Construction et évaluation pour la TA d'un corpus journalistique bilingue : application au français-somali

Houssein Ahmed Assowe

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

Houssein AHMED ASSOWE

Thèse dirigée par **Hervé BLANCHON**, UGA

préparée au sein du **Laboratoire d'Informatique de Grenoble** dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Construction et évaluation pour la TA d'un corpus journalistique bilingue : application au français-somali

Building and evaluating for MT a bilingual corpus : Application ton French-Somali

Thèse soutenue publiquement le **29 mai 2019**,
devant le jury composé de :

Monsieur HERVE BLANCHON
MAITRE DE CONFERENCES, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur MATHIEU LAFOURCADE
MAITRE DE CONFERENCES, UNIVERSITE DE MONTPELLIER, Rapporteur
Monsieur CHRISTOPHE ROCHE
PROFESSEUR, UNIVERSITE SAVOIE MONT BLANC, Président
Monsieur MAX SILBERZTEIN
PROFESSEUR, UNIVERSITE DE FRANCHE-COMTE, Rapporteur

UNIVERSITÉ GRENOBLE ALPES

N° attribué par la bibliothèque

/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

DOCTEUR ÈS SCIENCES

délivré par l'UNIVERSITÉ GRENOBLE ALPES

Spécialité : "INFORMATIQUE"

Thèse préparée au laboratoire LIG-GETALP (CNRS-INPG-UGA) dans le cadre de

l'École Doctorale "Mathématiques, Sciences et Technologies de l'Information, Informatique"

présentée et soutenue publiquement

par

Houssein AHMED ASSOWE

Le jj/mm/2019

CONSTRUCTION ET EVALUATION POUR LA TA D'UN CORPUS JOURNALISTIQUE BILINGUE : APPLICATION AU FRANÇAIS-SOMALI

JURY

M. Laurent Besacier	Examinateur
M. Mathieu Lafourcade	Rapporteur
M. Martin Benjamin	Rapporteur
M. Max Silberztein	Rapporteur
M. Christian Boitet	Examinateur
M. Hervé Blanchon	Directeur de thèse

Résumé en français

Dans le cadre des travaux en cours pour informatiser un grand nombre de langues « peu dotées », en particulier celles de l'espace francophone, nous avons créé plusieurs systèmes de traduction automatique français-somali dédiés à un sous-langage journalistique, permettant d'obtenir des traductions de qualité, à partir d'un corpus bilingue construit par post-édition des résultats de GOOGLE TRANSLATE (GT), à destination des populations somalophones et non francophones de la Corne de l'Afrique. Pour cela, nous avons constitué le tout premier corpus parallèle français-somali de qualité, comprenant à ce jour 98 912 mots (environ 400 pages standard) et 10 669 segments. C'est un corpus aligné, et de très bonne qualité. Nous l'avons construit en post-éditant les prétraductions de GT, qui combine pour cela son système de TA français-anglais et son système de TA anglais-somali. Ce corpus a fait l'objet d'une évaluation par 9 annotateurs bilingues qui ont donné un score de qualité à chaque segment du corpus, et corrigé éventuellement notre post-édition. À partir de ce corpus, en croissance, nous avons construit plusieurs versions successives d'un système de Traduction Automatique à base de fragments (PBMT), MosesLIG-fr-so, qui s'est révélé meilleur que GT sur ce couple de langues et ce sous-langage, en termes de mesure BLEU et du temps de post-édition. Nous avons fait également une première expérience de traduction automatique neuronale français-somali en utilisant OPENNMT, de façon à améliorer les résultats de la TA sans aboutir à des temps de calcul prohibitifs, tant durant l'entraînement que durant la traduction (le décodage).

D'autre part, nous avons mis en place une iMAG (passerelle interactive d'accès multilingue) qui permet à des internautes somaliens non francophones du continent d'accéder en somali à l'édition en ligne du journal « La Nation de Djibouti ». Les segments (phrases ou titres) prétraduits automatiquement par un système de TA fr-so en ligne disponible peuvent être post-édités et notés (sur une échelle de 1 à 20) par les lecteurs eux-mêmes, de façon à améliorer le système par apprentissage incrémental, de la même façon que ce qui a été fait pour le système français-chinois (PBMT) créé par [Wang, 2015].

Abstract

As part of on-going work to computerize a large number of "under-resourced" languages, especially those in the French-speaking world, we have created several French-Somali machine translation systems dedicated to a journalistic sub-language, allowing to obtain quality translations from a bilingual corpus built by post-editing GOOGLE TRANSLATE (GT) results, the final users being the Somali and non-French speaking populations of the Horn of Africa. For this, we have created the very first quality French-Somali parallel corpus, comprising to date 98,912 words (about 400 standard pages) and 10,669 segments. It is an aligned corpus of very good quality, built by post-editing pre-translations produced by GT, which uses a combination of its French-English and English-Somali MT language pairs. That corpus was evaluated by 9 bilingual annotators who assigned a quality score to each segment of the corpus and corrected our post-editing at some places. Using this growing corpus as training corpus, we have built several successive versions of a MosesLIG-fr-so statistical Phrase-Based Machine Translation System (PBMT), which has proven to be better than GoogleTranslate on this language pair and this sub-language, in terms of BLEU and post-editing time. We also used OpenNMT to build a first French-Somali neural MT system and experiment it.

On the other hand, we have set up an iMAG (interactive Multilingual Access Gateway) that allows non-French-speaking Somali surfers on the continent to access the online edition of the newspaper "La Nation de Djibouti" in Somali. The segments (sentences or titles), pre-automatically translated by any available fr-so MT system can be post-edited and rated (on a 1 to 20 scale) by the readers themselves, so as to improve the system by incremental learning, in the same way as has been done before for the French-Chinese PBMT system created by [Wang, 2015].

Résumé en somali

Iyadoo qayb ka ah shaqada socota ee lagu kombuyutargaraynayo luqado badan "oo aan aad uu kombuyutargaraynaysnayn », gaar ahaan kuwa dalalka afka Faransiiska ku hadla, waxaan sameynay dhowr nidaamyo oo turjumaadda mashiinka Faraansiis-Soomali kuwas oo lagu talagalay qoral saxaafadeedka, iyagoo suurtagalini karayo tarjumad taya fiican leh, oona laga abuuray koorbus laba afleh ah oo laga sameeyay tifaatirka iyo wanaajinta tarjumadda natiijoyinka aalada turjumadda ee GOOGLE TRANSLATE GT), arrintaas oo ujeedadiisu ay tahay in ay ka fa'ideystaan dadka af soomaliga ku hadla ee aan af faransiiska ku hadlin ee ku nool geeska afrika.

Sida darteed, waxaan abuurnay koorbuskii ugu horeyay ee Faransiis-Soomali oo tayo leh, kana kooban ilayo maanta 98 912 erey (Ku dhawaad 400 oo bog oo caadi ah) iyo 10 669 oo weedh. Waa koorbus labo afleh ah, oo tayo aad u ficaan leh. Waxaanu sameynay koorbuska innaka oo tifatirnay oo hagaajinay turjumaadyadii ugu horeyay ee GT, kaas oo si isku daar ah isticmalay nidaamkiisa turjumaada ee Soomali-Ingiriis iyo Ingiriis-Faransiis.

Koorbuskanii qayb ka midi ah waxa qiimeyn ku sameeyay sagal qiimeeyaal oo labo afleh ah iyagoo oo siiyay weedh kasta naatija tayadiisa la xidhidha, kadibna waay saxeen tifatirkeeni hore. Koorbuskanii si kordhaya waxaan ka sameynay dhowr nooc oo nidaamka turjumadda ee weedhaha ku salaysan, MosesLIG-fr-so, kaas oo noqday mid ka tayo ficaan GT marka la eego Faraansiis-Somali et iyo afhoosadka saxaafadeedka, marka la qiimeyo tayada BLEU iyo wakthiga la tifatirayo.

Waxaan kale oo sameynay tiijabinti ugu horeysay ee turjumadda mashiinka ee ku saleysan nerfiska ee labada af Faransiis-Soomali innago isticmalayna aalada OpenNMT, si aan uu hagaajino tayada turjumadda iyado oo naga qadan wakhti xisabin badan, marka tabobarka la siinayo iyo marka lagu turjumaayo.

Dhinaca kale, waxaan sameynay aalada iMAG, taas oo uu surtagalinaysa dadka shabakada isticmala ee soomalida aan ku hadlin af faraansiiska ee qaradda Afrika si ay af soomaliga ugu akhristaan qorallada boga internetka la soo geliyay ee wargeyska « La Nation ee Jibuuti ». Weedhaha (weedh ama cinwaan) ee marka u horeysa lagu turjumay nidamka mashiinka Faraansiis-Soomali, kaas oo la diyaariyay, ayaa akhristayaasha laftarkooda ay tifatiri karan, ayna qiimeyn karan, si ay uu hagaajiyan tayadiisa sida nidamka waxbarashada isku so noqnoqota, taas oo la mid ah sidii nidaamki turjumaadda mashiinka ee Faraansiis-Shiine (PBMT) ee uu abuuray [Wang, 2015].

À mes parents,
À Miski, ma femme
À Sirajudin, mon fils aîné

Remerciements

Je profite de cette occasion pour remercier et saluer très sincèrement les personnes qui ont croisé sur mon chemin durant ces cinq dernières années, et qui de près ou de loin ont contribué à la concrétisation de cette thèse.

Merci à l'administration de l'Université de Djibouti et aux responsables et membres de l'équipe GETALP du Laboratoire d'Informatique de Grenoble (LIG).

Un grand merci au Professeur **Laurent Besacier**, directeur de l'école doctorale **EDMSTII** et responsable de l'équipe GETALP du LIG pour les nombreuses dérogations et les soutiens indispensables qu'il m'accordés pour finir et soutenir cette thèse.

Je tiens à remercier également les membres du jury, en particulier les Professeurs **Martin Benjamin, Mathieu Lafourcade et Max Silberztein,** pour avoir accepté d'être rapporteurs.

Je remercie également le Professeur Emérite **Christian Boitet** pour avoir accepté d'être examinateur. Merci pour les relectures et les corrections successives de ma thèse, et pour toutes les améliorations et soutiens que vous m'avez apportés durant la fin de la rédaction de mon mémoire de thèse.

Je tiens à remercier très chaleureusement mon directeur de thèse, **Hervé Blanchon**, pour m'avoir fait confiance, puis m'avoir guidé, encouragé et conseillé tout en me laissant une grande liberté. Merci pour son soutien moral et sa compréhension sans faille, dont j'avais besoin durant toute cette thèse, et surtout durant les périodes difficiles.

Un grand merci au président de l'Université de Djibouti et à mon doyen pour leur soutien indéfectible aux doctorants, et les nombreuses facilités et congés qu'ils m'ont accordés tout au long de ma thèse.

Merci à mes amis grenoblois (**Osman**, **Mouhyadin**, **Andon**, **Ritesh**) et mes amis djiboutiens (**Said**, **Souleiman**, **Ilyas**, **Abdourahman**, **Dahir**) pour leur soutien moral, leurs conseils et leurs encouragements.

Mes dernières pensées vont à ma famille et à mes parents, qui m'ont toujours encouragé et soutenu moralement pour aller le plus loin possible dans mes études.

Je pense surtout à ma femme **Miski Souleiman**, qui m'a tant supporté et accepté mes voyages et absences répétées malgré son besoin que je sois présent auprès d'elle, surtout après l'arrivée de notre bébé.

Merci aussi à mes frères et sœurs (**Fozi**, **Kadra**, **Abdoukarim**, **Safia**, **Saredo**) qui m'ont apporté leur soutien et l'appui nécessaire durant cette longue période.

Enfin, merci à ceux et celles dont j'aurais omis de citer les noms ici, ils se reconnaîtront dans ces quelques lignes.

Table des matières

Table des illustrations (tableaux et figures)

Glossaire

Données parallèles. Les données alignées sont les éléments d'un corpus parallèle composé de deux langues ou plus. Chaque élément dans une langue correspond à l'élément correspondant dans l'autre langue. Les éléments, parfois appelés segments, peuvent être alignés par blocs, alignés sur les paragraphes, alignés sur les phrases ou alignés sur les items lexicaux.

Processus d'alignement. Il y a deux processus d'alignement. Dans la préparation du corpus, le processus d'alignement crée des données alignées. Pendant l'apprentissage, le processus d'alignement utilise un programme tel que MGIZA++ pour créer des fichiers d'alignement de mots.

Score BLEU. BLEU est l'abréviation de « BiLingual Evaluation Understudy ». Un score BLEU indique quelle est la similarité entre les séquences de mots et d'items lexicaux dans un ensemble de données, telles que la sortie de traduction automatique et celles d'un autre ensemble de données, comme une traduction humaine de référence.

Voir : processus d'évaluation.

Préparation du corpus. La préparation de corpus est le processus général d'extraction, de transformation, de catégorisation de divers documents en fonction de l'objectif initial et d'alignement des données résultantes dans un corpus parallèle pour l'apprentissage d'un modèle de traduction.

Processus d'évaluation. Le processus d'évaluation utilise un modèle de traduction de composants créés dans le processus d'apprentissage et configuré avec le processus d'optimisation pour traduire plusieurs milliers de phrases de langue source dans l'ensemble d'évaluation. Ce processus compare ensuite les traductions automatiques résultantes aux traductions de référence, également dans l'ensemble d'évaluation. Le dernier rapport d'évaluation de score BLEU montre à quel point les traductions automatiques correspondent aux traductions de référence.

Modèle hiérarchique. Modèle de traduction automatique statistique qui utilise un corpus d'apprentissage pour créer des alignements arborescents.

Données d'entraînement hiérarchiques. Un corpus d'apprentissage dans lequel chaque phrase est annotée avec une structure hiérarchique du langage, comme un arbre de constituants ou un arbre de dépendances fonctionnelles.

Modèle de langage. Un « modèle de langage » ou « ml » est une description statistique d'une langue qui donne les fréquences d'occurrences de N-grammes de mots dans un corpus. Le "ml" est formé à partir d'un grand corpus monolingue et enregistré sous forme de fichier. Le fichier de modèle de langage est un composant obligatoire de tout modèle de traduction. Le décodeur MOSES utilise un modèle de langage pour sélectionner la phrase de la langue cible la plus « probable » à partir d'un grand nombre de traductions « possibles » qu'il a généré en utilisant la table de traduction et la table de ré-ordonnancement.

Production des modèles de langage. Les fichiers d'un modèle de langage contiennent des données statistiques générées par des outils disponibles. Le décodeur MOSES peut utiliser plusieurs, notamment : KENLM, RANDLM et IRSTLM.

Fichier de configuration MOSES**.** Le fichier de configuration de MOSES est un fichier texte créé pendant le processus d'optimisation. Le fichier contient les chemins d'accès aux tables de

traductions, de ré-ordonnancement et du modèle de langage ainsi que d'autres codes et valeurs numériques qui contrôlent le fonctionnement d'un système MOSES.

N-grammes. Un n-gramme est une séquence d'éléments (1, 2, 3, etc.) figurant dans une séquence naturelle (phrase, titre) plus grande. Dans un ML, les n-grammes sont des séquences d'items lexicaux. Dans les tables de traductions et de ré-ordonnancement, les n-grammes sont des séquences de paires de mots appartenant aux langues source et cible.

Table de traductions. Une "table de fragments de traduction" est une description statistique d'un corpus parallèle de paires de phrases source-cible. Les fréquences que les n-grammes dans un texte en langue source coproduisent avec des n-grammes dans un texte en langue cible parallèle sont supposées correspondre à la probabilité que ces n-grammes appariés source-cible se reproduisent dans d'autres textes similaires au corpus parallèle. En termes pratiques, la table de fragments est un fichier créé pendant le processus d'entraînement et enregistré dans le dossier du modèle de traduction. Il fonctionne comme un dictionnaire sophistiqué entre les langues source et cible.

Les tables de traductions et de ré-ordonnancement sont des composants des modèles de traduction.

Pipeline. Un "pipeline" est une chaîne d'outils de processus connectés par des flux standard, de sorte que la sortie de chaque processus (*stdout*) nourrit directement l'entrée (*stdin*) du suivant.

Modèle de recassage. Un modèle de recassage est un modèle de traduction spécial qui change la casse des mots d'un texte pour obtenir la même casse que dans l'original (par exemple, première lettre en majuscule). Pour cela, on utilise l'alignement *a posteriori* entre chaque segment source et le segment cible produit.

Table de réordonnancement. Une « table de réordonnancement » contient les fréquences statistiques qui décrivent les changements dans l'ordre des mots entre les langues source et cible, tels que « green table » et « table verte ». En termes pratiques, une "table de réorganisation" est un fichier créé pendant le processus d'entraînement et enregistré en tant que fichier dans le dossier modèle. La table de réorganisation est un composant du modèle de traduction.

Langue source. La langue source est la langue du texte à traduire. Généralement, il s'agit de la langue d'origine du texte. La langue source est la même que la valeur de l'attribut "*srclang*" de la spécification TMX de la balise <*tu*>.

Langue cible. La langue cible est la langue dans laquelle le texte de la langue source doit être traduit.

Ensemble d'évaluation. Une paire de données de langue source et cible, contenant typiquement plusieurs milliers de paires utilisées dans le processus d'évaluation.

Tokenisation. La *Tokenisation* est le processus de séparation des items lexicaux.

Items lexicaux. Ce sont les mots-formes, les ponctuations non internes aux mots, les balises (XML en particulier), et les symboles spéciaux « hors texte », comme des marques de fabrique.

Chaîne d'outils. Une « chaîne d'outils » est une série d'outils de programmation liés ou « chaînés » utilisés dans une série où la sortie d'un outil en amont devient l'entrée d'un outil « en aval ».

Voir : Pipeline

Corpus d'entraînement. Un corpus linguistique avec des données parallèles, préparé pour construire la table de traduction et la table de réordonnancement des composants d'un modèle de traduction.

Processus d'entraînement. L'entraînement est un processus dans la branche d'apprentissage automatique du domaine de l'intelligence artificielle. Dans le processus d'apprentissage, un système "apprend" les relations entre les données parallèles. Dans la traduction automatique probabiliste, les textes en langue source sont considérés comme des stimuli qui génèrent le texte de la langue cible en réponse. Concrètement, l'apprentissage commence sur des bi-segments et crée la table de traductions et la table de réordonnancement, qui sont les composants d'un modèle de traduction.

Mémoire de traductions. Une mémoire de traductions (MT) est une donnée parallèle qui a été collectée dans le but d'aider à produire de futures traductions.

Modèle de traduction. Un « modèle de traduction » consiste en une ou plusieurs tables de traductions, zéro ou plusieurs tables de réordonnancement, un ou plusieurs modèles de langage et un fichier de configuration de MOSES, tous créés durant le processus d'apprentissage et d'optimisation.

Processus d'optimisation. Le réglage est un processus qui optimise les paramètres du fichier de configuration pour un modèle de traduction lorsqu'il est utilisé dans un but spécifique. Le processus d'optimisation traduit des milliers de phrases de langue source qui se trouvent dans les données d'optimisation avec un modèle de traduction, compare la sortie du modèle à un ensemble de traductions humaines de référence et ajuste les paramètres dans le but d'améliorer la qualité de la traduction. Ce processus se poursuit au cours de nombreuses itérations. À chaque itération, le processus de réglage répète les étapes jusqu'à ce qu'il atteigne un niveau optimal.

Ensemble d'optimisation. Une paire de données de langue source et cible, contenant généralement plusieurs milliers de paires utilisées dans le processus d'optimisation.

Aligneur de mots. Un aligneur de mots est un programme qui est chargé de créer des fichiers d'alignement de mots pendant le processus d'alignement des mots. Moses prend actuellement en charge les aligneurs de mots suivants : GIZA ++, MGIZA ++,BERKELEYALIGNER, *etc.*

Alignement de mots. Le processus d'alignement de mots utilise un aligneur de mots pour créer un fichier d'alignement de mots pendant le processus d'apprentissage.

Mots. Dans une langue naturelle, un mot est la plus petite unité de sens autonome. En traduction automatique, un mot est un item lexical créé dans le processus de création d'items qui n'est ni une ponctuation ni un symbole.

Introduction

L'objectif de cette thèse est la construction, la mise en place et l'évaluation d'outils et de ressources linguistiques pour la traduction automatique et plus généralement l'informatisation du somali. Cette langue est très faiblement dotée, et ne dispose actuellement d'aucune ressource linguistique suffisante pour réaliser des outils de traitement automatique du langage naturel (TALN) « empiriques » à l'état de l'art comme la traduction automatique statistique.

Notre travail se situe dans le cadre de grands travaux et mouvements internationaux qui souhaitent que chaque peuple ou communauté linguistique puisse disposer de tous les outils et ressources nécessaires pour utiliser les technologies de l'information et de la communication (TIC) dans leur langue maternelle.

Outre la réduction de la fracture numérique au niveau mondial, l'informatisation du plus grand nombre possible de langues permettra à chacun de profiter de tous les avantages et facilités qu'offrent les TIC dans la vie quotidienne de tous. Par exemple, dans certains pays d'Afrique de l'Est tels que le Kenya ou le Rwanda, l'utilisation des applications mobiles ou la localisation des logiciels informatique permettra l'accès aux services bancaires aux communautés rurales désenclavées et ayant de faibles revenus.

Pour une langue peu dotée comme le somali, la difficulté réside tant dans la maîtrise des techniques et des méthodes en vigueur dans le processus d'informatisation des langues que dans la constitution de ressources nécessaires pour construire un premier système de TA.

En effet, en dépit d'une présence abondante du somali sur la Toile, il n'existe pas de ressources linguistiques qui peuvent servir à mettre en place des systèmes à l'état de l'art de traduction automatique pour la paire de langues français-somali.

Le principal domaine de recherche de notre travail de thèse est la traduction automatique d'une langue peu dotée dans un sous-langage assez restreint pour obtenir des bons résultats, assez productif, et potentiellement intéressant pour des lecteurs ne maîtrisant pas la langue source.

A l'ère de la mondialisation et du web 2.0, la traduction automatique constitue un excellent moyen pour acquérir de nouvelles connaissances et assimiler des informations diffusées sur d'autres langues qu'il s'agisse de données multimédia ou écrites.

Grâce aux grandes capacités de stockage et de traitement de données des ordinateurs de nos jours et leur transfert sur les réseaux internet, on retrouve çà et là de plus en plus de données langagières, monolingues ou multilingues qui peuvent servir à amorcer des travaux de recherche sur la traduction automatique des langues africaines.

Cependant, les travaux menés jusqu'à ce jour sur des corpus de textes d'une grande quantité sur les langues bien informatisées ont mis en évidence que la traduction automatique ne produisait pas toujours des résultats de très bonne qualité. Ainsi une édition *a posteriori* des hypothèses de traduction peut améliorer la qualité des textes traduits automatiquement.

En outre, du fait de la diversité des domaines de données issues du web sur lesquelles les systèmes de traduction à base de segments ont été réalisés, la spécialisation du système sur un sous-langage couvrant largement le vocabulaire du domaine des textes à traduire et l'utilisation des techniques d'adaptation aux domaines de spécialité ont montré ces dernières années leur efficacité et ont grandement amélioré la qualité d'un système de traduction spécialisé par rapport à un système généraliste.

Pour une bonne compréhension de la problématique de notre travail de thèse, nous avons étudié dans un premier temps les différentes approches au problème difficile de la préparation et du déploiement des différents modules nécessaires pour informatiser et créer des ressources linguistiques pour les langues et couples de langues peu dotés, en général.

Nous avons ensuite approfondi cet état de l'art en étudiant la situation du somali dans le domaine du TALN : nous avons ainsi recensé toutes les ressources et outils disponibles pour cette langue, qu'il s'agisse de corpus monolingues ou bilingues, bruts ou annotés, de dictionnaires d'usage, de lexiques spécialisés, de grammaires, et d'outils informatisés automatiques ou semi-automatiques (analyseurs, correcteurs, étiqueteurs, parseurs, traducteurs). Parallèlement à nos travaux de collecte et de recensement des données langagières du somali, nous avons construit un ensemble d'outils de base pour le traitement automatique du somali. Ce sont un segmenteur de mots *(tokenizer)* et de phrases, un étiqueteur morphosyntaxique probabiliste, et un lemmatiseur.

Suite à cette introduction, le contenu de ce manuscrit est organisé comme suit.

Le chapitre 1 présente le contexte de la recherche et les problèmes abordés au cours de cette thèse, situe notre travail dans le cadre de l'informatisation des langues peu dotées de l'espace francophone africain, détaille le cas du somali, et définit les objectifs pratiques possibles pour cette thèse.

