Collecter, Transcrire, Analyser : quand la machine assiste le linguiste dans son travail de terrain

Elodie Gauthier

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Elodie GAUTHIER

Thèse dirigée par **Laurent BESACIER, Professeur des Universités, Université Grenoble Alpes**, et
co-encadrée par **Sylvie VOISIN, Maître de Conférences, Université Aix-Marseille**

préparée au sein du **Laboratoire d'Informatique de Grenoble (LIG)**
dans l'**École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Collecter, Transcrire, Analyser : quand la Machine Assiste le Linguiste dans son Travail de Terrain

Thèse soutenue publiquement le **30 mars 2018**,
devant le jury composé de :

Mme Martine ADDA-DECKER
Directrice de Recherche, CNRS, LPP, Université Paris III, Présidente
Mme Claire GARDENT
Directrice de Recherche, CNRS, LORIA, Rapporteur
M. Steven BIRD
Professeur des Universités, Charles Darwin University, Northern Institute, Rapporteur
Mme Sylvie VOISIN
Maître de Conférences, Université Aix-Marseille, DDL, Co-encadrante de thèse
M. Laurent BESACIER
Professeur des Universités, Université Grenoble Alpes, LIG, Directeur de thèse

À mon grand-père, concepteur de systèmes ingénieux et friand d'innovation

À mon père, persévérant et déterminé, pointilleux et perfectionniste

Remerciements

Lorsque je me suis engagée dans ce travail de thèse, je ne m'attendais pas à ce qu'il soit aussi intense et qu'il m'apporte autant. Ces années passées ont été extrêmement formatrices et m'ont enrichie, tant sur le plan professionnel que personnel.

Tout d'abord, je remercie Martine Adda-Decker, Claire Gardent et Steven Bird pour la considération portée à mon travail et pour avoir accepté de faire partie de mon jury, avec toutes les responsabilités que cela implique. La pertinence de vos commentaires et vos conseils m'ont permis de fournir un meilleur manuscrit.

Je souhaite également remercier mon directeur, Laurent Besacier, pour m'avoir proposé ce sujet et avoir cru en moi dès le départ, mais aussi pour la sagacité et la justesse de ses remarques et enfin pour m'avoir toujours poussée à donner le meilleur de moi-même.

Je remercie Sylvie Voisin, pour son co-encadrement qui m'a permis d'aborder ces recherches sous un angle différent mais complémentaire, ainsi que pour sa curiosité et sa minutie qui ont favorisé l'aboutissement de ce travail pluridisciplinaire et appronfondi.

Malgré vos plannings surchargés, vous m'avez toujours accordé le temps nécessaire et nos rencontres étaient, à chaque fois, constructives. Je me suis nourrie de votre enthousiasme, de votre ambition, de vos savoirs et de votre rigueur, et si j'avais à recommencer, je dirais oui sans aucune hésitation. Vous avez été des encadrants exemplaires.

Ce sujet de thèse m'a offert l'opportunité de découvrir de nouveaux pays et de nouvelles cultures. Mes deux voyages au Sénégal, notamment, ont été d'une richesse sans égale. J'ai rencontré, à l'UCAD, des personnes merveilleuses. Je remercie particulièrement Thierno Cissé pour avoir pris autant soin de moi lors de mon premier voyage. Je te suis extrêmement reconnaissante d'avoir été si disponible, accueillant et bienveillant. Je tiens aussi spécialement à remercier El Hadji Dièye et Alain-Christian Bassène pour leur accueil et leur gentillesse à mon égard, mais aussi pour la qualité de nos échanges et leur implication dans mon travail.

Je remercie aussi Mamour Dramé de nous avoir emmenées, avec Sylvie, dans son village, à Paka Thiare Secco, et Boubacar pour tout le temps qu'il m'a consacré durant mon terrain et pour l'attention qu'il a porté à mon travail alors qu'il n'est pas du tout linguiste. Sans ces études réalisées sur le terrain et sans leur participation, la qualité de ce travail aurait été moindre.

Je suis heureuse d'avoir passé ces années au sein de l'équipe GETALP, dans laquelle j'ai pu m'épanouir et mûrir mon projet de recherche. La multidisciplinarité de l'équipe m'aura clairement aidée à répondre aux problématiques variées de cette thèse. Je remercie tous les permanents de m'avoir aussi bien accueillie et accompagnée quand j'en ai eu besoin, notamment Benjamin Lecouteux et Solange Rossato qui m'ont permis, grâce leur expertise, de mener à bien certaines de mes expérimentations.

Merci aussi à tous les doctorants et post-doctorants, Alexis, Jérémy, David, Christophe, Fred et Sarah pour la bonne ambiance dans les bureaux et pendant les pauses repas et café !

Mention spéciale à mes fidèles comparses, Alexis et Jérémy, pour m'avoir supportée — dans le sens « soutenue », bien entendu !! — durant ces trois années ; merci pour votre présence, pour l'éclectisme des conversations menées, pour votre geektitude qui m'a fait découvrir énormément de choses, pour les parties de franches rigolades mais aussi pour m'avoir continuellement épaulée dans les moments plus durs. Et puis, Alex, je n'oublie pas que « Python, c'est la vie ! ». Aussi, je tiens à remercier Johann, tout d'abord pour m'avoir mis le pied à l'étrier à Python, pendant mon stage ! Mais surtout pour ta vivacité d'esprit, ton aide, ton soutien et tes réponses à mes questionnements que tu étais, parfois, le seul à saisir.

En dehors du laboratoire, je remercie de tout cœur tous mes amis qui m'ont suivie, qui se sont intéressés à mon travail, et même pour les plus courageux, qui ont essayé de comprendre mon sujet de thèse jusqu'à la fin ! Merci pour votre patience vis-à-vis de mes absences — de longue durée parfois — mais de toujours vous être rendus disponibles les rares fois où je l'ai été. Merci à Anaïs, Lucie, Kich, Nina, Stef, Marjo, Marie, Cécile, Pascale, Juliette, Coraline, Julia, Cyril, Blandine, Aurore, Babeth, Morgane, Marina de faire partie de ma vie et de l'égayer à chaque fois que l'on se retrouve.

Enfin, je remercie mes grands-parents, mes parents, mes oncles et tantes, mes soeurs et ma cousine, qui ont toujours cru en moi et qui, eux aussi, ont été extrêmement tolérants et compréhensifs. Je remercie tout spécialement ma petite sœur Julia, mon modèle de gaieté, d'enjouement et d'entrain, toujours présente pour me réconforter dans mes moments de doute.

Et, finalement, merci à toi, Candice, qui a été les deux jambes et les deux épaules dont j'ai eu besoin quand je flanchissais, et même parfois mon cerveau de substitution ! Tu m'as éclairée à bien des reprises au travers de tes judicieux conseils et habiles raisonnements. Tu as été l'oxygène et la force qui m'ont permis d'aller au bout de cette thèse. Merci pour l'intérêt que tu as toujours porté à mon travail, tes encouragements, ton indulgence et ton soutien sans faille. Définitivement, je ne te remercierai jamais assez d'ensoleiller et de réchauffer à tel point mon quotidien.

Table des matières

Résumé

Depuis quelques décennies, de nombreux scientifiques alertent au sujet de la disparition des langues qui ne cesse de s'accélérer. Face au déclin alarmant du patrimoine linguistique mondial, il est urgent d'agir afin de permettre aux linguistes de terrain, *a minima*, de documenter les langues en leur fournissant des outils de collecte innovants et, si possible, de leur permettre de décrire ces langues grâce au traitement des données assisté par ordinateur.

C'est ce que propose ce travail, en se concentrant sur trois axes majeurs du métier de linguiste de terrain : la collecte, la transcription et l'analyse.

Les enregistrements audio sont primordiaux, puisqu'ils constituent le matériau source, le point de départ du travail de description. De plus, tel un instantané, ils représentent un objet précieux pour la documentation de la langue. Cependant, les outils actuels d'enregistrement n'offrent pas au linguiste la possibilité d'être efficace dans son travail et l'ensemble des appareils qu'il doit utiliser (enregistreur, ordinateur, microphone, etc.) peut devenir encombrant.

Ainsi, nous avons développé Lig-Aikuma, une application mobile de collecte de parole innovante, qui permet d'effectuer des enregistrements directement exploitables par les moteurs de reconnaissance automatique de la parole (RAP). Les fonctionnalités implémentées permettent d'enregistrer différents types de discours (parole spontanée, parole élicitée, parole lue) et de partager les enregistrements avec les locuteurs. L'application permet, en outre, la construction de corpus alignés « parole source (peu dotée)-parole cible (bien dotée) », « parole-image », « parole-vidéo » qui présentent un intérêt fort pour les technologies de la parole, notamment pour l'apprentissage non supervisé.

Bien que la collecte ait été menée de façon efficace, l'exploitation (de la transcription jusqu'à la glose, en passant par la traduction) de la totalité de ces enregistrements est impossible, tant la tâche est fastidieuse et chronophage.

Afin de compléter l'aide apportée aux linguistes, nous proposons d'utiliser des techniques de traitement automatique de la langue pour lui permettre de tirer partie de la totalité de ses données collectées. Parmi celles-ci, la RAP peut être utilisée pour produire des transcriptions, d'une qualité satisfaisante, de ses enregistrements.

Une fois les transcriptions obtenues, le linguiste peut s'adonner à l'analyse de ses données. Afin qu'il puisse procéder à l'étude de l'ensemble de ses corpus, nous considérons l'usage des méthodes d'alignement forcé. Nous démontrons que de telles techniques peuvent conduire à des analyses linguistiques à granularité fine. En retour, nous montrons que la modélisation de ces observations peut mener à des améliorations des systèmes de RAP.

Abstract

In the last few decades, many scientists were concerned with the fast extinction of languages. Faced with this alarming decline of the world's linguistic heritage, action is urgently needed to enable field linguists, at least, to document languages by providing them innovative collection tools and to enable them to describe these languages. Machine assistance might be interesting to help them in such a task.

This is what we propose in this work, focusing on three pillars of the linguistic fieldwork : collection, transcription and analysis.

Recordings are essential, since they are the source material, the starting point of the descriptive work. Speech recording is also a valuable object for the documentation of the language. The growing proliferation of smartphones and other interactive voice mobile devices offer new opportunities for field linguists and researchers in language documentation. Field recordings should also include ethnolinguistic material which is particularly valuable to document traditions and way of living. However, large data collections require well organized repositories to access the content, with efficient file naming and metadata conventions.

Thus, we have developed Lig-Aikuma, a free Android app running on various mobile phones and tablets. The app aims to record speech for language documentation, over an innovative way. It includes a smart generation and handling of speaker metadata as well as respeaking and parallel audio data mapping. Lig-Aikuma proposes a range of different speech collection modes (recording, respeaking, translation and elicitation) and offers the possibility to share recordings between users. Through these modes, parallel corpora are built such as "under-resourced speech - well-resourced speech", "speech - image", "speech - video", which are also of a great interest for speech technologies, especially for unsupervised learning.

After the data collection step, the field linguist transcribes these data. Nonetheless, it can not be done — currently — on the whole collection, since the task is tedious and time-consuming. We propose to use automatic techniques to help the field linguist to take advantage of all his speech collection. Along these lines, automatic speech recognition (ASR) is a way to produce transcripts of the recordings, with a decent quality.

Once the transcripts are obtained (and corrected), the linguist can analyze his data. In order to analyze the whole collection collected, we consider the use of forced alignment methods. We demonstrate that such techniques can lead to fine-grained evaluation of linguistic features. In return, we show that modeling specific features may lead to improvements of the ASR systems.

Table des graphiques

Liste des tableaux

Introduction

Motivation

Depuis quelques décennies, de nombreux scientifiques alertent au sujet de la disparition des langues qui ne cesse de s'accélérer. Une des causes de cette extinction est la transmission orale qui ne s'opère plus entre les générations, car l'acquisition d'une langue dominante est privilégiée à la langue maternelle. De plus, ces langues — à tradition orale — sont dépourvues d'un système d'écriture. Cette absence d'écriture entraîne une disparition complète de la langue, au moment où le dernier locuteur de la communauté décède, puisque plus aucune trace permanente ne témoigne alors de l'existence de cette langue.

Cette thèse s'inscrit dans un contexte où, d'un côté, les langues ont besoin d'être collectées et documentées rapidement (une langue meurt tous les 14 jours environ selon Rymer (2012)) et, de l'autre, les technologies de traitement automatique de la langue (TAL) sont désormais abordables et applicables dans de nombreux domaines.

Face au déclin alarmant du patrimoine linguistique mondial, il est urgent d'agir afin de permettre aux linguistes, *a minima*, de documenter les langues en leur fournissant des outils de collecte innovants et, si possible, de leur permettre de décrire ces langues grâce au traitement des données assisté par ordinateur.

La reconnaissance automatique de la parole est un moyen qui peut être utilisé pour le traitement automatique des données. Mais cette technologie a une limite : l'entraînement d'un moteur de reconnaissance vocale a besoin d'une grande quantité de données (écrites et orales) pour être efficace. Afin de rendre l'utilisation de cet outil envisageable, un moyen doit être trouvé pour, d'une part, pallier l'indisponibilité des documents (numériques) écrits et, d'autre part, enregistrer les locuteurs dans un format directement utilisable par le système.

Des solutions doivent être trouvées afin de traiter les langues en danger. Néanmoins, les systèmes de transcription actuels ont besoin d'un minimum de données écrites afin d'être efficaces. Dans ce travail, nous nous sommes avant tout concentrés sur les problématiques liées aux langues très peu dotées qui, non seulement manquent de données numériques mais ont aussi besoin de technologies vocales afin d'être maintenues.

Cette thèse est motivée par la volonté de permettre aux linguistes d'utiliser les technologies comme la reconnaissance automatique de la parole pour transcrire et analyser les langues.

Cette vision est partagée par les partenaires du projet ANR ALFFA [1], par le biais duquel cette thèse a été financée. Ce projet ambitionne de rassembler linguistes de terrain et informaticiens afin de créer des outils profitables aux deux parties.

Nous avons adopté, dans ce travail de thèse, une approche pluridisciplinaire, dans laquelle intervient un travail de recherche aussi bien en informatique qu'en linguistique. Le co-encadrement est réalisé par des membres de différents laboratoires ayant chacun leur spécialité. D'une part, Laurent Besacier, directeur du Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole au sein du Laboratoire d'Informatique de Grenoble (LIG), spécialisé en traitement automatique des langues ; d'autre part, Sylvie Voisin, responsable du parcours Langues en Contact et Typologie de l'Université d'Aix Marseille et rattachée au laboratoire de Dynamique du Language de Lyon (DDL), linguiste de terrain spécialisée dans la description des langues d'Afrique de l'Ouest.
Nous avons appliqué les technologies de la parole aux problématiques des langues peu dotées (contribution à la recherche en informatique), mais aussi participé aux travaux de collecte, de documentation et de description du wolof (contribution à la recherche en linguistique). Bien que le wolof ne soit pas une langue en danger et soit décrit depuis longtemps, cette langue est toujours objet d'intérêt pour les linguistes et est encore en cours de description. De plus, le wolof a la particularité d'être peu doté (son orthographe est instable, sa présence est fortement limitée sur le Web et cette langue manque de ressources électroniques pour le traitement automatique des langues). Finalement, cette double approche en informatique et en linguistique nous a permis d'élaborer des méthodes et outils d'assistance au linguiste par la machine.

Le projet ALFFA

Le projet ANR Blanc ALFFA (*African Languages in the Field : speech Fundamentals and Automation*) [2] vise à mettre en place des méthodes innovantes pour aider à la description de langues africaines peu dotées, au moyen d'outils de traitement automatique de la parole tels que la reconnaissance de la parole (RAP) et la synthèse vocale (TTS). Pour cela, le projet rassemble des linguistes et des informaticiens, et fait intervenir laboratoires de recherche (le DDL à Lyon, le laboratoire d'Informatique d'Avignon (LIA), le LIG à Grenoble) et industrie (Voxygen, spécialisé dans la synthèse vocale et basé à Lannion et Rennes). Dans le projet, les tâches ont été attribuées en fonction des savoir-faire de chaque partenaire. Ainsi, le DDL s'est occupé des fondamentaux de la parole, des descriptions linguistiques et écologiques des langues ; Voxygen et le LIA se sont occupés des technologies vocales (respectivement, synthèse et reconnaissance) ; le LIG s'est occupé des collectes de données, du développement d'outils pour les linguistes (RAP et implémentation de mesures phonétiques).

1. Mais aussi par les partenaires du projet ANR/DFG BULB, projet auquel j'ai participé et qui sera décrit plus loin dans le manuscrit.
2. http://alffa.imag.fr

L'objectif, finalement, est de développer et de standardiser les ressources linguistiques dans de nombreuses langues, — pas uniquement celles qui disposent déjà de ressources —, pour introduire les langues nationales dans l'espace numérique, grâce à la création et à la diffusion de contenu dans ces langues, et pour fournir un accès multilingue aux ressources numériques.

Définition des problèmes

Cette thèse devra répondre aux problématiques suivantes :

1. Comment permettre au linguiste de terrain de travailler à grande échelle ?

   Le linguiste de terrain, actuellement, n'est pas en mesure d'exploiter l'intégralité de ses corpus (par manque de temps et de moyens financiers). Ses descriptions ne portent, par conséquent, que sur un sous-ensemble restreint des corpus collectés. Le premier objectif de cette thèse sera, donc, de se servir de la RAP. En haoussa, des corpus pour la construction d'un système sont disponibles mais en wolof de telles ressources n'existent pas. Nous devrons, dans un premier temps, constituer des corpus textuels et audio en wolof, pour construire un système de RAP pour cette langue.

2. Comment mettre en place des méthodes de collectes de données innovantes ?

   La collecte de données sur le terrain est une étape primordiale du travail du linguiste de terrain. Les outils d'enregistrement actuellement utilisés par le linguiste ne permettent pas de construire des ressources directement exploitables par le système de RAP ou les outils de TAL. Dans l'objectif d'augmenter l'efficacité du linguiste, nous développerons une application mobile, ergonomique et légère, qui permettra au linguiste de collecter de la parole utilisable par les moteurs de RAP, mais aussi dont les fonctionnalités permettront d'enregistrer différents types de discours.

3. Quelles adaptations des outils informatiques faut-il proposer à un non spécialiste afin de lui faire gagner du temps ?

   Un système de RAP est un outil compliqué à mettre en place et peu intuitif à utiliser pour un non spécialiste. Pourtant, le gain de temps engendré par l'utilisation de cette technologie, pour obtenir des transcriptions d'enregistrements vocaux ou pour mener des analyses à partir de données estimées par le système, est considérable. Cela signifie que nous devrons mettre en place des méthodes pour permettre au linguiste, dans un premier temps, de tirer profit de la RAP afin d'obtenir des transcriptions de ses corpus oraux ; dans un deuxième temps, d'accéder de manière simplifiée aux outils connexes à la RAP (comme les techniques d'alignement forcé), pour analyser automatiquement ses corpus.

4. Les méthodes automatiques issues de la RAP peuvent-elles aider le linguiste dans son travail d'analyse ?

Les systèmes de RAP sont intéressants pour analyser, de façon automatique, une langue à grande échelle et selon différentes granularités : phonétiques, morphologiques, syllabiques, lexicales, sémantiques ou encore prosodiques. Néanmoins, les mesures produites par les systèmes sont-elles justes ? Nous devrons montrer que les méthodes issues de la RAP peuvent être utiles pour vérifier des faits à grande échelle. Pour cela, nous effectuerons des analyses automatiques sur de grands corpus, au moyen de techniques d'alignement forcé. À travers cette méthodologie, nous déterminerons si des hypothèses linguistiques décrites dans la littérature sont constatées à l'aide de méthodes automatiques.

5. En retour, les analyses linguistiques à grande échelle peuvent-elles améliorer les systèmes de reconnaissance automatique de la parole ?

   Les résultats de nos analyses seront modélisés afin que le système de RAP prenne en compte les particularités linguistiques observées. De cette manière, nous montrerons si la modélisation de traits linguistiques précis peut rendre les systèmes de RAP plus efficaces et performants.

Plan du manuscrit

Le chapitre 1 de cette thèse sera consacré à l'état de l'art, dont le premier volet portera sur la linguistique de terrain. Nous présenterons ce domaine de recherche et discuterons de ce que constitue un terrain. Nous conclurons ce volet en présentant notre vision du linguiste de terrain assisté par la machine, vers laquelle nous avons voulu tendre tout au long de ces travaux de thèse. Le deuxième volet fera une brève revue des outils numériques existants pour la linguistique, notamment ceux répondant aux besoins des linguistes de terrain qui, dans un but de documentation et description, doivent exploiter et analyser des données multimodales, souvent atypiques. Le troisième volet portera sur la reconnaissance de la parole : nous rappellerons les principes généraux et nous introduirons les problématiques liées aux langues peu dotées.

Ensuite, nous entamerons le chapitre 2 de cette thèse, qui présentera les langues sur lesquelles nous avons principalement travaillé durant ces trois années : le haoussa et le wolof. Nous ferons un point sur le statut de ces deux langues d'Afrique sub-saharienne puis présenterons les règles linguistiques principales qui régissent leurs systèmes.

Cette thèse s'articule autour de 3 axes majeurs que sont la collecte, la transcription et l'analyse. Ils constitueront les trois derniers chapitres de ce manuscrit.

Tout d'abord, la première partie du chapitre 3 « Collecter », sera consacrée à l'application mobile de collecte de parole que nous avons développée : Lig-Aikuma. Nous rappellerons son origine et expliquerons les fonctionnalités implémentées. Une deuxième partie sera consacrée aux collectes menées au Sénégal, qui nous ont permis de constituer, en deux temps, des corpus de plusieurs types de parole et de plusieurs variétés du wolof.

Puis, nous continuerons avec le chapitre 4 « Transcrire », dans lequel nous introduirons les premiers systèmes de RAP que nous avons construits et déployés, en détaillant les approches acoustiques et linguistiques de modélisation adoptées.

Enfin, ce manuscrit de thèse se clôturera avec le chapitre 5 « Analyser ». Nous y aborderons la prise en compte de phénomènes linguistiques précis et, plus précisément, nous nous concentrerons sur l'opposition de longueur de voyelles. Nous expliquerons les analyses automatiques effectuées sur les corpus de parole en haoussa et en wolof, ainsi que les tendances observées sur la réalisation de l'opposition de longueur vocalique dans ces deux langues. Nous présenterons aussi la modélisation de ce phénomène dans nos systèmes de RAP et les gains apportés par cette prise en compte du contraste de durées sur nos systèmes.

En conclusion de ce manuscrit, nous résumerons les contributions apportées au terme de ces trois années de thèse puis dégagerons quelques perspectives.

CHAPITRE 1

État de l'art

1.1 La linguistique de terrain

La linguistique de terrain est indirectement née, à la fin des années 1900, grâce à l'anthropologue américain Franz Boas (1858–1942) pour qui l'élaboration d'une théorie devrait provenir d'observations empiriques. Plusieurs autres éminents anthropologues adhéreront à son courant de pensée, comme Edward Sapir ou encore Leonard Bloomfield.

1.1.1 Définition

Afin de comprendre ce qu'est la linguistique de terrain, commençons par définir les termes « linguistique » et « terrain ».

La linguistique. Étymologiquement, le mot « linguistique » semble avoir été emprunté à l'allemand *linguistik* et désignait, au 19è siècle, l'étude générale des langues et notamment leur classification (Balbi, 1826, page 61). Plus récemment, Martinet (1967, page 6), désigne la linguistique comme étant « l'étude scientifique du langage humain » et explique que cette discipline fait partie des sciences en cela qu'elle observe, mais ne prend jamais parti. La définition que l'on peut trouver actuellement de la linguistique, dans le Trésor de la Langue Française informatisé (TLFi), est la suivante :

> « une science qui a pour objet l'étude du langage, des langues envisagées comme systèmes sous leurs aspects phonologiques, syntaxiques, lexicaux et sémantiques. »

La linguistique est, donc, un domaine de recherche qui s'intéresse — au moyen de l'observation, de l'analyse et de l'expérience — d'une part à l'étude de l'acquisition du langage par l'humain et d'autre part, à l'étude de la structure et du fonctionnement des langues du monde.

Le terrain. La définition de ce qu'est un « terrain » n'est pas la même pour tous les linguistes. Hyman (2001, section 1.1) propose, avant tout, de discerner deux types d'approches : la linguistique de terrain prototypique et la linguistique de terrain non prototypique. L'approche est dite non prototypique lorsqu'il n'y a aucune interaction entre la communauté de la langue d'étude et le linguiste. Autrement dit, le linguiste qui ne rencontre pas les populations parlant la langue qu'il étudie ou le linguiste qui étudie sa propre langue (maternelle) ou

encore le linguiste qui analyse des matériaux recueillis par d'autres personnes (même s'il se trouve dans la communauté de locuteurs), sont des cas d'étude non prototypiques. À l'inverse, Hyman caractérise l'approche prototypique à travers trois propriétés : la distance, l'exotisme et la durée. Le linguiste adopte ainsi l'approche prototypique lorsqu'il entre en contact et se déplace durant un long séjour au cœur des populations qui parlent une langue éloignée de la sienne.

La vision de Sakel et Everett (2012, chap. 1) du terrain linguistique est moins manichéenne :

> "Fieldwork describes the activity of the researcher systematically analysing parts of a language, usually other than one's native language and usually within a community of speakers of that language."

Autrement dit, le terrain linguistique est à la fois prototypique et non prototypique. Sakel et Everett expliquent, à travers leur définition, que le terrain est, d'une part, prototypique, en ce que le linguiste se déplace dans la communauté de la langue d'étude qui lui est étrangère et d'autre part, non prototypique, car le linguiste se déplace dans cette communauté avec des connaissances et un système d'idées qui constituent déjà un *a priori* sur l'objet de son analyse. De plus, Sakel et Everett n'attachent pas d'importance à la distance qui sépare le lieu de vie du linguiste du lieu de son terrain, tant que le terrain est mené au sein de la communauté de locuteurs.

La linguistique de terrain. En ralliant les termes « linguistique » et « terrain », nous comprenons, donc, que la linguistique de terrain s'attelle à collecter des données au sein d'une communauté de locuteurs d'une langue. Ajoutons que, la plupart du temps, la langue d'étude est à tradition orale, c'est-à-dire que la langue est majoritairement parlée et non écrite. En allant directement à la rencontre des locuteurs, le linguiste s'intéresse à la collecte de données réelles. Le linguiste de terrain enregistre la langue en contexte : c'est l'usage par les locuteurs qui l'intéresse. Les discours enregistrés en contexte, mais aussi les photos, vidéos et prises de notes que capturent le linguiste, permettent de documenter la langue. Ces supports (multimodaux) constituent les fondements pour la description de la langue qui est faite par la suite.

Comme l'indique Bowern (2015, section 1.2.2), le linguiste de terrain doit posséder plusieurs qualités. Il doit être :

- **méthodique** : réfléchir et user d'une méthodologie de collecte adéquate qui lui permettra de collecter des données pertinentes, qui pourront être organisées ;
- **organisé** : stocker, nommer, trier intelligemment ses données pour pouvoir les retrouver et les traiter de façon efficace ;
- **capable de diriger/conduire un projet** : rechercher des informateurs, rechercher où loger et où se nourrir, être capable de communiquer avec les informateurs qui travailleront avec lui en amont du terrain et une fois sur place, gérer le budget qu'il allouera aux participants, réfléchir à ce qui sera partagé avec la communauté ;

- **éthique et responsable** : le linguiste de terrain est au cœur des populations et de leur quotidien. Il se doit de respecter les lois et les participants, ainsi que de leur demander leur consentement avant tout enregistrement et de les tenir informés de tout ce qui concerne les données acquises grâce à leur participation, tel que la diffusion des données.
- **respectueux et modeste** : le linguiste devra s'intégrer à la communauté et savoir s'accoutumer aux pratiques locales. La communauté qui l'accueille ne doit pas être envahie par la culture personnelle du linguiste ou par ses croyances propres.

Mais aussi doit avoir d'autres fonctions que celles de linguiste et devenir :

- **socio-anthropologue** : le linguiste étudie non seulement la langue de la communauté, mais aussi les interactions entre les personnes et les pratiques culturelles qu'il aura la chance de découvrir. La nature des données collectées est linguistique et extra-linguistique ;
- **ingénieur son et vidéo** : le linguiste, durant son terrain, doit connaître son matériel d'enregistrement audio pour saisir les faits de langue et, si possible, savoir manier son matériel vidéo pour filmer les participants lors des interviews ou des scènes du quotidien, dans un but de documentation.

Documentation et description

Les raisons pour le linguiste de se déplacer sur le terrain sont souvent liées à un travail de documentation (comprendre comment la langue fonctionne et comment elle est utilisée au quotidien par les locuteurs (natifs)) ou à un travail de description où l'approche adoptée peut être diachronique (observer l'évolution de la langue) ou synchronique (description de la langue à un moment donné). Documentation et description sont deux champs de recherche souvent amalgamés. Pourtant, Himmelmann (1998), dans *Documentary and Descriptive linguistics*, distingue clairement ces deux disciplines qui, bien que fortement corrélées, font appel à une méthodologie différente. Ainsi, la documentation correspond à tout ce qui touche à la collecte des données primaires tandis que la description correspond à tout ce qui touche à l'analyse de ces données.

Les langues en danger

Depuis presque une trentaine d'années maintenant est né un sous-domaine de la linguistique de terrain consacré aux « langues en danger ». Ce domaine de recherche de la linguistique de terrain s'est créé suite à un constat tragique : 25 langues par an, en moyenne, périssent. De plus, les linguistes ont noté une accélération de la disparition des langues et notamment ils estiment que 90% des langues d'Australie et d'Amérique du nord sont vouées à mourir dans moins de 100 ans. Différentes causes peuvent mener une langue à péricliter (Austin et Sallabank, 2011) : répression et apprentissage forcé, globalisation, choix politico-économiques, catastrophes naturelles, réchauffement climatique, massacres ethniques, guerres, etc. Cependant,

le défaut de transmission intergénérationnelle de la langue maternelle reste une cause majeure de ce déclin. Les langues dites en danger ne se trouvent pas que dans des pays lointains, inconnus du grand public et difficilement accessibles. En France aussi, nous avons des langues en danger, telles que le breton (Caisse, 2008) ou le franco-provençal (Bert, 2010). Les langues sont regroupées par famille (une famille peut être composée d'une seule langue). Lorsqu'une famille de langue meurt, c'est toute une richesse culturelle et linguistique que l'on perd : les pratiques culturelles, les rites et savoir-faire traditionnels, les idéologies, une philosophie de vie, mais aussi tous les savoirs scientifiques et environnementaux tels que la taxinomie adoptée [1]. Par ailleurs, l'UNESCO a écrit une convention pour protéger et sauvegarder cette diversité culturelle regroupée sous le nom de « patrimoine culturel immatériel » [2].

Le linguiste de terrain qui œuvre pour la sauvegarde des langues en danger s'affaire, ainsi, à recenser, documenter et décrire pour revitaliser et garder une trace de ces langues qui, inéluctablement, vont disparaître. À travers son métier, le linguiste de terrain transmet également sa passion pour les langues et informe les populations. À l'image de cet investissement, la motivation des locuteurs pour la préservation de leur langue se fait de plus en plus vive. Les locuteurs de communauté de langues en danger ont pris conscience de l'importance de maintenir leur langue. Ils sont demandeurs de formation et souhaitent contribuer à la documentation de leur langue. Grâce aux programmes de revalorisation des langues — comme celui du National Geographic (Anderson, Harrison et Rainier, 2007), celui pour la promotion des langues franco-provençales (Bert et Martin, 2012), ou encore celui porté par l'Université de Varsovie (*Endangered languages. Comprehensive models for research and revitalization* [3]) —, les linguistes de terrain de langues en danger peuvent aussi être formateurs et aider des locuteurs natifs à devenir experts linguistes (Anderson, 2011 ; Rice, 2011).

1.1.2 Matériel utilisé

Le linguiste de terrain a besoin au minimum d'un :

- **dictaphone**, pour enregistrer les séances de travail avec les locuteurs ;
- **carnet de terrain**, pour prendre des notes ;
- **ordinateur portable**, pour avant tout stocker les enregistrements audio et pouvoir les analyser, mais également pour pouvoir faire de l'élicitation à partir de vidéos.

Le linguiste peut également s'équiper d'un appareil photo et d'une caméra afin de filmer les rites et pratiques culturelles de la communauté. Ce type d'informations extra-linguistiques est précieux pour comprendre le fonctionnement de la langue.

1.1.3 Avant le terrain

Avant toute chose, le linguiste met en exergue les questions de recherche sur lesquelles il s'interroge et pour lesquelles il lui manque encore des données.

1. http://www.unesco.org/new/fr/culture/themes/endangered-languages/biodiversity-and-linguistic-diversity
2. https://ich.unesco.org/fr/qu-est-ce-que-le-patrimoine-culturel-immateriel-00003
3. http://www.revitalization.al.uw.edu.pl/eng/Page/About_the_project

L'établissement de ces hypothèses à vérifier lui permettra de réfléchir aux méthodes de recherche qu'il mettra en place durant son étude sur le terrain. Cela lui permettra aussi d'estimer un budget qui lui servira à financer son terrain (transport, logement, repas) et à gratifier les participants à ses enquêtes de terrain. La durée de son terrain, qui peut varier selon la disponibilité des locuteurs, les conditions climatiques ou autres imprévus, peut aussi dépendre du budget alloué ou être pris en compte en amont dans l'estimation du budget.
Dans certains cas, le linguiste doit également faire valider son projet par le comité d'éthique de son université/laboratoire.
Le linguiste ne doit pas oublier de se renseigner sur le pays, la région et la communauté dans laquelle il veut se rendre, afin d'éviter toute surprise d'ordre climatique, culturel, culinaire, etc. Le choix du lieu du terrain pourra amener le linguiste, selon les lois en vigueur dans son pays et dans le pays de destination, à se renseigner sur les examens médicaux, vaccins et médicaments obligatoires (traitement anti-paludisme par exemple), mais également sur les autorisations d'entrée sur le territoire. Ces procédures entraîneront un facteur temps non négligeable pour estimer la date de départ.
Enfin, la recherche d'un logement pour la durée du séjour est un autre point important dans la préparation du terrain, qui peut prendre du temps si le linguiste se rend pour la première fois sur le(s) lieu(x) du terrain.

1.1.4 Déroulement d'un terrain

Le déroulement d'un terrain dépend en grande partie de ce que veut observer le linguiste. La linguistique couvrant un large domaine, il devra se restreindre et se concentrer sur une seule thématique à la fois. Durant le terrain, le linguiste planifie ses rencontres avec les informateurs qu'il a déjà contactés en amont. Pendant les séances de travail, la prise de notes est essentielle, notamment concernant les informations extra-linguistiques. Ces informations, aussi appelées métadonnées, donnent des précisions sur le contexte et le déroulement de la séance. Sakel et Everett évoquent quelques éléments primordiaux à noter pour documenter les enregistrements liés à la séance de travail (tirés du projet EMELD [4] et de Samarin (1967)) :

- la date, le sujet et le lieu de la séance ;
- le nom du fichier audio relatif à la séance ;
- le nom, la nature (sons, textes, vidéos, images) et le sujet des documents utilisés pendant l'enregistrement de la séance ;
- les caractéristiques des documents travaillés pendant la séance ;
- le genre de discours enregistré (lecture, monologue, entretien, conversation, etc.) ;
- les personnes impliquées dans la séance et/ou l'enregistrement ;
- les linguistes présents et leurs contributions ;
- d'autres commentaires jugés utiles (intérieur/extérieur, domicile/lieu public, conditions climatiques (vent, pluie), environnement sonore (calme, bruyant), etc.).

4. http://emeld.org

Nous ajouterons à ces informations concernant les enregistrements effectués pendant la séance de travail, le recueil d'informations sociolinguistiques concernant les locuteurs enregistrés (âge, sexe, nombre et nature des langues parlées, niveau d'études, métier, etc.). De plus, le linguiste traduit, sur le terrain, ses enregistrements fraîchement récoltés, avec un locuteur natif ou bien avec un traducteur-interprète. Ces détails sont des indications essentielles pour l'analyse des données post-terrain, mais également pour la conservation et la pérennité des données.

Si la durée du terrain le permet, le linguiste peut commencer la transcription phonétique des enregistrements audio avec d'autres spécialistes de la langue (présents sur place) ou bien avec des locuteurs natifs. La transcription phonétique des enregistrements audio sera une étape incontournable avant d'entamer le processus d'analyse de la langue, une fois revenu du terrain.

Le linguiste entame généralement aussi un travail de glose des transcriptions avec un informateur principal. Les gloses participent, par exemple, à la construction de dictionnaires et lexiques spécialisés. Gloser son document signifie identifier les morphèmes, les décrire et les expliquer, dans des exemples. Cela permet de comprendre la construction morphologique et syntaxique de la langue. La convention standard la plus suivie pour gloser les mots des transcriptions est celle de Bickel, Comrie et Haspelmath (2008) qui proposent 10 règles à suivre.

1.1.5 Au retour du terrain

Au retour du terrain, le linguiste trie ses enregistrements et documents grâce aux métadonnées qu'il a pris soin de noter durant son terrain.

S'en suit une série d'analyses linguistiques des différents documents récoltés, qui dépendra des questions de recherche préalablement établies. Néanmoins, comme évoqué précédemment, les analyses des enregistrements effectués doivent nécessairement passer par une étape de transcription phonétique (et/ou orthographique). La transcription a pour but d'écrire avec des symboles phonétiques (et/ou orthographiques) ce que le linguiste perçoit à l'oreille. Ce travail est laborieux et chronophage : le temps de transcription d'une heure d'enregistrement peut prendre jusqu'à six heures, selon que le linguiste connaît déjà la langue d'étude ou non (Sakel et Everett, 2012).
Au retour du terrain, le linguiste peut aussi terminer le travail de glose préalablement entamé lorsqu'il était présent sur place, à l'aide des traductions réalisées et gloses déjà apposées.

Par la suite, le linguiste découvre l'agencement des mots au sein des phrases ainsi que leur relation et met ainsi en lumière la syntaxe de la langue. Il peut alors écrire une grammaire de la langue.

Des logiciels permettent de transcrire, traduire et gloser manuellement un enregistrement sonore. Ils offrent quelques avantages indéniables comme la construction d'un travail organisé, l'automatisation de quelques tâches, mais aussi l'interopérabilité des fichiers générés. Ces logiciels seront décrits à la section 1.2.

1.1.6 Vers un linguiste de terrain assisté par la machine

Nous avons vu dans les sections précédentes que les linguistes peuvent faire face dans leur travail de documentation, de description et d'analyse des langues à des tâches extrêmement coûteuses, que ce soit en terme de moyen financier (comme la traduction) ou de temps (comme la transcription).
Les linguistes sont de plus en plus nombreux à utiliser les logiciels développés pour les aider, tant les avantages sont nombreux. En effet, ces outils numériques permettent au linguiste de traiter rapidement un grand nombre de données, d'automatiser certaines tâches, mais également de construire des bases de données, des ontologies, des lexiques, etc. dans lesquels ils peuvent aisément naviguer et ainsi gagner du temps dans la recherche d'informations. Ces logiciels sont, pour la plupart, gratuits et, surtout, participent à la numérisation des données qui facilite ainsi le partage des données.

Néanmoins, ces logiciels ne sont que des interfaces en ce sens que le linguiste doit saisir, dans un premier temps, toutes les informations manuellement. Les récentes avancées techniques en traitement automatique du langage naturel (TALN) et l'augmentation des vitesses de calcul des machines et des débits de transfert offrent de nouvelles perspectives enthousiasmantes pour aider le linguiste de terrain lors de ses travaux de transcription, de traduction ou d'analyse phonétique par exemple.

Ainsi, nous imaginons un linguiste de terrain aidé par la machine, c'est-à-dire doté d'outils pratiques de collecte de données et tirant profit des progrès constants en traitement automatique du langage naturel. À titre d'exemple, les systèmes de reconnaissance automatique de la parole (RAP) pourraient être utilisés afin de transcrire automatiquement des enregistrements audio. Ces systèmes pourraient aussi aider le linguiste dans ses analyses pointues de la langue pour effectuer par exemple des analyses phonétiques sur la totalité de ses corpus.

Dans le même état d'esprit, Bird (2016) évoque la notion de « cyber-linguiste ». Il imagine un expert des langues qui, dans le futur, devra utiliser voire même développer ses propres algorithmes pour que la machine puisse déchiffrer les langues éteintes et pour lesquelles aucun travail documentaire n'a pu être accompli avant son extinction (mais dont nous possédons malgré tout des enregistrements audio).

1.1.7 Résumé

Le linguiste de terrain recherche la structure interne des langues (souvent à tradition orale et peu, voire non écrites), en analysant des données qu'il a lui-même recueillies auprès des locuteurs natifs vivant au sein de leur communauté. Le travail de description qu'il opère passe par la transcription phonétique de la langue, l'élaboration de son système phonologique, la recherche de sa structure morphologique, l'élaboration d'une orthographe, d'un lexique, d'un dictionnaire, d'une syntaxe et d'une grammaire.

Depuis plusieurs années maintenant, les linguistes de terrain se mobilisent pour décrire et documenter des langues dites en danger. Ces langues sont parfois parlées par moins de 1 000

locuteurs et la principale cause de déclin est la transmission de la langue entre les générations qui ne perdure plus.

Avant d'aller sur le terrain, le linguiste doit penser et gérer plusieurs choses. Il doit avant tout penser à ses questions de recherche, auxquelles il voudra répondre par ses enquêtes de terrain. Il doit également réfléchir au temps qu'il lui faudra rester auprès des locuteurs et, donc, au budget dont il aura besoin. De plus, le linguiste prendra soin de vérifier les procédures administratives à remplir avant son entrée sur le territoire étranger ainsi que les vaccins et autres traitements médicaux à suivre, mais aussi de se renseigner sur les coutumes du pays d'accueil afin d'éviter tout désagrément. Enfin, il devra réunir le matériel indispensable à la collecte de ses données, comme un dictaphone, un carnet de terrain et un ordinateur portable.

Une fois sur le terrain, le linguiste rencontre et interroge ses informateurs/locuteurs. Il est important que les enregistrements audio qu'il produit soient eux-mêmes décrits afin de ne perdre aucune information qui pourrait s'avérer cruciale dans le futur. Pour cela, le linguiste saisit des métadonnées qui permettent de détailler le contexte de l'enregistrement, le but de la séance, l'environnement dans lequel est effectué la séance, mais également de donner des informations supplémentaires sur le locuteur et les autres participants éventuels.

Au retour du terrain, le linguiste, dans un premier temps, transcrit ses enregistrements audio. Cette tâche peut être très longue, de même que la traduction, qui la suit généralement. Grâce à ces étapes, le linguiste peut ainsi déterminer le système phonologique de la langue, mais aussi découvrir les morphèmes, plus petites unités de sens d'une langue. Les morphèmes et la structure tout entière de la langue sont mis en lumière au travers de la glose, étape nécessaire à la compréhension de la construction de la langue. Ensuite, le linguiste dégage les structures syntaxiques de la langue. Ces travaux aboutissent généralement à des dictionnaires et des grammaires de la langue.

Tout ce processus de description, bien que passionnant, prend énormément de temps. Ce temps pourrait être utilisé pour faire plus de terrain et approfondir les connaissances que le linguiste possède déjà. Dans ce but, des logiciels à destination des linguistes ont été développés et permettent d'automatiser certaines tâches, à condition d'avoir, au préalable, rempli, à la main, un certain nombre de critères voire même d'étapes (comme la transcription par exemple). À l'heure où les avancées en apprentissage automatique et technologies vocales sont importantes, nous avons imaginé équiper le linguiste de terrain d'outils comme les systèmes de reconnaissance automatique de la parole pour que certaines tâches deviennent moins chronophages et qu'ils puisse effectuer des analyses fines de la langue.

1.2 Outils numériques pour la documentation, la description et l'analyse des langues

Le constat de la disparition, inéluctable et de plus en plus rapide, des langues à travers le monde, a mené à des réflexions au sein de nombreux domaines, tels que la linguistique, l'anthropologie, l'informatique, faisant naître des consortiums en faveur de la préservation

des langues. Ces consortiums se sont concentrés sur l'aide à la documentation des langues et se sont interrogés sur comment les nouvelles technologies numériques pouvaient participer à cette conservation des langues. La création de projets comme le projet DoBeS (de son nom original "Dokumentation Bedrohter Sprachen") [5], financé par la fondation Volkswagen en Allemagne, le « Programme pour la documentation des langues en danger » (de son nom original "Endangered Languages Documentation Programme", souvent abrégé ELDP) [6] engagé par l'université de Londres (SOAS), le projet américain du consortium LinguistList appelé « Méta-structure électronique pour la documentation des langues en danger » (de son nom original "Electronic Meta-structure for Endangered Languages Documentation", souvent abrégé EMELD) ou encore le programme « Documentation des langues en danger » (Wright, 2004) ("Documenting Endangered Languages", abrégé DEL) initié par un partenariat entre la NSF ("National Science Foundation") et la NEH ("National Endowment for the Humanities") ont permis l'émergence de nouvelles formes de collectes de données, de nouveaux outils de traitement, mais aussi de nouveaux standards en matière de structuration et d'organisation des données.

Bien que la plupart du temps, sur le terrain, le linguiste n'a pas le temps ou les moyens (manque d'électricité par exemple) d'utiliser de logiciels particuliers pour traiter ses données, il est courant désormais qu'à son retour il en utilise. Les logiciels les plus répandus dans la communauté linguistique sont FLEx [7] (développé par la Société internationale de linguistique (SIL)), ELAN [8] (développé par l'équipe TLA (*The Language Archive*) du MPI-PL (l'Institut Max Planck pour la psycholinguistique) et Praat [9] (développé à l'université d'Amsterdam). Ils sont dédiés au traitement des données audio enregistrées sur le terrain (communément appelées, « données primaires »). Le traitement de ces données s'applique aux deux sous-disciplines de la linguistique de terrain que sont la documentation des langues et la description des langues.

Dans cette partie, nous allons présenter chacun de ces sous-domaines de la linguistique de terrain et quelques uns des outils qui leurs sont dédiés.

1.2.1 Aide à la documentation des langues

La documentation concerne la récolte des données primaires que sont l'enregistrement des locuteurs d'une langue, la transcription et la traduction de ces enregistrements (Himmelmann, 1998), mais aborde aussi l'archivage, l'accès, l'utilisation et la publication des corpus de données récoltées (Thieberger, 2013).

Ces préoccupations entraînent de nouvelles réflexions concernant la méthodologie de collecte et de gestion des données à adopter, d'autant plus que la mobilisation des linguistes de terrain pour la sauvegarde, le maintien et la diffusion des langues, couplés à l'utilisation récente des outils informatiques, conduisent à la création de grands volumes de données. Pour aider le linguiste de terrain dans ses différentes tâches de documentation, des outils informatiques et mobiles ont été développés.

5. http://dobes.mpi.nl/dobesprogramme/?lang=fr
6. http://www.eldp.net
7. https://software.sil.org/fieldworks
8. https://tla.mpi.nl/tools/tla-tools/elan
9. http://www.fon.hum.uva.nl/praat

1.2.1.1 Collecte

La documentation d'une langue commence avant tout par la collecte de données, notamment d'enregistrements audio dans lesquels un ou plusieurs locuteurs interviennent.

Il existe plusieurs façons de collecter de la parole. Le linguiste peut enregistrer le locuteur en lui demandant de :

- Raconter un conte, une recette de cuisine, des histoires sur la tribu, etc. On parle alors d'un corpus de parole spontanée.
- Lire ou traduire des textes, mots ou phrases : le type de parole enregistré est dit élicité, car le discours est totalement dirigé et contraint ;
- Commenter des textes, des images ou des vidéos : la parole collectée est dite semi-dirigée, car le linguiste conduit les informateurs vers des points précis — des protocoles d'enquêtes —, mais la parole est totalement libre.

La collecte à partir de sources visuelles est complexe. Elle nécessite de trouver des supports pertinents, adaptés à l'environnement et à la culture du locuteur. Ulinski *et al.* (2014) ont développé un logiciel appelé WordsEye Linguistics Tool (WELT). Cet outils implémente le moteur de génération de scènes 2D et 3D nommé WordsEye (Coyne et Sproat, 2001). Grâce au logiciel WELT, le linguiste de terrain peut, simplement à partir de données textuelles, créer des scènes sur-mesure, culturellement appropriées. Ces scènes servent, par la suite, de matériau d'élicitation. De plus, l'élicitation à partir d'images/vidéos offre l'avantage de rapporter un corpus texte-image précieux pour la documentation.

Les problématiques de la recherche en linguistique sont semblables à celles d'autres disciplines, comme les humanités, les sciences sociales et les sciences dites « dures » (Gawne *et al.* 2017). Le développement de plates-formes collaboratives pour les humanités numériques a, donc, naturellement inspiré des outils pour la documentation des langues. À titre d'exemple, Albright et Hatton (2008) ont développé WeSay, un logiciel libre d'accès et de droits (licence MIT) qui a pour but de préserver les langues en danger, en faisant participer les locuteurs à la documentation de leur langue, à travers le développement d'un dictionnaire. Néanmoins ce logiciel est réservé aux utilisateurs possédant un ordinateur qui fonctionne sous le système d'exploitation Windows.

Avec la démocratisation des outils mobiles, notamment les *smartphones*, les plates-formes collaboratives peuvent aujourd'hui être utilisées par la grande majorité des populations (Drude *et al.* 2013). En effet, l'engouement pour les appareils mobiles permettent de mettre à contribution les locuteurs de façon ludique et est un moyen supplémentaire de documenter des langues minoritaires ou en voie d'extinction. Drude *et al.* indiquent aussi que, grâce aux technologies mobiles, les collectes peuvent désormais être multimodales (audio, mais aussi captures d'images et de vidéos), ubiquitaires (car elles peuvent s'effectuer sans la présence du linguiste), et de ce fait, être réalisées avec un plus grand nombre de locuteurs. La démocratisation et la mobilité de ces outils permettent finalement de collecter un très grand nombre de données et ainsi rendre les collectes bien plus conséquentes.

Face à cette évidence, Bird, Hanke *et al.* (2014) ont fait le pari d'une application mobile à destination des locuteurs de langues en danger, afin de réunir le maximum de ressources langagières le plus rapidement possible. Cette application, nommée Aikuma, permet la collecte de parole en masse. Les auteurs expliquent que l'enregistrement et le stockage des données sont les deux aspects primordiaux de la démarche de sauvegarde d'une langue. Leur traitement et leur analyse pourront toujours être accomplis par la suite.

Dans une perspective de développement de technologies ayant besoin d'une grande quantité de données, comme la reconnaissance automatique de la parole, De Vries, Davel *et al.* (2013) ont montré que collecter rapidement un large corpus dans plusieurs langues est désormais possible grâce une application mobile, en récoltant 800h de données dans 11 langues d'Afrique du Sud, à partir de leur application mobile Woefzela (De Vries, Badenhorst *et al.* 2011)

1.2.1.2 Méthodologie de collecte

Si l'enregistrement des locuteurs n'est plus un problème en soi, la méthodologie à adopter peut en devenir un. En effet, le travail sur le terrain du linguiste ne consiste pas qu'à l'enregistrement sonore de la réalité linguistique ; il nécessite également, au minimum, une prise de notes. Le linguiste, aussi, dessine ou filme parfois des scènes de vie ou l'environnement naturel dans lequel vivent ces locuteurs, afin de documenter les pratiques culturelles. Il use aussi de la méthode d'élicitation, technique pour faire parler les locuteurs et faire émerger des paradigmes (de conjugaison par exemple). C'est cet ensemble qui représente la collecte sur le terrain. L'enjeu est de réfléchir à une méthode qui rend le linguiste de terrain productif, en associant toutes ces ressources qui accompagnent l'enregistrement audio.

L'application mobile pour la collecte de parole, Aikuma (Bird, Hanke *et al.* 2014), en plus d'être novatrice dans sa façon de collecter la langue grâce au téléphone mobile, propose une nouvelle méthodologie de collecte. Puisqu'il ne suffit plus seulement d'enregistrer, l'application permet de créer des répétitions (*respeaking*) d'enregistrements [10] ainsi que leur traduction dans une langue dominante. Ces deux nouveaux types d'enregistrements permettent ainsi de rendre la langue analysable même après que les derniers locuteurs ont disparu. Cette application étant à l'origine de Lig-Aikuma, notre application mobile de collecte de parole sur le terrain, nous en ferons une présentation détaillée dans le chapitre 3.

Après l'expérience d'Aikuma, les auteurs ont décidé de développer un nouveau modèle d'applications mobiles. Cette fois, ils ne proposent plus une application découpée en fonctionnalités, mais plutôt un environnement général intégrant une fonctionnalité par micro-application. Cet ensemble de prototypes d'applications mobiles, développé par Bettinson et Bird (2017), est destiné à l'élaboration d'activités qui permettent de documenter et ainsi soutenir et promouvoir la préservation des langues (en danger). Jusqu'à présent, 3 prototypes

10. Le *respeaking* est un véritable concept introduit par Woodbury (2003), qui le définit comme la répétition articulée, dans un environnement dépourvu de nuisances sonores, d'un enregistrement audio brut qui lui est recueilli sur le terrain dans des conditions d'enregistrement rarement optimales.

ont été développés dans ce but : Aikuma-NG, AikumaLink et Zahwa. Aikuma-NG [11] permet d'écouter et de transcrire un enregistrement, mais également de le répéter (comme le mode *re-speaking* d'Aikuma présentée dans le paragraphe ci-dessus) et de le traduire de façon orale ou écrite. L'utilisateur peut insérer plusieurs entrées de texte afin d'aligner la transcription (dans la langue source) et la traduction écrite. AikumaLink est une plate-forme en ligne permettant de mettre en relation un linguiste avec un ou plusieurs informateurs, dans le but d'effectuer une tâche (linguistique). Enfin, Zahwa permet de prendre des photos ou une courte vidéo puis de les commenter. Elle a été initialement prévue pour la description de discours procéduraux comme des recettes de cuisine, mais peut finalement être appliquée à bien d'autre activités similaires (comme par exemple pour oraliser de courts tutoriels).

1.2.1.3 Gestion et organisation des données

La gestion et l'organisation des données, une fois celles-ci collectées, est la principale source de difficulté du linguiste. Une bonne organisation consiste d'abord en une bonne représentation de l'ensemble du terrain — avant, pendant et après chaque session de travail —, aussi idéale soit-elle.

Thieberger et Berez (2012) constatent que les descriptions linguistiques sont rarement accompagnées de leurs données alors que celles-ci sont d'une grande valeur pour la reproductibilité, mais également la pérennité des travaux de recherche. Dans leur manuel *Linguistic data management*, les auteurs mettent à disposition, pour pallier ces lacunes, des conseils méthodologiques pour la gestion des données, de leur collecte jusqu'à leur archivage.

Dans ce sens également, Bowern (2015, chap.4) propose un découpage en tâches du travail de terrain à travers un diagramme d'activités. Dans ce diagramme, la majorité du travail après une session est consacrée aux données : leur organisation, leur archivage et leur traitement. Pour organiser ses données, le linguiste a besoin d'informations sur celles-ci. Pour cela, et comme présenté dans le diagramme de Bowern, la prise de note sur le terrain, qu'elle soit faite directement sur ordinateur ou sur papier, est essentielle. Des données organisées et adéquatement annotées mènent nécessairement à un travail de description efficace et ainsi à un linguiste productif.

Les annotations des données digitales sont souvent appelées « métadonnées ». D'après la définition de l'association des professionnels de l'information et de la documentation (ADBS) [12], les métadonnées sont un :

> « ensemble structuré de données créées pour fournir des informations sur des ressources électroniques. Elles peuvent remplir différentes fonctions : a) gestion des ressources décrites (suivi du cycle de vie : création, modification, archivage) ; b) informations sur le contenu de la ressource pour en faciliter la découverte, la localisation, l'accès ; c) suivi de l'utilisation et du respect des droits et conditions d'utilisation associés à la ressource. »

11. https://github.com/aikuma/aikuma-ng
12. http://www.adbs.fr

Les métadonnées sont, donc, des informations qui permettent de documenter (si possible avec détails) le matériau collecté. Typiquement, un enregistrement audio d'un locuteur n'est qu'un fichier parmi tant d'autres dans l'ordinateur du linguiste. Que ce soit dès son retour de terrain ou plusieurs années après, se souvenir simplement du nom du locuteur interrogé ou des conditions dans lesquelles l'enregistrement a été mené, est compliqué. Les métadonnées servent à apporter des informations de ce type.

Par ailleurs, Austin (2013, section 3) indique que les métadonnées sont des « données à propos des données » et qu'elles « sont requises non seulement pour l'archivage, mais aussi pour la gestion, l'identification, la récupération et la compréhension des données dans le projet de documentation, une fois que le traitement et l'ajout de valeur sont effectués. »

Dans cette perspective, Hatton (2013) a mis au point SayMore, un logiciel qui permet au linguiste d'enregistrer directement avec l'application et d'organiser ses données après collecte (même si elles ont été collectées avec un autre outil). Le logiciel offre une interface permettant au linguiste de pouvoir renseigner des métadonnées dans un format standard (IMDI, développé par TLA) à propos des sessions d'enregistrements, des locuteurs, mais aussi des fichiers récoltés. De plus, il permet de réaliser des transcriptions et des traductions alignées avec le signal audio. La faiblesse de ce logiciel est qu'il est uniquement disponible pour les utilisateurs du système d'exploitation Windows et ne propose l'archivage des données qu'au format propriétaire du SIL.

1.2.1.4 Archivage et partage

Himmelmann (1998, section 2) indiquait déjà que la documentation d'une langue ne concerne pas que la récolte des données. Elle nécessite également, de la part du linguiste, le partage de ses notes de terrain, afin de rendre les données compréhensibles et accessibles par tous. Pourtant, Thieberger (2013) signalait encore un véritable manque d'outils standards pour le partage et l'archivage des données, entraînant un manque de descriptions et de partage des données collectées, malgré le travail pionnier, fournit par Bird et Simons (2000), de création de la plate-forme de ressources linguistiques OLAC (de son nom original "Open Language Archives Community", que l'on peut traduire par « Communauté des Archives Linguistiques Ouvertes »). Cette bibliothèque virtuelle, en plus d'offrir un hébergement des ressources linguistiques, propose un guide des bonnes pratiques pour la construction des archives.

Cet état des choses a suscité plusieurs initiatives, comme la *Collection Pangloss*[13] (Michailovsky *et al.* 2014), développé depuis 1994 au laboratoire CNRS du LACITO, à Paris et qui met à disposition, gratuitement, des ressources numériques de langues, majoritairement en danger. Ces ressources sont constituées d'enregistrements audio qui sont accompagnés de leur transcription et/ou de leur traduction (alignés dans le temps avec le signal) ainsi que de métadonnées répondant aux critères de la "Dublin Core"[14] et de l'OLAC. En août 2017, l'atlas regroupait 160 langues et 2 952 enregistrements. Le catalogue ELAR ("The Endangered

13. http://lacito.vjf.cnrs.fr/pangloss

14. Le standard *Dublin Core* est un ensemble de recommandations pour la description des ressources numériques.

Language Archive") [15] hébergé par SOAS, propose en libre accès (après inscription gratuite) la consultation ou le dépôt de ressources multimodales (vidéo, audio, images, textes). Néanmoins l'archive doit être construite et déposée au moyen de l'outil LAMUS qui s'occupe de la gestion des archives et de leur téléversement (Broeder *et al.* 2006). Très récemment, Hendery et Burrell (2016) ont développé *PARADISEC* (qui signifie "the Pacific And Regional Archive for Digital Sources in Endangered Cultures") [16], afin de conserver et partager les sources documentaires numériques collectées jusqu'à présent dans la région Pacifique (qui s'étend entre l'Océanie, l'Asie de l'Est et du Sud-Est). Cette région abrite la plus grande diversité culturelle et linguistique mondiale, avec environ un tiers des langues du monde qui y est parlé, mais dont la plupart sont vouées à disparaître d'ici la fin du siècle. Les documents archivés suivent les standards recommandés par l'OLAC et le "Dublin Core".

Enfin, en plus d'être utile à des fins de recherche, les archives ouvertes représentent un formidable atout pour le développement de la connaissance des langues du monde auprès du grand public. Elles permettent de découvrir la richesse linguistique de manière ludique grâce à la visualisation des langues sur une carte interactive, mais contribuent également à leur pérennisation. Finalement, les atlas en ligne participent aussi à cette exploration des langues même s'ils ne font pas office d'hébergeurs de ressources. Nous pouvons citer *WALS*, l'atlas des structures des langues (Haspelmath *et al.* 2005) dont la première édition, en 2005, était accompagné d'un CD-ROM offrant une recherche interactive. Ce livre et CD-ROM regroupait, à l'origine, 140 bases de données typologiques — c'est-à-dire qui fournissaient des descriptions sur les langues comme des traits phonologiques, morphologiques, syntaxiques, etc. — reliées par une structure métalinguistique commune. Cet atlas est en ligne depuis 2008 [17]. Bien que cette base de données soit d'une grande richesse pour la description des langues, elle ne propose pas d'importation des ressources et les données sources sur lesquelles les descriptions proviennent ne sont pas mentionnées. Sur le même modèle, Michaelis *et al.* (2013) ont aussi construit APICs [18], un atlas en ligne fournissant des informations sur la structure de 76 créoles, pidgin et langues mixtes, dont les sources sont cette fois indiquées. Nous pouvons également citer l'enrichissement continu, depuis 2005, de l'atlas linguistique Algonquian [19] à partir duquel nous pouvons écouter des locuteurs de dialectes canadiens ou encore la création de l'atlas linguistique de la région de Manang au Népal [20], qui réunit des photos et vidéos et participe ainsi à la documentation des langues en danger de cette région.

Dans un but collaboratif, Beermann, Hellan et Brindle (2006) ont développé un outil de partage et de travail collaboratif appelé TypeCraft. TypeCraft participe à l'échange de données linguistiques grâce à l'élaboration d'exemples de gloses qui peuvent ensuite être exportés au format XML pour être intégrés dans d'autres travaux (Beermann et Mihaylov, 2014).

15. https://elar.soas.ac.uk
16. http://www.paradisec.org.au
17. http://wals.info
18. http://apics-online.info
19. http://www.atlas-ling.ca
20. http://www.siue.edu/~shu/nepal8.html

1.2.2 Aide à la description des langues

Par rapport à la linguistique documentaire, la linguistique descriptive s'occupe de l'annotation des données collectées sur le terrain (après qu'elles ont été transcrites et traduites) ainsi que de leur analyse. Le travail de description permet de dégager la structure de la langue selon différentes granularités d'analyse (phonologie, morphologie, lexique, syntaxe, ...). Puisque la linguistique descriptive couvre plusieurs domaines, plusieurs outils existent, chacun dédié à un sous-domaine.

1.2.2.1 Transcription et annotation

Le premier travail du linguiste, une fois les entretiens effectués, consiste à mettre sous forme écrite le contenu de l'enregistrement audio. Cette étape est celle de la transcription. La transcription peut être définie comme la mise à l'écrit d'un flux. Pour le linguiste, la transcription permet d'écrire sous forme phonétique ou orthographique le contenu de ses entrevues de terrain. Pour ce travail, Barras *et al.* (1998) ont développé, en 1998, Transcriber. Cet outil permet de structurer les transcriptions : l'utilisateur découpe le signal acoustique et transcrit chaque segment. L'utilisateur peut aussi commenter l'enregistrement, en indiquant par exemple la présence de chevauchements dans les tours de parole, de silences ou bruits ambiants, de bruits de bouche et de respiration, etc. L'encodage unicode est supporté et les fichiers de transcriptions et d'annotations sont exportables au format XML. Le logiciel continue d'être maintenu et est disponible pour Windows, Linux et MacOS. Transcriber supporte les fichiers au format SGML (générés par *SCLITE*, un outil d'évaluation des systèmes de RAP). Ce format facilite la manipulation des données par les linguistes informaticiens. L'interface de Transcriber permet de visualiser, graphiquement, les différences entre une transcription manuelle (faite par un humain) et une transcription automatique (fournie par un système de reconnaissance).

Plus utilisé aujourd'hui, parce que plus complet, ELAN (acronyme signifiant "EUDICO Linguistic ANnotator") est un logiciel libre (licence GPL 2) pour l'annotation de corpus multimodaux (Wittenburg *et al.* 2006). L'annotation d'une vidéo est un atout majeur du logiciel, car cela facilite l'analyse de la langue des signes par exemple. L'annotation peut être réalisée à différents niveaux linguistiques (mot, glose, phrase, etc.), mais peut aussi correspondre à un commentaire ou à une traduction du fichier audio ou vidéo. Chaque niveau est représenté par une *tier* (qui correspond au champ dans lequel l'annotateur écrit). Cette structure permet d'aligner temporellement le son et les annotations. De plus, les *tiers* sont hiérarchiquement dépendantes les unes des autres. Avec ELAN, il est possible d'extraire les annotations, comme les listes de gloses et de morphèmes. Comme Transcriber, l'encodage unicode est supporté et les fichiers d'annotation sont structurés au format XML. ELAN peut être lancé sur Windows, Linux et MacOS. Dans la même catégorie, nous pouvons aussi citer le logiciel d'annotation de dialogue multimodal Anvil (Kipp, 2001). À l'origine, ce logiciel a été pensé pour l'étude du comportement, il est aujourd'hui utilisé dans bien d'autres domaines. La visualisation de l'enveloppe acoustique du signal audio ou de la hauteur tonale sont des fonctionnalités in-

téressantes pour le linguiste, tout comme la possibilité d'importer des fichiers créés avec le logiciel Praat. À l'instar d'ELAN, Anvil est compatible avec les principaux systèmes d'exploitation informatiques. La version 6 du logiciel est disponible gratuitement [21] depuis le mois d'août 2017.

Une version étendue de ELAN a été développée au laboratoire parisien LLACAN (acronyme de Langage, LAngues et Cultures d'Afrique Noire), appelée ELAN-CorpA [22]. Cette extension a été développée à l'origine dans le cadre du projet ANR CorpAfroAs, qui visait à la création de ressources audio pour les langues de la famille Afro-Asiatique (Adamou, 2016). ELAN-CorpA inclut une fonctionnalité d'interlinéarisation qui génère une nouvelle *tier* segmentée qui contient les gloses, à partir des *tiers* déjà saisies. Auparavant, il fallait passer par FLEx (ou Toolbox) pour réaliser ce travail, alourdissant les processus d'analyses linguistiques. L'interlinéarisation est réalisée à partir d'un dictionnaire qui peut être importé depuis FLEx (ou Toolbox).

1.2.2.2 Analyse lexicographique

Pour l'étude lexicographique et le découpage morphosyntaxique, les logiciels les plus utilisés par la communauté linguistique à l'heure actuelle sont Toolbox et FLEx.

