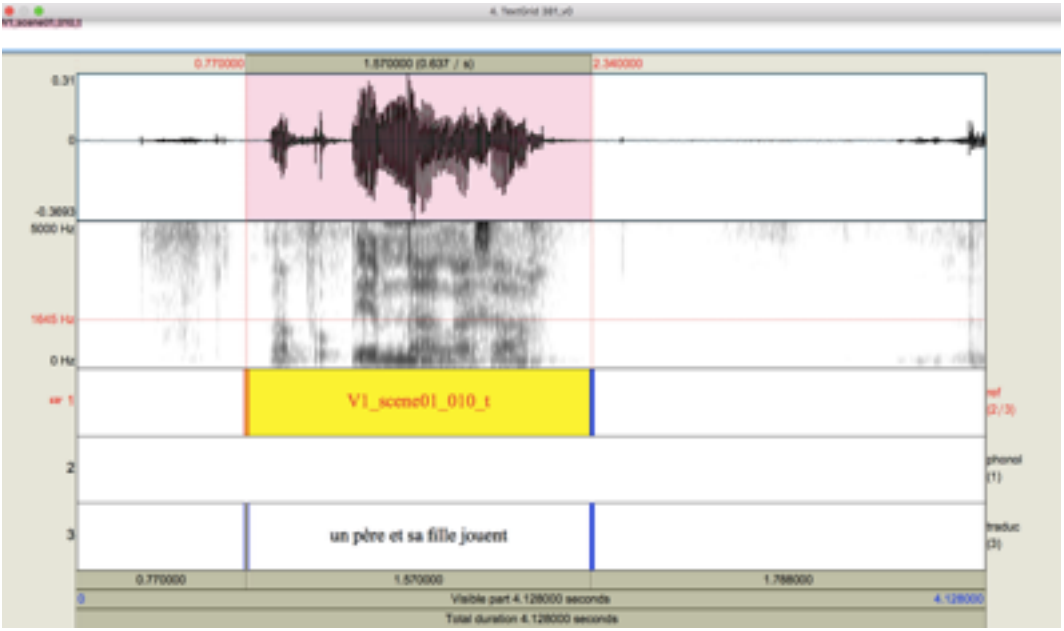

Figure 4: Exemple de transcription sous Praat de l'élicitation de parole en langue sereer

3 Elan_CorpA est un outil d'aide à l'annotation morpho-syntaxique. http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php

4 Praat: logiciel libre scientifique gratuit conçu pour la manipulation, le traitement et la synthèse de sons vocaux (phonétique) http://www.fon.hum.uva.nl/praat/

3.3. Tentative de transcription du sereer

Nous avons utilisé la base de données RefLex[5], qui donne accès à un lexique de sereer en ligne. Pour réaliser ces tentatives de transcriptions,nous nous sommes aidés de la traduction en français, puis du fichier audio en sereer, et du spectrogramme de ce fichier audio. Nous avons écouté le fichier audio en sereer plusieurs fois afin de distinguer les différents phonèmes. Nous nous référons ensuite à la transcription de la traduction en cherchant l'équivalent en sereer dans le lexique. Si nous trouvons une correspondance entre ce qui est entendu et le lexique, nous le notons dans la tier comme tentative de transcription. Sinon, nous nous référons également à la thèse de Marie Renaudier (2012). Une fois ces phases terminées, nous demandons confirmation à notre informateur, ou nous lui demandons des précisions.

La figure 5 montre un zoom de cette tentative de transcription du sereer, sous Elan_CorpA.