Le chapitre 2 propose un état de l'art assez complet et détaillé sur les ressources linguistiques statiques et dynamiques du somali et l'état de son informatisation. Il présente également une classification des outils et ressources et applications ou services disponibles à ce jour pour le somali, et précise, dans le cas du somali, les besoins à couvrir et les problèmes à résoudre.

Le chapitre 3 présente notre méthodologie et nos contributions. Nous décrivons notre stratégie de construction des ressources et corpus, ainsi que la méthode que nous avons utilisée pour construire plusieurs systèmes de TA spécialisés au sous-langage journalistique du couple de langue français-somali.

Le chapitre 4 présente notre contribution en matière de construction d'un premier corpus parallèle français-somali de haute qualité, LDJ-fr-so-A, par TA suivie de post-édition, dans le sous-langage des nouvelles journalistiques. Ce sous-langage est assez restreint pour obtenir de bons résultats, assez productif, et potentiellement intéressant pour des lecteurs ne maîtrisant pas la langue source.

Le chapitre 5 présente les 4 systèmes de TA construits, soit sur le corpus LDJ-fr-so-A seul, soit sur le corpus LDJ-fr-so-ABC, constitué du précédent, augmenté de bisegments extraits de deux corpus français-somali de moins bonne qualité et « hors domaine » (OPUS-fr-so-B et OPUS-fr-so-C).

Nous comparons les résultats de GT et ceux de nos 4 systèmes de TA sur ces deux corpus, évalués selon 2 mesures objectives, démontrant encore une fois que des systèmes spécialisés à des sous-langages assez restreints ont une qualité d'usage nettement meilleure que celle de systèmes généralistes, développés à partir de données de beaucoup plus grande taille.

Nous avons aussi effectué une évaluation subjective de la qualité purement linguistique de notre corpus avec des juges indépendants. Elle prouve que ce corpus est de très haute qualité, et que l'approche « TA+PE » est efficace. Il semble qu'une telle évaluation n'ait pas encore été faite sur des corpus bilingues concernant des couples de langues peu dotés.

Nous pouvons aussi conclure que, pour construire un système de TA de qualité d'usage supérieure à celle de GT, la taille de notre corpus LDJ-fr-so-A (environ 400 pages standard ou 100 000 mots) est suffisante, même s'il est clair que, pour ce type de sous-langage, il faudrait sans doute arriver à une taille 4 ou 5 fois supérieure pour que des lecteurs somalophones monolingues puissent lire *La Nation de Djibouti* via une TA fidèle et surtout fiable.

Le travail présenté dans ce mémoire débouche donc sur une conclusion optimiste : oui, il est possible de construire des systèmes de TA vers le somali, et plus précisément pour des sous-langages journalistiques, répondant à des besoins attestés et de qualité suffisante pour qu'ensuite les lecteurs eux-mêmes puissent corriger en ligne et contribuer (selon les modalités imaginées par Google) à l'amélioration du système.

On peut enfin tirer de ce travail quelques conclusions générales.

En ce qui concerne la qualité, on constate que les systèmes de TA (statistiques aussi bien que neuronaux) construits à partir de notre corpus « augmenté » sont légèrement meilleurs que ceux obtenus à partir de notre corpus « restreint », pourtant de meilleure qualité, mais 2 fois plus petit (environ 100 K mots contre 200 K mots). Notre approche est que nous obtiendrons une qualité bien supérieure quand nous disposerons d'un corpus spécialisé de 200 K mots environ (800 pages standard), ce qui devrait être possible si notre système est mis en ligne et si ses résultats sont améliorés (par PE) en continu par les lecteurs somalophones de *La Nation de Djibouti*.

Les outils et méthodes utilisés semblent pouvoir s'appliquer à un grand nombre de langues africaines « peu dotées ». Pour tempérer cet optimisme, il faut cependant noter qu'il faut un assez gros travail pour mettre en place ce genre d'opération, non seulement pour collecter et compléter (passage à l'échelle) les ressources langagières nécessaires, mais aussi pour susciter, fédérer et organiser les efforts d'une communauté virtuelle se constituant autour de chaque projet de ce type.

Chapitre I Contexte de la recherche et problèmes abordés

Introduction du chapitre I

Depuis l'antiquité, l'homme a toujours utilisé le langage naturel comme mode de communication par excellence, que ce soit à travers la parole, l'écriture ou les signes.

En raison de la diversité linguistique et de l'impossibilité de comprendre ou parler toute les langues humaines, les êtres humains ont toujours eu besoin de faire appel à des traducteurs ou interprètes pour communiquer entre eux.

Dans le passé, le travail de traduction et d'interprétation des langues étrangères était le fait de quelques lettrés ou spécialistes des langues étrangères, et cet effort avait un coût exorbitant. Depuis l'essor de l'informatique et des nouvelles technologies de l'information et de la communication (NTIC), on a vu l'apparition du concept de traduction automatique. Ce dernier constitue l'un des domaines les plus difficiles du traitement automatique des langues naturelles ; c'est aussi celui qui a fait l'objet de plus d'investissements humains et financiers, et mobilisé des scientifiques depuis le début des années 50.

Pour trouver des solutions au problème de la traduction, beaucoup d'efforts et d'investissement ont été consentis par la communauté des chercheurs dans ce domaine.

La traduction automatique constitue à n'en point douter l'une des tâches les plus importantes et les plus difficiles pour amorcer l'informatisation des langues naturelles et surtout celles du continent africain. Ces dernières, qui représentent plus du tiers des langues parlées dans le monde, souffrent d'un important handicap technologique et scientifique pour entamer leur informatisation.

Notre recherche se situe dans le cadre et dans le contexte de l'accès à l'information utile pour tous dans leurs langues maternelles, et en particulier aux nouvelles journalistiques, au bénéficie des populations de l'espace francophone africain.

Il s'agit dans ce chapitre de présenter le contexte et les problèmes abordés au cours de cette thèse et de mettre en évidence le lien entre la traduction automatique post-éditée et l'informatisation des langues africaines de la francophonie.

Après avoir présenté la problématique de l'informatisation d'une langue peu dotée de l'espace francophone africain, nous mettons l'accent sur le cas précis du somali, en faisant une description de sa typologie linguistique, de sa situation sociolinguistique actuelle et de sa géolinguistique, tout en la comparant avec les autres langues africaines francophones quant à leurs niveaux d'informatisation.

Nous présentons enfin les objectifs pratiques et théoriques possibles pour une première thèse en informatique sur l'informatisation du somali.

I.1 Problématique de l'informatisation d'une langue peu dotée dans l'espace francophone africain

I.1.1 Méthodes et outils pour informatiser une langue peu dotée

D'après la base de données « Ethnologue[1] », il y a 7 010 langues vivantes parlées dans plus de 200 pays au monde [SIL 2005]. Environ un tiers de ces langues, soit 2 100, sont parlées en Afrique et très peu (moins de 10) disposent des ressources linguistiques nécessaires pour mettre en place des technologies issues du traitement automatique du langage naturel.

Selon [Berment, 2004], en dépit de la richesse et de la diversité linguistique réelle et attestée en Afrique, les langues africaines, et en particulier celles de la francophonie, sont quasi-majoritairement des langues se situant dans la catégorie des langues peu dotées. Les causes de la faible informatisation des langues africaines sont très nombreuses et sont résumées ci-dessous.

- En Afrique, la culture et les savoirs ancestraux se transmettaient par voie orale. Jusqu'à tout récemment, il n'existait que peu d'œuvres littéraires, historiques ou folkloriques écrites dans les langues africaines, à l'exception de quelques langues qui ont adopté des alphabets étrangers, comme c'était le cas pour le haoussa du Niger et du Nigeria, du somali ou du sorabe[2] (variante du malagasy en alphabet arabe) qui utilisaient un système d'écriture fondé sur l'alphabet arabe.
- La colonisation européenne a imposé les langues indo-européennes dans presque tous les pays africains à l'exception de l'Éthiopie. Durant cette période, ce sont l'anglais, le français, l'espagnol et le portugais qui ont été les langues de communication et d'éducation en Afrique. Cela a abouti à ce qu'aujourd'hui, du point de vue linguistique, les Africains soient répartis entre Africains francophones, anglophones, lusophones et hispanophones. Cette situation persiste encore de nos jours, après plus de 5 décennies d'indépendance des pays africains. Les principales raisons sont que les élites politiques qui ont pris le pouvoir en Afrique après les indépendances ont préféré garder le système linguistique et éducatif hérité du colonialisme et ont orienté leur politique vers ces langues, malgré le fait que plus de 95% de leurs habitants aient comme langues maternelles des langues africaines.
- Certains pays africains, comme l'Ouganda, le Kenya, le Cameroun et l'Éthiopie, ont initié des programmes et des politiques ambitieuses en faveur du développement, de l'harmonisation et de la numérisation des langues africaines. Toutefois, par manque de moyens financiers, leur volonté politique n'a pas atteint leur objectif ni même abouti à des développements tangibles.
- La pauvreté et la faiblesse économique des pays africains, ainsi que le faible pouvoir d'achat moyen des Africains eux-mêmes, ont fait qu'il n'existe pas d'investissement important dans l'industrie des langues en Afrique. Par conséquent, aucun marché émergent n'est possible aujourd'hui pour les produits, les applications et les services utilisant la technologie langagière en Afrique.

[1] www.ethnologue.com

[2] Ne pas confondre avec la langue slave appelée aussi sorabe (https://fr.wikipedia.org/wiki/Sorabe)

- Dans la plupart des pays africains, il n'existe pas de réelle politique volontariste en faveur de l'informatisation des langues nationales, et le grand ou parfois très grand nombre de langues parlées dans un même pays[3] est considéré comme un obstacle important.

Au-delà de ces problèmes d'ordre politique et structurel, les langues africaines de la francophonie souffrent plus particulièrement d'un manque de compétences techniques et scientifiques, ce qui est un gros obstacle à leur informatisation.

I.1.1.1 Quelques définitions

Selon la définition communément admise et issue du Grand Robert de la Langue Française[4], « l'informatisation d'une langue est l'introduction dans une langue des méthodes informatiques, en mettant à la disposition de ses locuteurs tous les moyens dont ils ont besoin aussi bien à l'oral qu'à l'écrit, afin qu'ils puissent réaliser les activités ou services suivants à travers l'outil informatique :

1. Dialogue avec la machine : synthèse et reconnaissance vocale, traduction automatique,
2. Outils pour écrire ou lire un texte,
3. Envoyer un courrier électronique, etc. »

Dans le cadre de sa thèse de doctorat, Berment [Berment, 2004] a défini un ensemble de ressources et logiciels pouvant servir de cadre pour l'informatisation d'une langue.

Le tableau ci-dessous contient une description des éléments de ce cadre.

Ressources	Logiciels
Dictionnaires bilingues ou d'usage	- Logiciels de traitement de la langue écrite - Logiciels de traitement de l'oral - Logiciels de traduction automatique et d'aide à la traduction - Logiciels de reconnaissance optique des caractères (OCR) - Logiciels fournissant des services avancés (serveur vocal, etc.) - Logiciels existants adaptés (localisation)

Tableau 1 : Cadre d'informatisation d'une langue

L'existence ou non des ressources ou logiciels définis dans le tableau ci-dessus pour une langue quelconque permet de la classer dans l'une des différentes catégories des langues naturelles, eu égard à son niveau d'informatisation. De cette classification, il faut retenir quatre catégories de langues en fonction de son degré d'informatisation : langue peu dotée, langue moyennement dotée, et langue bien ou très bien dotée.

Par ailleurs, pour qu'une langue soit classée dans l'une ou l'autre des catégories de langues, Berment [Berment, 2004] a établi une mesure du niveau d'informatisation d'une langue appelée l'indice-π. Cette mesure permet d'évaluer quantitativement le degré d'informatisation d'une langue en suivant le protocole ci-dessous.

[3] D'après le site www.ethnologue.com, en Ethiopie et au Nigéria, il n'existe pas moins de 527 langues et dialectes parlés dans ces deux pays. Au Cameroun, il y en a 288.

[4] https://www.avanquest.com/France/logiciels/le-grand-robert-501724

Pour chacun des services ou ressources, un groupe d'utilisateurs locuteurs de la langue est invité à donner des scores sous forme de note et d'indice de criticité. C'est la moyenne pondérée de ces notes, appelée indice-π, qui reflète le niveau global de satisfaction de chacun de ces locuteurs. En fonction de la valeur de cet indice, Berment [Berment, 2004] a défini les 3 niveaux d'informatisation suivants :

- Langues-π : la moyenne pondérée des résultats des notations des locuteurs évaluateurs est comprise en 0 et 9,99 (peu dotée),
- Langues-μ : la moyenne pondérée des résultats est comprise entre 10 et 13,99 (moyennement dotée),
- Langues-τ : la moyenne pondérée est comprise entre 14 et 20 (bien et très bien dotée).

I.1.1.2 Méthodologie pour informatiser une langue ou un groupe de langues peu dotées

I.1.1.2.1 Les critères pour informatiser une langue peu dotée

[Berment, 2004] a proposé l'esquisse d'un livre blanc pour l'informatisation d'un groupe de langues peu dotées, qui pourrait être mise en œuvre selon lui par les Nations Unies.

Le premier critère qu'il propose pour choisir ou sélectionner la ou les langues à informatiser serait le nombre de locuteurs. Selon Breton [Breton, 2003], les langues parlées par moins d'un million de locuteurs[5] ont une existence précaire, alors que celles ayant entre 10 000 et 100 000 locuteurs sont sérieusement menacées et que les langues avec moins de 10 000 locuteurs sont presque déjà en phase de disparition. Ce critère est également celui retenu par les grandes sociétés de distribution de logiciels de bureautique tels que Microsoft et Apple pour localiser leurs suites bureautiques. De nos jours, seule une quarantaine de langues parmi les 400 langues parlées ou écrites par plus de 1 000 000 de locuteurs dans le monde sont prises en charge dans le composant linguistique des principales suites bureautiques.

Voici ci-dessous, selon [Berment, 2004], les différents critères ou facteurs qui peuvent influer sur le choix d'informatiser ou non un groupe de langues peu dotées :

- le nombre de locuteurs,
- le caractère officiel ou national de la langue,
- le caractère central de la langue,
- l'intérêt des populations pour des moyens informatiques dans leur langue,
- la motivation des bailleurs pour l'informatisation d'une langue,
- le niveau d'informatisation de la langue,
- l'existence d'une grammaire et d'un dictionnaire,
- l'existence d'une langue- π proche,
- la présence d'un bilinguisme permettant de faciliter la communication.

I.1.1.2.1 Choix des services et d'organisation d'un projet d'informatisation d'une langue peu dotée

Toujours selon [Berment, 2004] après avoir identifié le groupe de langues ou la langue à informatiser, le projet d'informatisation peut être effectué en deux phases. La première phase consiste à choisir les services les plus importants pour ces langues.

[5] D'après Ethnologue (https://www.ethnologue.com/statistics/size, consulté le 3/12/2018), il y a 401 langues ayant plus d'un million de locuteurs natifs.

Généralement, pour un premier projet d'informatisation d'une langue peu dotée, on retient les services de traitement de texte, les ressources dictionnairiques (avec au moins 5000 articles) et un premier système d'aide à la traduction humaine ou de traduction automatique. Ces trois services constituent les premières briques pour amorcer une informatisation progressive d'une langue ou d'un groupe de langues peu dotées.

Enfin, dans un souci d'efficacité et dans l'optique de maîtriser toute la complexité et les différentes facettes d'un projet d'informatisation, [Berment, 2004] a défini deux types d'organisation, dont l'un concerne l'organisation technique et le second l'organisation chronologique du projet, après avoir procédé à un inventaire des tâches élémentaires pour informatiser une langue. Pour plus de détails sur ces deux types d'organisation, voir [Berment, 2004].

I.1.1.3 L'informatisation des langues : un moyen pour réduire la fracture numérique

La fracture numérique est un terme qui désigne la disparité qui existe d'une part entre les pays développés et les pays du tiers monde, et d'autre part les zones de peuplement urbain et les zones rurales, en ce qui concerne l'accès aux technologies de l'information et de la communication (téléphonie mobile, ordinateur et Internet). Cette inégalité se manifeste aux trois niveaux suivants :

- l'inégalité dans l'accès et l'utilisation d'un ordinateur (de bureau ou portable),
- l'inégalité dans l'usage d'outils informatiques,
- l'inégalité dans l'accès aux informations issues de ces outils numériques.

Parmi toutes les régions du monde, le continent africain est celui qui accuse le plus grand retard en termes d'accès et d'utilisation du numérique.

Outre les difficultés d'ordre économique ou structurelles qui peuvent creuser le fossé numérique chez certains peuples ou pays, l'analphabétisme et l'absence d'outils adaptés aux langues et cultures endogènes des pays concernés sont les principaux facteurs qui créent l'inégalité dans l'accès et l'utilisation des TIC.

Ainsi, l'informatisation des langues peu dotées réalisée dans le cadre des stratégies de développement des TIC dans les pays du tiers monde est une condition nécessaire pour réduire cette fracture numérique. Elle devra passer, dans un premier temps, par la constitution de ressources langagières pour les langues peu dotées, par l'adaptation ou la localisation des systèmes d'exploitation ou des suites bureautiques pour le plus grand nombre des langues parlées au monde, et par la construction d'outils et de systèmes d'aide à la traduction pour les populations analphabètes, dans le cadre de services d'aide à la lecture active augmentée d'entrées et sorties orales.

I.1.2 Spécificités du cas des langues africaines de la francophonie

I.1.2.1 Aperçu global et spécificités des langues africaines francophones

I.1.2.1.1 Brève description des langues africaines de la francophonie

Selon (Grimes, éd. 1996), il existe environ 2 035 langues parlées en Afrique, soit le tiers des langues parlés dans le monde, même si ce chiffre est instable puisque certaines langues s'éteignent et que d'autres voient le jour où sont découvertes d'une période à une autre.

Si l'on exclut les langues extra-africaines ou celles venues de l'extérieur depuis 2 000 ans, comme les langues indo-européennes (français, anglais, portugais, espagnol, ourdou, hindi),

les langues sémitiques (arabe), et les langues malayo-polynésiennes il reste un peu plus de 2 000 langues africaines autochtones parlées par les Africains eux-mêmes.

La Figure 1 ci-dessous récapitule les branches ou familles de langues parlées en Afrique.

Figure 1 : Familles des langues africaines[6]

D'après la figure ci-dessus, les langues africaines des pays francophones du continent appartiennent surtout aux familles des langues afro-asiatiques (arabe, berbère, somali, oromo), les langues de la famille nigéro-congolaises A et B (foulani, haoussa, wolof, songhay, les langues bantoues comme le lingala, le congo, le swahili, les langues des Comores), les langues de la famille nilo-sahariennes (langues tchadiennes et langues soudanaises), et enfin les langues austronésiennes ou malayo-polynésiennes (malgache).

[6] https://en.wikipedia.org/wiki/File:African_language_families_en.svg

I.1.2.1.2 Problèmes et spécificités des langues africaines francophones

Les principaux problèmes auxquels sont confrontées les langues africaines de l'espace francophone sont entre autres le nombre pléthorique de langues et dialectes qui peuvent être parlés dans un même pays (par exemple au Cameroun, il existe au moins 288 langues ou dialectes).

Le manque de standardisation des langues africaines (adoption d'une orthographe fixe), et l'inexistence des ressources linguistiques qui pourraient faciliter leur informatisation ou leur localisation dans les outils informatiques et bureautiques, font partie des problèmes de ces langues.

Par ailleurs, il convient de préciser que la grande majorité des langues africaines n'avaient pas de systèmes d'écriture jusqu'à récemment et que les cultures et les traditions locales se transmettaient plutôt par la voie orale que par l'écriture. Cependant, depuis plusieurs siècles des communautés africaines en dehors de l'Afrique du Nord ont utilisé l'alphabet arabe ou latin pour écrire leurs langues. Ainsi, selon [D. Osborn, 2011], « les systèmes d'écriture des langues africaines peuvent être classés en trois grandes catégories :

- **Anciens systèmes d'écriture ayant évolué avec une langue ou une famille de langues.** L'arabe, l'alphasyllabaire éthiopien/guèze et le tifinagh en sont des exemples encore utilisés aujourd'hui.
- **Systèmes d'écriture inventés au cours des deux derniers siècles**, c'est à dire relativement récemment. Certains sont encore utilisés (comme le n'ko, le kikakui, le vaï et le mandombe), tandis que d'autres ont disparu.
- **Adaptation d'un autre alphabet**. L'exemple précolonial le plus répandu est l'utilisation des caractères arabes par les musulmans érudits pour écrire des langues africaines autres que l'arabe. Mais l'alphabet latin, introduit par les Européens lors de la colonisation, est sans aucun doute le plus répandu aujourd'hui. »

Aussi, depuis leur indépendance, les pays africains francophones de l'ex-empire colonial français ont adopté la langue française comme langue officielle de l'administration et du système éducatif. De plus, à cause de l'analphabétisme et de la pauvreté, beaucoup d'Africains et surtout ceux des communautés villageoises et rurales n'ont pas accès à l'information et sont donc victimes de la fracture numérique.

I.1.2.1.3 Lien entre l'informatisation du somali de Djibouti et les langues d'Afrique francophone

Nous avons choisi de faire le lien entre l'informatisation du somali de Djibouti et celle des langues des pays d'Afrique francophone pour les deux raisons suivantes.

- Comme nous le détaillerons dans la section I.2.1.2, la langue somalie est parlée dans 4 pays de la Corne d'Afrique. Nous avons ciblé pour notre étude la variante officielle, parlée par les Somalis de la République de Djibouti. Comme ce pays est francophone de par son passé colonial avec la France, et comme le français est la langue officielle du pays, les résultats de notre expérimentation de la traduction automatique de français en somali des articles du journal La Nation de Djibouti intéresseront en premier lieu les langues des pays d'Afrique francophone, comme le Sénégal, le Mali, le Burkina Faso et le Niger, etc.
- Pour la majorité des couples de langues offerts par GT, Google utilise l'anglais comme langue pivot avant de produire la traduction vers une quelconque langue cible (qu'elle soit africaine ou autre). En construisant un corpus bilingue français-somali à partir des prétraductions de GT, nous ouvrons une voie simple et efficace pour construire des

corpus parallèles spécialisés entre la langue française et les langues africaines ayant un lien avec le français afin de développer rapidement des systèmes de TA pour ces langues.

I.1.2.2 Répartition géographique et nombre de langues dans quelques pays africains

I.1.2.2.1 Géographie linguistique et nombre des familles de langues en Afrique

D'après la base de données Ethnologue[7], le continent africain connaît une très grande diversité linguistique par rapport aux autres continents. Le nombre de langues en Afrique représente le tiers des 7 097 langues vivantes parlées dans le monde, ce qui est estimé par Ethnologue à 2 143 langues pour une population africaine estimée en 2018 à environ 1,2 milliard d'habitants.

Toujours d'après Ethnologue, les langues africaines se répartissent en 5 grandes familles de langues qui sont les suivantes.

- **Les langues de la famille Niger-Congo.** Cette division regroupe 1 540 langues africaines réparties en 10 branches (Kordofanien, Mandé, Atlantique, Ijoïde, Kru, Gur, Adamawa-Oubanguien, Kwa, Bénoué-Congo et le Dogon) qui sont parlées en Afrique de l'Ouest, en Afrique de l'Est, en Afrique Australe et en Afrique Centrale. Elles sont principalement parlées dans les pays suivants : Burkina Faso, Côte d'Ivoire, Cameroun, Congo-Brazzaville, Congo-Kinshasa, Mali, Tchad, Nigéria, Libéria, Sierra-Leone, Kenya, Tanzanie, Ouganda, Somalie, Zambie, Mozambique, Zimbabwe, Malawi, Angola, Sénégal, Guinée-Bissau, Guinée-Conakry, Niger, Gambie, Bénin.
- Les langues de la famille Nilo-Saharienne. Cette division regroupe 205 langues africaines reparties en 6 branches (Songhay, Saharien, Maban, Fur, Chari-Nil, Koma) qui sont parlés en Afrique de l'Est, en Afrique du Nord et Afrique Centrale. Les langues appartenant à cette famille sont parlées dans les pays suivants : Ouganda, Soudan, Sud Soudan, Égypte, Érythrée, Éthiopie, Tchad, République Centrafrique, Niger, Nigeria, Mali, Benin, Algérie, Kenya, Tanzanie, Congo-Kinshasa.
- Les langues de la famille Afro-Asiatique. Cette division regroupe 377 langues africaines reparties en 5 branches (Sémitique, Lybico-berbère, Tchadique, Egypto-copte, Couchitique-omotique) qui sont parlées en Afrique du Nord, en Afrique de l'Est, en Afrique de l'Ouest et au Moyen-Orient. Les langues appartenant à cette famille sont parlées dans les pays africains suivants : Algérie, Égypte, Éthiopie, Érythrée, Somalie, Djibouti, Maroc, Mauritanie, Lybie, Tchad, Tunisie, Niger, Nigéria, Kenya, Tanzanie, Cameroun, Mali, Soudan.
- Les langues de la famille Khoïsan. Cette division regroupe 28 langues africaines réparties en 5 branches (Nord, Centre, Sud, Hadza et Sandawe) qui sont parlées en Afrique de l'Est, en Afrique Australe et en Afrique du Sud. Les langues appartenant à cette famille sont principalement parlées dans les pays suivants : Tanzanie, Namibie, Angola, Afrique du Sud, Botswana, Angola.
- Les langues de la famille Austronésien. Cette division regroupe 1 256 langues qui sont surtout parlées en Asie du Sud-est, dans l'Océan Pacifique et à Madagascar. C'est la deuxième famille de langues en nombre après celle du Niger-Congo. Seul le malagasy (malgache) parlé à Madagascar et sa variante shibushi parlée aux Comores et à Mayotte appartient aux langues africaines.