FLEx (acronyme provenant de Fieldworks Language Explorer) [23] et Toolbox sont des logiciels informatiques gratuits, développés par le SIL et destinés aux linguistes de terrain. Ils sont dédiés à l'édition de dictionnaires (Rogers, 2010). FLEx est le successeur direct de Toolbox [24]. Les fonctionnalités de ces deux logiciels sont similaires ; nous n'allons développer ici que les fonctionnalités du logiciel le plus récent, FLEx. Néanmoins, nous précisons que, même si Toolbox n'est plus maintenu par le SIL, ce logiciel est encore couramment utilisé par les linguistes qui ont commencé à construire leurs lexiques avec celui-ci. Bien que FLEx propose l'importation des ressources construites originellement avec Toolbox, des problèmes de compatibilité entre les deux logiciels ont été observés sur certaines fonctionnalités. Le manque d'interopérabilité entre les logiciels (et entre les systèmes d'exploitation) est d'ailleurs un gros problème et décourage les linguistes de terrain d'utiliser les outils numériques récents. FLEx rassemble plusieurs outils pour la création de lexiques monolingues. À partir de corpus textuels, il permet de découper les phrases en mots. Pour chaque mot, le linguiste peut définir le sens, le genre, indiquer si nécessaire les variantes dialectales, choisir une phrase qui illustrera le mot en contexte (au moyen d'un concordancier), … finalement, tout ce qui entre en jeu dans la création d'un dictionnaire. FLEx intègre également un module d'étiquetage morphosyntaxique et d'interlinéarisation des gloses. Plus le linguiste apportera d'informations sur une entrée, plus l'analyseur sera performant. Le linguiste a également la possibilité de conceptualiser chaque entrée de son lexique selon une liste de domaine de connaissances (Univers, êtres humains, langage, comportements sociaux, etc.). Cette classification permet en fin de compte de créer

21. http://www.anvil-software.org
22. http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php
23. https://software.sil.org/fieldworks
24. https://software.sil.org/toolbox

des ontologies de domaine. Bien qu'il soit possible de travailler à un niveau phrastique ou textuel, FLEx est plus orienté et adapté pour la lexicographie. Il ne propose pas, par exemple, d'étiquetage d'une phrase entière ou d'un ensemble de constituants, ni une recherche sur plusieurs étiquettes à la fois, dans un texte.
Ce logiciel délaisse les utilisateurs de Mac, car il ne fonctionne que sur les systèmes d'exploitation Windows et Linux. Il propose une interface en 13 langues (anglais, français, hindi, portugais, russe, espagnol, indonésien, malaisien, coréen, persan, chinois, turque, télougou). Pour le traitement automatique, Craig Farrow a développé une librairie appelée FlexApps [25], qui permet de manipuler grâce au langage de programmation Python, la base de donnée FLEx. Ce module peut conjointement être utilisé avec la boîte à outils NLTK (acronyme de Natural Language Toolkit) dédiée au traitement automatique de données linguistiques avec Python (Bird, 2006). À noter que les données au format Toolbox sont également exploitables par NLTK, grâce au module du même nom *toolbox* (voir Bird, Klein et Loper (2009, section 11.5)).

1.2.2.3 Analyse phonétique et prosodique

Pour cette expertise, le logiciel largement utilisé par la communauté des linguistes et phonéticiens est Praat (Boersma *et al.* 2002). Praat rassemble plusieurs outils dédiés à l'analyse fine de la parole, qui s'étend de l'annotation du signal acoustique (temporellement aligné) à l'analyse de la prosodie. Les forces principales de ce logiciel sont la visualisation du spectrogramme, l'affichage et le calcul automatique de paramètres acoustiques tels que les formants ou le pitch. Les annotations se font dans des *tiers*, rassemblées dans un fichier appelé *Textgrid*. Praat est multiplateforme et est distribué sous la licence publique générale GNU, ce qui le rend évolutif ; il bénéficie en conséquence d'une grande communauté de développement de scripts [26] et modules additionnels. Nous pouvons citer EasyAlign [27] (Goldman, 2011) qui aligne de façon automatique un signal et sa transcription orthographique et génère automatiquement de nouvelles segmentations des transcriptions à différents niveaux linguistiques (phonèmes, syllabes, mots) ou des programmes d'étiquetage semi-automatique ou automatique de la prosodie d'un signal tels que Prosogram (Mertens, 2004), MOMEL-INTSINT (Hirst, 2007), SLAM (Obin *et al.* 2014)).

Dans la même catégorie, Martin (2004) a développé WinPitch. Les fonctionnalités offertes par ce logiciel sont similaires à celles de Praat. WinPitch propose, en plus, l'import de vidéos, la visualisation en temps réel des annotations prosodiques sur le signal et l'alignement à la volée des transcriptions provenant d'autres applications (comme Transcriber par exemple). De plus l'interface est plus agréable et ergonomique que celle de Praat. L'inconvénient majeur est qu'il ne fonctionne que sous Windows et n'est gratuit que pour un usage personnel. Enfin, une version nommée WinPitch LTL est dédiée à l'enseignement de la prononciation.

Toujours destiné à l'annotation, EMU-SDMS ("The EMU Speech Database Management System") regroupe un ensemble d'outils permettant à la fois de créer, manipuler et analyser

25. https://github.com/cdfarrow/FLExTools/wiki
26. http://phonetics.linguistics.ucla.edu/facilities/acoustic/praat.html
27. http://latlcui.unige.ch/phonetique/easyalign.php

des bases de données vocales. Il peut être utilisé grâce à une interface en ligne appelée EMU-webApp [28]. L'utilisateur peut importer un fichier audio au format WAV seul ou accompagné de sa transcription au format annotJSON (généré avec EMU-SDMS) ou au format TextGrid (généré avec Praat) dans EMU-webApp. Le fichier audio importé est ensuite représenté sous forme de signal acoustique et de spectrogramme, qui permettent à l'utilisateur d'annoter ses segments. EMU-SDMS propose une classification hiérarchique des segments de parole, qui peut être visualisée sous forme d'arbre dans EMU-webApp. Chaque niveau de l'arbre correspond à une granularité linguistique (intonationnelle, lexicale, syllabique, phonémique, phonétique). Cette représentation offre la possibilité à l'utilisateur de rechercher des segments de parole en fonction de leur hiérarchie dans l'énoncé transcrit, via le moteur de recherche intégré dans EMU-SDMS.

Pour effectuer des analyses phonétiques quantitatives, FAVE (Rosenfelder *et al.* 2011), ProsodyLab (Gorman, Howell et Wagner, 2011) et SPPAS (Bigi et Hirst, 2012) sont des outils qui proposent des alignements automatiques (dits forcés) de phonèmes et des analyses statistiques de ces alignements. Nous ne les développerons pas dans cette thèse, mais ils méritent toutefois d'être cités.

Enfin, une liste non exhaustive, mais néanmoins étoffée, des logiciels à destination des linguistes peut être consultée sur le site Web de *The Linguist List* [29].

1.2.3 Résumé

Depuis quelques années maintenant, nous observons l'émergence d'applications, notamment mobiles, permettant d'obtenir des enregistrements de locuteurs des langues menacées d'extinction. La collecte de ces données vocales permet d'archiver, de préserver voire de revitaliser ces langues. Ces applications tirent profit de la démocratisation des technologies mobiles sur tablettes et *smartphones*. Les locuteurs peuvent désormais être acteurs dans la préservation de leur patrimoine culturel immatériel que sont leur(s) langue(s), leurs arts, leurs pratiques sociales, leurs rituels, leurs connaissances scientifiques et leurs savoir-faire traditionnels. L'émergence de ces applications rend maintenant les collectes et analyses linguistiques quantitatives en plus d'être qualitatives. Mais finalement, nous observons que les outils de collecte sont majoritairement destinés à la documentation des langues par les locuteurs et non par les linguistes. Il n'existe pas encore d'application (mobile) d'aide à la collecte de la parole destinée au linguiste et comportant une méthodologie de collecte réfléchie pour le travail de linguistique de terrain.

Après ce premier enjeu que représente la collecte des langues en danger, peu voire non connues, le second enjeu est de pouvoir les traiter *a posteriori* afin de les rendre exploitables dans le futur et ainsi préserver la langue en question. Cependant, le traitement des données ne peut être rendu possible qu'avec la coopération des linguistes, en rendant leurs corpus accessibles. Cette accessibilité passe indéniablement par la documentation des données (le renseignement des métadonnées), leur partage et leur archivage. Tout ce travail participe à la

28. http://ips-lmu.github.io/EMU-webApp
29. http://linguistlist.org/sp/GetWRListings.cfm?WRAbbrev=Software

pérennisation des langues qui bientôt vont s'éteindre. Toutefois, force est de constater qu'il existe toujours, au sein de la communauté linguistique, un grand manque d'outils pour la description et le partage des données collectées (Holton, Hooshiar et Thieberger, 2017) tout comme un manque de transparence dans les méthodologies adoptées (Gawne *et al.* 2017).

Deux freins majeurs ralentissent le travail du linguiste de terrain : la non-standardisation des outils numériques et des méthodes d'annotation, ainsi que le temps de traitement et d'analyse des données. En 2016 s'est tenue à Melbourne, en Australie, la conférence "Language Documentation Tools Summit" [30]. Cette conférence, qui a rassemblé linguistes et informaticiens autour de thématiques centrales en documentation et description des langues, a montré le besoin de réflexion autour du développement d'outils numériques pour la linguistique de terrain.

Une nouvelle discipline à part entière a vu le jour depuis quelques années, grâce aux linguistes et informaticiens œuvrant ensemble. Elle est appelée « linguistique informatique » ou parfois « linguistique computationnelle ». Puisqu'un ordinateur est capable de traiter un très grand nombre de données en un temps extrêmement réduit, l'idée est d'automatiser certaines tâches voire même d'apprendre à la machine comment repérer certains motifs répétitifs ou points saillants dans une langue à des fins de transcription, de traduction, mais aussi d'analyse. Les outils comme Elan, FLEx ou Praat restent indispensables pour des descriptions et analyses détaillées et l'élaboration *a posteriori* de lexiques et grammaires. Néanmoins de nouvelles techniques permettent aujourd'hui d'aider le linguiste dans son travail de transcription phonétique et orthographique, mais aussi de découverte morphologique et syntaxique de la langue (cette partie est détaillée au chapitre 4 et au chapitre 5).

En plus des logiciels existants, les techniques en linguistique computationnelle permettent désormais d'aller encore plus loin dans la recherche en typologie des langues, de leur origine jusqu'aux variations intralinguistiques.

Dans le domaine de la dialectologie, les outils pour l'analyse dialectométrique informatisée sont aussi en plein essor. Initiée par Séguy (1973), la dialectométrie — qui regroupe un ensemble de méthodes d'analyses statistiques et de classifications pour les données dialectales — a donné lieu à des avancées majeures en géolinguistique. Cependant ce n'est que depuis quelques années que les techniques réapparaissent, grâce à l'automatisation des calculs et aux progrès en matière de capacité de stockage, en informatique. Nerbonne *et al.* (2011) ont développé *Gabmap*, une application en ligne, libre d'accès et de droits, qui offre des outils pour aider le dialectologue néophyte en statistiques informatiques à analyser ses données et à en interpréter les résultats, notamment grâce à la construction de cartes et de graphes. Aurrekoetxea *et al.* (2013) proposent *DiaTech*, également un outil en ligne qui permet l'analyse et l'interprétation de données dialectales, mais qui intègre, en plus, une prise en charge des réponses multiples (fournies par les participants aux enquêtes linguistiques) dans les calculs statistiques ainsi que l'import de bases de données créées avec des outils externes.

30. https://sites.google.com/site/ldtoolssummit/home

Concernant les études en elles-mêmes, Burridge (2017) a analysé l'évolution des isoglosses [31] en Grande-Bretagne à partir d'un modèle mathématique. Appliqué à l'anglais britannique, il a montré que son modèle spatial est, d'une part, capable de reproduire les observations et les prédictions des dialectologues et d'autre part, capable de nouvelles prédictions sur l'évolution de la langue, notamment en fonction de l'âge des locuteurs. Rahimi, Cohn et Baldwin (2017) ont entraîné un modèle de géolocalisation à partir de textes en utilisant un perceptron multicouche. Ce modèle leur a permis de rendre compte de la variation diatopique lexicale aux États-Unis, à partir de la géolocalisation d'échanges de *tweets*. À ces fins, ils ont d'ailleurs créé leur propre jeu de données appelé DAREDS à partir du Dictionnaire de l'anglais régional américain [32] (abrégé *DARE*, de son nom original *Dictionary of American Regional English*) qui regroupe le vocabulaire régional américain. Ils s'en sont servis pour évaluer la performance de leur système à retrouver le lexique associé à une zone géographique.

1.3 La reconnaissance automatique de la parole (RAP)

1.3.1 Principes généraux de la RAP

La reconnaissance automatique de la parole permet de rendre le language humain compréhensible par la machine.

D'un point de vue plus technique, la reconnaissance automatique de la parole continue consiste à transcrire un signal acoustique en une suite de mots écrits. Un système de RAP est capable d'analyser et d'interpréter des séquences acoustiques puis, de prendre une décision sur la séquence phonétique ou lexicale qui représentera le mieux une séquence acoustique donnée. Les ressources nécessaires, pour la construction d'un système de reconnaissance de la parole, sont les suivantes (Pellegrini, 2008) :

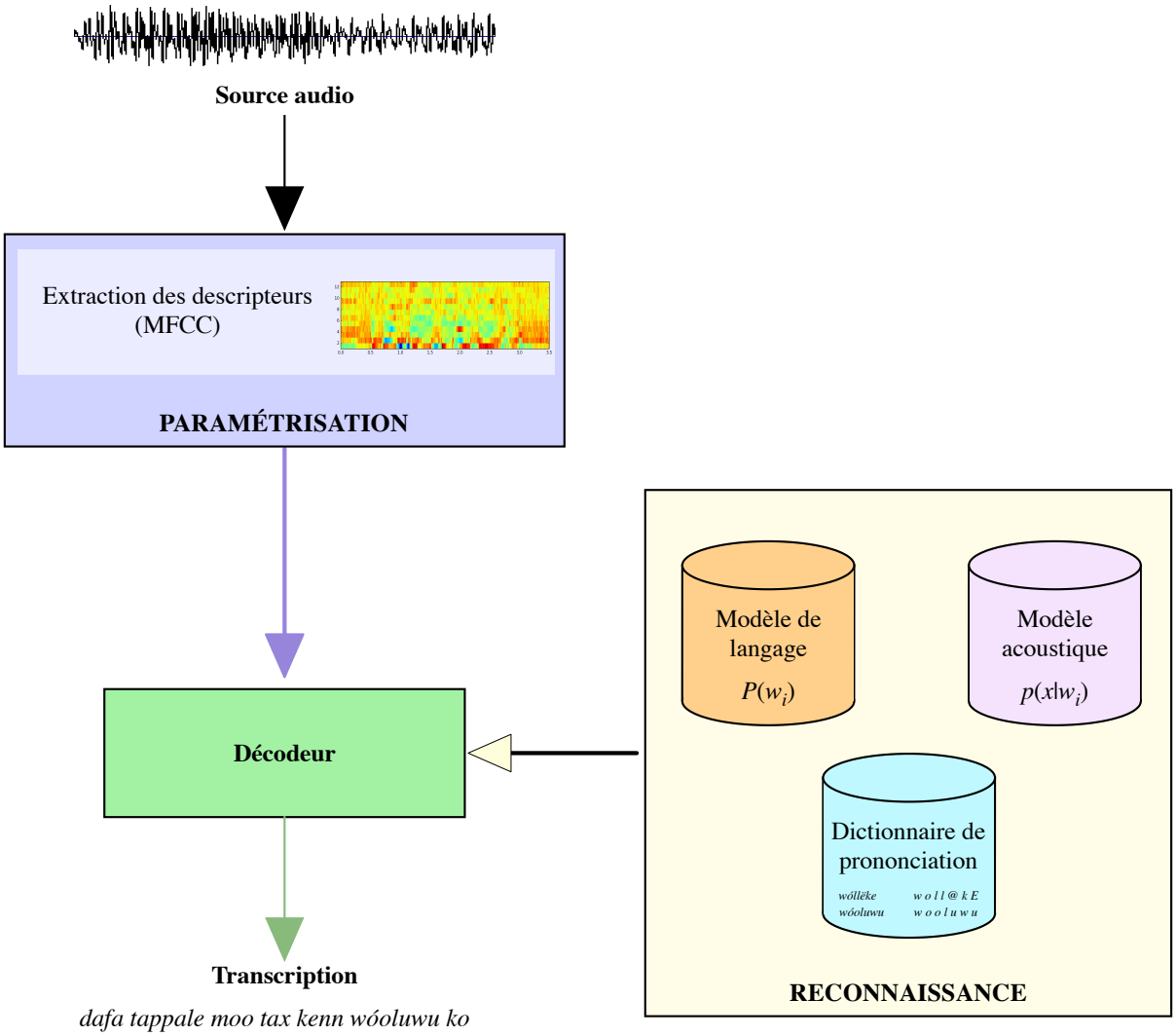
- un grand corpus de textes (~10M–100M mots), pour construire un modèle de langue ;
- un grand corpus audio (~10h–100h) transcrit, contenant des enregistrements audio qui seront transformés en vecteurs acoustiques, utilisés pour construire des modèles acoustiques ;
- un lexique de mots suivis de leur prononciation, qui permettra au système de faire le lien entre le modèle de langue et le modèle acoustique [33].

Pour que le système de RAP soit performant (c'est-à-dire qu'il reconnaisse correctement ce qui est dit dans l'enregistrement audio), ces 3 ressources doivent être représentatives de la langue à reconnaître. Cela signifie que le corpus textuel doit posséder la plus grande variété lexicale, morphologique et syntaxique possible. De la même manière, le lexique doit être le plus exhaustif possible, tant au niveau de la variété des lexèmes que dans la variété des prononciations. Enfin le corpus audio doit représenter, idéalement, le nombre de phonèmes présents

31. En linguistique, une isoglosse est la frontière qui délimite les aires géographiques d'un dialecte.
32. http://www.daredictionary.com
33. De récentes approches issues de l'apprentissage profond permettent désormais de se passer d'une telle ressource.

dans la langue ainsi que la totalité de leur réalisation phonétique (combinaisons phonétiques incluses) mais aussi doit contenir un panel de locuteurs variés.

Le processus de reconnaissance de la parole peut finalement être découpé en deux étapes : la paramétrisation et le décodage, illustrés sur le graphique 1.1.



Graphique 1.1 – Diagramme d'un système de reconnaissance de la parole.

1.3.1.1 La paramétrisation

La première étape d'un système de reconnaissance de la parole est d'extraire des caractéristiques pertinentes pour l'identification du contenu linguistique, en rejetant les autres informations contenues dans ce signal. En effet, le système de reconnaissance ne se sert pas directement du signal sonore brut : il exploite la représentation paramétrique de ce signal. Pour cela, des descripteurs sont extraits du signal acoustique et sont réunis dans des vecteurs (à n dimensions). Les descripteurs initialement exploités en RAP (après calcul de la transformée de Fourier discrète) sont les coefficients Mel cepstraux (abrégés "MFCC"). Ces coefficients sont généralement associés à d'autres (prédiction linéaire, énergie, ondelettes, delta et delta-delta, etc.) afin d'affiner la représentation du signal (Hacine-Gharbi, 2012).

1.3.1.2 Le décodage

Le décodage est l'étape durant laquelle le système décode le signal acoustique en le transcrivant. Auparavant, l'apprentissage, également appelé « entraînement », est l'étape où le système apprend à reconnaître un son et à l'associer avec un symbole, tandis que le décodage est l'étape où le système transcrit le signal de parole contenu dans un enregistrement inconnu.

Trois éléments clés interviennent dans la reconnaissance de la parole : un modèle acoustique, un dictionnaire de prononciation et un modèle de langue.

La phrase de décodage peut être représentée par la formule suivante :

$$\begin{aligned} w^* &= \arg\max_i \left[\frac{p(x|w_i).P(w_i)}{p(x)} \right] \\ &= \arg\max_i \left[p(x|w_i).P(w_i) \right] \end{aligned} \tag{1.1}$$

où x représente un vecteur acoustique et w_i un mot ou un phonème, $p(x|w_i)$ représente la probabilité du modèle acoustique et $P(w_i)$ la probabilité du modèle de langue.

≻ Le modèle acoustique

Il existe plusieurs techniques de modélisation acoustique. Les modèles statistiques étaient les modèles classiquement utilisés en RAP mais, depuis 2014, l'augmentation de la puissance de calcul des ordinateurs a entraîné un regain d'intérêt pour les réseaux de neurones. Ces réseaux sont aujourd'hui massivement étudiés, car les performances obtenues surpassent celles des systèmes à base de mélanges gaussiens.

Les techniques de modélisation acoustique varient selon l'architecture du réseau de neurones choisie. Il existe, par exemple, des modèles entraînés à partir de réseaux de neurones profonds (abrégé "DNN" pour *Deep Neural Networks*) (Hinton *et al.* 2012), les modèles à base de réseaux de neurones récurrents ("RNN" pour *Recurrent Neural Networks*, "LSTM" pour *Long Short-Term Memory*) (Graves, Mohamed et Hinton, 2013 ; Sak, Senior et Beaufays, 2014) ou encore les réseaux de neurones convolutifs (communément abrégé "CNN" pour *Convolutional Neural Networks*) (Abdel-Hamid *et al.* 2014).

Dans cette thèse, nous avons utilisé les DNN pour entraîner nos systèmes de RAP. Ce sont eux que nous évoquerons rapidement.

1. **Approche à base de mélange de gaussiennes (HMM/GMM)**. Ce modèle statistique est fondé sur des chaînes de Markov cachées (abrégé "HMM" pour *Hidden Markov Model*), généralement à 3 états, qui modélisent la variation de prononciation d'une même unité acoustique. À chaque état correspond un modèle multigaussien (abrégé "GMM") qui permet de modéliser la forme d'onde du signal de parole. Dans le cas des modèles acoustiques en contexte, chaque phonème possède un modèle par contexte existant. Ces premières modélisations du signal peuvent ensuite être affinées par plusieurs techniques d'adaptation au locuteur et par l'utilisation de SGMM ("Subspace Gaussian Mixture Models ") qui considèrent des sous-espaces de mélanges de gaussiennes contrairement aux GMM (Povey, Burget *et al.* 2011). Lors du décodage de nouveaux enregistrements par

le système, le modèle acoustique attribue, sur les nouvelles données, des scores de vraisemblance pour chaque phonème s'appuyant sur les modèles HMM appris.

Les scores permettent d'extraire des mots grâce à un dictionnaire de prononciation. Ce dictionnaire contient *a minima* les mots présents dans le corpus oral utilisé suivis de leur prononciation phonétique ainsi que toutes les variantes de prononciation possibles.

2. **Approche à base de réseaux de neurones profonds (DNN)**. Dans ce modèle, le réseau de neurones remplace le GMM (ou SGMM) pour modéliser la forme d'onde du signal. Les réseaux de neurones profonds sont composés de neurones pleinement interconnectés, répartis en couches. Ils sont dits « profonds » lorsque le réseau possède plusieurs couches cachées. Un réseau de neurones profond possède, donc, une couche d'entrée qui correspond aux valeurs pondérées du vecteur d'entrée, une couche de sortie qui rapporte les sorties du réseau et, entre elles, des couches cachées qui exécutent des transformations. Des biais sont aussi ajoutés à la couche d'entrée et aux couches intermédiaires.

   Dans les DNN utilisés pour la RAP, chaque sortie d'une couche l du réseau peut être définie comme suit :

$$\mathbf{x}_l = \sigma\left(\mathbf{b}_l + \mathbf{W}_l \mathbf{x}_{l-1}\right) \quad 1 \leq l < L \tag{1.2}$$

   où $\mathbf{b}_l$ symbolise le vecteur du facteur de biais, $\mathbf{W}_l$ est une matrice, de taille $(n \times |x_{l-1}|)$ — avec n le nombre de neurones dans la couche l —, qui représente les vecteurs de poids des neurones de la couche l. σ représente la fonction d'activation qui peut être une sigmoïde telle que :

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.3}$$

   La dernière couche (la *L-ième*) du réseau utilise la fonction softmax pour déterminer la probabilité *a posteriori* de chaque état HMM s, étant donné l'observation acoustique $\mathbf{o}_t$ à l'instant t :

$$p(s|\mathbf{o}_t) = \frac{e^{x_L}}{\sum_{s'} e^{x_L}} \tag{1.4}$$

   L'optimisation des couches cachées peut être réalisée par pré-entraînement du réseau, en utilisant des machines de Boltzmann réduites (en abrégé, "RBM" pour "Restricted Boltzmann Machines") (Hinton, 2010). Ce pré-entraînement en amont du DNN peut être effectué de manière non supervisée.

≻ Le dictionnaire de prononciation

Le dictionnaire de prononciation joue un rôle essentiel dans la reconnaissance automatique de la parole. En effet, c'est lui qui permet de faire le lien entre la forme de surface d'un mot

et sa représentation phonétique. Le dictionnaire est utilisé par deux fois. Une première fois, durant l'étape d'apprentissage : il permet de construire des modèles acoustiques en associant les représentations acoustiques de chaque unité lexicale avec leurs symboles phonétiques. Une deuxième fois, lors de l'étape de décodage : il est conjointement utilisé avec le modèle de langue et le modèle acoustique pour transcrire un signal inconnu.

≻ Le modèle de langue

Les suites de mots obtenues lors de la phase de décodage sont évaluées avec un modèle de langue. Ce modèle est entraîné à partir d'un corpus textuel (comprenant notamment les transcriptions du corpus d'apprentissage utilisées pour la génération du modèle acoustique). Les modèles communément utilisés sont des modèles dits empiriques. Différentes techniques de modélisation existent : les modèles probabilistes, traditionnellement utilisés, et les modèles à base de réseaux de neurones, de plus en plus employés, car montrant des performances supérieures aux modèles classiques.

1. **Les modèles de langue probabilistes.**

   Les modèles de langue traditionnellement utilisés en RAP sont des modèles statistiques. Ils sont élaborés à partir de lois de probabilités conditionnelles, par estimation de la vraisemblance maximum qu'une suite de mots se réalise effectivement dans la langue. Le modèle de langue dépend, donc, directement de la langue à reconnaître ainsi que de la grammaire de cette langue. La loi des chaînes est généralement utilisée pour représenter le calcul de la probabilité $P(W)$ d'une séquence de mots $W = w_1, w_2, w_3, \ldots, w_n$. Elle se décompose de cette façon :

$$\begin{aligned} P(W) &= P(w_1, w_2, w_3, \ldots, w_n) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_n|w_1, w_2, \ldots, w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|w_1, w_2, w_3, \ldots, w_{i-1}) \\ &= \prod_{i=1}^{n} P(w_i|h_i) \end{aligned} \tag{1.5}$$

   où w_i représente un mot, n est le nombre total de mots de la séquence et h correspond à l'historique du mot.

   L'équation 1.5 montre qu'un mot peut être déduit, en théorie, de l'intégralité de ses prédécesseurs. Cette formule n'est pas réaliste, car les modèles seraient bien trop longs à estimer, volumineux à stocker.

   Pour réduire ce contexte, un modèle d'ordre $n-1$ peut être utilisé, où n permet d'ajuster la taille de l'historique. Ce modèle est appelé *n-gramme*. En RAP, il est commun de modéliser la langue avec des trigrammes, pour obtenir un modèle d'une taille raisonnable qui estime de manière satisfaisante les probabilités d'apparition des mots dans la langue. Par réécriture de l'équation 1.5, ce modèle peut alors être représenté comme suit :

$$P(W) = \prod_{i=3}^{n} P(w_i | w_{i-2}, w_{i-1}) \tag{1.6}$$

La formule de l'équation 1.6 montre, par conséquent, qu'un modèle trigramme peut prédire un mot en fonction des deux mots le précédant. Ce modèle est aussi appelé modèle d'ordre 2.

2. **Les modèles de langue neuronaux.**

   Tout comme pour la modélisation acoustique, les réseaux de neurones sont apparus, ces dernières années, pour la modélisation du language. Ils font désormais partie des systèmes état de l'art de RAP à grand vocabulaire. Nous pouvons retrouver dans la littérature actuelle, plusieurs types de réseaux adoptés pour l'apprentissage linguistique : DNN (Arisoy *et al.* 2012), RNN (Mikolov, Karafiát *et al.* 2010) et LSTM (Sundermeyer, Schlüter et Ney, 2012) ou encore CNN, dont l'utilisation est récente (Pham, Kruszewski et Boleda, 2016). À titre informatif, les modèles de type LSTM sont actuellement ceux qui obtiennent de meilleurs résultats, mais nous ne les avons pas utilisés dans cette thèse.

   Les réseaux de neurones sont créés à partir de projection de mots dans des espaces vectoriels de faible dimension. Ces espaces rassemblent les mots qui partagent des similarités sémantiques et syntaxiques. Dans les modèles récurrents, l'information partagée entre les neurones permet au réseau de garder un historique complet de ce qui a déjà été vu auparavant. Contrairement aux modèles statistiques qui utilisent des lois de probabilités conditionnelles, les modèles neuronaux utilisent des algorithmes prédictifs sur les représentations vectorielles des mots. L'avantage des réseaux de neurones sur les modèles n-grammes réside dans le fait que l'historique d'un mot n'est plus vu comme une séquence exacte de mots précédents, mais plutôt comme une projection entière dans un espace de continu de grande dimension.
   Pour plus de détails sur le fonctionnement et la construction de ce type d'approches, le tutoriel réalisé par Neubig (2017) décrit avec précision les réseaux neuronaux appliqués aux modèles de langue.

Évaluation du modèle de langue

La qualité d'un modèle de langue est évaluée grâce à la mesure de perplexité. Cette mesure correspond à l'inverse de la probabilité d'un corpus d'évaluation, normalisé par le nombre de mots. La formule est la suivante :

$$PP(W) = P(W_i | w_1, w_2, \ldots, w_{i-1}))^{-\frac{1}{N}} \tag{1.7}$$

où N est le nombre total de mots, W une séquence de mots et w un mot.

Le test de perplexité donne une indication sur la performance d'un modèle de langue à trouver un mot nouveau, d'après un contexte. Théoriquement, plus la perplexité est basse, plus sa capacité de prédiction est élevée, donc meilleur est le modèle. Ainsi, le meilleur modèle

est celui qui prédit le mieux les mots d'un jeu de données inconnues (représenté en RAP par les corpus d'évaluation et de test). En général, on mesure également le taux de mots inconnus (dits hors vocabulaire), ce qui permet de donner une indication supplémentaire sur la qualité d'un modèle.

1.3.1.3 Évaluation de la performance d'un système de RAP

La performance d'un système de RAP s'évalue avec un score établi à partir du taux d'erreur de mots (communément abrégé "WER" pour *Word Error Rate*). Le WER sert à évaluer le taux de reconnaissance du système après décodage. Cette mesure compare la transcription du signal audio à reconnaître (appelée « référence ») avec la transcription effectivement produite par le système (appelée « hypothèse »). Le WER représente la somme du nombre d'opérations (substitution S, suppression D et insertion I) distinguant les deux transcriptions sur le nombre total N de mots de la référence, soit :

$$WER = \frac{S + D + I}{N} \times 100$$

À noter qu'il peut également être intéressant de calculer le taux d'erreur de caractères (abrégé "CER" pour *Character Error Rate*) qui se mesure de la même façon, mais sur les caractères.

1.3.2 Application aux langues peu dotées

1.3.2.1 Enjeu de la RAP pour les langues peu dotées

Depuis un peu plus d'une dizaine d'années, des travaux émergent pour adapter les techniques de traitement automatique du language naturel aux langues peu dotées. Cette appellation a été utilisée pour la première fois par Krauwer (2003). Berment (2004) a défini des critères pour qualifier une langue comme peu dotée : absence d'un système d'écriture ou orthographe instable, présence limitée sur le Web, manque d'études linguistiques, manque de ressources électroniques pour le traitement de la langue (comme les corpus monolingues, les dictionnaires électroniques bilingues, les systèmes de transcription automatiques, etc.). Comme expliqué précédemment, l'adaptation des systèmes de RAP aux langues peu dotées est, donc, un défi de taille, puisque les systèmes d'apprentissage automatique ont besoin d'un très grand nombre de ressources linguistiques (documents textuels numériques) et sonores pour être performants.

1.3.2.2 Travaux antérieurs

De plus en plus d'attention est portée sur le développement de ressources linguistiques et de technologies pour les langues peu dotées, comme en témoigne la création récente d'ateliers comme SLTU (*Spoken Language Technologies for Under-resourced Languages*) ou CCURL (*Collaboration and Computing for Under-Resourced Languages*), de groupes de travail comme SIGUL (*Special Interest Group on Under-resourced Languages*) ou encore de projets tels que ALFFA ou BULB.

Au cours des dernières décennies, plusieurs travaux ont été menés afin d'aider à l'identification, l'exploration et la légitimation de langues peu dotées. Besacier, Le *et al.* (2005) ont été les premiers à présenter des travaux de RAP pour les langues vietnamiennes et khmères à faible ressources. Plus tard, des travaux ont été publiés sur les langues africaines : somali (Nimaan, Nocera et Torres-Moreno, 2006), amharique (Pellegrini, 2008), langues sud-africaines (Barnard, Davel et Van Huyssteen, 2010).

Le vietnamien a continué à être étudié par d'autres équipes, si bien que Vu, Kraus et Schultz (2011) ont développé une méthode d'adaptation rapide d'un système de RAP pour cette langue. L'intérêt grandissant pour les langues peu dotées a ainsi mené à la publication d'un article d'introduction à la RAP pour ces langues (Besacier, Barnard *et al.* 2014). Cet article relate les travaux menés jusqu'alors dans le domaine du traitement automatique de la parole ainsi que les efforts réalisés pour la promotion des technologies vocales dans ces langues et pour ces langues. Cette étude a ainsi démontré que l'intérêt pour les langues peu dotées est réel, mais que de nombreux verrous sont encore à lever, notamment au niveau de l'accès et du partage des ressources. Enfin, des travaux récents ont montré que l'emploi de réseaux de neurones convolutifs profonds (plus de 5 couches convolutives) sur des données multilingues peut aussi s'avérer intéressant pour la RAP de langues à faible ressources (Sercu *et al.* 2016).

1.4 Conclusion

La linguistique de terrain est l'étude empirique du langage, c'est-à-dire qu'elle s'appuie effectivement sur les données récoltées auprès des locuteurs d'une langue. Le travail du linguiste de terrain est divisé en deux tâches principales : la documentation et la description des langues. D'un côté, la documentation apporte des informations sur les données brutes (collectées sur le terrain) et s'intéresse aux moyens à mettre en œuvre pour leur conservation et leur diffusion. D'un autre, la description est l'analyse relative à ces données brutes, à différentes échelles linguistiques, dans le but de dégager la structure interne de la langue.

Face à la diminution — de plus en plus rapide —, du nombre de langues dans le monde, la communauté scientifique se mobilise pour trouver des moyens de sauvegarder la diversité linguistique et réfléchir à des solutions pour aider le linguiste de terrain dans son travail quotidien. Les recherches et discussions interdisciplinaires ont fait émerger de nouveaux enjeux et de nouvelles opportunités. De nouvelles méthodologies et l'utilisation des dernières avancées en linguistique informatique offrent de nouvelles perspectives de travail au linguiste de terrain, tout en lui permettant de continuer à travailler avec les outils existants largement utilisés dans le domaine. Les progrès et la démocratisation des appareils mobiles, aussi, augmentent les possibilités de travail.

Enfin, les méthodes automatiques de traitement des données et l'apprentissage machine sont maintenant largement explorés pour améliorer la productivité du linguiste de terrain. Parmi les outils de traitement automatique de la langue, la reconnaissance automatique de la parole peut être un moyen utilisé pour aider le linguiste à traiter plus rapidement les données qu'il a collectées sur le terrain. Cette approche peut permettre d'automatiser certaines tâches

indispensables de la documentation et de la description, mais aussi permettre au linguiste de réaliser des études quantitatives à partir de ses données.

Par ailleurs, il nous semble important de mentionner qu'à plus large échelle, les recherches en traitement automatique des langues ont tiré parti des techniques d'analyse utilisées en linguistique historique (calculs de distance ou, inversement, de similarité par exemple) et, à leur tour, contribuent à l'exploration des théories sur la genèse des langues, comme la recherche des universaux du langage, la reconstruction phylogénétique des langues ou encore l'existence d'une potentielle proto-langue (une seule langue-mère qui serait à l'origine de toutes les langues du monde).

À titre d'exemple, le corpus parallèle libre d'accès, construit à partir de 100 traductions de la Bible par Christodouloupoulos et Steedman (2015), représente une ressource intéressante pour l'étude des structures linguistiques et la classification des langues. Dans le même état d'esprit, Rabinovich, Ordan et Wintner (2017) se sont servis du corpus Europarl (corpus parallèle contenant les traductions des rapports du Parlement Européen dans 21 langues européennes) pour montrer que les textes traduits sont imprégnés de certaines caractéristiques de la langue source, à tel point qu'ils ont été capable de reconstruire une classification de ces langues à travers un arbre.

Asgari et Schütze (2017) ont analysé automatiquement les marqueurs de temps à l'aide de leur méthode d'analyse cross-lingue appelée *SuperPivot*. Ils ont utilisé les traductions du nouveau testament dans 1 163 langues. Comme ces langues sont majoritairement peu dotées, l'intérêt de ce travail est l'annotation automatique des textes grâce à une méthode de projection.

En reconnaissance automatique de la parole, les mesures de similarité entre les langues sont investiguées pour améliorer l'apprentissage des systèmes à grand vocabulaire (Challa et Annabattula, 2017).

CHAPITRE 2

Langues Africaines Abordées

Dans la présente étude, nous avons principalement étudié le haoussa et le wolof, deux langues particulièrement utilisées en Afrique de l'Ouest. Néanmoins, au sein du projet ALFFA, d'autres langues africaines ont été considérées : le bambara, le peul et le fongbe pour l'Afrique de l'Ouest (couvrant ainsi plus de la moitié des 300 millions d'habitants d'Afrique de l'Ouest, en incluant le haoussa et le wolof) ; le swahili et l'amharique pour l'Afrique de l'Est (rassemblant près de 190 millions de locuteurs). Le choix des langues a principalement été gouverné par la couverture de la population ainsi que par les perspectives industrielles offertes par les langues et les pays où celles-ci sont parlées.

2.1 Langues abordées dans le projet ALFFA

Le haoussa. Le haoussa est la langue la plus parlée de la famille des langues tchadiques, qui font partie du phylum Afro-Asiatique (Vycichl, 1990). Le haoussa est parlé par environ 50 millions de locuteurs, en tant que première ou seconde langue, ce qui en fait la langue la plus utilisée en Afrique sub-saharienne. Jaggar (2001) explique qu'il existe, en haoussa, des variantes régionales qui se distinguent au niveau phonologique (notamment sur la tonologie), lexical, morphologique. Enfin, il existe un contraste de durées sur les phonèmes en haoussa.

Le wolof. Le wolof appartient à la famille des langues atlantiques, qui font partie du phylum Niger-Congo. Le wolof est parlé principalement au Sénégal et en Gambie, et aussi en Mauritanie. Le wolof est la langue la plus parlée au Sénégal, et elle est reconnue à la fois comme lingua franca et comme langue nationale alors que le français, unique langue officielle, est seulement parlé par une minorité de la population sénégalaise. Selon des estimations récentes, plus de 5 millions de personnes parlent le wolof en tant que langue maternelle.

Le bambara. Le bambara est la variété la plus parlée de la langue mandingue de la famille Mandé, aujourd'hui séparée du phylum Niger-Congo (Dimmendaal, 2008). Cette langue est la première langue d'environ 4 millions de locuteurs au Mali [1]. Le bambara est également utilisé en tant que seconde langue par près de 10 millions de locuteurs (au Mali) et est utilisée en tant que langue véhiculaire au Sénégal. On distingue le bambara « standard » parlé principalement à Bamako (bien que la forme se généralise de plus en plus à travers le Mali) et d'autres variantes

1. http://www.inalco.fr/langue/bambara-mandingue

du bambara parlées dans d'autres régions [2]. D'un point de vue morphologique, le bambara est une langue isolante. Son système d'écriture est basé sur l'alphabet latin. Enfin, comme la grande majorité des langues africaines, le bambara est une langue à tons.

Le peul. Le peul (aussi appelé pulaar ou fulfulde) regroupe un ensemble de dialectes parlés dans tous les pays de l'Afrique de l'Ouest par environ 70 millions de personnes. C'est aussi la langue maternelle des Toucouleurs, peuple habitant à la frontière avec la Mauritanie, dans la vallée du fleuve Sénégal. Dans le projet ALFFA, nous avons décidé de nous concentrer sur le pulaar qui est le dialecte occidental des langues peuls. Comme le wolof, le peul appartient aux langues atlantiques du phylum Niger-Congo. De plus, ce n'est pas une langue tonale. Au Sénégal, le pulaar a le statut de langue nationale et est parlé par près de 3,5 millions de locuteurs (Leclerc, 2013). Le dialecte pulaar utilise l'alphabet latin comme système d'écriture officiel.

Le fongbe. Le fongbe (ou fon) fait partie des langues kwa appartenant à la branche Niger-Congo (Greenberg, 1966). La langue est parlée au Bénin par un peu plus de la moitié de la population, au Togo et au Nigéria. Le fongbe est une langue à tons (Lefebvre et Brousseau, 2001). Son système d'écriture est basé sur l'alphabet latin ; parfois les diacritiques servant à représenter les tons sont notés mais ce n'est pas une règle systématique. Le fongbe possède 22 consonnes et 12 voyelles (Laleye *et al.* 2016).

Le swahili. Nous avons également couvert une partie de l'Afrique de l'Est dans le projet, en concevant un système de RAP pour le swahili. Le kiswahili appartient au phylum Niger-Congo et est la langue la plus parlée parmi les langues bantoues. C'est également la langue la plus répandue dans l'Est du continent : elle est parlée par plus de 100 millions de personnes [3]. Elle est utilisée comme langue maternelle mais aussi comme langue véhiculaire par une grande partie de la population de l'Afrique de l'Est. La langue a un statut officiel en Tanzanie et est une langue nationale dans de nombreuses autres régions d'Afrique centrale et orientale. Le système d'écriture arabe a été utilisé pendant des siècles pour écrire le swahili mais désormais le système d'écriture officiel est basé sur l'alphabet latin.

L'amharique. L'amharique a également été considéré dans le projet. C'est une langue sémitique qui fait partie du phylum Afroasiatique. C'est la deuxième langue la plus parlée parmi les langues sémitiques et une des langues les plus parlées sur le continent africain. Selon le dernier recensement du SIL datant de 2010, l'amharique est principalement parlé en Éthiopie par 22 millions de locuteurs [4] où la langue a un statut officiel. L'amharique utilise un système d'écriture alphasyllabaire (la séquence consonne-voyelle représente une unité discrète) nommé fidel (Comrie, 2009).

2. http://www.inalco.fr/langue/bambara-mandingue
3. http://swahililanguage.stanford.edu
4. https://www.ethnologue.com/language/amh

2.2 Le haoussa

Le haoussa fait partie des langues les plus parlées en Afrique. C'est la langue la plus utilisée en tant que langue véhiculaire en Afrique de l'ouest (Jaggar, 2001), rassemblant environ 50 millions de locuteurs (première et deuxième langue confondues) [5]. Environ un quart du vocabulaire haoussa provient de la langue arabe mais la langue a également été influencée par le français, dont on retrouve de nombreux emprunts. Il existe différentes variétés de haoussa, selon qu'il est parlé dans les régions de l'est (Kano, Zaria, Bauchi, Daura), de l'ouest (Sokoto, Gobir, Tahoua), du nord (Katsina, Maradi, Zinder) ou du sud (Zaria, Bauci). Le haoussa dit « standard » est celui parlé à Kano. C'est cette variété que nous étudierons dans les présents travaux.

2.2.1 Statut de la langue

Afin de juger du niveau d'utilisation d'une langue, Lewis et Simons (2010) ont mis au point l'échelle EGIDS (abréviation pour *Expanded Graded Intergenerational Disruption Scale*, en français *Échelle de perturbation intergénérationnelle graduée étendue*). Cette échelle comporte 13 niveaux qui permettent d'évaluer le degré de vitalité d'une langue. Le plus haut niveau est le degré 0 : la langue est étiquetée « Internationale », ce qui signifie qu'elle est largement utilisée dans le monde entier, comme langue de communication entre toutes les nations (échanges commerciaux, transmission des connaissances, relations internationales). À l'inverse, le plus bas niveau est le degré 10 : la langue est étiquetée « Éteinte », ce qui signifie qu'elle n'est plus du tout utilisée et que plus personne ne revendique d'attachement ethnique à cette langue.

Sur cette échelle, le haoussa est au niveau 2, ce qui signifie qu'elle est largement utilisée par la population, les institutions gouvernementales et administratives ainsi que par les médias. Le haoussa n'est donc pas en danger.

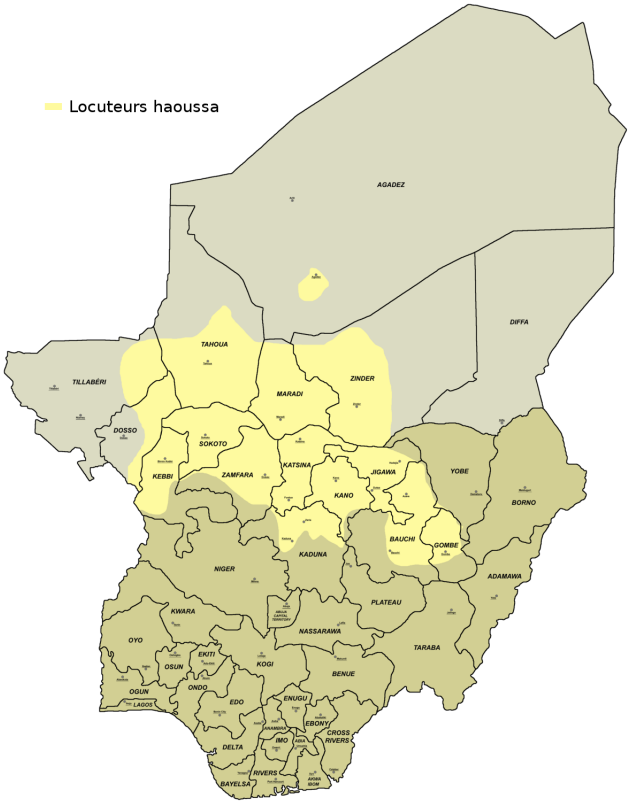
2.2.1.1 Situation géographique

Le haoussa est principalement parlé au Nigeria par environ 21% de la population (environ 30 millions de locuteurs) (Caron, 2000) et au sud du Niger par environ 47% de la population (environ 8 millions de locuteurs) en tant que langue maternelle [6], comme le montre le graphique 2.1. Des locuteurs natifs du haoussa vivent aussi au Cameroun, au Ghana et au Soudan (Koslow, 1995). Le haoussa est également parlé en tant que langue seconde dans tout le nord du Nigeria, au Niger, au Bénin, au Tchad et dans de nombreuses villes commerçantes d'Afrique de l'Ouest et d'Afrique centrale (telles que Dakar, Abidjan, Lomé, Ouagadougou ou Bamako)[5].

2.2.1.2 Situation sociopolitique

Le haoussa est largement utilisé dans la partie nord du Nigeria et au Niger, mais n'a pas le statut de langue officielle. Il est reconnu comme étant une des trois langues majeures du

5. https://www.ethnologue.com/language/hau
6. http://www.axl.cefan.ulaval.ca/afrique/niger.htm



Graphique 2.1 – Répartition des locuteurs de haoussa au Niger et Nigeria – Source : https://en.wikipedia.org/wiki/File:Hausa_language_map.png#filehistory.

pays par le gouvernement nigérian et fait partie des enseignements obligatoires du secondaire (Caron, 2000).

2.2.1.3 Classification linguistique

Le haoussa est une langue appartenant à la branche « ouest A » des langues tchadiques, famille de langues la plus vaste du phylum Afro-Asiatique. Parmi les langues tchadiques, le haoussa est de loin la langue la plus parlée (Vycichl, 1990).

2.2.2 Phonologie

Tel que décrit par Newman (2000), le haoussa est une langue tonale. Il existe 3 tons, portés par les voyelles, qui peuvent être bas, haut ou descendant. À l'écrit (notamment dans les dictionnaires), les tons sont transcrits par des diacritiques. Ainsi, le ton haut est représenté par un accent aigu (certains auteurs ne le marquent cependant pas toujours), le ton bas par un accent grave et le ton descendant par un accent circonflexe. Ce dernier ne peut se produire que dans une syllabe fermée. Le ton a une valeur lexicale et grammaticale.

2.2.2.1 Unités phonologiques

Comme la majorité des langues afro-asiatiques, le haoussa possède un grand nombre de consonnes, dont des consonnes non pulmoniques (injectives et éjectives) (Caron, 2011). Selon les variantes dialectales, le nombre de phonèmes consonantiques peut fluctuer et les points

d'articulation peuvent être différents. Les paragraphes ci-après inventorient les consonnes et les voyelles en haoussa, et les présentent selon les descriptions de la littérature.

Inventaire consonantique

Tableau 2.1 – Inventaire consonantique du haoussa

		Bilabiale	*Alvéolaire*	*Post-alvéolaire*	*Dorsale*			*Glottale*	
					pure	labial.	palatal.	pure	palatal.
Nasale		m	n						
Occlusive/Affriquée	sourde		t	tʃ	k	kʷ	kʲ	ʔ	ʔʲ
	voisée	b	d	dʒ	g	gʷ	gʲ		
	injective	ɓ	ɗ						
	éjective		tsʼ	(tʃʼ)	ƙ	ƙʷ	ƙʲ		
Fricative	sourde	ɸ	s	ʃ				h	
	voisée		z						
Roulée et battue			r	ɽ					
Approximante			l			w	j		

Le tableau 2.1 décrit le système phonologique des consonnes en haoussa, telles que réalisées à Kano (variante étudiée dans ces travaux). Il reprend les données de plusieurs sources, notamment l'inventaire phonologique de Newman (2000). Nous avons également pris en considération, dans ce tableau, l'inventaire de Schlippe *et al.* (2012) fournit dans la documentation de leurs ressources (après achat), puisque nous avons utilisé leur corpus audio pour nos expérimentations (décrites au chapitre 4).

32 consonnes sont dénombrées dans cet inventaire consonantique. Toutes les consonnes peuvent être géminées et sont alors prononcées plus longues. Ces géminées peuvent être opposées à leurs homologues brèves ; le changement de longueur entraîne une modification du sens du mot. Par exemple, « ku**l**è » *(chat)* s'oppose à « ku**ll**è » *(fermer à clé)*. Nous pouvons remarquer dans le tableau 2.1 que les consonnes du haoussa s'opposent par différents traits articulatoires. Ainsi, les occlusives et affriquées présentent 2 modes d'articulation pulmonique (sourde/voisée) et non pulmonique (injective/éjective). Les occlusives dorsales sourdes, voisées et éjectives peuvent être prononcées pures (prononciation type), labialisées (arrondissement des lèvres) ou palatalisées (la lame de la langue (l'apex) vient toucher le palais dur). La glottale /ʔ/ n'est jamais notée à l'initiale du mot mais est prononcée si celui-ci commence par une voyelle. Caron et Amfani (1997) expliquent que, devant les voyelles antérieures /o/ et /u/, que celles-ci soient brèves ou longues, les consonnes /b/, /k/, /g/, /ɓ/ et /ƙ/ sont labialisées tandis que, devant /i/ et /e/ (brefs ou longs), les consonnes /k/ et /g/ sont palatalisées. /n/ est prononcé [ŋ] devant /k/, /g/ et /ƙ/ ainsi qu'en position finale de mot.

Inventaire vocalique

Le haoussa comprend 5 timbres vocaliques de 3 degrés d'aperture, présentés dans le tableau 2.1, ainsi que 2 diphtongues /ai/ et /au/. Ces 5 voyelles peuvent varier en durée et peuvent ainsi également se réaliser longues.

Tableau 2.2 – Inventaire vocalique du haoussa – Source : Newman (2000)

	Antérieure	Centrale	Postérieure
Fermée	i		u
Mi-fermée	e		o
Mi-ouverte	ɛ		ɔ
Ouverte		a	

L'allongement de la voyelle dépend de plusieurs facteurs : sa position au sein du mot (initiale, médiane, finale), sa position dans la syllabe ou encore si la voyelle se trouve avant une pause.

Le timbre des voyelles peut être modifié par leur longueur et leur contexte environnant. La représentation phonétique que Caron et Amfani (1997) dressent est la suivante (nous ne décrivons que les brèves car les longues présentent les mêmes caractéristiques que décrites dans l'alphabet phonétique international (API)) :

- /a/ ⟶ [ʌ] (ouverte, centrale, lèvres neutres ; comme l'anglais *but, cut*)
- /e/ ⟶ [ɛ] (d'avant, mi-ouverte, lèvres neutres ; comme l'anglais *bet, get*)
- /i/ ⟶ [ɪ] (pas tout à fait fermée, pas tout à fait d'avant ; comme l'anglais *bit, lid*)
- /o/ ⟶ [ɔ] (demi-ouverte, d'arrière, arrondie ; comme l'anglais *god*)
- /u/ ⟶ [ʊ] (à mi-chemin entre fermé et demi-fermé, pas tout à fait arrière ; comme l'anglais *book*)

Ainsi, lorsque la voyelle ne précède pas une pause, les brèves sont raccourcies, relâchées et centralisées par rapport aux longues qui sont allongées et tendues. En revanche, lorsque la voyelle précède une pause, les voyelles brèves retrouvent leurs réalisations canoniques, comme décrites dans l'API (Jaggar, 2001). Aussi, la différence de longueur entre brèves et courtes est bien moins manifeste (Newman et Heuven, 1981).

En position finale de mot, la voyelle peut être courte ou longue et la différence entre les deux dépendra notamment de la présence d'une pause ou non. En position médiane, seules les voyelles longues /e/ et /o/ peuvent apparaître. Cependant, si ces voyelles se retrouvent dans une syllabe fermée, elles sont forcément raccourcies.

La variation de la longueur de la voyelle affecte directement la signification du mot. La longueur de la voyelle est difficile à identifier car elle n'est pas notée dans l'orthographe latine. Ainsi, le mot « daga » peut être prononcé [daga:] et signifie « bracelet à breloque » ou [da:ga:] et signifie dans ce cas « ligne de combat » (Newman, 2000).

Finalement, peu d'études existent concernant la mesure de la quantité vocalique et celle de Newman et Heuven (1981) reste une référence sur le sujet avec celle de Lindau-Webb (1985).

2.2.2.2 Syllabification

La syllabe en haoussa peut être fermée (structure CVC : la rime est composée d'un noyau et d'une coda) ou ouverte (structure CV ou CVV : la rime n'est constituée que d'un noyau qui peut être une voyelle brève, une voyelle longue ou une diphtongue). La consonne initiale peut être l'occlusion glottale (notée ʾ), mais elle n'est pas marquée dans l'orthographe. Le contraste de longueur n'apparaît que dans les syllabes ouvertes. Dans les syllabes fermées, les voyelles sont toujours brèves.

2.2.3 Morphologie

Le mot en haoussa est créé à partir d'un radical auquel on ajoute un affixe. L'affixe – qui peut être préfixe, infixe ou suffixe – est constitué d'au moins une voyelle finale, à laquelle est assignée un schéma tonal. Caron (2011) indique qu'il n'existe que deux préfixes, *ma-* et *ba-*, le premier pouvant servir dans le processus de dérivation et le deuxième dans la formation d'ethnonymes. Les infixes permettent de créer des noms pluriels. Les suffixes servent à catégoriser les noms mais aussi à la flexion et à la dérivation.

2.2.4 Système d'écriture

Environ un quart des mots du haoussa provient de l'arabe mais la langue a aussi été influencée par le français. Le haoussa s'écrit avec le système d'écriture arabe depuis le début du 17ème siècle : ce système d'écriture est appelé ʿajami. Cependant, le système d'écriture officiel est basé sur l'alphabet latin et est appelé boko. Ce système d'écriture a été imposé par les anglais lors de la colonisation, dans les années 30. Le boko est composé de 22 lettres de l'alphabet latin (A/a, B/b, C/c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, R/r, S/s, T/t, U/u, W/w, Y/y, Z/z), auxquelles ont été ajoutées 4 consonnes (Ɓ/ɓ, Ɗ/ɗ, Ƙ/ƙ et Ɣ/ɣ), ainsi que le signe diacritique ʾ représentant l'arrêt glottal et utilisé en position interne de mot (Caron, 2015). À l'écrit, les tons et longueurs des voyelles sont marqués en ʿajami, mais pas en boko. La construction de ressources informatisées est par conséquent plus délicate en boko car ces informations ne sont pas mentionnées dans l'écriture du mot.

2.3 Le wolof

Le wolof est la langue la plus parlée et la plus étudiée (et depuis longtemps) au Sénégal. Le premier dictionnaire, d'après la base de données RefLex [7] (Segerer et Flavier, 2013) remonte à 1825 (Dard et Savaresi, 1825). Les études ont principalement porté sur la syntaxe et la morphologie du wolof, tandis que celles portant sur la phonologie et phonétique sont encore anciennes ou peu approfondies.

7. http://reflex.cnrs.fr/Lexiques/webball/index.html

2.3.1 Statut de la langue

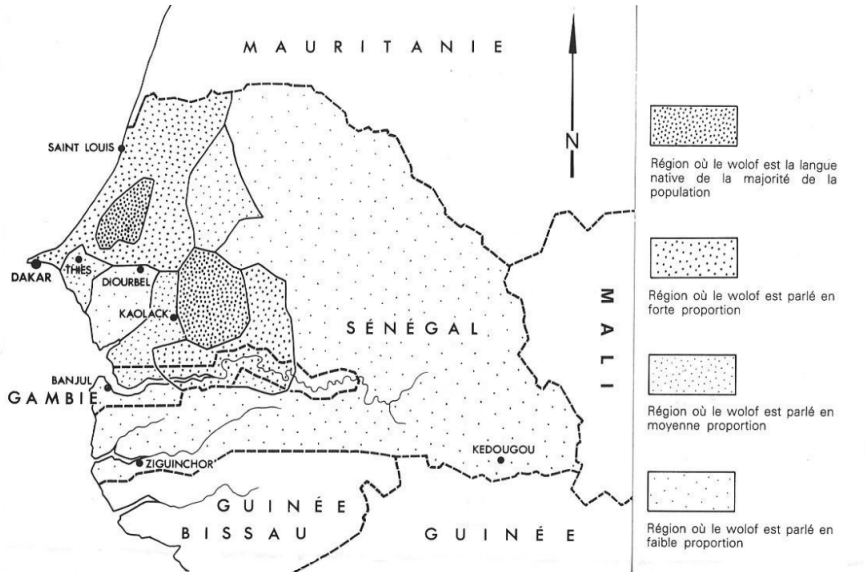
Le wolof n'est pas une langue en danger mais est une langue peu dotée. Elle est classée au niveau 4 sur l'échelle EGIDS. Cela signifie qu'elle est largement utilisée par la population, qu'elle est bien décrite et qu'elle est soutenue par le gouvernement.

2.3.1.1 Classification linguistique

Le wolof appartient à la branche nord de la famille des langues atlantiques, qui font partie du phylum Niger-Congo. 1 526 langues sont regroupées au sein de ce phylum, représentant ainsi la famille linguistique la plus étendue du monde (soit environ 21,5% des langues parlées à travers le monde) [8].

2.3.1.2 Situation géographique

Le wolof est principalement parlé par l'ethnie Wolof en Sénégambie, aire géographique regroupant le Sénégal, la Gambie et le sud de la Mauritanie. On dénombre environ 40% de locuteurs natifs wolof au sein de la population sénégalaise, 12% au sein de la population gambienne et 8% au sein de la population mauritanienne. C'est aussi une langue véhiculaire d'Afrique de l'Ouest, utilisée en tant que langue seconde par un grand nombre de locuteurs : rien qu'au Sénégal, 50% de la population parle le wolof en tant que deuxième langue. Le graphique 2.2 montre la répartition des wolofones en Sénégambie.



Graphique 2.2 – Répartition des wolophones en Sénégambie (Fal, Santos et Doneux, 1990).

Bien que le wolof soit la langue véhiculaire de Sénégambie, le wolof du Sénégal et le wolof de Gambie sont deux variétés distinctes. Chacune est représentée par un code ISO 639-3 unique, « WOL » pour le wolof sénégalais, « WOF » pour le wolof gambien. Bien que des différences se situent au niveau phonologique (dans l'accent) et au niveau lexical — notamment au niveau des emprunts — en raison du passé colonial distinct des pays (la Gambie a été

8. https://www.ethnologue.com/statistics/family

une colonie anglaise tandis que le Sénégal a été une colonie française), les linguistes décrivent également des différences au niveau morphologique et syntaxique (isoglosses sur les pronoms, pluriels de la classe nominale, marqueurs possessifs et marques flexionnelles sur le verbe, de temps, d'aspect et de mode (très souvent abrégé *TAM*), etc.).

Ce travail se focalise sur le wolof parlé au Sénégal seulement.

2.3.1.3 Situation sociopolitique

Le Sénégal reconnaît 21 langues nationales, parmi lesquelles le wolof. Bien que n'étant pas reconnue comme langue officielle, le wolof est de loin la langue la plus parlée au Sénégal, avec près de 90% de locuteurs l'utilisant au quotidien comme langue première ou seconde (Leclerc, 2013).

2.3.2 Variantes dialectales

Puisque la majorité de la population du Sénégal utilise le wolof comme langue d'interaction, des différences existent à Dakar mais également selon les régions (Guérin, 2016). Dans cette étude, nous utiliserons le terme « standard » pour le wolof parlé à Dakar par des locuteurs natifs de la langue et le terme « urbain » pour le wolof parlé par des locuteurs non natifs. Malgré l'existence de variétés dialectales, l'intercompréhension reste totale entre les personnes vivant dans les différentes régions (Dramé, 2012). Nous nous sommes intéressés, dans la présente étude, aux variantes que sont le faana-faana et le lébou. Le faana-faana, également appelé *wolof du Saloum* est parlé dans la région de Kaolack, à l'Est du Sénégal. Le lébou, quant à lui, est principalement parlé sur les côtes du Sénégal. Différentes variétés de lébou existent, selon les régions dans lesquelles il est parlé. Ainsi, nous distinguons par exemple le lébou parlé à Ouakam du lébou parlé à Yénn. Les différences observées dans ces langues par les linguistes, en comparaison avec le wolof « standard », sont des variations phonétiques ou morphophonologiques, mettant l'accent sur le vocalisme et sur certaines formes d'inflexion verbale (Robert, 2011). Très récemment, des variations morphologiques et syntaxiques ont été démontrées par Voisin et Dramé (en prép.) et Voisin (en prép.).

2.3.3 Phonologie

Les études descriptives de la phonologie et de la phonétique du wolof sont encore peu nombreuses et sont, à l'heure actuelle, toujours source de débats.

2.3.3.1 Unités phonologiques

Le wolof ne possède ni de tons ni de diphtongues. Au total, 60 phonèmes sont inventoriés en wolof (Guérin, 2011). Les phonèmes peuvent varier en longueur (Cissé, 2006). La prononciation brève ou allongée du phonème peut conduire à une signification différente du mot. Les paragraphes ci-après décrivent les inventaires des systèmes consonantiques et vocaliques du wolof, selon les descriptions les plus récentes de la littérature.

Inventaire consonantique

Le wolof comprend 45 consonnes, divisées en deux catégories : les consonnes dites « simples » (ou « faibles ») et celles dites « complexes » (ou « fortes »). Les consonnes simples (25 au total) regroupent les phonèmes constitués d'une seule unité, tandis que les consonnes complexes (au nombre de 20) regroupent les consonnes géminées et les consonnes prénasales. Les consonnes complexes peuvent être considérées de deux sortes, selon l'analyse de Maddieson dans la base de données LAPSyD (Maddieson *et al.* 2013) : les occlusives prénasales voisées constituent un élément unique tandis que les géminées et la séquence *nasale+occlusive sourde* constituent un groupe (addition de deux éléments). Ce dernier type de consonne complexe (considéré comme un groupe) n'apparaît d'ailleurs jamais en position initiale (mot ou syllabe). Toutes les consonnes simples peuvent devenir des géminées, excepté les fricatives. La gémination entraîne un allongement de la prononciation de la consonne.

Tableau 2.3 – Inventaire consonantique du wolof

	Labiale	**Dentale**	**Palatale**	**Vélaire**	**Uvulaire**	**Glottale**
Nasale	m	n	ɲ	ŋ		
Occl. prénasale	mb	nd	ɲɟ	ŋg		
Occlusive	p b	t d	c ɟ	k g		ʔ
Fricative	f	ɾ	s		χ	
Approximante		l	j	w		

Le tableau 2.3 présente l'inventaire phonologique des consonnes du wolof. Il provient de plusieurs sources, notamment Maddieson *et al.* (2013) et Guérin (2016). Les phonèmes à gauche de la cellule correspondent aux consonnes sourdes tandis que ceux de droite correspondent aux consonnes voisées. Nous pouvons remarquer que seules les occlusives connaissent une opposition de voisement.

Robert (2011) explique que les consonnes voisées, en position finale, se dévoisent et sont prononcées implosées. Les occlusives (simples) deviennent géminées en position médiane et finale de radical. Enfin, la présence de consonnes complexes en début de mot ou devant une voyelle longue est impossible.

Nous avons pris le parti de reprendre la description de Ka (1994) et de ne représenter, dans le tableau 2.3, que les occlusives prénasales voisées comme consonnes complexes, estimant que les prénasales sourdes relèvent plus de la paire de phonèmes. Nous avons choisi de présenter un tableau reflétant la prononciation du wolof. Notre inventaire est ainsi constitué de la consonne fricative uvulaire non voisée /χ/ — et non de la fricative vélaire non voisée /x/ —; la consonne /s/ est classée en tant que palatale et non comme alvéolaire; la consonne /ɾ/ est définie comme une fricative et non comme une battue.

Tableau 2.4 – Inventaire vocalique du wolof

(a) Voyelles brèves

	Antérieure	Centrale	Postérieure
Fermée	i		u
Mi-fermée	e		o
		ə	
Mi-ouverte	ɛ		ɔ
Ouverte		a	

(b) Voyelles longues

Antérieure	Centrale	Postérieure
iː		uː
eː		oː
ɛː		ɔː
	aː	

Inventaire vocalique

En wolof, le système vocalique est composé de 15 voyelles (8 phonèmes brefs et 7 phonèmes longs). Le tableau 2.4 représente l'inventaire vocalique du wolof que nous avons utilisé dans nos travaux. Cet inventaire est largement tiré de celui de Diagne (1971), mais nous avons pris le parti de ne pas représenter la voyelle /à/ (afin d'être cohérent avec notre dictionnaire de prononciation construit pour les expérimentations décrites au chapitre 4). Nous pouvons remarquer que les voyelles brèves peuvent toutes être allongées, excepté /ə/.
Le wolof est une langue à harmonie vocalique, dans laquelle les voyelles peuvent se distinguer par l'avancée de la racine de la langue (trait souvent désigné par le terme anglais *ATR (Advanced Tongue Root)*).
Il existe, selon les régions, des variations de prononciation en wolof. Les voyelles sont les plus touchées par ces variations. Par exemple, /ə/ peut ne pas être réalisé dans certains parlers wolof. Ces variations de prononciation entraînent des désaccords dans les études descriptives. La voyelle /a/ est la plus controversée (ceci fait l'objet d'un paragraphe à la page 40).

Enfin, comme pour les consonnes, l'allongement est représenté orthographiquement par une duplication du graphème. Ainsi, la prononciation du mot « fit » (signifiant « bravoure », « courage ») et celle du mot « fiit » (qui veut dire « piège ») ne diffère qu'au niveau de la longueur de la voyelle. De la même façon, les verbes « boroos » (« brosser ») et « boros » (« brocher »), ou encore les verbes « set » (être propre) et « seet » (chercher) sont des paires minimales qui ne se distinguent, respectivement, que par la durée de la voyelle /o/ et /e/.