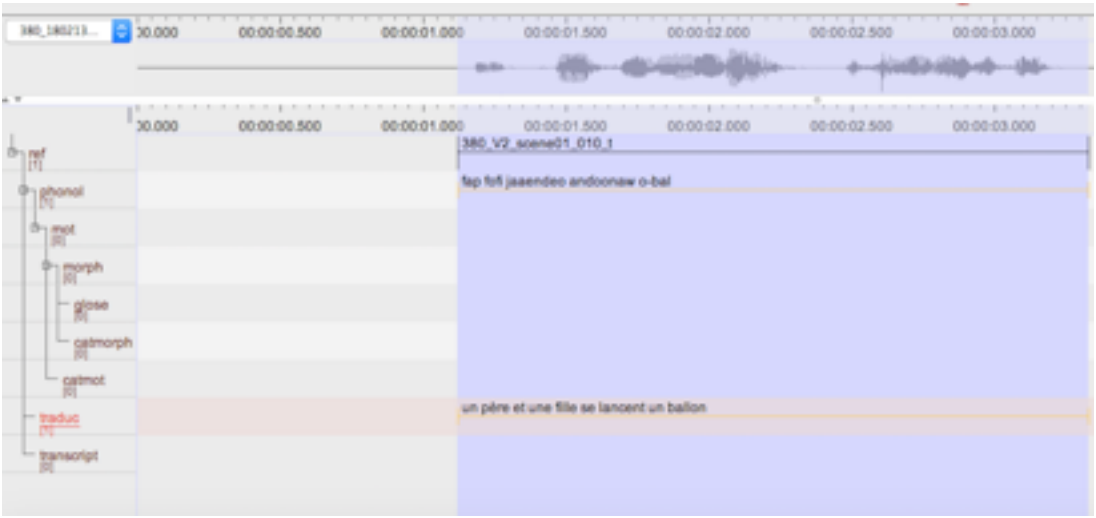

Figure 5: Zoom d'une tentative de transcription du sereer, sous Elan_CorpA

5 http://reflex.cnrs.fr/database/

Les données collectées en sereer ont été, pour quelques unes, contrôlées par un informateur. Malheureusement, il n'était pas assez disponible pour nous permettre de vérifier chaque enregistrements, que ce soit en sereer ou en français. On peut cependant remarquer la forte ressemblance entre certains enregistrements pour une même vidéo. On peut alors en déduire que les propos des locuteurs sont semblables, donc il y a de fortes chances pour qu'ils traduisent effectivement la vidéo pour la langue sereer. Concernant les quelques essais de transcriptions, elles ont également été contrôlées par ce même informateur qui les a validées. L'informateur ayant lui aussi été enregistré, il connaissait par conséquent les vidéos auxquelles se rattachaient les segments audios. Lors d'enregistrements futurs, et si nous disposons de suffisamment de locuteurs, il serait intéressant qu'un locuteur soit essentiellement informateur, ignorant le contenu des vidéos, afin d'avoir un recul suffisant pour la validation des données en sereer, leurs traductions en français, ainsi que pour les transcriptions proposées.

4. L'outil Persephone[6]

La transcription manuelle s'avère longue et fastidieuse. Selon les besoins, pour une minute de discours, la transcription manuelle peut prendre plus de trente minutes pour un linguiste. Il s'agit de temps considérable. Tout les corpus audios existants ne peuvent donc pas être tous annotés, par manque de temps et de moyen. C'est dans un souci de faciliter la transcription pour le linguiste, et d'éviter à certains corpus de finir oublié, que l'outil Persephone a vu le jour.
Persephone est un outil de transcription automatique des phonèmes. Contrairement aux autres outils de reconnaissance vocale, il a été conçu pour fonctionner dans des situations où les données enregistrées sont limitées, c'est à dire que le corpus ne contient qu'une heure, voire moins de discours transcrit. Ce qui peut s'avèrer être le cas lorsque nous travaillons sur des langues peu dotées ou des langues en danger. Cet outil permet, grâce à de faibles quantités, de créer un modèle de transcription afin de la faciliter. L'objectif de cet outil est de réaliser de la phonémie de pointe. Grâce à sa simplicité d'utilisation, toute personne peut l'utiliser. Cet outil est implémenté en Python et en Tensorflow. Il est possible d'y avoir recours via un lien docker. Toutes les étapes de manipulation sont expliquées dans un fichier annexe bien détaillé. Afin de nous familiariser avec Persephone, ils nous proposent dans un premier temps de nous exercer avec un de leur corpus: le corpus Na. Nous nous sommes donc familiarisé avec le Na. Nous testons actuellement l'outil grâce à un corpus en wolof, car il s'agit d'un corpus entièrement transcrit, grâce auquel nous pourrons ensuite élaborer un modèle, peut être applicable pour le sereer.

[6] Récupération de l'outil: https://pypi.org/project/persephone/#files; ADAMS, Oliver and COHN, Trevor and NEUBIG, Graham and CRUZ, Hilaria and BIRD, Steven and MICHAUD, Alexis, 2018. Proceedings of LREC 2018. *Evaluating phonemic transcription of low-resource tonal languages for language documentation*

Conclusion

Cet article présente nos travaux en cours sur la collecte de données en sereer avec une application mobile (LIG-AIKUMA). La transcription manuelle est une étape qui demande énormément de temps et de rigueur au linguiste. Cette étape peut être facilitée grâce à un nouvel outil: Persephone.

Remerciements

Cet article a été réalisé lors d'un stage de Master 2 IDL (Industries de la Langue) réalisé au DDL (Dynamique du Langage) à Lyon et au LIG (Laboratoire d'Informatique de Grenoble) à Grenoble. Ce stage est co-encadré par Sylvie Voisin et Laurent Besacier. Il est financé par le CNRS grâce au laboratoire DDL.

Références

FAYE, Waly Coly. 1979. Etude morphosyntaxique du sereer singandum (région de Jaxaaw-Ñaaxar). Grenoble Thèse de 3e cycle, sous la direction de Denis Creissels.

MC LAUGHLIN, F., 1992. Noun classification in Seereer-Siin. University of Texas, PhD thesis.

RENAUDIER, Marie. 2012. Dérivation et valence en sereer. Lyon, sous la direction de PHILIPPSON Gérard: Université Lumière - Lyon2 Thèse de Doctorat de troisième cycle.

POZDNIAKOV, Konstantin & SEGERER Guillaume. 2006. Les alternances consonantiques du sereer : entre classification nominale et dérivation. *Linguistique du Musée royal de l'Afrique centrale* Africana Linguistica(XXII). 137–162.

GAUTHIER, Elodie, 2018. Collecter, Transcrire, Analyser: quand la machine assiste le linguiste dans son travail de terrain. Grenoble, sous la direction de VOISIN Sylvie et BESACIER Laurent: Université Grenoble Alpes, Thèse.

RENAUDIER, Marie, 2015. Les classes nominales en sereer. *Les classes nominales dans les langues atlantiques,* 486-520.

ADAMS, Oliver and COHN, Trevor and NEUBIG, Graham and CRUZ, Hilaria and BIRD, Steven and MICHAUD, Alexis, 2018. Proceedings of LREC 2018. *Evaluating phonemic transcription of low-resource tonal languages for language documentation*