[7] https://www.ethnologue.com/browse/families

I.1.2.2.2 Situation linguistique des quelques pays d'Afrique francophone

Les pays d'Afrique francophone représentent environ la moitié des pays du continent africain (26/54 pays) et le nombre de locuteurs ayant la langue française comme langue officielle ou d'enseignement dans leurs pays, à côté des langues locales et nationales, est estimée selon l'OIF[8] à 274 millions. Par ailleurs, à l'exception des locuteurs de la famille des langues Khoïsan, toutes les autres familles de langues répertoriées en Afrique (afro-asiatiques, nilo-sahariennes, nigéro-congolaises, malayo-polynésiennes) sont parlées par des populations également francophones. En effet, l'influence de la langue française, date de la période coloniale, et tous ces 26 pays ont comme langue officielle le français.

Cependant ces dernières années, les gouvernements de certains pays d'Afrique francophone (Niger, Sénégal, Maroc) ont décidé de développer linguistiquement les langues nationales en leur accordant un statut de langue d'enseignement dans l'enseignement élémentaire, et en finançant des programmes de construction de ressources linguistiques pour ces langues majoritairement très faiblement dotées.

C'est ainsi que le Royaume du Maroc a créé le 17 octobre 2001 un institut royal pour la culture amazighe, l'IRCAM dont l'objectif est « la promotion de la culture et le développement amazighe »[9] afin de développer l'usage et l'étude linguistique de la deuxième langue en termes de locuteurs de ce pays.

Le Sénégal, quant à lui, a institué dans sa constitution du 7 janvier 2001 que les langues nationales du pays sont le diola, le malinké, le pular, le sérère, le soninké, le wolof. Des programmes d'enseignement de ces langues dans les écoles maternelles et primaires ont été conçus, ainsi que des actions de recherche pour amorcer l'informatisation de ces langues et surtout le wolof qui est lingua franca du Sénégal à côté du français. Depuis le 9 décembre 2014, les interventions des députés sénégalais à l'hémicycle sont traduites simultanément dans les six langues nationales (diola, maliné, pular, sérère, soninké et le wolof). En conséquence, chaque député peut s'exprimer dans sa langue maternelle, ainsi qu'en français.

Enfin au Niger, l'état a mis en place des arrêtés définissant et fixant les systèmes d'écriture des quatre langues nationales suivantes : haoussa, kanouri, tamajaq et zarma. A l'instar du Sénégal, les députés nigériens peuvent s'exprimer dans leurs langues maternelles grâce à l'existence d'un système d'interprétation simultané des débats parlementaires.

I.2 Le cas du somali

I.2.1 Histoire et situation socio-politique du somali

I.2.1.1 Brève histoire de la langue somalie

La langue somalie, à l'instar des langues africaines, est une langue munie d'un système d'écriture officiel très récent, car il date du début du XXème siècle. En effet, même s'il a existé un alphabet indigène basé sur l'alphabet arabe depuis le XIVème siècle[10], la plupart des érudits somalis écrivaient en arabe jusqu'au début du siècle dernier.

[8] https://www.francophonie.org/Estimation-des-francophones.html

[9] http://www.ircam.ma/

[10] Des légendes orales racontent qu'un certain Yusuf Al Kawnayn qui est connu au moyen âge sous le nom de Abu Barakaat Yusuf al-Barbari c'est à dire originaire de Berbera (Ville du nord de la Somalie) a été le premier somalien qui a écrit un alphabet basé sur l'arabe pour la langue somalie.

Outre cet alphabet arabe du somali, communément appelé « alphabet *wadaad* », il y a eu plusieurs tentatives de création d'alphabets autochtones de la part de plusieurs savants ou linguistes somaliens : l'alphabet *Osmanya* du nom de son auteur Osman Yusuf Kenadid (cf. Figure 2), l'alphabet *Borama* du nom de la ville d'origine (Borama) de son auteur Sheikh Abdurahman Sheikh Nuur, et enfin l'alphabet *Kaddare* de Hussein Sheikh Ahmed Kaddare.

L'alphabet latin est celui utilisé actuellement par tous les somalophones de par le monde et c'est celui que le gouvernement de Somalie a désigné comme alphabet d'écriture de la langue officielle du pays. L'adoption de cet alphabet a eu lieu le 21 octobre 1972, jour du troisième anniversaire du gouvernement révolutionnaire de la Somalie sous l'égide du président somalien Mohammad Siyaad Barre. C'est la proposition du linguiste somalien Shire Jama Shire qui a été retenue suite à un appel à proposition lancé par le gouvernement avec le soutien de l'UNESCO sous la direction du professeur de linguistique africaine B. W. Andrzejewski [B. W. Andrzejewski, 1974].

Paradoxalement, la Somalie ayant adhéré en 1974 à la Ligue Arabe sous l'influence de l'Egypte et de l'Arabie Saoudite, deux ans après avoir choisi l'alphabet latin, elle l'a toujours maintenu, malgré la pression et le financement conséquent qu'elle obtenait pour promouvoir la langue arabe de la part de cette instance.

Figure 2 : Système d'écriture autochtone de l'alphabet Osmanya[11]

I.2.1.2 La situation socio-politique actuelle du somali

Malgré sa spécificité d'être, avec l'amharique (langue officielle de l'Éthiopie) et du swahili (Kenya), l'une des trois premières langues africaines ayant eu un statut de langue officielle dans un état du continent africain, le développement linguistique du somali a beaucoup souffert de la guerre civile somalienne qui fait rage dans ce pays depuis janvier 1990.

Par ailleurs, selon [Morin, 1986], depuis son adoption comme langue officielle, « la politique linguistique de la Somalie, qui a été caractérisée par une scolarisation et une alphabétisation massives, a eu un double effet :

- la généralisation de l'emploi du somali dans tous les domaines : administratif, économique, scolaire, scientifique, et le recul de toutes les autres langues étrangères parlées traditionnellement en Somalie, à l'exception notable de l'arabe ;
- l'hostilité croissante des États voisins principalement de l'Éthiopie, inquiète de l'apparition d'une langue africaine écrite, rivale régionale de l'amharique. »

[11] https://fr.wikipedia.org/wiki/Somali#/media/File:Ciismaniya.jpg

Comme indiqué dans la Figure 4 ci-dessous, le somali est parlé dans 3 pays de la Corne de l'Afrique, qui partagent tous leurs frontières avec la Somalie (Kenya, Djibouti, Ethiopie). Cette proximité géographique peut parfois créer des confusions en termes de nationalité concernant les personnes ayant comme langue maternelle le somali, et démontre l'imbrication des situations sociolinguistiques de la Somalie avec ses voisins. Ainsi, il convient de distinguer entre les Somalis qui sont les habitants de la Corne d'Afrique et ethniquement somalis, les Somaliens, qui sont ceux disposant de la citoyenneté de la Somalie (ex Somalia Italia et British Somaliland), et enfin les somalophones, qui parlent ou écrivent en somali et qui peuvent être aussi des non-Somalis, comme les Oromos somalisés des régions Bale et Sidamo de l'Éthiopie, les Afars et les Arabes[12] djiboutiens ayant le somali comme seconde langue, et qui vivent surtout dans les villes de Djibouti et de Dikhil en République de Djibouti.

I.2.2 Typologie et famille linguistique du somali

I.2.2.1 Origine et typologie

La langue somalie fait partie de la branche des langues couchitiques, et appartient à une sous-division de la grande famille des langues dites afro-asiatiques ou chamito-sémitiques, avec l'omotique, le berbère, le sémitique et l'égyptien ancien. La sous-branche des langues couchitiques est composée d'une trentaine de langues. La langue somalie est la deuxième langue couchitique en nombre de locuteurs, après l'oromo[13]. Elle appartient également à la sous branche est-couchitique ou au LEC (Lowlands East Cushitic) avec le rendille et le boni.

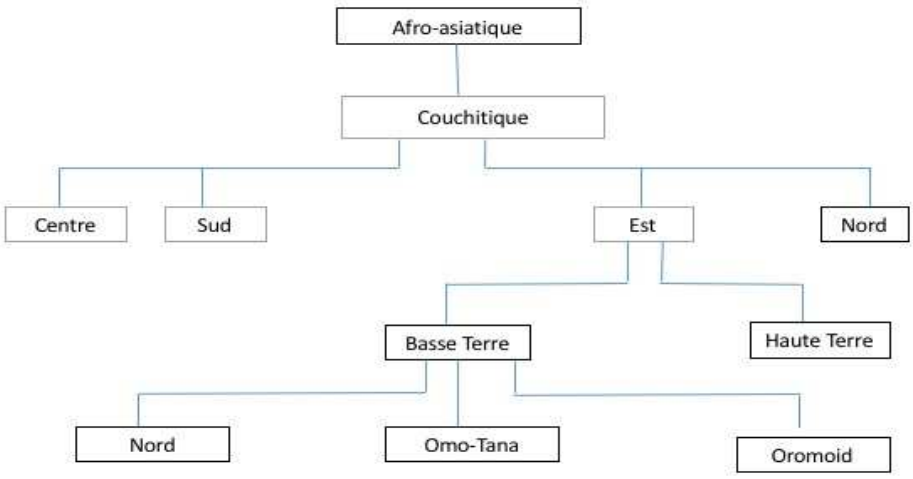

Figure 3 : Famille des langues afro-asiatique couchitiques

L'aire géographique de la langue somalie s'étend entre le sud de Djibouti, le sud-est de l'Éthiopie, la Somalie, et le nord-est du Kenya, comme le montre la Figure 4 suivante :

[12] Environ 10 à 15% de la population djiboutienne est d'origine arabe yéménite.

Voir aussi : http://www.axl.cefan.ulaval.ca/afrique/djibouti.htm

[13] L'oromo est la première langue parlée en Éthiopie, avec au moins 40 millions de locuteurs.

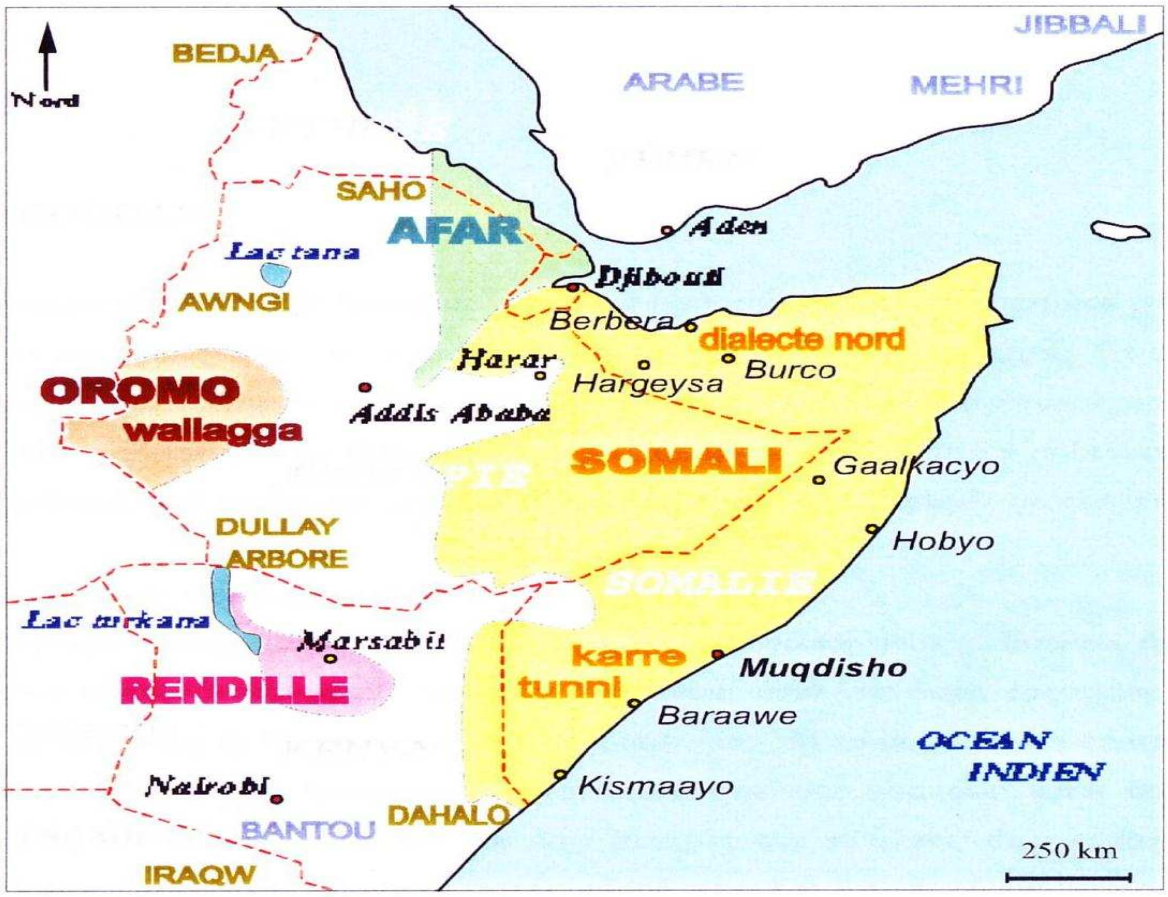

Figure 4 : Aire géographique du somali[14]

La position de la langue somalie dans les sous-branches des langues couchitiques et des familles afro-asiatiques est donnée dans la Figure 3 ci-dessus.

[14] Source : [Barillot, X., 2002]

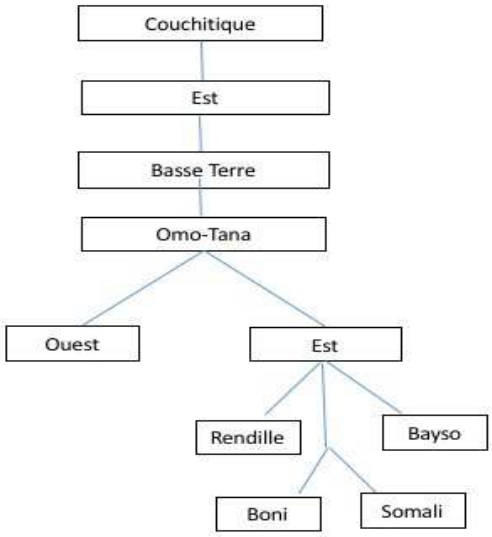

Figure 5 : Schéma des langues est-couchitiques

I.2.2.2 Structure phonologique et phonétique du somali

Comme dans toutes les langues naturelles, une syllabe est composée en somali d'un ou plusieurs phonèmes, et chaque mot de la langue somalie contient au moins une syllabe. La structure syllabique du somali est simple ; elle est constituée comme suit :

- **v** : une seule voyelle, qu'elle soit brève ou longue, ou bien une diphtongue.
  Exemple : *ú* (à), *oo* (et), *èy* (chien)
- **cv** : une consonne suivie d'une voyelle (brève ou longue) ou d'une diphtongue.
  Exemple : *kú* (dans), *síi* (donner), *cáy* (insulte).
- **vc** : une voyelle (brève ou longue) suivi d'une consonne.
  Exemple : *ul* (bâton), *il* (œil).
- **cvc** : une consonne suivie d'une voyelle (brève ou longue), suivie d'une consonne.
  Exemple : *nin* (garçon).

Les quatre structures ci-dessus avec des voyelles brèves forment la structure syllabique de base du somali. Il faut leur ajouter 4 autres formes de syllabes, en remplaçant chaque voyelle brève par une voyelle longue. Ainsi, au total, il existe 8 structures syllabiques en somali.

Le Tableau 2 ci-dessous présente la structure phonétique des consonnes de la langue somali.

	Labiales	Labio-dentales	Dentales	Alveo-laires	Retro-flexes	Palatales	Vélaires	Uvulaires	Pharyn-gales	Glot-tales
Occlusives voisées	b		d		dh		g	Q		'
Occlusives non voisées		t				k				
Nasales	m			n						
Fricatives voisées		f		s		sh		Kh	x	h
Fricatives non voisées						j			c	
Roulées				r						
Latérales				l						
Approximantes	w					y				

Tableau 2 : Structure phonétique des consonnes du somali[15]

I.2.3 Dialectes, locuteurs et répartition géographique

I.2.3.1 Les différents dialectes du somali

Contrairement à d'autres langues africaines, et bien qu'il existe plusieurs dialectes parlés actuellement, la langue somalie jouit d'une relative homogénéité. Il existe trois variétés ou dialectes dans cette langue : le dialecte du nord (ou somali standard), le dialecte benadiri (ou somali littoral) et enfin le may.

Le dialecte officiel qui a été choisi pour l'adoption de l'écrit du somali est celui du nord, qui est considéré comme le somali standard par toutes les communautés somalies de par le monde [Saeed 1999], [Abdullahi 1995]. Voici quelques précisions sur les trois principaux dialectes du somali repérés par les études linguistiques [Saeed 1999], [Abdullahi, 1995].

- Le may est la langue des Rahanweins, l'une des grandes tribus somalies habitant dans le sud-ouest du pays. La ville de Baidoa (Baydhaba), qui se trouve à 250 km de Mogadiscio, est le chef-lieu de ce dialecte. La plupart des locuteurs de ce dialecte habitent dans cette ville et ses alentours. C'est un dialecte incompréhensible pour les Somalis du nord et de Djibouti.
- Le benadiri est le dialecte des habitants de Mogadiscio, ainsi que de la région côtière dite Benaadir de la Somalie allant de la ville de Marka à Mogadiscio. On appelle aussi ce dialecte, le somali du littoral.
- Le somali du nord est le dialecte parlé dans le centre et dans le nord-est (Puntland), dans le nord-ouest (Somaliland) ainsi qu'en République de Djibouti. Il fait office de langue commune et officielle de tous les somalis.

I.2.3.2 Les locuteurs du somali et leur répartition géographique

Le nombre de locuteurs du somali ainsi que leur répartition géographique ou leur pays de résidence sont récapitulés dans le tableau ci-dessous. Comme on le voit, du fait d'une importante diaspora somalienne qui est installée dans plus 20 pays depuis la guerre civile en Somalie, la langue somalie est l'une des rares langues africaines à être parlée et utilisée dans

[15] Source : [Abdillahi N. et al., 2007]

27 pays répartis dans les 5 continents de la planète (Afrique, Europe, Asie, Amériques et Océanie). Avec un nombre de locuteurs total estimé entre 16 224 000 [Ethnologue, 2018] et 22 796 942 (Tableau 3) locuteurs, le somali fait partie des dix langues les plus parlées dans le continent africain.

Classement	Pays	Nombres de locuteurs	Répartition (en %)
1	Somalie	14 500 000	62,81
2	Ethiopie	4.5 million	20,18
3	Kenya	2.4 million	10,53
4	Djibouti	524 000	2,3
5	Yémen	200 000	0,88
6	Etats-Unis d'Amériques	135 266	0,6
7	Royaume-Uni	98 000	0,43
8	Emirats Arabes Unis	90 900	0,4
9	Suède	66 369	0,29
10	Canada	62 550	0,27
11	Norvège	43 196	0,19
12	Afrique du Sud	40 000	0,17
13	Pays-Bas	39 465	0,17
14	Arabie Saoudite	34 000	0,15
15	Allemagne	33 900	0,15
16	Egypte	22 709	0,1
17	Danemark	21 210	0,09
18	Finlande	19 059	0,08
19	Australie	16 169	0,07
20	Italie	8 228	0,04
21	Suisse	7 025	0,03
22	Autriche	6 161	0,03
23	Belgique	2 627	0,01
24	Pakistan	2 500	0,09
25	Lybie	2 500	0,09
26	Nouvelle-Zélande	1 617	0,007
27	Irlande	1 495	0,006
	Total	22 796 942	100

Tableau 3 : Répartition géographique des locuteurs somalis

I.2.4 Le somali et les langues africaines de la francophonie

I.2.4.1 L'informatisation du somali par rapport aux autres langues africaines de la francophonie

Selon la classification de [Berment, 2004], la langue somalie fait partie des langues faiblement dotées, à l'instar de la majorité des 2000 langues africaines. Cependant, depuis une dizaine d'années, le niveau d'informatisation du somali commence à croître puisqu'il existe déjà des corpus linguistiques du somali issus du Web, et qui ont servi à réaliser des outils ou des applications qui sont de nature à faciliter l'utilisation de cette langue dans les technologies de l'information et de la communication (TIC).

Parmi ces outils, on trouve des correcteurs orthographiques réalisés par des sociétés privées[16], ou par des chercheurs et particuliers passionnés par le traitement automatique de la langue, tel que l'ingénieur somalien Mohamed I. Mursal qui a collaboré avec le professeur Kevin Scannel pour construire un premier correcteur orthographique libre du somali[17].

Récemment, dans le cadre de notre travail de thèse, nous avons construit un premier prototype d'étiqueteur grammatical du somali [H.A Assowe, 2013].

D'autres ressources et applications de TALN ont été développées pour le somali au cours de ces dix dernières années. Des descriptions détaillées sur ces ressources linguistiques sont donnés dans le second chapitre de ce mémoire.

I.2.4.2 La traduction en somali des sources d'information écrites en français à Djibouti

En dépit de la présence abondante du somali sur le web [Van der Verken et al., 2003] et de la croissance des sites web journalistique somaliens sur la toile [Scannell K., 2007], il n'existe à ce jour pratiquement aucun système ou outil de traduction automatique des sources d'information écrites depuis et vers le somali.

L'exception notable est Google, qui a inclus le somali parmi les 5 langues africaines ajoutées dans son outil de traduction gratuite Google Translate le 28 août 2013.

Pour accéder en somali aux nouvelles journalistiques publiées en langue anglaise, comme celles des médias du Kenya, de l'Éthiopie ou d'autres pays, les somalophones du continent peuvent utiliser GT-en-so, dont les traductions sont très loin d'être bonnes, mais donnent une idée du contenu. En revanche, à cause du passage par l'anglais, ce n'est pas le cas pour GT-fr-so. Au début de notre travail, aucun système de TA français-somali capable de traduire de façon utilisable les articles du journal francophone La Nation de Djibouti n'existait à ce jour. Malgré le faible poids des Somalis (cf. **Erreur ! Source du renvoi introuvable.**) de Djibouti, la République de Djibouti occupe une place centrale dans la résolution du conflit somalien, et plusieurs conférences de réconciliation entre Somaliens ont eu lieu à Djibouti depuis 1991. L'existence d'un système de TA spécialisé au sous-langage des nouvelles journalistiques de *La Nation de Djibouti* permettrait de faciliter l'accès des Somalis du continent aux informations diffusées dans ce quotidien.

En République de Djibouti, l'article 1 de la constitution[18] stipule que les deux langues officielles du pays sont le l'arabe et le français. Par conséquent, toute l'administration publique, l'éducation, les médias, la presse écrite et audiovisuelle sont écrits et lus en français et en arabe.

Par ailleurs, il existe deux journaux de la presse écrite, qui sont d'une part un quotidien, le journal francophone *La Nation de Djibouti*[19], et un hebdomadaire, le journal arabophone *Al Qarn*[20]. Malheureusement, à part les Djiboutiens éduqués et jouissant d'un niveau d'éducation scolaire ou universitaire suffisant en français et en arabe, la majorité des Djiboutiens, et surtout les femmes et les personnes âgées non scolarisées, n'ont pas accès aux informations diffusées dans ces deux journaux, d'où une inégalité en termes d'accès à l'information officielle et utile en République de Djibouti. Cette situation persiste encore de nos jours à

[16] www.somitek.com

[17] http://www.somaliaonline.com/community/topic/somali-born-engineer-develops-spell-checker/

[18] https://fr.wikisource.org/wiki/Constitution_de_Djibouti

[19] http://www.lanationdj.com

[20] http://www.alqarn.dj/Open_Pages.php?P=4

Djibouti, en raison de l'inexistence d'une traduction professionnelle ou automatique des contenus de ces deux journaux en langues nationales, que ce soit en somali ou en afar.