2.3.3.2 Syllabification

La structure canonique d'une syllabe en wolof est la suivante : CV(:)C(C). L'attaque de la première syllabe du mot ne peut être composée que d'une consonne simple ou d'une occlusive prénasale voisée, tel que présenté dans le tableau 2.3 (et stipulé à la page 38). Pour un mot polysyllabique, les attaques suivantes peuvent être constituées de n'importe quel type de consonne (simple ou complexe). En ce qui concerne le noyau de la syllabe, il peut être construit avec une voyelle brève ou longue. En théorie, sauf pour /aː/, une voyelle longue ne peut qu'apparaître devant une consonne simple ou une occlusive prénasale voisée (Ka, 1994), bien que Diouf (2003) démontre qu'il existe de nombreux mots où la règle n'est pas respectée, tel que le verbe

intransitif « jebbi » [jɛːbbi] (se dit pour une plante, signifie *repousser, reprendre vie*) ou le nom commun « kócc » qui se prononce [k**o**ːcc] (signifie, en français, *mérou commun*).

2.3.3.3 Morphophonologie

Le processus de suffixation peut entraîner un phénomène d'alternance consonantique. Deux types d'alternances existent : (1) Une consonne simple devient complexe en position initiale (« **b**aax » *(être bon)* ⟶ « **mb**aax » *(bonté)*) ou lors de l'ajout d'un inversif ou d'un correctif en final de radical (« le**m** » *(plier)* ⟶ « le**mm**i » *(déplier)* ; « sa**f** » *(être agréable)* ⟶ « sa**pp**i » *(être désagréable)*). (2) Une consonne complexe devient simple lors du processus de dérivation verbale ou nominale (« là**gg**i » *(infirme)* ⟶ « laa**g**o » *(infirmité)*).

En wolof, une suite de voyelles différentes est impossible. Un phénomène d'assimilation vocalique se produit alors. En conséquence, dans le cadre de l'affixation, les voyelles fusionnent pour en créer une nouvelle. Ceci est courant dans les mots polysyllabiques se terminant par un phonème vocalique. La règle de coalescence [9] est la suivante :

- /i/ + /a/ ⟶ /eː/
- /u/ + /a/ ⟶ /oː/
- /e/ + /a/ ⟶ /eː/
- /u/ + /e/ ⟶ /oː/
- /oː/ + /a/ ⟶ /oː/
- /o/ + /a/ ⟶ /oː/
- /a/ + /a/ ⟶ /aː/

Cas particulier du phonème /a/

Le phonème bref /a/ en wolof possède un statut particulier, sujet à controverse. Sauvageot (1965), par exemple, décrit un système vocalique où la paire /a/~/aː/ s'oppose uniquement sur le trait de longueur. De son côté, Diagne (1971) propose l'ajout d'un phonème bref, noté /à/, plus ouvert que son homologue bref /a/. Dans sa représentation, les deux phonèmes ouverts et brefs /a/ et /à/ s'opposent à un seul phonème ouvert et long /aː/.

Le statut phonologique du phonème /a/ suscitant toujours des discussions, nous avons opté, dans ces travaux, pour opposer ce phonème uniquement sur le trait de longueur. Ainsi, comme le montre le tableau 2.4, nous opposons /a/ (bref) et /aː/ (long).

Étant donné la règle, /a/ devant consonne géminée devrait être prononcé bref, et d'un point de vue orthographique prendre un accent grave. Pourtant, des exceptions existent, comme nous pouvons l'observer à travers les exemples suivants, tirés du dictionnaire de Diouf (2003) : « bàkku » [b**a**ːkku] *(faire son propre éloge)*, « wàll » [w**a**ːllə] *(contaminer)*, « sàdd » [s**a**ːddə] *(orner, embellir)*, « ràgg » [r**a**ːggə] *(être chétif)*.

9. En linguistique, la coalescence définit deux unités qui se combinent pour en former une nouvelle.

2.3.4 Morphologie

Le wolof est une langue à classes nominales. La dérivation verbale et nominale est très fréquente et ne fonctionne que par suffixation (il n'y a pas de préfixation en wolof). 8 consonnes servent à classifier les noms au singulier (b-, k-, l-, w-, m-, g-, s-, j-) et 2 consonnes au pluriel (y- et ñ-). La dérivation peut être construite de 5 manières (Guérin, 2016) :

- par suffixation ;
- par alternance consonantique à l'initiale ;
- par réduplication ;
- par composition ;
- par conversion (les classificateurs nominaux peuvent s'alterner pour créer un nouveau nom).

2.3.5 Orthographe

Le premier système d'écriture en usage était le wolofal (Cissé, 2006). C'est un adjami (couramment utilisé en Afrique de l'Ouest, son alphabet utilise l'alphabet arabe) composé de 18 caractères empruntés à l'alphabet arabe (Ngom, 2010). Le wolofal a été utilisé jusqu'à la période pré-coloniale, notamment à travers la diffusion de l'Islam (Lüpke et Bao-Diop, 2014). Ensuite, l'alphabet latin est entré en vigueur. Le wolofal est toujours en usage dans la littérature religieuse mais l'orthographe officielle du wolof est, depuis 1971, fondée sur l'alphabet latin (Robert, 2011). 29 lettres sont utilisées en wolof : a, à, b, c, d, e, é, ë, f, g, i, j, k, l, m, n, ñ, ŋ [10], o, ó, p, q, r, s, t, u, w, x, y. La plupart d'entre elles peuvent devenir des digraphes pour former des géminées (aa, bb, cc, dd, ee, ée, gg, ii, jj, kk, ll, mm, nn, ññ, ŋŋ, oo, óo, pp, tt, uu, ww, yy) ou des prénasales (mb, nd, nj, ng).

La langue officielle du Sénégal est le français. Par définition, le français est donc la langue en usage dans les institutions gouvernementales comme les administrations, les tribunaux, les écoles et universités. C'est la langue de diffusion publique, qui doit être utilisée par les politiciens et les médias. Par conséquent, le wolof n'est pas la langue apprise dans l'enceinte scolaire ce qui en rend son orthographe instable. Bien que le Centre de Linguistique Appliquée de Dakar (CLAD) [11] coordonne la standardisation orthographique du wolof, les locuteurs s'approprient la langue à leur façon et, de ce fait, l'écrivent sans suivre de règles particulières. Cette variabilité dans l'écriture du wolof en fait un enjeu pour le traitement automatique de la langue naturelle, notamment en reconnaissance automatique de la parole.

2.4 Synthèse

Le haoussa et le wolof sont les deux langues principalement abordées dans ces travaux de thèse. Ces deux langues nous ont intéressé pour leur opposition de longueur sur les consonnes et sur les voyelles, affectant la signification du mot. Cette caractéristique est un enjeu majeur

10. "ŋ" est parfois orthographié "N"
11. http://clad.ucad.sn

pour créer des systèmes de RAP fiables, et rendre leur utilisation envisageable par les linguistes (de terrain). Les ambiguïtés liées à l'opposition de longueur pourrait être soulevées grâce au contexte (syntaxique par exemple), mais les langues peu dotées ne possèdent pas une quantité de données suffisante pour cela. Par ailleurs, d'un point de vue technologique, la bonne reconnaissance du mot (et donc des contrastes) par les systèmes est importante pour des outils de recherche par mots-clés par exemple, qui utilisent des mots isolés.

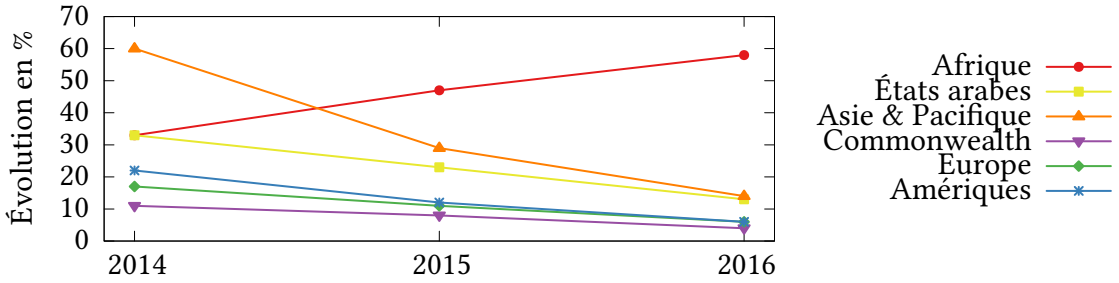
Enfin, d'un point de vue opérationnel sur le projet ALFFA, nous avons également étudié d'autres langues africaines. Des ressources ont été recueillies pour 3 langues supplémentaires : l'amharique, le fongbe et le swahili. Ces collectes ont permis de doter ces langues de technologies vocales (ce travail est détaillé plus loin, aux chapitre 3 et chapitre 4).

CHAPITRE 3

Collecter : LIG-AIKUMA, une application mobile de collecte de parole sur le terrain

Dans leur dernière étude publiée en 2010, l'UNESCO estime à environ 3 000 le nombre de langues à travers le monde qui disparaîtront d'ici la fin du siècle (MOSELEY, 2010). Il est de plus en plus urgent de préserver le patrimoine culturel immatériel que sont la langue, les arts, les pratiques sociales, les rituels, les connaissances scientifiques et les savoir-faire traditionnels.

Parallèlement, les nouvelles technologies émergent. Durant les dernières décennies, les appareils mobiles n'ont cessé de gagner en part de marché. D'après l'union internationale des télécommunications (UIT) [1], 3,6 milliards d'abonnements aux réseaux mobiles avec connexion internet seraient souscrits à la fin 2016 dans le monde, correspondant à un taux de pénétration global de presque 50%. CHÉNEAU-LOQUAY (2012) observe que c'est en Afrique (hors États arabes) que la progression est la plus forte. Cette tendance se poursuit comme indiqué sur le diagramme du graphique 3.1, généré à partir des chiffres de l'UIT [1]. Cette croissance s'explique grâce à une forte démocratisation du téléphone mobile depuis une dizaine d'années et par la valeur sociétale que revêt son usage pour les populations africaines (CHÉNEAU-LOQUAY, 2010).



GRAPHIQUE 3.1 – Évolution du nombre de souscriptions aux réseaux mobiles cellulaires, donnant accès aux communications de données à débit « large bande », entre 2014 et 2016.

De plus, les *smartphones* et tablettes sont de plus en plus puissants et performants, permettant désormais d'en faire des outils de travail à part entière. Face aux nombreux avantages des appareils mobiles (légèreté, autonomie, rapidité, multitâche, encombrement minime, etc.), les opportunités offertes aux linguistes de terrain et aux chercheurs en documentation et en description des langues sont multiples, d'autant plus lorsqu'il s'agit de langues en danger ou de langues peu dotées qui nécessitent des collectes dans des conditions souvent difficiles.

1. https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/ITU_Key_2005-2016_ICT_data.xls

3.1 Contexte du développement de Lig-Aikuma

3.1.1 Le projet BULB

Le projet BULB (qui signifie “Breaking the Unwritten Language Barrier”) est un projet ANR-DFG franco-allemand, d’une durée de 3 ans, qui a débuté en mars 2015. 7 partenaires sont associés dans ce projet, répartis comme suit :

- 4 laboratoires de linguistique qui sont, côté français, le Laboratoire de Phonétique et Phonologie (LPP, UMR 7018) et le laboratoire Langage, Langues et Cultures d’Afrique Noire (LLACAN, UMR 8135) basés à Paris, ainsi que, côté allemand, le laboratoire de linguistique générale (Zentrum für Allgemeine Sprachwissenschaft également appelé ZAS) basé à Berlin et l’Institut de linguistique basé à Stuttgart (Institut für Linguistik, Universität Stuttgart) ;
- 3 laboratoires d’informatique qui sont, côté français, le Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur (LIMSI, UPR 3251) à Paris et le laboratoire d’Informatique de Grenoble (LIG, UMR 5217) ainsi que, côté allemand, l’Institut des technologies de Karlsruhe (Karlsruhe Institut for Technologies également appelé KIT).

BULB vise à développer une méthodologie pour la préservation de langues non écrites, au moyen de collectes et de technologies innovantes. Le projet se concentre sur trois langues bantoues typo-génétiquement différentes — d’après les critères de classification définis par Guthrie (1948) — qui sont le bàsàá (classe A43), le myènè (classe B11) et le mbochi (classe C25). Même si ces langues sont relativement bien décrites et documentées en comparaison de toutes les langues bantoues (il en existe environ 500), elles restent considérées comme très peu dotées. Certaines sont même évaluées comme étant en danger, d’après l’échelle EGIDS (Lewis et Simons, 2010).

3.1.1.1 Les langues du projet

Le bàsàá est une langue parlée dans le centre et sur les côtes du Cameroun. La SIL dénombre environ 300 000 locuteurs en 2005 [2]. Les premiers travaux sur cette langue datent du début des années 1900 (Rosenhuber, 1908). Depuis, un dictionnaire bilingue bàsàá-français est apparu (Lemb, De Gastines et Hebga, 1973) et la langue continue d’être décrite comme l’attestent les récents écrits de Hamlaoui et Makasso (2015) et Makasso, Hamlaoui et Lee (2017).

La langue myènè, quant à elle, est parlée au Gabon. Elle regroupe six variantes dialectales où l’inter-compréhension des locuteurs (environ 45 000 locuteurs sont recensés en 2007 [3]) est totale. Le dictionnaire bilingue mpongwè-français (Raponda-Walker et Tardy, 1934) constitue le premier document linguistique publié sur une variante de ce groupe. Ambouroue, Vanhoudt et Grégoire (2007) ont fourni pour la première fois une description globale de

2. https://www.ethnologue.com/language/bas
3. https://www.ethnologue.com/language/mye

la variante orungu (du groupe myènè). Des analyses plus spécifiques sur certaines propriétés linguistiques ont par la suite complété les études sur la langue, comme celle de Van de Velde et Amboroué (2011).

Enfin, le mbochi est une langue parlée au Congo par environ 161 000 personnes [4]. Le premier dictionnaire bilingue a été publié au début du 21ème siècle et comporte 31 775 entrées (Kouarata, 2000). L'étude de cette langue a fait l'objet de plusieurs thèses, notamment celles de Amboulou (1998) et Embanga Aborobongui (2013). La dernière étude en date concerne la relation entre l'intonation et les tons de cette langue (Rialland et Aborobongui, 2017).

Comme expliqué par Adda, Stüker *et al.* (2016), ces trois langues partagent des propriétés linguistiques complexes de la famille des langues bantoues, notamment au niveau de la morphologie (nominale et verbale), de la phonologie (lexicale et post-lexicale) et des tons (qui entraînent un changement de sens au niveau lexical, mais aussi au niveau grammatical). De plus, ces langues disposent de très peu de documents numériques et souffrent d'une convention orthographique instable. Ces particularités en font des enjeux majeurs pour le traitement automatique des langues.

3.1.1.2 Objectifs du projet

L'enjeu majeur de BULB est de produire des ressources (écrites) pour combler le manque de corpus (écrits) dans les langues peu voire non écrites. De ce fait, plusieurs objectifs nourrissent le projet. L'objectif premier est de collecter de manière massive des données langagières sur ces langues, de les faire répéter et de les traduire en français. Cet objectif en engendre, donc, un second : celui de concevoir des outils et méthodes automatiques, pour aider à la collecte et à la documentation de langues non écrites ou en danger, à travers une collaboration entre laboratoires spécialisés en linguistique de terrain et laboratoires spécialisés en traitement automatique de la langue. Une fois ces données orales récoltées, l'objectif suivant est d'utiliser des méthodes de traitement automatique du langage afin de segmenter, classifier et transcrire de manière non supervisée les enregistrements.

Ainsi, la première phase du projet a été consacrée à la collecte de données en bàsàá, en myènè et en mbochi. Dans ce but, le développement d'une application mobile pour la collecte sur le terrain a été envisagée. En effet, les appareils mobiles tels que le *smartphone* ou la tablette sont des outils aux atouts considérables. La légèreté, le faible encombrement, l'autonomie de l'appareil, mais aussi l'accessibilité du produit (les *smartphones* et tablettes sont de moins en moins coûteux et de plus en plus performants) sont des avantages non négligeables pour une mission sur le terrain où l'équipement du linguiste se doit d'être sommaire.

3.1.1.3 Méthodologie

La méthodologie utilisée dans le projet s'inspire des travaux antérieurs réalisés par Mark Liberman et Steven Bird, qui ont participé à la conception du projet.

4. https://www.ethnologue.com/language/mdw

Afin de créer des modèles efficaces et de tester différentes configurations, suffisamment de données doivent être collectées. Dans le projet, la cible envisagée est de 100 heures d'enregistrement par langue.

Ces enregistrements seront ensuite répétés (c'est-à-dire qu'un second enregistrement sera effectué dans la même langue que l'enregistrement initial, mais environnement et/ou locuteur différent). En effet, les conditions d'enregistrement, la variabilité introduite par les locuteurs, les disfluences, élisions et autres phénomènes liés à la parole spontanée, dégradent les performances des systèmes automatiques de traitement de la parole. Pour atténuer tous ces problèmes qui peuvent être présents dans les enregistrements spontanés ou qui peuvent survenir suite à de mauvaises conditions d'enregistrement, les enregistrements initiaux seront répétés (c'est le concept de *"respeaking"* expliqué un peu plus loin) par un ou deux locuteurs natifs de la langue source.

Ensuite, ces enregistrements répétés dans de meilleures conditions seront traduits oralement dans une langue pour laquelle des techniques de pointe en matière de traitement automatique du langage existent. Cette langue sera idéalement la langue véhiculaire du pays dans lequel s'effectuera les collectes.

Suite au constat alarmant que la plupart des 7 000 langues aujourd'hui répertoriées dans le monde seront éteintes d'ici une centaine d'années (Rymer, 2012), Bird (2016) indiquait dans un article paru dans le média en ligne *The Conversation* que « les cyberlinguistes du futur devront concevoir des algorithmes pour décrypter les enregistrements qui ont été réalisés avant cet événement d'extinction de masse ». Ainsi indique-t-il qu'il est urgent de collecter en masse les langues en danger. Ensuite, les progrès des méthodes computationnelles pourraient permettre de décrypter ces langues (comme Champollion avait décrypté les hiéroglyphes grâce à la pierre de Rosette).

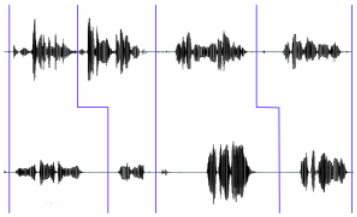
Métaphore de la pierre de Rosette.

Le projet s'inspire ainsi de cette métaphore. Le but est d'utiliser des méthodes automatiques et des techniques d'apprentissage probabiliste et d'apprentissage profond (réseaux de neurones) afin d'obtenir :

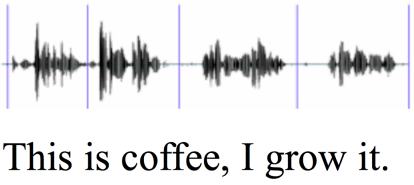
- des alignements entre l'enregistrement d'origine répété et sa traduction orale, comme illustré au graphique 3.2a.
- des alignements entre l'enregistrement d'origine répété et sa traduction écrite, comme illustré au graphique 3.2b.
- des alignements entre de la parole élicitée et des images (celles utilisées pour ladite élicitation), comme illustré au graphique 3.2c.

3.1.2 L'application initiale Aikuma

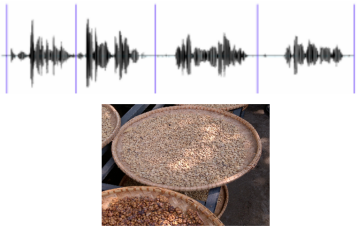
Aikuma est une application mobile, fonctionnant sur le système d'exploitation Android, développée dès 2013 par Hanke et Bird (2013). Partant du constat alarmant que les langues



(a) Illustration d'un alignement entre fichier re-parlé et fichier traduit oralement.



(b) Illustration d'un alignement parole/texte.



(c) Illustration d'un alignement parole/image.

Graphique 3.2 – Illustration des données parallèles collectées selon la méthodologie du projet.

sont menacées d'extinction, mais aussi que la majorité des langues ne possède aucune ressource écrite, Bird, Hanke *et al.* (2014) ont imaginé une application mobile collaborative permettant de récolter en masse des données orales accompagnées de leur traduction (également orale), afin de sauvegarder le plus rapidement possible ces langues en danger. Aikuma a été testée pour la première fois dans des villages éloignés au Brésil et au Népal (Bird, 2014) afin de recueillir des enregistrements audio en langues Tembé, Nhengatu et Kagate. Les auteurs ont distribué des téléphones mobiles disposant d'Aikuma, afin que les locuteurs s'enregistrent eux-mêmes.

L'application proposait alors trois fonctionnalités :

- **L'enregistrement de parole spontanée**, implanté pour que les locuteurs d'une langue en danger puissent s'enregistrer en toute autonomie et partager leurs enregistrements à travers la plateforme. Toutefois, puisqu'il n'est pas rare, sur le terrain, de se trouver dans un environnement bruité où des événements extérieurs dégradent la qualité acoustique de l'enregistrement (animaux, véhicules, conditions climatiques, etc.), les développeurs ont proposé une seconde fonctionnalité, appelée *Respeaking*.

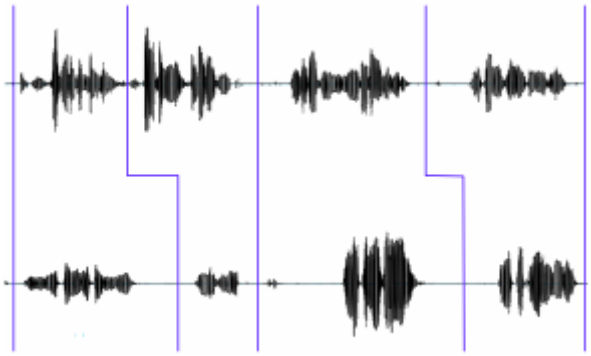
- **Le *respeaking* d'enregistrements vocaux**, a ainsi été conçu pour que d'autres locuteurs de la même langue répètent l'enregistrement initial dans de meilleures conditions acoustiques. Ce concept a été introduit pour la première fois en 2003 par Woodbury (2003). Il s'agit d'un nouvel enregistrement du fichier audio d'origine, dans la même langue que la langue source, mais prononcé de façon plus intelligible et dans un endroit le plus calme possible, afin d'être transcrit plus facilement *a posteriori* par un linguiste ou par une machine. Bird (2010) a été le premier à mettre en place cette technique, lors d'un terrain en Papouasie-Nouvelle-Guinée. La Papouasie-Nouvelle-Guinée est le pays du monde où se concentre le plus de langues : 820 langues indigènes. Nombre de ces langues sont en danger voire très proches de l'extinction, car elles sont parlées par moins de 1 000 habitants [5]. Par ailleurs, durant ce terrain Bird (2010) a élaboré "a model of Basic Oral Language Documentation" (c'est-à-dire « un modèle de documentation

5. https://www.cia.gov/library/publications/the-world-factbook/geos/pp.html

basique du langage oral »), adapté pour être utilisé lors de collectes dans des villages reculés et où les langues et cultures sont menacées de disparition.

- **La traduction orale**, nécessaire pour que ces enregistrements puissent être traduits dans une langue dominante, par des locuteurs bilingues, et ainsi permettre l'accès au sens de ce qui a été initialement exprimé par le locuteur de la langue en voie d'extinction.

L'intérêt majeur des modes *Respeaking* et *Traduction* réside dans le fait que chaque segment reparlé ou traduit est aligné avec l'enregistrement d'origine, grâce à un fichier de *mapping* qui recense l'horodatage de chaque segment audio reparlé ou traduit. Une illustration de cet alignement est présentée au graphique 3.3.
Posséder ces informations d'alignement est très précieux pour des tâches de traitement automatique du langage telles que la transcription automatique ou la traduction automatique.



Graphique 3.3 – Illustration de l'alignement des segments entre un fichier source et sa traduction. Source : Bird, Hanke *et al.* (2014).

Aikuma offrait également la possibilité d'être utilisée comme un réseau social local. Les utilisateurs pouvaient, s'ils possédaient une connexion réseau suffisante, déposer, synchroniser et partager leurs enregistrements avec les autres utilisateurs. De cette manière, les locuteurs monolingues pouvaient s'enregistrer dans leur langue maternelle, déposer leurs enregistrements sur la plateforme puis un autre utilisateur, bilingue, pouvait le récupérer et le traduire, pour qu'il soit accessible à une plus large communauté. Grâce à cette traduction — ou à ce *respeaking* —, les langues en danger étaient doublement documentées et pouvaient ainsi être exploitées ultérieurement (par des linguistes ou des machines). Enfin, les auteurs ont choisi une interface dénuée de texte pour que l'application soit utilisable par tout un chacun.

Néanmoins, Aikuma possédait quelques limites d'utilisabilité. À titre d'exemple, les modes étaient difficilement identifiables, l'import de médias tels que des images ou des vidéos n'était pas supporté ou encore aucune correction n'était possible, lors d'un *respeaking* ou d'une traduction, si le locuteur se trompait (il ne pouvait pas revenir sur ce qu'il venait de dire).

Depuis, le développement d'Aikuma s'est arrêté. Les développeurs ont choisi de développer une suite de micro-applications web, basée sur AngularJS et accessible depuis le Chrome Store, afin de rendre l'utilisation multi-plateforme (Bettinson et Bird, 2017). Ce modèle permet que chaque application gravite autour d'un projet général, la documentation des langues, tout en

étant indépendantes les unes des autres. Cette suite de micro-applications mobiles hybrides a été détaillée précédemment, à la sous-sous-section 1.2.1.2.

3.2 L'évolution d'Aikuma vers Lig-Aikuma

Aikuma a, originellement, été conçue pour répondre à la nécessité de sauvegarder la diversité linguistique mondiale et faciliter la récolte de données langagières. Cependant, les collectes de données à grande échelle nécessitent une organisation optimale des ressources, telle que des conventions strictes de nommage de fichiers et de métadonnées, pour favoriser l'accès aux données.

Aikuma possèdait des atouts évidents mais nous avons voulu des collectes dirigées par le linguiste de terrain, dont les fonctionnalités se conforment à la façon de travailler du linguiste de terrain, tout en conservant les 3 fonctionnalités d'origine : l'enregistrement de parole libre, la répétition de fichier audio, la traduction orale d'un fichier audio. Par ailleurs, ces fonctionnalités, initialement développées dans Aikuma, sont d'un intérêt indéniable en reconnaissance de la parole (découverte non supervisée par exemple) ou encore la traduction automatique « parole à parole ». C'est pourquoi nous avons décidé d'étendre Aikuma et non pas de créer une toute nouvelle application.

Lig-Aikuma est donc une extension d'Aikuma. Nous avons ajouté de nouvelles fonctionnalités afin de répondre aux besoins des linguistes de terrain qui veulent documenter et décrire une langue.

Cette évolution permet d'enregistrer les locuteurs dans leur environnement naturel, mais aussi de collecter des données ethno-linguistiques liées aux enregistrements. De plus, l'application a été réalisée de manière à ce que la collecte de parole soit efficace, grâce à un découpage en modes d'utilisation bien distincts. Mais aussi, deux nouveaux modes ont été implémentés par rapport à l'application initiale : d'un côté, le mode *Élicitation* permet désormais au linguiste de faire éliciter de la parole à un locuteur au moyen de texte, d'image ou encore de vidéo ; de l'autre, le mode *Vérification* permet au linguiste de corriger rapidement et simplement du texte (erreurs orthographiques, syntaxiques, fautes de prononciation, etc.).

3.2.1 Motivations

Les fonctionnalités d'Aikuma étaient prometteuses pour collecter des corpus parallèles de parole. Néanmoins, afin d'équiper le linguiste de terrain de manière utile et efficace, des améliorations ergonomiques et fonctionnelles étaient indispensables.

Nous avons tiré parti de cette application et lui avons ajouté des fonctionnalités qui nous paraissaient essentielles, pour que l'utilisateur final soit le linguiste de terrain et ainsi orienter l'application pour un travail de documentation et de description. Au lieu d'une application destinée à la sauvegarde et la pérennisation des langues, nous avons imaginé une application-outil

assistant le linguiste de terrain et lui proposant une nouvelle méthodologie de travail.

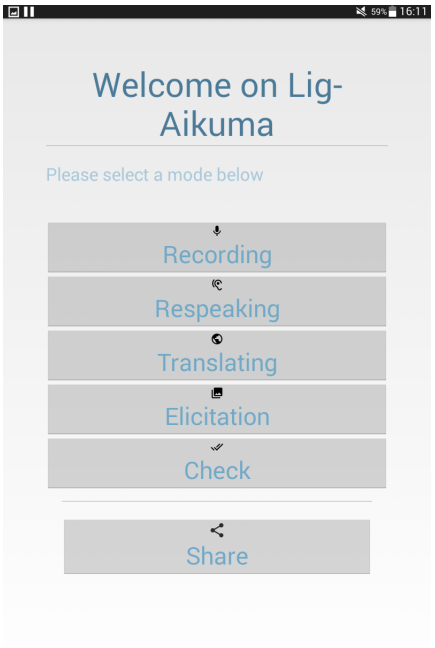
Au fil des échanges avec les linguistes de terrain ainsi que les partenaires du projet BULB, nous avons construit un cahier des charges contenant des cas d'usage complets qui ont abouti au développement de Lig-Aikuma.

Chaque mode (hormis les modes *Enregistrement*, *Respeaking* et *Traduction*) provient de discussions multiples et de rencontres régulières au cours desquelles les linguistes ont exprimé leur besoin et les informaticiens les limites de certaines implémentations. Finalement, Lig-Aikuma représente une formalisation de la façon de travailler du linguiste de terrain.

D'un point de vue technique, le développement de l'application s'est déroulé sur 3 ans. Une première version a été déployée en 2015, puis deux mises à jour majeures ont été publiées suite à l'encadrement de deux stagiaires. Le développement de Lig-Aikuma s'est déroulé en plusieurs phases. L'application mobile résulte de discussions menées entre linguistes et informaticiens mais également des expériences de terrain.

L'application est disponible sous licence AGPL (licence publique générale Affero). La dernière version en date peut être téléchargée depuis le site web http://lig-aikuma.imag.fr ou depuis le Play Store de Google. L'historique des versions stables de Lig-Aikuma est également publié sur la Forge du LIG [6] (la plus récente apparaissant en premier). Le code source est déposé sur le gitlab de l'Université Grenoble Alpes [7].

3.2.2 Modes d'utilisation



Graphique 3.4 – Écran d'accueil de Lig-Aikuma (capture d'écran depuis une tablette).

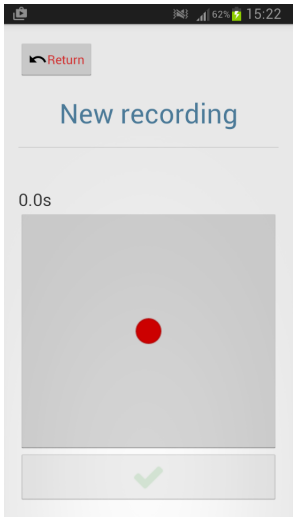
6. https://forge.imag.fr/frs/?group_id=805
7. https://gricad-gitlab.univ-grenoble-alpes.fr/besaciel/lig-aikuma.git

Lig-Aikuma propose six modes d'utilisation qui sont : l'Enregistrement, le *Respeaking*, la Traduction, l'Élicitation, la Vérification, le Partage. Les trois premiers modes sont hérités de l'application initiale Aikuma, mais ont été adaptés pour nos besoins. Chacun de ces modes est accessible directement à l'ouverture de l'application via un menu les différenciant les uns des autres, de façon à rendre l'application plus ergonomique. Les boutons, tout comme l'ensemble de l'application, sont illustrés par du texte, mais également par des icônes. Nous avons voulu rendre l'application simple d'utilisation en privilégiant ce type d'interface et ainsi en simplifier l'usage, car un tel outil peut paraître déroutant de prime abord pour un néophyte. Une copie de l'écran d'accueil à l'ouverture de Lig-Aikuma est présentée au graphique 3.4.

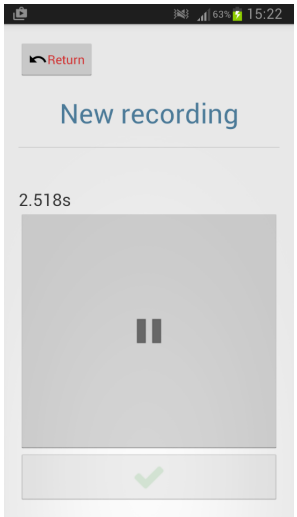
Enregistrement

Le mode *Enregistrement* permet au linguiste d'enregistrer le locuteur de manière libre, spontanée, sans consigne particulière. C'est un mode qui peut s'apparenter au dictaphone. Ce mode est légèrement plus évolué que celui de l'application initiale Aikuma, grâce au formulaire de métadonnées qui est à remplir avant l'enregistrement. Ce formulaire permet de collecter des informations sociolinguistiques sur le locuteur, mais aussi sur les conditions d'enregistrement. Ces informations sont très importantes pour le linguiste, au retour de son terrain. Elles permettent de dater l'enregistrement, de se rappeler dans quelles conditions acoustiques le linguiste et son informateur se trouvaient au moment de l'enregistrement, etc. Ce formulaire fait l'objet d'un paragraphe détaillé à la sous-section 3.2.3.

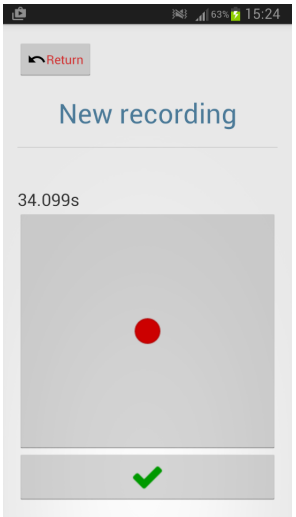
Le graphique 3.5 présente chacune des étapes lors d'un enregistrement. Le graphique 3.5a montre l'activité sur laquelle arrive l'utilisateur après avoir validé le formulaire de métadonnées renseigné précédemment. Le graphique 3.5b présente l'interface lorsque l'enregistrement est en cours et le graphique 3.5c l'interface une fois l'enregistrement terminé.



(a) Vue de l'activité **avant** l'enregistrement.



(b) Vue de l'activité **pendant** l'enregistrement.



(c) Vue de l'activité **après** l'enregistrement.

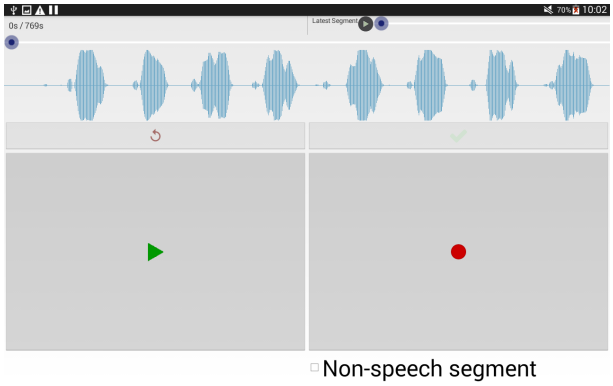
Graphique 3.5 – Illustration du mode Enregistrement de Lig-Aikuma (capture d'écran depuis un *smartphone*).

Il arrive fréquemment qu'un enregistrement ait été initialement produit avec du bruit environnant et devienne ainsi difficilement post-traitable, ou encore que le linguiste se rende compte, lors de la phase de réécoute, que l'enregistrement n'est pas analysable, car le locuteur a parlé trop faiblement, avec une mauvaise articulation ou bien un débit trop rapide. Pour pallier ce problème, l'utilisateur peut utiliser le mode *Respeaking*, implémenté à l'origine dans Aikuma, mais nettement enrichi et stabilisé dans Lig-Aikuma.
Inspiré de la possibilité de traduire un enregistrement dans Aikuma, le mode *Traduction* de Lig-Aikuma offre la même interface améliorée que le mode *Respeaking* et les mêmes fonctionnalités. Ce n'est que la finalité du mode qui est différente : cette fois, la langue d'enregistrement est différente de la langue du fichier source.

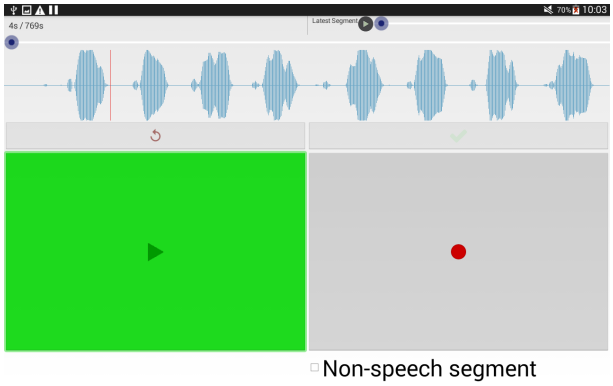
Respeaking et Traduction

Les modes *Respeaking* et *Traduction* représentent des tâches quasiment similaires, à ceci près que la langue d'enregistrement cible ne sera pas la même selon le mode. Ces modes font tous deux appel à un fichier d'origine contenant de la parole. L'import d'un fichier enregistré avec un autre outil que Lig-Aikuma est possible, tant que le fichier respecte les caractéristiques de codage audio suivantes : il doit être codé sur un seul canal sonore (c'est-à-dire être monocanal), avoir une fréquence d'échantillonnage fixée à 16kHz, être réglé sur une échelle de quantification à 16bit et être au format WAV. Dans ce cas, l'utilisateur doit, dans un premier temps, renseigner des informations concernant le fichier qu'il vient d'importer, inconnu par Lig-Aikuma, au moyen d'un premier formulaire de métadonnées. Cette étape n'est pas nécessaire si le fichier importé a été enregistré, à l'origine, avec Lig-Aikuma. Une fois le fichier d'origine à reparler ou traduire sélectionné — cette étape vaut aussi bien pour l'import d'un fichier audio initialement enregistré avec Lig-Aikuma que pour un fichier externe à l'application —, l'utilisateur doit remplir un formulaire (de métadonnées) concernant le locuteur et la langue de l'enregistrement qui va être fait. Ainsi, pour un *respeaking*, la langue saisie dans le formulaire sera la même que celle de l'enregistrement d'origine alors que, pour une traduction, la langue saisie sera différente. Pour faciliter le remplissage et éviter les confusions, la langue de l'enregistrement (qui va être produit) est automatiquement remplie dans le formulaire de métadonnées s'il s'agit d'un *respeaking* alors que ce champ est vierge lorsqu'il s'agit d'une traduction.

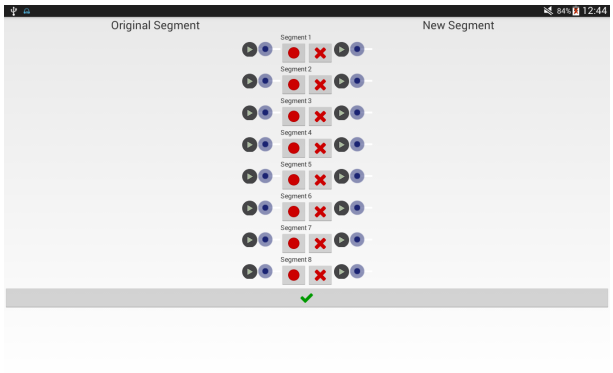
Le graphique 3.6 présente le déroulement d'un *respeaking*. Le principe et la disposition des boutons sont hérités de l'application Aikuma : l'utilisateur appuie sur le bouton de lecture ▶ à gauche de l'écran et reste appuyé jusqu'à ce qu'il juge qu'il a assez de matière à reparler ou traduire. Il peut relâcher et appuyer sur le bouton de lecture autant de fois que nécessaire. Une fois qu'il se sent prêt et qu'il a relâché le bouton de lecture, il peut presser le bouton ● situé à droite de celui-ci pour s'enregistrer. L'utilisateur recommence les deux actions (lecture-enregistrement) jusqu'à la fin du fichier d'origine.
Nous avons apporté à cette activité Android quatre améliorations majeures, visibles sur le graphique 3.6a et le graphique 3.6b :



(a) Vue de l'activité **avant** *respeaking*/traduction.



(b) Vue de l'activité **pendant** *respeaking*/traduction.



(c) Vue de l'activité **après** *respeaking*/traduction.

Graphique 3.6 – Illustration des modes *Respeaking* et *Traduction* de Lig-Aikuma - capture d'écran à partir d'une tablette.

- L'utilisateur a la possibilité d'écouter, grâce au bouton de lecture situé en haut à droite de l'écran, le segment qu'il vient de reparler/traduire. Dans le cas où ce morceau enregistré ne lui convient pas, il peut le refaire en appuyant sur le bouton dédié.
- Une barre de progression lui indique sa position dans la lecture du fichier audio original.
- Un visualiseur de l'enveloppe du signal audio d'origine lui offre un moyen supplémentaire de se repérer durant la lecture du fichier.
- Une case à cocher permet à l'utilisateur d'indiquer les segments qui ne contiennent pas de parole pertinente.

Une fois que l'utilisateur a terminé la tâche — en appuyant sur le bouton de validation ✔ —, un résumé de tous les segments créés s'affiche. Cette activité est présentée sur le graphique 3.6c. L'utilisateur peut ainsi réécouter n'importe quel segment original, couplé au même segment reparlé/traduit, et le ré-enregistrer autant de fois qu'il le souhaite. Les boutons de lecture ▶ situés à gauche de l'écran permettent de réécouter un segment du fichier original tandis que ceux de droite permettent de réécouter un segment qui a été répété/traduit durant la séance. Le bouton ● sert à ré-enregistrer de nouveau le segment original, dans le cas où le nouveau segment qui a été répété/traduit n'est finalement pas satisfaisant. Enfin, le bouton ✖ est utile lorsque le segment qui vient d'être répété/traduit sur l'instant ne convient pas : ce bouton supprime ce qui vient d'être fait et permet de revenir à la toute première version du segment répété/traduit.

Dans le cas où l'utilisateur est contraint de terminer la tâche en cours (par manque de temps par exemple), il peut à tout moment valider ou appuyer sur le bouton de retour arrière de son appareil mobile et choisir d'enregistrer sa session. Il pourra ensuite relancer l'application et récupérer sa séance à l'endroit où il s'était arrêté. Cette fonctionnalité de sauvegarde est présentée à la sous-sous-section 3.2.4.6 et au graphique 3.14.

Comme indiqué précédemment, nous avons ajouté la possibilité de notifier lorsque le contenu du fichier original n'est pas de la parole.
À l'origine, dans l'application Aikuma, l'horodatage (en nombre d'échantillons) de chaque couple de segment écouté-répété/traduit était sauvegardé dans un fichier texte appelé fichier de *mapping*. Ces informations d'alignement entre les fichiers audio sont précieuses pour la construction de corpus parallèles en traduction automatique par exemple. Nous avons bien entendu conservé cette fonctionnalité, mais nous nous en sommes également inspirés pour implémenter une sauvegarde des segments de non-parole. Ainsi, au moyen d'une simple case à cocher, l'utilisateur peut indiquer les segments qui ne contiennent pas de parole pertinente. Par exemple, au moment de l'écoute, s'il entend du bruit ou de la parole non exploitable ou encore du silence, il peut le notifier au moyen de cette case à cocher (visible sur le graphique 3.6). L'utilisateur s'enregistre alors en indiquant que ce qu'il vient d'entendre n'est pas exploitable. En résumé, si la case est décochée, l'application se comporte comme à l'origine : l'horodatage de tous les segments est enregistré dans un fichier de *mapping*[8]. En revanche, si la case est cochée, les informations de temps (le début et la fin) concernant le segment qui vient d'être entendu ainsi que le commentaire qui lui aura été associé sont également enregistrées, mais dans un fichier à part[9].

Nous avons voulu faire le lien entre ces fichiers de *mapping* enregistrés par l'application et l'utilisation que le linguiste (de terrain) peut en faire. Dans ces fichiers de *mapping*, ce sont le nombre d'échantillons audio qui sont recensés. Nous avons eu l'idée de transformer ces fichiers pour qu'ils soient lus par des logiciels qui permettent de faire de la transcription. Ainsi,

8. Le nom du fichier est construit de la manière suivante :
`<nomFichierOriginal>_<rspk/trsl>.map`
9. Le nom du fichier est construit de la manière suivante :
`<nomFichierOriginal>_<rspk/trsl>_non-speech.map`

l'application génère désormais un fichier au format CSV présentant 5 colonnes (délimitées par une virgule) qui indiquent pour chaque segment reparlé/traduit :

- le nom de la *tier* (original, rspk (pour *respeaking*) ou trsl (pour *traduction*) ;
- le temps de début (en millisecondes) ;
- le temps de fin (en millisecondes) ;
- la durée totale de ce segment (en millisecondes) ;
- une annotation (typiquement, « non-speech » pour les segments spécifiés depuis l'application et rien pour les autres).

Ce fichier peut être directement importé dans le logiciel Elan. Il suffit de l'importer via l'onglet *Fichier*, puis *Importer*, et sélectionner « CSV » [10].

L'import de ce fichier dans Elan crée des *tiers* automatiquement découpées. Celles-ci proviennent des segmentations effectuées avec Lig-Aikuma par l'utilisateur durant des sessions de *respeaking* et de traduction. Ainsi, le linguiste peut visualiser l'alignement des segments produits avec Lig-Aikuma entre un fichier d'origine et son *respeaking* ou entre un *respeaking* et sa traduction. Le découpage en unités plus petites, lors d'un travail d'analyse linguistique, en est facilité. De plus, les segments qui auront été signalés comme « non parole » dans l'application seront automatiquement annotés.

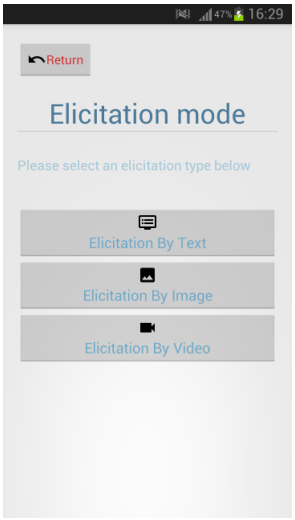
Les tâches de *respeaking* et de traduction ne sont pas des activités triviales : elles font appel à des opérations cognitives complexes. Ces modes demandent une allocation conséquente des ressources du locuteur. C'est pourquoi nous avons intégré un visualiseur de l'enveloppe du signal audio du fichier d'origine, visible au graphique 3.6a et au graphique 3.6b, pour assister le locuteur dans sa tâche de *respeaking* ou traduction. La visualisation de l'enveloppe acoustique du signal est un moyen lui permettant de répéter de manière plus fluide ce qu'il vient d'écouter. La compréhension faisant appel à des processus mentaux complexes, inhérents à chacun, il est important de laisser à l'utilisateur le libre contrôle de la durée d'écoute. De plus, la visualisation est un moyen pratique, permettant à la tâche de *respeaking*/traduction de s'adapter aux ressources cognitives de chacun. Enfin, cette fonctionnalité permet d'orienter l'utilisateur en lui suggérant des points d'arrêt dans la lecture du fichier audio. En lui livrant des indices sur l'enregistrement audio qu'il doit traiter, la tâche de réécoute-répétition/traduction en est facilitée. Son utilisation est détaillée à la sous-sous-section 3.2.4.4.

Le mode d'enregistrement (et les deux modes *respeaking* et traduction qui en découlent) sont adaptés à un enregistrement source de parole libre, où les propos du locuteur ne sont influencés d'aucune sorte. Le locuteur n'est aidé d'aucun support pour produire du contenu. Afin d'aider à la production de contenu ou d'orienter le locuteur sur un sujet précis, nous avons

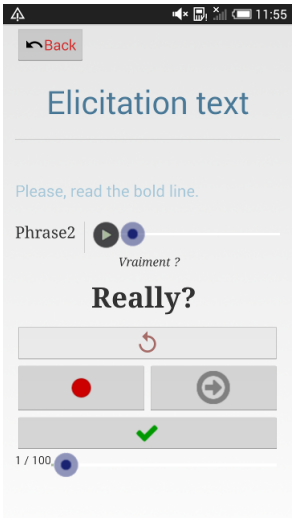
10. pour plus d'informations sur l'import d'un fichier CSV dans Elan, se référer à la documentation officielle : http://www.mpi.nl/corpus/html/elan/ch01s04s02.html (Section 1.4.2.5)

implémenté un mode d'élicitation. Ce mode nommé *Élicitation* offre la possibilité d'importer 3 types de médias à partir desquels le locuteur devra s'exprimer.

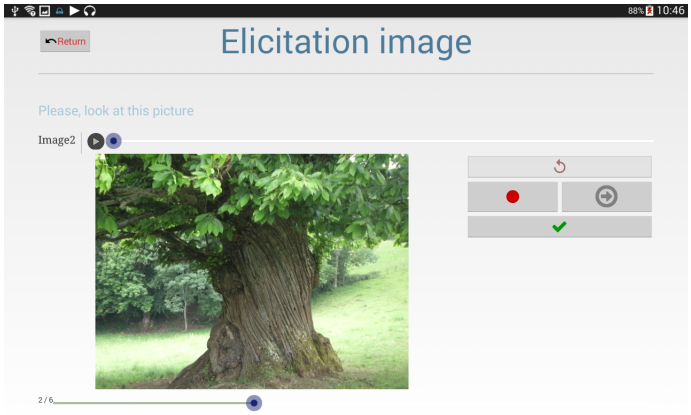
Élicitation



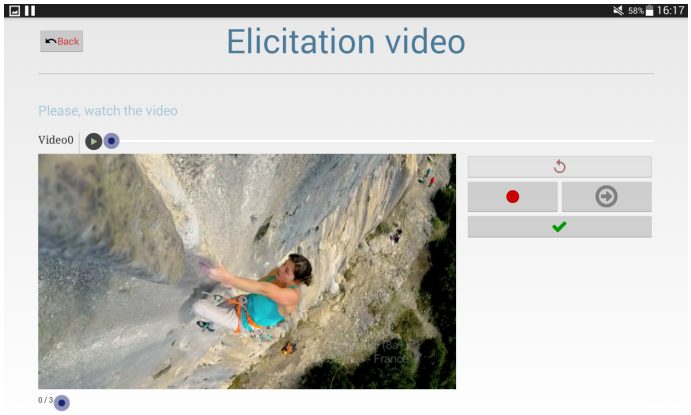
(a) Vue de l'activité présentant les choix possibles dans le mode *Élicitation*.



(b) Vue de l'activité *Élicitation par texte*.



(c) Vue de l'activité *Élicitation par image*.



(d) Vue de l'activité *Élicitation par vidéo*.

Graphique 3.7 – Illustrations du mode *Élicitation* de Lig-Aikuma.

Le mode *Élicitation* permet de faire parler les locuteurs à l'aide d'un support. Nous avons inclus dans ce mode la possibilité d'éliciter de la parole à partir de trois stimuli : du texte, des images ou des vidéos. Le graphique 3.7 illustre les 4 activités implémentées pour ce mode d'élicitation. Après sélection du mode à l'écran d'accueil, l'utilisateur a le choix entre différents sous-modes : *Élicitation par texte*, *Élicitation par image* et *Élicitation par vidéo*. Cette activité est présentée sur le graphique 3.7a.
L'élicitation à partir de texte est illustrée par le graphique 3.7b. Nous pouvons voir en gras, la phrase à lire par le locuteur et en italique une traduction (optionnelle) de cette phrase. En dessous à gauche se trouve le bouton permettant de lancer l'enregistrement et par simple (nouvelle) pression d'arrêter celui-ci. À droite du bouton d'enregistrement se trouve le bouton permettant de passer à la phrase suivante. Au dessus de ces deux boutons se situe le bouton permettant de recommencer l'enregistrement si l'utilisateur n'est pas satisfait de ce qu'il vient de dire alors qu'en dessous se trouve le bouton de validation qui enregistre et met fin à la session. Enfin, il y a, en haut à gauche, l'habituel bouton retour qui permet de revenir en arrière et sortir du mode. Ensuite, les illustrations du graphique 3.7c et du graphique 3.7d présentent respectivement l'élicitation à partir d'image et l'élicitation à partir de vidéo. Les mêmes boutons (des illustrations aux fonctionnalités) que ceux de l'élicitation par texte sont présents dans ces modes. Pour finir, les illustrations du graphique 3.7a et du graphique 3.7b montrent la disposition de l'interface en mode portrait sur un téléphone mobile alors que celles du graphique 3.7c et du graphique 3.7d montrent l'affichage en mode paysage sur une tablette.

Le principe de ce mode est le suivant : le linguiste place dans un dossier de l'appareil mobile les fichiers qui lui serviront de support pour l'élicitation (texte, images ou vidéos). En sélectionnant le mode qu'il désire, l'application se charge d'importer les médias puis de les afficher à l'écran. Chaque phrase/image/vidéo peut-être élicitée ou passée. À la fin de la session (que toutes les ressources aient été vues ou non), le linguiste peut valider les fichiers audio enregistrés dans l'appareil. Un fichier (nommé *linker.txt*), permettant de relier le chemin absolu du fichier WAV avec le chemin absolu de l'image associée, est également généré au même endroit que les enregistrements.

Les modes présentés jusqu'à présent permettent de faire de l'enregistrement de parole. Nous avons doté Lig-Aikuma de deux autres fonctionnalités : un mode de vérification imaginé dans un premier temps pour du post-traitement de signaux audio, ainsi qu'un mode de partage des fichiers audio enregistrés avec Lig-Aikuma pour que le linguiste puisse transmettre les enregistrements au locuteur, si celui-ci le désire.

Vérification

Le mode *Vérification* propose deux sous-modes nommés *Vérification de mots* et *Vérification de transcription*. Ces modes ont été implémentés afin de produire des vérifications textuelles simples. Les usages peuvent être multiples ; ils dépendront du fichier chargé par l'utilisateur. L'avantage principal de ce mode est que la tâche demandée de vérification peut être effectuée de n'importe où et de manière totalement autonome. L'utilisateur n'est pas contraint

par son environnement ou par le support sur lequel il doit travailler. Par exemple, si nous avions développé une application sur ordinateur, notre informateur aurait été contraint par le lieu, l'encombrement de l'outil et le délai d'exécution de celui-ci (temps au démarrage, etc).

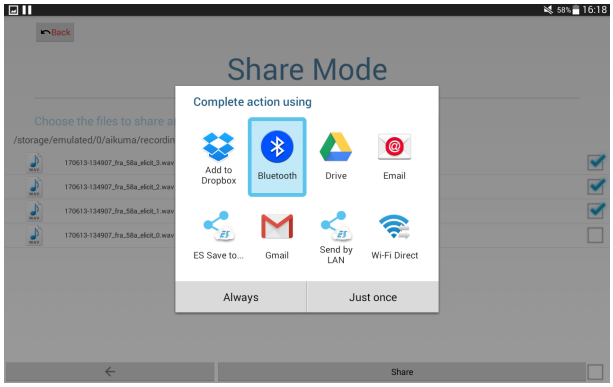
Le mode *Vérification de mots* est une activité qui prend 1 seule entrée : un fichier texte (format CSV dont les séparateurs sont la virgule et le point-virgule). Le mode affiche un mot principal (le premier mot du fichier), en gras, suivi d'une liste d'autres mots précédés d'une case à cocher. Dans le cas de notre étude, nous avons utilisé des mots, mais la granularité n'est pas limitée : ce peut être des phrases ou des caractères. Dans le cadre de notre travail, le mode *Vérification de mots* nous a permis de vérifier des variantes orthographiques du wolof (pour rappel, comme explicité à la sous-section 2.3.5, l'orthographe du wolof est instable). Pour ce faire, nous avons extrait les lemmes des dictionnaires bilingues en notre possession (Diouf, 2003) et (Fal, Santos et Doneux, 1990) et leurs variantes orthographiques — lorsqu'elles étaient indiquées —. Puis, nous avons demandé à nos informateurs experts du wolof si les différentes orthographes tirées des dictionnaires étaient couramment usitées et si les mots pouvaient s'employer dans le même contexte. Voici quelle était la consigne : *Cocher la/les variante(s) qui a/ont un sens identique au mot affiché en gras. La ou les variantes sélectionnée(s) ne doivent entretenir aucune relation de polysémie.* Notre but était ainsi d'obtenir les mots qui ont le même sens afin de pouvoir les substituer dans n'importe quel contexte (selon un axe paradigmatique).

L'autre mode de vérification, appelé *Vérification de transcriptions*, autorise en entrée 2 médias : un fichier sonore et un fichier texte. L'activité se charge simplement d'importer un fichier audio et affiche du contenu textuel. L'association entre le fichier audio et le contenu textuel s'effectue à travers le fichier texte qui contient le nom du fichier audio. Le principe est le suivant : l'utilisateur écoute le fichier audio et indique si le texte affiché correspond à ce qu'il vient d'entendre, au moyen d'une case à cocher.
Dans le cadre de notre travail, nous nous sommes servis de ce mode afin de vérifier l'exacte correspondance entre des fichiers audio en notre possession et leurs transcriptions. Cela nous a permis, durant la phase de construction de notre système de RAP, d'éliminer de nos corpus d'évaluation et de test des phrases qui avaient mal été prononcées ou lues lors de notre campagne de collecte.

Partage

Le mode *Partage* offre la possibilité au linguiste d'échanger, avec le locuteur, les enregistrements effectués durant la séance. Plusieurs choix sont possibles pour le transfert. Ainsi, si les deux terminaux bénéficient d'un accès à Internet, l'envoi peut se faire par e-mail. Ou bien, si les appareils en sont dotés, le partage peut s'effectuer au moyen de la technologie Bluetooth. Cette fonctionnalité est représentée au graphique 3.8.

(a) Explorateur de fichier permettant de sélectionner le ou les fichiers audio à partager.



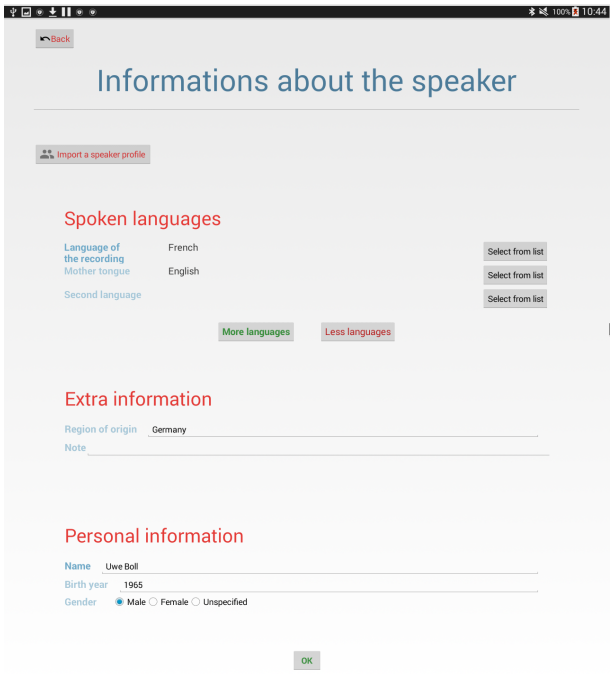
(b) Boîte de dialogue permettant de sélectionner l'application avec laquelle transférer le ou les fichiers audio.

Graphique 3.8 – Illustration du mode de partage de Lig-Aikuma.

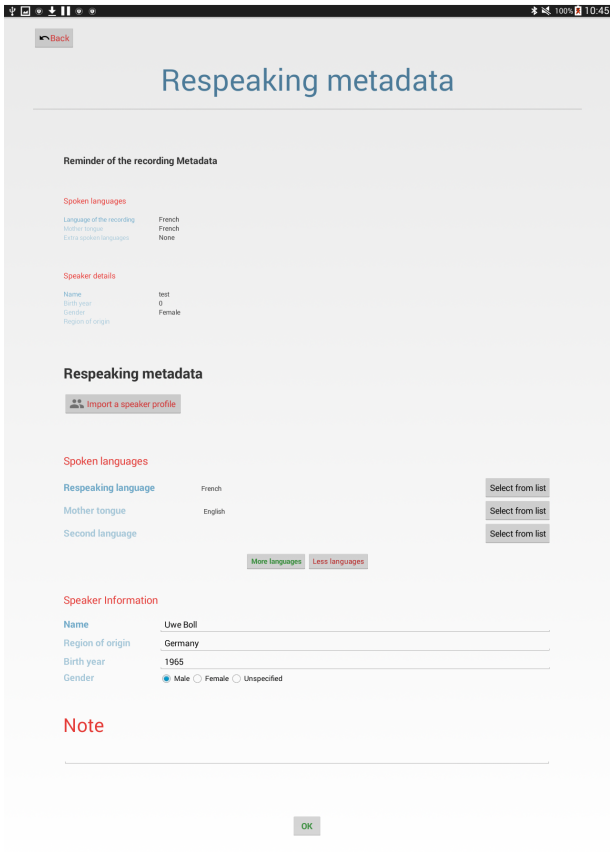
3.2.3 Métadonnées

L'application initiale Aikuma enregistrait déjà quelques métadonnées correspondant aux fichiers audio enregistrés. Pour les linguistes de terrain, les informations relatives au locuteur interrogé sont cruciales. Traditionnellement, le linguiste de terrain remplit un questionnaire sociolinguistique avant toute séance. Des informations comme les langues parlées par le locuteur, la langue de communication avec ses parents, sa région d'origine, sa durée de scolarisation, sont des renseignements nécessaires au linguiste. À partir de ce constat, il nous a semblé évident d'introduire dans l'application mobile un formulaire complet concernant le locuteur.

Ainsi, avant tout enregistrement, le linguiste de terrain remplit un des formulaires illustrés au graphique 3.9 portant sur les langues du locuteur et les informations le concernant. Le graphique 3.9a présente le formulaire à remplir dans le mode *Enregistrement* tandis que le graphique 3.9b montre le formulaire à remplir lorsque l'utilisateur se trouve dans le mode *Respeaking*. Le formulaire à remplir dans le mode *Traduction* est équivalent à celui du mode *Respeaking*, excepté le nom des champs (« translating » en lieu et place de « respeaking »).
La différence entre les deux formulaires du graphique 3.9 réside dans la présence d'un en-tête récapitulant les informations du fichier audio original qui va être répété (ou traduit). Ces infor-



(a) Formulaire de métadonnées dans le mode *Enregistrement* de Lig-Aikuma.



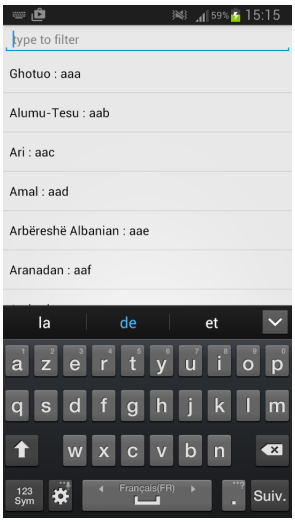
(b) Formulaire de métadonnées dans le mode *Respeaking* de Lig-Aikuma.

Graphique 3.9 – Illustrations du formulaire de métadonnées dans Lig-Aikuma.

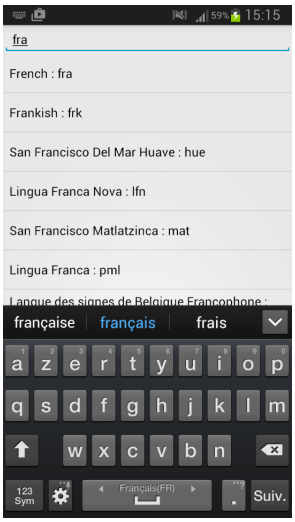
mations sont celles qui ont été saisies dans le formulaire illustré au graphique 3.9a, au moment de l'enregistrement (libre) initial.

Les champs sont, donc, les suivants (ceux accompagnés d'un astérisque sont obligatoires) :

- la langue dans laquelle l'enregistrement va être produit* ;
- la langue maternelle du locuteur ;
- la langue seconde du locuteur (s'il en parle une autre) ;
- les autres langues parlées (si plus que 2) ;
- le nom du locuteur* ;
- la région d'origine du locuteur ;
- l'année de naissance du locuteur ;
- le genre du locuteur ;
- un champ « note » permettant de renseigner des informations complémentaires.



(a) Liste de langues pré-chargée dans l'application.



(b) Utilisation du filtre pour réduire les langues affichées à l'écran.

Graphique 3.10 – Processus de sélection d'une langue dans Lig-Aikuma.

Enfin, les langues peuvent être renseignées de deux façons, présentées au graphique 3.10 : l'utilisateur peut sélectionner la langue du locuteur dans une liste prédéfinie de langues. Il peut, soit faire défiler (exemple au graphique 3.10a) cette liste, soit filtrer les langues en tapant les premiers caractères du nom de la langue (exemple au graphique 3.10b).

Lors de la validation du formulaire, tout ce qui a été saisi dans les champs est sauvegardé et conservé dans un fichier au format JSON (un exemple est fourni en annexe B). Ce fichier permet de structurer les informations liées au locuteur. C'est ce dernier qui permet la récupération des informations du fichier audio d'origine dans les modes *Respeaking* et *Traduction*, qui sont affichés en tant que « Résumé des informations » (visible sur le graphique 3.9b).

Les informations renseignées dans le formulaire de métadonnées, mais également les enregistrements qui sont produits lors des séances sont des données personnelles. Afin de garantir au locuteur l'intégrité et la confidentialité de ces données, nous avons intégré à l'application un formulaire de consentement qui doit être lu (à voix haute par le linguiste si le locuteur n'est pas en capacité du faire) et signé par les deux parties (linguiste et locuteur).

3.2.3.1 Formulaire de consentement

Lig-Aikuma est une application de collecte de parole. Autrement dit, des données personnelles sont récoltées, ce qui nous a poussé à prendre en compte des considérations éthiques. Un formulaire de consentement est un document important puisqu'il permet au linguiste d'informer concrètement le participant sur le but de ces enregistrements ainsi que ce à quoi celui-ci adhère en acceptant d'être enregistré. Du point de vue de l'informateur, ce document lui permet d'être pleinement informé, rassuré, considéré et de se sentir propriétaire des données recueillies sur lui au travers de l'autorisation qu'il émettra sur ce qui va être collecté. Enfin, ce formulaire permet d'évoquer la question de la confidentialité et de la propriété des données.

Si le linguiste de terrain travaille avec des informateurs alphabétisés, Sakel et Everett (2012) suggèrent de lire à haute voix une explication des travaux qui vont être effectués (contexte des travaux de recherche, objectifs, méthodologie adoptée, financement, partenaires, profit potentiel, rémunération, gain personnel retiré par le linguiste) — qui auront été mis par écrit au préalable — puis de faire signer l'informateur pour accord.
Dans le cas où le linguiste travaille avec des informateurs ne sachant pas lire, les auteurs proposent de filmer l'informateur en train de consentir aux travaux de recherche qui vont être menés (enregistrements, études, diffusion des données, etc).

Ainsi, l'intégration d'un formulaire de consentement généré à la validation du formulaire de métadonnées nous a paru être évidente. Lors de la validation du formulaire de métadonnées, une nouvelle activité affiche le formulaire de consentement (au format PDF), pré-rempli automatiquement avec les informations renseignées en amont dans le formulaire de métadonnées. Il précise l'objectif de ces enregistrements et permet à l'informateur d'indiquer qu'il consent à être enregistré. Un exemplaire de ce formulaire est présenté en annexe E. Idéalement, la signature de ce celui-ci est faite électroniquement, à partir d'un logiciel dédié, mais elle peut évidemment, aussi, se faire manuellement après impression du formulaire.