I.3 Objectifs pratiques et théoriques possibles pour une première thèse en informatique sur l'informatisation du somali

I.3.1 Objectifs pratiques

I.3.1.1 Les besoins principaux pour l'informatisation du somali

I.3.1.1.1 Répondre aux besoins d'accès en somali du journal local francophone de Djibouti

Pour une diffusion plus large et inclusive de l'information publiée actuellement en langue française par l'unique organe de presse écrite en République de Djibouti, *La Nation de Djibouti*, le besoin de traduire vers les langues nationales le contenu de la version numérique de ce journal s'est fait de plus en plus sentir ces derniers temps. Avec l'arrivée des smartphones bon marché et la technologie Internet 3G et 4G, de plus en plus de jeunes Djiboutiens utilisent l'Internet pour chercher des informations ou suivre l'actualité du pays à l'aide de leurs téléphones portables.

Outre les Djiboutiens somalophones, il existe une grosse communauté de somalophones établis dans les pays occidentaux ou dans les pays voisins (Somalie, Éthiopie ou Kenya) qui ne peuvent suivre ou accéder aux informations publiques et à l'actualité du pays à cause de la barrière linguistique. Ainsi, il existe à Djibouti et dans son voisinage immédiat un vrai besoin d'accès en somali à l'information véhiculée par les articles et dépêches publiées quotidiennement sous forme électronique sur le site web de *La Nation de Djibouti*.

I.3.1.1.2 Constitution d'une communauté Web pour l'amélioration de la TA journalistique par contribution volontaire

En l'absence d'un système de traduction automatique français-somali des nouvelles journalistiques, il faudra constituer une communauté de bénévoles et de passionnés de la traduction journalistique pour amorcer un premier travail de post-édition et de traduction corrigée en utilisant dans un premier temps Google Translate (qui passe par un pivot textuel anglais).

Dans le domaine de la traduction professionnelle, il existe une grande communauté de traducteurs anglais-somali [21] grâce à l'importante diaspora somalienne établie depuis longtemps aux USA, au Royaume Uni, au Canada, et dans les pays anglophones. Toutefois, il n'existe peu ou pas de traducteurs professionnels pour le couple français-somali.

Selon Wikipédia, « La production participative, l'externalisation ouverte ou le crowdsourcing, est l'utilisation de la créativité, de l'intelligence et du savoir-faire d'un grand nombre de personnes, en sous-traitance, pour réaliser certaines tâches traditionnellement effectuées par un employé ou un entrepreneur. »

Ainsi, pour pallier le manque de traducteurs français-somali, il faudra créer un groupe ou une communauté de traducteurs/post-éditeurs bénévoles des articles du journal La Nation de Djibouti, dans le cadre d'un travail de contribution bénévole, à l'aide d'une plateforme dédiée et libre, dans une optique d'amélioration continue de la TA journalistique à partir de Google Translate et d'autres systèmes construits dans le cadre de cette thèse.

[21] https://www.proz.com/freelance-translators/english-to-somali/

I.3.1.2 Les applications

Les travaux qui doivent être réalisés dans le cadre de cette thèse ont des objectifs pratiques qui vont contribuer à l'informatisation du somali. Ces objectifs pratiques répondent à un besoin spécifique, à savoir offrir l'accès en somali au journal local francophone *La Nation de Djibouti* aux populations somalophones du continent africain, et plus particulièrement des pays voisins de Djibouti (Somalie, Éthiopie, Kenya et le Yémen).

En effet, il s'agit de construire et de déployer un système d'accès en somali au site de *La Nation de Djibouti*. Ci-dessous, nous allons décrire les applications informatiques qui ont un lien avec notre thème de recherche et qui ont été ou qui doivent être développées pour répondre à des besoins attestés, dans le cadre de l'informatisation du somali.

I.3.1.2.1 Système de traduction de dialogue oraux (TAD) bilingue pour les activités sanitaires en zone de conflit ou camps de réfugiés

À cause de la diversité linguistique et de la difficulté permanente de trouver des traducteurs en langues africaines dans certaines situations critiques comme les catastrophes naturelles, les missions humanitaires ou les opérations de maintien de la paix, les technologies issues du traitement automatique du langage naturel, telles que la traduction et la reconnaissance automatique de la parole ou la synthèse vocale, peuvent fournir des solutions adéquates et efficaces pour faciliter la communication et l'intercompréhension entre les humanitaires et les populations victimes des guerres ou des famines.

C'est ainsi que, dans le cadre d'un projet pilote entre plusieurs laboratoires de recherche et l'université de Carnegie-Mellon (CMU) [Carbonell et al., 2004], un système de traduction de dialogues (TAD) oraux bilingues (anglais-somali) pour les activités humanitaires (sanitaires et contre la famine) a été expérimenté en utilisant PTRANS (Portable TRANSlator).

I.3.1.2.2 Construction d'un système de TA améliorable en continu

À cause de l'indisponibilité de corpus parallèles de bonne qualité pour construire un système de TA de bonne qualité français-somali, il faut s'inspirer des méthodes qui ont été appliquées pour d'autres couples de langues, afin de surmonter cette difficulté [Wang L., X, 2015].

À l'aide des boites à outils disponibles et libres de droit telles que MOSES [Koehn et al., 2007] et d'autres, on peut aujourd'hui apprendre un système de TA à partir d'un corpus bilingue construit par post-édition de traductions de GOOGLE TRANSLATE. Ensuite, en utilisant des méthodes d'apprentissage incrémental et avec le soutien d'une communauté de bénévoles traducteurs/post-éditeurs, la TA français-somali issue de ces méthodes pourra s'améliorer en continu.

I.3.1.2.3 Localisation des logiciels grand public

La localisation est la tâche qui consiste à traduire et adapter culturellement les messages contenus dans les interfaces d'utilisation des applications et logiciels, la création de contenus Internet en langues locales. L'objectif principal de la localisation est de faciliter l'accès et l'utilisation de l'informatique, et les nouveaux outils des TIC en visant le plus grand nombre possible de personnes, et surtout celles des pays pauvres, en leur offrant des contenus traduits dans leurs langues maternelles.

Par ailleurs, à travers la localisation des logiciels grand public, les populations africaines pourront alors tirer pleinement profit des atouts des TIC et réduire ou combler leur retard technologique et leur fracture numérique.

I.3.1.3 Les ressources

I.3.1.3.1 Corpus monolingue, bilingue, brute ou annoté

Un corpus monolingue est un ensemble de données textuelles écrit dans une certaine langue. Il peut être généraliste, c'est-à-dire contenant des données de divers sujets ou thèmes, ou bien spécialisé c'est-à-dire constitués de textes d'un seul domaine ou sujet.

Généralement, les corpus monolingues servent surtout en TAL à construire des modèles statistiques de la langue, et à créer des correcteurs orthographiques ou bien un lexique et un dictionnaire monolingue.

Un corpus bilingue est un ensemble de données textuelles alignées en niveau segments (phrases, titres ou libellés) et qui sont écrites dans deux langues différentes, sachant que pour chaque paire l'un est la traduction de l'autre. Comme un corpus monolingue, un corpus bilingue peut être généraliste ou spécialisé. Les corpus bilingues sont surtout utilisés pour créer les tables ou modèles de traduction dans un système de traduction automatique, ou, pour créer des dictionnaires bilingues.

Qu'il soit monolingue ou bilingue, un corpus peut être brut (il contient seulement des données textuelles sans aucune autre information), ou bien annoté (il contient outre les textes, des informations supplémentaires de nature à enrichir les données textuelles à des fins de traitements postérieurs du TAL). Ces applications sont principalement l'étiquetage morphosyntaxique, la désambiguïsation lexicale ou sémantique, la TA, etc.). Les annotations linguistiques d'un corpus peuvent être de type externe ou interne.

I.3.1.3.2 Extraction et constitution d'un dictionnaire bilingue spécialisé

Des travaux de recherche récents ont montré l'intérêt des dictionnaires bilingues spécialisés pour l'amélioration de la qualité de la TA à base d'exemples [Nasredine et al., 2016].

Il est également prouvé que l'utilisation de petits corpus parallèles spécialisés produisait de meilleures performances que l'utilisation de données généralistes ou hors domaine, au moins dans le cas de système de TA statistique [Hajlaoui & Boitet, 2008].

Ainsi, dans le cadre du développement de systèmes de TA spécialisés pour des langues peu dotées, comme c'est notre cas avec le couple français-somali, les données parallèles construites à partir de la post-édition pourront être utilisées pour extraire et constituer un dictionnaire bilingue spécialisé qui plus tard pourra améliorer la TA statistique. Ce dictionnaire devra couvrir une partie ciblée du vocabulaire général et spécifique, par exemple celui du journal francophone *La Nation de Djibouti*.

I.3.1.4 Les outils

I.3.1.4.1 Outils d'aide à l'apprentissage de la lecture

Parmi les outils qui peuvent faciliter l'apprentissage d'une langue par la lecture active, il y a la possibilité de créer et de mettre en ligne des dictionnaires munis de synthèse vocale. En outre, on peut dynamiser l'apprentissage d'une langue étrangère pour des populations analphabètes ou non scolarisées, en combinant la reconnaissance de la parole, la traduction automatique et la synthèse vocale entre le texte à lire, écrit souvent dans une langue bien dotée et non africaine, et la langue maternelle ou locale que l'apprenant africain maîtrise ou pratique quotidiennement.

Ainsi, le développement d'outils d'aide à l'apprentissage de la lecture issus du TALN pourrait grandement participer à l'alphabétisation des communautés villageoises ou rurales de certains pays d'Afrique francophone.

I.3.1.4.2 Outils d'aide à l'écriture

À l'instar des outils de TALN d'aide à l'apprentissage de la lecture, le développement d'outils de base comme les correcteurs orthographiques et grammaticaux, même basiques, peut faciliter et accompagner l'apprentissage de l'écriture en langues africaines.

Inclus dans toutes les versions du Pack Office de Microsoft Word et dans Open Office pour certaines langues bien dotées, les correcteurs orthographiques et grammaticaux sont devenus de nos jours un moyen pour améliorer les productions textuelles ou apprendre à corriger les erreurs d'écriture. Malheureusement, comme nous l'avons rappelé, la majorité des langues de l'Afrique francophone est dépourvue de ces outils, à l'exception de quelques-unes comme le wolof, l'afrikans[22], le tswana[23] , le haoussa, le berbère ou le swahili.

I.3.2 Objectifs théoriques

I.3.2.1 Estimation de la qualité linguistique d'un corpus bilingue sans traduction professionnelle

I.3.2.1.1 Évaluation de la qualité d'un corpus bilingue post-édité

L'évaluation d'un corpus bilingue construit par post-édition des pré-traductions de Google Translate est un excellent moyen pour avoir un aperçu global de la qualité dudit corpus, qui servira par la suite à entraîner un système de TA statistique. Les éléments ci-dessous sont généralement utilisés dans la littérature pour évaluer un corpus post-édité.

- **Le temps de post-édition (par page source) des pré-traductions automatiques**. Plus on passe du temps à post-éditer un segment, plus la post-édition se rapproche d'une retraduction complète,
- **La distance d'édition** (mixte ou non) entre les segments pré-traduits et les segments de la post-édition. On observe une corrélation qui, si on arrive à la préciser, permettra d'estimer l'effort de post-édition sur une autre traduction brute que celle qui a effectivement été post-éditée : plus cette distance est grande, plus on a passé de temps à effectuer cette post-édition.
- Le nombre de mots ou de segments post-édités par minute. Dans le métier de traducteur professionnel, les traducteurs passent une heure à traduire une page standard (de 1400 signes, ou 250 mots en français ou en anglais), soit environ 4 mots/minute. Pour arriver à une qualité professionnelle, il faut encore 20 mn de révision (par un traducteur senior).

I.3.2.1.2 Evaluation de la qualité d'un système de TA spécialisé à partir d'un corpus post-édité

Un système de TA spécialisé construit à partir d'un corpus post-édité est évalué de deux façons.

1. **Évaluation avec références.** Dans cette évaluation, on choisit un corpus de test pour évaluer à l'aide des mesures d'évaluation objectives telles que le score BLEU, NIST et TER : plus les scores BLEU ou NIST seront élevés et plus le score TER sera faible, mieux sera évaluée la traduction.
2. **Évaluation sans références.** Cette évaluation permet d'estimer la qualité de la TA en fonction du temps de post-édition, lui-même estimable à partir de la distance d'édition

[22] https://spel.co.za/product/african_spelling_checkers/

[23] https://www.webspellchecker.net/samples/additional-languages-demo.html#lang-code=tn_ZA

mixte. Un travail antérieur [Wang Haozhou, 2015] a permis d'estimer que, pour le couple français-chinois, une unité de distance mixte de post-édition correspondait à 2 secondes de temps total de post-édition. Nous évaluerons donc ce facteur dans notre cas.

I.3.2.2 Amélioration du temps de post-édition de la TA entre différents types de systèmes de TA

I.3.2.2.1 Détermination et choix d'un système de TA spécialisé minimisant le temps de post-édition des prétraductions français-somali

Comme indiqué dans la partie I.3.2.1, le meilleur système de TA spécialisé construit avec un corpus bilingue post-édité sera celui qui nécessitera le moins de temps de post-édition des prétraductions d'un article ou d'un segment français vers le somali.

Ainsi, pour pouvoir déterminer et choisir le meilleur système de TA, nous allons expérimenter plusieurs types de système de TA avec les mêmes données : TA statistique avec MOSES, TA statistique avec adaptation des nouvelles données parallèles avec MOSES, et enfin un système de TA neuronale français-somali avec OPENNMT.

Une évaluation comparative en termes du temps de post-édition par page standard à partir d'un certain nombre de phrases ou par des articles sera effectuée sur les résultats de TA de ces différents systèmes. À la fin, le meilleur système de TA français-somali se démarquera parmi tous ces systèmes par sa capacité à réduire le temps de post-édition, plutôt que par les scores BLEU, NIST et TER.

I.3.2.2.2 Détermination de la corrélation entre distance d'édition mixte et temps de post-édition par page standard

Pour évaluer la qualité d'un segment post-édité, on calcule la distance d'édition qui selon (Levenshtein 1966 ; Wagner & Fischer 1974) est le coût minimal des opérations telles que l'insertion, la suppression et la substitution, nécessaires pour transformer un segment pré-traduit en un segment post-édité.

Depuis 2004, au laboratoire d'informatique de Grenoble, on utilise aussi la distance d'édition mixte [Pineau & Boitet 2004], qui combine la distance d'édition basée sur les mots, D_{mot} et la distance d'édition basée sur les caractères, D_{car}.

Par ailleurs, des travaux récents [Wang, 2015] menés au sein du laboratoire LIG-GETALP ont mis en évidence le lien ou la corrélation qui peut exister entre la distance d'édition mixte d'un segment ou d'un document post-édité et son temps de post-édition, sur le couple de langue français-chinois : une unité de cette distance mixte correspondait dans cette expérience à 2 secondes de post-édition.

Plus concrètement, le coût d'un échange entre 2 éléments e et f sera noté X (e, f) (e→f), le coût d'une insertion d'un élément f sera noté I(f) = X (ε, f), et le coût d'une suppression d'un élément e sera noté S(e) = X (e, ε).

Au niveau des caractères, si A = $a_1 \ldots a_m$, et B = $b_1 \ldots b_n$, alors :

$D_{car.}$ (A, B) = min Coût (A→B), avec ($\forall$ a, b 2 caractères) [X (a, b) = S(a) = I(b) = 1].

Au niveau des mots, si A = $u_1 \ldots u_m$ et B = $v_1 \ldots v_n$ (u_i et v_j sont des mots), alors :

$D_{mot.}$ (A, B) = min Coût (A→B), avec ($\forall$ u, v 2 mots) [X (u, v) = $D_{car.}$ (u, v)].

Pour $\alpha \in [0, 1]$, on a alors : $D_{mixte,\alpha}$ (A, B) = α $D_{car.}$ (A, B) + (1- α) $D_{mot.}$ (A, B).

Dans l'étude de [Wang Haozhou, 2015], la meilleure valeur de α s'est trouvée être α= 0,3.

Il s'agit dans le cadre de notre thèse de déterminer le cas échéant si cette corrélation existe aussi dans le cadre de la TA d'un corpus spécialisé post-édité français-somali.

Conclusion du chapitre I

Dans ce premier chapitre, nous avons exploré la problématique de l'informatisation d'une langue peu dotée, qui a déjà été traitée par [Berment, 2004] pour les langues d'Asie du Sud-Est.

À travers la langue somalie de Djibouti, qui appartient à l'espace francophone, nous avons également présenté les prémisses d'une informatisation par étape des langues d'Afrique francophone, tout en nous focalisant sur plusieurs besoins critiques.

Dans le cas précis du somali de Djibouti, il s'agit de l'accès en langue somali à des informations publiées par le journal *La Nation de Djibouti*, en utilisant un système de TA empirique construit puis incrémentalement amélioré avec des données issues de la post-édition.

Dans le prochain chapitre, nous présenterons de façon assez détaillée l'état d'informatisation du somali.

Chapitre II État de l'art de l'informatisation du somali

Introduction du chapitre II

À l'instar des autres langues africaines de l'espace francophone, la langue somalie dispose de peu de ressources linguistiques, d'où son statut de langue faiblement dotée.

Cependant, depuis quelques décennies, un certain nombre de ressources, d'outils de base et d'applications TAL a été construit ou réalisé pour cette langue. Nous en avons utilisé certains dans le cadre de nos travaux de thèse. Ces ressources et outils se répartissent entre des ressources dictionnairiques numérisées ou au format papier, des corpus linguistiques monolingues ou bilingues, des grammaires descriptives, et des outils de TALN de base tels que des segmenteurs, des racineurs et enfin des applications ou services conçus dans le cadre de projets de TALN ou de localisation pour le somali.

Le but de ce chapitre est de présenter premièrement l'ensemble des ressources, outils de base et applications pour informatiser une langue, puis de les recenser et de faire un état récapitulatif sur le somali.

II.1 Classification des ressources, outils de base et applications (ou services) pour une langue à informatiser

II.1.1 Ressources

Il s'agit des ressources linguistiques « statiques », qui sont des données exploitables par des humains, des programmes, ou les deux : dictionnaires, corpus, et grammaires descriptives (par opposition aux grammaires dynamiques comme les grammaires transformationnelles ou les grammaires procédurales).

II.1.1.1 Dictionnaires et bases lexicales

Un dictionnaire est un ouvrage de référence qui a pour but de recueillir l'ensemble des mots d'une langue d'une manière générale ou dans un domaine précis. Il est généralement présenté par ordre alphabétique et doit contenir la définition, les synonymes, la traduction ou d'autres informations comme la catégorie grammaticale, l'étymologie etc.

Une base lexicale ou terminologique est une sorte de dictionnaire abordant le lexique ou la terminologie, c'est à dire le langage spécifique d'une ou plusieurs spécialités, généralement disponible sur Internet.

Ainsi, il existe plusieurs types de dictionnaires en fonction de leur nature : un ou des dictionnaires d'usage de la langue ou bien un dictionnaire terminologique. La plupart des dictionnaires sont disponibles seulement en format papier. Toutefois, ces dernières décennies, on assiste au développement de dictionnaires informatisés, comme le TLFI (Trésor de la Langue Française Informatisé[24]). À côté des dictionnaires d'usage, numérisés ou pas, il existe des dictionnaires terminologiques qui ciblent un domaine ou une terminologie précise d'une ou plusieurs langues, mais, pour les langues-π, ils sont rares et de petite taille.

[24] *TLFi : Trésor de la langue Française informatisé*, http://www.atilf.fr/tlfi, ATILF - CNRS & Université de Lorraine.

Par contre, pour les pays de l'Union Européenne, il existe une grande base terminologique multilingue, IATE (Inter Active Terminology for Europe). Cette base terminologique renferme 8,4 millions de termes (pour plus de 800 000 concepts) et contient des termes pour les 24 langues officielles de l'Union Européenne. Cependant, sa couverture est assez faible pour les langues baltes, les langues slaves, le hongrois, le finnois et le grec.

Mais IATE est loin de couvrir les 24 langues pour chaque concept. Il y a environ 800 000 concepts, ce qui donnerait 19,2 M de termes, si la couverture était totale. Or il n'y en a que 8,4 M, un peu moins de 44 %.

C'est une base terminologique ouverte et en évolution puisque chaque traducteur professionnel d'une institution européenne peut mettre à jour et modifier son contenu après que ses contributions ont été validées et passées par une étape de vérification et de validation automatique.

Parmi les autres bases lexicales bilingues ou multilingues existant dans ce domaine, on retrouve celle développé dans le cadre du projet Papillon [Tomokiyo, M. et al., 2000]. Ce projet qui a débuté au mois de juillet de l'année 2000, avait pour objectif de réaliser un environnement de développement coopératif à travers internet et la création d'une grande base lexicale multilingue et plus spécialement vers le français, le japonais et l'anglais. À la suite de ce projet, une nouvelle plateforme Jibiki [Mangeot, M. et al., 2003 ; Sérasset, G., 2004] ayant pour but de faciliter la création de bases de données lexicales de type variés, a été créée.

La plateforme JIBIKI permet de traiter toutes les ressources lexicales au format XML et contenant des microstructures et macrostructures différentes. En outre, elle offre de nombreuses fonctionnalités prêtes à l'emploi, telles que l'import d'un dictionnaire ou d'un volume, la création et l'édition de nouvelles entrées, et enfin la gestion des contributions des utilisateurs par l'administrateur et la recherche dans les bases lexicales.

Dans le cadre du projet ANR Scientext, [Falaise, A. et al., 2011] ont développé un outil appelé Scientext qui permet à des non-informaticiens d'exploiter des corpus linguistiques annotés syntaxiquement. Pour utiliser ce logiciel, il faut passer par les trois étapes suivantes : choix et sélection d'un sous-corpus, recherche des phénomènes linguistiques et enfin affichage des réponses selon la méthode KWIC (*Key Word In Context*). L'utilisation de Scientext a dépassé actuellement le seul cadre de la recherche académique ou scientifique, puisqu'il est également utilisé comme un outil de didactique du FLE.

II.1.1.2 Corpus

Comme nous l'avons évoqué au I.3.1.1, les corpus linguistiques utilisés en TAL se distinguent par leur type (monolingue, bilingue), leur caractéristique (bruts, annotés) ou leur nature (écrits ou oraux).

Parmi les corpus linguistiques les plus aboutis et utilisés souvent dans les applications du TAL, il y a ceux distribués par le *Linguistic Data Consortium* (LDC) qui est un consortium ouvert, composé d'universités, de sociétés et de laboratoires de recherche gouvernementaux. Il est hébergé au sein de l'université américaine de Pennsylvanie.

Le but de ce consortium est de créer, collecter et distribuer des bases de données textuelles, des lexiques et d'autres ressources linguistiques à des fins de recherche et de développement. Le programme LDC a été initié en 1992 dans le cadre d'un projet de la DARPA (Agence américaine de recherche avancée) et est partiellement financé par la fondation nationale de la Science (NSF) des USA. L'ensemble des ressources développées et conçues par le LDC est

accessible à travers une plateforme web sous la forme d'un catalogue dédié[25]. Ce catalogue contient des ressources linguistiques textuelles ou audio dans plusieurs dizaines de langues.

Au niveau européen, c'est l'ELRA/ELD [26], l'association européenne des ressources linguistiques fondée en 1991, dont le quartier général est à PARIS, qui est chargée de promouvoir la collecte de ressources linguistiques ainsi que leur évaluation. Pour la dissémination et la vulgarisation des atouts des ressources linguistiques dans le développement du TAL et des TIC, elle organise une conférence internationale sur les ressources linguistiques et l'évaluation (LREC) tous les 2 ans.

Dans une optique plus globale ou universelle, le projet OPUS [Jörg Tiedemann, 2012] avait pour but de collecter des textes gratuits à partir du web en vue de créer des corpus parallèles librement accessibles pour le maximum de langues possibles. À ce jour, la plateforme OPUS[27] contient des corpus bilingues concernant plus de 300 langues ou variétés de langues différentes parlées et écrites dans les 5 continents de la planète. À notre connaissance, c'est la plateforme qui a atteint le plus de langues et qui contient des corpus monolingues et parallèles pour amorcer des applications ou travaux TAL tels que la traduction automatique ou les outils de bases tels que les correcteurs orthographiques ou grammaticaux.

Le corpus JRC-Acquis [Steinberger et al. 2006] est l'un des premiers corpus parallèles alignés au niveau des segments avec des prétraitements réalisés en 2006 et distribués par la commission européenne. Il contient un grand nombre de documents légaux et administratifs écrits dans les 24 langues officielles de l'Union Européenne. Les ressources multilingues collectées dans le corpus JRC-Acquis, avec 1 milliard de mots, constitue l'un des meilleurs corpus parallèles alignés pour faire des applications TAL telles que la traduction automatique (statistique ou neuronale), la désambiguïsation sémantique multilingue, ou la production de ressources lexicales ou sémantiques multilingues telles que des dictionnaires ou des ontologies.

II.1.1.3 Grammaires descriptives

Pour l'apprentissage et l'écriture d'une langue écrite et parlée, il existe deux types de grammaires : la grammaire normative (prescriptive) et la grammaire descriptive. La première est celle enseignée dans les écoles en vue d'apprendre aux jeunes élèves les normes et les règles à respecter pour parler et écrire correctement une langue vivante. Le but de la grammaire prescriptive est d'étudier les règles qui régissent une langue donnée, afin de produire des énoncés ou phrases corrects et reconnus par les locuteurs natifs de cette langue.

Par contraste, une grammaire descriptive décrit comment la langue est, telle qu'elle est parlée par ses locuteurs, sans y apporter un quelconque jugement.

Par ailleurs, il existe des grammaires formelles pour certaines langues, comme les grammaires hors-contexte et leurs dérivées (grammaires attribuées, GPSG, HPSG, LFG, TAG) ou les types de grammaire de dépendances, qui se basent sur les théories des langages formels, et associent des structures aux énonces d'une langue.

À l'instar des langues bien dotées des pays occidentaux, il existe çà et là des ouvrages de référence contenant des grammaires prescriptives et descriptives pour les langues de l'Afrique francophone, et plus précisément pour la langue somalie ([Saeed 1999], [Cabdalla, 1999]).

[25] https://catalog.ldc.upenn.edu/

[26] http://www.elda.org/en/

[27] http://opus.nlpl.eu

II.1.2 Outils de base

II.1.2.1 Segmenteurs

Un segmenteur est un outil de base du TALN, qui permet de segmenter un texte sous forme de phrases, de mots, de syllabes ou de documents web en vue de procéder à un traitement ultérieur.

Le niveau de segmentation d'un texte dépend du type d'application qui sera utilisé par la suite, et de la structure des mots du texte à segmenter. Par exemple, pour certaines langues asiatiques comme le chinois ou le japonais, un ou plusieurs symboles textuels peuvent être considérés comme un item lexical. Pour les langues d'Asie du Sud-Est comme le khmer, le lao ou le thaï, les items lexicaux de la langue sont segmentés au niveau des syllabes comme l'outil SYLLA de [Berment, 2004].

Un segmenteur peut servir également à segmenter du contenu textuel dans une page, et dans ce cas son rôle sera de faire une segmentation par document. [R. Kalitvianski, 2013] a récemment développé en Java un outil appelé SEGNORM pour segmenter des documents XML avant leurs traitements ultérieurs. Les documents sont segmentés en blocs de texte en fonction d'une liste de balises dépendant du format utilisé (HTML ou ODT) puis en phrases à l'aide de règles au format SRX (Segmentation Rules eXchange) dépendantes de la langue.

Actuellement l'outil SegNorm segmente les documents en français et en anglais (cf. Figure 6). Parmi les autres outils de base entrant dans le domaine de la segmentation, il y a les outils de césure ou d'hyphénation qui permettent de déterminer l'endroit où les mots d'une langue seront coupés en deux. Le plus souvent, l'hyphénation a lieu lorsqu'il y a un tiret entre deux mots qui sont une même unité lexicale ou qui sont séparés par une ligne, car la coupure est obligatoire pour embellir le texte. Tous les logiciels de traitement de texte disposent de l'option d'hyphénation dans leurs fonctionnalités standard pour les langues bien dotées.

Figure 6 : Exemple de page segmentée par SegNorm[28]

[28] Source : https://pro.aiakide.net/?what=demos&lang=en&section=segdoc

II.1.2.2 Racineurs

Un racineur est un outil basique du traitement automatique d'une langue. Son but est de retrouver la racine d'un item lexical (mot-forme en général) et cela à travers l'identification et la suppression des différents préfixes, suffixes qui peuvent s'ajouter à un radical du lemme (mot abstrait ou forme de citation) pour produire la forme en question.

Une racine est différente d'un lemme, qui est par définition la « forme de citation » d'un lexème dans un dictionnaire, par exemple « aller » pour la forme « irons ».

Par ailleurs le terme *radical* est à différencier également du terme *racine,* puisqu'il désigne la plus petite unité lexicale permettant de former des mots apparentés qui ont la même origine et étymologie.

Il existe plusieurs méthodes et algorithmes pour construire un racineur pour une langue. Les plus utilisés et connus sont les algorithmes de Porter [Porter, 1980], Lovins [Lovins, 1968] et Paice [Paice & Chris D., 1990].

Selon une étude comparative effectuée par [Anjali Ganesh Jivani & al., 2011], les différents algorithmes de racinisation peuvent être classés en trois catégories : les algorithmes basés sur la troncation, les algorithmes fondés sur des méthodes statistiques, et les algorithmes utilisant des méthodes mixtes.

Les algorithmes à troncation procèdent à la suppression successive des suffixes et des préfixes d'un mot, jusqu'à trouver le radical du mot. Par exemple, pour le mot « inversion », le radical est « vers » et la racine « vrt » (comme pour vortex et vertèbre).

Dans cette catégorie, on trouve entre autres ceux de Julie Beth Lovins[29], Porter [Porter, 1980], Paice et Husk [Paice &Chris D., 1990] et enfin l'algorithme de Dawson qui n'est autre qu'une extension de l'algorithme de Porter : il couvre environ 1200 affixes pour l'anglais.

Dans la catégorie des algorithmes dits statistiques, on trouve les racineurs à base de n-grammes, les racineurs à base de Modèles de Markov Cachés (HMM) [Melucci & Orio, 2003] et les racineurs YASS (*Yet Another Suffix Striper)* [Prasenjit Majmunder & al., 2007].

Enfin, dans la catégorie d'algorithmes dits mixtes ou à base de corpus, il y a, pour ne citer que les plus connus, l'algorithme de Krovetz [Krovetz Robert, 1993] et celui de Xu et Croft [Xu & Croft, 1998].

Pour réaliser notre racineur pour le somali, nous avons utilisé un algorithme à troncation avec la méthode de Porter.

II.1.2.3 Supports aux dictionnaires et bases lexicales

Les dictionnaires et bases lexicales utilisés souvent en TALN sont accompagnés d'autres outils qui peuvent servir de support en vue d'enrichir leur lexique ou construire des dictionnaires et bases lexicales multilingues. C'est ainsi que, dans le cadre du projet BabelNet, [Navigli, R. & Ponzetto, S. P., 2010] ont construit l'un des plus grands réseaux lexicaux multilingues, avec un nombre d'entrées estimé actuellement à plus de 13 millions, sur 284 langues, ces entrées étant reliées entre elles par une grande quantité de relations sémantiques.

[29] Lovins JB (1968) Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11: 22-31. C'est le tout premier article parlant de « racineur », sachant qu'il existait des analyseurs morphologiques depuis longtemps, en particulier pour le russe (projet GAT dès 1954, CETA dès 1962, etc.).

BabelNet a été créé en intégrant automatiquement les entrées lexicales multilingues de la plus grande encyclopédie en ligne qu'est Wikipédia, avec le lexique de la langue anglaise de WordNet. Pour établir la correspondance avec des mots manquants dans d'autres langues, la traduction automatique a été utilisée.

II.1.2.4 Supports aux corpus monolingues et multilingues parallèles

Il existe çà et là des outils et applications qui sont destinés à servir de support aux corpus monolingues, bilingues ou parallèles. Le système SECTra_w (Système d'Exploitation de Corpus de Traductions) développé dans le cadre de la thèse de [C. Phap Huynh, 2010] est le plus connu, car il a été utilisé pour construire de nouveaux corpus bilingues, pour améliorer la traduction automatique et pour évaluer la qualité de la TA dans le cadre de plusieurs campagnes d'évaluation au sein du groupe de recherche GETALP du laboratoire d'Informatique de Grenoble.

II.1.3 Applications et services

Les outils ou les applications et services issus du traitement automatique du langage naturel se répartissent entre ceux qui peuvent être utilisés par des professionnels non-développeurs ou le grand public, et ceux destinés à des contextes et des utilisateurs particuliers.

II.1.3.1 Outils utilisés par le grand public ou des professionnels non-développeurs

II.1.3.1.1 Correcteurs typographiques

Pour faciliter la lecture et l'esthétique d'impression d'un texte écrit et composé sur un ordinateur, il convient de respecter certaines règles typographiques concernant la taille des caractères, le type de police de caractère utilisé, l'espacement entre les mots et paragraphes etc.

Le rôle d'un correcteur typographique est de veiller au respect des règles typographiques de chacune des langues écrites dans un alphabet donné. C'est un outil connu de tous, connectable ou intégrable à toutes les applications comme des services.

La majorité des correcteurs typographiques actuels est dédiée aux langues bien dotées des pays développés telles que le français, l'anglais, l'espagnol, le japonais et le chinois etc.

Par exemple, pour le français, il existe des règles typographiques concernant l'insertion en français d'un quadratin (petit blanc insécable) avant les signes de ponctuations double (: ; ! ?) et entre les guillemets doubles (« »), ou encore l'obligation d'une majuscule en début de chaque phrase, et une aide à l'entrée ou à la saisie de caractères spéciaux, comme ™ © π μ → ↔ etc.

Cependant, il existe peu ou pas d'outil de ce type pour les langues africaines de l'espace francophone.

II.1.3.1.2 Correcteurs orthographiques et grammaticaux

a. ***Correcteurs orthographiques***

Les premiers travaux sur les correcteurs orthographiques datent des années 60 [Kukich K., 1992]. L'objectif d'un correcteur orthographique est de détecter les erreurs orthographiques dans un texte électronique, de sélectionner et de proposer les solutions possibles à ces erreurs, de classer les solutions pour l'utilisateur par ordre de préférence, et enfin de corriger les erreurs du texte en fonction du choix de l'utilisateur.

Un correcteur orthographique est un outil de base TALN, indispensable à l'informatisation d'une langue peu dotée puisqu'il peut servir à faciliter l'utilisation de l'outil informatique et l'alphabétisation des locuteurs natifs d'une langue peu dotée.

Les correcteurs orthographiques sont devenus aujourd'hui des outils utilisés par le grand public, car tous les éditeurs de logiciels de traitement de texte ou de bureautique (Microsoft, Open Office) ont intégré dans leurs solutions bureautiques des correcteurs orthographiques couvrant pratiquement toutes les langues bien dotées, et très utilisées par leurs utilisateurs.

b. ***Correcteurs grammaticaux***

Un correcteur grammatical est un outil de base spécialisé du TALN. Contrairement à un correcteur orthographique, il est chargé de vérifier la conformité des mots d'un texte aux règles de grammaire (accords, ordre des mots, utilisation des bonnes prépositions, des cas, des modes, etc.) et aux règles de la sémantique (sens de la phrase correcte, erreurs sur la confusion de mots homophones, etc.) d'une langue donnée.

La majorité des éditeurs de logiciels de traitement de texte ont inclus des correcteurs grammaticaux dans leur solution bureautique. Ces derniers temps, leur utilisation s'est généralisée dans les outils de messagerie électronique, dans les forums de discussions, et dans les navigateurs web. La correction grammaticale s'effectue au fur et à mesure que l'utilisateur saisit son texte, et des corrections lui sont proposées dans un menu contextuel ou bien sont disponibles d'un seul coup à sa demande.

Pour la langue française les principaux correcteurs grammaticaux propriétaires sont ANTIDOTE, PROLEXIS, CORDIAL, LE ROBERT CORRECTEUR, et LANGUAGETOOL, un logiciel libre de correction grammaticale.

Malgré sa vulgarisation et son utilisation standardisée pour les langues bien dotées, à notre connaissance il n'existe aucun correcteur grammatical pour le somali.

II.1.3.1.3 Correcteurs terminologiques

La terminologie est définie selon la norme [ISO 1087-1] comme étant « l'étude scientifique des notions et des termes en usage dans les langues de spécialité ». Plus généralement, un dictionnaire terminologique contient la liste des termes recommandés officiellement dans une langue donnée par rapport aux termes équivalents dans une langue étrangère.

Ainsi, le rôle d'un correcteur terminologique est de proposer à ses utilisateurs le remplacement des termes étrangers par les termes officiels dans la langue d'écriture dans un document administratif, technique ou scientifique pour une diffusion plus large de la terminologie de la langue. Le correcteur utilise pour cela une base de données terminologiques mise à jour régulièrement. C'est un outil qui peut être utilisé d'une manière autonome à partir d'une application web, ou intégré dans les principaux logiciels de traitement de texte.

La suite bureautique LIBRE OFFICE inclut un correcteur terminologique français développé dans le cadre d'un projet interministériel entre le ministère de la Culture et du Budget français.

Les termes étrangers et leurs correspondants en français sont extraits de la base de données FranceTerme[30].

[30] www.culture.fr/franceterme

Malheureusement, à notre connaissance, il n'existe pas de correcteur terminologique informatisé pour les langues africaines peu dotées et encore moins pour le somali de Djibouti[31].

II.1.3.2 Applications destinées à des contextes et des utilisateurs particuliers

Parmi les applications concrètes utilisées dans des contextes particuliers, tels que la veille, le renseignement, la compréhension d'un texte écrit en langue étrangère, sa diffusion ou sa dissémination, il y a la traduction des documents multilingues écrits ou oraux.

II.1.3.2.1 Reconnaissance et synthèse de la parole

a. ***Reconnaissance automatique de la parole***

La reconnaissance automatique de la parole est une technique issue du TALN et de l'intelligence artificielle (IA) qui a pour but d'analyser la voix humaine capturée à l'aide d'un microphone pour réaliser sa transcription sous forme de texte dans une langue donnée afin de réaliser des applications de type interface homme-machine.

Les premiers travaux de recherche sur la reconnaissance automatique de la parole ou RAP datent de 1952. Ils ont été réalisés au sein du laboratoire Bell Labs et avaient pour objet de reconnaître des chiffres dans un dispositif électronique câblé [H. Davis et al., 1952].

Par la suite, plusieurs autres laboratoires privés ou universitaires ont poursuivi la recherche en RAP, ce qui a permis d'améliorer les performances des systèmes de reconnaissance vocale. Avec la croissance rapide des capacités de traitement des ordinateurs et l'arrivée des téléphones intelligents (smartphones), la reconnaissance automatique de la parole est devenue de nos jours un moyen de faire communiquer les machines avec les humains pour exécuter certaines tâches telles que la réservation de billets d'avion ou de train, la mise en place de serveurs vocaux interactifs ou l'aide et l'assistance aux personnes handicapées ou âgées, etc.

b. ***Synthèse de la parole***

La synthèse de la parole est une technique informatique faisant partie du TALN et de l'IA et qui consiste à créer artificiellement des sons compréhensibles par les humains à partir d'un texte écrit dans une langue donnée. Pour créer une parole artificielle, la synthèse de la parole utilise des techniques issues d'une part de la linguistique pour convertir les mots d'un texte en phonèmes prononçables, et d'autre part le traitement numérique du signal pour transformer cette version phonétique en un son numérique écoutable sur un haut-parleur. Ce faisant, la synthèse vocale effectue l'opération inverse de la reconnaissance vocale ou RAP.

Il existe de nos jours plusieurs applications technologiques qui utilisent la synthèse vocale comme la vocalisation des écrans d'ordinateurs pour les personnes malvoyantes ou aveugles, ou bien l'utilisation des serveurs vocaux interactifs pour la vocalisation des annuaires téléphoniques ou d'adresses, etc.

II.1.3.2.2 Traduction de l'écrit et de l'oral

a. ***Traduction automatique***

La traduction automatique est une branche de la linguistique computationnelle qui est apparue au milieu du XXème siècle. Le but de la traduction automatique (TA) est d'utiliser des

[31] Par « somali de Djibouti », nous entendons la variante du somali dite Somali du nord qui est parlée en Somaliland, Puntland, la région Somali de l'Ethiopie et enfin à Djibouti.

programmes informatiques afin de traduire un texte ou une parole dans une langue source vers une langue cible sans l'intervention ou l'assistance d'un humain.

Elle se distingue de la traduction assistée par ordinateur (TAO) puisque dans celle-ci une partie de la tâche de traduction est réalisée manuellement par un traducteur humain.

b. ***Aide à la traduction humaine***

Il existe de nos jours une multitude d'outils informatiques qui sont conçus pour aider le traducteur professionnel ou amateur dans sa tâche de traduction. Ces logiciels, qui sont entre autres les dictionnaires informatisés, les glossaires spécialisés, les bases de données terminologiques, les concordanciers ou encore les mémoires de traductions et les systèmes de traduction automatique, sont généralement mentionnés dans la littérature comme faisant partie des outils de traduction assistée par ordinateur (TAO). En effet, cette dernière est définie comme étant tout outil informatique mis à la disposition des professionnels de la traduction, qu'ils soient gratuits ou payants, en vue de faciliter leur travail.

Par ailleurs, dans les différents cursus de formation aux métiers de la traduction, la maîtrise des outils de TAO fait partie des compétences que doit posséder tout traducteur professionnel.

II.1.3.2.3 Traitement de l'information textuelle

a. ***Recherche d'information***

La recherche d'information est un domaine de recherche qui se situe entre la bibliothéconomie et les sciences de l'information. Elle utilise des techniques informatiques, issues du TALN pour rechercher automatiquement des informations dans un corpus. Ce dernier est constitué d'un ensemble de documents de types différents (texte, audio et vidéo) portant sur un ou plusieurs domaines, et sont stockés dans une base de données qui peut être relationnelle, structurée ou non structurée.

b. ***Résumé automatique***

Le résumé automatique d'un texte est un domaine du TALN qui permet de produire une version condensée d'un texte écrit en utilisant des techniques informatiques. Cette version résumée du texte original doit être une représentation abrégée et exacte du contenu textuel du document.

Dans la littérature, on recense plusieurs techniques pour réaliser un résumé automatique, mais trois approches sont les plus utilisées de nos jours dans ce domaine.

La première, dite approche par abstraction, a pour but de générer un résumé composé de phrases qui ne se trouvent pas forcément dans le texte original.

La seconde approche, dite résumé par extraction, consiste à extraire des phrases complètes considérées comme étant les plus pertinentes du texte, et qui sont rassemblées par la suite pour produire le résumé du texte.

La dernière approche, dite résumé par compression de phrases, génère le résumé du texte à l'aide de phrases extraites qui sont par la suite compressées entre elles afin d'en faire le résumé du contenu textuel.

C'est la seconde approche qui est de loin la plus utilisée actuellement dans les logiciels de résumé automatique.

c. ***Indexation***

A l'instar de la recherche d'information, l'indexation automatique des documents est un domaine de l'informatique qui utilise des méthodes issues du TALN et des sciences de

l'information et des bibliothèques ou bibliothéconomie pour organiser un ensemble de documents de natures différents (texte, audio, vidéo etc.) afin de faciliter la recherche de contenus dans cette collection. Pour ce faire, l'indexation automatique utilise un index qui est une liste de descripteurs, auquel est associée une partie ou une liste de documents de la collection que chaque descripteur doit renvoyer.

L'intérêt d'indexer automatiquement une collection de documents est motivé par le besoin de pouvoir retrouver facilement plus tard les documents ou une partie de leurs contenus textuels, qui pourront intéresser un utilisateur, sans qu'il doive parcourir toute la collection.

Il existe plusieurs méthodes pour indexer le contenu d'une collection en fonction de la nature des documents à indexer. Les méthodes scientifiques actuellement en vigueur pour effectuer l'indexation automatique sont l'extraction des caractéristiques de chaque document, le clustering ou le partitionnement des données de la collection, la quantification et enfin la recherche d'information.

II.2 Le cas du somali

II.2.1 Ressources pour le somali

II.2.1.1 Dictionnaires bilingues

II.2.1.1.1 Dictionnaire français-somali & somali-français

Le premier dictionnaire français-somali a été réalisé par le linguiste et enseignant francophone somalien Abdulghani Gourré Farah aux éditions l'Harmattan en 2008 [Farah, A.G., 2008].

Ce premier dictionnaire bilingue a pour but de présenter aux lecteurs somalophones, désirant comprendre et pratiquer le français dans la vie courante, un vocabulaire essentiel et portant sur tous les aspects de la vie en français. Ce dictionnaire est composé de 10 000 entrées, avec de nombreux exemples d'usage pour chacune, une définition claire et précise de chaque mot français en somali, une description sur son information grammaticale, et une transcription phonétique du mot français basée sur l'alphabet somali afin de faciliter sa prononciation.

Ce dictionnaire contient également des exemples de conjugaison des verbes réguliers et irréguliers. Ce même auteur a conçu un autre dictionnaire, somali-français de 9 000 mots, destiné aux somalophones qui souhaitent pratiquer et comprendre la langue française dans leur vie quotidienne.

II.2.1.1.2 Dictionnaire anglais-somali & somali-anglais

Il existe deux dictionnaires (anglais-somali & somali-anglais) réalisés par Nicholas Awde [C. Quadir & N. Awde, 1999] en 1999. Comme les dictionnaires français-somali, ces derniers sont également destinés aux lecteurs anglophones qui souhaitent comprendre et pratiquer l'anglais dans la vie courante.

Ce dictionnaire est composé de 9 000 mots et contient également pour chaque mot une description succincte de sa catégorie grammaticale, sa prononciation et des exemples.

II.2.1.1.3 Dictionnaire trilingue somali-anglais-italien

La fondation culturelle redse-online.com a développé un dictionnaire trilingue entre le somali, l'anglais et l'italien. Comme indiqué dans la page d'accueil du projet[32], ce dictionnaire est

[32] http://www.redsea-online.com/modules.php?name=dictionary

librement accessible et utilisable et ses entrées sont en constante augmentation puisque les utilisateurs peuvent eux-mêmes ajouter de nouvelles entrées après validation et révision du comité du dictionnaire.

À ce jour, ce dictionnaire trilingue contient 12 373 entrées de base et 34 282 nouvelles entrées qui ont été récemment soumises par les visiteurs/contributeurs du site web et sont en cours de révision et validation par le comité du dictionnaire.

II.2.1.2 Corpus

II.2.1.2.1 Corpus monolingue somali à partir du Web

Le corpus monolingue du somali que nous avons utilisé pour nos différentes expérimentations est un corpus issu du web et qui a été construit par [Scannell K., 2007] dans le cadre du projet Crúbadán destiné à construire des ressources linguistiques pour les langues peu dotées.

II.2.1.2.2 Corpus bilingue somali-anglais

Le corpus bilingue somali-anglais que nous avons utilisé pour notre première expérience de TA somali-anglais est un corpus que nous avons téléchargé depuis la plate-forme OPUS [Jörg Tiedemann, 2012]. Il est composé de trois sous-corpus issus de domaines différents.

Le tableau ci-dessous résume la description de ce corpus.

Nom corpus	Type de données	Taille en segments (milliers)	Taille en items source (milliers)	Taille en items cible (milliers)
TANZIL	Religieux	93,9	2.800	1.900
GNOME	Informatique	0,8	5,1	3,1
UBUNTU	Informatique	28	2	0,2
	Totaux	94,6	2.807,1	1.903.3

Tableau 4 : Caractéristiques du corpus OPUS (anglais-somali)

II.2.1.2.3 Corpus bilingue français-somali

À l'instar des données parallèles de la paire anglais-somali, que nous constituées à partir de la plateforme OPUS, il n'existait pas de corpus bilingue français somali et plus particulièrement dans le domaine des nouvelles journalistiques. Outre les données bilingues extraites à partir du web (environ 7 000 segments), il faut ajouter les 10 669 segments parallèles que nous avons construits par post-édition de la TA de Google et 3 457 segments français-somali constitués à partir du corpus TED. Le tableau ci-dessous récapitule le corpus bilingue français-somali.

Nom corpus	Type de données	Taille en segments (milliers)	Taille en items source (milliers)	Taille en items cible (milliers)
LaNationDJ_fr-so	Journalistique	10,67	100	99
TED_fr_so	Autre	3,46	36	32
TANZIL	Religieux	6,3	2800	1.900
GNOME	Informatique	0,8	5,1	3,1
UBUNTU	Informatique	28	2	0,2
	Totaux	21,3	2 943,1	2 035.3

Tableau 5 : Caractéristiques du corpus bilingue français-somali

II.2.2 Outils de base pour le somali

II.2.2.1 Premier étiqueteur morphosyntaxique du somali

Un étiqueteur morphosyntaxique sert à attribuer automatiquement des catégories morphosyntaxiques ou parties du discours (part of speech) aux mots d'un corpus textuel brut d'une langue donnée.