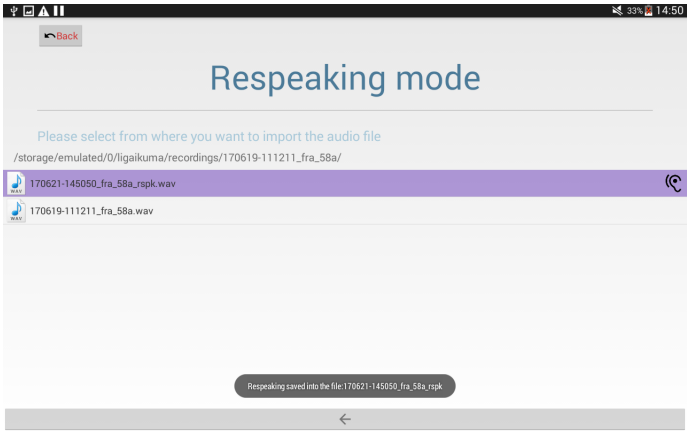
3.2.4 Autres améliorations

Plusieurs efforts ont été menés afin de répondre à des critères ergonomiques élémentaires. Nous avons ajouté de nombreuses améliorations visuelles afin de fournir à l'utilisateur un meilleur confort d'utilisation et une application compréhensible, dans laquelle il est facile de naviguer.

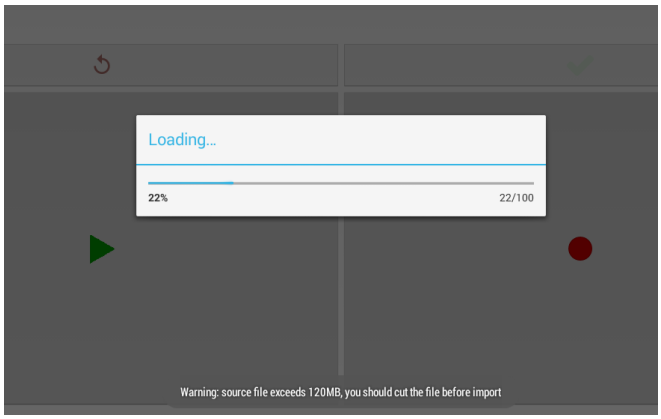
3.2.4.1 Retours utilisateurs

Nous avons insisté sur la clarté de l'application en affichant des messages d'information afin que l'utilisateur soit toujours informé des traitements effectués par l'application. Par exemple, un message explicite sera affiché en cas d'erreur en indiquant à l'utilisateur ce qui s'est mal déroulé, ou bien un message annonçant le nom du dossier des fichiers sauvegardés sera affiché lorsqu'une session est terminée (exemple au graphique 3.11) ou encore un avertissement sera déployé si un fichier audio trop long (d'une taille supérieure à 120Mo) est sélectionné pour un *respeaking*/traduction (exemple au graphique 3.12).

Aussi, nous avons inséré, dans chaque activité de l'application faisant appel à un import de fichier, le chemin absolu de l'endroit où se trouve l'utilisateur. Ceci lui permet de savoir instantanément où il se situe dans l'arborescence de ses fichiers. Cet indicateur est visible sur le graphique 3.13.



GRAPHIQUE 3.11 – Message d'information indiquant à l'utilisateur le nom du dossier dans lequel les fichiers audio ont été enregistrés.



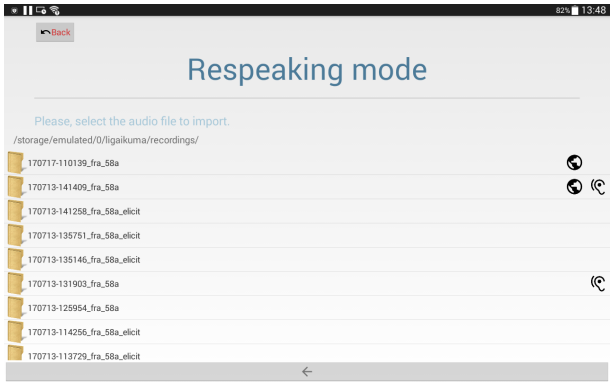
GRAPHIQUE 3.12 – Message avertissant l'utilisateur que le fichier importé est très long et qu'il devrait le découper, par mesure de précaution.

3.2.4.2 Barre de progression

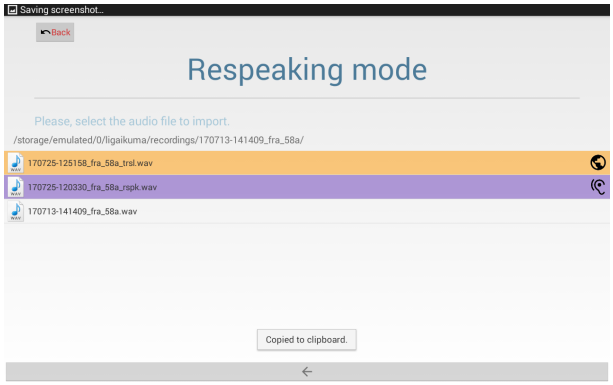
Tous les modes intègrent un objet représentant la progression de la session. Par exemple, dans le mode *Enregistrement*, la progression est représentée par l'affichage du temps d'enregistrement qui défile dès l'appui sur le bouton et s'arrête lors d'une nouvelle pression. Dans le mode *Élicitation*, l'utilisateur a connaissance de l'état d'avancement de la tâche grâce à une barre qui progresse au fur et à mesure du déroulement de l'activité, mais aussi grâce à l'indication du nombre d'éléments traités sur le nombre total d'éléments. Cette fonctionnalité est visible sur l'illustration graphique 3.7c, en dessous de l'image et sur l'illustration graphique 3.7d, en dessous de la vidéo.

3.2.4.3 Colorisation et icônes

Dans Lig-Aikuma, de nombreux types de fichiers sont manipulés : sons, textes, images, vidéos. Par conséquent, il a fallu trouver des moyens pour que la navigation dans l'application reste commode et fluide. À titre d'exemple, dans les modes *Respeaking* et *Traduction*, l'utilisateur doit choisir un fichier audio (original) à traiter. Puisque les fichiers reparlés ou traduits se situent tous dans le même répertoire (celui du fichier parent), le menu de sélection des fichiers à importer devient rapidement chargé et les différents types de fichiers ne sont plus distinguables, malgré le suffixe ajouté à la fin du nom.



(a) Lors de l'importation d'un fichier, une icône rappelle quel type de fichier audio est contenu dans les dossiers listés.



(b) Lors de l'importation, les fichiers sont colorés en fonction de leur type (enregistrement libre, *respeaking* ou traduction).

Graphique 3.13 – Différenciation des fichiers à sélectionner dans Lig-Aikuma.

Pour aider à la reconnaissance de tous les types de fichiers, différentes icônes ont été insérées dans l'application. 3 illustrations ont été mises en place :

- Dans la boîte de dialogue d'importation de fichiers, une icône représentant un dossier ( ) et une autre représentant un fichier sonore ( ) ont été ajoutées pour discerner le type de l'élément. Ces icônes sont illustrées sur la Figure 3.13.
- Dans le mode *Élicitation*, l'icône qui représente un dossier contenant des images tandis que l'icône qui représente un dossier contenant des vidéos ont été intégrées, facilitant ainsi la sélection d'un média en particulier.

- Les icônes illustrant les modes *Respeaking* et *Traduction* sont rappelées dans l'arborescence des dossiers, si ceux-ci contiennent ces types de fichier audio. Ainsi, l'icône indique que le dossier contient un fichier audio reparlé tandis que l'icône représente un fichier audio traduit. La Figure 3.13a illustre l'implémentation de ces icônes.

  Une fois dans le dossier, un fond coloré permet de clairement différencier les fichiers, en plus des icônes toujours présentes. Les fichiers originaux, pour leur part, ne sont pas colorés. La Figure 3.13b montre l'implémentation des fonds colorés et des icônes.

3.2.4.4 Enveloppe du signal

Dans notre version initiale, pendant un *respeaking* ou une traduction, l'utilisateur écoutait et stoppait la lecture du fichier audio d'origine de manière instinctive, à l'aveugle. Pour rendre ce découpage plus intuitif, nous avons intégré à ces deux modes la visualisation de l'enveloppe du signal audio. Cette fonctionnalité est illustrée sur le graphique 3.6a et le graphique 3.6b. Lorsque l'utilisateur appuie sur le bouton de lecture, un marqueur rouge lui indique la position de la tête de lecture. L'enveloppe du signal défile au fur et à mesure de l'écoute de l'enregistrement. La visualisation de l'enveloppe du signal montre les passages de silence (le signal est plat), ce qui permet à l'utilisateur d'anticiper l'arrêt de la lecture du fichier original. D'emblée, la segmentation des fichiers originaux est plus propre et entraîne une répétition ou une traduction plus fluide et mieux construite.

3.2.4.5 Géolocalisation

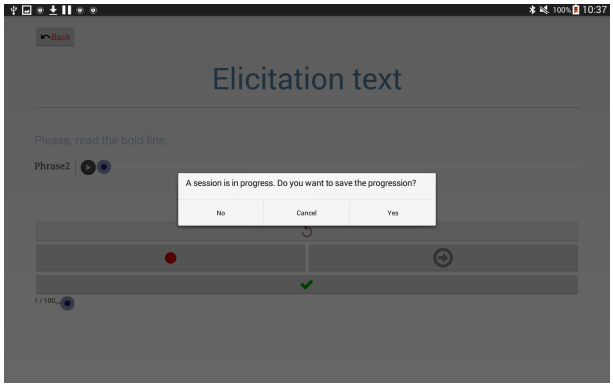
Lig-Aikuma géolocalise automatiquement l'appareil mobile sur lequel l'application est installée. Cette information est enregistrée dans le fichier de métadonnées généré lors de chaque session d'enregistrement. Cette fonctionnalité était implémentée à l'origine dans Aikuma, mais n'était plus fonctionnelle. La géolocalisation de l'appareil peut être obtenue de 3 façons. Classées selon leurs priorités, celles-ci sont :

- localisation à partir d'une application tierce qui a déjà récolté les coordonnées géographiques (longitude et latitude) de l'appareil ;
- localisation par triangulation des points d'accès Wi-Fi et des antennes relais situés à proximité ;
- localisation par réseau satellite (appelé *Géo-Positionnement par Satellite* ou plus couramment « GPS »).

Les priorités ont été établies selon la consommation de la batterie (de la moins à la plus gourmande en énergie). La géolocalisation dans Lig-Aikuma n'est disponible que si le terminal mobile est muni d'un récepteur GPS ou relié à un abonnement mobile.

3.2.4.6 Sauvegardes

Le temps étant précieux lorsque le linguiste est sur le terrain, trois types de sauvegardes ont été pensés pour optimiser les séances d'enregistrements.



(a) Message demandant à l'utilisateur s'il souhaite sauvegarder la session en cours.

(b) Boîte de dialogue s'affichant à la ré-ouverture de l'application, lorsqu'une session a été sauvegardée.

Graphique 3.14 – Sauvegarde et reprise de session dans Lig-Aikuma.

Sauvegarde des champs. Les saisies dans le formulaire de métadonnées sont enregistrées avant tout enregistrement audio de l'utilisateur. Ceci permet de faire gagner du temps au linguiste lorsqu'il travaille avec le même locuteur et veut enchaîner des modes différents. Ainsi, lorsque le linguiste a terminé sa tâche et veut passer à une autre, le formulaire est automatiquement pré-rempli avec les informations renseignées pour la tâche précédente.

Sauvegarde de session. Lorsque l'utilisateur doit mettre prématurément fin à sa tâche (par exemple, une élicitation), il peut à tout moment s'arrêter et quitter l'application. Il lui suffit de valider ou d'appuyer sur le bouton de retour arrière de son appareil afin qu'une boîte de dialogue (illustrée par le graphique 3.14a) lui propose de sauver la progression en cours. Plus tard, au moment de la ré-ouverture de Lig-Aikuma, un message indiquant à l'utilisateur qu'une session est en cours s'affiche. 3 choix s'offrent alors à lui : reprendre la session où il s'était arrêté, ne pas reprendre la session, mais sauvegarder le travail qui avait été effectué

ou ne pas reprendre la session et effacer ce qui avait précédemment été accompli. La boîte de dialogue proposant ces choix est illustrée par le graphique 3.14b. Cette fonctionnalité est également utile en cas de crash inopiné, de manque de batterie ou de fermeture involontaire de Lig-Aikuma.

Sauvegarde des profils. Toutes les informations saisies dans le formulaire de métadonnées — avant tout premier enregistrement audio — sont automatiquement enregistrées dans un profil utilisateur. Ainsi, lors d'une collecte sur le terrain, le linguiste qui travaille plusieurs fois avec un locuteur peut importer un profil déjà fourni auparavant. Le linguiste a également la possibilité de supprimer un profil.

3.2.4.7 Échantillons

Nous fournissons avec l'application des exemples de fichier afin que l'utilisateur puisse tester chaque mode de Lig-Aikuma. Cela facilite la prise en main de l'application. Ces fichiers sont accessibles dès l'installation de l'application, dans un dossier nommé *examples*.

3.2.5 Synthèse

Le tableau 3.1 récapitule les fonctionnalités majeures implémentées dans Aikuma et celles ajoutées ou améliorées dans Lig-Aikuma.

Toutes les fonctionnalités d'Aikuma ont été conservées, mais la majeure partie d'entre-elles a été enrichie ou adaptée selon nos besoins, comme les modes *Respeaking* et *Traduction*. En outre, le renforcement de l'autonomie a été une priorité dans le développement de Lig-Aikuma, tout comme sa robustesse. D'autre part, l'application bénéficie d'une meilleure identification des modes, grâce à une présentation claire de ceux-ci à l'ouverture de Lig-Aikuma.

Lig-Aikuma intègre deux nouveaux modes d'enregistrement nommés *Élicitation* et *Vérification* ainsi qu'un mode *Partage*. Le premier mode offre la possibilité au linguiste de faire parler son informateur sur un sujet particulier, sans influence aucune de sa part, mais avec l'aide d'un support. Ainsi, l'élicitation offre au linguiste la possibilité d'utiliser du texte, des images ou encore des vidéos comme stimuli pour faire parler le locuteur. Ce dernier peut, de cette façon, décrire une image ou une vidéo sur un thème tout en étant libre de s'exprimer de la façon qu'il souhaite. Ce type de recueil de corpus spontané permet au linguiste de capturer des phénomènes linguistiques recherchés (précis) tout en étant émis de manière naturelle. Le second mode, *Vérification*, permet au linguiste de corriger du contenu sous forme de texte. Par exemple, ce mode peut lui permettre de valider si des transcriptions correspondent au signal audio associé, très simplement et rapidement.

Le troisième mode, *Partage* permet d'envoyer grâce à une technologie sans fil (par mail ou Bluetooth par exemple) les fichiers audio enregistrés pendant une session.

Aussi, l'application offre désormais la possibilité de sauvegarder une session en cours, de visualiser l'enveloppe du signal lors d'un *respeaking* ou d'une traduction, de renseigner de

Tableau 3.1 – Comparatif des fonctionnalités d'Aikuma et de Lig-Aikuma.

Fonctionnalités	Aikuma	Lig-Aikuma
Enregistrement et documentation	✔	✔
Respeaking et traduction orale	✔	✔
Extras : Sync. et partage, géolocalisation, interface graphique sans texte	✔	✔
Identification des modes simples	✘	✔
Amélioration des modes *Respeaking* et traduction orale	✘	✔
Élicitation de parole	✘	✔
Vérification rapide de textes	✘	✔
Sauvegarde des sessions	✘	✔
Nouveau type de profil utilisateurs	✘	✔
Interface multilingue	✘	✔
Formulaire de métadonnées	✘	✔
Génération d'un formulaire de consentement	✘	✔
Échantillon de fichiers	✘	✔
Export des segments alignés vers Elan	✘	✔
Extras : Multi-résolution, différentiation de la nature des fichiers, convention de nommage des fichiers et visualisation de l'enveloppe du signal	✘	✔

manière plus détaillée un formulaire de métadonnées et possède une convention de nommage des fichiers propre et universelle. L'application géolocalise automatiquement l'appareil s'il est lié à un compte Google et gère maintenant toutes les tailles d'appareils mobiles, du petit écran de *smartphone* aux plus larges écrans de tablettes. De plus, chaque profil de locuteur est enregistré, faisant ainsi gagner du temps au linguiste s'il travaille de nouveau avec un locuteur avec lequel il a déjà travaillé. Lig-Aikuma possède une interface multilingue : la langue de l'application s'adapte à la langue du terminal, et est, pour l'instant, traduite en français, anglais et allemand. Nous avons également inclus des fichiers d'exemple dans l'application pour que l'utilisateur puisse tester, directement après installation, chacune de ses fonctionnalités et ainsi rendre la prise en main plus rapide et aisée. Enfin, Lig-Aikuma génère de façon automatique un fichier au format CSV qui peut être importé dans le logiciel d'annotation Elan. Ce fichier est créé à la fin du mode *Respeaking* ou *Traduction*, à partir des segments créés par l'utilisateur après qu'il a répété ou traduit une partie du fichier audio original.

Nous avons commencé à développer Lig-Aikuma au premier trimestre 2016. Nous sommes partis au Sénégal tester l'application en fin d'année 2016. Elle nous a permis de collecter diffé-

rents types de discours (parole spontanée, parole élicitée, parole lue) et plusieurs langues (wolof et variantes dialectales). Toutefois, avant cette date, nous avions déjà collecté un corpus de parole lue, durant l'année 2015, pour construire le premier système de RAP du wolof, en collaboration avec Voxygen.

Spécifications

Lig-Aikuma est le fruit de la collaboration entre linguistes et informaticiens. Au sein du projet BULB, de nombreuses interactions ont eu lieu lors des réunions avec les laboratoires de linguistique tels le LPP, le LLACAN et le laboratoire berlinois ZAS. De plus, les linguistes qui ont utilisé l'application ont toujours pris le temps de nous faire des retours sur leur utilisation lors des collectes sur le terrain, comme les bogues rencontrés, la praticité de l'outil, les défauts à corriger et les améliorations qui pouvaient être implémentées. Cette boucle vertueuse a donné lieu à une application ergonomique, pensée pour les linguistes et en partie par eux.
Enfin, le développement de Lig-Aikuma a également bénéficié des conseils et suggestions avisés de Sylvie Voisin du DDL, émis lors des réunions et des collectes de terrain effectuées ensemble.

3.3 Les campagnes de collecte

3.3.1 Première collecte au Sénégal

Lors de notre première campagne de collecte d'enregistrements de parole, nous n'avions pas encore développé Lig-Aikuma. La collecte s'est déroulée à Dakar, au Sénégal, en collaboration avec Voxygen MBOA présent sur place. 18 participants ont été sélectionnés, d'un niveau scolaire élevé, pour la tâche de lecture. En effet, le wolof est une langue avant tout parlée (très rarement lue). La langue ayant le statut de langue nationale (et pas officielle), elle n'est enseignée qu'en 2ème année universitaire, ce qui réduit drastiquement le nombre de locuteurs pouvant participer à la tâche demandée.

Tout d'abord, nous avons décidé de collecter de la parole lue, dans le but de construire le premier système de RAP à grand vocabulaire du wolof. Cependant, très peu de documents électroniques sont disponibles en wolof. Nous avons, donc, construit notre propre corpus textuel, afin de rassembler des éléments écrits qui seront lus par les locuteurs lors des sessions d'enregistrement.

Corpus textuel

Nous avons utilisé, dans un premier temps, un corpus de textes écrit en wolof (standard), collecté originellement par Nouguier Voisin (2002). Ce corpus rassemble 6 types de documents différents, dans des formats variés (PDF, MS Word/Excel, HTML) : des proverbes (Kobès et Abiven, 1922), des contes (Kesteloot et Dieng, 1989), des transcriptions (faites à la main) de débats à propos de guérisseurs, une chanson intitulée « Baay de Ouza » et deux dictionnaires : le « Dictionnaire wolof-français » écrit par Aram Fal, Rosine Santos, Jean-Léonce

Doneux (Fal, Santos et Doneux, 1990) — extrait de la base de donnée RefLex [11] élaborée par Segerer et Flavier (2013) — et le « Dictionnaire wolof-français et français-wolof » de Jean Léopold Diouf (Diouf, 2003).

Tous ces documents ont été extraits au format TXT afin d'être post-traités de façon automatique. Nous avons résolu les problèmes d'encodage dans les fichiers (par exemple, des symboles comme « 🍃 » ou « 🎁 » encodaient, respectivement, la lettre « l » et la lettre « k »), avons supprimé toute source non écrite en wolof (comme la suppression des numérotations de sections, les listes numérotées, les notes en français, les caractères spéciaux, etc.) et avons converti la casse en minuscule. Nous avons également normalisé l'orthographe de certains textes (comme les contes de Kesteloot et les proverbes de Kobes et Abiven) selon la convention officielle établie par le Centre de Linguistique Appliquée de Dakar (CLAD). Par exemple, l'accent circonflexe sur les voyelles signifiait l'allongement de la voyelle, autrement dit le redoublement de celle-ci selon la norme actuelle. Par exemple, le mot « *bât* » a été normalisé « *baat* » (c'est un nom qui signifie 1. « Voix ; mot ; parole ; phrase ; propos » ; 2. « Parole d'honneur » ; 3. « Cou ; encolure ; col »), le nom « *dôm* » a été réécrit « *doom* » (et signifie 1. « Fruit » ; 2. « Enfant » ; 3. « Clé ») ou encore le verbe « *jîtu* » a été ré-orthographié « *jiitu* » (signifiant « Précéder ; devancer »). Enfin, pour la construction des modèles de langue, nous avons aussi débarrassé les textes de toute marque de ponctuation.

En fin de compte, ce corpus textuel représente un ensemble de 20 162 syntagmes de 147 801 mots exploitables. À partir de cet ensemble, nous avons aléatoirement extrait un corpus de 6 000 syntagmes d'une longueur comprise entre 6 et 12 mots. Nous avons choisi cette taille de syntagmes pour deux raisons : la première, afin de réduire le coût cognitif que représente la tâche de lecture et ainsi éviter le plus possible d'erreurs de lecture au cours des sessions d'enregistrement ; la deuxième, pour pouvoir construire des corpus d'évaluation et de test avec des enregistrements audio relativement courts. Puis, à partir de cet ensemble homogène, nous avons créé 18 sous-corpus, composés de 1 000 syntagmes chacun, qui seront lus par nos participants lors de la séance d'enregistrement. Le détail de la composition de ce sous-corpus est présenté dans le tableau 3.2, accompagné par la proportion de chaque genre littéraire dans ce sous-corpus au graphique 3.15.

Les diagrammes du tableau C.1 montrent la répartition des genres littéraires contenus dans chacun de nos trois corpus qui serviront, par la suite, à l'entraînement et à l'évaluation de notre système de RAP.

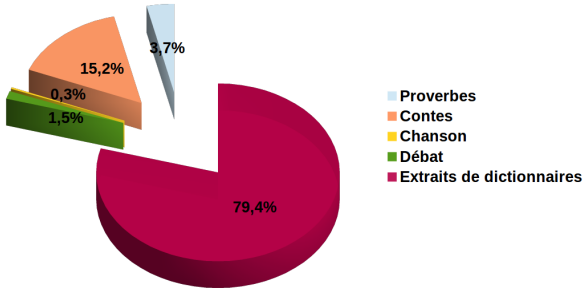
Corpus oral

Nous avons profité de la campagne d'enregistrement de voix de notre partenaire de projet Voxygen pour collecter notre corpus de lecture. 18 locuteurs de wolof (10 hommes et 8 femmes) de différentes catégories socio-professionnelles (journaliste, étudiant, gestionnaire, enseignant, télé-opérateur), âgés de 24 à 48 ans, ont été enregistrés. Chacun des locuteurs a lu les 1 000 énoncés qui lui ont été présentés et a été enregistré avec un microphone Samson

11. www.reflex.cnrs.fr

Tableau 3.2 – Répartition des genres littéraires dans le corpus textuel

	Nombre de lignes
Proverbes	657
Contes	2 739
Chanson	55
Débat	262
Extraits de dictionnaires	14 287
TOTAL	**18 000**



Graphique 3.15 – Proportion des genres littéraires dans le corpus textuel final.

G-track dans un environnement propre, dénué le plus possible de bruits environnants et de réverbération. Un questionnaire sociolinguistique a également été recueilli pour chaque locuteur.
Les 18 000 énoncés enregistrés représentent 21h23 de signal. Le corpus de parole collecté est détaillé en annexe D, dans le tableau D.1.

3.3.1.1 Problèmes rencontrés

Locuteurs

Les locuteurs sélectionnés par les partenaires MBOA ont tous remplis un questionnaire sociolinguistique. Nous nous sommes rendus compte après la campagne que pour quatre de ces locuteurs le wolof n'était pas leur langue maternelle. Des enregistrements comportaient ainsi des accents créoles ou encore des phrases prononcées différemment de leur graphie. De plus, un locuteur avait des difficultés d'articulation (sigmatisme interdental), ce qui a posé problème car, pour une tâche de lecture, il est important que les participants soient de bons lecteurs n'ayant aucun trouble du langage. Néanmoins, nous n'avons retiré aucun locuteur du corpus de parole : nous avons jugé que trois des locuteurs non natifs pouvaient faire partie du corpus d'entraînement afin que celui-ci possède de la variété et le locuteur le plus problématique (ayant un trouble d'articulation) a été placé dans notre corpus de test, ceci pour évaluer notre système sur ce type de parole atypique.

Lecture

Finalement, les tâches de lecture furent plus compliquées que nous l'avions prédit. La lecture est un travail difficile, même pour des lettrés. En effet, la langue officielle étant le français, les wolofones n'ont pas du tout l'habitude de lire du wolof. De plus, les langues nationales au Sénégal ne sont enseignées qu'en 2ème année de faculté et pour une durée d'un an seulement. Par conséquent, les étudiants qui choisissent le wolof n'apprennent à le lire que tard dans leur cursus. Mais aussi, certaines phrases soumises aux lecteurs étaient parfois trop littéraires ou écrites dans un vieux wolof (tels que les contes de Kesteloot et Dieng (1989) ou encore les proverbes de Kobès et Abiven (1922)), complexifiant d'autant plus la tâche de lecture.

De ce fait, de nombreux locuteurs ont eu des difficultés à lire le texte présenté et nous nous sommes aperçus, après la campagne, que de nombreux enregistrements ne correspondaient pas à la transcription proposée. Nous avons supposé que cela provenait d'un problème de structure syntaxique ou de vocabulaire non maîtrisé et inconnu : parfois, le locuteur ne prononçait pas du tout le mot à lire, parfois il était déstabilisé par le sens de la phrase et pouvait inventer un mot nouveau à substituer au mot écrit qu'il ne connaissait pas ou dont la signification lui était obscure, etc.

3.3.1.2 Quelques recommandations

Environnement

Lors d'une collecte audio, plusieurs recommandations sont à prendre en compte. En effet, au vu des technologies et des performances informatiques actuelles, il est fondamental de prendre soin de la qualité de l'enregistrement afin que les données collectées puissent être analysables par l'expert et par la machine dans des conditions optimales. De nombreux domaines ont émergé pour lesquels la qualité des fichiers audio est essentielle, comme la transcription (automatique) fine, l'analyse phonologique et phonétique (tons, traits distinctifs, etc.), l'analyse prosodique, le traitement automatique du langage, la synthèse et la reconnaissance vocale. Afin d'obtenir des enregistrements de qualité, plusieurs dispositions doivent être prises afin d'éviter des désagréments qui pourraient entraver la qualité d'un enregistrement audio.

En intérieur, afin de prévenir tout phénomène de réverbération et d'écho, il est préférable de ne pas se placer trop près d'un mur et, si la pièce est pourvue de peu de mobilier, de tapisser les murs avec un tissu lourd pour étouffer le son. De plus, une attention particulière est à prêter aux bruits domestiques. Pour cela, mettre le micro sur un pied peut être utile pour éviter les bruits de tables, de chocs, etc. ou utiliser un micro-cravate, pour stabiliser l'amplitude sonore (typiquement, en cas de mouvements de tête). Enfin, il est important d'être attentif aux bruits parasites comme les ventilations et climatiseurs, mais aussi de choisir une pièce à l'écart des lieux fréquentés afin de se prémunir des bruits extérieurs (discussions, bruits de pas, animaux, etc.).

Si l'enregistrement dans un intérieur isolé et à l'abri est impossible, les mêmes précautions que celles citées précédemment concernant les bruits ambiants en intérieur sont applicables, d'autant plus que les bruits d'animaux, de véhicules ou de passants ne sont pas atténués par des cloisons. Aussi, l'utilisation d'un micro-cravate est vivement conseillée afin d'être au plus proche de la source à enregistrer. Enfin, il faut penser à se protéger des conditions météorologiques telles que le vent, la pluie, l'ensoleillement et la poussière (si des prises ou visionnages de vidéos sont prévus).

Outre l'environnement, le dispositif d'enregistrement doit également être réglé avec soin. En effet, il convient de vérifier et régler les niveaux de volume afin d'éviter le phénomène de distorsion, de désactiver le gain automatique, mais aussi de faire attention au format d'enregistrement : le MP3 est à bannir — à cause de la compression et du masquage qui dégradent la qualité du signal — au profit d'un format PCM comme le WAV (réglé au minimum à une fréquence d'échantillonnage de 44.1kHz et une profondeur de 16bits.).

Exploitation et Interprétation du questionnaire sociolinguistique

L'enquête sociolingue a été pensée et réalisée, mais n'a pas été exploitée de manière optimale. Une attention toute particulière aurait dû être portée sur le questionnaire avant toute session d'enregistrement. Puis, un essai avec les locuteurs sélectionnés suivi d'un envoi d'échantillons aurait dû être présenté aux prestataires de la collecte avant de commencer. Enfin, les enregistrements manquaient de métadonnées documentant les sessions d'enregistrement (telles que des informations sur le lieu exact d'enregistrement, une description précise du matériel utilisé, des indications sur les conditions sonores, des précisions sur l'état du locuteur durant l'enregistrement, etc.). Ces métadonnées nous auraient permis de détecter plus rapidement les erreurs faites dans la collecte (accents, bruits environnants, hésitations, mauvaise articulation, etc.)

Ce qu'il ne faut pas reproduire

Pour la collecte, il est recommandé que l'opérateur connaisse et comprenne la langue dans laquelle va parler le locuteur interrogé. Pour une collecte de parole lue où ce qui incombe le plus est que l'enregistrement audio corresponde au texte lu, il faut absolument faire attention à ce que lit le locuteur et ne pas hésiter à le faire relire si nécessaire.
De plus, avant chaque début de session, les téléphones portables doivent être éteints afin d'éviter tout phénomène d'interférence avec le matériel audio.
Enfin, documenter chaque session d'enregistrement, *in situ* devrait être un réflexe : dire quelles sont les conditions d'enregistrements (bruit ambiant), informer sur l'endroit où sont effectués les enregistrements (intérieur/extérieur, pièce vide/meublée ou chambre sourde, taille de la pièce, bruits ambiants/environnants, type de microphone utilisé, etc.), communiquer sur l'état de santé du locuteur (maladie qui peut avoir une influence sur l'enregistrement telle que le rhume par exemple qui peut engendrer une nasalisation des voyelles orales).

Ces informations peuvent être un atout considérable pour l'analyste. *A fortiori*, dans notre cas d'usage, il en va de l'évaluation efficace du système de reconnaissance vocale.

3.3.2 Deuxième collecte au Sénégal

Cette seconde campagne de collecte a eu pour but principal de collecter de la parole spontanée en wolof ainsi que deux de ses variétés. Scientifiquement, l'intérêt de cette collecte était double : du point de vue informatique (traitement automatique et technologies vocales), collecter de la parole spontanée et des dialectes wolof permettait d'améliorer et de tester notre système de reconnaissance de la parole, mais aussi d'effectuer des analyses automatiques validant ou infirmant des hypothèses linguistiques (dont les expérimentations étaient plus ou moins anciennes ou effectuées manuellement, sur très peu données) ; du point de vue linguistique, cette collecte permettait une recherche ciblée sur l'expression de l'inaccompli et l'expression de la trajectoire [12]. De plus, la prise en compte des variantes dialectales faisait partie des points mentionnés dans le projet ALFFA. Enfin, cette collecte a permis de renforcer le lien

12. Même si l'intercompréhension est avérée entre les variétés de wolof, des différences — dont l'inaccompli fait partie — existent néanmoins.

entre traitement automatique et collecte linguistique, un des objectifs principaux du projet ALFFA.

Cette collecte a également été l'occasion de tester Lig-Aikuma dans des conditions réelles. Le mode *Élicitation* est une fonctionnalité intéressante pour le linguiste de terrain, car il permet de récolter plusieurs variétés/langues tout en proposant le même stimuli à chacun des locuteurs. L'application a été installée et utilisée sur un téléphone mobile HTC Desire 820 (5,5") et sur une tablette Samsung Galaxy Tab 4 (10"). Tous les modes de Lig-Aikuma ont été testés, excepté le sous-mode *Vérification de mots*.

- Le mode *Enregistrement* a été utilisé pour enregistrer de la parole spontanée ;
- Le mode *Respeaking* a permis de faire répéter des enregistrements bruités ;
- Le mode *Traduction* a servi pour la traduction du discours spontané ;
- Le mode *Élicitation* a été utilisé pour faire émerger des réalités linguistiques, à partir de textes, d'images et de vidéos.
- Le mode *Vérification de transcriptions* a permis de vérifier des transcriptions émises par notre système de RAP du wolof sur des signaux dont nous ne possédions pas de transcriptions (manuelles).

Concernant l'élicitation à partir d'images, nous avons utilisé comme support *L'histoire de la grenouille* (de son nom original "*Frog story*" (Mayer, 1969)), une histoire constituée d'une succession de 26 images, dans laquelle aucun mot n'est présent. Le but est de reconstruire et raconter l'histoire verbalement. À l'origine, cette histoire était utilisée pour étudier l'acquisition du langage chez les enfants. Plusieurs études ont été menées à partir de ce support, telles Cameron et Wang (1999) ou Reilly *et al.* (2004), afin de découvrir le développement des structures du discours et l'évolution des constructions syntaxiques. Nous avons utilisé ce protocole, car il est adapté aux langues non écrites, mais aussi aux personnes analphabètes. Il est ainsi accessible à toute personne voyante et permet d'obtenir de la parole spontanée sans aucun biais ni influence textuelle.

Un autre matériau très utilisé pour l'élicitation est le court-métrage *L'histoire de la poire* (le nom original étant "*The Pear story*"), créé par Chafe (1980). C'est un film muet (mais contenant des effets sonores), en couleur, d'une durée de 6 minutes. L'hypothèse de départ de l'auteur est qu'une grande partie de la connaissance humaine est emmagasinée de manière non verbale, en tant que représentations mentales. À travers l'élaboration de cette histoire, il a tenté d'explorer l'étendue des moyens d'expression verbale de ces représentations, mais également d'analyser la façon dont le discours narratif était produit dans différentes langues. Néanmoins, nous n'avons pas utilisé *L'histoire de la poire* car ce protocole est particulièrement marqué culturellement. Il a été tourné en Californie du Nord et est inadapté pour la population visée du Sénégal (locuteurs de villages reculés, poires rares, végétation non cohérente, etc.).

Concernant l'élicitation à partir de vidéos, nous avons utilisé comme matériau les clips élaborés dans le cadre du projet Trajectoire [13]. Ce projet a pour but de « décrire la variation

13. http://www.ddl.ish-lyon.cnrs.fr/trajectoire/ProjetGP.html

inter-langues dans l'expression de la trajectoire, en mettant l'accent sur la sémantique spatiale et ses différents moyens d'expression (morpho)syntaxiques. » (Grinevald, 2011). Ce corpus est composé de 76 séquences d'action, d'une durée variant de 5 à 11 secondes, dans lesquelles une ou plusieurs personnes se déplacent dans différents types d'environnement. 3 versions du corpus ont été créées, comportant les mêmes clips, mais apparaissant dans des ordres différents. De cette manière, tous les participants peuvent rester observer le déroulement d'une séance d'enregistrement sans être influencés par ce qui est dit par les locuteurs précédents. Bien que ces clips soient, tout comme *L'histoire de la poire*, culturellement marqués (présence de grotte, par exemple, alors qu'il n'y en a pas au Sénégal), nous en avons tiré parti, car ils ont été conçus dans le but d'orienter le discours en utilisant des verbes de mouvement, des expressions d'orientation dans l'espace, des constructions faisant appel à la formulation de la direction et du déplacement, etc., concepts que nous voulions faire émerger dans la langue.

Par précaution, nous avons également utilisé pendant les sessions d'enregistrement un dictaphone ZOOM Handy, réglé en mode stéréo, à une fréquence d'échantillonnage de 44.1kHz avec une précision de 16bits et enregistrant au format WAV. Nous avons interrogé 21 locuteurs. Au total, avec Lig-Aikuma nous avons collecté 2 heures et 27 minutes de parole alors qu'avec la méthode d'enregistrement traditionnelle — c'est-à-dire avec un enregistreur allumé en continu durant la totalité de la séance –, nous avons récolté 6 heures et 37 minutes de signal audio. La méthode traditionnelle contient par conséquent beaucoup de bruit, non seulement d'un point de vue acoustique, mais également parce que l'enregistrement comprend aussi des discussions pendant ou après la tâche de collecte, qui ne constituent pas directement des « données » (dans le sens « contenu pertinent pour l'analyse »).

Finalement, les données collectées dans ce cadre visaient à recueillir auprès d'informateurs parlant soit faana-faana, soit lébou un même type d'événement (vidéos *Trajectoire* dans lesquelles des personnes se déplacent), afin d'analyser les marques d'inaccompli dans ces différentes variétés, et les comparer au wolof standard. Dans le même temps, la collecte sur les vidéos de trajectoire permettait de compléter des données déjà recueillies sur ce point lors d'un terrain précédent. L'utilisation des vidéos avait l'avantage de permettre ce type d'enquête et de tester les outils développés (Lig-Aikuma).

3.3.2.1 Collecte de wolof

Durant cette campagne, nous avons enregistré 7 locuteurs wolofs (3 locuteurs de wolof *standard* et 4 locuteurs de wolof *urbain*). En terme de type d'énoncés collectés, nous avons enregistré du discours spontané, mais semi-guidé par élicitation d'images (3 minutes de parole) et de vidéos (25 minutes de parole) ainsi que du discours lu, enregistré à partir du mode *Élicitation de texte* (35 minutes de parole).

L'élicitation à partir de texte a ainsi été la plus productive en parole. Nous avons ainsi utilisé ce mode pour trois tâches différentes :

- La première tâche était de lire 12 phrases écrites en français et de les traduire en wolof (voir annexe F). Les phrases ont été créées afin de repérer les répartitions lexicales entre les différentes variétés dialectales du wolof pour un même sémantisme et tenter de trouver, dans au moins une variété, l'utilisation d'un lexème disparu dans l'utilisation du wolof standard. Deux locuteurs ont participé à cette tâche : un locuteur de wolof *standard* et un locuteur de wolof *urbain*. Cette tâche nous a permis de récolter environ 6 minutes de parole.

- La seconde tâche consistait également en la lecture de phrases (103 au total, jointes en annexe F) qui se focalisaient cette fois sur la production du /a/ en wolof standard et urbain. Pour rappel (le débat sur la voyelle /a/ est expliqué au chapitre 2), le postulat le plus largement accepté actuellement est que la voyelle /a/ peut se réaliser différemment selon sa graphie. Phonologiquement, les graphèmes « a » et « à » correspondent au même phonème, mais le dernier est réalisé plus allongé et plus ouvert que le premier. « aa » représente, quant à lui, un phonème à part entière qui peut être opposé aux deux premiers, avec une prononciation plus longue. En vue de produire nos propres analyses phonétiques, nous avons sélectionné des paires minimales s'opposant sur la graphie de la voyelle /a/ et avons construit des phrases contenant un mot de la triade « à, a, aa » à chaque fois que cela a été possible. Nous avons essayé de faire en sorte que la position des mots dans les phrases soit similaire pour que les comparaisons acoustiques soient le moins biaisées possible. Nous avons ensuite traduit ces phrases en français et les avons mélangées afin que les lecteurs ne découvrent pas l'objet de notre étude et ne soient, de ce fait, pas influencés par leur propre prononciation. Finalement, 5 locuteurs ont participé à cette tâche (3 locuteurs de wolof *standard* et 2 locuteurs de wolof *urbain*) qui nous ont permis de recueillir environ 25 minutes de parole avec cette tâche d'élicitation.

- La dernière tâche a été accomplie par un seul locuteur. C'était une tâche particulière, mise en place pour analyser la formation du collectif en wolof. 67 phrases (jointes en annexe F) ont été lues par un locuteur de langue wolof native et ceci a permis de récolter près de 4 minutes de parole.

Nous avons aussi fait de l'élicitation par images, à partir de *L'histoire de la grenouille ("Frog story")*, avec un locuteur. Cette tâche correspond à près de 3 minutes de parole. Par manque de temps, nous n'avons pas poussé cette tâche avec d'autres locuteurs durant la campagne et avons préféré faire en priorité d'autres tâches. Enfin, nous avons fait éliciter de la parole à partir des vidéos de trajectoire (pour rappel, le principe de cette tâche est évoqué en sous-section 3.3.2). Les 76 clips ont tous été vus et commentés par 3 locuteurs au total (de wolof *urbain*).

3.3.2.2 Variantes dialectales : faana-faana et lébou

Au cours de cette campagne, nous avons également enregistré des locuteurs de faana-faana et de lébou, langues reconnues comme variantes dialectales du wolof. En tout, 6 locuteurs de

faana-faana (tous originaires de la région du Saloum, dont notre informateur qui a reparlé et traduit en français des enregistrements libres) et 5 locuteurs de lébou (4 locuteurs du lébou de Yénn et 1 locuteur du lébou de Ouakam) ont été enregistrés. Durant les séances d'enregistrement, nous avons essayé à chaque fois de collecter de la parole à partir des vidéos de trajectoire et de la parole spontanée à partir d'enregistrement libre. Pour faire parler les locuteurs durant un temps assez long pendant les séances d'enregistrement libre, nous leur avons demandé de nous raconter par exemple des contes, des histoires sur leur village ou encore de nous expliquer des recettes de cuisine, comment construire une case, leur métier, etc.

41 minutes de parole ont été récoltées à partir de l'élicitation par vidéos en lébou et 24 minutes en faana-faana. 7 minutes de parole spontanée ont été collectées en lébou et 15 minutes en faana-faana. Nous avons également utilisé une fois le mode *Élicitation par images* en lébou (représentant 3 minutes de parole) et trois fois le mode *Élicitation par texte* en faana-faana (représentant 8 minutes de parole).

Pour résumer, 57 minutes de parole ont été collectées en lébou et 53 minutes en faana-faana.

En plus de cette collecte auprès des locuteurs, notre informateur a utilisé le mode *Traduction* de Lig-Aikuma lorsque nous étions à Paka Thiare Secco, en pays Faana-Faana. Il a traduit en français les enregistrements libres qui venaient d'être effectués par les locuteurs, ce qui représente 13 minutes de traduction. Il a également reparlé certains enregistrements qui avaient été récoltés dans un environnement bruyant, représentant ainsi 5 minutes de *respeaking*.

Le tableau 3.3 récapitule le corpus collecté durant ce terrain.

3.3.2.3 Difficulté de la collecte sur le terrain

Sur le terrain, dans des villages éloignés aux équipements sommaires, il est difficile de trouver un endroit bien isolé des bruits environnants. Les collectes se font en intérieur si possible, à l'extérieur sinon. Cependant, même dans un intérieur, de nombreuses nuisances sonores peuvent apparaître. Les maisons ne disposent pas de portes ou de fenêtres bien fermées, réduisant par conséquent fortement l'isolation acoustique ; les pièces sont spacieuses, dépourvues d'ameublement et sont des lieux de passage fréquents ; certaines toitures sont en zinc, créant des craquements sous l'effet de la chaleur. En extérieur, les bruits parasites notables, qui pouvaient gêner l'enregistrement, venaient d'animaux (poules, chèvres, vaches, ânes, coqs, chats, oiseaux, grillons), de véhicules (voitures, motos), des habitants (discussions, pleurs de nouveaux-nés, rires d'enfants) mais aussi des conditions météorologiques (vent, soleil).

Avec Lig-Aikuma, nous n'avons pas eu de problème particulier mis à part la réflexion du soleil sur l'écran de la tablette ou du téléphone qui a rendu le visionnage des vidéos délicat par moment, d'autant plus que certains locuteurs avaient des problèmes de vue dus à la vieillesse.

Tableau 3.3 – Récapitulatif des enregistrements de wolof, lébou et faana-faana, collectés sur le terrain avec Lig-Aikuma durant la seconde campagne de collecte au Sénégal (novembre-décembre 2016).

Langues	Wolof	Lébou	Faana-faana	TOTAL
Élicitation texte	35'10	*3'*†	8'	46'10
Élicitation image *(frog story)*	2'57	5'38		8'35
Élicitation vidéo *(trajectoire)*	25'07	41'48	24'39	1h31'34
Enregistrement		9'52	15'48	25'40
Respeaking			5'10	5'10
Traduction			13'25††	13'25
TOTAL	1h03'14	57'18	1h07'02	**3h07'34**
		1h00'18†		***3h10'34***

† Locutrice lébou, mais langue de l'enregistrement plus proche du wolof urbain (fortes influences).

†† Traduction en français du faana-faana : non compté dans la durée totale du temps de parole en faana-faana.

3.3.3 Autres données collectées avec Lig-Aikuma

Au sein du projet ALFFA, l'application mobile Lig-Aikuma a aussi servi à collecter des corpus oraux en fongbe et en amharique.
Laleye *et al.* (2016) ont utilisé le mode d'enregistrement libre de Lig-Aikuma pour enregistrer 28 locuteurs natifs de fongbe, au Bénin. Près de 10h de parole ont pu être collectées, permettant, par la suite, de construire un modèle acoustique de la langue, ceci dans le but d'apprendre un système de RAP du fongbe. Les autres modes de l'application n'ont pas été utilisés.
En amharique, langue parlée en Éthiopie, 8 000 phrases provenant du corpus BTEC ont été traduites oralement, représentant 7h30 d'enregistrements de parole (Melese, Besacier et Meshesha, 2016).

Au sein du projet BULB, Lig-Aikuma a également été utilisée par les linguistes de terrain. L'application mobile a ainsi servi à collecter, jusqu'à présent, près de 239h de parole au sein de ce projet, en bàsàá, mbochi, myènè. Le tableau 3.4 représente la quantité de données collectées jusqu'à maintenant avec Lig-Aikuma dans le projet BULB.

En bàsàá, environ 47 heures de parole spontanée et répétée en utilisant les modes *Enregistrement* et *Respeaking*, 34 heures de parole traduite oralement en français avec le mode *Traduction* et 8 heures, de parole élicitée à partir d'images, ont été enregistrées.

En mbochi, environ 33 heures de parole spontanée ont été enregistrées. La totalité de ce corpus d'enregistrements a été répété ("respeaking") et traduit oralement en français. De plus, 5h45 de parole contrôlée (lecture de tables de conjugaison) et 7h45 de parole lue ont aussi été enregistrées. Enfin, 1 500 images ont été capturées et commentées.

Tableau 3.4 – Enregistrements audio reparlés et traduits jusqu'à présent avec Lig-Aikuma dans le projet BULB (en dehors de notre propre collecte en wolof).

Langue	Enregistré	Reparlé	Traduit	Élicitation
Bàsàá	23h	24h	34h	8h
Mbochi	33h	30h	30h	14h
Myènè	9h	29h	5h	*non prévu*

La campagne de collecte en myènè a été la première où Lig-Aikuma a été utilisée en situation réelle. Dès lors, l'application était encore en développement et a souffert de quelques dysfonctionnements. Elle n'a, donc, pas été utilisée pour tous les enregistrements. Ainsi, seulement 9 heures de parole ont été enregistrées avec le mode *Enregistrement* de Lig-Aikuma. 36 heures supplémentaires de parole spontanée, de parole élicitée à partir d'images, de parole lue et de discours oratoires ont été collectées à partir d'un dictaphone (TASCAM DR100 et FOSTEX FR2). 29 heures de ces enregistrements ont été répétées avec Lig-Aikuma et 5 heures ont été traduites oralement en français jusqu'à présent.

En plus des collectes de parole, les linguistes ont tiré profit de l'usage du *smartphone* ou de la tablette. Spontanément, d'autres fonctionnalités du *smartphone* ou de la tablette ont été utilisées comme par exemple l'appareil photo, qui a permis de capturer des images et des vidéos des us et coutumes des peuples de la langue d'intérêt. À titre d'exemple, 1 500 images ont été collectées en mbochi et ont directement permis de faire de l'élicitation.

3.4 Résumé

Ce chapitre récapitule les collectes effectuées en vue de construire — rapidement — des systèmes de transcriptions, mais aussi le travail accompli pour développer une méthodologie de collecte de corpus oraux innovante.

Plusieurs collectes de corpus oraux ont été réalisées, de deux façons différentes : une manière classique (avec un enregistreur) et de manière originale (avec une application mobile).

L'ensemble des données collectées durant le projet ALFFA est disponible sur notre dépôt Github [14]. Ces corpus sont également disponibles sur la plateforme d'hébergement OpenSLR [15], initiative de l'Université John Hopkins (JHU). OpenSLR a pour but de rassembler de très nombreuses ressources dédiées à la reconnaissance vocale. Ces données sont libres de droit et d'accès.

La méthodologie adoptée et l'application mobile Lig-Aikuma développée sont le fruit d'une collaboration avec les partenaires du projet BULB, qui répond à des problématiques similaires à celles du projet ALFFA. Nous avons proposé, à travers Lig-Aikuma, un outil mobile équipant le linguiste de terrain, lors de ses missions auprès de populations parlant des langues en voie

14. https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR
15. http://www.openslr.org/25/

d'extinction et/ou peu, voire non, écrites.

Nous avons choisi de développer une application mobile, notamment, pour son faible encombrement : ce moyen permet d'accompagner le linguiste de terrain — constamment en déplacement — dans son travail de collecte, sans le surcharger. Nous avons aussi voulu un outil mobile d'une grande autonomie, simple d'utilisation et produisant des données naturellement exploitables (par la machine et par lui-même) à son retour de terrain. En effet, le *smartphone* est :

- un outil léger, peu ostentatoire et doté de batteries de plus en plus puissantes. Il en fait par conséquent l'outil adéquat pour le linguiste de terrain qui peut être mené à étudier des langues dans des environnements aux ressources très faibles.
- un outil de nos jours performant, qui peut permettre de collecter des données directement exploitables par les logiciels informatiques déjà existant (tels qu'Elan, FLEx, etc.). Toutes les gammes sont désormais dotées d'un appareil photo aux performances accrues, d'un microphone de qualité, mais aussi d'un écran et des haut-parleurs de mieux en mieux définis.

Lig-Aikuma dérive d'une application existante nommée Aikuma. Cette application Android, développée par Hanke et Bird (2013), permettait de collecter de la parole dans une langue en danger et/ou non écrite auprès de populations habitant dans des villages isolés. L'objectif principal d'Aikuma était de rendre pérenne les données audio collectées afin que, dans le futur, elles puissent encore être exploitables. Pour cela, l'application offrait à l'utilisateur la possibilité de partager ses enregistrements afin que d'autres personnes parlant la langue les commentent ou les traduisent. Bien que l'application initiale soit dotée de remarquables atouts, quelques fonctionnalités essentielles ont dû être implémentées afin de répondre aux besoins des linguistes de terrain.

Afin d'équiper de façon utile, pratique et avantageuse le linguiste de terrain, des fonctionnalités supplémentaires étaient, donc, nécessaires. Ainsi, nous avons repensé Aikuma et créé Lig-Aikuma.

Nous avons intégré à Lig-Aikuma un mode *Élicitation* à partir de 3 types de médias différents ainsi qu'un mode *Vérification*. De plus, un formulaire sociolinguistique détaillé est associé à chaque type d'enregistrement (libre, *respeaking*, traduction et élicitation) et un formulaire de consentement est généré pour signature avec tout nouveau participant.

Les modes *Respeaking/Traduction* ont largement été enrichis : désormais, il est possible de visualiser l'enveloppe du signal du fichier audio d'origine, mais également de marquer les segments qui ne contiennent pas de parole (pertinente). Ces deux améliorations majeures apportent confort d'utilisation et précision dans la découpage du fichier original, et lui font gagner du temps sur l'étiquetage de ses données.

Enfin, l'application enregistre automatiquement le profil du locuteur, rendant ainsi possible l'import ultérieur des informations renseignées une première fois dans le formulaire des

métadonnées. Lig-Aikuma génère automatiquement un fichier au format CSV qui peut être importé directement dans le logiciel d'annotation Elan. Ce fichier répertorie tous les segments créés avec l'application, lors de l'utilisation du mode *Respeaking* ou *Traduction*.

Lig-Aikuma s'adapte à la résolution du terminal sur lequel elle est lancée. L'application a été installée et utilisée avec succès sur différents *smartphones* (HTC Desire 820, Samsung Galaxy S3, Google Nexus 6 et Wiko Rainbow) et tablettes (Samsung Galaxy Tab 4, 7" et 10"). Lors de son installation, des échantillons de fichiers sont déployés pour pouvoir immédiatement la tester et comprendre les fichiers pris en charge. La documentation de l'application est disponible en version électronique (en français, anglais et allemand) ainsi qu'en version vidéo (en français, sous-titré en anglais) sur notre site web https://lig-aikuma.imag.fr/.

Lig-Aikuma a fait l'objet de plusieurs publications, en français et en anglais, dans des conférences aussi bien d'informatique que de linguistique. Ainsi, nous l'avons présentée lors de l'atelier francophone TALAF — organisé dans le cadre de la conférence JEP-TALN-RECITAL 2016 —, qui s'intéresse au traitement automatique des langues africaines (Blachon *et al.* 2016). Nous l'avons également présenté lors d'une session de démonstration à la conférence Interspeech 2016 (Gauthier, Blachon *et al.* 2016). Une publication est aussi actuellement en cours, dans les actes du colloque international organisé pour les 40 ans du LACITO (Gauthier, Besacier et Voisin, 2018), le laboratoire des langues et civilisations à tradition orale. Conjointement avec les partenaires du projet BULB, nous avons communiqué au sujet de Lig-Aikuma et de la méthodologie adoptée dans le projet, durant l'atelier CCURL 2016, qui promeut la collaboration et l'informatique pour les langues à faibles ressources (Adda, Adda-Decker *et al.* 2016), ainsi qu'à l'atelier SLTU 2016, dédié aux technologies de la parole pour les langues peu dotées (Blachon *et al.* 2016).

Plusieurs campagnes de collecte ont été menées avec Lig-Aikuma, au sein du projet BULB et du projet ALFFA. Jusqu'à présent, presque 260h de données audio ont été récoltées via l'application, dans 8 langues africaines et variétés différentes.
Dans le cadre du projet ALFFA, 18 locuteurs différents ont été enregistrés (6 locuteurs de faana-faana, 5 locuteurs de lébou, 7 locuteurs de wolof), permettant ainsi de collecter 3 heures de parole lue et spontanée avec Lig-Aikuma. Grâce à ces corpus, des améliorations futures de Lig-Aikuma ont émergé. Des méthodes innovantes de traitement automatique ont pu être appliquées comme la découverte non supervisée de mots (Godard *et al.* 2016), la segmentation non supervisée en phonèmes (Vetter *et al.* 2016), la détection de frontières de phonèmes (Franke *et al.* 2016), l'analyse phonologique (Rialland *et al.* 2015 ; Cooper-Leavitt *et al.* 2017b) ou encore l'analyse phonétique fine (Rialland *et al.* 2016 ; Gauthier, Besacier et Voisin, 2017).

Initialement conçu à des fins de documentation des langues, Lig-Aikuma aura aussi permis de créer des corpus de données très facilement, selon un même protocole, qui ont donné

lieu à l'évaluation d'outils (Rialland, Adda-Decker *et al.* 2018).

Finalement, le développement de Lig-Aikuma aura permis :

1. d'effectuer des collectes massives
2. d'accélérer la transition d'un corpus brut de parole à un jeu de données utilisable par des systèmes de traitement automatique
3. d'exploiter rapidement des données pour développer des outils technologiques
4. de partager simplement les collectes à la communauté

CHAPITRE 4

Transcrire : Automatisation à l'aide de la reconnaissance automatique de la parole

Dans ce chapitre, nous allons présenter, dans un premier temps, les principes de base d'un système de reconnaissance automatique de la parole (abrégé « RAP ») ainsi que l'enjeu que représente le développement de ce genre de système pour les langues peu dotées. Ensuite, nous présenterons les systèmes de RAP que nous avons construit pour le haoussa et le wolof, ainsi que les expérimentations mises en place pour faire face au problème du manque de ressources numériques pour la construction des systèmes. Enfin, nous indiquerons les outils implémentés pour la reproductibilité de nos expériences.

Les travaux présentés dans ce chapitre ont été publiés dans les actes de la conférence internationale *LREC* ("Conference on Language Resources and Evaluation") (Gauthier, Besacier, Voisin *et al.* 2016) et ont fait l'objet d'une présentation orale à la conférence internationale *Interspeech* (Gauthier, Besacier et Voisin, 2016b).

4.1 Construction des systèmes de RAP pour le haoussa et le wolof

4.1.1 Ressources utilisées

4.1.1.1 Le corpus Globalphone pour le haoussa

Afin de construire un système de RAP du haoussa, nous avons acheté les ressources nécessaires auprès d'ELRA (acronyme de "European Language Resources Association"). Ce corpus — récupérable sous l'identifiant ELRA-S0347 [1] —, est composé d'un corpus audio (8 heures et 44 minutes au total), d'un modèle de langue (41 435 mots) et d'un dictionnaire de prononciation contenant des variantes (42 659 entrées) [2]. Ces données ont été collectées et construites par Schlippe *et al.* (2012) dans le cadre du projet GlobalPhone. Ce projet — qui pour but la construction de systèmes de reconnaissance de la parole continue à grand vocabulaire —, rassemble une collection de ressources multilingues (jusqu'à présent, 20 des langues les plus répandues dans le monde).

1. http://catalog.elra.info/product_info.php?products_id=1177
2. http://catalog.elra.info/product_info.php?products_id=1203

Le corpus textuel. Schlippe *et al.* ont récupéré la version hors ligne de cinq journaux publiés sur le Web écrits en boko (le système d'écriture officiel fondé sur l'orthographe latine). Pour obtenir un corpus exploitable pour la RAP, le contenu textuel extrait a été nettoyé (balises HTML, caractères spéciaux, lignes vides et doublons, contenu autre qu'en langue haoussa). Cette collection a ensuite été utilisée comme matériau de lecture pour l'enregistrement du corpus audio ainsi que pour la création du modèle de langue.

Le corpus de parole. Le corpus audio a été collecté dans 5 régions du Cameroun (Maroua, Douala, Yaoundé, Bafoussam, Ngaoundéré et Nigeria) et contient, donc, différents accents. Il est composé de 102 locuteurs de langue maternelle haoussa, âgés de 16 à 60 ans qui ont lu au total 7 895 phrases. Les enregistrements ont été réalisés avec un micro-casque Sennheiser 440-6, dans différents environnements et contiennent quelquefois du bruit. Les données ont été échantillonnées à 16kHz, en 16-bit et encodées au format PCM. Nous les avons converties au format WAV pour le traitement, par la suite, par la boîte à outils Kaldi.

Tableau 4.1 – Présentation du corpus de parole lue en haoussa.

Corpus	#Homme	#Femme	#Phrase	#Mot	Durée
Apprentissage *(train)*	24	58	5 863	39 566	6 h 36 min
Évaluation *(dev)*	4	6	1 021	6 293	1 h 02 min
Test *(test)*	5	5	1 011	6 198	1 h 06 min
Total	**33**	**69**	**7 895**	**52k**	**8 h 44 min**

Ce corpus de parole a été divisé en trois sous-corpus qui serviront à l'apprentissage du système de RAP et en l'évaluation de sa performance. Le tableau 4.1 présente ces trois sous-corpus. Nous avons gardé la même subdivision du corpus audio que celle adoptée par Schlippe *et al.* (2012).

4.1.1.2 Le corpus collecté pour le wolof

Nous avons utilisé le corpus textuel (rassemblant environ 148k mots) ainsi que le corpus de parole récolté au Sénégal (représentant au total 21 h 21 minutes de signal audio). Les signaux ont été convertis en mono-canal et échantillonnés à 16kHz, 16bits. Ces corpus ont été évoqués au chapitre précédent, en sous-section 3.3.1. Le tableau 4.2 présente la répartition du corpus de parole lue que nous avons adoptée pour construire les corpus d'apprentissage, d'évaluation et de test de nos systèmes de RAP.

Nous avons sélectionné 14 locuteurs pour le corpus d'apprentissage, 2 locuteurs pour le corpus d'évaluation et 2 locuteurs pour le corpus de test. Nous avons vérifié que les 3 corpus contenaient une proportion de genres littéraires équivalente (illustration au tableau C.1). Plus précisément, les extraits de dictionnaires constituent environ 80% de chaque corpus, les contes représentent environ 15%, les proverbes environ 3%, les débats 1,5% et finalement les paroles de la chanson représentent 0,3%.

Tableau 4.2 – Présentation du corpus de parole lue en wolof.

Corpus	#Homme	#Femme	#Phrase	#Mot	Durée
Apprentissage *(train)*	8	6	13 998	132 963	16 h 49 min
Évaluation *(dev)*	1	1	2 000	18 790	2 h 12 min
Test *(test)*	1	1	2 000	18 843	2 h 20 min
Total	**10**	**8**	**17 998**	**171k**	**21 h 21 min**

Recueil de données textuelles sur le Web

Notre corpus textuel initialement construit est constitué de 147 801 mots. Pour entraîner un modèle de langue stochastique, cet ensemble représente très peu de données. Nous avons alors décidé de recueillir plus de données textuelles écrites en wolof, en parcourant le Web. Très peu de documents écrits en wolof sont disponibles en ligne et il est difficile de trouver des données correctement structurées (conformes aux règles syntaxiques). Finalement, nous avons trouvé des fichiers électroniques (au format PDF) de sites éducatifs ainsi que des textes religieux. Ainsi, nous avons extrait — au format TXT pour qu'il soit facilement exploitable par la machine —, des contenus de la Déclaration Universelle des Droits de l'Homme, de la Bible et d'un livre écrit par un humaniste. En terme de post-traitement, nous avons supprimé les symboles, les signes de ponctuation et le texte non significatif (comme la numérotation des sections, les listes numérotées, etc.) des textes collectés. Enfin, nous avons converti le texte en minuscules, ceci dans le but de ne pas faire de distinction sur la casse. Au total, nous avons obtenu 197 430 mots supplémentaires.

Étant donné les données limitées trouvées manuellement, nous avons décidé d'explorer la base de données de Wikipédia pour collecter une plus grande quantité de données en wolof. Nous avons récupéré tous les articles indexés en wolof en utilisant l'outil *Wikipedia Extractor* (Attardi et Fuschetto, 2013). Comme ce type de base de données ouverte n'est que légèrement supervisé, certains articles peuvent être multilingues. Pour supprimer le texte non écrit en wolof, nous avons appliqué l'outil de détection de langue *Google Compact Language Detector (CLD2)*[3]. Comme CLD2 ne peut pas reconnaître le wolof, mais peut détecter les langues les plus utilisées, nous avons utilisé l'outil pour filtrer les langues détectées comme « non wolof » et ainsi avons supposé que les documents restants étaient effectivement écrits en wolof. Enfin, pour améliorer la précision de la récupération de texte en wolof, nous avons appliqué l'outil de sélection de données *Xenc* (Rousseau, 2013). Après ces deux passes de filtrage, nous avons nettoyé le contenu en le convertissant en petite casse, en éliminant les balises HTML/XML et toute marque de ponctuation. Finalement, nous avons recueilli environ 311k mots depuis la base de données Wikipédia.

Le tableau 4.3 résume les données textuelles finalement acquises depuis le Web.

3. https://github.com/CLD2Owners/cld2

Tableau 4.3 – Données textuelles supplémentaires écrites en wolof, récupérées en ligne.

Texte	**#Phrase**	**#Mot**
Déclaration Universelle des Droits de l'Homme	112	1 923
Le Message de Silo	602	10 443
La Bible	14 474	185 064
Wikipedia	10 738	311 995
Total	**25 926**	**509 425**

4.1.2 Systèmes de référence pour le haoussa et le wolof

4.1.2.1 Dictionnaires de prononciation

Comme évoqué dans la première section de ce chapitre, les étapes d'apprentissage et de décodage d'un système de RAP nécessitent un dictionnaire de prononciation.

Le dictionnaire de prononciation créé par Schlippe *et al.* (2012) pour le haoussa, contient 42 660 entrées dont 583 variantes de prononciation et une entrée *SIL* qui modélise l'unité de silence. Nous avons, en plus, ajouté l'entrée *<UNK>* afin de correspondre au format demandé par Kaldi. Ce lexique répertorie 33 phonèmes [4] (26 consonnes, 5 voyelles et 2 diphtongues). De plus les voyelles possèdent des marques de ton ou de longueur dans la transcription phonétique [5]. Dans le but d'élaborer un premier système, qui fera office de référence pour nos futures expérimentations, nous avons décidé d'éliminer ces étiquettes de notre lexique de prononciation. Notre premier système de RAP pour le haoussa — pour lequel nous rapportons les performances dans le tableau 4.7 —, a donc été entraîné avec un dictionnaire de prononciation ne tenant pas compte de ces caractéristiques phonologiques.

Pour le wolof, nous avons construit notre propre dictionnaire de prononciation, en utilisant, comme point de départ un ensemble de 8 724 entrées, qui provenait de la concaténation des transcriptions phonétiques des mots fournis par les dictionnaires de Fal, Santos et Doneux (1990) et de Diouf (2003). Ces dictionnaires utilisent les symboles de l'alphabet phonétique international (API) pour transcrire les phonèmes du wolof. Au lieu d'utiliser ces symboles, nous avons fait le choix d'opter pour le jeu de caractères phonétiques SAMPA (ou "Speech Assessment Methods Phonetic Alphabet") (Wells, 1997) pour transcrire les phonèmes du wolof dans le dictionnaire de prononciation. Ce jeu de caractères phonétiques représente les symboles de l'API en utilisant la norme de codage de caractères ASCII sur 7 bits. L'atout du SAMPA est, donc, sa robustesse aux problèmes d'encodage puisqu'il est lisible par tout type de machine, contrairement aux caractères de l'API qui nécessitent d'être encodés en UTF-8. Pour une correspondance entre les symboles SAMPA et les symboles de l'IPA, le site web du Laboratoire de Phonétique Expérimentale « Arturo Genre » de l'Université de Turin propose

4. Schlippe *et al.* considèrent le phonème /p/ comme existant en haoussa, contrairement à notre description de l'inventaire phonologique de la langue, au chapitre 2.

5. Pour rappel, la sémantique du mot diffère selon le ton et la longueur vocalique porté par la voyelle.

des tableaux interactifs [6]. À partir de ces 8 724 entrées, nous avons entraîné un modèle de prononciation 7-grammes pour le wolof, en utilisant Phonetisaurus (Novak, 2011), un système de conversion graphème-à-phonème (G2P) [7]. Cet outil nous a permis de transcrire automatiquement, sous forme de symboles phonétiques, le vocabulaire encore non phonétisé de notre corpus textuel. En effet les entrées des dictionnaires, phonétiquement transcrites, ne couvraient pas la totalité de notre corpus textuel. Finalement, le dictionnaire de prononciation associé au modèle de langue `LM_initial` est composé de 15 575 entrées (dont 302 variantes). Grâce à l'augmentation de notre corpus en récupérant des ressources écrites sur le Web, nous avons pu accroître la taille de notre lexique de prononciation, qui comporte désormais 32 039 entrées.

4.1.2.2 Apprentissage des modèles acoustiques

Pour construire nos systèmes de RAP du haoussa et du wolof, nous avons utilisé la boîte à outils pour la reconnaissance automatique de la parole Kaldi (Povey, Ghoshal *et al.* 2011).

Nous avons utilisé la procédure standard fournie par Kaldi pour entraîner nos systèmes de RAP du haoussa et du wolof.

Pour chaque langue, nous avons construit deux systèmes qui ont été entraînés selon différentes techniques de modélisation acoustique. Le premier système utilise un modèle probabiliste, dépendant du contexte, entraîné à partir de chaînes de Markov cachées et de mélange gaussien (nous l'appellerons CD-HMM/SGMM) ; le deuxième système est fondé sur l'apprentissage d'un modèle acoustique entraîné à partir d'un réseau de neurones profond (nous l'appellerons CD-HMM/DNN).

Le modèle CD-HMM/SGMM. Pour les deux systèmes, des vecteurs de 13 coefficients MFCC sont extraits toutes les 10ms sur les données d'entraînement (6h36min pour le haoussa, 16h49min pour le wolof). Nous avons construit des modèles à base de triphones en employant 2 887 états dépendant du contexte pour le haoussa et 3 401 états dépendant du contexte pour le wolof, ainsi que 40k gaussiennes pour les deux langues. De plus, nous avons appliqué les coefficients delta-delta sur les MFCC, puis une estimation LDA ("Linear Discriminant Analysis") avec une transformation MLLT ("Maximum Likelihood Linear Transform") (Gopinath, 1998), une approche SAT ("Speaker Adaptative Training") (McDonough, Schaaf et Waibel, 2002) avec transformation fMLLR ("Maximum Likelihood Linear Regression") (Gales, 1998) pour l'adaptation du locuteur et enfin une estimation MMI ("Maximum Mutual Information") (Chow, 1990) et fMMI ("feature-space Maximum Mutual Information") (Povey, Kanevsky *et al.* 2008) qui permettent de créer des modèles plus robustes à la variabilité inter-locuteurs. Ces modèles ont, par la suite, été affinés en employant des SGMM (Povey, Burget *et al.* 2011) — dont les gains face aux GMM sont significatifs —, et une estimation MMI.

6. http://www.lfsag.unito.it/ipa/index_fr.html

7. Nous considérons le phonème plutôt que le graphème puisque notre objectif est, à terme, l'analyse de la réalisation phonétique en haoussa et en wolof. Comme nous ne nous concentrons pas seulement sur la performance des systèmes de RAP, mais que nous voulons aussi mesurer la durée des voyelles et son impact sur le système de RAP, une modélisation graphémique n'a pas été prise en compte dans ces travaux.

Le modèle CD-HMM/DNN. Kaldi propose deux types d'entraînement pour la construction de réseaux neuronaux. Nous avons choisi d'utiliser la méthode dite « de Karel » (Vesely *et al.* 2013) qui implémente l'utilisation de *RBM*. Pour les deux langues, nous avons construit un réseau à 7 couches, adapté avec les critères sMBR (*state-level Minimum Bayes Risk*) (Kingsbury, 2009). La couche d'entrée et chacune des cinq couches cachées sont composées de 1 024 unités cachées. Le réseau a été entraîné à partir de 11 trames consécutives (la trame courante, les 5 trames précédentes et les 5 trames suivantes) des mêmes MFCC que dans le système à base de HMM/GMM. En outre, les mêmes états HMM ont été utilisés comme cibles du réseau de neurones profond. Les poids initiaux pour le réseau ont été obtenus en utilisant des RBM, tandis que la méthode de descente de gradient stochastique ("Stochastic Gradient Descent") a été utilisée pendant la rétro-propagation pour les optimiser. Pour entraîner nos systèmes à base de réseaux de neurones, nous avons utilisé un serveur équipé de processeurs graphiques (communément abrégé *GPU* pour "Graphics Processing Units") Nvidia et de la boîte à outils CUDA (délivrée par Nvidia) pour accélérer les calculs.

4.1.2.3 Modélisation linguistique

En ce qui concerne le modèle de langue (ML), nous avons utilisé un modèle de trigrammes, dans le processus de décodage, pour chaque système de RAP du haoussa et du wolof.

Pour le haoussa, nous avons utilisé le modèle de langue fournit dans le corpus Globalphone. C'est un modèle composé de 81 528 trigrammes au total (pour 41 435 mots uniques), que nous avons converti en petite casse et transformé au format ARPA. La perplexité du modèle de langue du haoussa, calculée sur le corpus *dev* et sur le corpus *test* (mentionnés au tableau 4.1) est, respectivement, de 88 (0.19% de mots inconnus) et de 90 (0.21% de mots inconnus).

Pour le wolof, nous avons construit deux modèles de langue, avec la boîte à outils SRILM (Stolcke *et al.* 2002). SRILM est développé par le laboratoire STAR ("The Speech Technology and Research Laboratory") du SRI International. Il permet la création et l'utilisation de modèles de langue à base de n-grammes pour représenter et traiter des textes. Le premier modèle est constitué de 7 243 trigrammes. Il a été entraîné sur 106 206 mots (11 065 mots uniques). Ces données d'apprentissage ont été générées à partir de la collection de textes dont nous disposions initialement (les quelques documents électroniques, mentionnés au chapitre précédent, formant 20 162 occurrences représentées par 147 801 mots). Nous avons retiré, de cette collection, les textes lus par les locuteurs pour les enregistrements vocaux (41 595 mots supprimés). Sa perplexité, mesurée sur le corpus *dev* (mentionné au tableau 4.2) est de 294 (7.1% de mots hors vocabulaire) tandis qu'elle est de 301 sur le corpus *test* (7,2% de mots hors vocabulaire). Nous appellerons ce modèle `LM_initial` par la suite. Pour améliorer la qualité du ML, nous avons entraîné un modèle à base de réseaux de neurones récurrents en utilisant la boîte à outils RNNLM (Mikolov, Kombrink *et al.* 2011), que nous avons combiné avec notre modèle n-gramme. Cette interpolation nous a permis d'atteindre un score de perplexité de 262 sur le corpus *dev* et un score de 267 sur le corpus de *test*.

Le deuxième modèle de langue trigramme, entraîné avec SRILM, est une interpolation entre le modèle `LM_initial` et un second construit à partir des données récupérées sur le Web et nettoyées (509k, présentées précédemment dans le tableau 4.3). Ce modèle interpolé, que nous appellerons `LM_initialPlusWeb` a finalement été appris à partir de 601 609 mots (29 148 mots uniques) et est constitué de 83 971 trigrammes. Sa perplexité est de 314 (5,4% de vocabulaire inconnu) sur le corpus *dev* et de 323 (5,1% de vocabulaire inconnu) sur le corpus *test*. Nous avons aussi construit un modèle de langue à base de RNN et l'avons interpolé avec `LM_initialPlusWeb`. L'interpolation du modèle probabiliste avec un modèle à base de réseau de neurones récurrent nous a permis de faire baisser la perplexité à 292 sur le corpus *dev* et à 297 sur le corpus *test*.

Le tableau 4.4 récapitule les modèles de langue construits à ce stade.

Tableau 4.4 – Présentation des deux modèles de langue construits pour le wolof.

Modèle de langue	**#Mots du corpus textuel**	**Perplexité**		**Mots hors vocabulaire (%)**	
		dev	test	dev	test
`LM_initial`	~106k	294	301	7,1	7,2
`LM_initial+RNN`	~106k	262	267	7,1	7,2
`LM_initialPlusWeb`	~600k	314	323	5,4	5,1
`LM_initialPlusWeb+RNN`	~600k	292	297	5,4	5,1

4.1.2.4 Résultats

Pour évaluer nos systèmes, nous avons utilisé la boîte à outils SCTK (acronyme de "SCoring ToolKit"). Cette boîte à outils est développée au NIST (acronyme de "National Institute of Standards and Technologies") et fournit les programmes les plus utilisés pour évaluer et analyser les performances des système de RAP. Kaldi intègre d'ailleurs SCTK et mesure le WER des systèmes grâce au programme *sclite*. *sclite*[8] produit un alignement entre la référence et l'hypothèse d'un enregistrement audio, calcule le WER et procure également des informations comme le WER par locuteur.

Pour le haoussa. Nous avons entraîné deux systèmes de RAP : un système avec modélisation HMM et SGMM (CD-HMM/SGMM) et un système hybride avec modélisation par réseaux de neurones (CD-HMM/DNN), tous deux utilisant le modèle de langue et les données précédemment cités. Nous avons obtenu un WER de 9,5% sur le corpus de développement (appelé *dev*, détaillé dans le tableau 4.1) avec un système à base de SGMM. La reconnaissance est améliorée avec des DNN, avec un score de 8,0%. En décodant sur le corpus de test (appelé *test*, détaillé dans le tableau 4.1), nous avons obtenu un WER de 19,3% contre 11,3% avec des DNN.

8. http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Pour le wolof. Tout d'abord, nous avons voulu mesurer l'impact des deux modèles de langue du wolof `LM_initial` et `LM_initialPlusWeb`, afin de décider plus tard quel modèle nous utiliserions pour modéliser les longueurs de voyelles. Le tableau 4.5 reporte la performance du premier système de RAP du wolof, entraîné à partir des approches HMM/SGMM et DNN. Nous observons que notre premier système de RAP du wolof peut atteindre un score WER de 26%.

Tableau 4.5 – Performance des systèmes de RAP du wolof, utilisant différents modèles acoustiques et modèles de langue.

Modèle acoustique	**WER (%)**			
	LM_initial (~106k)		**LM_initialPlusWeb (~600k)**	
	dev	test	dev	test
CD-HMM/SGMM	30,4	35,2	28,6	33,6
sans diacritique	*28,4*	*33,5*	*26,6*	*31,7*
CD-HMM/DNN	28,1	34,0	**26,0**	32,0
sans diacritique	*26,0*	*32,1*	*23,9*	*30,1*

Nous supposons que les performances de `LM_initialPlusWeb` ne sont pas aussi élevées que nous l'espérions du fait du modèle de langue qui a été augmenté à partir de données provenant majoritairement du Web, impliquant un vocabulaire et une syntaxe très différents de ceux du corpus de parole (similaire à `LM_initial`). Nous observons néanmoins une légère amélioration de la performance avec `LM_initialPlusWeb` par rapport à `LM_initial`. De plus, nous avons analysé les résultats des systèmes de RAP et constaté que de nombreuses erreurs sont dues à des problèmes de normalisation orthographique dans le texte. Le wolof est une langue morphologiquement complexe et certaines erreurs peuvent apparaître à ce stade. De plus, comme nous l'avons dit au chapitre 2, le wolof n'a pas une orthographe réellement fixée ; un même mot peut avoir plusieurs formes de surface, surtout s'il contient des signes diacritiques. Par exemple, le mot « jél » peut aussi être écrit « jël » et signifie « pris » ou « voler ». En outre, le mot « r**a**ndal » peut être écrit « r**à**ndal » et signifier la même chose : « garder loin », mais peut aussi être orthographié « **d**andal ». Par ailleurs, à propos de ce dernier, on remarque que la variabilité dans l'orthographe peut également se manifester au sein des consonnes. Comme autre exemple de ce type, nous pouvons citer le mot « **c**éetal » qui peut s'écrire « **s**étal » et signifie « organiser le mariage ». Ce constat nous a poussé à évaluer le taux d'erreur de mots en éliminant les signes diacritiques, à la fois des hypothèses et des références. Ces résultats apparaissent dans le tableau 4.5. La différence entre les deux scores (WER avec ou sans diacritique) correspond à des erreurs sur les diacritiques. Nous remarquons que le score WER est légèrement inférieur lorsque nous ne tenons pas compte des signes diacritiques.

Nous pouvons constater que la performance de notre système de RAP du wolof est faible, que celui-ci soit appris à partir d'un modèle acoustique hybride à base de réseaux de neurones ou d'un modèle à base de HMM et de SGMM. Cette faible performance peut s'expliquer par le score de perplexité élevé du modèle de langue et par le fait que nous n'avons que deux locu-

teurs dans chacun de nos corpus d'évaluation et de test (alors que les corpus d'évaluation du haoussa sont composés de 10 locuteurs). En effet, nous observons jusqu'à 9 points de différence entre les scores WER des locuteurs wolof. Les scores par locuteur sont joints en annexe G et en annexe H. De plus, comparé au haoussa où nous avions 0,19% de mots inconnus dans le modèle de langue, pour le wolof nous en avons 5,4%. Enfin, comme expliqué précédemment, la normalisation dans l'écriture des mots est un problème qui pénalise à la fois le modèle de langue et le score WER du système. C'est d'ailleurs pour cette raison que nous afficherons également, par la suite, le taux d'erreur de caractères (CER) dans nos tables de résultats, car le CER est moins sensible aux problèmes de normalisation orthographique.

Malgré les explications avancées précédemment, ces scores WER nous ont paru élevés. En analysant de plus près les scores, nous avons notamment été étonnés par le taux élevé de substitutions (20.9% sur le *dev* et 25.6% sur le *test*). De plus, les différences de score inter-locuteurs, dans chacun de ces deux corpus d'évaluation, nous a fait nous interroger sur la qualité des enregistrements collectés à Dakar. Nous nous sommes demandés si les locuteurs enregistrés ont prononcé correctement les phrases qui leur ont été soumises. En effet, le fait que le wolof ne soit pas appris à l'école peut poser problème dans une tâche de lecture, qui est une activité peu courante en wolof, puisque cette langue est à tradition orale. Les wolofones sont bien moins entraînés et habitués que nous à lire. Il était, donc, possible de posséder, dans nos corpus, des phrases lues différemment de leur transcription originale. Pour vérifier cette hypothèse, nous avons extrait des sorties du système de RAP les transcriptions erronées et les avons soumises à l'expertise d'un linguiste et locuteur natif de wolof.

Bien que les transcriptions imparfaites soient un problème bien connu pour l'entraînement d'un système de RAP, nous avons, tout d'abord, décidé de nous concentrer sur le nettoyage des transcriptions utilisées pour l'évaluation.

Nettoyage du corpus de parole initialement collecté en wolof

Pour cette tâche de contrôle, nous avons utilisé le mode *Vérification* de notre application mobile, Lig-Aikuma. La tâche consistait, donc, à écouter les enregistrements et à confirmer ou non la correspondance parfaite entre une transcription et l'enregistrement audio associé. Il a résulté de cette analyse manuelle qu'environ la moitié des phrases était lue de manière incorrecte : pour le corpus *dev*, 880 phrases prononcées sur 2 000 divergent de leur transcription (représentant ainsi près de 44% du corpus) ; pour le corpus *test*, ce sont 1 154 phrases sur 2 000 qui ont été mal prononcées (représentant ainsi près de 58% du corpus).

Ainsi, nos nouveaux corpus *dev* et *test* nettoyés peuvent être décrits comme suit :

- *dev* : 1 120 phrases - 1 heure et 12 minutes de durée totale de signaux audio
- *test* : 846 phrases - 55 minutes de durée totale des enregistrements sonores

Ces tailles sont finalement similaires aux tailles des corpus *dev* et *test* du haoussa.

Nous pouvons nous attendre à la même tendance sur les données d'entraînement mais — la vérification des énoncés prenant beaucoup de temps — nous n'avons pas appliqué cette

méthodologie à notre corpus *train*. De plus, l'apprentissage est plutôt robuste pour les transcriptions imparfaites, donc nous avons gardé la totalité des données collectées initialement pour apprendre nos modèles de RAP.

Nous avons ré-évalué notre système de RAP du wolof (qui utilise le modèle de langue `LM_initialPlusWeb`) en utilisant les corpus d'évaluation nettoyés. Les résultats sont présentés dans le tableau 4.6.

Tableau 4.6 – Comparaison des performances des systèmes CD-HMM/DNN du wolof utilisant les transcriptions initiales (avec diacritiques) ou les transcriptions nettoyées (avec diacritiques) – modèle de langue `LM_initialPlusWeb`.

Corpus	**WER** (CER) **(%)**	
	dev	test
Initial	26,0 (9,5)	32,0 (13,0)
Nettoyé	19,0 (6,9)	23,0 (9,3)

Au vu des résultats présentés dans le tableau 4.6, nous pouvons constater que les performances du système sont nettement supérieures lorsqu'il est évalué sur des données nettoyées. Nous obtenons une réduction du WER de 7% sur le corpus *dev* et 9% sur le corpus *test*. Nous avons également observé une réduction du taux de substitution de mots dans chacun des corpus d'évalution (4% sur *dev* et 6% sur *test*). La différence entre les résultats des deux corpus d'évaluation montre, finalement, que le nettoyage des données est une étape supplémentaire, mais indispensable lorsqu'il s'agit de collecte de parole lue dans des langues peu alphabétisées.

Dans la suite de nos travaux pour le wolof, nous n'utiliserons plus que le modèle de langue `LM_initialPlusWeb` pour l'apprentissage de nos systèmes de RAP et le décodage de nouveaux signaux, ainsi que les corpus d'évaluation (*dev* et *test*) nettoyés des enregistrements possédant des transcriptions discordantes.

4.1.2.5 Conclusion sur les systèmes construits

Finalement, nous pouvons résumer les premiers résultats obtenus pour nos systèmes de RAP de référence du haoussa et du wolof. La performance des premiers systèmes de RAP pour le haoussa et wolof, entraînés à partir des approches statistiques (CD-HMM/SGMM) et neuronales (CD-HMM/DNN), est présentée dans le tableau 4.7 ci-dessous. À ce stade, les inventaires de phonèmes (pour les deux langues) ne prennent en compte aucun contraste de longueur de voyelle.

Nous observons, dans le tableau 4.7, que les systèmes à base de réseaux de neurones profonds offrent de meilleures performances que les systèmes à base de modèles markoviens et de mélange gaussien. Ces premiers résultats montrent que notre premier système de RAP pour le haoussa peut atteindre un score WER inférieur à 10%, même sans modélisation spécifique de

la longueur de la voyelle. En ce qui concerne le wolof, notre premier système de RAP [9] peut atteindre un score WER proche des 20%.

Tableau 4.7 – Résultats des systèmes de RAP de référence selon différents modèles acoustiques, pour le haoussa et le wolof – avec adaptation du locuteur.

Langue	WER (CER) (%)			
	CD-HMM/SGMM		CD-HMM/DNN	
	dev	test	dev	test
Haoussa	9,5 (2.7)	19,3 (5,9)	8,0 (2,1)	11,3 (3,7)
Wolof (nettoyé)	22,0 (8,1)	25,1 (10,1)	19,8 (6,9)	23,9 (9,3)

4.1.3 Augmentation de la taille du corpus par perturbation du signal

Pour combler le manque de ressources, nous avons considéré la technique d'augmentation de données, à travers la perturbation de la vitesse du signal acoustique.

L'augmentation de données consiste à augmenter artificiellement la quantité de données d'apprentissage. Cette méthode a été largement utilisée dans le traitement d'images (Paulin *et al.* 2014) et le traitement de la parole contre le bruit (Hannun *et al.* 2014).

Récemment, une simulation consistant à perturber la longueur du conduit vocal (abrégé *VTLP* pour "Vocal Tract Length Perturbation") a montré des résultats encourageants pour la RAP à base de réseaux de neurones profonds (Jaitly et Hinton, 2013). Cette technique a également été appliquée avec succès par Ragni *et al.* (2014) qui a réussi à améliorer la RAP pour l'assamais et le créole haïtien, deux langues peu dotées du programme IARPA Babel [10].

Ko *et al.* (2015) ont présenté une technique d'augmentation du corpus audio avec faible coût d'implémentation. Cette méthode, basée sur la perturbation de la vitesse, a été appliquée sur plusieurs corpus de données vocales. Elle a conduit à une amélioration plus élevée du score WER que la technique VTLP, tout en étant très simple à mettre en œuvre. Cependant, dans l'étude proposée par Ko *et al.* (2015), la perturbation de la vitesse ne s'appliquait qu'aux langues anglaises et mandarines (disposant de ressources suffisantes). Nous avons voulu, par conséquent, contribuer à l'élargissement de cette méthode en l'appliquant aux langues à faible ressources (le haoussa et le wolof).

Nous avons augmenté artificiellement la taille de nos corpus d'apprentissage en modifiant la vitesse des signaux. Pour cela, nous avons testé deux techniques différentes sur nos corpus de parole. La première technique implique une modification du spectre du signal. Nous l'appellerons α-sampling. Tout d'abord, nous avons converti nos fichiers au format brut (*RAW*). Ensuite, nous avons légèrement modifié le taux d'échantillonnage original f par αf (avec α fixé à 0.9 ou α fixé à 1.1). Enfin, nous avons généré un fichier WAV, à partir des échantillons

9. Apprentissage effectué avec le modèle de langue interpolé de 600k mots et décodage sur les données nettoyées.

10. https://www.iarpa.gov/index.php/research-programs/babel

bruts, avec le nouveau taux d'échantillonnage αf. Contrairement à la technique VTLP, ce traitement simple produit un signal de temps déformé qui est plus rapide ou plus lent (en fonction de la valeur de α) que le signal initial. De plus, dans le domaine fréquentiel, cela conduit à une contraction spectrale ou à une dilatation (en fonction de α). Perceptivement, la voix paraît plus aiguë ou plus grave.

Ce qui est décrit ci-dessus est probablement très similaire à la fonction `speed` de l'outil de manipulation audio *Sox*[11] (et utilisé par Ko *et al.* (2015)). Cependant, nous avons remarqué que la fonction `speed`, utilisée avec un même facteur α, conduit à une longueur de signal différente par rapport à la méthode présentée ci-dessus. De plus, nous n'avons pas réussi à trouver une description claire de ce que fait la fonction `speed` de *Sox*. Par conséquent, nous avons aussi décidé de tester les deux techniques dans nos expériences (plus tard, nous ferons référence à *α-sampling* et à *sox/speed* pour identifier chacune des approches). D'un point de vue perceptif, nous n'avons pas relevé de différences majeures entre *α-sampling* et *sox/speed*. Les deux approches modifient la hauteur et l'enveloppe spectrale du signal. Nous pouvons, donc, considérer que cette transformation des signaux permet de simuler de nouveaux locuteurs. Enfin, comme notre objectif était d'augmenter artificiellement la variabilité des locuteurs, il est important de mentionner que nous n'avons pas utilisé la fonction `tempo` de *Sox*. En effet, cette fonction ne produit qu'un changement de vitesse de la parole, tout en essayant de conserver l'enveloppe spectrale et la même hauteur du signal.

En prévision d'un nouvel apprentissage du système de RAP, nous avons créé de nouveaux corpus d'entraînement à partir des signaux originaux et de leurs transformations. Ainsi, chaque méthode a permis de créer un corpus audio de 3 à 5 fois plus grand que le corpus initial. En effet, chaque corpus d'entraînement du haoussa et du wolof a été augmenté en utilisant la méthode *α-sampling* (taille multipliée par 3) ou *sox/speed* (taille multipliée par 3) ou la combinaison des deux (taille multipliée par 5). Puisque chaque transformation (modification du spectre ou modification de la vitesse) entraîne une modification du signal, nous avons attribué de nouveaux identifiants de locuteurs aux signaux transformés.

De cette manière, nous avons pu augmenter notre corpus d'entraînement en haoussa de 166 « nouveaux » locuteurs, pour chacune des techniques de perturbation du signal utilisées (*α-sampling* et *sox/speed*). Ainsi, pour chaque technique de perturbation du signal (*α-sampling* ou *sox/speed*), nous avons deux nouveaux corpus d'entraînement composés chacun de 249 locuteurs (83 locuteurs pour les signaux originaux et 166 locuteurs simulés pour les signaux transformés) constituant un total de 17 589 enregistrements sonores (5 863×3). Le corpus d'entraînement créé à partir de la combinaison des deux techniques (*α-sampling+sox/speed*) est, pour sa part, composé de 415 locuteurs (83 locuteurs pour les signaux originaux et 332 locuteurs simulés à partir des signaux transformés) représentant un total de 29 315 enregistrements audio (5 863 ×5).

Quant au wolof, cette méthodologie nous a permis d'augmenter notre corpus d'entraînement de 28 « nouveaux » locuteurs (pour chacune des deux techniques, *α-sampling* ou *sox/s-*

11. http ://sox.sourceforge.net/sox.html

peed) et de 56 « nouveaux » locuteurs (pour la combinaison des deux, *α-sampling+sox/speed*). Ainsi, pour chaque technique de perturbation du signal (*α-sampling* ou *sox/speed*), nous avons deux nouveaux corpus d'entraînement composés chacun de 42 locuteurs (14 locuteurs pour les signaux originaux et 28 locuteurs simulés pour les signaux transformés) constituant un total de 41 994 enregistrements sonores (13 998×3). Le corpus d'entraînement créé à partir de la combinaison des deux techniques (*α-sampling+sox/speed*) est, pour sa part, composé de 70 locuteurs (14 locuteurs pour les signaux originaux et 56 locuteurs simulés pour les signaux transformés) pour un total de 69 990 enregistrements sonores (13,998×5).

Les résultats des performances de ces systèmes de RAP, sur nos corpus *dev* et *test*, sont reportés dans le tableau 4.8.

Résultats obtenus

Nous avons réitéré le processus d'apprentissage avec les corpus augmentés et avons évalué l'impact sur la performance du système de RAP. Le tableau 4.8 montre les performances obtenues avec les différentes techniques de perturbation du signal, sur le wolof et le haoussa.

Tableau 4.8 – Résultats selon les systèmes de RAP CD-HMM/DNN du haoussa et wolof - avec augmentation des données (par perturbation de la vitesse).

WER (CER) **(%)**	**Haoussa**		**Wolof (nettoyé)**	
	dev	test	dev	test
sans perturbation	*8,0* (2,1)	*11,3* (3,7)	*19,8* (6,9)	*23,9* (9,3)
α-sampling	8,6 (2,3)	12,0 (3,8)	19,8 (7,0)	23,1 (8,9)
sox/speed	8,6 (2,2)	12,7 (3,9)	19,7 (6,8)	23,4 (9,1)
α-sampling+sox/speed	10,2 (2,7)	14,6 (4,6)	20,4 (7,3)	25,4 (9,7)

Le tableau 4.8 permet d'observer que la différence entre les techniques *α-sampling* et *sox/speed* est faible. Sur la haoussa, aucune des deux techniques n'est satisfaisante, en comparaison des performances du système de référence. Sur le wolof, en revanche, chacune des deux méthodes apporte un gain. La méthode *sox/speed* fournit une très légère amélioration sur le corpus *dev* (0,1% de gain) tandis que la méthode *α-sampling* est plus efficace sur le corpus *test* (0,8% de gain).

La dernière observation que nous pouvons faire sur le tableau 4.8 est que l'utilisation des deux techniques à la fois (α-sampling+sox/speed) dégrade les performances des systèmes (jusqu'à 3,3% supplémentaires sur le corpus *test* du haoussa et 1,5% sur le corpus *test* du wolof).

Nous pensons que l'utilisation d'un réseau de neurones profond (dans notre cas, à 7 couches dont 5 cachées) est capable de modéliser correctement les unités acoustiques et la variation inter-locuteur, même avec des modèles acoustiques dont le nombre de locuteurs est limité (comme pour le wolof).

Au regard de nos expérimentations, les résultats de la technique d'augmentation de données sont mitigés. Aucune amélioration n'a été observée sur le système de RAP du haoussa et les gains observés sur celui du wolof restent minimes.

4.2 Autres systèmes de RAP développés pour les langues africaines

Dans le projet ALFFA, nous avons aussi entraîné des systèmes de reconnaissance de la parole pour trois autres langues africaines : l'amharique (Melese, Besacier et Meshesha, 2016), le fongbe (Laleye *et al.* 2016) et le swahili (Gelas, Besacier et Pellegrino, 2012)).

Tableau 4.9 – Performance des systèmes de RAP pour le swahili l'amharique et le fongbe – modèle acoustique à base de CD-HMM/SGMM

Task	**WER (%)**
Swahili (radiodiffusion)	20,7
Fongbe (lecture)	16,6
Hausa (lecture)	10,0
Amharic (lecture)	8,7

Les données nécessaires à l'entraînement d'un système de RAP du swahili sont incluses dans le paquet d'installation Kaldi. Nous pouvons remarquer qu'en comparaison avec le wolof, ces systèmes sont plus performants, mais ces meilleurs résultats sont certainement dus aux différences entre les modèles de langue estimés sur des volumes divers de données (swahili : 28M de mots ; fongbe : 995k de mots ; haoussa : 8M de mots ; amharic : 4M de mots).

4.3 Reproductibilité

4.3.1 Partage sur Github

Nous distribuons toutes les ressources nécessaires à la construction des systèmes de RAP du wolof et du haoussa, mais également pour les autres langues du projet ALFFA (amharique, fongbe et swahili) sur notre dépôt Github [12]. Ces ressources rassemblent les enregistrements accompagnés de leurs transcriptions, les modèles de langue ainsi que les scripts dont a besoin Kaldi pour le développement et l'évaluation d'un système de RAP. Pour le haoussa, nous fournissons les scripts pour Kaldi, mais les ressources pour la construction du système sont à se procurer auprès d'ELRA.

4.3.2 Machines virtuelles

Grâce à l'aide de Eric Riebling et Florian Metze (Université de Carnegie Mellon), nous avons pu développer une machine virtuelle (VM), pour le fongbe [13] et pour le wolof [14], qui permet d'utiliser les systèmes de RAP de référence développés pour ces deux langues, sans avoir de connaissances poussées en informatique (et apprentissage machine).

12. https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR
13. https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR/FONGBE/FONGBE-vm
14. https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR/WOLOF/WOLOF-VM

La VM réalise automatiquement la récupération des ressources depuis notre dépôt Github, le téléchargement et l'installation de Kaldi, puis la construction du système de RAP à partir de nos scripts. Tout ceci est réalisé automatiquement.

Nous utilisons l'outil Vagrant [15] afin d'automatiser le déploiement de la VM. Vagrant utilise un fichier appelé *Vagrantfile* (écrit en Ruby) pour décrire et configurer la VM. Enfin, la machine virtuelle Vagrant s'exécute à l'aide du logiciel VirtualBox [16].

Avantages

La virtualisation permet de recréer un environnement logiciel ou matériel dans un système hôte de même architecture. Ainsi, il est possible de lancer un système d'exploitation Windows sur une machine fonctionnant sous Linux et ainsi de pouvoir lancer des logiciels développés pour la plate-forme Windows uniquement. Tout l'intérêt de la virtualisation est de reconstituer, de façon automatique, une configuration reproductible.

Pour le linguiste et la communauté scientifique en générale, la virtualisation a de nombreux avantages :

- La VM est interopérable, donc elle peut être lancée par tous, peu importe le système d'exploitation utilisé ;
- Les non informaticiens n'ont pas besoin de maîtriser les procédures d'installation et tous les problèmes qui peuvent en découler (numéros de version adéquats, dépendances manquantes, etc.) car la VM se charge de tout ;
- La VM se lance dans un environnement indépendant, stable et optimisé. Autrement dit, il n'y a pas de risque d'infection du système hôte, aucun d'impact si un problème survient dans la VM et celle-ci n'utilise que les ressources allouées dans la configuration ;
- Le linguiste peut reproduire le contexte de ses expérimentations autant de fois qu'il le souhaite et sur différentes machines, car la configuration de la VM est contenue dans un simple fichier.

Les VM que nous avons déployées, en fongbe et en wolof, sont des outils « clé en main ». Elles sont directement utilisables et permettent au linguiste d'obtenir des transcriptions automatiques fournies par le système de RAP, en toute autonomie.

4.4 Résumé

Nous avons présenté, dans ce chapitre, la construction de différents systèmes de RAP pour le haoussa et pour le wolof. Nous avons, tout d'abord, construit des systèmes avec les données à notre disposition, sans modélisation de traits linguistiques particuliers, afin de disposer d'un score de référence.

15. https://www.vagrantup.com
16. https://www.virtualbox.org

Alors que pour le haoussa nous avons utilisé des ressources audio et textuelles déjà disponibles pour la RAP, nous avons, pour le wolof, utilisé nos propres données récoltées au Sénégal. Pour la composition de notre base de connaissance acoustique du wolof, nous nous sommes servis des enregistrements de parole lue, réalisés au Sénégal par Voxygen MBOA (collecte détaillée au chapitre 3, sous-section 3.3.1). En ce qui concerne la constitution de notre base de connaissance linguistique du wolof, nous avons utilisé des documents numériques mis à notre disposition, que nous avons nettoyés. Néanmoins, après un premier apprentissage, nous nous sommes aperçus que les documents textuels utilisés initialement en wolof représentaient trop peu de données pour constituer notre base de connaissance linguistique (le modèle de langue). Nous avons alors récolté près de 500k énoncés supplémentaires à partir du Web. L'augmentation de ce corpus textuel nous a permis d'améliorer la qualité de notre modèle de langue du wolof et, en conséquence, la performance de notre système de RAP.

Par la suite, nous avons considéré la méthode de l'augmentation des données pour accroître la variabilité des locuteurs au sein de notre corpus audio et renforcer notre modèle acoustique. Pour ce faire, nous avons testé deux approches : la première, que nous avons nommée *sox-speed*, utilise la commande `speed` de l'outil *Sox* qui transforme les enregistrements audio en modifiant la vitesse du signal (accélération/décélération) ; la deuxième, que nous avons appelée *α-sampling*, transforme le signal acoustique par compression/dilatation du spectre. Ces deux méthodes modifient le *pitch* et l'enveloppe spectrale du signal acoustique, et ainsi nous permet de simuler de nouveaux locuteurs. Finalement, ces techniques ne nous ont pas apporté de gain important sur la performance du système de transcription.

CHAPITRE 5

Analyser : La machine au service du phonéticien

L'amélioration de la RAP à travers des études phonétiques fines, effectuées en employant des méthodes automatiques elles-mêmes issues des technologies de traitement de la parole, est investiguée depuis longtemps. Les motivations sont doubles : d'une part, les transcriptions fournies automatiquement offrent un gain de temps considérable au linguiste pour effectuer ses analyses ; d'autre part, les études linguistiques menées ont montré des améliorations significatives de la RAP. L'optimisation des systèmes à travers l'appréhension des particularités de la langue peut mener à de meilleures transcriptions, et être pertinente pour l'annotation et l'analyse linguistique de corpus oraux.

Ahmed M. Abdelatty, Van der Spiegel et Mueller (2001) ont développé un système de classification automatique des consonnes plosives en anglais américain. Pour construire leur modèle acoustico-phonétique, ils se sont appuyés sur des descriptions linguistiques et sur leurs propres connaissances, mais aussi sur des outils statistiques comme l'analyse discriminante, les arbres de décision ou encore l'analyse d'histogrammes. Leur système a abouti à une reconnaissance plus précise et robuste de ces consonnes.

Adda-Decker, Boula de Mareüil *et al.* (2005) se sont intéressés, en vue d'améliorer la RAP spontanée du français, aux variations de prononciation des syllabes. Adda-Decker, Boula de Mareüil *et al.* ont utilisé la RAP en tant qu'outil pour leurs analyses linguistiques, en exploitant d'abord les descriptions phonémiques, syllabiques et lexicales de la langue. En retour, leurs analyses ont conduit à de nouvelles observations sur la réalisation des syllabes et leurs structures. Un peu plus tard, Nemoto (2011) s'est attelée à l'analyse acoustico-prosodique des mots et des syntagmes dans deux styles de parole, toujours dans le but de modéliser de façon plus précise les variations de prononciation en français. Les paramètres acoustico-prosodiques ont été extraits automatiquement, à partir d'un système d'alignement en phones [1]. Aussi, Vasilescu, Vieru et Lamel (2014) ont analysé les transcriptions erronées produites par leur système de RAP du roumain contemporain afin de comprendre les particularités de la langue et améliorer l'efficacité de leurs modèles acoustiques.

Les méthodes d'alignement forcé (issues de la RAP) ont également été investiguées en tant qu'outil de traitement automatique de la parole par Adda-Decker, Lamel et Adda (2014) pour analyser la construction phonémique du luxembourgeois, langue peu dotée d'Europe, et sa correspondance avec des langues proches comme l'allemand, l'anglais ou encore le français.

1. Contrairement au phonème qui est la représentation phonémique (théorique) d'un son d'une langue, le phone est la réalisation phonétique (concrète) d'un phonème.

Des analyses linguistiques du mbochi — langue peu dotée d'Afrique — ont aussi été réalisées très récemment à partir d'alignements forcés (Cooper-Leavitt *et al.* 2017a).

Ces nombreuses études montrent que la caractérisation des phénomènes linguistiques est utile, à la fois pour améliorer la RAP et pour entreprendre des analyses sur de grandes quantités de données. De plus, ces analyses menées empiriquement viennent compléter les descriptions linguistiques en apportant des conclusions tangibles.

Les phénomènes linguistiques oraux, parmi lesquels l'opposition de longueur, peuvent impliquer des erreurs d'ordre sémantique ou syntaxique. Ils doivent donc nécessairement être considérés dans des technologies comme la reconnaissance automatique de la parole. En effet, même si certaines ambiguïtés lexicales peuvent être évitées grâce au modèle de langue et au dictionnaire de prononciation, — qui prennent en compte le contexte d'apparition du mot —, la modélisation à une granularité très fine (phonétique) est indispensable pour des technologies vocales incluant de la recherche par mot-clés où les mots sont prononcés de façon isolés.

Pour autant, les phénomènes phonétiques particuliers ne sont pas facilement identifiables et, par conséquent, modélisables. C'est ce que nous allons montrer dans ce chapitre, en prenant pour exemple l'opposition de durée vocalique.

5.1 Cas de l'opposition de longueur dans les langues

Contrairement au français *standard* qui a perdu ce trait distinctif, de nombreuses langues possèdent des phonèmes qui s'opposent sur la base du trait de longueur. Autrement dit, dans ces langues, les phonèmes peuvent être réalisés brefs ou longs. Cette opposition de durée, pour un même son, joue un rôle majeur dans le décodage du mot et la compréhension de la phrase, car la façon dont le mot sera prononcé et perçu modifiera son sens.

Comme mentionné au chapitre 2, le haoussa et le wolof sont des langues dans lesquelles il existe des paires minimales qui varient selon la durée de réalisation de la voyelle.

En haoussa, la prononciation du mot « fito » peut être soit [fi:to:] (la voyelle /i/ est allongée dans la première syllabe) et, dans ce cas, signifie « sifflement » ; soit [fito:] (/i/ est cette fois réalisé bref) qui signifie alors « transport ». Dans cet exemple, nous pouvons constater que c'est la durée de la voyelle qui affecte la signification du mot. De la même manière, le mot « baya » peut être prononcé [ba:ya:] (le /a/ est réalisé long dans les deux syllabes) qui signifie « dos », ou [ba:ya] (le /a/ est allongé dans la première syllabe puis bref dans la seconde) qui veut alors dire « au fond ».

En wolof, les mots « fit » [fit] et « fiit » [fi:t] se distinguent grâce à la durée de prononciation de la voyelle /i/. Si la voyelle est prononcée brève, le mot signifie « courage » ; si la voyelle est prononcée allongée, le mot signifie « piège ». De la même façon, les mots « wall » (*sauver*) et « waal » (*profiter*), ou encore « set » (*être propre*) et « seet » (*rechercher*) ne se différencient,

à l'oral, que par la prononciation (brève ou longue) de leur voyelle. Cette distinction entre voyelles courtes et voyelles longues est aussi décrite dans toutes les variétés de wolof. Comme illustré par les exemples ci-dessus, c'est la réduplication de la voyelle, dans l'orthographe du wolof, qui encode la durée.

D'un point de vue linguistique, l'identification de l'opposition de longueur vocalique, dans ces deux langues, nous permet de vérifier, sur un corpus conséquent de voyelles, si ce contraste décrit au niveau phonologique est effectivement observé au niveau phonétique.

5.1.1 Contributions

La première contribution de ce chapitre est l'analyse à grande échelle du contraste de la longueur des voyelles, pour le haoussa et le wolof. Les voyelles, dans ces deux langues, ont la particularité de s'opposer sur le trait de la longueur. Il est important pour les outils de traitement automatique de la parole de modéliser de façon correcte ce phénomène, car il met en jeu la compréhension de la phrase prononcée. Afin de juger de l'importance de la prise en compte des durées des voyelles dans les systèmes de RAP, nous commencerons par une étude pilote sur le haoussa. Nous verrons que la modélisation de la longueur vocalique en haoussa n'est pas une tâche aisée et nous proposerons un étiquetage des voyelles en fonction de la structure vocalique dans laquelle elle se réalise. Cet étiquetage nous a permis d'analyser phonétiquement les durées des voyelles prononcées en parole lue et de mettre en exergue les contrastes observés. La deuxième contribution de ce chapitre émane de cette analyse, qui montre que l'opposition de longueur n'est pas aussi marquée pour toutes les voyelles. Compte tenu de nos observations sur le haoussa, nous avons également voulu modéliser l'opposition de longueur vocalique existant en wolof. La troisième contribution de ce chapitre est la proposition de plusieurs paramètres pour juger du degré de bimodalité dans la distribution (des durées) pour une voyelle donnée. Ces paramètres montrent différents degrés de contraste selon la voyelle considérée. Les mesures ont été effectuées sur les distributions d'environ 14k voyelles en wolof. Nous montrons également, dans une quatrième contribution que, dans le cas du discours lu, le besoin de transcriptions manuelles peut être relâché pour l'analyse du contraste de durée, car l'utilisation de la RAP peut conduire à des mesures très similaires et ainsi mener aux mêmes conclusions. Enfin, la cinquième contribution de ce chapitre est une application de notre méthodologie assistée par ordinateur pour étudier l'opposition de longueur vocalique dans un discours plus spontané pour le wolof ainsi que pour l'une de ses variantes dialectales, le faana-faana.

5.1.2 Études précédentes

5.1.2.1 L'opposition de longueur dans les langues

Concernant la description et l'analyse des caractéristiques du langage, Gelas, Besacier, Rossato *et al.* (2010) a montré la pertinence de modèles acoustiques multilingues pour étudier, à grande échelle, des phénomènes particuliers de langues (comme le contraste de voyelles en

punu, langue bantoue parlée au Gabon). Plus précisément, son travail a évalué la pertinence d'un alignement automatique pour l'étude des profils de durée, même dans une langue jusqu'alors inconnue de l'outil d'alignement automatique. Les résultats ont indiqué que la distribution bimodale attendue des durées de voyelles pourrait être détectée correctement (et automatiquement). De plus, l'analyse phonémique étant une partie fondamentale de la description et de la documentation d'une langue, l'utilisation de méthodes automatiques a été étudiée pour identifier les unités de sons contrastives dans la langue. Ce processus consiste à prendre une transcription phonétique et à vérifier le contraste entre les paires de phonèmes. Néanmoins, ce travail est fastidieux et, comme investigué par Kempton et Moore (2009), Gelas, Besacier, Rossato *et al.* a utilisé des méthodes automatiques afin de mettre en exergue ces contrastes phonémiques.

La durée de la voyelle est une mesure largement utilisée en phonétique acoustique. De nombreux facteurs affectent la durée de la voyelle comme son emplacement dans l'espace vocalique (Lindblom, 1967 ; Lehiste, 1970), la position et la longueur du mot (Lindblom, Lyberg et Holmgren, 1981), le contexte environnant de la voyelle (House, 1961 ; Maddieson, 1984), le débit de la parole (Gay, 1981 ; Magen et Blumstein, 1993) et la position de la voyelle au sein du mot (Myers, 2005).

Magen et Blumstein (1993) ont étudié les effets du débit de parole sur la distinction (au niveau perceptif) des voyelles en coréen. Leurs résultats ont montré que les voyelles longues et brèves varient selon le débit de parole, de sorte que les effets symétriques trouvés pour les consonnes ne sont pas trouvés pour les voyelles. Les voyelles brèves ne fournissent pas une ancre phonétique : les durées de voyelles brèves produites à un rythme lent chevauchent presque toujours celles des voyelles longues produites à un rythme rapide.

Plus tard, Myers (2005) a étudié l'opposition de longueur des voyelles en kinyarwanda, langue bantoue où la longueur est un élément fondamental de la compréhension, car elle permet de distinguer deux mots. À l'instar du haoussa et contrairement au wolof, la longueur n'est pas marquée dans l'orthographe. En kinyarwanda, les voyelles longues n'apparaissent jamais en position initiale ou finale de mot. Myers a finalement observé dans son étude les mêmes caractéristiques inhérentes aux voyelles que décrit dans la littérature de phonétique acoustique : les voyelles dites brèves ont effectivement une durée inférieure à celle des voyelles dites longues et l'aperture de la voyelle affecte sa durée (les voyelles fermées sont plus courtes que les voyelles ouvertes). Mais il a aussi montré que l'allongement de la voyelle, lorsqu'elle est située dans l'avant-dernière syllabe, est réalisée en kinyarwanda. Cette observation avait déjà été faite pour d'autres langues bantoues, mais n'avait encore pas été reportée pour cette langue. Ce phénomène, commun à de nombreuses langues bantoues, qui s'expliquerait par l'accent porté sur l'avant-dernière syllabe, est par ailleurs toujours source d'intérêt auprès des linguistes (Hyman, 2013).

Comme soulevé par Adi *et al.* (2016), les principales études antérieures sur la durée des voyelles ont été effectuées à l'aide d'annotations manuelles. L'annotation manuelle est une tâche chronophage. En effet, le temps passé au découpage, à l'écoute et à l'annotation des voyelles, occasionne une analyse partielle de la durée des voyelles, car opérée sur un nombre

limité de mots. Nous pensons que l'utilisation d'outils automatiques peut conduire à des mesures plus objectives et reproductibles, à plus grande échelle.

Dans la continuité des travaux effectués par Magen et Blumstein, Lee et Shin (2016) ont étudié la production et la perception du contraste de longueur sur les voyelles, en coréen, avec des méthodes automatiques. Ils ont remarqué que tous les locuteurs coréens de l'étude produisaient l'opposition de longueur existant au sein des voyelles, mais ils ont également conclu que l'opposition brève *versus* longue est plus faible en discours spontané.

L'opposition de longueur des voyelles a également été étudiée pour mieux comprendre l'acquisition du langage. Bion *et al.* (2013) ont analysé 11 heures de discours à destination du nouveau-né, en utilisant des méthodes statistiques, pour explorer comment les nourrissons apprennent à discriminer le contraste de longueur de voyelles existant en japonais. Ils ont découvert que la distribution statistique de durée pour une voyelle donnée n'est pas clairement bimodale (les distributions se chevauchent), ceci à cause du fait que les voyelles longues sont beaucoup moins fréquentes que les voyelles brèves.

En wolof, très peu d'études phonétiques ont été publiées, notamment sur le contraste de la longueur des voyelles. Une exception est le travail de Sock (1983) qui a étudié une variante dialectale du wolof gambien, proche du faana-faana étudié dans ce chapitre. Sock a expliqué que la durée des voyelles est inhérente à chacune d'elles (par exemple, la voyelle /a/ est toujours plus longue que la voyelle /i/). Dans son étude — effectuée sur de la parole lue —, il a comparé 3 paires minimales, chacune contenant les voyelles /i/, /a/ et /u/ et a remarqué que l'opposition de longueur était plus important pour la voyelle /a/ que pour /i/ et /u/. Sock a également relevé que le débit de parole joue un rôle sur la production de la longueur vocalique : en débit rapide, le contraste était moins marqué qu'en débit normal.

Finalement, en 2006, Cissé (2006) a souligné qu'une analyse exhaustive de la phonétique du wolof faisait défaut et, à notre connaissance, cet état de fait est encore d'actualité.

5.1.2.2 Modélisation de la durée phonémique dans les systèmes de RAP

Les informations de durée sont encore peu utilisées dans la modélisation des systèmes de RAP, en raison de l'utilisation de modèles de Markov cachés (HMM) qui ne permettent pas de modéliser de façon adéquate les durées de phonèmes (Pylkkönen et Kurimo, 2004). En 2000, Gadde (2000) s'intéresse à la prise en compte de l'information de durée au niveau acoustique, en modélisant la durée des mots. Dans ces travaux, Gadde a amélioré le taux d'erreur de mots de son système de RAP en attribuant un nouveau score aux N-meilleures listes, avec les modèles de durée construits. Quelques années plus tard, Povey (2004) a travaillé, lui aussi, sur la modélisation de la durée au niveau des mots, mais également au niveau des phonèmes. Povey a utilisé la méthode de la réévaluation de treillis de mots (en anglais, "lattice rescoring") mais la réduction du taux d'erreur de mots était limitée. Néanmoins, la technique mise en œuvre permet d'améliorer la qualité de la transcription en combinant les systèmes. La même année, Pylkkönen et Kurimo (2004) ont étudié le phénomène de contraste longueur phonémique existant en finnois, langue dans laquelle la durée est importante pour la compréhension du sens d'un mot. Différentes techniques de modélisation (variantes de HMM) sont comparées

par Pylkkönen et Kurimo. Ils obtiennent, en plus d'une réduction du temps de décodage, une réduction relative du taux d'erreur de caractères de 8%, par rapport au système de référence (qui ne tient pas compte de la durée des phones). Plus tard, Alumäe et Nemoto (2013) se sont aussi intéressés à la modélisation de la durée des phonèmes en estonien car, dans cette langue, il existe trois longueurs différentes d'un même phonèmes (court, long et très long). Alumäe et Nemoto ont pris en compte la durée du phonème en contexte (le phone), ainsi que les connaissances linguistiques et phonétiques, pour former des groupes de phones ayant des durées similaires (à l'aide d'un arbre de décision). Ils ont réévalué la liste des N-meilleures hypothèses de leur système de RAP avec leur modèle de durée, mais le taux d'erreur de mots n'était, encore une fois, que légèrement amélioré.

Finalement, toutes ces approches prennent pour acquis l'opposition de longueur des phonèmes alors que nous pensons que, pour les langues peu dotées, la vérification empirique de leur réalisation concrète est une étape importante, avant de décider de séparer un phonème en deux phones distinctifs (bref *versus* long) dans les modèles de RAP.

5.2 Prise en compte des longueurs de voyelles dans les modèles de RAP en haoussa

Dans le but d'analyser, de façon empirique, la production du contraste de longueur en haoussa et en wolof, nous avons tout d'abord construit de nouveaux systèmes de RAP prenant en compte ces distinctions de durées. Pour ce faire, nous avons construit de nouveaux dictionnaires de prononciation, dans lesquels nous avons attribué une étiquette aux voyelles afin de différencier leur durée, d'après les descriptions phonologiques que l'on peut retrouver dans la littérature. Ainsi, nous avons étiqueté chaque voyelle de nos dictionnaires de prononciation selon qu'elle est décrite comme se réalisant brève ou longue. Nous décrivons, ci-dessous, notre méthodologie d'annotation des voyelles.

Afin d'étiqueter les voyelles, nous avons envisagé l'utilisation des alignements forcés. L'alignement forcé consiste en l'alignement temporel de la transcription avec le signal audio qui lui correspond. Nous avons appliqué l'algorithme d'alignement phonémique fournit par la boîte à outils Kaldi, qui utilise l'algorithme de Viterbi (Viterbi, 1967) pour décoder les phonèmes modélisés avec des HMMs. Les ressources nécessaires à l'obtention d'un alignement forcé sont un dictionnaire de prononciation, un modèle acoustique et des fichiers audio accompagnés de leurs transcriptions (phonétiques ou orthographiques).

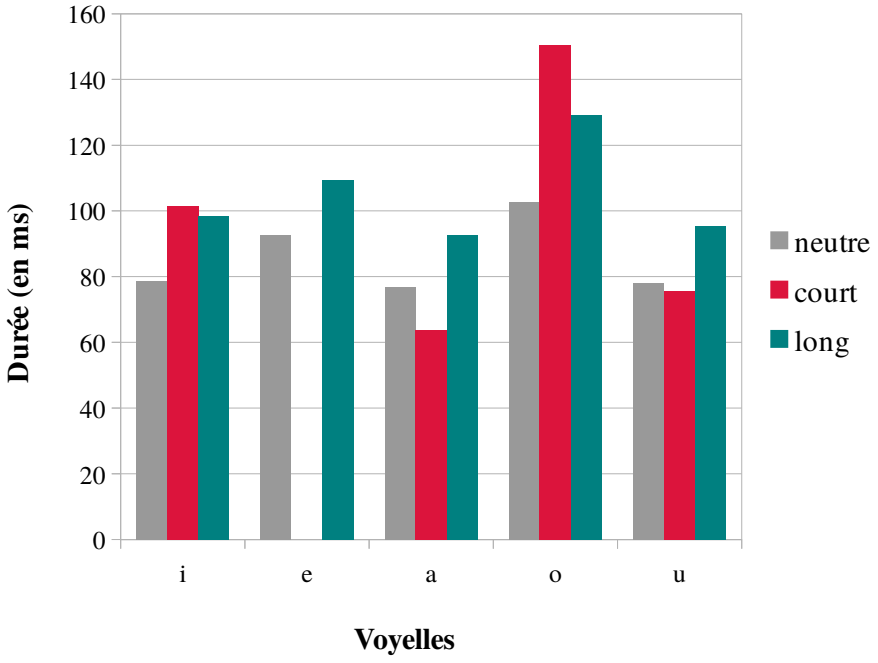
5.2.1 Étiquetage des voyelles du haoussa

5.2.1.1 Le dictionnaire de prononciation du corpus Globalphone

Dans un premier temps, nous avons utilisé le dictionnaire de prononciation construit par Schlippe *et al.* (2012) afin de construire le modèle acoustique discriminant les longueurs de voyelles. Dans ce dictionnaire, des étiquettes sont apposées sur les phonèmes vocaliques. Ces

étiquettes fournissent des informations sur les tons et les longueurs en haoussa. Nous ne nous sommes intéressés qu'à l'opposition de longueur en haoussa, et avons par conséquent retiré les informations tonales fournies.
Deux étiquettes apportent des informations sur la nature de la longueur vocalique dans ce dictionnaire : une étiquette indiquant que le phonème se réalise bref (notée `_S`) et une étiquette indiquant que le phonème se réalise long (notée `_L`). De plus, il subsiste des voyelles non annotées (nous considérerons ces voyelles comme portant une étiquette *neutre*). Ces 3 informations fournies sont mutuellement exclusives.

Performance du système entraîné. Le système de RAP CD-HMM/DNN entraîné avec ce dictionnaire obtient un score de 8,3% sur le corpus *dev* (contre 8,0% avec le système de référence) et 11,4% sur le corpus *test* (contre 11,3% avec le système de référence). Aucun gain n'est donc observé et les performances sont même légèrement amoindries. Pour avoir une idée de la manière dont sont modélisées les durées des voyelles par le système, nous avons forcé l'alignement des enregistrements utilisés pour l'apprentissage du système. Nous avons ensuite calculé la durée moyenne de chaque voyelle, en fonction de l'étiquette qui lui est attribuée dans le dictionnaire (bref, long, neutre). Le graphique 5.1 montre que les durées estimées par le système ne correspondent pas aux étiquettes apposées dans le dictionnaire. La voyelle /i/ étiquetée « brève » a une durée moyenne plus élevée que celle étiquetée « longue », tout comme la voyelle /o/ pour laquelle l'écart est encore plus flagrant. De plus, la voyelle /e/ dans sa version « brève » n'existe pas.



Graphique 5.1 – Durée moyenne des cinq voyelles du haoussa étiquetées dans le dictionnaire Globalphone.

Ces incohérences au niveau de l'étiquetage des informations de durée sur les voyelles nous ont incité à modifier le dictionnaire acoustique.

5.2.1.2 Nouvelle approche d'étiquetage des voyelles du haoussa

Nous avons imaginé un nouvel étiquetage des longueurs vocaliques en haoussa, qui s'appuie de manière plus précise sur les descriptions phonologiques de la langue. Ainsi, afin de représenter au mieux les descriptions phonologiques du haoussa pour ce qui concerne les longueurs vocaliques, nous avons décidé de développer notre propre algorithme d'étiquetage. Cet algorithme a permis d'associer à la voyelle une étiquette de longueur en fonction du contexte syllabique dans lequel elle se situe. Cet étiquetage est un étiquetage *a priori*, établi selon les descriptions phonologiques détaillées dans la littérature descriptive du haoussa.

Pour rappel, en haoussa, le contraste de longueur n'est pas marqué dans l'orthographe standard. La longueur de la voyelle varie en fonction de plusieurs facteurs qui peuvent être le type de syllabe dans lequel la voyelle se situe, la position de la voyelle au sein du mot (début, milieu, fin) ou encore la position de la voyelle dans l'énoncé (avant une pause ou non) (Newman et Heuven, 1981). Dans ce travail de thèse, nous nous sommes uniquement concentrés sur la position de la voyelle au sein de la syllabe.

En haoussa, la syllabe peut être fermée (c'est-à-dire que la syllabe se termine par une consonne, sa forme est CVC) ou ouverte (la syllabe se termine par une voyelle simple ou une diphtongue, sa forme est CV ou CVV). De plus, Newman (2000) explique que, dans une syllabe fermée, les voyelles sont toujours courtes alors que, dans une syllabe ouverte (et lorsque la voyelle est en position pré-pausale), un contraste de longueur peut apparaître. Plus précisément encore, Jaggar (2001) spécifie que les voyelles /e/ et /o/ ont une durée plus longue dans un contexte syllabique ouvert que dans un contexte fermé. Par conséquent, pour les syllabes fermées, nous nous attendons à avoir des voyelles plus courtes que pour les syllabes ouvertes. Finalement, ces descriptions phonologiques de la langue nous ont permis d'attribuer aux voyelles différentes étiquettes en fonction du contexte syllabique.

Avant de considérer l'entraînement d'un nouveau système de RAP avec un dictionnaire de prononciation tenant compte de notre nouvelle approche d'étiquetage des longueurs vocaliques, nous avons de nouveau analysé, empiriquement, la variation de durée des voyelles au sein de la langue. Nous avons utilisé les alignements forcés, dans lesquels l'information concernant la place du phonème dans le mot (initiale, milieu, finale) est indiquée. Nous nous sommes servis de ces indications pour découper les mots en syllabes [2]. Ce découpage nous a permis, ensuite, d'attribuer une étiquette indiquant dans quel type de structure syllabique (ouverte ou fermée) la voyelle apparaît. Pour ce faire, nous avons utilisé une fenêtre de 4 phonèmes (le phonème précédent, le phonème courant et les deux phonèmes suivants). En effet, pour le mot « akuya » par exemple, si le phonème courant est /u/, nous avons besoin de savoir quel est le phonème qui suit /y/, s'il y en a un. Dans cet exemple, il y en a un (le phonème /a/). Donc, cela signifie que /u/ apparaît dans une syllabe de type ouverte, parce que nous pouvons décomposer le mot de cette manière : V-CV-CV (*a-ku-ya*). Au contraire, s'il n'y avait rien après

2. En effet, les alignements forcés fournis par Kaldi contiennent une suite continue de phonèmes ; sans ces informations sur la frontière des mots, nous n'aurions pas su quand le mot se terminait ou commençait (en conséquence, nous n'aurions pas pu segmenter les mots en syllabes).

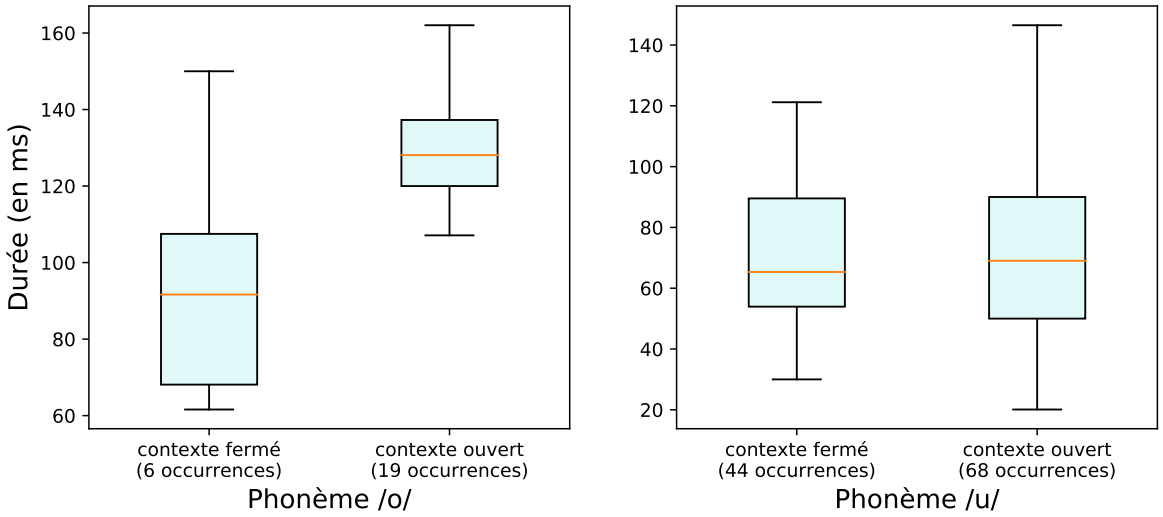
/y/, la structure syllabique aurait été V-CVC, et une étiquette indiquant que la syllabe est fermée aurait été assignée au phonème/u/. En outre, la raison pour laquelle nous avons besoin de connaître le phonème précédant la voyelle courante est que nous pouvons avoir une structure CVV (c'est-à-dire une syllabe constituée d'une consonne suivie d'une diphtongue à considérer comme un tout).

De cette façon, nous avons attribué — d'après les critères phonologiques de la langue —, soit une étiquette _C (pour "*Closed*") indiquant que la voyelle est le noyau d'une syllabe fermée, soit une étiquette _O (pour "*Open*") indiquant que la voyelle est le noyau d'une syllabe ouverte.

Finalement, cet étiquetage (voyelle fermée *versus* ouverte) effectué en considérant le contexte syllabique de la voyelle, nous a permis de la catégoriser et d'analyser sa durée.

5.2.2 Résultats d'analyse sur le haoussa

5.2.2.1 Analyse d'un échantillon de 102 énoncés réalignés manuellement



Graphique 5.2 – Diagrammes de Tukey illustrant les distributions des durées du phonème /o/ et du phonème /u/ du haoussa – Annotation manuelle des frontières.

Tout d'abord, nous avons sélectionné de manière aléatoire, dans notre corpus d'apprentissage, 102 énoncés prononcés par tous les locuteurs (82 au total). Nous avons forcé l'alignement de ces énoncés avec notre modèle acoustique CD-HMM/SGMM [3]. Puis, nous avons corrigé manuellement ces alignements forcés, en réalignant chaque frontière de voyelles avec l'outil Praat [4]. Cet outil permet d'écouter un fichier audio tout en visualisant, conjointement, le spectrogramme et les transcriptions associés.

3. Au début de ces expérimentations, nous n'avions pas encore entraîné de modèle acoustique employant une approche neuronale.

4. Pour que les alignements forcés fourni par Kaldi soit lisibles sous Praat, nous avons dû les convertir au format TextGrid. En effet, Kaldi les fournit au format *stm*. Ce format de fichier est communément utilisé, avec le format *ctm*, par les outils de RAP pour marquer les segments temporels reconnus par la machine dans un enregistrement audio. Pour une description détaillée de ce format, se référer au site web http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm

Ces 102 alignements corrigés manuellement nous serviront de référence et seront systématiquement comparés aux alignements obtenus automatiquement dans le reste de cette analyse pilote sur le haoussa.

Nous avons, donc, extrait les durées des voyelles provenant de ces annotations manuelles et avons calculé leur distribution statistique en fonction de l'étiquette « fermée » (_C) ou « ouverte » (_O). Pour rappel, cette étiquette est attribuée selon l'environnement syllabique dans lequel la voyelle se situe.

Nous avons choisi les diagrammes de Tukey pour représenter les distributions de durées de chaque phonème vocalique. Nous avons également représenté les distributions au moyen d'un histogramme. Ces représentations sont consultables en annexe I.

Ces représentations graphiques ont mis en évidence le fait que l'opposition de longueur n'est pas facilement observable pour toutes les voyelles du haoussa. La tendance qui se dégage est que le contraste de durée semble fortement marqué pour les voyelles /e/ et /o/ seulement. Les distributions de durées se chevauchent pour les autres voyelles (/i/, /a/ et /u/).

Afin d'illustrer ce résultat, nous présentons sur le graphique 5.2 les distributions de durées des voyelles /o/ et /u/ (obtenues à partir des alignements forcés corrigés manuellement). Pour plus de clarté, les valeurs extrêmes ne sont pas représentées sur ce graphique, mais le sont sur ceux joints en annexe. Cette figure montre finalement la différence de contraste que nous avons pu constater au sein des voyelles du haoussa.

Nous pouvons observer que la voyelle /o/ dans une syllabe ouverte est réalisée plus longue que dans une syllabe fermée, avec des quartiles qui se distinguent. Au contraire, les quartiles quasiment alignés de la voyelle /u/ dénotent une opposition de longueur quasiment inexistante. Les graphiques illustrant les distributions des autres voyelles sont jointes à l'annexe I.

5.2.2.2 Analyse des 5 863 énoncés alignés automatiquement

Dans cette section, nous avons voulu voir si les tendances observées précédemment, à partir d'alignements corrigés par l'homme, se vérifiaient sur les alignements effectués par le système de RAP, non corrigés.

Comparaison entre alignements automatiques et alignements corrigés sur 102 énoncés

Dans un premier temps, nous avons validé la qualité des alignements forcés (automatiques) du système de RAP en comparant les durées estimées avec celles mesurées après réalignement manuel sous Praat.

Pour cela, nous avons comparé le contraste de durée, de chaque voyelle, obtenu par alignement automatique pour les 102 énoncés sélectionnés aléatoirement, et celui obtenu pour ces mêmes 102 énoncés pour lesquels nous avons corrigés manuellement les alignements. Ces alignements forcés ont été obtenus en utilisant notre modèle acoustique CD-HMM/SGMM [5].

5. Au moment de ces expérimentations, nous n'avions pas encore entraîné de modèle acoustique à base de réseaux de neurones.

Cette mesure du contraste de durée est représentée par le delta de leur durée moyenne en contexte syllabique fermé et en contexte syllabique ouvert. Un delta élevé nous indiquera que l'opposition de longueur est fortement marquée.

Tableau 5.1 – Delta des durées moyennes des cinq voyelles du haoussa prononcées dans une syllabe ouverte et fermée, calculé à partir des données obtenues par alignement automatique du système CD-HMM/SGMM de référence ou les réalignements manuels.

Δ (en ms)	**/i/**	**/e/**	**/a/**	**/o/**	**/u/**
Alignement manuel (102 énoncés)	-7,3	34,0	15,4	36,9	0,2
Alignement automatique (102 énoncés)	-0,5	43,4	8,6	39,2	7,2

Le tableau 5.1 présente le delta des durées moyennes mesuré pour les voyelles du haoussa, provenant des 102 alignements forcés non corrigés et ces mêmes 102 alignements forcés corrigés avec l'outil Praat.

Les deux lignes du tableau 5.1 montrent le delta des durées moyennes pour chaque voyelle.

La ligne « alignement automatique » correspond aux durées extraites des 102 alignements forcés qui ont été sélectionnés aléatoirement pour être corrigés.