Le somali étant dépourvu de cet outil, notre premier travail a consisté dès lors à trouver une approche rapide et à moindre coût pour effectuer ce traitement. Parmi les approches les plus utilisées dans ce domaine, on distingue l'approche à base de règles ou experte et l'approche probabiliste. Par ailleurs un étiqueteur morphosyntaxique fait partie des premiers outils à réaliser pour amorcer le traitement automatique d'une langue naturelle.

Toutefois, avant tout travail d'automatisation de la tâche consistant à attribuer une catégorie morphosyntaxique aux mots d'un texte, il faut définir un jeu d'étiquettes sur lequel se basera l'étiqueteur morphosyntaxique. C'est ainsi que nous avons dans un premier temps identifié et listé les différentes parties du discours de la langue somalie à partir d'une étude linguistique approfondie et grammaticale de cette langue.

Nous avons ainsi réussi à définir des jeux d'étiquettes (ou parties du discours) de 3 niveaux de granularité basés sur une classification grammaticale des mots somalis. Pour notre premier étiqueteur morphosyntaxique du somali, nous avons utilisé seulement un jeu d'étiquettes d'un seul niveau de granularité (cf. Tableau 6 ci-dessous).

`	Catégorie
N	Nom commun
V	Verbe
A	Adjectif
J	Adverbe
D	Déterminant
R	Pronom personnel
F	Marqueur de focus
H	Marqueur de type de phrase
C	Conjonction
O	Apposition verbale
I	Idéophones
N	Interjections
X	Autre
P	Ponctuation

Tableau 6 : Catégories grammaticales du somali (1er niveau)

La Figure 7 ci-dessous contient la phrase suivante : « ***Dowlada Jibuuti waxay bilowday qorshee ay ku kombuyutargaraynayso afafka wadanka*** » (Le gouvernement de Djibouti a commencé un plan pour informatiser les langues nationales*),* avec les catégories grammaticales de 1er niveau attribuées automatiquement par notre étiqueteur morphosyntaxique.

Dowlada	N
Jibuuti	N
waxay	F
bilowday	V
qorshee	N
ay	R
ku	O
kombuyutargaraynayso	V
afafka	N
wadanka	N
.	P

Figure 7 : Exemple d'étiquetage grammatical d'une phrase somalie

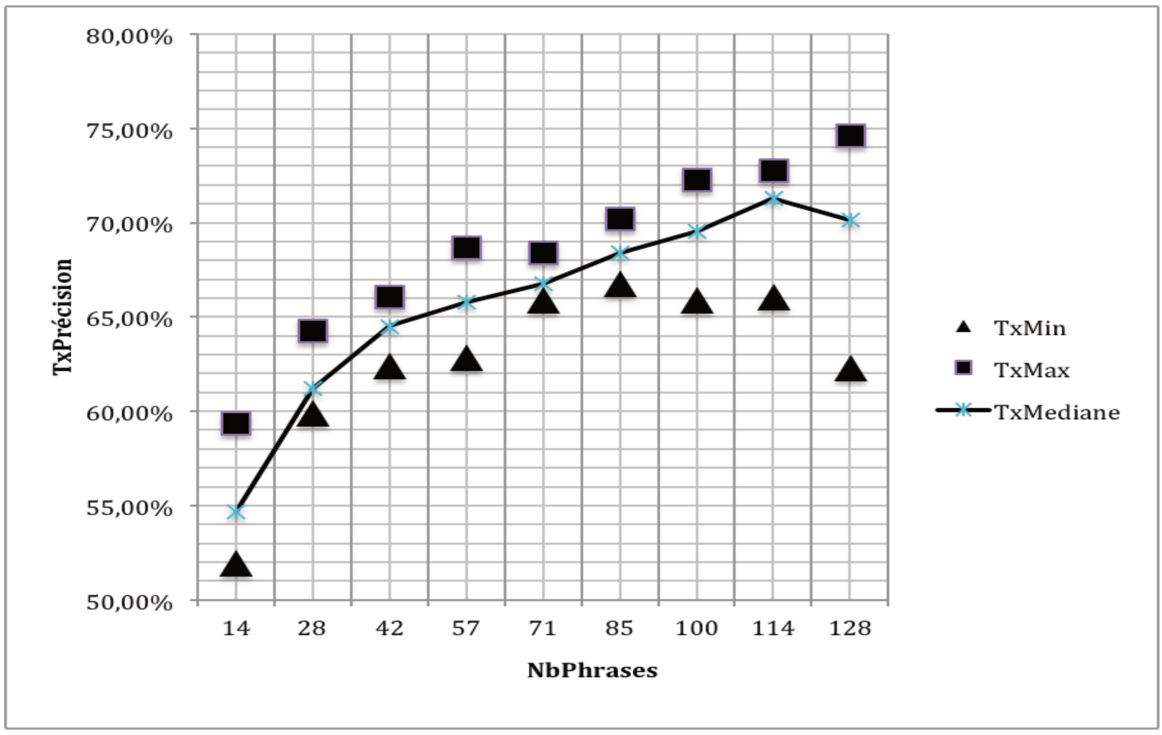

Figure 8 : Graphique des performances de l'étiqueteur

L'analyse du taux de précision (voir Figure 8) de notre étiqueteur en fonction de la quantité de données du corpus d'apprentissage utilisé pour l'entraînement indique que le taux maximum de précision (74,6%) est atteint au cours d'une des expériences de la phase 9 et cela avec le plus faible pourcentage de mots inconnus (23,83%). Dans cette configuration, le taux de dispersion des phrases du corpus (l'écart-type) représente 3,37% (le second écart-type de toute l'évaluation), ce qui indique un manque d'homogénéité des données utilisées durant cette phase.

Par ailleurs, pour que notre étiqueteur atteigne le meilleur taux de précision du somali avec n'importe quelle taille et type de données, il faudra augmenter la taille de notre corpus

d'apprentissage pour atteindre au moins deux fois sa taille actuelle, de telle sorte que le pourcentage des mots inconnus soit le plus faible possible.

La Figure 9 ci-dessous contient un état récapitulatif des différentes évaluations que nous avons effectuées sur les données d'apprentissage de notre étiqueteur.

Phase	Phrases	Tx-Moy	Ecart-Type	TxMin	TxMax	TxMed	IncMed	ConMed
1	14	**56,72%**	**4,15%**	**51,94%**	**59,36%**	**54,68%**	**52,29%**	**47,71%**
2	28	**63,10%**	**2,30%**	**59,90%**	**64,27%**	**61,22%**	**42,74%**	**57,26%**
3	42	**64,88%**	**0,55%**	**62,39%**	**66,05%**	**64,50%**	**36,45%**	**63,55%**
4	57	**65,58%**	**1,96%**	**62,85%**	**68,69%**	**65,78%**	**32,47%**	**67,53%**
5	71	**68,51%**	**0,16%**	**65,92%**	**68,40%**	**66,77%**	**30,93%**	**69,07%**
6	85	**68,96%**	**1,28%**	**66,75%**	**70,22%**	**68,38%**	**27,36%**	**72,64%**
7	100	**70,87%**	**0,83%**	**65,89%**	**72,27%**	**69,53%**	**26,07%**	**73,93%**
8	114	**71,94%**	**0,62%**	**66,07%**	**72,75%**	**71,28%**	**23,34%**	**76,66%**
9	128	**72,36%**	**3,37%**	**62,31%**	**74,60%**	**70,12%**	**23,83%**	**76,17%**

Figure 9 : Récapitulatif des évaluations de l'étiqueteur du somali

II.2.2.2 Un racineur du somali

Pour le besoin de notre premier étiqueteur morphosyntaxique, nous avons mis en place un module de racinisation basé sur l'algorithme de Porter [Porter, 1980].

L'intérêt de mettre en place un tel outil est motivé par les deux contraintes suivantes.

- D'une part, il est nécessaire de disposer d'un corpus étiqueté morpho-syntaxiquement avec le lemme de chaque mot afin de pouvoir utiliser les principaux étiqueteurs (*taggers*) existants, comme *Brill Tagger* ou *Tree Tagger*.
- D'autre part, la lemmatisation s'apparente aussi à une analyse morphologique des mots d'un corpus, dans la mesure où elle permet de constituer un lexique avec pour chaque morphème lexical ou grammatical ses différentes formes fléchies ou paradigmes. Cela pourrait plus tard servir pour la construction d'un lexique des formes fléchies du somali.

II.2.2.3 Un segmenteur du somali

Le segmenteur que nous avons développé consiste surtout à séparer les unités linguistiques ou mots d'un fragment ou d'une phrase. On l'utilise souvent durant l'étape de prétraitement dans le domaine du TALN : avant l'apprentissage du modèle de langage et de la table de traductions pour la TA, et avant l'attribution d'une catégorie grammaticale par un étiqueteur morphosyntaxique.

Généralement, les segmenteurs séparent les mots d'un segment ou d'une phrase à l'aide des différentes formes de ponctuation, à l'exception des mots composés ou liés qui doivent faire l'objet d'exceptions. Le segmenteur développé pour le somali dans le cadre de notre travail préliminaire à cette thèse [Assowe, 2011] a été surtout utilisé en TA et en étiquetage morphosyntaxique.

Le contenu d'un article journalistique, extrait à partir d'un site web journalistique somalien (www.puntlandpost.com), ainsi que sa segmentation à l'aide de notre segmenteur sont donnés dans la Figure 10 et la Figure 11 ci-dessous.

Figure 10 : Extrait d'un article journalistique somalien

www.puntlandpost.com

NEWS - WARARKA Qaramada Midoobay oo walaac ka muujisay xayiraadda wali saran duulimaadyadii Soomaaliya. By Elmi. Posted to the web 25-06-2003,15:04:03 Muqdisho : Haweenay u hadashay hay'adda UNDP ee fadhigeedu yahay Nayroobi ayaa sheegtay in haddii ay sii socdaan xayiraadda saaran duulimaadkiiSoomaaliya iyo Kenya, in ay keeni karto dhibaato xagga banii'aadinimada ah.

NEWS	the	ay
-	web	sii
WARARKA	25/06/2003	socdaan
Qaramada	,	xayiraadda
Midoobay	15 :04 :03	saaran
oo	Muqdisho	duulimaadkii
walaac	:	Soomaaliya
ka	Haweenay	iyo
muujisay	u	Kenya
xayiraadda	hadashay	,
wali	hay'adda	in
saran	UNDP	ay
duulimaadyadii	ee	keeni
Soomaaliya	fadhigeedu	karto
.	yahay	dhibaato
By	Nayroobi	xagga
Elmi	ayaa	banii'aadinimada
.	sheegtay	ah
Posted	in	.

Figure 11 : Segmentation de l'extrait de l'article somalien

II.2.2.4 Un analyseur morphologique à base de HFST pour le somali

Un analyseur morphologique à base de transducteurs à états finis (HFST) pour le somali a été développé en 2015 au sein de l'université arctique du royaume de Norvège dans le cadre du projet *Giellatekno*. Le framework *HFST* (*Helsinki Finite-State Technology*) a été conçu pour créer et compiler des analyseurs morphologiques. Ce framework contient des outils libres qui sont une duplication du framework *XFST*, comme les outils *lexc* et *twolc,* qui sont très connus et bien documentés, pour la construction d'analyseurs morphologiques sous forme de transducteurs en cascade.

Outre le développement d'analyseurs morphologiques, le framework *HFST* est utilisé aussi pour construire des correcteurs orthographiques, des parseurs et d'autres outils de TALN pour les langues naturelles. Le framework inclut les outils tels que *hfst-lex*, *hfst-twolc* et le *hfst-xfst* qui fournit une compatibilité avec les outils de Xerox.

L'analyseur morphologique du somali développé dans le cadre de ce projet utilise le framework HFST, et l'ensemble des documentations nécessaires pour compiler et utiliser cet analyseur se trouve en ligne sur le site du projet *Giellatekno*. Les données et les programmes

de lancement de cet analyseur morphologique pour le somali étaient librement accessibles en ligne[33] jusqu'à août 2015. Nous les avons téléchargées à l'époque, et sommes toujours en mesure de les utiliser.

II.2.3 Applications (services) linguistiques pour le somali

II.2.3.1 Un correcteur orthographique de base pour le somali

Un premier correcteur orthographique appelé « Higgaad-saxe[34]» a été développé par un linguiste somalien en utilisant les outils *Spell*. Ce correcteur est basé sur un lexique de 17 000 mots somalis, et il y a aussi une version plug-in pour l'intégrer dans Open Office et les applications Firefox.

Ce correcteur est opérationnel, mais il souffre d'un manque de couverture du vocabulaire somalien. Un travail d'amélioration et d'augmentation de son lexique de base serait nécessaire pour le rendre plus efficace et performant.

II.2.3.2 Un prototype de reconnaissance automatique de la parole somalie

Dans le cadre de sa thèse de doctorat, Nimaan [Nimaan, 2007], a réalisé le tout premier système à l'état de l'art de reconnaissance automatique de la parole somalie. En utilisant des corpus textuels monolingues issus du web, il a construit un premier sous-corpus audio somali appelé « Asaas » de 10 heures de parole spontanée à une fréquence d'échantillonnage de 16 Khz, codé sur 16 bits, à l'aide de l'outil de transcription audio *Transcriber*. Ce corpus audio a fait l'objet d'une modélisation acoustique en faisant une correspondance entre les phonèmes français et somalis.
À côté de ce corpus audio transcrit, Nimaan [Nimaan, 2007], a construit un modèle trigramme de la langue somali, basé sur le corpus textuel de 2,8 M de mots récolté sur le web. Il en a ensuite développé un second, basé sur 4 400 racines extraites du modèle du langage.

Les premières expérimentations de reconnaissance de la parole somalie ont été effectuées avec le moteur Speeral du laboratoire d'informatique d'Avignon (LIA) avec un taux d'erreur sur les mots (WER) de 20,9 % sans normalisation orthographique, et de 32% avec normalisation.

Une seconde expérience a été effectuée sur les mêmes données, mais avec une décomposition en racines du système de transcription. Cette seconde expérimentation a amélioré les différents taux d'erreur de la reconnaissance automatique de la parole somalie. Le Tableau 7 et le Tableau 8 ci-dessous contiennent les deux résultats de ces expérimentations.

	Corrects	Sub	Dél	Ins	WER
Non normalisé	75,2	19,2	5,6	7,1	32,0
Normalisé	84,2	13,2	1,9	5,2	20,9

Tableau 7 : Taux d'erreur en RAP du somali avec et sans normalisation [Nimaan, 2007]

	Corrects	Sub	Dél	Ins	Taux erreur
WRER	**87,8**	**8,0**	**4,2**	**1,9**	**14,2**
RER	**83,3**	**10,8**	**5,9**	**1,7**	**18,3**

Tableau 8 : WRER et taux d'erreur racines (RER) en RAP du somali ([Nimaan, 2007])

[33] www.qaamuus.so

[34] http://www.somaliaonline.com/community/topic/somali-born-engineer-develops-spell-checker/

Selon Nimaan [Nimaan, 2007], l'ensemble des données et outils de traitement de la parole somalie construits dans le cadre de ses travaux servira à effectuer un mode de représentation des données audio pour leur indexation. Ces travaux se situaient dans le cadre de la préservation et de la sauvegarde du patrimoine immatériel africain, et plus exactement, de la numérisation, l'archivage, l'indexation et l'exploitation des archives audiovisuelles de la République de Djibouti.

II.2.3.3 Ressources et outils TALN de la fondation culturelle REDSEA-ONLINE.COM

Depuis une dizaine d'années, la fondation culturelle REDSEA-ONLINE.COM, basée à Hargeisa dans le nord de la Somalie [Jama, M., 2012], développe des ressources et des outils pour l'informatisation du somali.

Parmi ses contributions concrètes, il y a la construction d'un logiciel libre de droits appelé UBBO pour la correction orthographique et le traitement de textes en somali. Ce logiciel contient un peu plus que les 180 000 mots les plus fréquemment utilisés en somali. Il est librement téléchargeable en ligne et peut être utilisé dans un ordinateur équipé d'un système d'exploitation Microsoft Windows pour l'instant. Enfin, il contient une version web qui permet à un utilisateur de corriger automatique son texte en ligne via une interface web dédiée. Le même corpus a servi également pour développer un premier prototype d'analyseur morphosyntaxique, qui identifie automatiquement les parties du discours et des informations sur la morphologie ou la dérivation des mots d'un texte en somali.

Dans cette même institution, a été développé un outil de synthèse vocale appelé « Waa Kuma ? » dont l'objectif principal est d'aider les locuteurs somalis, qui souffrent de problème de vision, à identifier les appels entrants dans leur téléphone à domicile. Cette application utilise un moteur de synthèse vocale libre de droit pour prononcer en somali le nom de l'appelant si ce dernier est enregistré dans le répertoire du téléphone, sinon elle annonce seulement le numéro de l'appelant en langue somali.

La fondation REDSEA-ONLINE.COM a développé également un analyseur (parseur) automatique de la poésie somalienne en utilisant les paradigmes et les règles d'altération des sons du somali en s'inspirant de l'analyse structurelle des poésies occidentales.

Enfin, dans le cadre d'un projet de collaboration entre l'université L'orientale de Naples (Italie) et la fondation Redsea, [Jama, M., 2016] a construit un grand corpus annoté d'environ 1,5 million d'items lexicaux du somali. Ce corpus a été divisé en 10 sous-corpus en fonction du sous-langage du texte (poésie, chants traditionnels, proverbes et énigmes, contes traditionnels, fiction, faits réels, journalisme, science).

II.2.3.4 La traduction automatique du somali avec Google

Selon un article publié dans le quotidien britannique The Guardian[35] le 28 août 2013, le système de traduction en ligne et gratuit du géant de l'informatique Google a ajouté 5 nouvelles langues africaines dans son système de traduction, parmi lesquelles se trouve le somali.

Comme nous le verrons en détails dans le chapitre V de ce mémoire, nous avons calculé le score BLEU de la traduction de notre corpus d'évaluation avec GT. Le tableau ci-dessous contient les scores des mesures d'évaluation BLEU, METEOR et TER. Le score BLEU, très faible, n'est pas vraiment corrélé à la qualité d'usage. En effet, un test de M. Benjamin indique surtout qu'il est possible de comprendre les résultats de GT-FR-SO.

[35] https://www.theguardian.com/world/2013/aug/29/google-translate-african-languages

Mesure d'évaluation	Score
BLEU	0,18
METEOR	0,31
TER	0,73

Tableau 9 : Différents scores d'évaluation de GT-FR-SO

Dans les sections ci-dessous, nous allons présenter quelques exemples de traduction d'articles journalistiques du somali vers le français ou vers l'anglais et vice-versa.

II.2.3.4.1 Exemple de traduction d'un article du journal La Nation de Djibouti du français vers le somali avec Google Translate

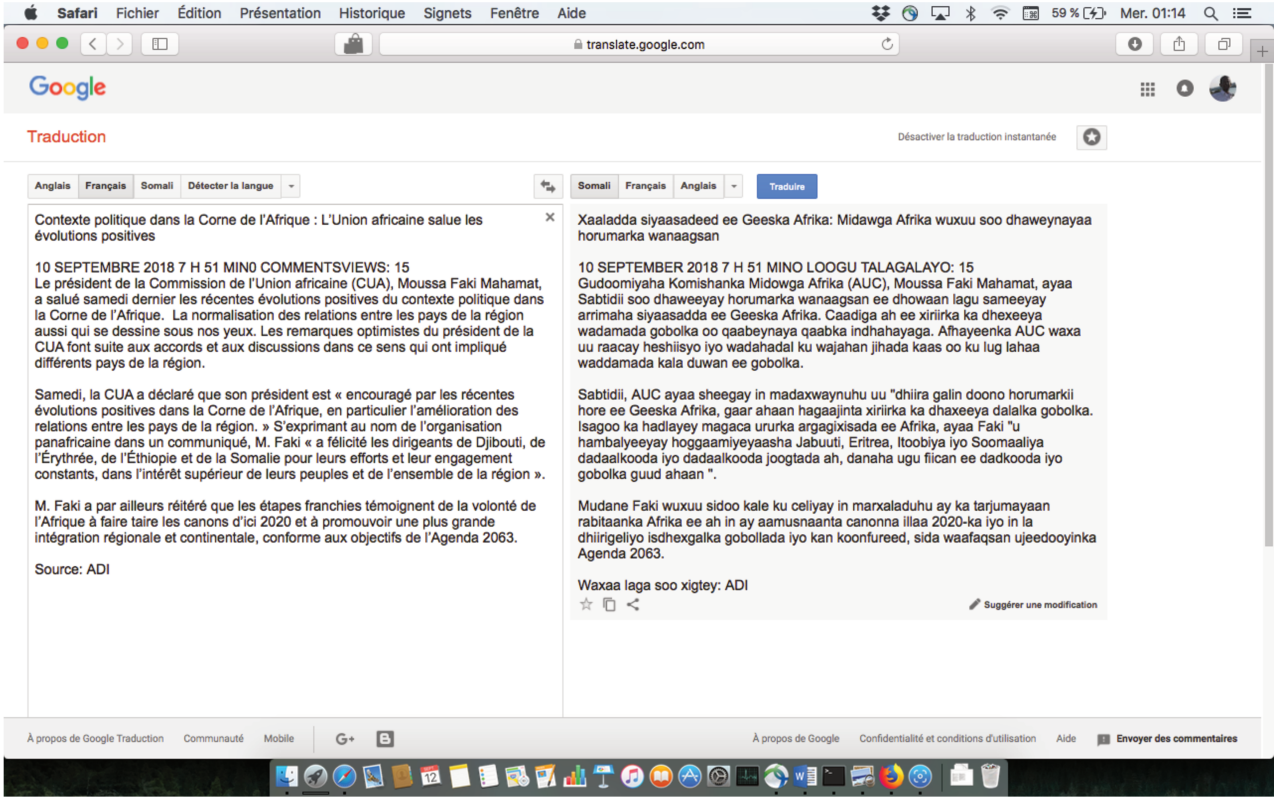

Figure 12 : Exemple de traduction avec Google d'un article de La Nation de Djibouti du français vers le somali

La Figure 12 contient la traduction d'un article du journal La Nation de Djibouti à l'aide de l'interface de traduction de GT. En observant le résultat de la TA avec GT, on peut faire les remarques ci-dessous.

- Les dates, les heures, les sigles et les abréviations ont été directement affichés avec leurs traductions vers l'anglais et non pas vers le somali. Par exemple ; « SEPTEMBRE » a gardé sa traduction en anglais « SEPTEMBER » au lieu de « Sebtember » qui est son équivalent en somali. Le sigle « CUA » qui correspond à la « Commission de l'Union Africaine » n'a pas été traduit vers le sigle équivalent en somali à partir du segment « Komishanka Midowga Afrika » qui pourrait être « KMA » au lieu de « AUC » pour « African Union Commission ».
- Plusieurs mots ou termes ont été directement traduits vers le somali sans tenir compte du contexte et du thème de l'article qui traite ici une information de politique régionale africaine. Par exemple, le terme « normalisation » a été traduit par « Caadiga ah », qui

veut dire normal, alors que les termes « hagaajinta » ou « wanaajinta » étaient plus adéquats comme traductions pour ce terme.

- En dépit des erreurs citées plus haut, comme l'a expérimenté dans ses tests préliminaires Martin Benjamin[36], la traduction français-somali produite par GT est intelligible et correctement compréhensible.

II.3 Besoins à couvrir et problèmes à résoudre (ce qu'on voudrait faire)

II.3.1 Besoins à couvrir

Comme nous l'avons présenté dans la section I.3.1.1.1, nous avons ciblé des besoins précis à combler, qui sont entre autres la possibilité d'accéder en langue somali à la principale source d'information journalistique en République de Djibouti pour les locuteurs de cette langue dans le continent.

Par ailleurs, du fait des faibles ressources multilingues disponibles pour le somali, la constitution d'une communauté de traducteurs/post-éditeurs/volontaires nous a paru indispensable pour amorcer un travail de traduction automatique du somali vers le français.

À travers cette application de traduction des sources journalistiques francophone de Djibouti, nous espérons poser les premières briques pour réaliser l'informatisation du somali, réduire la fracture numérique et faire de l'utilisation des TIC dans le domaine linguistique un premier pas vers l'inclusion numérique et l'alphabétisation des communautés rurales et non scolarisées des pays de la Corne de l'Afrique. Il n'en demeure pas moins que cela impliquera des difficultés et des problèmes techniques et pratiques à surmonter dans le cadre de ce travail.