La ligne « alignement manuel » correspond aux durées extraites des 102 alignements forcés qui ont été corrigés manuellement.

Ce tableau montre finalement que les deltas mesurés à partir des alignements automatiques sont proches de ceux mesurés à partir des alignements corrigés manuellement. De plus, les delta positifs montrent que les voyelles étiquetées « ouvertes » (donc phonologiquement décrites comme se réalisant allongées) ont une durée de réalisation, en moyenne, effectivement plus longue que celles étiquetées « fermées ». En revanche, cette tendance n'est pas observée pour la voyelle /i/ dont le delta est négatif.

Par ailleurs, une fois notre modèle CD-HMM/DNN entraîné, nous avons réalisé à nouveau ces mesures sur les alignements forcés fournis par cette approche, dans le but de comparer la justesse de l'alignement entre les deux modèles.

Ceci nous a montré que les différences de durée entre les alignements SGMM et DNN ne sont pas significatives et que les deux approches de modélisation acoustique semblent appropriées pour analyser le contraste vocalique. Le tableau exposant ces mesures entre les deux approches CD-HMM/SGMM et CD-HMM/DNN est accessible en annexe (tableau K.1).

Puisque ces mesures préliminaires montrent que les durées calculées à partir des alignements réalignés manuellement sont finalement proches des durées calculées à partir des alignements automatiques bruts, nous avons procédé à la vérification du contraste pour la totalité du corpus d'apprentissage (représentant 5 863 énoncés). Par souci de cohérence, et puisque les approches SGMM ou DNN semblent équivalentes pour l'estimation des durées, nous avons continué d'utiliser les alignements forcés SGMM pour la suite de cette étude sur le haoussa.

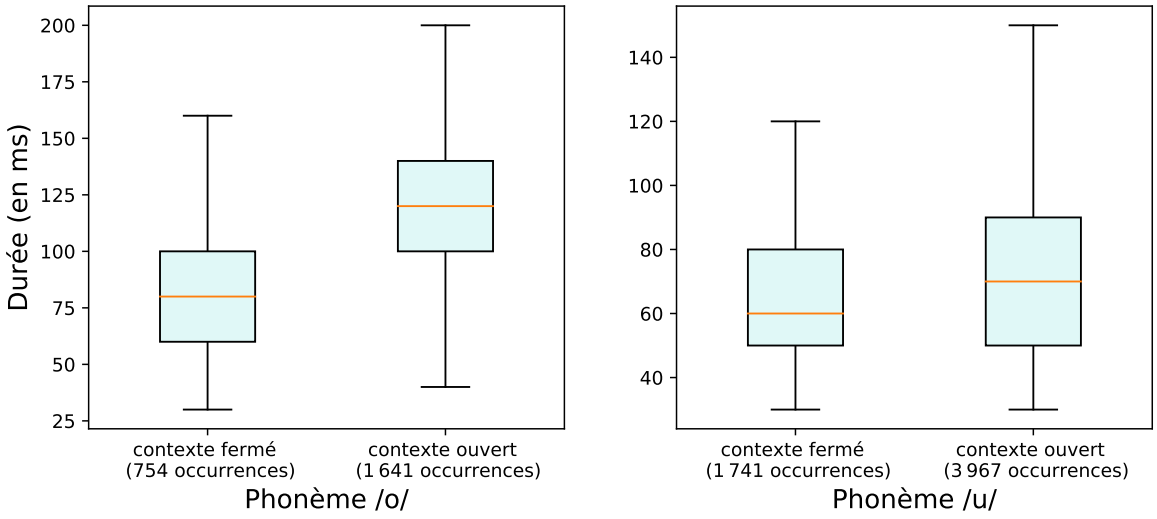
Nous avons appliqué la même méthodologie que dans la section précédente, en utilisant, donc, cette fois-ci, les sorties des alignements forcés du système de RAP obtenues sur les 5 863

énoncés du corpus d'apprentissage du haoussa, tout en vérifiant au préalable que la proportion de chaque voyelles était la même entre les deux ensembles de données (manuel et automatique). Cette répartition est indiquée sur le graphique J.1, en annexe J.

Analyse des alignements automatiques sur 5 863 énoncés

Ainsi, nous avons forcé l'alignement des 5 863 énoncés prononcés par les 82 locuteurs de haoussa du corpus d'apprentissage, avec le modèle acoustique CD-HMM/SGMM. Puis, nous avons calculé les durées de chaque voyelle et tracé leurs distributions statistiques.

Le graphique 5.3 illustre la distribution des durées, pour le phonème vocalique /o/ et le phonème vocalique /u/, estimées de façon automatique par le système de RAP. Les distributions statistiques de toutes les voyelles sont jointes en annexe L. Nous pouvons constater, en le comparant avec le graphique 5.2 que les distributions de durées sont similaires. Ceci signifie que le système a été capable d'estimer les frontières des phonèmes proches de ce qu'aurait pu réaliser un humain.



Graphique 5.3 – Diagrammes de Tukey illustrant les distributions des durées du phonème /o/ et du phonème /u/ du haoussa – Annotation des frontières par alignement forcé.

Les conclusions de cette analyse, réalisée sur plus de 45k de voyelles, sont similaires à celles tirées des alignements manuels. Le contraste de longueur n'est pas marqué pour toute les voyelles : seules les voyelles /e/ et /o/ présentent un contraste de durée fort. Notre hypothèse, concernant ces résultats, est que la longueur des voyelles /e/ et /o/ est plus prévisible en fonction de leur contexte syllabique. En effet, comme évoqué plus haut en sous-sous-section 5.2.1.2, Jaggar (2001) explique que les voyelles /e/ et /o/ répondent à des règles strictes : elles se réalisent brèves dans une syllabe fermée, tandis qu'elle se réalisent allongées dans une syllabe ouverte. Pour les trois autres voyelles /a/, /i/ et /u/, les linguistes ne sont pas toujours d'accord sur leurs réalisations phonétiques et nos résultats empiriques confirment que le contexte syllabique (ouvert/fermé) seul ne permet pas de prédire leur durée. D'autres règles sont nécessaires, telle que la prise en compte de la position de la voyelle au sein du mot par exemple. De plus, nous avons remarqué un écart entre le nombre d'occurrences de voyelles étiquetées

« fermée » ou « ouverte ». Les voyelles en contexte syllabique ouvertes sont plus fréquentes que leur contrepartie apparaissant en contexte fermé, ce qui rend l'observation du contraste délicate.

Cette analyse assistée par ordinateur du contraste de longueur de voyelles montre que sa réalisation concrète dépend de la voyelle considérée. Finalement, ces mesures empiriques, à grande échelle (45k de voyelles considérées), nous ont incité à essayer de modéliser l'opposition de longueur des voyelles dans nos systèmes de RAP. C'est ce que nous proposons dans la section suivante.

5.3 Modélisation de la durée des voyelles dans l'apprentissage des systèmes de RAP

Au vu des résultats obtenus sur les analyses effectuées à partir des alignements forcés du haoussa, nous avons également considéré l'étiquetage des voyelles du wolof en fonction de leur longueur.

5.3.1 Étiquetage des voyelles du wolof

Contrairement au haoussa, l'étiquetage des voyelles du wolof a été relativement simple. En effet, la longueur (phonologique) est marquée par la duplication de la voyelle dans l'orthographe du mot. Lorsque la voyelle est simple, nous lui avons attribué une étiquette « _S » (pour *Short*), qui indique que la voyelle est phonologiquement considérée brève ; de la même manière, lorsque la voyelle apparaît doublée dans le mot, nous avons rassemblé les deux phonèmes vocaliques correspondants en un seul et lui avons assigné une étiquette « _L » (pour *Long*), qui indique que la voyelle est phonologiquement considérée allongée. Toutes les voyelles de notre inventaire phonétique possèdent ainsi deux réalisations dans notre dictionnaire de prononciation, sauf le phonème /ə/ (orthographié "ë") dont l'allongement n'est pas phonologique.

Tout comme pour le haoussa, nous avons d'abord entraîné un système modélisant le contraste de durée pour la totalité des voyelles du wolof (pour lesquelles les descriptions phonologiques de la langue expliquent qu'une opposition de longueur existe).

Performance du système entraîné avec étiquetage des longueurs sur toutes les voyelles du wolof. Nous avons relancé la procédure complète pour construire un système de RAP avec Kaldi. Les meilleurs scores pour chacun de nos corpus d'évaluation et de test proviennent du modèle acoustique CD-HMM/DNN. Nous avons obtenu un score WER de 20,0% sur notre corpus *dev* (contre 19,8% avec le système de référence) et un score de 24,6% sur notre corpus de *test* (contre 23,9% avec le système de référence). Nous n'avons, par conséquent, observé aucun gain sur les performances du système.

Pour autant, nous avons forcé l'alignement des énoncés du corpus d'apprentissage, en utilisant le modèle acoustique CD-HMM/DNN, afin de vérifier si l'opposition de durée était observable au niveau des voyelles. Nous avons tracé les distributions (statistiques) des durées pour chaque phonème vocalique annoté. Finalement, ces distributions nous ont montré que, comme pour le haoussa, le contraste de durée au sein d'une même paire de voyelle est observable.

Au vu de ces résultats, nous avons décidé de modéliser dans notre système de RAP l'opposition de longueur dans les deux langues en sélectionnant les voyelles pour lesquelles le contraste observé était le plus marqué. Nous espérons qu'un système modélisant plus finement le phénomène d'opposition de longueur vocalique fournisse de meilleures performances.

5.3.2 Nouvel étiquetage des dictionnaires du haoussa et du wolof

Nous avons formé de nouveaux modèles acoustiques du haoussa et du wolof, en forçant le système à représenter ce contraste. Nous avons ainsi étiqueté de façon automatique, comme indiqué à la section 5.2, toutes les voyelles de nos dictionnaires de prononciation du haoussa et du wolof. Autrement dit, nous avons modélisé deux unités phonétiques (phones) différentes pour un même phonème vocalique.

5.3.3 Nouveaux systèmes construits

Pour chaque langue, nous avons entraîné un système de RAP pour lequel le modèle acoustique et le dictionnaire de prononciation utilisés modélisent la longueur pour seulement un sous-ensemble sélectionné de voyelles. Ces systèmes tiennent compte du contraste existant *a priori* entre les longueurs des voyelles.

Le tableau 5.2 montre les étiquettes attribuées aux sous-ensembles de voyelles. Nous pouvons voir que, pour le haoussa, nous avons également dans notre dictionnaire une étiquette « `_unk` ». En effet, il reste des voyelles non étiquetées dans notre dictionnaire du haoussa, puisque nous n'avons tenu compte que de la position des voyelles au sein des syllabes, et non de la grande variabilité de leur réalisation phonétique en fonction de leur position dans le mot.

Tableau 5.2 – Résumé des étiquettes attribuées à un sous-ensemble de voyelles pour lesquelles le contraste de longueur observé a été le plus fort, pour chaque système de RAP entraîné.

Nouveaux systèmes	RAP du haoussa	RAP du wolof
#Contrastes	2 ([e], [o])	5 ([a], [e], [ɛ], [o], [ɔ])
Étiquettes	`_C` / `_O` / `_unk` Fermé / Ouvert / Inconnu	`_S` / `_L` Bref / Long

À ce stade, le dictionnaire de prononciation du haoussa est ainsi composé de 37 phonèmes (au lieu de 33 initialement) et celui du wolof est constitué de 39 phonèmes (au lieu de 34 initialement). En utilisant ces nouveaux dictionnaires de prononciation, nous avons entraîné, pour

chaque langue, un nouveau système de RAP initié par le même protocole que détaillé au chapitre précédent. Le modèle à 3 états du haoussa est entraîné en utilisant 2 969 états dépendants du contexte, tandis que celui du wolof a été entraîné avec 3 406 états (dépendants du contexte), ainsi que 40k gaussiennes.

Le tableau 5.3 expose les résultats obtenus en modélisant la longueur comme expliqué ci-dessus. Nous rappelons également le score WER obtenu sans modélisation de l'opposition de longueur vocalique et avec une modélisation de cette opposition sur la totalité des voyelles.

Tableau 5.3 – Résultats des systèmes de RAP CD-HMM/DNN avec modélisation de la longueur vocalique, pour le haoussa et le wolof – avec adaptation du locuteur.

Langue	**Type de modélisation**	**CD-HMM/DNN WER (%)**	
		dev	test
Haoussa	sans modélisation	8,0	11,3
	avec modélisation sur toutes les voyelles	8,3	11,2
	avec modélisation sur un sous-ensemble de voyelles	7,9	10,6
Wolof (nettoyé)	sans modélisation	19,8	23,9
	avec modélisation sur toutes les voyelles	20,0	24,6
	avec modélisation sur un sous-ensemble de voyelles	20,0	24,5

Nous pouvons voir dans le tableau 5.3 que nous avons légèrement amélioré les performances du système de RAP du haoussa (-0,1 % de WER pour le corpus *dev* et -0,7 % de WER pour le corpus *test*). À propos du système de RAP du wolof, nous observons en revanche une légère dégradation des performances à ce stade.

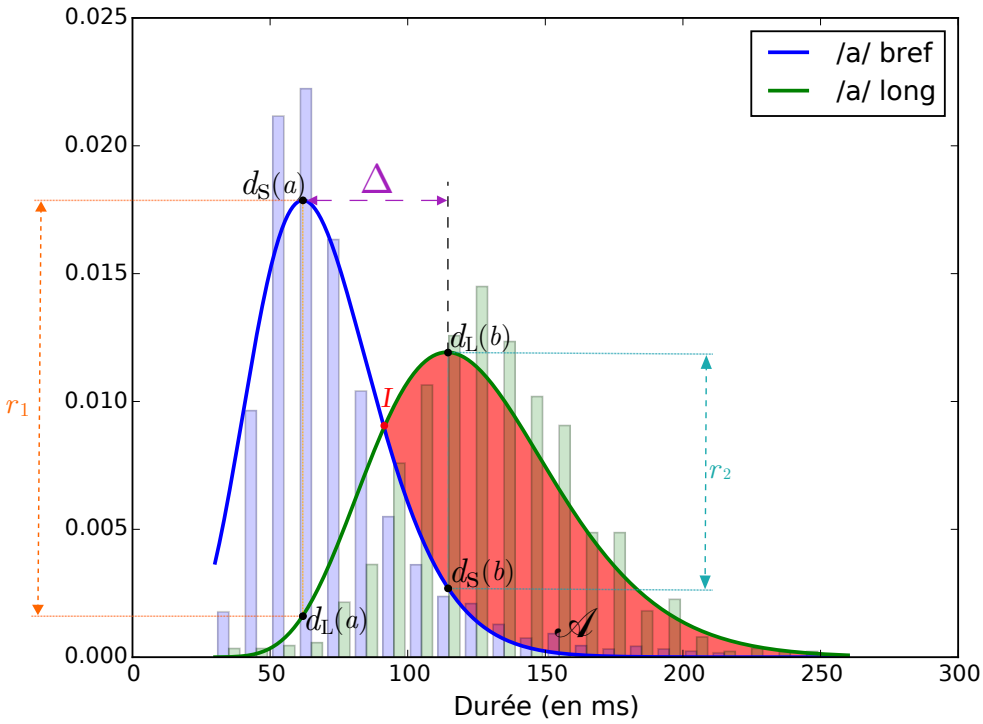
Bien que les performances des systèmes entraînés soient similaires, nous pensons que les deux systèmes de RAP (modélisation de la durée ou non) sont complémentaires. Pour chaque langue, nous avons combiné le modèle acoustique CD-HMM/DNN ne distinguant pas la longueur vocalique des unités phonétiques avec le modèle modélisant cette distinction. Ces résultats sont présentés dans le tableau 5.4. La colonne *Gain* montre que comparé au meilleur WER, la combinaison permet de gagner jusqu'à 0,3% pour le système de RAP du haoussa et 0,9 % sur chaque corpus d'évaluation du wolof.

Pour conclure cette section sur la modélisation de la durée des voyelles au sein du système de RAP, l'étiquetage des voyelles n'augmente pas les performances du système de RAP. Néanmoins, cette modélisation apporte une information supplémentaire aux linguistes qui veulent annoter leur corpus (dans un scénario d'analyse de la parole). Grâce à l'alignement forcé par exemple, l'étiquetage des voyelles peut permettre d'effectuer des analyses linguistiques à grande échelle telles que des mesures phonétiques. C'est ce que nous allons voir dans les sections suivantes.

Tableau 5.4 – Performance des systèmes de RAP du haoussa et du wolof, en combinant le système sans modélisation des longueurs vocaliques avec le système modélisant les longueurs des voyelles. La colonne "Gain" montre la réduction absolue du taux d'erreur de mots par rapport au meilleur des deux systèmes.

Langue	**WER (Gain) (%)**	
	dev	**test**
Haoussa	7,8 (*0,1*)	10,3 (*0,3*)
Wolof (nettoyé)	18,9 (*0,9*)	23,0 (*0,9*)

5.4 Sélection de paramètres pour une mesure fine du contraste de durée des voyelles



Graphique 5.4 – Histogramme et distribution gamma de la voyelle /a/ en wolof lu - fort contraste

L'analyse objective du degré de bimodalité de la distribution de durée d'un voyelle n'est pas simple. Une des raisons est que - pour certaines voyelles - il peut y avoir beaucoup plus d'occurrences courtes que de longues (Sauvageot, 1965) : une distribution simple de la fréquence de la voyelle ne permet pas de faire émerger le contraste de longueur. Comme l'ont montré Bion *et al.* (2013) dans leurs travaux sur le japonais, regarder les distributions est une possibilité, mais des caractéristiques plus objectives sont nécessaires si nous voulons une évaluation fine du degré de contraste entre différents styles de discours et dialectes.

Cette section propose différents critères pour estimer le degré de bimodalité de la distribution des durées d'une voyelle donnée. Ces caractéristiques ne sont pas extraites des vraies distributions des voyelles courtes et longues, mais de leurs approximations gamma normalisées. Nous avons préféré les distributions gamma aux gaussiennes pour leur asymétrie. En effet, la loi gamma est particulièrement adaptée pour modéliser des données qui évoluent dans le temps (représentation de réels positifs uniquement avec $x \in [0, +\infty[$).

Cette étude ne porte que sur les distributions de durées des voyelles du wolof.

Nous définissons la distribution des voyelles étiquetées « brèves » (_S) par $d_S(x)$ et la distribution des voyelles étiquetées « longues » (_L) par $d_L(x)$. À titre d'exemple, la distribution des durées de la voyelle /i/ étiquetée « brève » est définie par $d_{/i/}(x)$ et la distribution des durées de la voyelle /i/ étiquetée « longue » est définie par $d_{/ii/}(x)$).

Conformément à cette définition, r_1 est défini par l'équation 5.1 et représente le ratio entre $d_S(a)$ et $d_L(a)$, avec a représentant le maximum de $d_S(x)$. Plus la valeur de r_1 est élevée, plus l'écart entre les deux distributions est élevé, au maximum de la distribution $d_S(x)$.
Si r_1 est supérieur à 1, cela signifie que le nombre de valeurs au point $d_S(a)$ est supérieur au nombre de valeurs relevées au point $d_L(a)$. Autrement dit, il y a plus de voyelles étiquetées « brèves » dans la classe de valeurs dans laquelle se situe $d_S(a)$ que de voyelles étiquetées « longues ». Au contraire, si r_1 est inférieur à 1, cela signifie que la distribution $d_L(a)$ contient plus de valeurs que la distribution $d_S(a)$; donc qu'il y a plus de voyelles étiquetées « longues » dans l'intervalle de valeur dans lequel se situe $d_S(a)$ que de voyelles étiquetées « brèves ».

De la même façon, r_2, défini par l'équation 5.2, est le ratio entre $d_L(b)$ et $d_S(b)$, lorsque b représente le maximum de $d_L(x)$. La signification des valeurs de r_2 est la même que cité précédemment pour r_1 mise à part que r_2 permet de relever, au maximum de $d_L(x)$, l'amplitude entre la distribution des voyelles notées « longues » par rapport à celles notées « brèves ».

$$r_1 = \frac{d_S(a)}{d_L(a)} \tag{5.1}$$

où $a = \arg\max_x (d_S(x))$.

$$r_2 = \frac{d_L(b)}{d_S(b)} \tag{5.2}$$

où $b = \arg\max_x (d_L(x))$.

$\mathscr{A}$ correspond à l'aire calculée entre les deux courbes lorsque $d_S(x) < d_L(x)$, comme indiqué dans l'équation 5.3. Plus l'aire est importante, plus le contraste de durée devrait être fort.

$$\mathscr{A} = \int_I^{\infty} d_L(x) - d_S(x)\, \mathrm{d}x \tag{5.3}$$

Nous avons également calculé le paramètre Δ qui représente la différence entre les deux modes [6] de $d_S(x)$ et $d_L(x)$, comme défini par l'équation 5.4. Plus la valeur de Δ est élevée, plus le contraste est important.

$$\Delta = \arg\max_x (d_L(x)) - \arg\max_x (d_S(x)) \tag{5.4}$$

Le graphique 5.4 illustre les histogrammes de durée ainsi que les notations relatives aux courbes gamma associées. Cette figure prend pour exemple les distributions statistiques calculées pour le phonème /a/. Nous avons observé un contraste fort pour cette voyelle dans notre corpus de wolof lu.

En rassemblant les paramètres précédemment cités, nous pouvons classer les voyelles (selon le degré d'aperture et le degré d'antériorité) dans un tableau et relever les contrastes pour chacune d'elles.

Enfin, il est important de noter que nous n'avons pas utilisé le Dip test de Hartigan (Hartigan et Hartigan, 1985) pour mesurer l'unimodalité d'une distribution, car nos mesures préliminaires ont montré que ce test conclut toujours à la bimodalité de notre distribution - même pour des contrastes extrêmement faibles.

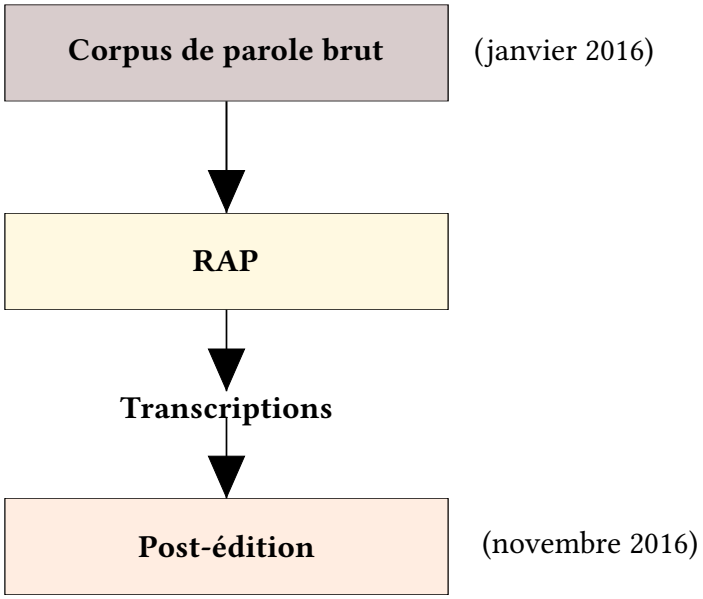
5.5 Analyse du contraste de durée des voyelles en wolof

Dans cette section, nous avons analysé différents types de discours et dialectes.

Pour cela, plusieurs corpus ont été utilisés :

- le corpus nettoyé de parole lue *dev*, détaillé au chapitre 4 (tableau 4.2), qui nous a servi à évaluer nos systèmes de RAP du wolof ;
- un corpus de parole semi-dirigée, en wolof et en faana-faana, collecté en janvier 2016 sur le terrain ;

Contrairement au corpus de parole lue, nous ne possédions pas de transcriptions du corpus de parole semi-dirigée. De ce fait, nous avons utilisé notre meilleur système de RAP CD-HMM/DNN du wolof lu pour les obtenir. Néanmoins, des mesures préliminaires nous ont montré que les sorties brutes du système étaient trop bruitées pour être utilisées directement, compte tenu de la nature plus spontanée des données. De ce fait, nous avons profité d'un nouveau terrain, en novembre 2016, pour faire corriger ces transcriptions automatiques par deux linguistes. Cette méthodologie est illustrée sur le graphique 5.5.



Graphique 5.5 – Schéma des corpus utilisés.

5.5.1 Corpus utilisés

Le corpus de parole brut

En plus de notre corpus de wolof lu collecté en début de thèse (Gauthier, Besacier, Voisin *et al.* 2016), Sylvie Voisin a collecté d'autres enregistrements durant un terrain qui s'est déroulé en janvier 2016. La collecte a porté sur l'élicitation de parole en wolof *standard*[7] et dans deux variantes dialectales : le faana-faana parlé dans la région du Saloum (Kaolack), et la variante lébou parlée à Ouakam.

22 locuteurs différents ont été enregistrés (6 locuteurs de faana-faana, 2 locuteurs de lébou, 3 locuteurs de wolof *urbain* and 11 locuteurs de wolof *standard*), permettant ainsi de collecter 1h30 de parole semi-dirigée. Chaque locuteur devait regarder une série de 76 clips vidéos conçus dans le but de faire émerger le concept de trajectoire dans une langue (ces vidéos ont été mentionnées au chapitre 3). Nous qualifierons ces enregistrements comme étant de la parole semi-spontanée. Les analyses présentées dans les sections suivantes proviendront de ces corpus oraux.

Utilisation de la RAP pour l'obtention de transcriptions

Nous avons utilisé notre meilleur système de RAP CD-HMM/DNN du wolof pour décoder ces nouveaux enregistrements. Chaque locuteur a été enregistré en continu, car la collecte a été réalisée avec un dictaphone (Lig-Aikuma n'était pas encore disponible en version stable). Nous avons alors, dans un premier temps, post-traité les enregistrements pour éliminer le bruit qui pouvait exister comme les commentaires du linguiste, les questions du locuteur ou encore

6. Le mode représente, en statistiques, la valeur la plus fréquente d'une distribution.
7. Le wolof dit *standard* désigne la langue parlée par des locuteurs natifs, à Dakar. Pour cette raison, il est parfois appelé « wolof de Dakar ». Il est souvent comparé au wolof dit *urbain*, parlé à Dakar par des locuteurs dont le wolof n'est pas la langue maternelle.

les plages de silence. Puis, nous avons segmenté ces enregistrements afin d'obtenir plusieurs fichiers audio de plus courte durée. Pour ce faire, nous avons utilisé un outil développé dans l'équipe, appelé *WavAutoSegmentor* [8]. Cet outil implémente un algorithme de détection d'activité sonore basé sur les changements du niveau d'énergie. Ceci nous a permis de découper automatiquement les enregistrements originaux au moment des pauses et ainsi éviter la troncation d'un mot. Parfois, la segmentation s'est réalisée au milieu d'une phrase, ce qui aurait pu entraîner un mauvais décodage du système de RAP. Mais un réglage du niveau d'énergie, au cas par cas, nous a permis de limiter ce phénomène. Enfin, nous avons décodé chaque segment audio avec notre meilleur système de RAP pour avoir des transcriptions.

Post-édition manuelle des transcriptions automatiques

En novembre 2016, Sylvie et moi-même sommes parties au Sénégal pour recueillir de nouveaux enregistrements de lébou et de faana-faana au moyen de Lig-Aikuma. Nous avons pu tester l'application mobile en condition réelle. Ainsi, nous avons utilisé le même protocole que lors de la collecte précédente,— effectuée par Sylvie seule —, pour l'élicitation de parole : nous avons utilisé le mode « Élicitation à partir de vidéos » pour faire visionner, à chaque locuteur, la série de 76 clips permettant l'émergence du concept de trajectoire dans la langue. Au total, nous avons enregistré 10 locuteurs, répartis ainsi : 3 locuteurs de faana-faana (variante de Kaolack), 4 locuteurs de lébou (variantes de Ouakam et de Yénn) et 3 locuteurs de wolof *urbain*. Ces enregistrements représentent environ 1h30 de signal audio.

Aussi, durant ce terrain, nous avons fait écouter et corriger les transcriptions automatiquement générées, par deux linguistes : un locuteur natif de wolof et un locuteur natif de faana-faana. Ces séances de post-édition nous ont permis de réaliser 5 transcriptions de faana-faana (sur 6) et 3 transcriptions de wolof *standard* (sur 11) [9]. Par la suite, nous avons forcé l'alignement de ces transcriptions exactes avec notre modèle acoustique CD-HMM/DNN de wolof lu. Les analyses de parole semi-spontanée présentées dans la suite de ce travail proviennent des durées estimées par le système.

Résumé des corpus sur lesquels porte notre analyse

Le tableau 5.5 récapitule les corpus de parole sur lesquels nous allons mesurer le contraste de longueur de voyelles dans les sections suivantes. Le corpus de wolof lu correspond à notre corpus *dev*.

Nous comparerons, dans un premier temps, le wolof *standard* lu (corpus *dev*) au wolof *standard* semi-spontané, collecté en janvier 2016. Dans un second temps, nous comparerons le wolof *standard* semi-spontané à la variante faana-faana enregistrée dans les mêmes conditions d'élicitation collectés en janvier 2016.

8. https://github.com/FredericAMAN/WavAutoSegmentor

9. Nous n'avons pas pu faire vérifier les transcriptions automatiques des enregistrements en wolof *urbain* ou celles des enregistrements en lébou.

Tableau 5.5 – Nouveaux corpus de parole du wolof (*standard* et variantes régionales).

Corpus	#Homme	#Femme	#Phrase	#Mot	Durée
Wolof (lu)	1	1	1 120	10 461	1h12 min
Wolof (semi-spontané)	2	1	254	2 825	14 min
Faana-Faana (semi-spontané)	5	0	454	3 365	19 min

5.5.2 Analyse du contraste en parole lue

5.5.2.1 Alignement forcé de transcriptions manuelles

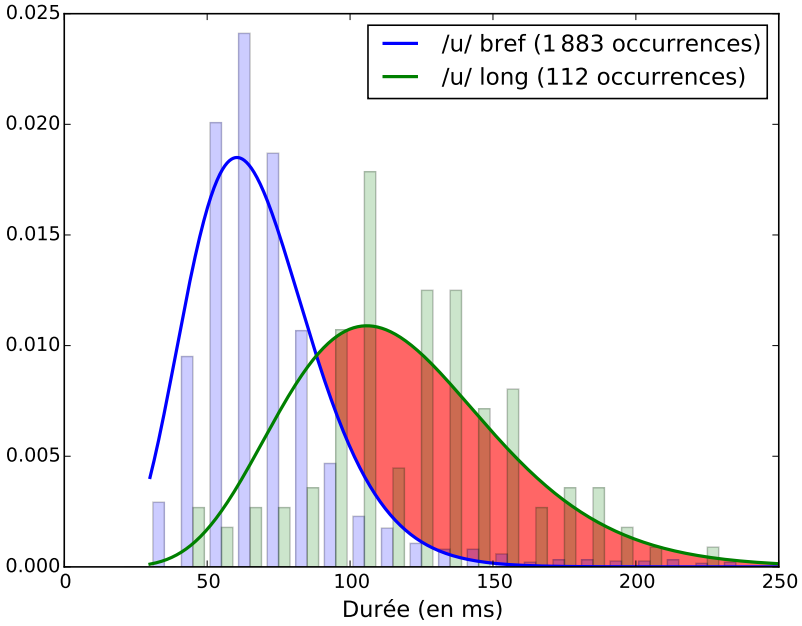
Dans un premier temps, nous avons extrait les durées en forçant l'alignement des transcriptions de nos corpus de développement (*dev*). L'alignement forcé a été effectué avec notre modèle acoustique CD-HMM/DNN du wolof prenant en compte l'opposition de longueur sur la totalité des voyelles.

Les données sont partitionnées dans différents ensembles désignés par $\mathscr{D}_l^v$ où $v \in \mathscr{V} = \{i, e, \varepsilon, a, \mathfrak{o}, o, u\}$ correspond à la voyelle étudiée et $l \in \mathscr{L} = \{S, L\}$ correspond à la durée attendue de la voyelle (« S » pour *Short* ou « L » pour *Long*). Nous avons calculé les durées de chaque voyelle et dessiné leur histogramme après la suppression des valeurs aberrantes (nous avons conservé les observations x comprises entre $\mu - 3\sigma < x < \mu + 3\sigma$). Nous avons également approximé nos distributions réelles par la densité de probabilité d'une distribution gamma.

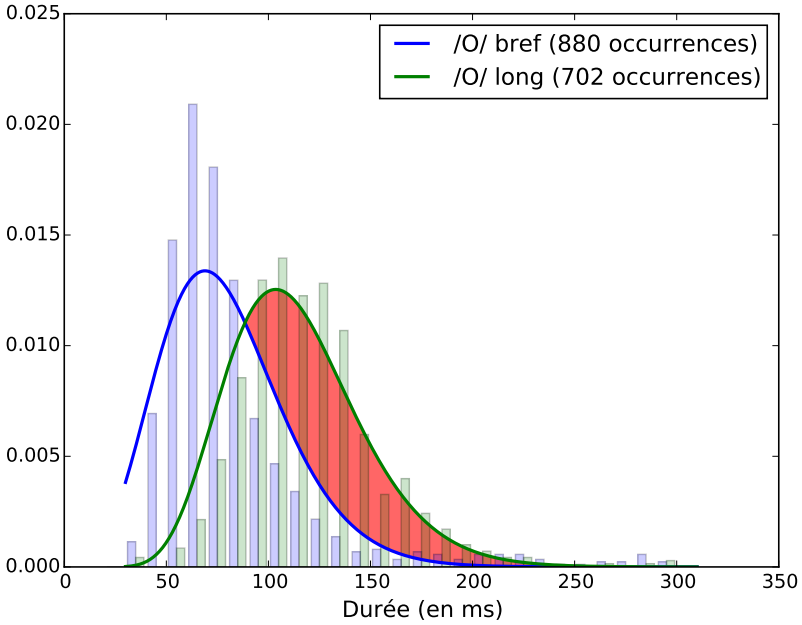
Les distributions normalisées pour chaque voyelle confirment que la bimodalité existe pour chacune d'elles. Cependant, le degré de contraste diffère pour chaque voyelle. Par exemple, un fort contraste de durée est observé pour la voyelle /u/ (graphique 5.6) tandis qu'un contraste faible est observé pour la voyelle /ɔ/ (graphique 5.7).

Le tableau 5.6 montre les mesures de l'opposition de longueur. Les voyelles sont triées selon leur degré d'aperture. En plus des caractéristiques de contraste décrites en section 5.4, nous avons également affiché, dans la troisième colonne, la durée moyenne μ (en ms) de chaque réalisation (brève ou longue) du phonème. La voyelle /a/ est celle qui apparaît le plus fréquemment (à la fois dans sa réalisation brève et longue) tandis que la voyelle /o/ est celle qui apparaît le plus rarement. Ceci peut facilement être expliqué par le fait que la voyelle /a/ apparaît plus fréquemment dans le vocabulaire wolof que la voyelle /o/. Nous pouvons aussi observer que les voyelles étiquetées brèves sont bien plus fréquentes que les voyelles étiquetées longues (sauf pour /o/ pour laquelle les proportions sont quasiment équivalentes).

Nous observons que la caractéristique articulatoire qui semble affecter la durée de la voyelle est l'aperture. Sock (1983) avait décrit, pour le wolof de Gambie, que la durée moyenne des voyelles augmentait avec l'ouverture de la mâchoire. Sur notre corpus, cette règle s'observe en effet sur les voyelles brève postérieures (/ɔ/, /o/ et /u/) mais pas sur les autres. Elle ne s'observe pas non plus sur les voyelles longues. Le tableau 5.6 montre aussi que les voyelles brèves d'un même niveau d'aperture (/i/~/u/, /e/~/o/, /ɛ/~/ɔ/) ont des durées moyennes semblables. Cela



Graphique 5.6 – Histogramme et distribution gamma de la voyelle /u/ en wolof lu (corpus *dev*) - fort contraste



Graphique 5.7 – Histogramme et distribution gamma de la voyelle /ɔ/ en wolof lu (corpus *dev*) - faible contraste

peut être lié au fait que ces paires entretiennent entre elles une relation phonologique qualifiée d'opposition équipollente [10].

Globalement, Δ varie entre 34 ms et 52 ms et $\mathscr{A}$ entre 0,39 et 0,67. Nous observons que le contraste de longueur est plus important pour la voyelle /a/ que pour la voyelle /i/ et /u/,

10. Se dit de deux phonèmes qui possèdent un ou plusieurs traits en commun (Saussure, 1952, Définition 8, page 31)

Tableau 5.6 – Paramètres de contraste de durée extraits de parole lue en wolof (corpus *dev*).

Phonème bref long	**#Occurrence**	μ (en ms)	$\mathbf{r_1}$	$\mathbf{r_2}$	$\mathscr{A}$	Δ (en ms)
/i/ **/i:/**	2 157 131	86 133	**2,02**	**1,52**	**0,39**	**47**
/e/ **/e:/**	224 179	75 115	**3,81**	**2,03**	**0,53**	**35**
/ɛ/ **/ɛ:/**	1 263 556	78 121	**2,74**	**1,63**	**0,46**	**40**
/a/ **/a:/**	4 678 883	69 125	**10,95**	**4,31**	**0,67**	**52**
/ɔ/ **/ɔ:/**	880 702	78 114	**2,40**	**1,62**	**0,43**	**34**
/o/ **/o:/**	60 67	74 114	**3,87**	**2,08**	**0,55**	**34**
/u/ **/u:/**	1 883 112	67 117	**5,13**	**2,80**	**0,60**	**45**

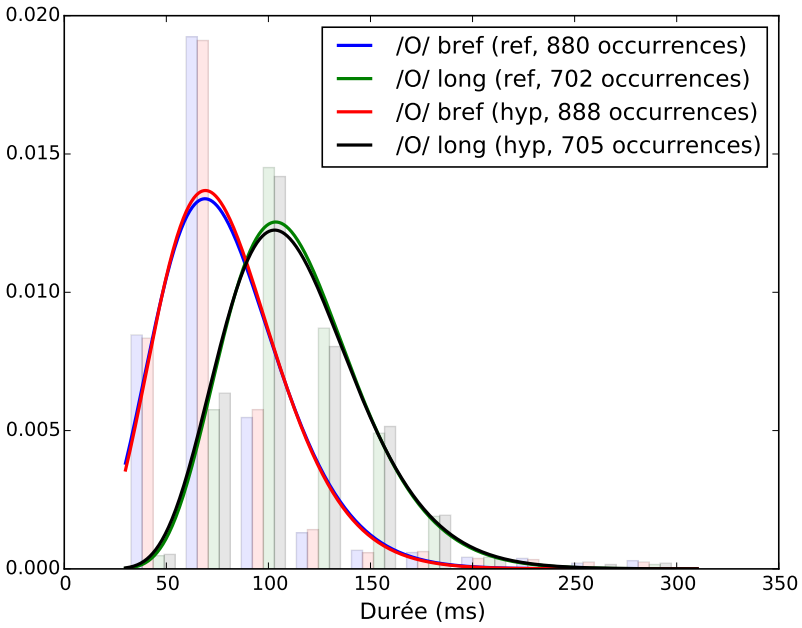
comme décrit par Sock (1983) en ce qui concerne le wolof de Gambie. La voyelle /a/ présente des ratios r_1 et r_2 très élevés, de même qu'une valeur importante de l'aire $\mathscr{A}$ et un grand Δ. Les voyelles mi-ouvertes /ɛ/ et /ɔ/ sont celles pour lesquelles l'opposition est la moins observée, avec des ratios r_1 et r_2 moins élevés que les autres voyelles, une aire $\mathscr{A}$ et un Δ modérés. Étonnamment, la voyelle /i/ possède un Δ fort au regard des autres paramètres (r_1, r_2 et $\mathscr{A}$) plus bas que toutes les autres voyelles.

Finalement, lorsque nous nous concentrons sur les valeurs des paramètres présentés dans le tableau, nous pouvons relever que les voyelles d'un même niveau d'aperture (/i/~/u/, /e/~/o/, /ɛ/~/ɔ/) montrent des valeurs très similaires. Aussi, ces paramètres révèlent que toutes les voyelles sont contrastées en longueur. Les histogrammes et distributions gamma associées sont fournis en annexe M. Le tableau 5.6 montre enfin que les caractéristiques de contraste sont corrélées, mais qu'elles sont aussi complémentaires pour décrire la forme des distributions de durée des voyelles.

Pour conclure sur cette sous-sous-section, cette analyse (réalisée de manière entièrement automatique sur presque 14k voyelles) montre que les caractéristiques proposées peuvent mettre en évidence différents degrés de contraste pour chaque voyelle considérée et confirmer — à plus grande échelle — les analyses précédentes.

5.5.2.2 Alignement forcé de transcriptions automatiques (avec RAP)

Dans cette sous-sous-section, nous essayons de voir si les transcriptions manuelles peuvent être remplacées par des hypothèses émises par le système de RAP tout en conservant les mêmes tendances/conclusions. Dans ce cas, nous relâchons la contrainte d'avoir une transcription manuelle de l'ensemble de données. Nous avons calculé les durées de voyelles à partir de l'alignement forcé obtenu avec des transcriptions de la RAP (produites avec notre système de RAP de



Graphique 5.8 – Histogramme et distribution gamma de la voyelle /ɔ/ en wolof lu (corpus *dev*) - en utilisant la transcription manuelle (ref) ou la transcription automatique (hyp)

référence du wolof pour lequel nous obtenons des performances atteignant 20% de WER sur de la parole lue) et calculé les distributions gamma comme indiqué dans la section 5.5.2.1 précédente. Pour chaque voyelle, nous avons comparé les deux distributions (transcription manuelle *vs* transcription automatique) en utilisant le test statistique de Kolmogorov-Smirnov (Massey Jr, 1951) [11]. L'hypothèse nulle H_0 est que les distributions calculées à partir des transcriptions manuelles et celles calculées à partir des transcriptions émises par le système de RAP sont similaires. Pour chaque voyelle v, aucune différence significative n'a été relevée (p-value $p > 0.1$). Afin d'illustrer ce résultat, le graphique 5.8 montre les histogrammes de durées ainsi que leur courbe gamma associées, pour la voyelle /u/, lorsque nous avons forcé l'alignement des transcriptions manuelles et automatiques. Les courbes sont semblables, ce qui confirme que, pour la parole lue, le besoin de transcriptions manuelles peut être assoupli puisque l'utilisation de la RAP conduit à des mesures très similaires et aux mêmes conclusions. Les autres courbes sont jointes en annexe N. Pour les sous-sections suivantes (analyse du discours semi-spontané), la RAP sera également utilisée pour produire des transcriptions. Cependant, comme annoncé au début de cette section, nous les avons faites corriger par un expert linguiste, en raison de la nature plus spontanée des données [12].

5.5.3 Analyse du contraste en parole semi-spontanée

Pour comparer la durée de la voyelle en contexte de parole lue à la production en contexte plus spontané, nous avons pris les transcriptions de nos corpus *dev* et les avons comparées

11. Ce test permet de vérifier si deux échantillons suivent une même loi.
12. Des mesures préliminaires nous ont montré que les transcriptions de la RAP sur la parole spontanée sont trop bruitées pour être utilisées directement. Nous avons obtenu environ 31% de WER pour le wolof et 66% de WER pour le faana-faana.

aux durées des transcriptions automatiques que notre système de RAP a produit. Comme mentionné au début de cette section, nous avons collecté 11 enregistrements en Wolof standard, mais seulement 3 de leurs transcriptions produites par la RAP ont pu être vérifiées par un expert. Cela représente 14 minutes de signal audio. En ce qui concerne la parole semi-spontanée, le test de Kolmogorov-Smirnov a révélé que les hypothèses du système de RAP dont nous avons forcé l'alignement sont significativement différentes des alignements forcés des transcriptions manuelles ($p < 0.01$). Nos analyses suivantes seront, donc, produites en utilisant un alignement forcé des transcriptions manuelles.

Tableau 5.7 – Paramètres de contraste de durée extraits de parole semi-spontanée en wolof.

Phonème bref long	**#Occurrence**	μ (en ms)	$\mathbf{r_1}$	$\mathbf{r_2}$	$\mathscr{A}$	Δ (en ms)
/i/ **/iː/**	863 84	77 94	**1,04**	**1,19**	**0,18**	**18**
/e/ **/eː/**	77 97	70 98	**2,88**	**1,39**	**0,47**	**23**
/ɛ/ **/ɛː/**	323 50	61 93	**2,51**	**1,30**	**0,43**	**28**
/a/ **/aː/**	908 136	60 110	**41,77**	**5,18**	**0,71**	**48**
/ɔ/ **/ɔː/**	162 77	75 94	**1,37**	**1,23**	**0,28**	**18**
/o/ **/oː/**	37 70	53 112	**N/C**	**N/C**	**N/C**	**55**
/u/ **/uː/**	419 15	58 104	**N/C**	**4,69**	**0,71**	**45**

Nous avons calculé, pour le corpus de parole semi-spontanée en wolof, les mêmes paramètres que présentés précédemment. Le tableau 5.7 les illustre et les histogrammes accompagnés des distributions gamma sont fournis en annexe O. Nous observons ainsi, pour la parole semi-spontanée, qu'il y a beaucoup plus de voyelles courtes que de longues, sauf pour /e/ où les proportions sont quasiment similaires, mais où nous pouvons observer un plus grand effectif de voyelles longues (tout comme pour /o/). Aussi, nous ne possédons pas assez de voyelles /o/ dans leur version brève et /u/ dans leur version longues pour effectuer un calcul des valeurs à partir de leurs distributions statistiques.

En observant la durée moyenne μ des voyelles, nous remarquons que cette valeur est moins élevée que sur le corpus de parole lue (pour chaque réalisation brève ou longue). Autrement dit, les voyelles, prononcées en contexte semi-spontané, ont tendance à être raccourcies, comparé à leur réalisation en contexte lu. Ces conclusions étaient attendues, mais elles confirment que notre méthodologie assistée par la machine permet des mesures utilisables à plus grande échelle. Lorsque nous comparons μ en contexte lu et en contexte semi-spontané, nous observons que les voyelles longues sont plus affectées par le style de parole, surtout les voyelles antérieures (/iː/, /eː/ et /ɛː/) ainsi que la voyelle postérieure semi-ouverte /ɔː/, tandis que les ver-

sions brèves des voyelles sont les moins impactées, à part pour la voyelle postérieure fermée /o/.

Les résultats pour la voyelle /u/ doivent être considérés avec prudence, car ils proviennent de seulement 15 longues occurrences, tout comme pour la voyelle /o/ pour laquelle nous n'avons que 37 occurrences brèves. Les paramètres calculés montrent que l'opposition de longueur est nettement réduite en contexte semi-spontané pour les paires /i/~/iː/ et /ɔ~/ɔː/, en comparaison à ce que nous avons observé sur leur production en contexte lu. Les ratios sont fortement réduits (respectivement, r_1 perd 1,02 et 1,03 ; r_2 perd 0,33 et 0,39), tout comme leur aire sous la courbe $d_L(x)$ ($\mathscr{A}$ perd 21% pour la voyelle /i/ et 15% pour la voyelle /ɔ/) et leur delta (Δ perd 29 ms pour /i/ et 16 ms pour /ɔ/). Finalement, la voyelle la moins impactée est la voyelle /a/, pour laquelle le contraste est toujours fortement marqué.

Concernant l'analyse de ces tendances observées, il ne faut pas oublier que le contenu lexical dans le corpus de lecture (constitué d'un ensemble d'histoires, de chansons, de débats, de proverbes) diffère de celui des corpus spontanés (exprimant la trajectoire, idiolecte du locuteur). Par conséquent, d'autres analyses sont nécessaires pour expliquer ces résultats. Malheureusement, aucune étude n'a été menée sur le discours spontané en wolof pour le moment, de sorte que ces observations ne peuvent être comparées à des travaux précédents.

5.5.4 Analyse du contraste sur une variante dialectale en parole semi-spontanée

Pour aligner les transcriptions en faana-faana, nous avons utilisé notre modèle acoustique CD-HMM/DNN appris sur le corpus de wolof *standard* lu, comme pour les autres alignements forcés.

Tableau 5.8 – Paramètres de contraste de durée extraits de parole semi-spontanée en faana-faana.

Phonème bref long	**#Occurrence**	μ (en ms)	$\mathbf{r_1}$	$\mathbf{r_2}$	$\mathscr{A}$	Δ (en ms)
/i/ **/iː/**	887 170	81 75	**0,72**	**1,39**	**-0,09**	**-1**
/e/ **/eː/**	79 114	80 93	**1,07**	**1,28**	**0,22**	**14**
/ɛ/ **/ɛː/**	200 178	71 93	**1,43**	**0,99**	**0,24**	**19**
/a/ **/aː/**	927 186	68 101	**1,85**	**1,21**	**0,35**	**30**
/ɔ/ **/ɔː/**	195 113	67 76	**1,32**	**0,81**	**0,13**	**6**
/o/ **/oː/**	25 50	61 88	**5,93**	**3,52**	**0,62**	**23**
/u/ **/uː/**	331 2	60 100	**N/C**	**N/C**	**N/C**	**N/C**

Le tableau 5.8 montre les valeurs calculées pour les mêmes caractéristiques que celles présentées précédemment pour le wolof lu (tableau 5.6) et le wolof semi-spontané (tableau 5.7), pour notre corpus semi-spontané en faana-faana. Les histogrammes accompagnés des distributions gamma sont rapportés en annexe P. Les caractéristiques de la voyelle /u/ n'ont pas pu être calculées en raison du manque d'occurrences de la voyelle dans sa version longue qui empêche de calculer sa distribution.

Tout comme pour le corpus de parole en wolof semi-spontané, nous observons dans le tableau 5.8 que les voyelles longues /eː/ et /oː/ apparaissent toujours plus fréquemment que leur version brève.

Ce tableau montre également que la seule voyelle pour laquelle nous observons un contraste évident est /a/, comme observé sur les corpus de wolof lu et semi-dirigé. Pour /o/, nous observons aussi un fort contraste (comme en wolof *standard* semi-spontané) ; néanmoins, cette évaluation est faite sur un nombre bien moins élevé d'occurrences.

Les valeurs négatives de $\mathscr{A}$ et de Δ, calculées pour la voyelle antérieure fermée /i/, montrent que les distributions des voyelles étiquetées « longues » sont inférieures à celles étiquetées « brèves » ou se superposent.

En regardant la valeur des paramètres calculés dans le tableau, nous notons que la distinction entre brèves et longues est faible, contrairement au wolof *standard* semi-dirigé. Notamment, l'opposition de longueur sur la voyelle /ɔ/ est fortement réduite ; la distribution des durées pour la voyelle /i/ montre des temps de prononciation similaires, quelque soit la nature de sa réalisation (brève ou longue).

Ces résultats ne permettent pas de démontrer qu'il existe en faana-faana une forte opposition de la longueur des voyelles, comme nous l'avons observé en wolof *standard*. Néanmoins, nous ne pouvons pas non plus affirmer que le contraste de longueur vocalique n'existe pas en faana-faana. Dans les descriptions phonologiques de cette variante, comme dans le wolof gambien, l'opposition brève/longue est décrite ; nous pouvons, donc, supposer que les différences dialectales en wolof ne sont pas fondées sur ce manque de contraste. En tout cas, le test de Kolmogorov-Smirnov à deux échantillons a révélé que les distributions des durées des voyelles /e/ et /eː/, /ɛː/, /u/ et /uː/ ainsi que /oː/ du corpus semi-spontané en faana-faana n'étaient pas significativement différentes du corpus semi-spontané en wolof, alors que les distributions des autres voyelles (/i/ et /iː/, /ɛ/, /a/ et /aː/, /ɔ/ et /ɔː/) l'étaient. Enfin, comme cette variante a été peu étudiée, nous espérons que notre analyse représente une première étape dans l'étude du contraste phonémique dans les dialectes wolofs.

5.6 Résumé

Nous avons proposé, dans ce chapitre, de modéliser l'opposition de longueur existant au sein des voyelles pour optimiser la RAP pour deux langues d'Afrique subsaharienne : le haoussa et le wolof. Cette prise en compte de l'opposition de longueur par les systèmes est importante pour la bonne reconnaissance du mot, car la durée de prononciation d'une voyelle, à elle seule, peut affecter le sens du mot.

Pour l'analyse du haoussa, nous avons procédé à un étiquetage *a posteriori* des sorties de l'alignement forcé du système, avant d'entraîner un nouveau système de RAP, car la durée de la voyelle n'est pas facilement identifiable dans cette langue : non seulement elle n'est pas marquée dans l'orthographe, mais en plus elle peut varier selon plusieurs facteurs (position de la voyelle dans la syllabe, dans le mot, dans le groupe syntagmatique, ...). Dans l'étude présentée, nous nous sommes concentrés sur la réalisation de la voyelle dans son contexte syllabique. Procéder ainsi nous a permis de vérifier que notre étiquetage selon le contexte syllabique était pertinent pour découvrir l'existence de l'opposition de longueur vocalique dans la langue.

De ce fait, nous avons entraîné un nouveau système de RAP du haoussa qui tient compte, dans le modèle acoustique, de l'opposition de longueur pour toutes les voyelles. Les performances ont été très légèrement améliorées, mais seulement pour l'un des deux corpus d'évaluation, par rapport au système de référence. Nous avons alors décidé d'entraîner un nouveau système de RAP à partir d'une modélisation de la durée des voyelles, mais pour un sous-ensemble uniquement. Ce sous-ensemble a été sélectionné à partir des analyses statistiques évoquées précédemment : nous avons choisi les voyelles du haoussa pour lesquelles le contraste était le plus probant (à savoir, les voyelles /e/ et /o/). Ce mode opératoire a permis d'améliorer les performances du système de RAP du haoussa pour chaque corpus d'évaluation. Les résultats de ces systèmes de RAP sont rappelés dans le tableau 5.9.

Puisque cette étude préliminaire s'est avérée concluante pour l'optimisation du système de RAP du haoussa et pour la mise en évidence de l'opposition de durée au sein d'un grand échantillon de voyelles, nous avons décidé d'appliquer la même méthodologie au wolof. Ainsi, nous avons directement (sans étiquetage *a posteriori* des alignements forcés du système de RAP de référence) fait la différence entre voyelles brèves et voyelles longues dans notre dictionnaire de prononciation du wolof, et avons lancé l'apprentissage d'un nouveau système de RAP. En fin de compte, la distinction des durées entre les unités vocaliques, dans notre modèle acoustique, n'a pas permis cette fois d'optimiser le système de RAP pour le wolof (voir le tableau 5.9). Néanmoins, les techniques d'alignement forcé — permises par la RAP —, a rendu possible le calcul du contraste de durée d'un volume important de voyelles en wolof.

Les travaux, portant sur l'analyse préliminaire du contraste de longueur des voyelles en haoussa et la modélisation de ce contraste dans nos systèmes de RAP pour les deux langues, ont été présentés lors du 5ème atelier international pour les langues peu dotées *SLTU* ("Spoken Language Technologies for Under-resourced languages") (Gauthier, Besacier et Voisin, 2016a).

Le tableau 5.9 rappelle les différentes performances des systèmes de RAP, pour chaque langue, selon la modélisation acoustique adoptée (sans modélisation de l'opposition de longueur, avec modélisation sur toutes les voyelles ou sur un sous-ensemble de voyelles). La colonne gain montre la différence de performance avec le système de référence (sans modélisation).

Tableau 5.9 – Résultats des systèmes de RAP CD-HMM/DNN avec modélisation de la longueur vocalique, pour le haoussa et le wolof – avec adaptation du locuteur.

Langue	Type de modélisation	WER (Gain) (%)	
		dev	test
Haoussa	sans modélisation	8,0	11,3
	avec modélisation sur toutes les voyelles	8,3 (-0,3)	11,2 (+0,1)
	avec modélisation sur un sous-ensemble de voyelles	7,9 (+0,1)	10,6 (+0,7)
Wolof (nettoyé)	sans modélisation	19,8	23,9
	avec modélisation sur toutes les voyelles	20,0 (-0,2)	24,6 (-0,7)
	avec modélisation sur un sous-ensemble de voyelles	20,0 (-0,2)	24,5 (-0,6)

L'analyse préliminaire de l'opposition de longueur au niveau des voyelles nous a montré que la bimodalité d'une distribution statistique n'était pas facilement décelable, à première vue, sur un histogramme. En effet, les voyelles courtes apparaissent bien plus fréquemment que les voyelles longues. À l'inverse, le *Dip test* de Hartigan — qui permet de déterminer l'unimodalité d'une distribution statistique —, s'est montré négatif sur toutes nos distributions. Afin d'approfondir cette étude sur le contraste de durée des voyelles, nous avons alors choisi de sélectionner des paramètres permettant l'analyse fine de la bimodalité d'une distribution de durées. Ces paramètres ont été extraits de la densité de probabilité de la loi gamma, approximée à partir de nos distributions de durées. Nous nous sommes appuyés sur ces paramètres pour comparer la durée de réalisation des voyelles dans différents contextes : discours lu *versus* discours semi-spontané, wolof *versus* faana-faana. Ces comparaisons ont montré que l'opposition de longueur vocalique existe en wolof et dans sa variante dialectale faana-faana, à divers degrés. De même, l'opposition de longueur perdure en discours plus spontané, même si la durée des voyelles longues est largement impactée, en comparaison des brèves dont la durée semble plus stable.

La méthode décrite dans ce chapitre de thèse, permettant de juger le degré de bimodalité des distributions de durées des voyelles, a fait l'objet d'une publication (Gauthier, Besacier et Voisin, 2017) dans les actes de la conférence Interspeech 2017 qui s'est déroulée à Stockholm. Toutefois, ce manuscrit présente de nouveaux résultats en raison d'une mise à jour de nos systèmes de RAP depuis la publication.

Pour conclure ce chapitre, la modélisation de la longueur des voyelles s'est révélée utile dans les deux langues puisque, d'une part, elle nous a permis de légèrement améliorer la performance de nos systèmes de référence pour le haoussa (les résultats sont rappelés dans le tableau 5.9). D'autre part, les modèles acoustiques générés nous ont permis d'effectuer des analyses phonétiques, de façon automatique, sur les productions véritables et actuelles de la durée des voyelles par les locuteurs en wolof et en faana-faana (langues pour lesquelles les études concernant la durée des voyelles sont peu nombreuses et anciennes).

Cette étude représente une analyse préliminaire de l'observation de la réalisation phonétique [13] de la longueur vocalique en wolof. Ce travail vient enrichir les analyses phonétiques en wolof, dont il manque encore de nombreuses études. Par ailleurs, Cissé (2006) soulignait que « la différenciation entre voyelles brèves et voyelles longues est un point parmi bien d'autres sur lesquels les résultats de l'analyse phonétique sont urgents. ». Nous avons réalisé une preuve de concept qui a montré que des analyses quantitatives, d'une granularité très fine, sont réalisables au moyen de la machine. D'une part, les résultats, produits automatiquement, se sont avérés corrects, car similaires aux descriptions de la littérature. D'autre part, ils ont permis de nouvelles découvertes sur des styles de parole jusqu'alors non étudiés et sur des variantes dialectales.

Pour aller plus loin dans les analyses phonétiques proposées, il pourrait être intéressant de calculer les caractéristiques extraites de distributions plus fines.

À ce titre, effectuer une analyse de la durée des voyelles en fonction de leur contexte consonantique adjacent permettrait de compléter les études menées par Calvet (1966) sur le wolof et, de manière générale, les études sur la réalisation des unités phonémiques dans les langues. Ces mesures pourraient également porter sur les paires minimales, mais ce type d'observation se limite alors aux données disponibles (qui sont fortement restreintes pour les langues peu dotées).

Pour finir, l'opposition de longueur de consonnes est aussi décrite en wolof. L'investigation de ce trait et de son impact sur le décodage des mots seraient tout aussi intéressants à prendre en compte.

13. Par opposition à la théorie phonologique.

Conclusion

Contributions

Durant ces 3 années de recherche, nous avons finalement mis au point des outils de collecte innovants pour les linguistes de terrain. À travers une étroite collaboration, nous avons développé des méthodes pour l'aide à la collecte, à la transcription et à l'analyse des langues. Nous avons appliqué notre méthodologie au wolof, parlé au Sénégal, afin de :

1. Collecter de la parole
2. Créer des technologies vocales
3. Obtenir des transcriptions de façon automatique
4. Utiliser ces transcriptions afin d'effectuer des analyses phonétiques, selon plusieurs facteurs (style de parole et dialectes)

Dans le but d'aider le linguiste de terrain dans son travail de collecte de données, nous avons développé une application Android appelée Lig-Aikuma. Cette application est une extension d'Aikuma (Hanke et Bird, 2013) mais a été repensée pour la linguistique de terrain, en permettant de collecter de la parole lors des études sur le terrain.

Cette application est le résultat de nombreuses discussions au cours desquelles les linguistes ont fait part de leurs besoins et attentes concernant ce type d'outils mais également durant lesquelles les informaticiens ont été à l'écoute tout en indiquant les limites d'implémentation de certaines fonctionnalités (dans le but de conserver une application en adéquation avec les usages sur le terrain).

L'application a été testée avec succès sur le continent africain, et a permis de collecter — à ce jour — 6 langues dont 3 langues très peu dotées (260h de parole a été récolté au total). Nous avons personnellement testé l'application, durant un terrain au Sénégal. Différents types de discours (spontané, reparlé, traduit, lu, semi-dirigé) et deux variétés dialectales du wolof (faana-faana et lébou) ont pu être collectés.

Autre particularité des collectes effectuées avec Lig-Aikuma, le rassemblement de corpus parallèles « parole source (peu dotée)-parole cible (bien dotée) », « parole-image », « parole-vidéo » qui présentent un intérêt fort pour les technologies de la parole, notamment pour l'apprentissage non supervisé. Lig-Aikuma a permis de collecter rapidement et massivement des corpus de parole pour le bàsàá, le mbochi et le myéné (cette dernière est par ailleurs une langue menacée).

Dans l'optique de faire gagner du temps au linguiste, nous avons développé des systèmes de transcription. En effet, il faut compter de 3 heures à 6 heures de travail au linguiste (selon sa

maîtrise de la langue) pour transcrire 1 heure d'enregistrement de parole. L'automatisation de la tâche de transcription peut donc représenter un gain de temps considérable pour le linguiste.

La construction d'un système de transcription, pour les linguistes qui étudient des langues peu dotées, présente deux principaux défis : puisque ce système doit bénéficier au linguiste, il a fallu, d'une part, le rendre accessible et utilisable, même sans connaissances fortes en informatique ; d'autre part, les langues peu dotées sont, par définition, des langues pour lesquelles les ressources disponibles manquent cruellement.

Lorsque la langue n'est pas dotée d'un système d'écriture, le linguiste de terrain commence par transcrire phonétiquement la langue. En particulier, dans le cas des langues en danger, les linguistes ont besoin de transcriptions phonétiques puisque l'orthographe n'est pas disponible. Cependant, pour les langues dont l'alphabet a été établi, le linguiste de terrain transcrit aussi orthographiquement ces langues, et les transcriptions peuvent alors également servir à analyser la langue.

Le haoussa et le wolof, étudiés dans ces travaux, sont des langues dont l'orthographe est fixée officiellement. Par conséquent, nous avons considéré le développement de systèmes de transcription orthographique.
Alors que pour le haoussa nous avons utilisé les ressources du projet Globalphone, nous avons en revanche rassemblé nos propres données pour la construction du système de reconnaissance de la parole du wolof. Ce système de RAP à grand vocabulaire du wolof constitue une première pour cette langue. Ce système atteint un taux d'erreur de mots de 19%. Ce résultat reste élevé par rapport à d'autres langues peu dotées, mais il est fortement dû au modèle de langue qui a été estimé sur un très faible volume de données. Finalement, le développement d'un système de transcription a aussi été nécessaire pour obtenir des alignements automatiques « parole-texte » et pouvoir réaliser des analyses phonétiques.

Ainsi, grâce aux techniques d'alignement forcé permises par le développement de systèmes de RAP, nous avons pu effectuer des analyses phonétiques, de façon totalement automatique. Nous avons vérifié — sur la totalité des voyelles de nos corpus — que l'opposition de longueur décrite en haoussa et en wolof était concrètement produite par les locuteurs (de nos corpus). Ces analyses, bien que préliminaires, ont montré que la RAP peut être utilisée pour analyser des phénomènes subtils tel que le contraste de durée qui peut exister au niveau du phonème. De plus, nous avons utilisé les distributions statistiques des durées (fournies par l'alignement forcé) afin de sélectionner des paramètres qui servent à rapidement contrôler l'existence du contraste phonémique.

Pour conclure ce travail, nous avons développé des ressources et des solutions pour les langues peu dotées, durant ces 3 ans de thèse.

Nous avons répondu aux problématiques de départ, à savoir :

- Mettre en place des méthodes de collecte de données innovantes au moyen de Lig-Aikuma ;

- Adapter les outils informatiques pour un non spécialiste afin de lui faire gagner du temps en ayant :
  - créé une application mobile simple et intuitive ;
  - déployé une machine virtuelle pour la construction et l'utilisation d'un système de RAP de façon simple et rapide ;
- Permettre au linguiste de terrain de travailler à grande échelle au moyen de la RAP ;
- Améliorer un système de RAP grâce aux analyses linguistiques, en montrant qu'un étiquetage précis des voyelles peut optimiser un système de RAP tel que nous l'avons observé avec le système de RAP du haoussa.

Perspectives

Wolof

Du côté de la modélisation acoustique, nous avons montré la robustesse de nos modèles de parole lue en wolof, à travers des mesures comparant des alignements forcés avec leurs versions post-éditées manuellement. En effet, ces mesures ont mis en évidence que l'utilisation de transcriptions vérifiées peut être assouplie pour une analyse phonétique de la parole lue, lorsqu'elle relève d'une grande quantité de données. Néanmoins, sur de la parole spontanée, les transcriptions humaines sont encore à privilégier. De ce fait, les recherches concernant les approches non supervisées (Dunbar *et al.* 2017) pourraient être un moyen de relâcher cette contrainte et de permettre des analyses sur des annotations obtenues de façon complètement automatique, pour de la parole moins contrôlée.

Du côté de le modélisation linguistique, nous n'avons pas pu pousser les investigations pour améliorer le modèle de langue. Afin de l'optimiser, un effort pourrait être mené à la normalisation des textes en wolof. Un premier filtrage pourrait être effectué, comme étudié par Adda et Adda-Decker (1997) par exemple, puis des méthodes pourraient être appliquées pour pallier les différentes orthographes d'un même mot. Ceci restreindrait les choix possibles, lors de la construction du modèle de langue, ou encore, durant la phase de construction de l'arbre de décision. Cette méthode pourrait permettre d'obtenir une réduction du taux d'erreur sur les mots, au moment du décodage. Nous avions également pensé à la mise en place d'une plateforme collaborative afin de faire participer les locuteurs sur des choix orthographiques de variantes mais, même dans les titres d'émissions télévisées, de journaux, de noms de partis politiques ou encore sur les affiches publicitaires, l'orthographe officielle n'est pas toujours connue ou utilisée.

Nous avons démontré que des analyses phonétiques d'un phénomène linguistique subtil peuvent être réalisées, à grande échelle, avec l'aide de méthodes automatiques. Néanmoins, ces analyses restent préliminaires et pourraient être affinées en contextualisant les observations : par exemple, la prise en compte du contexte d'apparition des phonèmes pourrait me-

ner à des analyses de la durée combinatoire [14], plutôt qu'à des analyses de la durée inhérente comme nous l'avons fait. De même, observer l'opposition de longueur sur des paires minimales mènerait à des analyses encore plus précises de la réalisation de ce trait linguistique, mais cette approche est contrainte par la quantité de données disponibles. Enfin, une mesure de la charge fonctionnelle des voyelles pourrait être intéressante afin de compléter les analyses, comme l'ont fait Surendran et Niyogi (2003) ; Pellegrino, Marsico et Coupé (2012) ; Oh *et al.* (2013), dans le but d'évaluer à quel point l'information communiquée par le contraste vocalique, au sein d'une paire minimale, est importante et utile pour la compréhension du mot.

Enfin, du point de vue de l'analyse fonctionnelle de la RAP construite pour le wolof, il pourrait être pertinent d'envisager l'entraînement d'un modèle acoustique bilingue wolof-français. Puisque le français est la langue officielle au Sénégal, des emprunts réciproques aux deux langues existent, du fait de leur contact. Comme l'avait déjà remarqué Dumont (1983, chapitre 3), des vocables des deux langues se retrouvent, autant dans les documents écrits que dans le discours des locuteurs.

Revalorisation de corpus oraux existants

Avec l'application mobile Lig-Aikuma, d'autres utilisations peuvent être imaginées. Les modes « Répétition » et « Traduction » peuvent être utilisés par exemple, pour revaloriser voire revitaliser des corpus oraux existants. De nombreux enregistrements effectués avant l'ère du numérique, sur bande magnétique, sont conservés mais inexploités. Ces données ont une valeur inestimable et pourtant vont disparaître du fait de la déterioration des supports sur lesquels elles sont enregistrées, mais aussi parce qu'un jour, inévitablement, le matériel nécessaire à la lecture de ces supports n'existera — ou ne fonctionnera — plus.
Ces modes d'utilisation peuvent aussi être envisagés pour commenter oralement des enregistrements audio. À ce titre, Lig-Aikuma pourrait trouver un intérêt dans le domaine de l'ethnographie : en ethnomusicologie par exemple, qui étudie les musiques du monde et qui oeuvre également à la préservation et à la valorisation d'anciennes formes musicales, les scientifiques pourrait utiliser l'application afin d'annoter oralement leurs enregistrements de terrain.

Apprentissage non supervisé

Les collectes effectuées avec Lig-Aikuma participent aux recherches actuelles qui ont pour but de ne plus utiliser (ou peu) les transcriptions — orthographiques ou phonétiques — dans le processus de décodage d'une langue. À cet effet, le corpus de wolof collecté durant ces travaux de thèse a été utilisé pendant le *Zero Resource Speech Challenge* (Dunbar *et al.* 2017) et le corpus de mbochi collecté avec Lig-Aikuma a été exploité pour réaliser de la découverte de mots (Godard *et al.* 2016 ; Zanon Boito *et al.* 2017) mais aussi de la découverte d'unités linguistiques par le projet Rosetta (Scharenborg *et al.* 2018).
Ces méthodes peuvent ainsi apporter des solutions aux langues peu dotées voire, à terme, aux langues en danger.

14. Qui tient compte de la nature des phonèmes adjacents.

Bibliographie

Chapitre 1 – État de l'art

Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn et Dong Yu (2014). « Convolutional neural networks for speech recognition ». Dans : *IEEE/ACM Transactions on audio, speech, and language processing* 22.10, p. 1533–1545 (page 22).

Adamou, Evangelia (2016). *A corpus-driven approach to language contact : Endangered languages in a comparative perspective*. T. 12. Walter de Gruyter GmbH & Co KG (page 16).

Albright, Eric et John Hatton (2008). « WeSay : A tool for engaging native speakers in dictionary building ». Dans : *Documenting and revitalizing Austronesian languages* 1, p. 189 (page 10).

Anderson, Gregory D.S. (2011). « Language Hotspots : what (applied) linguistics and education should do about language endangerment in the twenty-first century ». Dans : *Language and Education* 25.4, p. 273–289. doi : 10.1080/09500782.2011.577218 (page 4).

Anderson, Gregory D.S., David K. Harrison et Chris Rainier (2007). *Enduring Voices Project*. url : https://www.nationalgeographic.org/archive/projects/enduring-voices/about/ (page 4).

Arisoy, Ebru, Tara N Sainath, Brian Kingsbury et Bhuvana Ramabhadran (2012). « Deep neural network language models ». Dans : *Proceedings of the NAACL-HLT 2012 Workshop : Will We Ever Really Replace the N-gram Model ? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, p. 20–28 (page 25).

Asgari, Ehsaneddin et Hinrich Schütze (2017). « Past, Present, Future : A Computational Investigation of the Typology of Tense in 1000 Languages ». Dans : *arXiv preprint arXiv :1704.08914*. url : https://arxiv.org/pdf/1704.08914.pdf (page 28).

Aurrekoetxea, Gotzon, Karmele Fernandez-Aguirre, Jesús Rubio, Borja Ruiz et Jon Sánchez (2013). « 'DiaTech' : A new tool for dialectology ». Dans : *Literary and linguistic computing* 28.1, p. 23–30 (page 19).

Austin, Peter K (2013). « Language documentation and meta-documentation ». Dans : *Keeping languages alive : Documentation, pedagogy and revitalization*, p. 3–15 (page 13).

Austin, Peter K et Julia Sallabank (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press (page 3).

Balbi, Adrien (1826). *Introduction à l'Atlas ethnographique du globe*. T. 1. Rey et Gravier (page 1).

Barnard, Etienne, Marelie H. Davel et Gerhard B. Van Huyssteen (2010). « Speech Technology for Information Access : a South African Case Study. » Dans : *AAAI Spring Symposium : Artificial Intelligence for Development* (page 27).

Barras, Claude, Edouard Geoffrois, Zhibiao Wu et Mark Liberman (1998). « Transcriber : A Free Tool For Segmenting, Labeling And Transcribing Speech ». Dans : *First international conference on language resources and evaluation (LREC)*, p. 1373–1376 (page 15).

Beermann, Dorothee, Lars Hellan et Jonathan Brindle (2006). « TypeCraft : A Natural Language Database ». Dans : *Legon-Trondheim Linguistics Project Meeting in Accra* (page 14).

Beermann, Dorothee et Pavel Mihaylov (2014). « TypeCraft Collaborative Databasing and Resource Sharing for Linguists ». Dans : *Language resources and evaluation* 48.2, p. 203–225 (page 14).

Berment, Vincent (2004). « Méthodes pour informatiser les langues et les groupes de langues «peu dotées» ». Thèse de doctorat. Université Joseph-Fourier-Grenoble I (page 26).

Bert, Michel (2010). « Qui parle une langue en danger ? Locuteurs du francoprovençal et de l'occitan en Rhône-Alpes, France ». Dans : *Faits de langues* 35-36 (page 4).

Bert, Michel et Jean-Baptiste Martin (2012). « Genèse d'une politique linguistique régionale : le projet FORA ». Dans : *Langues de France, langues en danger : aménagement et rôle des linguistes (textes rassemblés par Anne-laure Dotte, Valelia Muni Toke et Jean Sibille)*. Editions Privat, p. 65–78. url : https://halshs.archives-ouvertes.fr/halshs-01179241 (page 4).

Besacier, Laurent, Etienne Barnard, Alexey Karpov et Tanja Schultz (2014). « Automatic speech recognition for under-resourced languages : A survey ». Dans : *Speech Communication* 56, p. 85–100. doi : 10.1016/j.specom.2013.07.008 (page 27).

Besacier, Laurent, Viet-Bac Le, Eric Castelli, S Sethserey et Ludovic Protin (2005). « Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer ». Dans : *TALN 2005* (page 27).

Bettinson, Mat et Steven Bird (2017). « Developing a suite of mobile applications for collaborative language documentation ». Dans : *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 156–164 (pages 11, 48).

Bickel, Balthasar, Bernard Comrie et Martin Haspelmath (2008). « The Leipzig Glossing Rules. Conventions for interlinear morpheme by morpheme glosses ». Dans : *Revised version of February* (page 6).

Bigi, Brigitte et Daniel Hirst (2012). « SPeech Phonetization Alignment and Syllabification (SPPAS) : a tool for the automatic analysis of speech prosody ». Dans : *Speech Prosody*. Shanghai, China, p. 1–4. url : https://hal.archives-ouvertes.fr/hal-00983699 (page 18).

Bird, Steven (2006). « NLTK : the natural language toolkit ». Dans : *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, p. 69–72 (page 17).

— (2016). « Computing gives us tools to preserve disappearing languages ». English. Dans : *The Conversation*. URL : https://theconversation.com/computing-gives-us-tools-to-preserve-disappearing-languages-60235 (pages 7, 46).

Bird, Steven, Florian R Hanke, Oliver Adams et Haejoong Lee (2014). « Aikuma : A mobile app for collaborative language documentation ». Dans : *ACL 2014*, p. 1 (pages 11, 47, 48).

Bird, Steven, Ewan Klein et Edward Loper (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. " O'Reilly Media, Inc." (page 17).

Bird, Steven et Gary Simons (2000). « White paper on establishing an infrastructure for open language archiving ». Dans : *Workshop on Web-Based Language Documentation and Description, Philadelphia, PA*, p. 12–15 (page 13).

Boersma, Paulus Petrus Gerardus *et al.* (2002). « Praat, a system for doing phonetics by computer ». Dans : *Glot international* 5 (page 17).

Bowern, Claire (2015). *Linguistic fieldwork : A practical guide*. Palgrave Macmillan (pages 2, 12).

Broeder, Daan, Andreas Claus, Freddy Offenga, Romuald Skiba et Paul Trilsbeek (2006). *LAMUS–the Language Archive Management and Upload System* (page 14).

Burridge, James (2017). « Spatial evolution of human dialects ». Dans : *arXiv preprint arXiv :1703.00533*. URL : https://arxiv.org/pdf/1703.00533.pdf (page 20).

Caisse, Peter (2008). « Le breton : une langue en danger ». Dans : *Philadelphia : Drexel University* (page 4).

Challa, Krishnaveer Abhishek et Jawahar Annabattula (2017). « Computational Linguistic Distances and Big Data ». Dans : *Exploring the Convergence of Big Data and the Internet of Things*, p. 55 (page 28).

Christodouloupoulos, Christos et Mark Steedman (2015). « A massively parallel corpus : the Bible in 100 languages ». Dans : *Language Resources and Evaluation* 49.2, p. 375–395. ISSN : 1574-0218. DOI : 10.1007/s10579-014-9287-y (page 28).

Coyne, Bob et Richard Sproat (2001). « WordsEye : An automatic text-to-scene conversion system ». Dans : *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, p. 487–496 (page 10).

De Vries, Nic J., Jaco Badenhorst, Marelie H Davel, Etienne Barnard et Alta De Waal (2011). « Woefzela–an open-source platform for ASR data collection in the developing world ». Dans : Conference paper (page 11).

De Vries, Nic J., Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard et Alta De Waal (2013). « A smartphone-based ASR data collection tool for under-resourced languages ». Dans : *Speech Communication Journal* 56, p. 119–131 (page 11).

Drude, Sebastian, Bruce Birch, Daan Broeder, Peter Withers et Peter Wittenburg (2013). « Crowd-sourcing and apps in the field of linguistics : Potentials and challenges of the coming technology ». Dans : *The Language Archive* (page 10).

Gawne, Lauren, Barbara F Kelly, Andrea L Berez-Kroeker et Tyler Heston (2017). « Putting practice into words : The state of data and methods transparency in grammatical descriptions ». Dans : *Language Documentation & Conservation* 11, p. 157–189. url : http://hdl.handle.net/10125/24731 (pages 10, 19).

Goldman, Jean-Philippe (2011). *EasyAlign : An automatic phonetic alignment tool under Praat* (page 17).

Gorman, Kyle, Jonathan Howell et Michael Wagner (2011). « ProsodyLab-Aligner : A tool for forced alignment of laboratory speech ». Dans : *Canadian Acoustics* 39.3, p. 192–193 (page 18).

Graves, Alex, Abdel-rahman Mohamed et Geoffrey Hinton (2013). « Speech recognition with deep recurrent neural networks ». Dans : *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, p. 6645–6649 (page 22).

Hacine-Gharbi, Abdenour (2012). « Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole ». Thèse de doctorat. Université d'Orléans (page 21).

Haspelmath, Martin, Matthew S Dryer, David Gil et Bernard Comrie (2005). *The world atlas of language structures* (page 14).

Hatton, John (2013). « SayMore : Language documentation productivity ». Dans : *Proceedings of the 3rd International Conference on Language Documentation and Conservation* (page 13).

Hendery, Rachel et Andrew Burrell (2016). « Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) ». Dans : *Memory of the World in Canberra (12/11/2016 : Canberra Museum and Gallery)* (page 14).

Himmelmann, Nikolaus P. (1998). « Documentary and Descriptive linguistics ». Dans : *Linguistics* 36, p. 161–195 (pages 3, 9, 13).

Hinton, Geoffrey E. (2010). *A practical guide to training restricted Boltzmann machines*. UTML TR 2010-003. Dept. Computer Science, University of Toronto (page 23).

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath *et al.* (2012). « Deep neural networks for acoustic modeling in speech recognition : The sha-

red views of four research groups ». Dans : *Signal Processing Magazine, IEEE* 29.6, p. 82–97 (page 22).

Hirst, Daniel J (2007). « A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation ». Dans : *Proceedings of the XVIth International Conference of Phonetic Sciences*. T. 12331236 (page 17).

Holton, Gary, Kavon Hooshiar et Nick Thieberger (2017). « Developing collection management tools to create more robust and reliable linguistic data ». Dans : *Language Documentation & Conservation* (page 19).

Hyman, Larry M (2001). « Fieldwork as a state of mind ». Dans : *Linguistic fieldwork*, p. 15–33 (pages 1, 2).

Kipp, Michael (2001). « Anvil-a generic annotation tool for multimodal dialogue ». Dans : *Seventh European Conference on Speech Communication and Technology* (page 15).

Krauwer, Steven (2003). « The basic language resource kit (BLARK) as the first milestone for the language resources roadmap ». Dans : *Proceedings of SPECOM 2003*, p. 8–15 (page 26).

Martin, Philippe (2004). « WinPitch LTL II, A Multimodal Pronunciation Software ». Dans : *InSTIL/ICALL Symposium 2004* (page 17).

Martinet, André (1967). *Eléments de linguistique générale*. A. Colin (page 1).

Mertens, Piet (2004). « The prosogram : Semi-automatic transcription of prosody based on a tonal perception model ». Dans : *Speech Prosody 2004, International Conference* (page 17).

Michaelis, Susanne, Philippe Maurer, Martin Haspelmath et Magnus Huber (2013). *The atlas of pidgin and creole language structures online*. Max Planck Institute for Evolutionary Anthropology Leipzig (page 14).

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François et Evangelia Adamou (2014). « Documenting and researching endangered languages : the Pangloss Collection ». Dans : *Language Documentation & Conservation* 8, p. 119–135 (page 13).

Mikolov, Tomas, Martin Karafiát, Lukas Burget, Jan Cernockỳ et Sanjeev Khudanpur (2010). « Recurrent neural network based language model. » Dans : *Interspeech*, p. 3 (page 25).

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg et Therese Leinonen (2011). « Gabmap-a web application for dialectology ». Dans : *Dialectologia : revista electrònica*, p. 65–89 (page 19).

Neubig, Graham (2017). « Neural Machine Translation and Sequence-to-sequence Models : A Tutorial ». Dans : *arXiv preprint arXiv :1703.01619* (page 25).

Nimaan, Abdillahi, Pascal Nocera et Juan-Manuel Torres-Moreno (2006). « Boites a outils TAL pour les langues peu informatisees : Le cas du Somali ». Dans : *JADT'06 : actes des 8es Journées internationales d'analyse statistique des données textuelles : Besançon, 19-21 avril 2006* 3, p. 697 (page 27).

Obin, Nicolas, Julie Beliao, Christophe Veaux et Anne Lacheret (2014). « SLAM : Automatic stylization and labelling of speech melody ». Dans : *Speech Prosody*, p. 246–250 (page 17).

Pellegrini, Thomas (2008). « Transcription automatique de langues peu dotées ». Thèse de doctorat. Université Paris Sud-Paris XI (pages 20, 27).

Pham, Ngoc-Quan, German Kruszewski et Gemma Boleda (2016). « Convolutional Neural Network Language Models. » Dans : *EMNLP*, p. 1153–1162 (page 25).

Povey, Daniel, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz et Samuel Thomas (2011). « The subspace Gaussian mixture model—A structured model for speech recognition ». Dans : *Computer Speech & Language* 25.2, p. 404–439 (pages 22, 87).

Rabinovich, Ella, Noam Ordan et Shuly Wintner (2017). « Found in Translation : Reconstructing Phylogenetic Language Trees from Translations ». Dans : *arXiv preprint arXiv :1704.07146*. url : https://arxiv.org/pdf/1704.07146.pdf (page 28).

Rahimi, Afshin, Trevor Cohn et Timothy Baldwin (2017). « A Neural Model for User Geolocation and Lexical Dialectology ». Dans : *arXiv preprint arXiv :1704.04008* (page 20).

Rice, Sally (2011). « Applied field linguistics : delivering linguistic training to speakers of endangered languages ». Dans : *Language and Education* 25.4, p. 319–338. doi : 10.1080/09500782.2011.577216 (page 4).

Rogers, Chris (2010). « Fieldworks Language Explorer (FLEx) 3.0 ». Dans : *Language Documentation & Conservation* 4 (page 16).

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini et Jiahong Yuan (2011). « FAVE (Forced Alignment and Vowel Extraction) program suite ». Dans : URL http ://fave.ling.upenn.edu (page 18).

Sak, Haşim, Andrew Senior et Françoise Beaufays (2014). « Long short-term memory recurrent neural network architectures for large scale acoustic modeling ». Dans : *Fifteenth Annual Conference of the International Speech Communication Association* (page 22).

Sakel, Jeanette et Daniel L. Everett (2012). *Linguistic fieldwork : a student guide*. Cambridge University Press (pages 2, 5, 6, 62).

Samarin, William J (1967). *Field linguistics : A guide to linguistic field work*. Holt, Rinehart et Winston (page 5).

Séguy, Jean (1973). *La dialectométrie dans l'Atlas linguistique de la Gascogne*. Sous la dir. de RLiR. T. 37. Société de linguistique romane, p. 1–24 (page 19).

Sercu, Tom, Christian Puhrsch, Brian Kingsbury et Yann LeCun (2016). « Very deep multilingual convolutional neural networks for LVCSR ». Dans : *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, p. 4955–4959 (page 27).

Sundermeyer, Martin, Ralf Schlüter et Hermann Ney (2012). « LSTM neural networks for language modeling ». Dans : *Thirteenth Annual Conference of the International Speech Communication Association* (page 25).

Thieberger, Nicholas (2013). « Digital humanities and language documentation ». Dans : *Selected Papers from the 44th Conference of the Australian Linguistic Society*, p. 144–59 (pages 9, 13).

Thieberger, Nicholas et Andrea L Berez (2012). *Linguistic data management*. New York : Oxford University Press, p. 90–120 (page 12).

Ulinski, Morgan, Anusha Balakrishnan, Daniel Bauer, Bob Coyne, Julia Hirschberg et Owen Rambow (2014). « Documenting endangered languages with the WordsEye Linguistics Tool ». Dans : *ACL 2014*, p. 6 (page 10).

Vu, Ngoc Thang, Franziska Kraus et Tanja Schultz (2011). « Rapid Building of an ASR System for Under-Resourced Languages Based on Multilingual Unsupervised Training. » Dans : *Interspeech*. Citeseer, p. 3145–3148 (page 27).

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann et Han Sloetjes (2006). « Elan : A professional framework for multimodality research ». Anglais. Dans : *Proceedings of LREC*. T. 2006, 5th (page 15).

Woodbury, Anthony C. (2003). « Defining documentary linguistics ». English. Dans : *Language Documentation and Description*. Sous la dir. de Peter K. Austin. T. 1. London, p. 35–51 (pages 11, 47).

Wright, E (2004). « Documenting Endangered Languages ». Dans : *Documenting Endangered Languages* (page 9).

Chapitre 2 – Langues Africaines Abordées

Caron, Bernard (2000). « Les langues au Nigeria ». Dans : *Notre Librairie. Revue des littératures du Sud* 141, p. 8–15 (pages 31, 32).

— (2011). « Haoussa ». Dans : *Dictionnaire des Langues*. Sous la dir. de Joëlle Busuttil et Alain Peyraube Emilio Bonvini. Presses Universitaires de France, p. 263–269. url : https://halshs.archives-ouvertes.fr/halshs-00643960 (pages 32, 35).

— (2015). « Hausa grammatical sketch ». Dans : (page 35).

Caron, Bernard et Ahmed H Amfani (1997). *Dictionnaire français-haoussa : suivi d'un index haoussa-français*. KARTHALA Editions (pages 33, 34).

Cissé, Mame Thierno (2006). « Problèmes de phonétique et de phonologie en wolof ». Dans : *Revue électronique internationale de sciences du langage SudLangues* 6, p. 1–41 (pages 37, 41, 103, 128).

Comrie, Bernard (2009). *The world's major languages*. Routledge (page 30).

Dard, Jean et Antonio Savaresi (1825). *Dictionnaire français-wolof et français-bambara, suivi du dictionnaire wolof-français ; par m. J. Dard.*. a l'Imprimerie Royale (page 35).

Diagne, Pathé (1971). *Grammaire de wolof moderne*. Présence africaine (pages 39, 40).

Dimmendaal, Gerrit J (2008). « Language ecology and linguistic diversity on the African continent ». Dans : *Language and Linguistics Compass* 2.5, p. 840–858 (page 29).

Diouf, Jean Léopold (2003). *Dictionnaire wolof-français et français-wolof*. KARTHALA Editions (pages 39, 40, 58, 70, 86).

Dramé, Mamour (2012). « Phonologie et Morphosyntaxe Comparées de trois dialects Wolof ». Thèse de doctorat. UCAD, Dakar. (page 37).

Fal, Arame, Rosine Santos et Jean Léonce Doneux (1990). *Dictionnaire wolof-français : suivi d'un index français-wolof*. KARTHALA Editions (pages 36, 58, 70, 86).

Greenberg, Joseph Harold (1966). *The languages of Africa*. T. 25. Indiana University (page 30).

Guérin, Maximilien (2011). *Le syntagme nominal en wolof : une approche typologique* (page 37).

Guérin, Maximilien (2016). « Les constructions verbales en wolof : vers une typologie de la prédication, de l'auxiliation et des périphrases ». Thèse de doctorat. Sorbonne Paris Cité (pages 37, 38, 41).

Jaggar, Philip J (2001). *Hausa*. T. 7. John Benjamins Publishing (pages 29, 31, 34, 106, 110).

Ka, Omar (1994). *Wolof phonology and morphology*. University Press of Amer (pages 38, 39).

Koslow, Philip (1995). *Hausaland : the fortress kingdoms*. Chelsea House Pub (page 31).

Laleye, Fréjus A A, Laurent Besacier, Eugène Ezin et Cina Motamed (2016). « First Automatic Fongbe Continuous Speech Recognition System : Development of Acoustic Models and Language Models ». Dans : *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. T. 8. Gdansk, Poland, p. 477–482. doi : 10.15439/2016F153. url : https://hal.archives-ouvertes.fr/hal-01436788 (pages 30, 78, 96).

Leclerc, Jacques (2013). « Sénégal ». Dans : *http ://www.axl.cefan.ulaval.ca/afrique/senegal.htm (consulté le 30 janvier 2018)* (pages 30, 37).

Lefebvre, Claire et Anne-Marie Brousseau (2001). *A grammar of Fonge*. De Gruyter Mouton, p. 608 (page 30).

Lewis, M Paul et Gary F Simons (2010). « Assessing endangerment : expanding Fishman's GIDS ». Dans : *Revue roumaine de linguistique* 55.2, p. 103–120 (pages 31, 44).

Lindau-Webb, Mona (1985). « Hausa vowels and diphthongs ». Dans : *UCLA working papers in phonetics* 60, p. 75–92 (page 34).

Lüpke, Friederike et Sokhna Bao-Diop (2014). *Beneath the surface ? Contemporary ajami writing in West Africa, exemplified through Wolofal* (page 41).

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé et François Pellegrino (2013). « LAPSyd : lyon-albuquerque phonological systems database. » Dans : *INTERSPEECH*, p. 3022–3026 (page 38).

Newman, Paul (2000). « The Hausa Language ». Dans : *An Encyclopedic Reference Grammar*. url : http://tocs.ub.uni-mainz.de/pdfs/09159748X.pdf (pages 32–34, 106).

Newman, Roxana Ma et Vincent J van Heuven (1981). « An acoustic and phonological study of pre-pausal vowel length in Hausa ». Dans : *Journal of African Languages and Linguistics* 3.1, p. 1–18 (pages 34, 106).

Ngom, Fallou (2010). « Ajami scripts in the Senegalese speech community ». Dans : *Journal of Arabic and Islamic Studies* 10, p. 1–23 (page 41).

Robert, Stéphane (2011). « Le wolof ». Dans : *Dictionnaire des langues*. Sous la dir. de Joëlle Busuttil & Alain Peyraube Emilio Bonvini. Dicos Poche. Quadrige/P.U.F., p. 23–30. url : https://hal.archives-ouvertes.fr/hal-00600630 (pages 37, 38, 41).

Sauvageot, Serge (1965). *Description synchronique d'un dialecte wolof : le parler du Dyolof*. 73. Institut Français d'Afrique Noire, Dakar. (pages 40, 114).

Schlippe, Tim, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs et Tanja Schultz (2012). « Hausa large vocabulary continuous speech recognition. » Dans : *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU)*, p. 11–14 (pages 33, 83, 84, 86, 104).

Segerer, Guillaume et Sébastien Flavier (2013). « Base de données RefLex : Reference Lexicon ». Dans : url : http://reflex.cnrs.fr/Lexiques/webball/index.html (pages 35, 70).

Voisin, Sylvie (en prép.). « Le wolof et ses variantes. » Dans : *JWAL* (page 37).

Voisin, Sylvie et Mamour Dramé (en prép.). « Inaccompli et complexe verbal dans différentes variantes du wolof. » Dans : *Africana Linguistica* (page 37).

Vycichl, Werner (1990). « Les langues tchadiques et l'origine chamitique de leur vocabulaire ». Dans : *Relations interethniques et culture matérielle dans le bassin du lac Tchad : actes du IIIème Colloque MEGA-TCHAD, Paris, ORSTOM, 11-12 septembre 1986*. IRD Editions, p. 33 (pages 29, 32).

Chapitre 3 – Collecter : Lig-Aikuma, une application mobile de collecte de parole sur le terrain

Adda, Gilles, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Sebastian Stüker, Mark Van de Velde, François Yvon et Sabine Zerbian (2016). « Innovative technologies for under-resourced language documentation : The BULB Project ». Dans : *Workshop CCURL 2016 - Collaboration and Computing for Under-Resourced Languages - LREC*. Portoroz, Slovenia. url : https://hal.archives-ouvertes.fr/hal-01350124 (page 81).

Adda, Gilles, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon et Sabine Zerbian (2016). « Breaking the Unwritten Language Barrier : The BULB Project ». Dans : *Procedia Computer Science* 81, p. 8–14. issn : 1877-0509. url : http://www.sciencedirect.com/science/article/pii/S1877050916300370 (page 45).

Amboulou, Celestin (1998). « Le Mbochi : langue bantoue du Congo Brazzaville (zone C, groupe C20) ». Thèse de doctorat. INALCO, Paris (page 45).

Ambouroue, Odette, Bettie Vanhoudt et Claire Grégoire (2007). « Eléments de description de l'orungu : langue bantu du gabon (B11b) ». Thèse de doctorat (page 44).

Bettinson, Mat et Steven Bird (2017). « Developing a suite of mobile applications for collaborative language documentation ». Dans : *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 156–164 (pages 11, 48).

Bird, Steven (2010). « A scalable method for preserving oral literature from small languages ». Dans : *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries* (page 47).

— (2014). « Collecting Bilingual Audio in Remote Indigenous Communities ». Dans : COLING (page 47).

— (2016). « Computing gives us tools to preserve disappearing languages ». English. Dans : *The Conversation*. url : https://theconversation.com/computing-gives-us-tools-to-preserve-disappearing-languages-60235 (pages 7, 46).

Bird, Steven, Florian R Hanke, Oliver Adams et Haejoong Lee (2014). « Aikuma : A mobile app for collaborative language documentation ». Dans : *ACL 2014*, p. 1 (pages 11, 47, 48).

Blachon, David, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker et Annie Rialland (2016). « Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app ». Dans : *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*. Yogyakarta, Indonésie. doi : https://doi.org/10.1016/j.procs.2016.04.030. url : https://www.sciencedirect.com/science/article/pii/S1877050916300448 (page 81).

Cameron, Catherine Ann et Min Wang (1999). « Frog, where are you ? Children's narrative expression over the telephone ». Dans : *Discourse Processes* 28.3, p. 217–236. doi : 10.1080/01638539909545082 (page 74).

Chafe, Wallace L (1980). *The pear stories : Cognitive, cultural, and linguistic aspects of narrative production* (page 74).

Chéneau-Loquay, Annie (2010). « Modes d'appropriation innovants du téléphone mobile en Afrique ». Dans : *Conférence des plénipotentiaires de l'union africaine des télécommunications*. Brazzaville, France. url : https://halshs.archives-ouvertes.fr/halshs-00565328 (page 43).

Chéneau-Loquay, Annie (2012). « La téléphonie mobile dans les villes africaines. Une adaptation réussie au contexte local ». Dans : *L'Espace géographique* 41, p. 82–93. doi : 10.3917/eg.411.0082 (page 43).

Cooper-Leavitt, Jamison, Lori Lamel, Annie Rialland, Martine Adda-Decker et Gilles Adda (2017b). « Developing an Embosi (Bantu C25) Speech Variant Dictionary to Model Vowel Elision and Morpheme Deletion ». Dans : *Proceedings of Interspeech 2017*. Stockholm, Suède (page 81).

Diouf, Jean Léopold (2003). *Dictionnaire wolof-français et français-wolof*. KARTHALA Editions (pages 39, 40, 58, 70, 86).

Embanga Aborobongui, Georges Martial (2013). « Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi C25) ». Thèse de doctorat. Université de la Sorbonne nouvelle - Paris III. url : https://tel.archives-ouvertes.fr/tel-01064356 (page 45).

Fal, Arame, Rosine Santos et Jean Léonce Doneux (1990). *Dictionnaire wolof-français : suivi d'un index français-wolof*. KARTHALA Editions (pages 36, 58, 70, 86).

Franke, Joerg, Markus Mueller, Fatima Hamlaoui, Sebastian Stueker et Alex Waibel (2016). « Phoneme Boundary Detection using Deep Bidirectional LSTMs ». Dans : *Speech Communication*. VDE, p. 1–5 (page 81).

Gauthier, Elodie, Laurent Besacier et Sylvie Voisin (2017). « Machine Assisted Analysis of Vowel Length Contrasts in Wolof ». Dans : *Interspeech 2017*. Stockholm, Suède (pages 81, 127).

Gauthier, Elodie, Laurent Besacier et Sylvie Voisin (2018). « LIG-AIKUMA, une application mobile pour la collecte de parole sur le terrain ». Dans : *Du terrain à la théorie : Les 40 ans du Lacito* (page 81).

Gauthier, Elodie, David Blachon, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda et Grégoire Bachman (2016). « LIG-AIKUMA : a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies ». Dans : *Interspeech 2016 (Show and Tell)* (page 81).

Godard, Pierre, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland et François Yvon (2016). « Preliminary Experiments on Unsupervised Word Discovery in Mboshi ». Dans : *Interspeech 2016*. San Francisco, California, USA (pages 81, cxxxii).

Grinevald, Colette (2011). « On constructing a working typology of the expression of path ». Dans : *Faits de langues* 3, p. 43–70 (page 75).

Guthrie, Malcolm (1948). *The classification of the Bantu languages* (page 44).

Hamlaoui, Fatima et Emmanuel-Moselly Makasso (2015). « Focus marking and the unavailability of inversion structures in the Bantu language Bààsá(A43) ». Dans : *Lingua* 154, p. 35–64. issn : 0024-3841. doi : http://dx.doi.org/10.1016/j.lingua.2014.10.010. url : http://www.sciencedirect.com/science/article/pii/S0024384114002605 (page 44).

Hanke, Florian R. et Steven Bird (2013). « Large-Scale Text Collection for Unwritten Languages ». Dans : *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, p. 1134–1138 (pages 46, 80, 129).

Kesteloot, Lilyan et Bassirou Dieng (1989). *Du tieddo au talibé*. T. 2. Editions Présence Africaine (pages 69, 71).

Kobès, Mgr et R.P. Abiven (1922). *Dictionnaire volof-français*. Mission catholique, Dakar (pages 69, 71).

Kouarata, Guy-Noël (2000). *Dictionnaire Mbochi - Français*. SIL-Congo, Brazzaville (page 45).

Laleye, Fréjus A A, Laurent Besacier, Eugène Ezin et Cina Motamed (2016). « First Automatic Fongbe Continuous Speech Recognition System : Development of Acoustic Models and Language Models ». Dans : *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. T. 8. Gdansk, Poland, p. 477–482. doi : 10.15439/2016F153. url : https://hal.archives-ouvertes.fr/hal-01436788 (pages 30, 78, 96).

Lemb, Pierre, François De Gastines et Meinrad Pierre Hebga (1973). *Dictionnaire basaa-français*. Collège Libermann (page 44).

Lewis, M Paul et Gary F Simons (2010). « Assessing endangerment : expanding Fishman's GIDS ». Dans : *Revue roumaine de linguistique* 55.2, p. 103–120 (pages 31, 44).

Makasso, Emmanuel-Moselly, Fatima Hamlaoui et Seunghun J Lee (2017). « Aspects of the intonational phonology of Bàsàá ». Dans : *Intonation in African Tone Languages* 24, p. 167 (page 44).

Mayer, Mercer (1969). *Frog, where are you ?* Dial Press New York (page 74).

Melese, Michael, Laurent Besacier et Million Meshesha (2016). « Amharic Speech Recognition for Speech Translation ». Dans : *Atelier Traitement Automatique des Langues Africaines (TALAF). JEP-TALN 2016* (pages 78, 96).

Moseley, Christopher (2010). « Atlas des langues en danger dans le monde (3 e éd) ». Dans : *Paris : Éditions Unesco. Version en ligne : http ://www. unesco. org/culture/en/endangeredlanguages/atlas* (page 43).

Nouguier Voisin, Sylvie (2002). « Relations entre fonctions syntaxiques et fonctions sémantiques en wolof ». Thèse de doctorat. Lyon 2 (page 69).

Raponda-Walker, André et Louis Tardy (1934). *Dictionnaire mpongwe-français, suivi d'éléments de grammaire*. Metz, La Libre Lorraine (page 44).

Reilly, Judy, Molly Losh, Ursula Bellugi et Beverly Wulfeck (2004). « "Frog, where are you ?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome ». Dans : *Brain and language* 88.2, p. 229–247 (page 74).

Rialland, Annie, Georges Martial Embanga Aborobongui, Martine Adda-Decker et Lori Lamel (2015). « Dropping of the Class-Prefix Consonant, Vowel Elision and Automatic Phonological Mining in Embosi (Bantu C 25) ». Dans : *Selected Proceedings of the 44th Annual Conference on African Linguistics* (page 81).

Rialland, Annie et Martial Embanga Aborobongui (2017). « How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift ». Dans : *Intonation in African Tone Languages* 24, p. 195 (page 45).

Rialland, Annie, Martial Embanga Aborobongui, Martine Adda-Decker et Lori Lamel (2016). « Phonologie et traitement automatique de la parole : le cas de l'embosi ». Dans : (page 81).

Rialland, Annie, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Elodie Gauthier, Pierre Godard et Jamison Cooper-Leavitt (2018). « Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville) ». Dans : *11th edition of the Language Resources and Evaluation Conference (LREC 2018)*. Miyazaki, Japan. url : https://hal.archives-ouvertes.fr/hal-01710043 (page 82).

Rosenhuber, S. (1908). *Die Basa-Sprache*. T. 11. MSOS, p. 219–306 (page 44).

Rymer, Russ (2012). « Vanishing voices ». Dans : *National Geographic* 222.1, p. 60–93 (pages xiii, 46).

Sakel, Jeanette et Daniel L. Everett (2012). *Linguistic fieldwork : a student guide*. Cambridge University Press (pages 2, 5, 6, 62).

Segerer, Guillaume et Sébastien Flavier (2013). « Base de données RefLex : Reference Lexicon ». Dans : url : http://reflex.cnrs.fr/Lexiques/webball/index.html (pages 35, 70).

Van de Velde, Mark et Odette Ambouroué (2011). « The grammar of Orungu proper names ». Dans : *Journal of African Languages and Linguistics* 23, p. 113–141 (page 45).

Vetter, Marco, Markus Müller, Fatima Hamlaoui, Graham Neubig, Satoshi Nakamura, Sebastian Stüker et Alex Waibel (2016). « Unsupervised phoneme segmentation of previously unseen languages ». Dans : *Proceedings of the Interspeech 2016* (page 81).

Woodbury, Anthony C. (2003). « Defining documentary linguistics ». English. Dans : *Language Documentation and Description*. Sous la dir. de Peter K. Austin. T. 1. London, p. 35–51 (pages 11, 47).

Chapitre 4 – Transcrire : Automatisation à l'aide de la reconnaissance automatique de la parole

Attardi, Giuseppe et Antonia Fuschetto (2013). « Wikipedia extractor ». Dans : *Medialab, University of Pisa* (page 85).

Chow, Y-L (1990). « Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm ». Dans : *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, p. 701–704 (page 87).

Diouf, Jean Léopold (2003). *Dictionnaire wolof-français et français-wolof*. KARTHALA Editions (pages 39, 40, 58, 70, 86).

Fal, Arame, Rosine Santos et Jean Léonce Doneux (1990). *Dictionnaire wolof-français : suivi d'un index français-wolof*. KARTHALA Editions (pages 36, 58, 70, 86).

Gales, M. (1998). « Maximum likelihood linear transformations for hmm-based speech recognition ». Dans : *Computer Science and Language*. T. 12, p. 75–98 (page 87).

Gauthier, Elodie, Laurent Besacier et Sylvie Voisin (2016b). « Speed perturbation and vowel duration modeling for ASR in Hausa and Wolof languages ». Dans : *Interspeech 2016*. San Francisco, California, USA (page 83).

Gauthier, Elodie, Laurent Besacier, Sylvie Voisin, Michael Melese et Uriel Pascal Elingui (2016). « Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition : a Case Study of Wolof ». Dans : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovénie : European Language Resources Association (ELRA). isbn : 978-2-9517408-9-1 (pages 83, 117).

Gelas, Hadrien, Laurent Besacier et Francois Pellegrino (2012). « Developments of Swahili resources for an automatic speech recognition system ». Dans : *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*. Cape-Town, Afrique Du Sud. url : http://hal.inria.fr/hal-00954048 (page 96).

Gopinath, R. A. (1998). « Maximum likelihood modeling with gaussian distributions for classification ». Dans : *Proc. of ICASSP*, p. 661–664 (page 87).

Hannun, Awni Y., Carl Case, Jared Casper, Bryan C. Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates et Andrew Y. Ng (2014). « Deep Speech : Scaling up end-to-end speech recognition ». Dans : *CoRR* abs/1412.5567. url : http://arxiv.org/abs/1412.5567 (page 93).

Jaitly, Navdeep et Geoffrey E Hinton (2013). « Vocal tract length perturbation (VTLP) improves speech recognition ». Dans : *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, p. 625–660 (page 93).

Kingsbury, Brian (2009). « Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling ». Dans : *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, p. 3761–3764 (page 88).

Ko, Tom, Vijayaditya Peddinti, Daniel Povey et Sanjeev Khudanpur (2015). « Audio augmentation for speech recognition. » Dans : *INTERSPEECH*, p. 3586–3589 (pages 93, 94).

Laleye, Fréjus A A, Laurent Besacier, Eugène Ezin et Cina Motamed (2016). « First Automatic Fongbe Continuous Speech Recognition System : Development of Acoustic Models and Language Models ». Dans : *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. T. 8. Gdansk, Poland, p. 477–482. doi : 10.15439/2016F153. url : https://hal.archives-ouvertes.fr/hal-01436788 (pages 30, 78, 96).

McDonough, John, Thomas Schaaf et Alex Waibel (2002). « On maximum mutual information speaker-adapted training ». Dans : *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. T. 1. IEEE, p. I–601 (page 87).

Melese, Michael, Laurent Besacier et Million Meshesha (2016). « Amharic Speech Recognition for Speech Translation ». Dans : *Atelier Traitement Automatique des Langues Africaines (TALAF). JEP-TALN 2016* (pages 78, 96).

Mikolov, Tomas, Stefan Kombrink, Anoop Deoras, Lukar Burget et Jan Cernocky (2011). « RNNLM-Recurrent neural network language modeling toolkit ». Dans : *Proc. of the 2011 ASRU Workshop*, p. 196–201 (page 88).

Novak, Josef R (2011). « Phonetisaurus : A wfst-driven phoneticizer ». Dans : *The University of Tokyo, Tokyo Institute of Technology*, p. 221–222 (page 87).

Paulin, Mattis, Jérôme Revaud, Zaid Harchaoui, Florent Perronnin et Cordelia Schmid (2014). « Transformation pursuit for image classification ». Dans : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3646–3653 (page 93).

Povey, Daniel, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz et Samuel Thomas (2011). « The subspace Gaussian mixture model—A structured model for speech recognition ». Dans : *Computer Speech & Language* 25.2, p. 404–439 (pages 22, 87).

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer et Karel Vesely (2011). « The Kaldi speech recognition toolkit ». Dans : *IEEE 2011 workshop on automatic speech recognition and understanding*. EPFL-CONF-192584. IEEE Signal Processing Society (page 87).

Povey, Daniel, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon et Karthik Visweswariah (2008). « Boosted MMI for model and feature-space discriminative training ». Dans : *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, p. 4057–4060 (page 87).

Ragni, Anton, Kate M Knill, Shakti P Rath et Mark JF Gales (2014). « Data augmentation for low resource languages ». Dans : *Fifteenth Annual Conference of the International Speech Communication Association* (page 93).

Rousseau, Anthony (2013). « Xenc : An open-source tool for data selection in natural language processing ». Dans : *The Prague Bulletin of Mathematical Linguistics* 100, p. 73–82 (page 85).

Schlippe, Tim, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs et Tanja Schultz (2012). « Hausa large vocabulary continuous speech recognition. » Dans : *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU)*, p. 11–14 (pages 33, 83, 84, 86, 104).

Stolcke, Andreas *et al.* (2002). « SRILM-an extensible language modeling toolkit. » Dans : *INTERSPEECH* (page 88).

Vesely, Karel, Arnab Ghoshal, Lukáš Burget et Daniel Povey (2013). « Sequence-discriminative training of deep neural networks. » Dans : *Proceedings of the Interspeech 2013*, p. 2345–2349 (page 88).

Wells, John (1997). « SAMPA computer readable phonetic alphabet ». Dans : *Handbook of standards and resources for spoken language systems* 4 (page 86).

Chapitre 5 – Analyser : La machine au service du phonéticien

Adda, Gilles et Martine Adda-Decker (1997). « Normalisation de textes en français : une étude quantitative pour la reconnaissance de la parole ». Dans : *1eres JST FRANCIL* (page 131).

Adda-Decker, Martine, Philippe Boula de Mareüil, Gilles Adda et Lori Lamel (2005). « Investigating syllabic structures and their variation in spontaneous French ». Dans : *Speech Communication* 46.2, p. 119–139 (page 99).

Adda-Decker, Martine, Lori Lamel et Gilles Adda (2014). « Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois ». Dans : *XXXe Journées d'Études sur la Parole (JEP'14)*, p. 284–292 (page 99).

Adi, Yossi, Joseph Keshet, Emily Cibelli, Erin Gustafson, Cynthia Clopper et Matthew Goldrick (2016). « Automatic measurement of vowel duration via structured prediction ». Dans : *The Journal of the Acoustical Society of America* 140.6, p. 4517–4527 (page 102).

Ahmed M. Abdelatty, Ali, Jan Van der Spiegel et Paul Mueller (2001). « Acoustic-phonetic features for the automatic classification of stop consonants ». Dans : *IEEE Transactions on Speech and Audio Processing* 9.8, p. 833–841. issn : 1063-6676. doi : 10.1109/89.966086 (page 99).

Alumäe, Tanel et Rena Nemoto (2013). « Phone duration modeling using clustering of rich contexts. » Dans : *INTERSPEECH*. Citeseer, p. 1801–1805 (page 104).

Bion, Ricardo AH, Kouki Miyazawa, Hideaki Kikuchi et Reiko Mazuka (2013). « Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech ». Dans : *PloS one* 8.2, e51594 (pages 103, 114).

Calvet, Maurice (1966). « Étude phonétique des voyelles du wolof ». Dans : *Phonetica* 14.3, p. 138–168 (page 128).

Cissé, Mame Thierno (2006). « Problèmes de phonétique et de phonologie en wolof ». Dans : *Revue électronique internationale de sciences du langage SudLangues* 6, p. 1–41 (pages 37, 41, 103, 128).

Cooper-Leavitt, Jamison, Lori Lamel, Annie Rialland, Martine Adda-Decker et Gilles Adda (2017a). « Corpus based linguistic exploration via forced alignments with a "lightweight" ». Dans : *8th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics* (page 100).

Dumont, Pierre (1983). *Le français et les langues africaines au Sénégal*. T. 3. KARTHALA Editions (page cxxxii).

Dunbar, Ewan, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera et Emmanuel Dupoux (2017). « The zero resource speech challenge 2017 ». Dans : *arXiv preprint arXiv :1712.04313* (pages 131, cxxxii).

Gadde, VR Rao (2000). « Modeling word duration for better speech recognition ». Dans : *Proceedings of NIST Speech Transcription Workshop* (page 103).

Gauthier, Elodie, Laurent Besacier et Sylvie Voisin (2016a). « Automatic Speech Recognition for African Languages with Vowel Length Contrast ». Dans : *Procedia Computer Science* 81, p. 136–143 (page 126).

— (2017). « Machine Assisted Analysis of Vowel Length Contrasts in Wolof ». Dans : *Interspeech 2017*. Stockholm, Suède (pages 81, 127).

Gauthier, Elodie, Laurent Besacier, Sylvie Voisin, Michael Melese et Uriel Pascal Elingui (2016). « Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition : a Case Study of Wolof ». Dans : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovénie : European Language Resources Association (ELRA). isbn : 978-2-9517408-9-1 (pages 83, 117).

Gay, Thomas (1981). « Mechanisms in the control of speech rate ». Dans : *Phonetica* 38.1-3, p. 148–158 (page 102).

Gelas, Hadrien, Laurent Besacier, Solange Rossato et Francois Pellegrino (2010). « Using automatic speech recognition for phonological purposes : Study of Vowel Lenght in Punu (Bantu B40) ». Dans : *Laphon 12*. New Mexico (US), p. x–x. url : https://hal.archives-ouvertes.fr/hal-00959185 (pages 101, 102).

Godard, Pierre, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland et François Yvon (2016). « Preliminary Experiments on Unsupervised Word Discovery in Mboshi ». Dans : *Interspeech 2016*. San Francisco, California, USA (pages 81, cxxxii).

Hanke, Florian R. et Steven Bird (2013). « Large-Scale Text Collection for Unwritten Languages ». Dans : *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, p. 1134–1138 (pages 46, 80, 129).

Hartigan, John A et PM Hartigan (1985). « The dip test of unimodality ». Dans : *The Annals of Statistics*, p. 70–84 (page 116).

House, Arthur S (1961). « On vowel duration in English ». Dans : *The Journal of the Acoustical Society of America* 33.9, p. 1174–1178 (page 102).

Hyman, Larry M (2013). « Penultimate lengthening in Bantu ». Dans : *Language typology and historical contingency : In honor of Johanna Nichols*, p. 309–330 (page 102).

Jaggar, Philip J (2001). *Hausa*. T. 7. John Benjamins Publishing (pages 29, 31, 34, 106, 110).

Kempton, Timothy et Roger K Moore (2009). « Finding allophones : An evaluation on consonants in the TIMIT corpus ». Dans : *Tenth Annual Conference of the International Speech Communication Association* (page 102).

Lee, Goun et Dong-Jin Shin (2016). « An acoustic and perceptual investigation of the vowel length contrast in Korean ». Dans : *Journal of the Korean society of speech sciences* 8.1, p. 37–44 (page 103).

Lehiste, Ilse (1970). *Suprasegmentals*. MIT Press, Cambridge, MA (page 102).

Lindblom, Björn (1967). « Vowel duration and a model of lip mandible coordination ». Dans : *Speech Transmission Laboratory Quarterly Progress Status Report* 4, p. 1–29 (page 102).

Lindblom, Björn, Bertil Lyberg et Karin Holmgren (1981). *Durational patterns of Swedish phonology : do they reflect short-term motor memory processes ?* T. 3. Indiana University Linguistics Club (page 102).

Maddieson, Ian (1984). « Phonetic cues to syllabification ». Dans : *UCLA Working papers in phonetics* 59, p. 85–101 (page 102).

Magen, Harriet S et Sheila E Blumstein (1993). « Effects of speaking rate on the vowel length distinction in Korean ». Dans : *Journal of Phonetics* 21, p. 387–410 (pages 102, 103).

Massey Jr, Frank J (1951). « The Kolmogorov-Smirnov test for goodness of fit ». Dans : *Journal of the American statistical Association* 46.253, p. 68–78 (page 122).

Myers, Scott (2005). « Vowel duration and neutralization of vowel length contrasts in Kinyarwanda ». Dans : *Journal of Phonetics* 33.4, p. 427–446 (page 102).

Nemoto, Rena (2011). « Large-scale acoustic and prosodic investigations of french ». Thèse de doctorat. Paris 11 (page 99).

Newman, Paul (2000). « The Hausa Language ». Dans : *An Encyclopedic Reference Grammar*. url : http://tocs.ub.uni-mainz.de/pdfs/09159748X.pdf (pages 32–34, 106).

Newman, Roxana Ma et Vincent J van Heuven (1981). « An acoustic and phonological study of pre-pausal vowel length in Hausa ». Dans : *Journal of African Languages and Linguistics* 3.1, p. 1–18 (pages 34, 106).

Oh, Yoon Mi, François Pellegrino, Christophe Coupé et Egidio Marsico (2013). « Cross-language comparison of functional load for vowels, consonants, and tones ». Dans : *Interspeech*. Citeseer, p. 3032–3036 (page cxxxii).

Pellegrino, François, Egidio Marsico et Christophe Coupé (2012). « La typologie des systèmes vocaliques revisitée sous l'angle de la charge fonctionnelle ». Dans : *Proc. of the Joint Conference JEP-TALNRECITAL*. T. 1, p. 617–624 (page cxxxii).

Povey, Daniel (2004). « Phone duration modeling for LVCSR ». Dans : *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. T. 1. IEEE, p. I–829 (page 103).

Pylkkönen, Janne et Mikko Kurimo (2004). « Duration modeling techniques for continuous speech recognition. » Dans : *Proceedings of Interspeech 2004* (pages 103, 104).

Saussure, Ferdinand de (1952). *Cahiers Ferdinand de Saussure*. T. 10. Société genevoise de linguistique (page 120).

Sauvageot, Serge (1965). *Description synchronique d'un dialecte wolof : le parler du Dyolof*. 73. Institut Français d'Afrique Noire, Dakar. (pages 40, 114).

Scharenborg, Odette, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel *et al.* (2018). « Linguistic unit discovery from multi-modal inputs in unwritten languages : Summary of the" Speaking Rosetta" JSALT 2017 Workshop ». Dans : *arXiv preprint arXiv :1802.05092* (page cxxxii).

Schlippe, Tim, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs et Tanja Schultz (2012). « Hausa large vocabulary continuous speech recognition. » Dans : *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU)*, p. 11–14 (pages 33, 83, 84, 86, 104).

Sock, Rudolph (1983). « L'organisation temporelle de l'opposition de quantité vocalique en Wolof de Gambie. Sa résistivité aux conditions de durée segmentales et suprasegmenales. » Thèse de doctorat (pages 103, 119, 121).

Surendran, Dinoj et Partha Niyogi (2003). « Measuring the usefulness (functional load) of phonological contrasts ». Dans : *University of Chicago* (page cxxxii).

Vasilescu, Ioana, Bianca Vieru et Lori Lamel (2014). « Exploring pronunciation variants for romanian speech-to-text transcription ». Dans : *Spoken Language Technologies for Under-Resourced Languages* (page 99).

Viterbi, Andrew (1967). « Error bounds for convolutional codes and an asymptotically optimum decoding algorithm ». Dans : *IEEE transactions on Information Theory* 13.2, p. 260–269 (page 104).

Zanon Boito, Marcely, Alexandre Bérard, Aline Villavicencio et Laurent Besacier (2017). « Unwritten Languages Demand Attention Too ! Word Discovery with Encoder-Decoder Models ». Dans : *arXiv preprint arXiv :1709.05631* (page cxxxii).

Index

ANNEXE A

Bibliographie personnelle

Gauthier, Elodie, Laurent Besacier et Sylvie Voisin (février 2018). « Machine Assistance for Large Scale Phonetics ». Dans : *Computational Methods for Endangered Language Documentation and Description*. Poster. Paris, France.

— (mars 2018). « LIG-AIKUMA, une application mobile pour la collecte de parole sur le terrain ». Dans : *Du terrain à la théorie : Les 40 ans du Lacito*. Paris, France. *À paraître.*

— (2017). « Machine Assisted Analysis of Vowel Length Contrasts in Wolof ». Dans : *Interspeech 2017*. Stockholm, Suède. URL : http://www.isca-speech.org/archive/Interspeech_2017/pdfs/0268.PDF

— (2016). « Speed Perturbation and Vowel Duration Modeling for ASR in Hausa and Wolof Languages ». Dans : *Interspeech 2016*. San Francisco, USA. URL : http://www.isca-speech.org/archive/Interspeech_2016/pdfs/0461.PDF

Gauthier, Elodie, David Blachon, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda et Grégoire Bachman (2016). « LIG-AIKUMA : a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies ». Dans : *Interspeech 2016 (Show and Tell)*. San Francisco, USA. URL : http://www.isca-speech.org/archive/Interspeech_2016/pdfs/2003.PDF

Gauthier, Elodie, Laurent Besacier, Sylvie Voisin, Michael Melese et Uriel Pascal Elingui (2016). « Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition : a Case Study of Wolof ». Dans : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovénie. URL : http://www.lrec-conf.org/proceedings/lrec2016/pdf/460_Paper.pdf

Gauthier, Elodie, Laurent Besacier, et Sylvie Voisin (2016). « Automatic Speech Recognition for African Languages with Vowel Length Contrast ». Dans : *Procedia Computer Science 81*, p. 136–143. Yogyakarta, Indonésie. URL : https://hal.archives-ouvertes.fr/hal-01350040/document

Blachon, David, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker et Annie Rialland (2016). « Parallel Speech Collection for Under-Resourced Language Studies using the LIG-Aikuma Mobile Device App ». Dans : *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*. Yogyakarta, Indonésie. url : https://www.sciencedirect.com/science/article/pii/S1877050916300448/pdf?md5=c3f4c1d33c04f6e17e068c4a9abec35e&pid=1-s2.0-S1877050916300448-main.pdf

Blachon, David, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker et Annie Rialland (2016). « Collecte de parole pour l'étude des langues peu dotées ou en danger avec l'application mobile Lig-Aikuma ». Dans : *Atelier Traitement Automatique des Langues Africaines (TALAF). JEP-TALN 2016*. Paris, France. url : http://talaf.imag.fr/2016/Actes/BLACHON_ET_AL%20-%20Collecte%20de%20parole%20pour%20l'%C3%A9tude%20des%20langues%20peu%20dot%C3%A9es%20ou%20en%20danger%20avec%20l'application%20mobile%20Lig-Aikuma.pdf

Besacier, Laurent, Elodie Gauthier, Mathieu Mangeot, Philippe Bretier, Paul Bagshaw, Olivier Rosec, Thierry Moudenc, François Pellegrino, Sylvie Voisin, Egidio Marsico et Pascal Nocera (2015). « Speech Technologies for African Languages : Example of a Multilingual Calculator for Education ». Dans : *Interspeech 2015 (Show and Tell)*. Dresde, Allemagne. url : http://www.isca-speech.org/archive/interspeech_2015/papers/i15_1886.pdf

ANNEXE **B**

Exemple de fichier JSON généré par LIG-AIKUMA

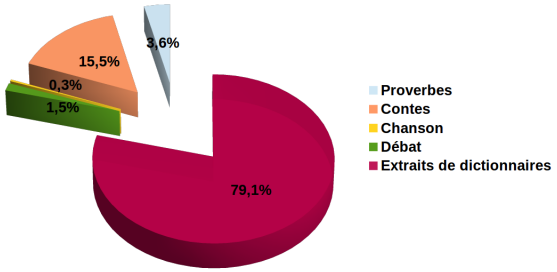
{”speaker_name” :“mamadou cisse”,
“speaker_birth_year” :1986,
“Format” :“vnd.wave”,
“mother_tongue” :“”,
“location” :null,
“recording_lang” :“wol”,
“origin” :“yénn (lébou)”,
“androidID” :“58abbdb5a1eb5867”,
“speakers” :[],
“date” :“2016-12-04T11 :16 :02Z”,
“suffix” :“161204-111602_wol_58a_elicit”,
“file_type” :null,
“version” :“v01”,
“sampleRate” :16000,
“languages” :[],
“source” :“v01-tnrxjuwkbbma”,
“BitsPerSample” :16,
“item_id” :“tnrxjuwkbbma”,
“speaker_gender” :“Male”,
“name” :“161204-111602_wol_58a_elicit_0”,
“device” :“SAMSUNG-SM-T530”,
“NumChannels” :1,
“user_id” :“1524218440”,
“durationMsec” :4968}

ANNEXE C

Corpus textuel du wolof

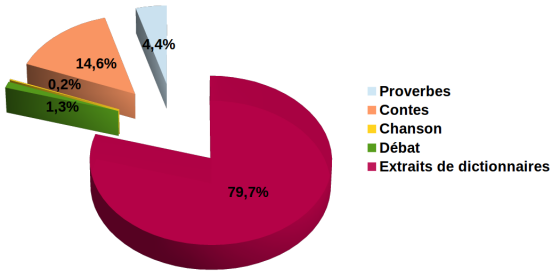
TABLEAU C.1 – Répartition du corpus textuel dans chaque corpus du système de RAP.

	Nb. de lignes
Proverbes	507
Contes	2 168
Chanson	46
Débat	204
Extraits de dictionnaires	11 075
TOTAL	**14 000**



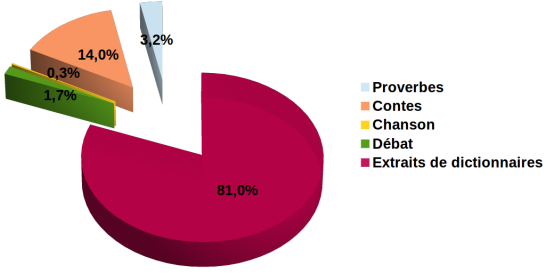
(a) Répartition dans le corpus d'entraînement (*train*).

	Nb. de lignes
Proverbes	87
Contes	291
Chanson	4
Débat	25
Extraits de dictionnaires	1 593
TOTAL	**2 000**



(b) Répartition dans le corpus de développement (*dev*).

	Nb. de lignes
Proverbes	63
Contes	280
Chanson	5
Débat	33
Extraits de dictionnaires	1 619
TOTAL	**2 000**



(c) Répartition dans le corpus d'évaluation (*test*).

ANNEXE D

Corpus de parole du wolof

Tableau D.1 – Récapitulatif des locuteurs wolof du corpus de parole récolté.

ID	Locuteur	Genre	Profession	Langue maternelle	Âge	Durée totale
01	GF	H	Étudiant	SER	29	01 :12 :07.68
04	RB	H	Étudiant	FRA	24	01 :19 :53.17
05	RLF	H	Journaliste	FRA	30	01 :16 :30.90
06	PID	H	Journaliste	WOL	34	01 :07 :01.89
08	MD	H	Enseignant	WOL	48	01 :11 :14.33
11	OD	H	Étudiant	SER	29	01 :18 :17.14
15	TD	H	Journaliste	WOL	44	01 :10 :37.14
16	MN	H	Gérant	WOL	32	01 :04 :54.48
17	OS	H	Étudiant	WOL	28	01 :03 :16.24
18	CT	H	Étudiant	WOL	27	01 :05 :19.58
RÉSUMÉ				**6/10 wolof natifs**	**33**	**11 :49 :10**
02	SS	F	Journaliste	WOL	33	01 :10 :26.23
03	HG	F	Étudiante	WOL	28	01 :05 :26.70
07	MSN	F	Étudiante	WOL	28	01 :08 :53.46
09	NCN	F	Téléoperatrice	WOL	27	01 :21 :43.53
10	SD	F	Journaliste	WOL	38	01 :04 :27.14
12	OL	F	Étudiante	WOL	28	01 :16 :09.62
13	SDI	F	Étudiante	WOL	32	01 :20 :03.69
14	NAF	F	Étudiante	WOL	29	01 :07 :04.34
RÉSUMÉ				**8/8 wolof natives**	**30**	**09 :34 :11**

Tableau D.2 – Durée totale des corpus de parole utilisés pour entraîner et tester le système de RAP à grand vocabulaire du wolof.

	Durée totale
TRAIN	16 :49 :52
DEV	02 :12 :28
TEST	02 :20 :58
Total	**21 :23 :27**

ANNEXE **E**

Formulaire de consentement

Formulaire de consentement

Je soussigné(e) Nath déclare accepter librement, et de façon éclairée, de participer comme sujet à l'étude dans le cadre du .....
.....

Le linguiste, ..., en charge de la collecte, s'engage à protéger l'intégrité du participant tout au long de l'étude. Le participant donne l'autorisation au linguiste de partager les enregistrements avec la communauté des chercheurs ou des locuteurs de la langue pour des fins culturels et scientifiques. Le linguiste s'engage à ne pas diffuser des contenus potentiellement sensibles. Le participant peut décider à tout moment de restreindre l'accès aux enregistrements.

Le consentement pour poursuivre la recherche peut être retiré à tout moment sans donner de raison et sans encourir aucune responsabilité ni conséquence. Les réponses aux questions ont un caractère facultatif et le défaut de réponse n'aura aucune conséquence pour le sujet.

Le participant a la possibilité d'obtenir des informations supplémentaires concernant cette étude auprès du linguiste, et ce dans les limites des contraintes du plan d'étude. Toutes les informations concernant le participant sont conservées de façon anonyme et confidentielle.

Le linguiste s'engage à préserver absolument la confidentialité pour toutes les informations concernant le participant.

Le participant : Nath

Le linguiste :

Fait le 25/05/2017 à:

Fait le 25/05/2017 à:

Signature :

Signature :

ANNEXE **F**

Corpus d'élicitation par texte

Variantes dialectales

Quelle est sa dimension ? ~ Quelle est sa taille ? ##
Veux-tu que je te donne sa taille ? ##
Je ne peux supporter de telles injures. ##
C'est ici que je m'arrête. ##
Il a vendu à peu près 100 kg de poissons. ##
J'ai trié 10kg d'arachides et ma sœur autant. ##
Il est passé à peu près 10 fois devant chez moi ce matin. ##
Je te donne 1 000 CFA, ni plus ni moins. ##
Il est de ma taille. ##
Mon jardin est plus grand que le tien. ##
Ce poisson est énorme, je n'en ai jamais vu de cette taille. ##
Je n'irais pas plus loin ! Je m'arrête là. ## (dans le sens de ça suffit)

Collectif

Les poissons se font de plus en plus rares sur les côtes. ##
Cette chose est grande. ##
Les pêcheurs connaissent bien la mer. ##
Des montons sont dans la rue. ##
Des jeunes filles s'amusent dans la cour. ##
Les abeilles « butinent »les fleurs. ##
Les souris ont des queues plus petites que celles du rat. ##
les abeilles font du miel. ##
L'abeille m'a piqué sur le bras. ##
Les pêcheurs sont partis en la pirogue ce matin. ##
Ces personnes vivent ici. ##
Le poisson est de plus en plus cher. ##
Les choses en plastique n'existaient pas avant. ##
L'invité est reparti. ##
Le mouton s'est échappé. ##
Ces choses sont rouges. ##

L’ânier garde les ânes. ##
Le vacher grade les vaches. ##
Le berger garde les moutons. ##
Deux abeilles m’ont piqué le pied. ##
Les femmes portent souvent des pagnes. ##
Le berger rentre ses montons. ##
Les hommes doivent subvenir aux besoins de leur famille. ##
Les deux souris se sont enfuies par ce trou. ##
Les familles se sont retrouvées. ##
L’esclave a été battu. ##
Cette personne est bavarde. ##
Les bergers ne sont pas tous des peuls. ##
les jeunes filles doivent trouver de bon mari. ##
Deux hommes s’approchent. ##
En été, les mangues sont mûres. ##
Ma sœur aime les mangues qui sont vertes. ##
Les deux arbres de ce champs seront coupés. ##
Les esclaves sont en train de cultiver. ##
La jeune fille va à l’école. ##
Les beaux moutons font de belles offrandes. ##
Les bergers sont des peuls. ##
les moutons mangent de l’herbe. ##
Les arbres ont besoin d’eau. ##
L’arbre est tombé. ##
il m’a dit des choses gentilles. ##
L’homme est monté sur l’échelle. ##
La famille est sacrée. ##
Les invités sont partis. ##
Il me disait toujours des choses qui étaient gentilles. ##
Des hommes sont entrés dans cette maison. ##
Les deux esclaves se sont échappés. ##
Les deux moutons ont été retrouvés. ##
il me dit toujours des choses qui sont gentilles. ##
Les enfants vont à l’école. ##
Les mangues dans la cuisine sont prêtes à manger. ##
Les enfants jouent au foot. ##
Toute la famille est venue pour le baptême. ##
Les abeilles sont en train de fabriquer leur ruche. ##
Des femmes travaillent. ##
Les deux mangues ont été mangées. ##
les abeilles de forêt font un meilleur miel. ##

Deux choses restent à faire. ##
Les gens ont besoin d'eau pour vivre. ##
Deux jeunes filles s'en vont. ##
La mangue est mûre. ##
Cette femme est belle. ##
Les souris sont dans le champs. ##
Les esclaves doivent obéir. ##
Des bergers sont venus en ville. ##
Les arbres ont été planté sur ce terrain. ##
La souris grignote cette graine. ##

Triade « à, a, aa »

Je veux mordre. ##
Je veux frotter cette tache. ##
Je veux un terrain nu. ##
Je veux un melon. ## (une pastèque)
J'ai une infirmité. ##
Je veux une branche. ##
Je veux décorer cette clôture. ##
Je veux rouler une cigarette. ##
Je veux un sac. ##
Je veux une partie de ce territoire. ##
Je veux un grand-père. ##
Je veux être pied nu. ##
Je veux projeter une lance. ##
Je veux un clou. ##
Je veux préparer une bouillie de mil à l'arachide. ##
J'aime les obstacles ~ difficultés. ##
Je veux faire un croc en jambe. ##
Je veux une pirogue. ##
Je veux tisser. ##
Je veux obtenir un gain. ##
Je veux une sœur. ##
Je veux tomber dans ce trou. ##
Je veux viser. ##
Je suis surexcité. ##
Je veux qu'il prenne feu, qu'il s'allume. ##
Je veux descendre. ##
Je veux présenter des condoléances. ##
Je veux une scarification. ##

Je veux tomber. ##
(wadd) Je veux du bois de chauffe. ##
Je suis de mauvaise humeur. ##
Je veux brûler ce bâton. ##
Je veux du maïs. ##
Je veux enlever les croûtes d'une plaie. ##
Je veux vomir. ##
Je veux qu'il se déchire. ## (qu'il craque, pour un sac par exemple)
Je veux créer. ##
Je veux aller avec eux. ##
Je veux griffer. ##
Je veux faire l'éloge des ancêtres. ##
Je veux boucher le trou. ##
Je veux couper. ##
Je veux choisir. ##
Je veux investir. ## (placer de l'argent)
J'ai la nostalgie de ma jeunesse. ##
Je veux abandonner. ##
Je veux une part. ##
Je veux porter ce bracelet. ##
Je veux malmener cet enfant. ## (maat-maate)
J'aime la brousse. ##
Je veux couper en petit. ## (daggati)
Je suis enflé. ## (boursoufflé, bouffi)
Je veux quitter. ##
Je veux jurer. ##
Je veux une clé. ##
Je veux piler sommairement. ##
Je veux écoper la pirogue. ##
Je veux des chaussures. ##
Je veux une grande cuillère en bois. ##
J'ai mal aux épaules. ##
Je veux dépasser les bornes. ##
Je veux une palissade. ## (clôture)
J'ai échouer à une épreuve. ##
J'ai un abcès. ##
Je veux barboter. ##
Je veux réussir en quelque chose. ##
Je veux parler une langue étrangère. ##
Je veux porter sur l'épaule. ##
Je veux qu'il s'arrête je veux que ça cesse. ##

Je veux épouser cet homme. ##
Je veux me courber. ## (me baisser)
Je veux un chapeau de paille. ##
Je veux un nid d'oiseau. ##
Je veux injurier. ##
Je veux être normal. ##
Je veux changer de direction. ##
Je veux galoper. ##
Je veux une grappe. ##
Je veux prêter main forte. ##
J'aime cette voix. ##
Je veux demander. ##
Je veux battre l'arachide. ##
Je suis malingre je suis chétif ## (ràgg)
Je veux un brûle encens. ##
Je veux une gourde en calebasse. ##
Je veux somnoler. ##
Je veux frotter. ##
Je veux passer. ##
Je veux un petit tas d'arachide. ##
J'ai paré le coup de poing. ##
Je veux une corde en fibre. ##
J'ai une croûte. ##
Je veux du cuir brut. ##
Je veux couvrir le toit d'une maison. ##
Je veux attiser le feu. ##
Je veux un bâton. ##
Je veux aiguiser le couteau. ##
J'aime faucher les jambes. ## (pour un lutteur)
Je veux démonter cette clôture. ##
Je veux un champ de manioc. ##
Je veux un collier de perles.##

ANNEXE G

Performance par locuteur du système de RAP de référence du haoussa

Le tableau G.1 expose les scores brut et en pourcentage obtenus pour chaque locuteur du corpus évalué.