II.3.2 Problèmes pratiques à résoudre

Pour répondre aux besoins cités précédemment, il convient de s'attaquer à la résolution des trois problèmes pratiques fondamentaux ci-dessous dans le cadre notre travail de recherche :

- Le premier problème auquel nous devrons trouver une solution est celui de la création d'un système de traduction automatique pour cette paire de langues dans un sous-langage spécialisé. En effet, il a été maintes fois démontré dans la littérature du domaine qu'un système de TA spécialisé construit avec des données d'un domaine précis avait plus de performances et produisait de bons résultats sur la qualité de traduction par rapport à un système de TA construit à partir d'un corpus plus généraliste.
- La résolution de ce premier problème implique de créer des ressources linguistiques (corpus bilingue, corpus monolingue) et des outils de base du TALN tels qu'un lemmatiseur ou un segmenteur pour effectuer les prétraitements nécessaires sur les données qui serviront à entrainer le système de TA direct français-somali. Ainsi, du fait du statut de langue très faiblement dotée du somali, il nous faudra trouver une solution nous permettant de construire les premières données parallèles spécialisées pour cette paire de langues pour expérimenter la TA d'une langue peu dotée de la francophonie.
- Une fois les données construites et le système de TA conçu, il nous faudra effectuer une évaluation à la fois subjective et objective sur les données de test. Outre l'utilisation des mesures d'évaluation communément admises dans la TA, nous effectuerons une auto-évaluation annotée d'un petit échantillon de segments parallèles auprès des locuteurs

[36] http://bit.ly/gt_scores

bilingues, et aussi une comparaison entre plusieurs systèmes de TA concernant les temps de post-édition et l'ampleur des modifications nécessaires pour aboutir à une bonne qualité de traduction.

II.3.3 Questions à traiter

Les trois problèmes pratiques que nous avons évoqués au II.3.2, nous amènent à tenter de traiter les deux questions scientifiques ci-dessous.

- La première question que nous allons aborder dans ce travail sera celle de l'approche à adopter pour amorcer un premier travail pour réduire la fracture numérique dans les pays de l'espace somalophone en Afrique de l'Est. Pour cela, nous avons décidé de nous focaliser sur un besoin précis et réel, celui de l'accès en langue maternelle à des informations diffusées en langue française.
- La seconde question que nous allons traiter dans cette thèse est celle de la méthodologie et des outils à utiliser pour concevoir un système de TA indépendant entre une langue peu dotée et la langue française, et surtout la stratégie de pérennisation de ce travail via la constitution d'une communauté de bénévoles qui construira les ressources linguistiques nécessaires à cette tâche et réaliseront également son évaluation.

Conclusion du chapitre II

Dans ce chapitre, nous avons fait un état de l'art de toutes les ressources linguistiques réalisées ou conçues à ce jour pour la langue somalie. Bien que cette dernière soit une langue très peu dotée, un assez grand nombre d'outils de base du TALN nécessaire à la poursuite du projet d'informatisation du somali ont été déjà construits partiellement.

Dans la suite, nous présentons notre contribution et la méthodologie que nous avons adoptée pour résoudre les problèmes pratiques et les questions scientifiques soulevées dans le présent chapitre.

Chapitre III Méthodes et contributions scientifiques

Introduction du chapitre III

Les outils et méthodes pour la construction de ressources multilingues et de systèmes de traduction automatique pour les langues peu dotées sont largement étudiés ces dernières années au sein de la communauté du TALN. En témoignent les nombreuses publications dans les conférences internationales dédiées aux ressources linguistiques et leurs applications.

Ainsi en fonction de l'application TALN visée et de la nature des ressources linguistiques à construire, plusieurs méthodes ont été adoptées et constituent de nos jours l'état de l'art dans ce domaine.

Par exemple, lorsqu'il existe une langue voisine ou proche de la langue peu dotée et que celle-ci est bien dotée, l'approche dite de « voisinage » est privilégiée afin de tirer profit de la ressemblance morphologique et lexicale qui existe entre les deux langues. D'autres ont utilisé le web [Do D., 2011] pour construire des ressources linguistiques telles qu'un corpus monolingue ou bilingue. L'utilisation des réseaux sociaux et de la technique dite de « *crowdsourcing* » a permis également de constituer des corpus parallèles ou d'améliorer les pré-traductions résultant d'un système de traduction, dans le cadre de travaux portant sur les langues peu dotées.

Dans le cadre de ce chapitre, nous présenterons la méthodologie que nous avons suivie tout au long de notre travail pour répondre au besoin précis et réel existant en République de Djibouti et son voisinage, à savoir la construction d'un système de traduction automatique améliorable en continu, dédié au journal en ligne *La Nation de Djibouti*.

Avant cela, nous présentons les objectifs pratiques et les problèmes théoriques inhérents à la construction et au déploiement d'un système de TA français-somali, et en particulier à la construction des ressources nécessaires.

III.1 Choix des objectifs pratiques et des problèmes théoriques

Nous avons choisi de centrer notre travail sur un objectif pratique principal, ainsi que sur les sous-objectifs qui en découlent, et de nous attaquer seulement aux points théoriques sur lesquels nos développements permettront d'avancer. Nous détaillons maintenant (1) la construction et le déploiement d'un système de TA français-somali, (2) la construction de ressources (corpus, dictionnaires), et (3) la stratégie choisie.

III.1.1 Construction et déploiement d'un système de TA français-somali

III.1.1.1 Objectif pratique

III.1.1.1.1 Accès par une iMAG pour améliorer la qualité de prétraductions automatiques par post-édition contributive d'un corpus spécialisé

a. ***Prétraduction d'articles journalistiques avec Google Translate***

À notre connaissance, le seul système de TA gratuit qui permet de traduire un texte entre la paire de langues français-somali est à ce jour *Google Translate* (GT). Ce dernier utilise la langue anglaise comme langue pivot pour traduire un texte depuis le français vers le somali et vice-versa. Par conséquent, dans le cadre de la création d'un corpus bilingue à partir des

articles du journal *La Nation de Djibouti*, il nous faudra utiliser GT pour avoir une première version de chaque article traduit en somali. Ensuite, à l'aide de la plateforme SECTra_w/iMAG et avec le soutien d'un ou plusieurs post-éditeurs bénévoles, nous avons procédé à la post-édition des segments pré-traduits, et à une auto-évaluation des résultats, en leur attribuant un score de qualité allant de 0 à 20.

La Figure 13 et la Figure 14 ci-dessous montrent respectivement un article du journal *La Nation de Djibouti* prétraduit en somali, avec à droite sa version française, et à gauche la traduction produite par GT sous la plateforme SECTra_w/iMAG.

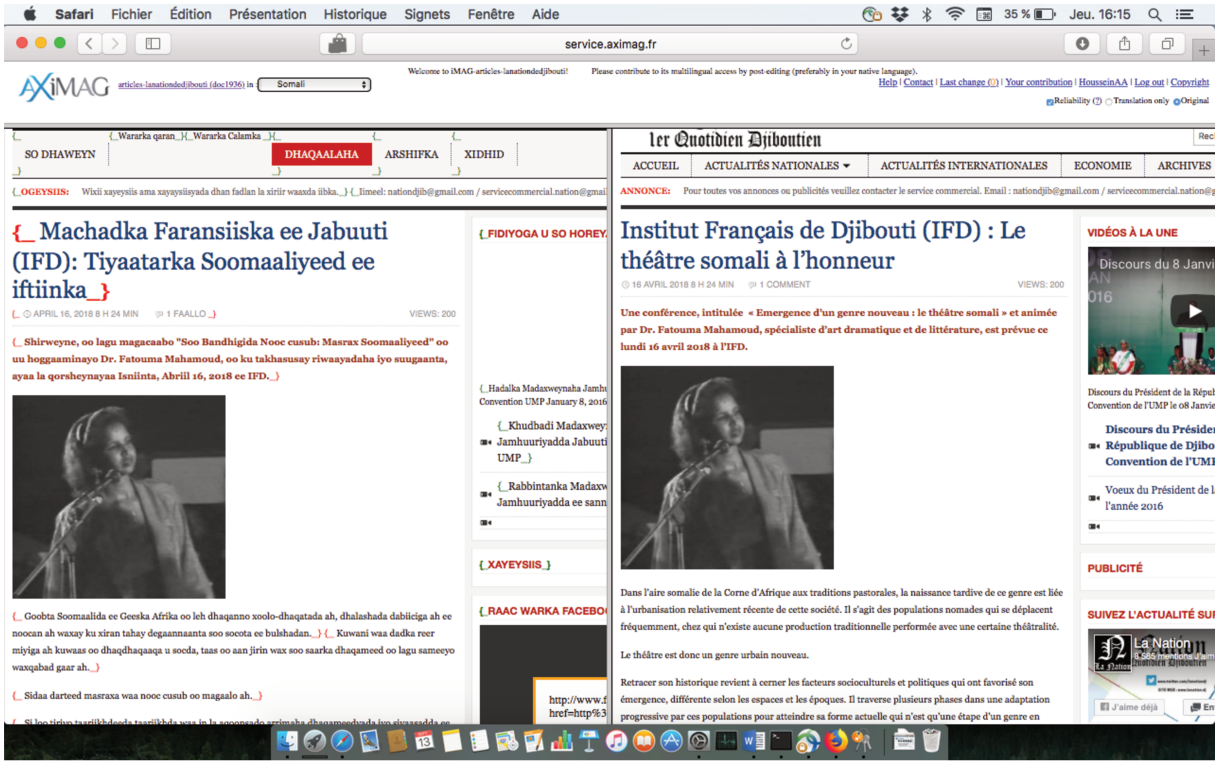

Figure 13 : Article en français du journal La Nation de Djibouti traduit en somali avec GT sous SECTra_w/iMAG

Figure 14 : Version traduite en somali avec GT d'un article du journal La Nation

b. ***Post-édition des prétraductions pour améliorer la qualité de la TA par un ou plusieurs contributeurs bénévoles***

Comme un locuteur du somali peut le voir, dans les deux figures ci-dessus, les prétraductions produites par Google Translate (GT) présentent de nombreuses et graves erreurs de traduction. Par exemple, certains mots des segments sont traduits sans tenir compte de leurs contextes syntaxiques dans la phrase, ou directement mot à mot.

Cependant, les prétraductions produites par GT peuvent aider à la compréhension globale et générale du contenu journalistique de l'article, et peuvent être utilisables (sans correction) dans un premier temps en l'état par des locuteurs somalophones souhaitant comprendre les grandes lignes des articles de ce journal.

D'autre part, GT utilise lui-même un système de TA pivot pour le cas de la paire de langues français-somali en passant par l'anglais à cause de l'inexistence dans leur système d'un grand corpus parallèle et abondant français-somali, contrairement à la paire somali-anglais.

Par conséquent, une ou plusieurs étapes de post-édition par des contributeurs bénévoles seront nécessaires pour améliorer la qualité de traduction et obtenir une seconde version publiable.

La Figure 15 ci-dessous présente une version post-éditée par un post-éditeur bilingue de l'article précédent.

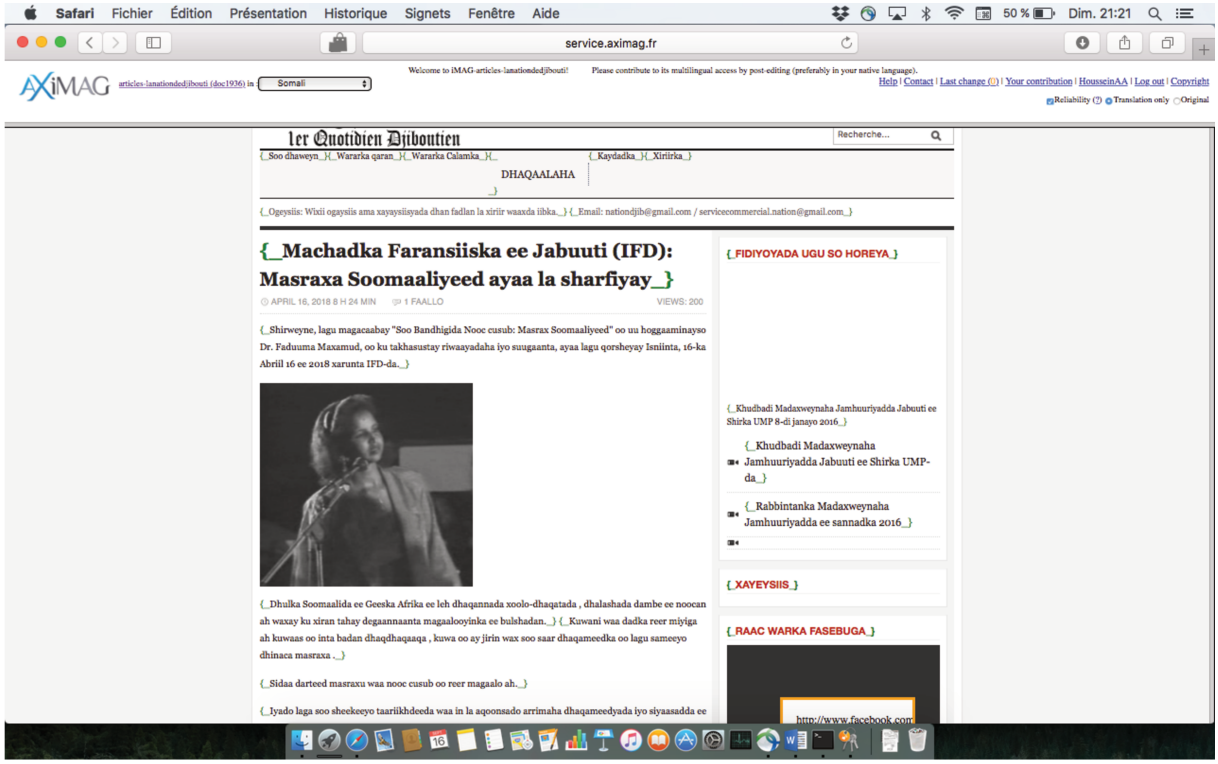

Figure 15 : La version post-éditée de l'article de La Nation

III.1.1.1.2 Amélioration de la qualité de la TA par rapport à Google en réduisant le temps de post-édition

a. ***Calcul de la distance d'édition mixte et du temps de post-édition par page standard (***

Une tâche de post-édition comme celle réalisée sur l'article prétraduit ci-dessus nécessite de procéder à des modifications sur les segments pré-traduits (insertion, suppression, substitution, etc.) afin d'aboutir à une traduction assez bonne pour être mise à la disposition du public cible de notre travail.

Ces modifications réalisées par un post-éditeur nécessairement bilingue, ont un coût en temps qu'on cherche à estimer à partir des différentes opérations faites sur les segments pré-traduits. Nous cherchons donc à relier la distance d'édition (mixte) et le temps de post-édition par page standard de 250 mots.

La première mesure permet de calculer le nombre de modifications ou changements effectués au niveau des mots et des caractères sur un article du journal prétraduit, par rapport à sa version après la post-édition et elle est calculable a *posteriori*, sans effort humain supplémentaire, entre la PE et tout résultat de TA, qu'on soit parti ou non de lui pour obtenir la traduction (post-éditée) considérée (« la PE ») . La seconde mesure permet de calculer le temps que le post-éditeur a passé pour post-éditer une page standard d'un article journalistique. Dans notre cas, il s'agit de calculer la distance d'édition mixte (Dmix) et le temps de post-édition par page standard entre les segments pré-traduits et la post-édition. Contrairement aux mesures classiques d'évaluation de la TA, le temps passé à post-éditer est une bonne mesure de la qualité d'usage.

b. ***Utilisation d'un système de TA spécialisé pour les prétraductions afin de réduire le temps de post-édition par rapport à Google Translate***

Pour améliorer la qualité d'usage des résultats de TA des articles journalistiques, nous construirons un système de traduction automatique spécialisé à partir des données parallèles construites par post-édition de GT.

Rappelons qu'une expérience similaire à la nôtre a été réalisée par [Wang, 2015] sur des données franco-chinoises dans le domaine économique et industriel, et qu'elle a prouvé une très nette amélioration, puisqu'on est passé de 17 min/page avec GT à 11 min/page avec MosesLIG/fr-zh.

III.1.1.2 Questions et problèmes théoriques

III.1.1.2.1 Problèmes de la création d'un service d'accès en langue peu dotée d'un journal francophone en ligne

Le premier problème que nous devons régler dans le cadre de cette thèse est celui de l'accès en langue somali au journal francophone *La Nation de Djibouti,* afin que les locuteurs somalophones de la Corne de l'Afrique puissent comprendre son contenu. En l'absence de traducteurs bilingues franco-somali et d'un système de TA de bonne qualité, nous avons tenté de résoudre ce problème en mettant en place un système de TA par postédition des prétraductions de *Google Translate* à l'aide d'une plateforme dédiée à cet effet.

Le but de notre expérimentation était d'impliquer des Djiboutiens bilingues qui pourraient corriger les résultats de la TA de quelques articles journalistiques de *La Nation de Djibouti* sélectionnés préalablement par nous-même. À partir de leurs postéditions, nous avons constitué un corpus bilingue spécialisé qui a servi de base à la création d'un système de TA autonome français-somali du journal Djiboutien.

III.1.1.2.2 Problèmes de la taille des données bilingues pour créer un système de TA statistique d'une meilleure qualité que Google Translate

Le second problème que nous avons rencontré dans le cadre de ce travail a été la faiblesse de la taille des données bilingues français-somali à notre disposition avant et après l'expérience de post-édition des articles du journal *La Nation de Djibouti*. En effet, étant donné la nature statistique des systèmes de TA à l'état de l'art, la qualité de la traduction d'un nouveau système de TA pour un nouveau couple de langues dépend à la fois de la quantité des données bilingues utilisées pour entraîner le modèle d'apprentissage de la traduction, et de la spécialisation des ces données au sous-langage visé.

Dans notre cas, nous avons pu constituer un corpus bilingue de seulement 12 863 segments (194 135 mots). C'est pourquoi nous avons combiné notre corpus parallèle avec d'autres données bilingues hors du sous-langage journalistique, collectées à partir de sources ouvertes et libres de droits. Nous verrons que cela a un peu amélioré nos résultats.

III.1.1.2.3 De l'amélioration de la qualité en termes de post-édition des prétraductions avec différentes approches et systèmes de TA statistique

Après avoir construit un corpus parallèle à partir des postéditions de bénévoles bilingues, il nous fallait trouver une méthode pour améliorer et évaluer la qualité des prétraductions de GOOGLE TRANSLATE, mesurée à partir du temps de postédition total. Il nous fallait donc pouvoir estimer les temps qu'auraient mis les post-éditeurs s'ils avaient post-édité les résultats de nos systèmes. Il en faudrait sans doute au moins 2 fois plus comme pour le français-chinois de [Wang, 2015] pour dépasser GT.

[Wang, 2015] et [HaoZhou, 2016] ont dans le cadre de leurs travaux sur la TA français-chinois, développé un bon estimateur de qualité (QE) afin de choisir le meilleur système de TA. Les expérimentations qu'ils ont réalisées ont permis de mettre en évidence une corrélation linéaire entre la distance d'édition mixte (Dmix) et le temps de post-édition de la TA. Ce dernier temps peut être transformé en un score qualitatif par la formule de Boitet.

III.1.1.2.4 De l'estimation de la qualité pour choisir le système de TA avec moins de temps de post-édition

La principale question scientifique à laquelle nos expérimentations doivent donner des réponses est de déterminer quelle est la quantité de données parallèles suffisante pour créer un système de TA qui sera meilleur que GT, et dont les lecteurs somalophones du journal trouveront les traductions (brutes) utilisables et compréhensibles. Si, à l'issue de la construction du premier système de TA autonome français-somali, la qualité de traduction n'est pas au rendez-vous, nous aurons 2 possibilités non exclusives :

1) créer sur les mêmes données bilingues un ou plusieurs autres systèmes de TA basés sur d'autres approches, comme la TA neuronale ou la TA statistique factorielle à base de fragments,
2) augmenter la taille de notre corpus bilingue spécialisé et de bonne qualité (en post-éditant les meilleurs résultats de TA)

Cette diversité d'approches de TA et la multiplication des systèmes de TA, quoique basés sur des données parallèles de relativement faible quantité, nous permettre d'effectuer une évaluation et une analyse comparative de ces différents systèmes sur un petit échantillon de données de test. Cette analyse devra nous permettre de trouver les voies et les moyens pour améliorer la TA français-somali, et aussi d'identifier le système de TA avec lequel nos contributeurs post-éditeurs passeront le moins de temps à post-éditer les prétraductions. La réduction du temps de post-édition constituera en effet dans notre cas le facteur discriminant le plus important entre les différents systèmes de TA spécialisés.

III.1.2 Construction de ressources

III.1.2.1 Corpus

III.1.2.1.1 Objectif pratique

a. ***Construction d'un corpus bilingue pour un premier système de TA direct français-somali***

La première ressource linguistique et la plus importante lorsqu'on souhaite mettre en place un système indépendant de TA est un corpus bilingue contenant des centaines de milliers ou des millions de segments de texte alignés au niveau des phrases de la paire de langues à traduire automatiquement. Généralement, dans la littérature, la méthode la plus utilisée consiste à récupérer et extraire des données parallèles à partir du Web ou d'archives numériques d'institutions pratiquant le multilinguisme. Par exemple, pour la commission européenne, un système de traduction statistique multilingue a été construit à partir d'un corpus de 20 M de segments (ou phrases) multilingues qui était déjà disponible.

[Pouliquen et al., 2013] ont également utilisé des mémoires de traductions issues d'un long travail de traduction professionnelle qui s'est étalé durant les 11 années précédentes, pour son système de TA multilingue pour le compte de l'OMPI[37]. D'autres travaux, tels que ceux de

[37] L'Organisation Mondiale de la Propriété Intellectuelle (OMPI), ou WIPO fait partie des agences internationales du système des Nations Unies, et son siège est à Genève.

[Do D., 2011] ont utilisé des corpus comparables pour extraire sous certaines conditions un corpus parallèle pour construire un système de TA.

Toutefois, aucune de ces méthodes ne convenait à notre cas, puisqu'il n'existait avant notre thèse aucun corpus comparable et encore moins parallèle dans le domaine des nouvelles journalistiques de la paire français-somali.

La méthode que nous avons adoptée et qui correspond à notre cas et notre besoin est celle utilisée ces dernières années au sein du groupe GETALP du laboratoire d'Informatique de Grenoble(LIG). Elle consiste à construire un corpus bilingue spécialisé pour un couple de langue français-langue_X peu doté, en tirant profit de l'existence d'un système de TA langue anglaise-langue_X à partir des prétraductions produites par un système de TA généraliste, librement accessible et utilisable dans un environnement dédié afin de gagner du temps, sans passer par des traductions professionnelles qui seraient trop coûteuses en temps et ressources financières.

b. ***Constitution d'un corpus brut monolingue du somali (modèle du langage) à partir du web***

La seconde ressource linguistique indispensable à la construction d'un système de TA indépendant est un corpus brut monolingue. Il est constitué de millions de segments ou phrases écrites dans une seule langue, sans aucune autre information supplémentaire. Les corpus bruts monolingues sont surtout utilisés dans le domaine du TALN pour créer des modèles de langue.

Dans le cas précis de la construction d'un système de TA statistique, le corpus monolingue sert à construire le modèle de la langue cible qui est utilisé par le décodeur.

Pour avoir le plus possible de n-grammes et mieux estimer leurs probabilités, le corpus monolingue doit être le plus volumineux possible.

Grâce à la présence croissante de corpus langagiers sur la Toile, on utilise souvent les méthodes par aspiration ou par extraction à partir du web pour construire des corpus monolingues. Dans le cadre de notre travail, nous avons utilisé un corpus monolingue d'environ 4 M de mots aspiré du web pour créer le modèle de la langue du somali.

III.1.2.1.2 Questions théoriques : comment évaluer la qualité de la TA et celle du corpus bilingue ?

a. ***Evaluer subjectivement le corpus post-édité***

a.i **Auto-notation du corpus post-édité**

Le système de traduction GT et la plateforme SECTra_w/iMAG nous ont permis de construire le premier corpus bilingue français-somali spécialisé sur les nouvelles journalistiques grâce à la post-édition de nos contributeurs bénévoles.

Outre une évaluation objective (temps de PE, distance Dmix), nous avons voulu réaliser une évaluation subjective portant sur la perception de quelques locuteurs somalophones bilingues sur la qualité de la TA français-somali. Pour cela, il devrait suffire de soumettre un échantillon de segments d'un article, tiré de notre corpus post-édité, à des annotateurs bilingues afin qu'ils puissent attribuer de nouvelles notes de qualité (allant de 0 à 20) et proposer le cas échéant des corrections sur les segments déjà post-édités.