Tableau G.1 – Sortie de l'outil *sclite* des résultats par locuteurs, du système de RAP de référence, à base de DNN, des corpus d'évaluation *dev* et *test* du haoussa.

Locuteur	Type	#phrase	#Mot	Corr	Sub	Ins	Del	Err	P.Err
018	brut	102	622	594	24	8	4	36	23
018	%	102	622	95,50	3,86	1,29	0,64	5,79	22,55
031	brut	104	710	671	35	6	4	45	26
031	%	104	710	94,51	4,93	0,85	0,56	6,34	25,00
034	brut	96	607	565	37	7	5	49	32
034	%	96	607	93,08	6,10	1,15	0,82	8,07	33,33
038	brut	105	610	576	30	8	4	42	24
038	%	105	610	94,43	4,92	1,31	0,66	6,89	22,86
046	brut	100	587	555	27	8	5	40	23
046	%	100	587	94,55	4,60	1,36	0,85	6,81	23,00
047	brut	96	577	510	56	4	11	71	38
047	%	96	577	88,39	9,71	0,69	1,91	12,31	39,58
050	brut	100	585	526	48	3	11	62	37
050	%	100	585	89,91	8,21	0,51	1,88	10,60	37,00
055	brut	110	699	652	37	3	10	50	31
055	%	110	699	93,28	5,29	0,43	1,43	7,15	28,18
058	brut	108	668	629	33	7	6	46	30
058	%	108	668	94,16	4,94	1,05	0,90	6,89	27,78
072	brut	100	628	570	44	7	14	65	35
072	%	100	628	90,76	7,01	1,11	2,23	10,35	35,00
TOTAL	brut	1021	6293	5848	371	61	74	506	299
TOTAL	%	1021	6293	92,93	5,90	0,97	1,18	8,04	29,29

(a) Scores obtenus par locuteurs du corpus *dev*.

Locuteur	Type	#phrase	#Mot	Corr	Sub	Ins	Del	Err	P.Err
002	brut	104	752	710	37	26	5	68	38
002	%	104	752	94,41	4,92	3,46	0,66	9,04	36,54
014	brut	70	421	390	24	2	7	33	20
014	%	70	421	92,64	5,70	0,48	1,66	7,84	28,57
025	brut	105	617	547	65	11	5	81	50
025	%	105	617	88,65	10,53	1,78	0,81	13,13	47,62
028	brut	110	677	621	50	8	6	64	41
028	%	110	677	91,73	7,39	1,18	0,89	9,45	37,27
030	brut	111	679	539	121	10	19	150	67
030	%	111	679	79,38	17,82	1,47	2,80	22,09	60,36
052	brut	105	627	576	43	8	8	59	32
052	%	105	627	91,87	6,86	1,28	1,28	9,41	30,48
053	brut	101	609	554	47	7	8	62	37
053	%	101	609	90,97	7,72	1,15	1,31	10,18	36,63
062	brut	104	616	563	42	10	11	63	25
062	%	104	616	91,40	6,82	1,62	1,79	10,23	24,04
070	brut	101	593	539	46	5	8	59	33
070	%	101	593	90,89	7,76	0,84	1,35	9,95	32,67
088	brut	100	607	552	42	4	13	59	32
088	%	100	607	90,94	6,92	0,66	2,14	9,72	32,00
TOTAL	brut	1011	6198	5591	517	91	90	698	375
TOTAL	%	1011	6198	90,21	8,34	1,47	1,45	11,26	37,09

(b) Scores obtenus par locuteurs du corpus *test*.

ANNEXE H

Performance par locuteur du système de RAP de référence du wolof

Le tableau H.1 expose les scores brut et en pourcentage obtenus pour chaque locuteur du corpus évalué.

Tableau H.1 – Sortie de l'outil *sclite* des résultats par locuteurs, du système de RAP de référence, à base de DNN, des corpus d'évaluation *dev* et *test* originaux du wolof.

Locuteur	Type	#phrase	#Mot	Corr	Sub	Ins	Del	Err	P.Err
03	brut	1000	9390	7098	2051	275	241	2567	868
03	%	1000	9390	75,59	21,84	2,93	2,57	27,34	86,80
06	brut	1000	9414	7326	1873	242	215	2330	862
06	%	1000	9414	77,82	19,90	2,57	2,28	24,75	86,20
TOTAL	brut	2000	18804	14424	3924	517	456	4897	1730
TOTAL	%	2000	18804	76,71	20,87	2,75	2,43	26,04	86,50

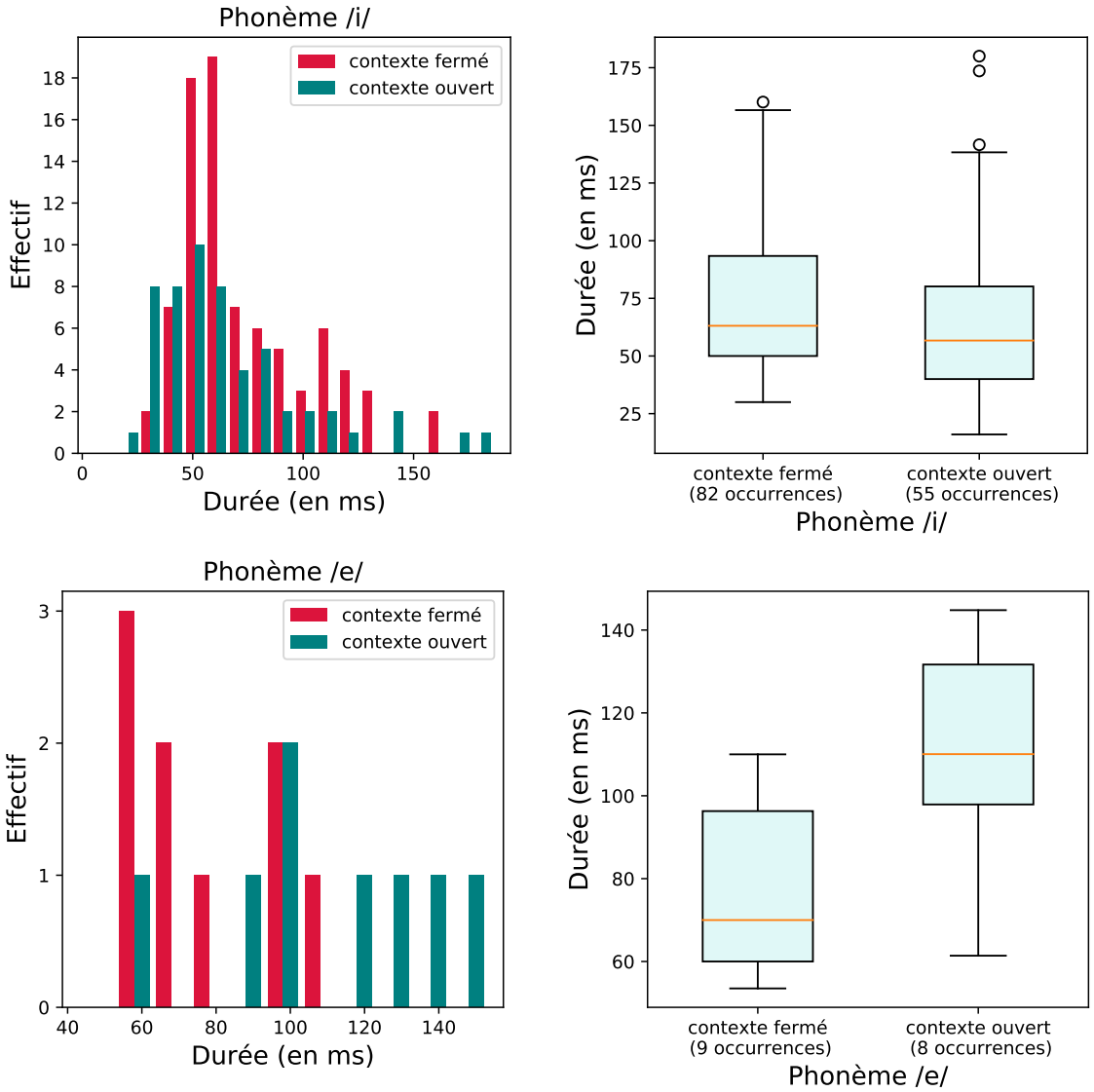
(a) Scores obtenus par locuteurs du corpus *dev*.

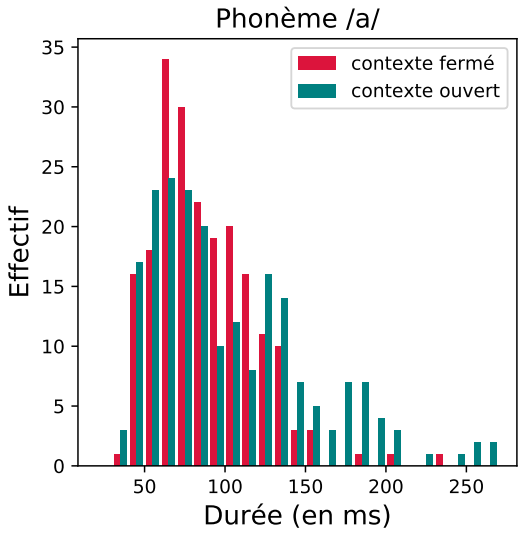
Locuteur	Type	#phrase	#Mot	Corr	Sub	Ins	Del	Err	P.Err
05	brut	1000	9373	6473	2632	326	268	3226	949
05	%	1000	9373	69,06	28,08	3,48	2,86	34,42	94,90
10	brut	1000	9482	6861	2197	199	424	2820	904
10	%	1000	9482	72,36	23,17	2,10	4,47	29,74	90,40
TOTAL	brut	2000	18855	13334	4829	525	692	6046	1853
TOTAL	%	2000	18855	70,72	25,61	2,78	3,67	32,07	92,65

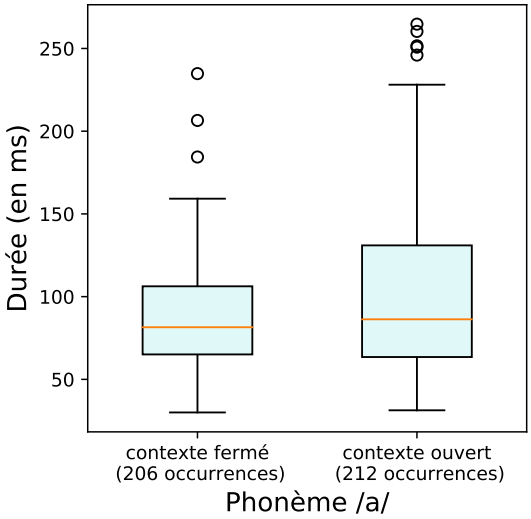
(b) Scores obtenus par locuteurs du corpus *test*.

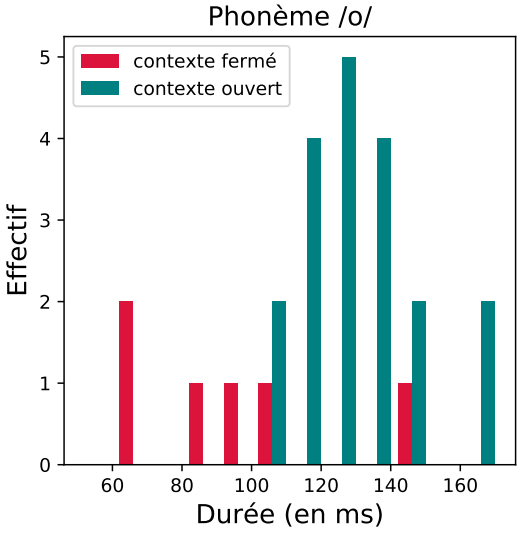
ANNEXE I

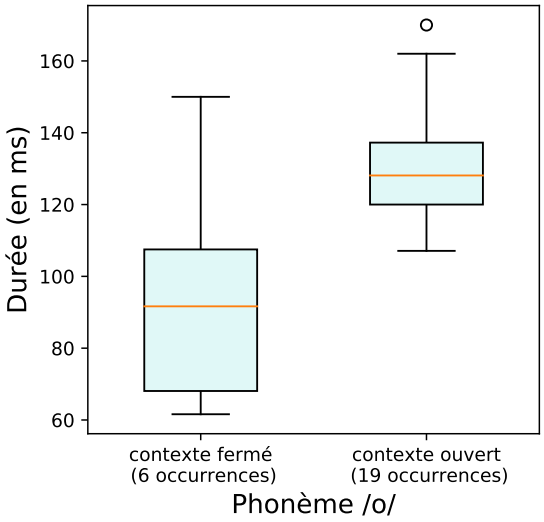
Alignements manuels de 102 fichiers en haoussa

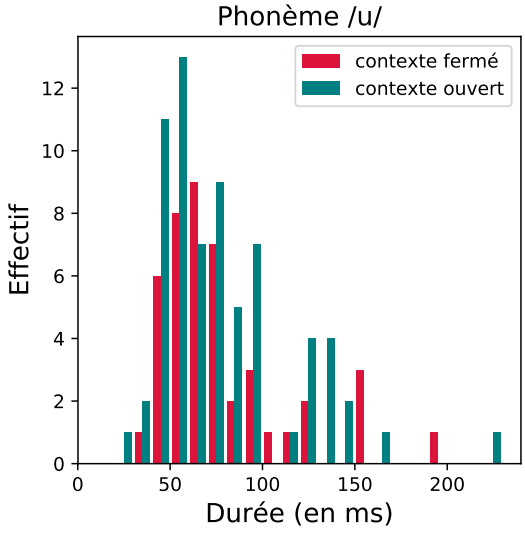


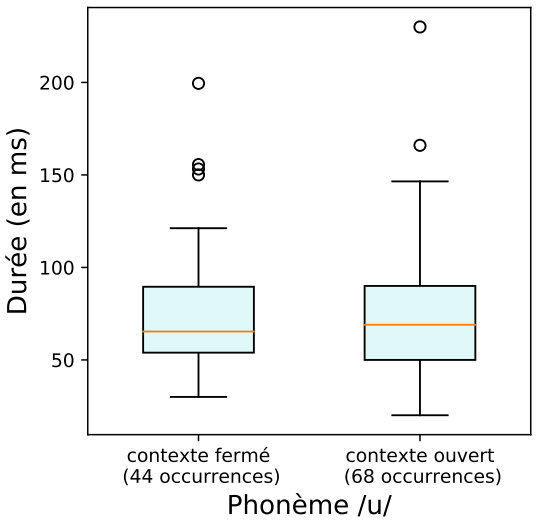

Phonème /a/
contexte fermé
contexte ouvert
Effectif
Durée (en ms)


Durée (en ms)
contexte fermé
(206 occurrences)
contexte ouvert
(212 occurrences)
Phonème /a/


Phonème /o/
contexte fermé
contexte ouvert
Effectif
Durée (en ms)


Durée (en ms)
contexte fermé
(6 occurrences)
contexte ouvert
(19 occurrences)
Phonème /o/

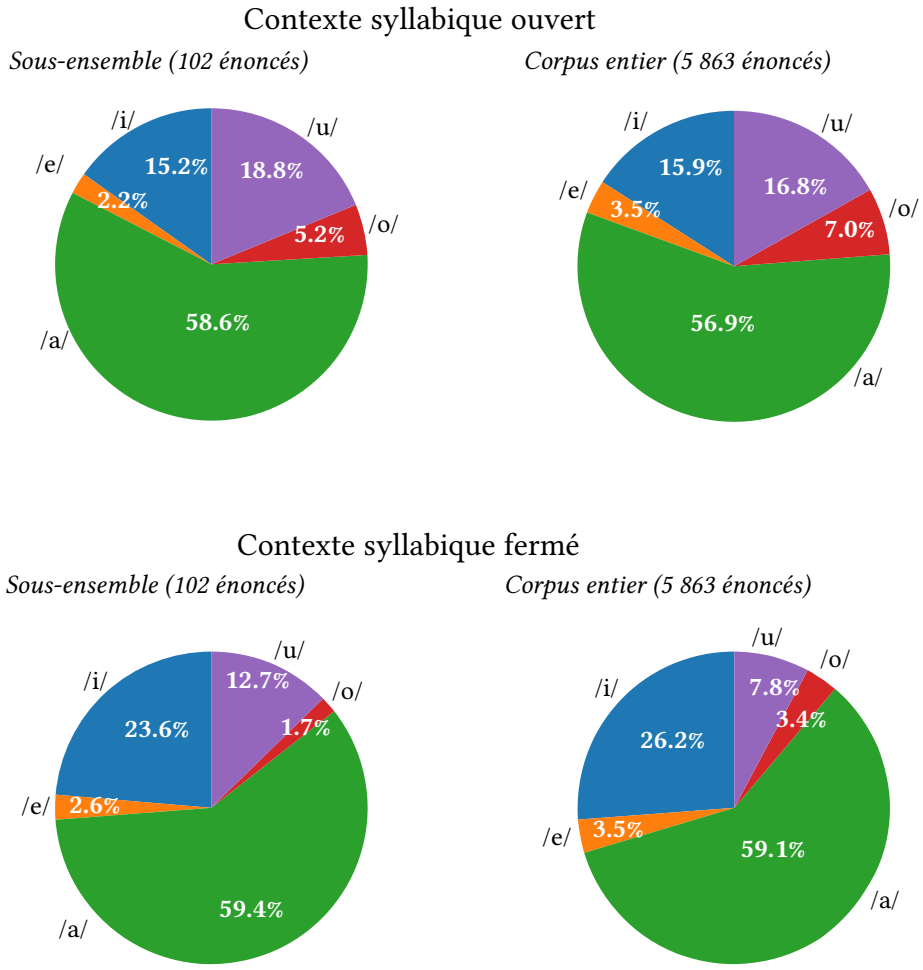

Phonème /u/
contexte fermé
contexte ouvert
Effectif
Durée (en ms)


Durée (en ms)
contexte fermé
(44 occurrences)
contexte ouvert
(68 occurrences)
Phonème /u/

ANNEXE J

Répartition des voyelles dans les corpus du haoussa

Corpus d'apprentissage du haoussa



GRAPHIQUE J.1 – Répartition des voyelles du haoussa analysées dans le corpus d'apprentissage entier (5 863 énoncés) et dans le sous-ensemble manuellement réaligné (102 énoncés).

Le GRAPHIQUE J.1 présente la répartition des voyelles dans le corpus d'apprentissage (contenant 5 863 énoncés) et dans le sous-ensemble de 102 énoncés sélectionné dans ce corpus. Les voyelles sont réparties selon leur contexte syllabique de réalisation. Finalement, cette figure montre que la répartition des voyelles dans chaque corpus est homogène. Les tendances analysées au niveau du contraste de durées des voyelles dans le corpus total sont comparables à celles observées dans le sous-corpus réaligné manuellement.

ANNEXE K

Comparaison des approches SGMM et DNN pour l'analyse des durées de voyelles en haoussa

TABLEAU K.1 – Comparaison des approches SGMM et DNN pour l'analyse des durées de voyelles en haoussa. Les durées sont obtenues par alignement forcé des données d'apprentissage, à partir d'une modélisation acoustique sans prise en compte des caractéristiques de longueur vocalique. L'étiquetage des voyelles en fonction de leur contexte syllabique a été effectué *a posteriori*. Les mesures présentées sont les durées moyennes de chaque voyelle, en millisecondes, selon le contexte syllabique ouvert ou fermé dans lequel elle est réalisée. Les deltas sont calculés sur ces durées moyennes.

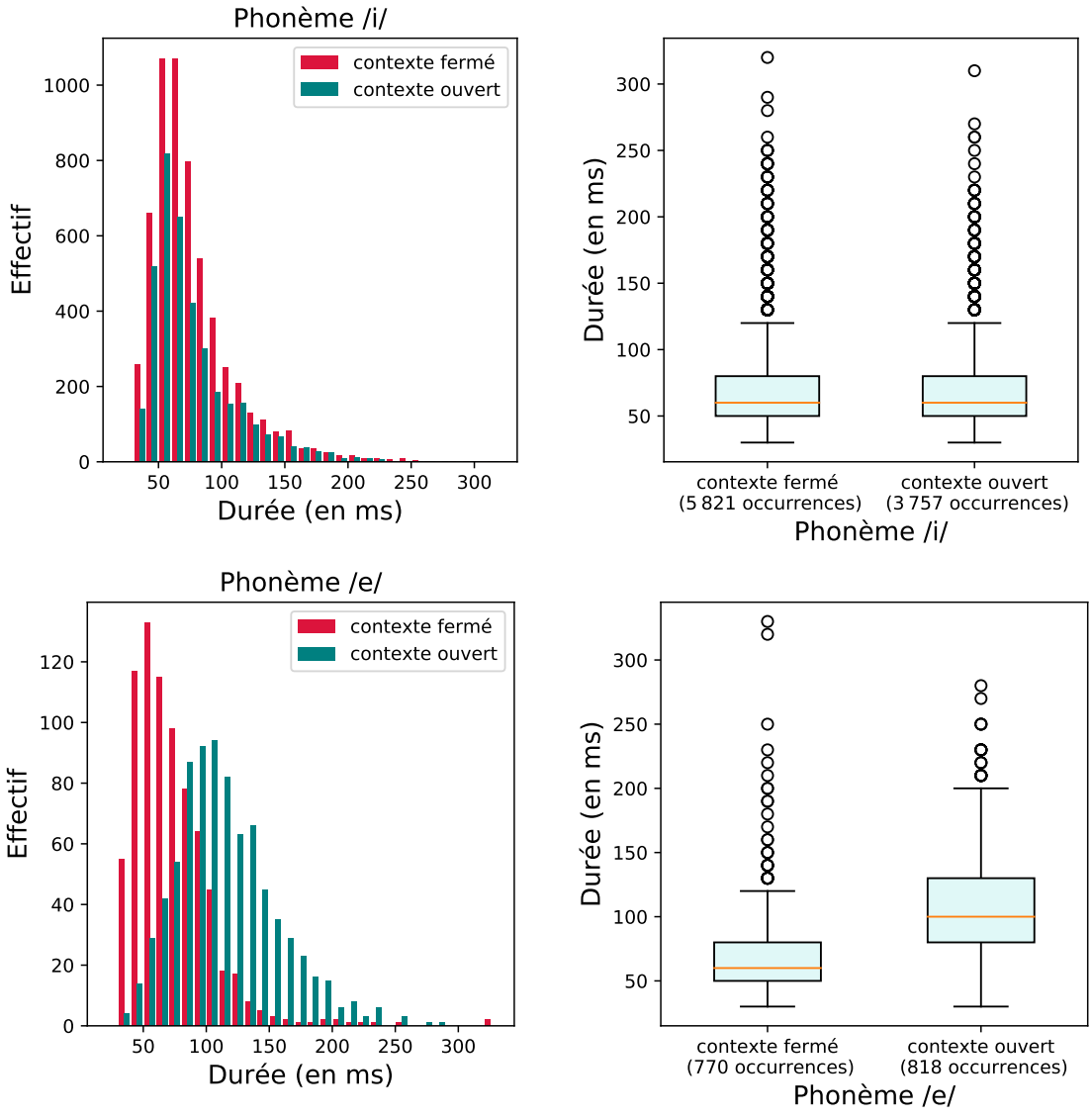
Modèle acoustique	**Contexte syllabique**	**Durée moyenne (en ms)**				
		i	**e**	**a**	**o**	**u**
CD-HMM/SGMM	ouvert	71,7	111,5	83,1	123,0	78,5
	fermé	72,2	68,1	74,5	83,7	71,3
	delta (Δ)	*-0,5*	*43,4*	*8,6*	*39,2*	*7,2*
CD-HMM/DNN	ouvert	71,2	111,0	83,6	121,5	78,5
	fermé	72,7	69,3	74,6	84,6	72,4
	delta (Δ)	*-1,5*	*41,7*	*9,1*	*36,9*	*6,1*
Delta des modèles ($\lvert SGMM - DNN \rvert$)		**1,0**	**1,6**	**0,4**	**2,3**	**1,0**

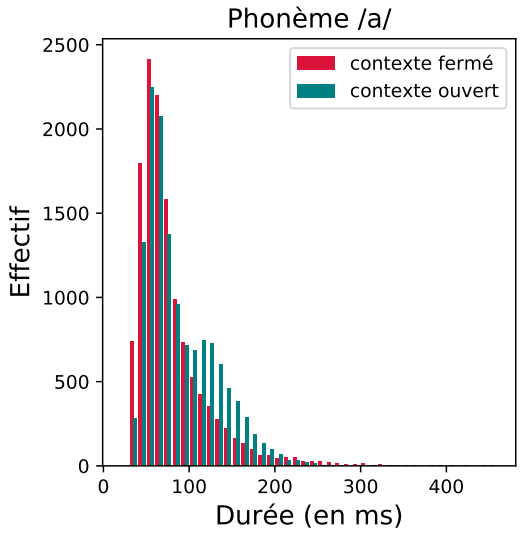
Le tableau K.1 conclut que les deux approches explorées apportent des estimations de durées similaires. L'opposition de longueur est fortement marquée pour les voyelles /e/ et /o/ par les deux approches, tandis que les autres voyelles sont moins contrastées.

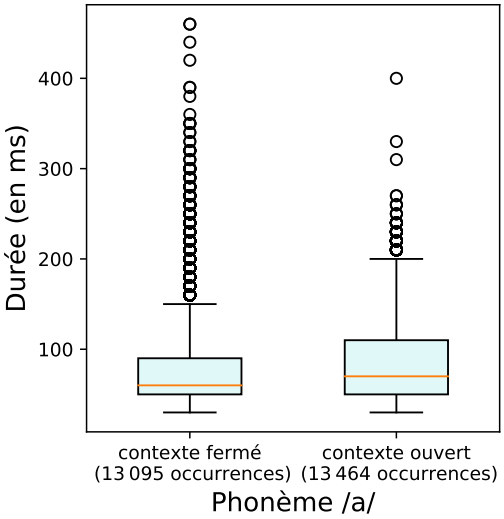
ANNEXE L

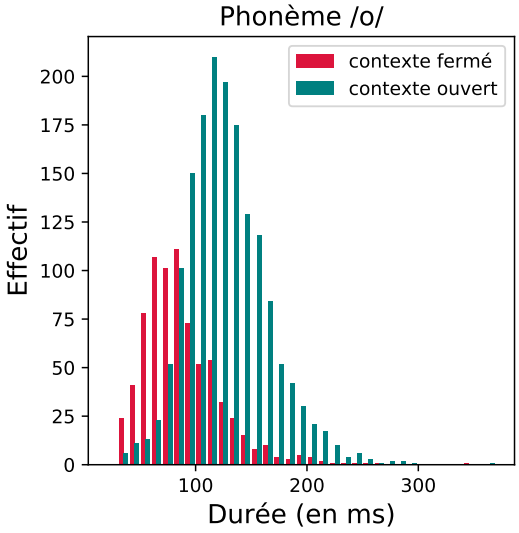
Alignements forcés du corpus d'apprentissage du haoussa (5 863 fichiers)

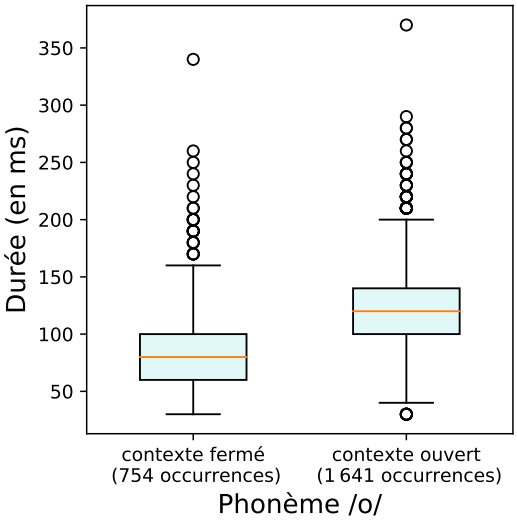
Les figures ci-après présentent les distributions statistiques des durées (en millisecondes) des cinq voyelles /i/, /e/, /a/, /e/ et /u/ du haoussa selon le contexte syllabique de la voyelle prononcée (parole lue). Ces durées ont été extraites des alignements forcés émis par approche CD-HMM/SGMM (sans modélisation des longueurs) et l'information sur le contexte syllabique a été annotée *a posteriori*.

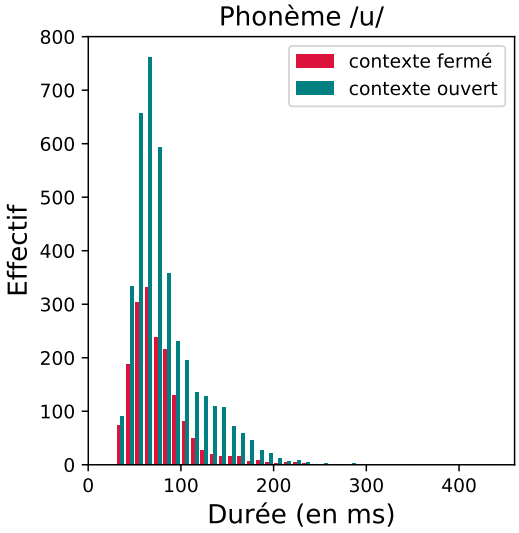


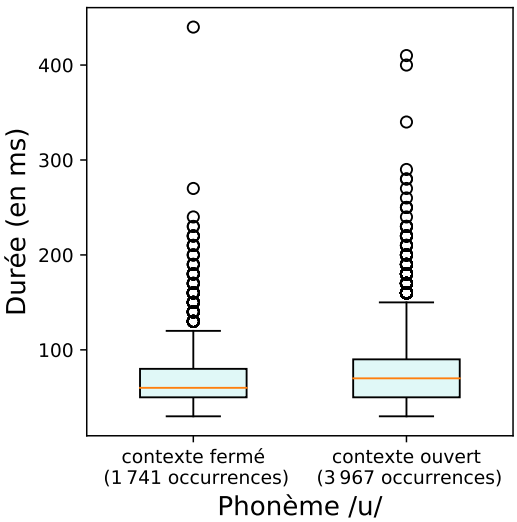

Phonème /a/
contexte fermé
contexte ouvert
Effectif
Durée (en ms)


Durée (en ms)
contexte fermé
(13 095 occurrences)
contexte ouvert
(13 464 occurrences)
Phonème /a/


Phonème /o/
contexte fermé
contexte ouvert
Effectif
Durée (en ms)

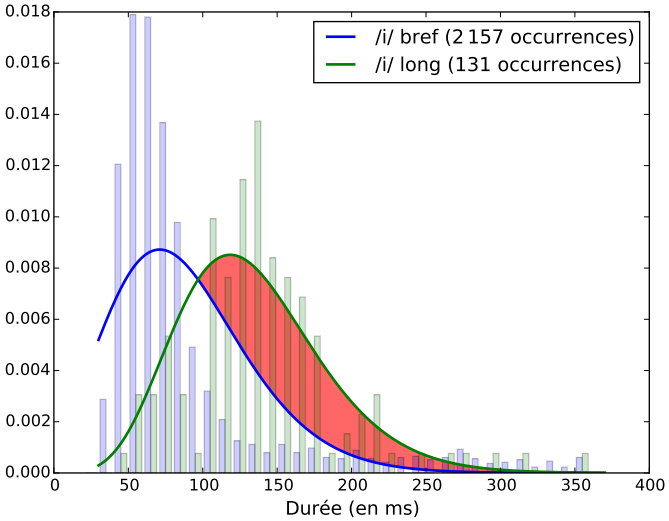

Durée (en ms)
contexte fermé
(754 occurrences)
contexte ouvert
(1 641 occurrences)
Phonème /o/

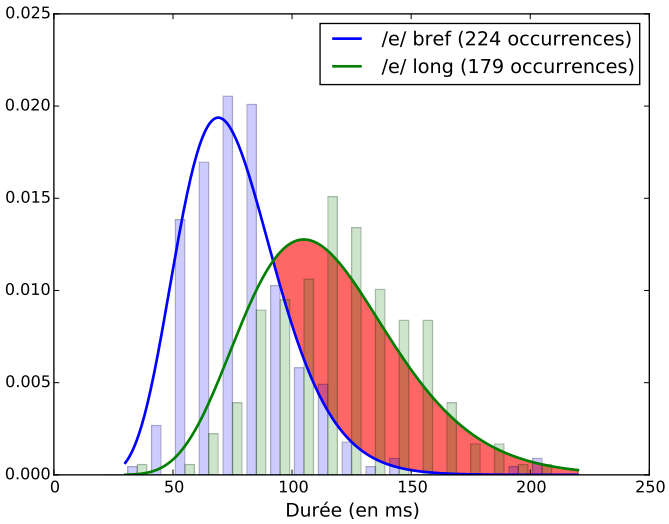

Phonème /u/
contexte fermé
contexte ouvert
Effectif
Durée (en ms)

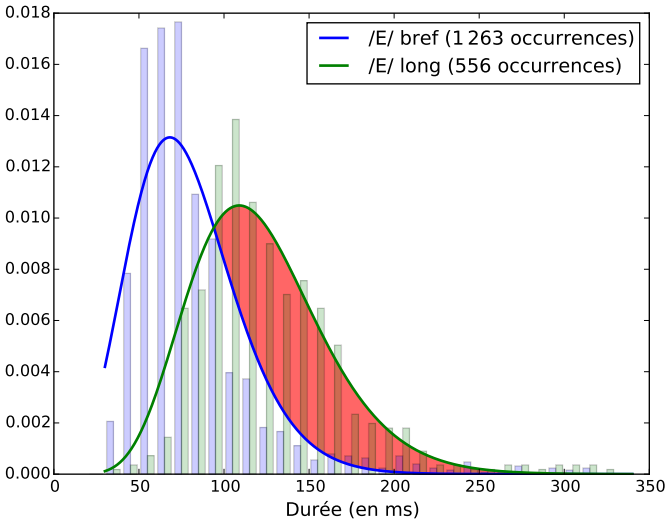

Durée (en ms)
contexte fermé
(1 741 occurrences)
contexte ouvert
(3 967 occurrences)
Phonème /u/

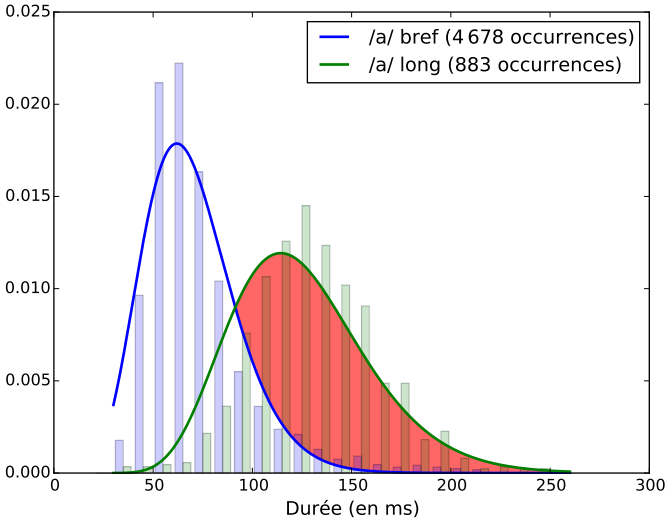
ANNEXE M

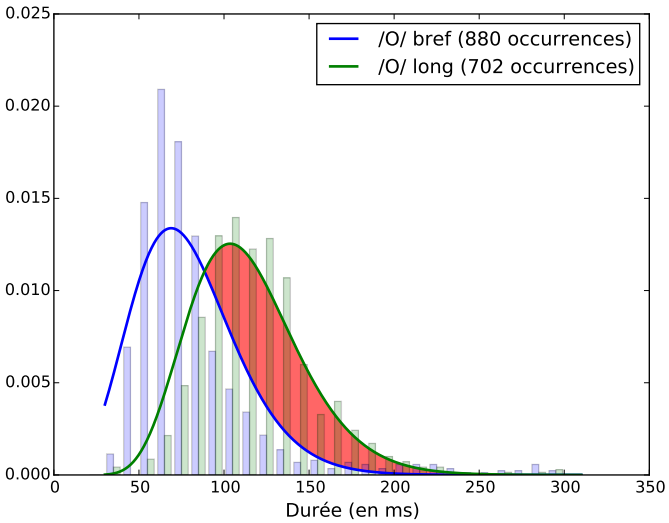
Histogrammes normalisés et distributions gamma des voyelles en wolof lu (corpus *dev*)

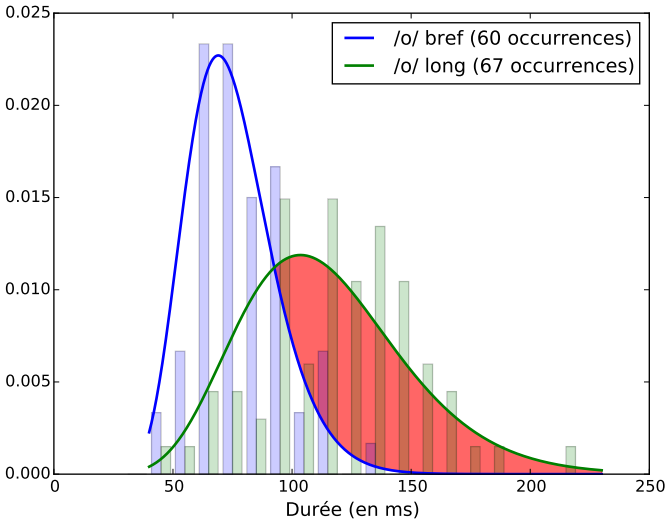


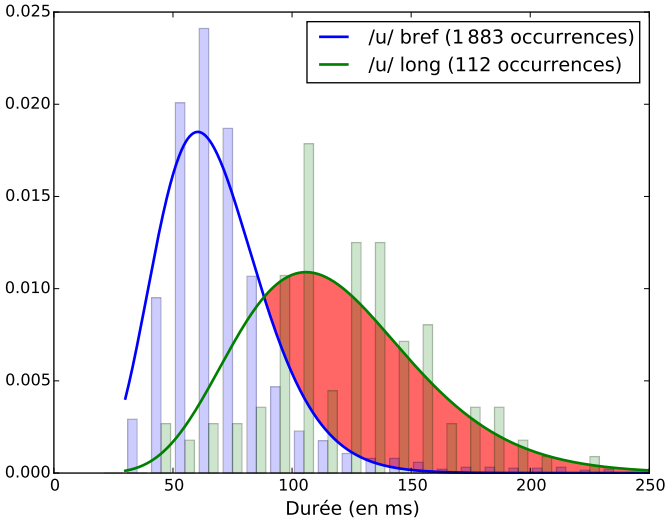



/E/ bref (1 263 occurrences)
/E/ long (556 occurrences)
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
300
350
Durée (en ms)


/a/ bref (4 678 occurrences)
/a/ long (883 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
Durée (en ms)


/O/ bref (880 occurrences)
/O/ long (702 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
350
Durée (en ms)

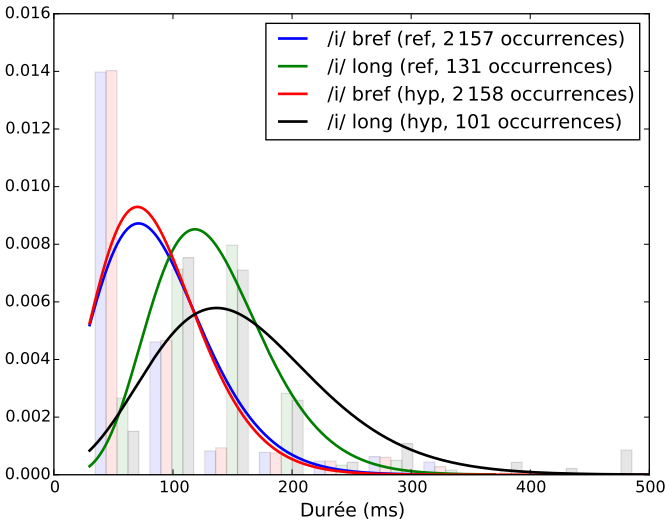

/o/ bref (60 occurrences)
/o/ long (67 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (en ms)

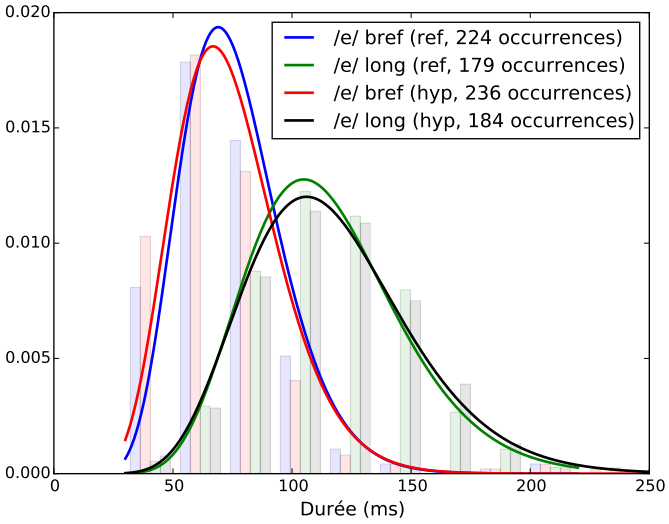

/u/ bref (1 883 occurrences)
/u/ long (112 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (en ms)

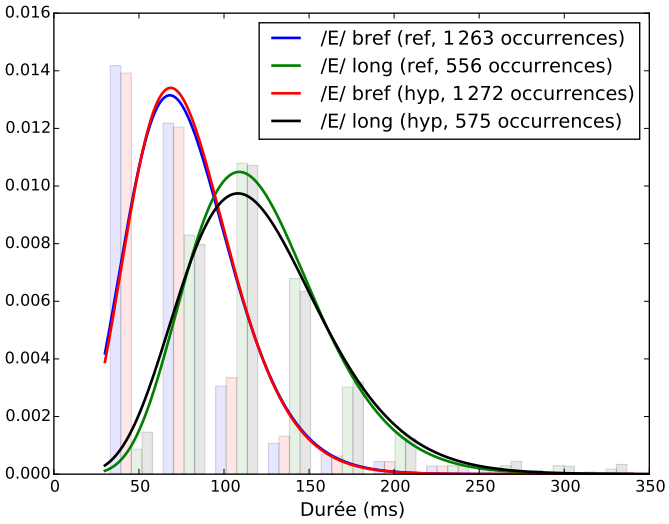
ANNEXE N

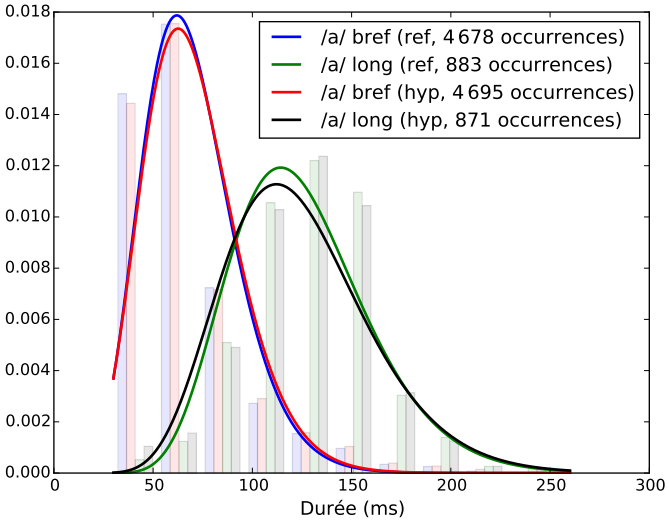
Comparaison des alignements forcés à partir de transcriptions manuelles ou automatiques du wolof

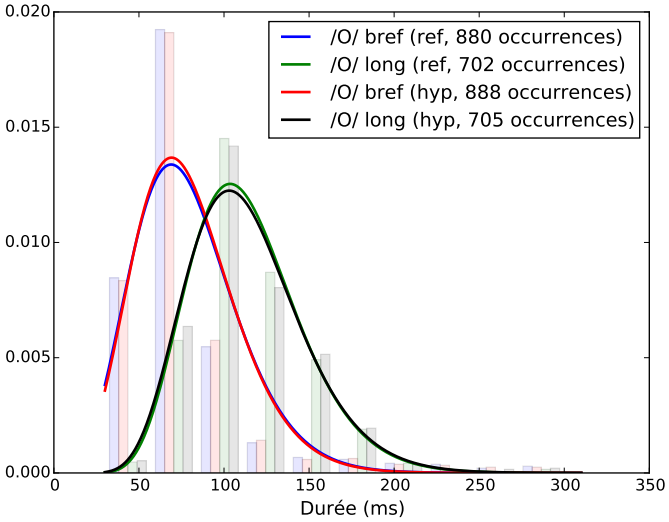
Histogrammes normalisés et distributions gamma des voyelles en wolof lu (corpus *dev*) - en utilisant la transcription humaine (ref) ou la transcription de la RAP (hyp)

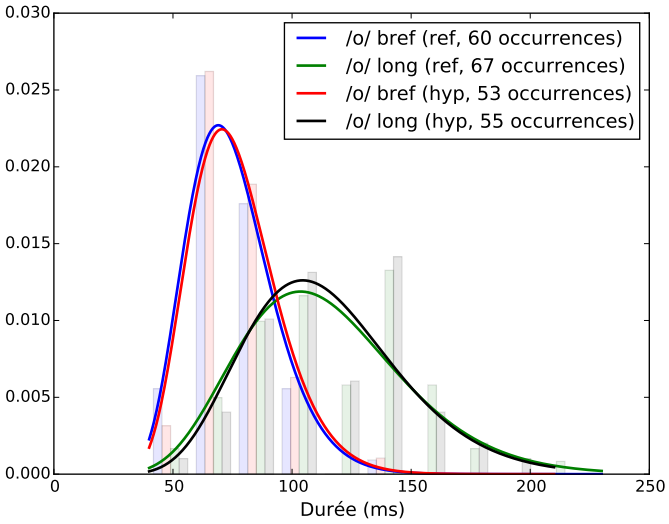


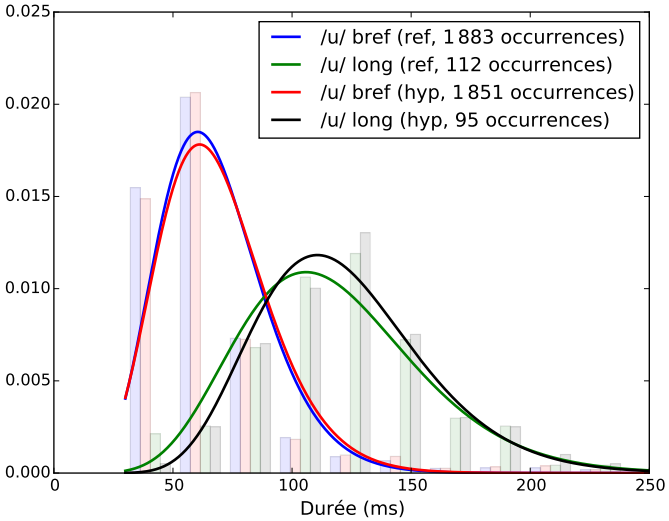



/E/ bref (ref, 1 263 occurrences)
/E/ long (ref, 556 occurrences)
/E/ bref (hyp, 1 272 occurrences)
/E/ long (hyp, 575 occurrences)
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
300
350
Durée (ms)


/a/ bref (ref, 4 678 occurrences)
/a/ long (ref, 883 occurrences)
/a/ bref (hyp, 4 695 occurrences)
/a/ long (hyp, 871 occurrences)
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
300
Durée (ms)

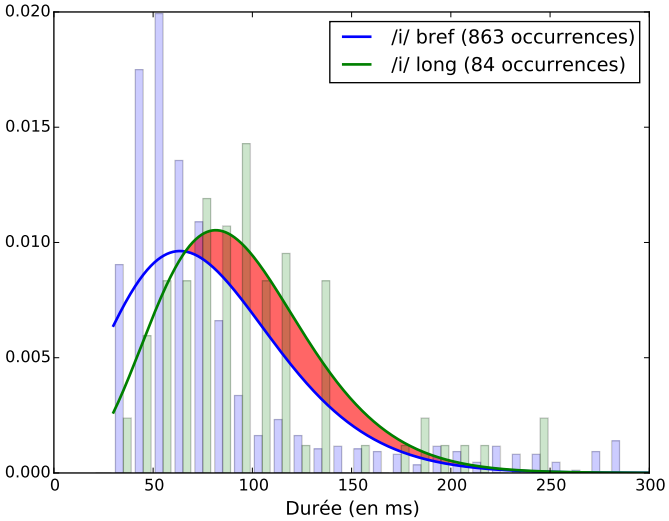

/O/ bref (ref, 880 occurrences)
/O/ long (ref, 702 occurrences)
/O/ bref (hyp, 888 occurrences)
/O/ long (hyp, 705 occurrences)
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
350
Durée (ms)

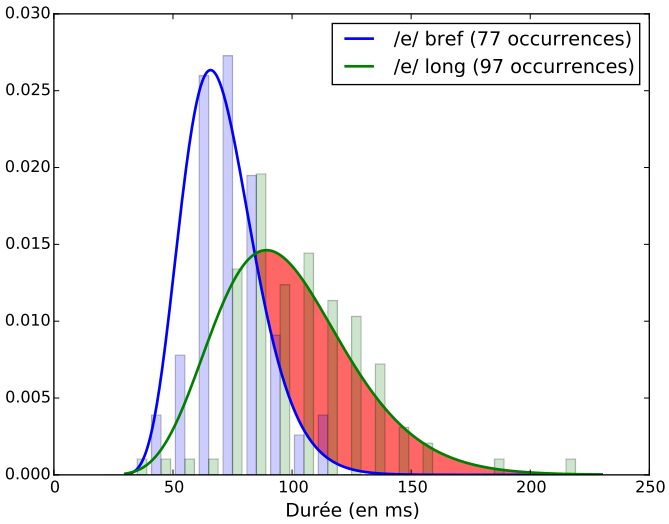

/o/ bref (ref, 60 occurrences)
/o/ long (ref, 67 occurrences)
/o/ bref (hyp, 53 occurrences)
/o/ long (hyp, 55 occurrences)
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (ms)

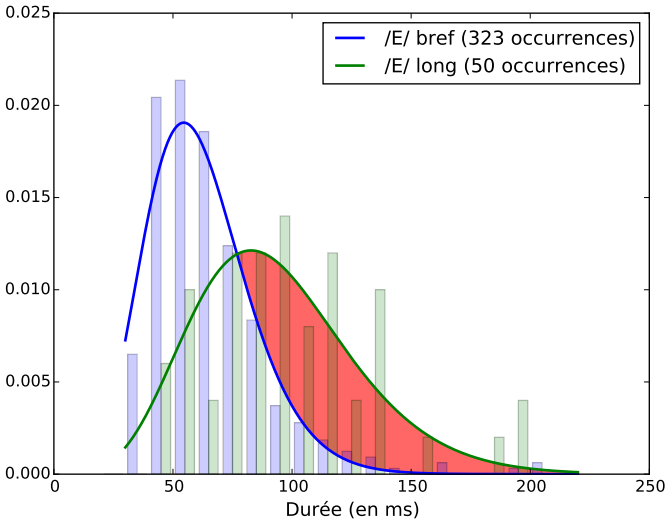

/u/ bref (ref, 1 883 occurrences)
/u/ long (ref, 112 occurrences)
/u/ bref (hyp, 1 851 occurrences)
/u/ long (hyp, 95 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (ms)

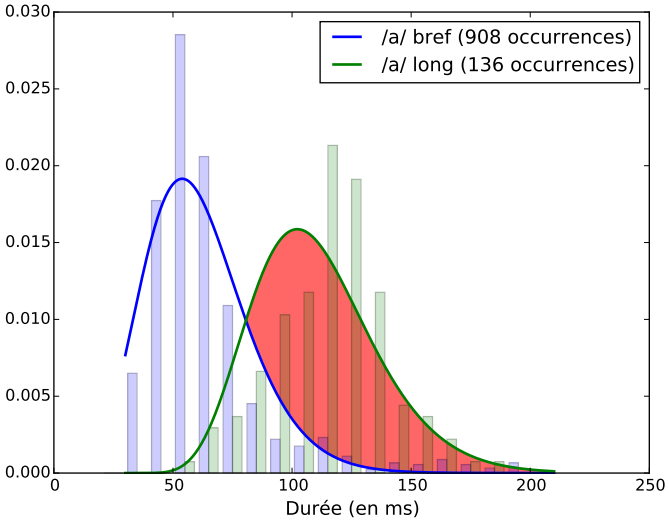
ANNEXE O

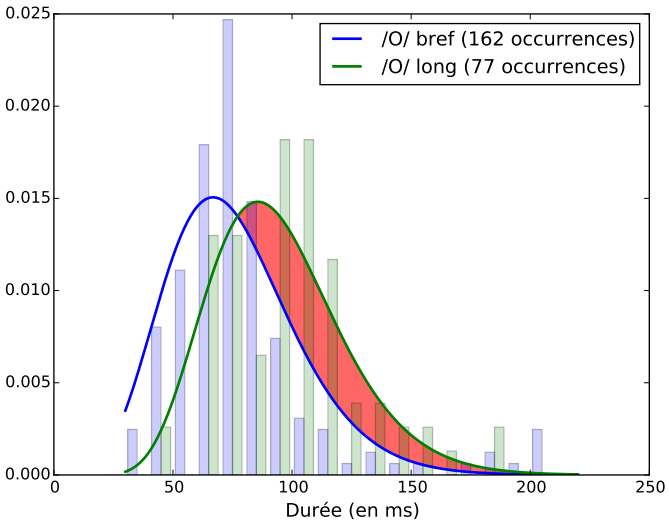
Histogrammes normalisés et distributions gamma des voyelles en wolof semi-spontané

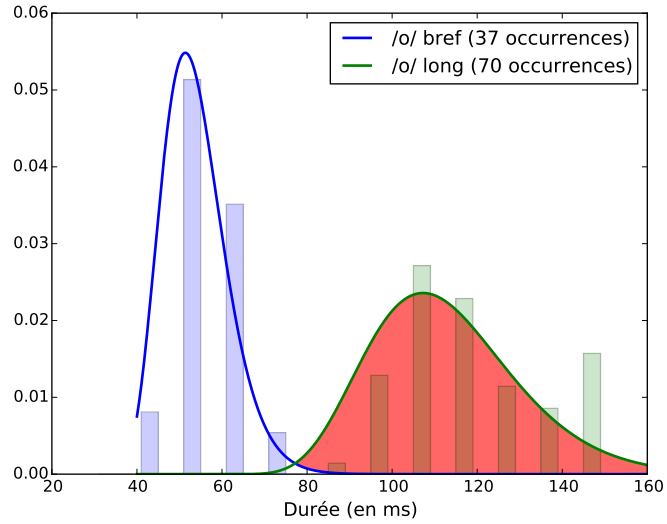


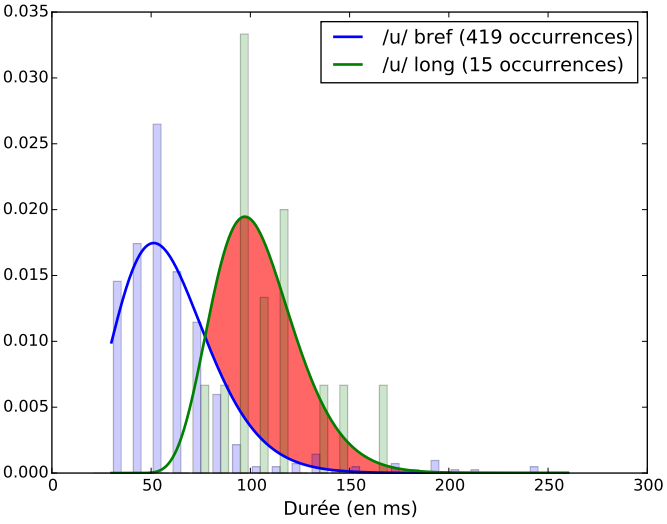



/E/ bref (323 occurrences)
/E/ long (50 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (en ms)


/a/ bref (908 occurrences)
/a/ long (136 occurrences)
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (en ms)

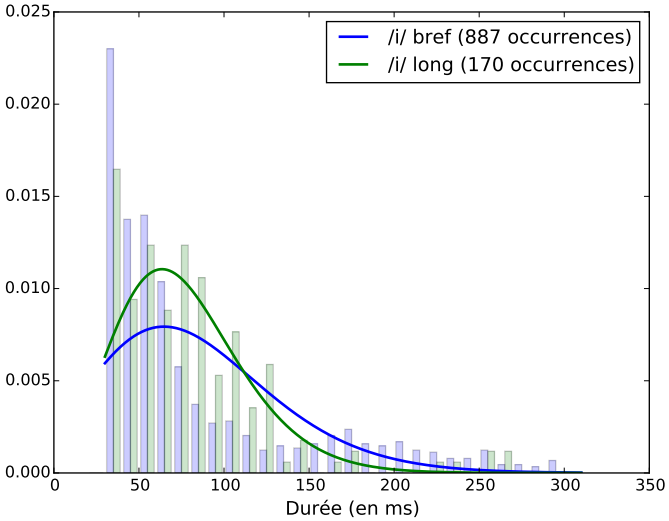

/O/ bref (162 occurrences)
/O/ long (77 occurrences)
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (en ms)

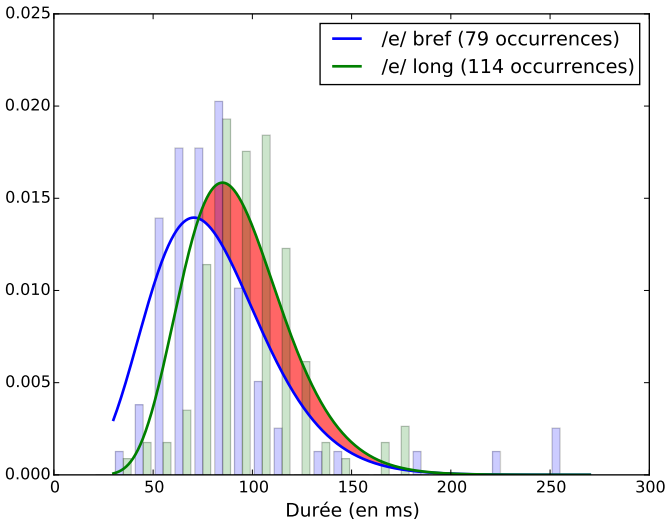

/o/ bref (37 occurrences)
/o/ long (70 occurrences)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
120
140
160
Durée (en ms)

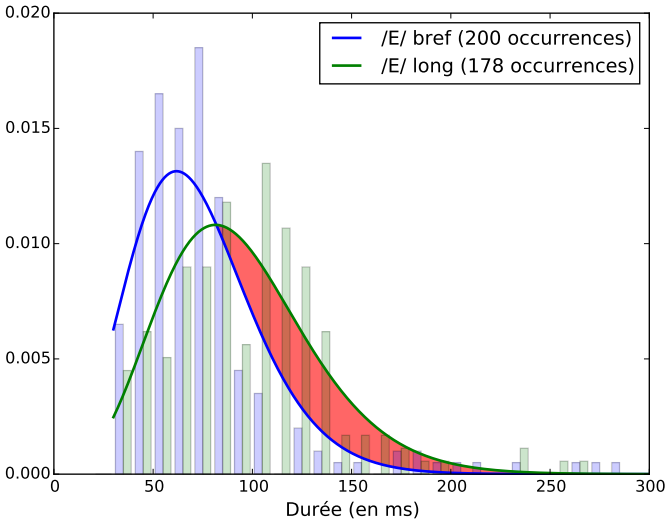

/u/ bref (419 occurrences)
/u/ long (15 occurrences)
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
Durée (en ms)

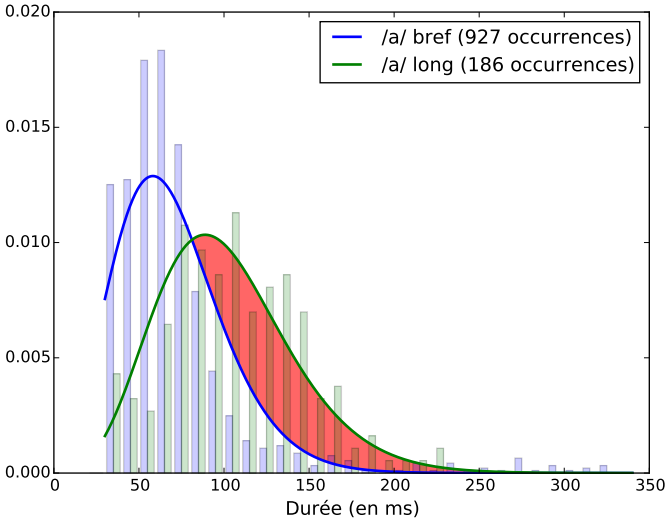
ANNEXE P

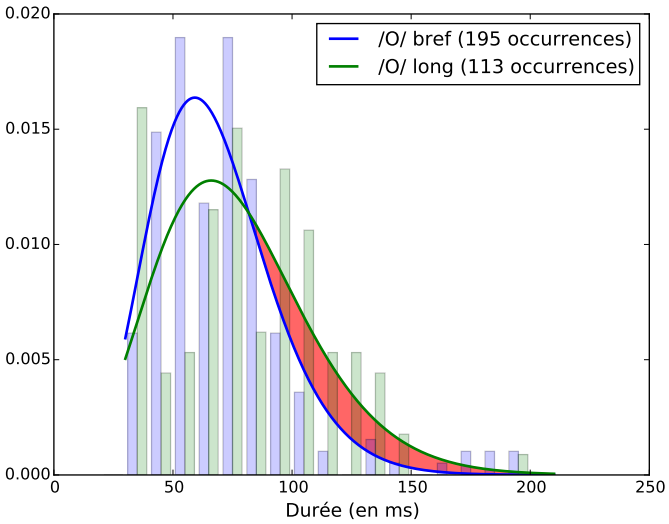
Histogrammes normalisés et distributions gamma des voyelles en faana-faana semi-spontané

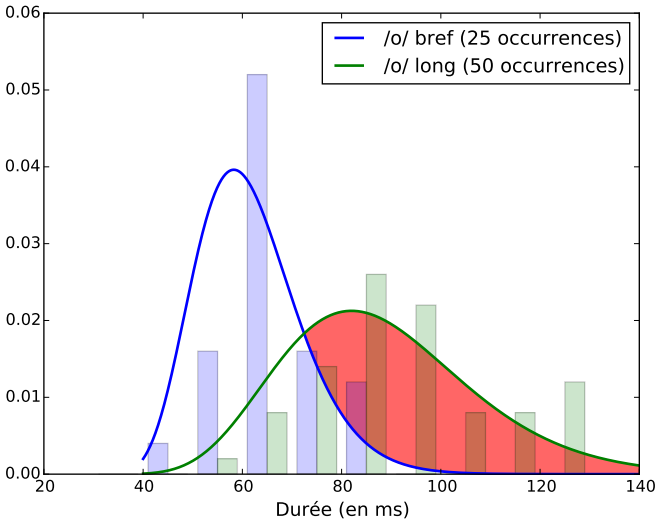


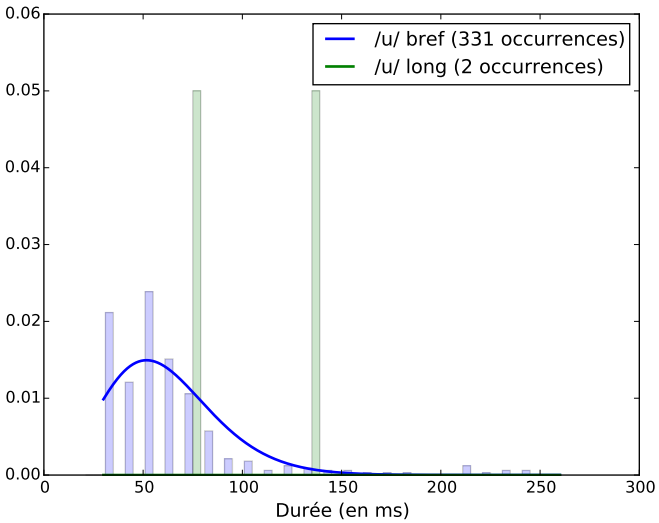



/E/ bref (200 occurrences)
/E/ long (178 occurrences)
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
Durée (en ms)


/a/ bref (927 occurrences)
/a/ long (186 occurrences)
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
350
Durée (en ms)


/O/ bref (195 occurrences)
/O/ long (113 occurrences)
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Durée (en ms)


/o/ bref (25 occurrences)
/o/ long (50 occurrences)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
120
140
Durée (en ms)


/u/ bref (331 occurrences)
/u/ long (2 occurrences)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
Durée (en ms)