Cette évaluation par notation nous permettra d'avoir un aperçu global de la qualité de notre post-édition, mais pas d'évaluer le système de TA utilisé, car, pour nos évaluateurs les traductions (post-éditées par nous), qu'ils évaluent pourraient aussi bien être des premiers jets de traduction 100% humaine.

a.ii Notation par des juges bilingues

Dans le métier de la traduction, il est toujours important de savoir en amont l'impression et de la perception qu'auront les futurs utilisateurs sur la qualité de la traduction.

Par exemple, dans le cadre de la localisation de logiciels utilitaires ou grand public, les traductions des messages à localiser faites par des traducteurs professionnels avec ou sans l'aide d'un système de TA sont envoyés à des réviseurs/valideurs afin qu'ils puissent valider et corriger avant la publication des versions définitives des messages traduits.

Ainsi, il est parfois primordial d'effectuer une évaluation subjective sur la sortie des systèmes de TA pour améliorer la qualité de ces systèmes ou comparer les traductions de plusieurs systèmes de TA sur les mêmes segments.

Dans notre cas, nous allons soumettre un échantillon de segments bilingues de notre corpus à une dizaine d'annotateurs djiboutiens maîtrisant parfaitement le français et le somali, pour qu'ils attribuent des notes à nos segments post-édités et proposent éventuellement une seconde post-édition des segments jugés mal traduits (ce qui est déjà indiqué par l'auto-évaluation du post-éditeur ayant fait la première passe).

a.iii Cohérence de l'auto-notation avec les temps de post-édition et les notes des juges bilingues

Une fois les tâches d'auto-notation et d'évaluation par des annotateurs bilingues finalisées, nous effectuerons un travail d'analyse et de synthèse qui portera sur les deux aspects suivants.

- la comparaison entre le temps de post-édition de tous les annotateurs et le temps de nos post-éditions sur le corpus initial,
- l'analyse et l'étude de la cohérence des notes attribuées par les juges annotateurs bilingues à chacun des segments, et les notes que nos post-éditeurs avaient attribuées auparavant. Il s'agit aussi dans ce cas précis de calculer l'accord inter-annotateur selon la formule du Kappa de Fleiss.

Nous effectuerons également d'autres calculs statistiques pour enrichir notre évaluation subjective et faire des commentaires et des remarques sur les aspects les plus intéressants de cette enquête.

b. *Evaluer objectivement la qualité de la TA*

b.i Temps de post-édition total avec Google Translate

Durant la phase de construction de notre corpus bilingue post-édité, nous avions pris l'habitude de relever, après la fin de la post-édition de chaque article, le temps en minutes que nous avions passé pour réaliser cette post-édition en partant de la prétraduction de *Google Translate*.

Nous avons ensuite calculé le temps moyen total de post-édition sur l'ensemble de tous les articles journalistiques de notre corpus. Ce dernier est appelé temps de post-édition totale avec GT et nous servira de référence pour comparer la post-édition avec GT et d'autres systèmes de TA. Son mode de calcul repose sur le quotient entre le temps de post-édition global de chaque article journalistique et le nombre de pages dont il est composé durant la phase de post-édition sous l'interface avancée de post-édition de SECTra_w/iMAG.

b.ii Temps de post-édition total avec les différents systèmes de TA

À l'instar des post-éditions avec GT, nous relèverons et calculerons aussi le temps de post-édition sur un petit échantillon de segments d'évaluation issus du corpus post-édité avec les différents systèmes de TA construit sur le corpus post-édité.

Ainsi nous aurons un temps de post-édition total pour chacun des systèmes de TA.

Par la suite, nous allons effectuer une comparaison entre le temps de post-édition total avec GT et avec chacun des autres systèmes de TA spécialisés. Le but de cette évaluation objective est d'identifier le système de TA demandant le moins de temps de post-édition car c'est celui qui devra être utilisé pour poursuivre la construction et l'amélioration du corpus bilingue post-édité.

III.1.2.1.3 Évaluation

a. ***Objectifs pratiques***

a.i **Estimation de la qualité linguistique du corpus bilingue post-édité sans traduction professionnelle**

La principale difficulté dans le cadre du développement de systèmes de TA pour les langues peu dotées est l'absence de corpus parallèles issues des traductions professionnelles. Cela s'explique par le nombre insuffisant des traducteurs professionnels pour les langues africaines, et le coût trop élevé que cette tâche pourrait engendrer si on voulait faire appel à des professionnels du métier.

a.ii **Évaluation de la qualité de traduction d'un système de TA spécialisé à base de corpus post-édité avec différentes approches et outils de TA**

Comme il est prévisible, à cause de la petitesse de notre corpus bilingue post-édité, la qualité de la TA français-somali avec notre premier système risque d'être mauvaise.

Par manque de temps et à cause de l'impossibilité pour nous de mobiliser des moyens financiers pour payer des post-éditeurs bilingues, même en *crowdsourcing* (comme l'a fait [Potet, 2013] afin d'augmenter largement la taille de son corpus bilingue), nous avons décidé d'expérimenter plusieurs approches différentes de TA afin d'améliorer la qualité de traduction de ces systèmes par rapport à celle de GT.

b. ***Objectifs théoriques***

b.i **Estimation de la corrélation entre la distance mixte de post-édition et le temps total de PE**

Parmi les objectifs théoriques que nous voudrions atteindre dans le cadre de notre travail, il y a la question la corrélation qui peut exister entre le temps total de post-édition d'un document prétraduit et post-édité, et la distance d'édition mixte. Une telle corrélation a déjà été mise en évidence dans le cadre d'un travail expérimental sur la paire de langue français-chinois [Haozhou, 2015].

b.ii **La recherche sur l'estimation de qualité pour choisir les meilleures prétraductions pour la post-édition**

Les méthodes d'évaluation objective des résultats des systèmes de TA actuellement en vigueur dans ce domaine sont basées sur la comparaison avec des traductions de référence, généralement produite par des traducteurs professionnels.

Toutefois ces dernières années des travaux comme ceux de [Gamon et al. 2005 ; Negri et al. 2012 ; Avramidis 2012 ; Gupta et al. 2013] ont été entrepris pour trouver des méthodes pouvant estimer la qualité des segments pré-traduits avec un système de TA, qu'il soit empirique, analogique, neuronale ou à base de règles mais sans références.

Il s'agit, étant donné un segment source Seg_L1 et une traduction proposée Seg_L2, ainsi que diverses informations (par exemple la terminologie bilingue à respecter) de calculer un score QE (Seg_L1, Seg_L2, infos).

L'estimateur de qualité ne peut pas utiliser la distance avec la PE, puisqu'il s'agit justement de créer cette PE à partir du résultat de TA ayant le meilleur score QE. On cherche en fait à *prédire* l'effort de PE. Parmi les informations utilisées, on peut faire intervenir, outre une terminologie bilingue, la longueur Seg_L2 par rapport à celle de Seg_L1, les scores de correction orthographique et grammaticale de Seg_L2, le score du modèle de langue Seg_L2 pour le modèle de langue du sous-langage spécialisé, etc.

III.1.2.2 Dictionnaire(s)

III.1.2.2.1 Objectifs pratiques

a. ***Premier dictionnaire français-somali du sous langage journalistique***

Comme nous l'avons évoqué au II.1.1.1, la majorité des dictionnaires monolingues ou bilingues des langues peu dotées est éditée au format papier. Il est difficile dans ces conditions d'utiliser ces ressources dictionnairiques dans le cadre de projets TALN.

Pour le somali, il existe plusieurs dictionnaires de référence mais à l'exception de quelques projets de dictionnaires informatisés somali ou langue somali-X[38], il n'existe à ce jour aucun dictionnaire informatisé dédié au somali et entièrement opérationnel.

De grands projets de recherche tels que *Jibiki* et *Papillon* ont utilisé des dictionnaires informatisés pour construire des bases lexicales multilingues mais elles ne contiennent pas le somali.

L'un des objectifs basiques et sous-jacentes à notre ambition d'amorcer l'informatisation du somali est d'utiliser notre premier corpus bilingue construit par post-édition des articles de La *Nation de Djibouti* pour produire une version informatisée et librement accessible du premier dictionnaire français-somali de sous-langage nouvelles journalistiques.

Ce dictionnaire pourrait servir non seulement à améliorer la TA français-somali mais aussi comme outil d'aide à la traduction pour des professionnels des médias en République de Djibouti et ailleurs.

Il existe des outils libres de droits comme *Giza++* ou *Apertium* pour créer des dictionnaires bilingues à partir d'un corpus parallèle[39].

Par ailleurs, des travaux récents tels que celui de [Nasredine et al., 2016] ont montré l'impact positif d'un dictionnaire bilingue spécialisé sur la performance d'un système de TA.

b. ***Second dictionnaire combiné avec les données déjà existantes d'un dictionnaire français-somali***

À cause de la taille insuffisante de notre corpus bilingue post-édité et la présence d'un nombre élevé de mots ou d'expressions qui se répètent dans tous les articles du journal, le premier dictionnaire bilingue spécialisé qui sera créé avec nos données initiales n'atteindra même pas le millier de mots différents. C'est pourquoi, pour enrichir et augmenter la taille de notre dictionnaire informatisé, nous combinerons entre ce premier dictionnaire et la version numérique du dictionnaire français-somali de [Farah A. G, 2008] qui contient 9000 entrées, manuellement traduites et couvrant divers domaines.

Au total, notre second dictionnaire combiné et informatisé contiendra environ 10 000 mots de la paire français-somali. Ce deuxième dictionnaire bilingue pourra servir à améliorer la qualité

[38] https://www.lexilogos.com/somali_dictionnaire.htm et http://hooyo.web.free.fr/F_dico_01.html

[39] http://wiki.apertium.org/wiki/Utiliser_GIZA%2B%2B

de la TA français-somali, et pourra être utilisé par nos post-éditeurs ou les traducteurs professionnels djiboutiens comme outil de TAO.

III.1.2.2.2 Questions théoriques

a. ***Utilisation d'un corpus bilingue pour enrichir un dictionnaire déjà existant***

Le développement des ressources linguistiques comme un corpus bilingue post-édité peut participer non seulement à la constitution de ressources pour une langue peu dotée, mais aussi à l'enrichissement et à l'amélioration d'un système de TA, ainsi que d'un dictionnaire déjà existant pour cette langue.

Pour cela, nous appliquerons différents outils pour extraire de notre corpus bilingue de qualité, en croissance continue, des termes et entités nommées monolingues et bilingues.

Cette réflexion nous permettra de voir le lien entre développement de ressources spécialisés pour la TA d'une langue peu dotée et l'informatisation de cette dernière afin de réduire la fracture numérique dans les pays d'Afrique francophone.

b. ***Enrichir une base lexicale de type PIVAX-3 avec termes et entités nommées (EN)***

La constitution de dictionnaires bilingues informatisés dont le premier est spécialisé sur les mots issus de notre corpus journalistique et le second est une forme combinée entre le premier et d'autres mots et termes provenant d'un dictionnaire plus généraliste du français, nous a amené à nous poser les trois questions ci-dessous.

- Pourra-t-on augmenter la taille de notre dictionnaire bilingue après avoir combiné les deux premiers dictionnaires ?
- À cause de la nature journalistique du contenu du premier dictionnaire, combien d'entités nommées, d'acronymes ou d'abréviations pourra-t-on trouver dans les deux dictionnaires bilingues ?
- La réponse à ces deux premières questions nous ramène à la question de savoir s'il sera possible d'enrichir facilement une base lexicale français-somali dans PIVAX-3 avec le vocabulaire général, les termes et les entités nommées, à l'instar du travail effectué dans [Zhang, Y. & Mangeot, M., 2013].

III.2 Stratégie choisie

III.2.1 Stratégie de construction des corpus

III.2.1.1 Choix et sélection des données du sous-langage

III.2.1.1.1 Définition et choix d'un sous-langage

Comme déjà dit, nous nous sommes fixé comme objectif de mettre en ligne un « bon » système de TA français-somali dédié au journal francophone *La Nation de Djibouti*. Outre sa version papier éditée et distribuée quotidiennement dans la capitale et dans les régions de l'intérieur, il a un site web ou sont publiées les versions numériques des 10 au 20 article les plus pertinents de la version papier.

Comme indiqué dans sa page d'accueil[40], le journal *La Nation de Djibouti* est composé de deux rubriques principales qui contiennent tous les articles publiés en ligne. La première

[40] http://www.lanationdj.com

rubrique concerne l'actualité nationale et a pour sous-thèmes l'économie, le sport et la santé. La seconde rubrique contient les actualités internationales.

Dans le cadre de notre travail, nous avons choisi de post-éditer les articles à post-éditer majoritairement dans la rubrique « actualité nationale » et majoritairement ses sous-thèmes « économie » et « santé ».

III.2.1.1.2 Sélection des articles journalistiques en ligne du site Lanationdj.com

Avant de soumettre les articles à la plateforme SECTra_w/iMAG pour la tâche de post-édition, nous les avons sélectionnés depuis le site web du journal http ://lanationdj.com/ en tenant compte des critères ci-dessous.

- Le contenu de l'article doit traiter des sujets tels l'économie, la santé et parfois l'éducation ou l'enseignement supérieur. Nous nous sommes limité à ces trois thèmes car nous pensons que ce sont les articles les plus consultés et lus par les visiteurs du journal.
- Le second critère était celui de la taille du contenu de l'article. Ce critère nous a permis d'éliminer les articles dont les contenus étaient trop longs car le système de TA que nous utilisons pour les prétraductions limite les fichiers à traduire à 5000 mots. Nous avions fait l'expérience d'un ou deux articles trop longs et nous avons remarqué que le découpage en paragraphe était très mal fait à cause de leur taille et cela s'est répercuté dans les prétraductions. Certains segments étaient très mal traduits, car ils étaient coupés de leur contexte et la traduction mot-à-mot est très mauvaise pour le somali avec GT.
- Le troisième critère était que leurs contenus ne devraient pas être trop difficile à comprendre ou traduire pour nos futurs post-éditeurs. Les articles qui contenaient des parties de texte littéraire ou trop technique ont été exclus en amont de notre sélection.

III.2.1.2 Choix d'un ou plusieurs systèmes de TA pour les prétraductions

III.2.1.2.1 Google Translate, premier système de TA par défaut

Pour réaliser les prétraductions de notre sélection d'articles journalistiques, il nous fallait choisir un ou plusieurs systèmes de TA. À l'exception de *Google Translate* qui permet de traduire un texte depuis ou vers le somali en passant par l'anglais, nous n'avons trouvé aucun autre système de TA traitant le somali.

Depuis ses travaux de recherche en TA probabiliste, [Och F., J., 2002] est considéré comme étant le pionnier ou le fondateur du système de TA Google Translate.

Ses travaux ont permis de construire le premier prototype de traduction automatique de Google en 2005. Outre des textes ou pages web, qui ne doivent pas dépasser les 5000 mots (deux fois plus si on appelle GT depuis Chrome), GT peut traduire aussi des documents sous forme de fichier au format doc, txt, pdf, etc.

Dans sa première version, GT traduisait seulement entre l'anglais, le français, l'arabe, l'espagnol et l'allemand.

Aujourd'hui, GT est capable de traduire automatiquement entre plus de 103 langues parlées[41] dans les 5 continents (Afrique, Asie, Amérique, Europe, Océanie). À ce jour, c'est le système de TA gratuit le plus utilisé et le plus populaire.

Outre le somali, GT traduit actuellement depuis et vers les 11 langues africaines suivantes : le zoulou, le yoruba, le xhosa, le swahili, le shona, le malagasy, l'igbo, le haoussa, le chichewa, l'amharique. Pour ces onze langues, il est donc possible de construire des corpus parallèles en

[41] https://translate.google.com/intl/fr/about/languages/

post-éditant les prétraductions comme nous l'avons nous-même fait, et de tirer profiter de nos résultats en TA statistique et neuronale dans un sous-langage spécialisé.

Seules, le haoussa du Niger et le malagasy de Madagascar, qui sont parlés dans deux pays de l'espace francophone, pourraient dans un premier temps faire l'objet d'une expérimentation similaire à la nôtre (en TA français-haoussa et français-malagasy) dans un domaine spécialisé, car un réel besoin de traduction existe dans ces deux pays.

III.2.1.2.2 Un ou plusieurs systèmes de TA Moses comme module de TA

Pour ne pas nous limiter seulement à GT pour construire notre corpus bilingue français-somali, nous avons construit plusieurs systèmes de TA avec nos premières données bilingues initiales, en utilisant les boites à outils MOSES [Koehn et al., 2007] et OPENNMT [Klein et al., 2017].

En effet, comme nous voudrions doter le somali d'un système de TA indépendant, nous avons prévu dès le début d'utiliser GT seulement comme système de prétraduction.

Notre stratégie pour cela est, à partir de notre premier corpus bilingue post-édité, de construire plusieurs systèmes de TA indépendants, non seulement afin d'améliorer la qualité de la TA français-somali, mais aussi pour déterminer le type de système de TA qui permettra à nos contributeurs et post-éditeurs bénévoles de passer le moins de temps possible à post-éditer.

III.2.1.3 Une passerelle web iMAG avec une interface pour la post-édition

III.2.1.3.1 Première interface de post-édition

La plateforme de post-édition SECTra_w/iMAG utilisée dans ce travail dispose de deux interfaces de post-édition.

La première interface qui est visible dans la figure 16 ci-dessous affiche dans un cadre dédié la page web pré-traduite avec tous ses segments, tels que le segmenteur de GT les a segmentés.

Figure 16 : Première interface de post-édition d'un article de La Nation de Djibouti

III.2.1.3.2 Interface de post-édition avancée

La première interface de post-édition vu dans la section ci-dessus est surtout destinée aux utilisateurs qui veulent améliorer les prétraductions proposés par le système de TA afin de mieux assimiler et comprendre le contenu de l'article. Elle pourrait être utilisée pour mettre en place rapidement un service d'accès en langue somali au journal *La Nation de Djibouti*. Ce service permettra donc de proposer une première version de l'article traduit en somali dans sa mise en page initiale avec la possibilité pour l'utilisateur de corriger les parties qui l'intéressent sans passer beaucoup de temps.

La seconde interface de post-édition, dite « avancée » (cf. Figure 17), permet au contributeur de réaliser proprement la post-édition. À travers une interface web dans laquelle est indiqué le nom du pseudo-document à post-éditer (ici Doc1937), on accède à la mémoire de traductions à laquelle appartient l'article et une indication sur le nombre de segments de l'article à post-éditer avec une précision sur le pourcentage restant de la tâche de post-édition. Le contributeur pourra facilement faire complètement sa tâche de post-édition.

Outre les informations citées précédemment, le post-éditeur dispose d'un cadre à 3 colonnes dont la première contient le texte source, par groupes de 20 segments, un champ textuel contenant les segments pré-traduits avec GT à post-éditer, et enfin une dernière colonne qui contient différentes propositions de post-édition sous forme de mémoire de traductions.

Enfin, dans l'interface de post-édition avancée, le post-éditeur dispose d'une zone de texte sous forme de liste déroulante ou il peut choisir un score différend du score par défaut associé à son profil pour noter sa post-édition.

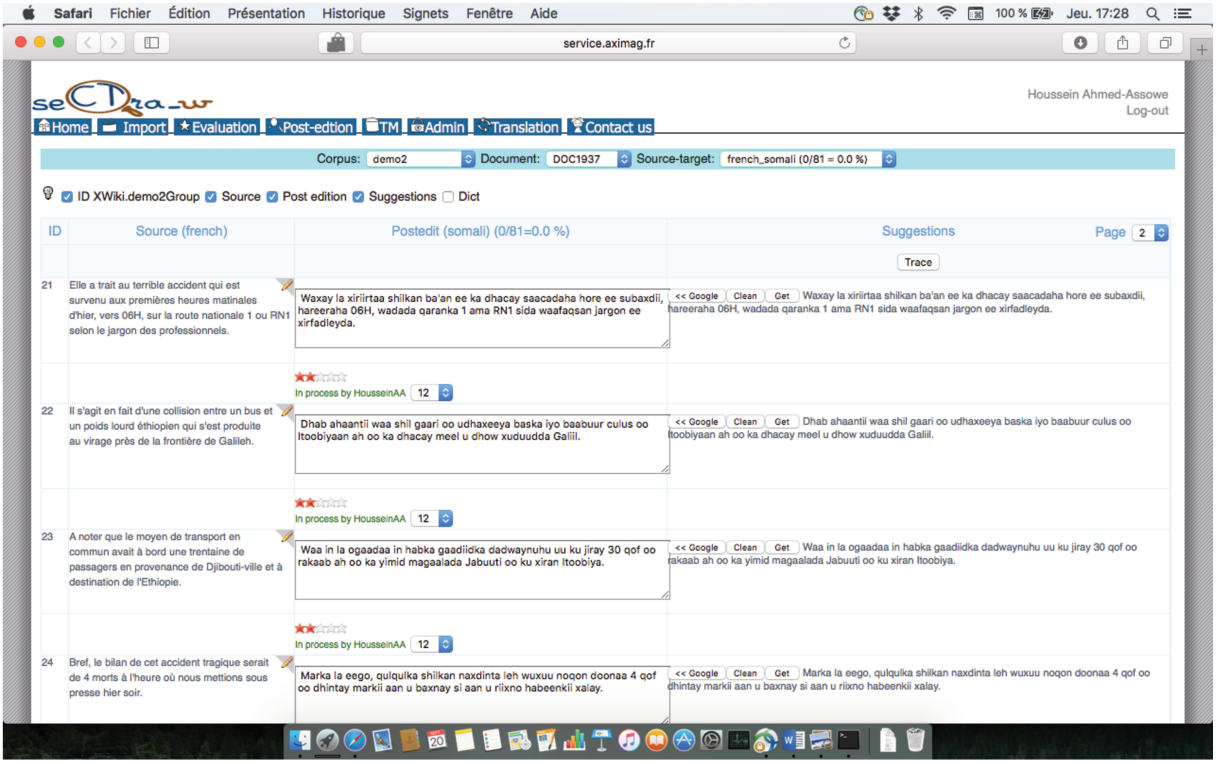

Figure 17 : Interface avancée de post-édition d'un article de La Nation de Djibouti

III.2.1.4 Recueil et analyse du corpus post-édité

III.2.1.4.1 Méthode de recueil des articles post-édités

Notre méthode pour recueillir les articles post-édités consiste à prélever des informations telle que le temps de post-édition par document ou page standard de 250 mots, ou encore le nombre de segments et de mots de chaque document post-édité.

Voici les différentes étapes que nous avons suivies pour construire notre corpus post-édité en utilisant SECTra_w/iMAG.

- Sélection et stockage dans un espace web de tous les articles à post-éditer. Cette étape est primordiale car, étant donné que le système SECTra_w/iMAG utilise des pseudo-documents pour faire la post-édition, pour éviter et surmonter les difficultés qui peuvent résulter des changements ou modifications qui seront faites sur les données de notre corpus qui sont de type journalistique, nous avons créé un espace web dédié pour stocker d'abord les articles de notre sélection afin d'avoir une copie locale non modifiable,
- Création et utilisation d'une passerelle iMAG dans SECTra_w pour corriger les prétraductions de GT. Dans un premier temps, on utilise seulement GT pour produire les prétraductions.
- À la fin de chaque post-édition, recueil dans un fichier Excel de toutes les informations que nous avons jugées utiles pour faire des analyses et des études d'évaluation sur la qualité de notre corpus post-édité.

III.2.1.4.2 Méthode d'évaluation du corpus post-édité

Une fois les articles à post-éditer finis et notre corpus bilingue créé, nous devons effectuer une évaluation portant sur la qualité linguistique de ce corpus avant son utilisation comme corpus d'apprentissage d'un système de TA.

Notre méthode d'évaluation a été la suivante.

Premièrement, nous avons recueilli dans un tableau récapitulatif toutes les informations concernant le temps de post-édition total de chaque document, ainsi que les caractéristiques des données post-éditées comme le nombre de segments et mots post-édités.

Ces informations nous ont permis d'estimer le temps de post-édition de notre corpus avec GT.

Plus tard, nous utiliserons cette information pour faire une comparaison en temps de post-édition total entre les autres systèmes de TA et GT.

Deuxièmement, nous avons fait une évaluation dite « subjective » car elle est basée sur le score de qualité associé à chacun des segments de notre corpus. Nous avons comparé ces scores de qualité dans un échantillon de 54 segments post-édités avec ceux attribués par des annotateurs bilingues.

III.2.2 Stratégie de construction du système de TA

III.2.2.1 Premier essai des systèmes de TA avec les données bilingues initiales

III.2.2.1.1 Prétraitements et nettoyage du corpus post-édité

Comme nous l'avons expliqué dans la section III.2.1.4, nous avons recueilli à partir de la plateforme SECTra_w/iMAG l'ensemble des articles post-édités. Ces données bilingues initiales nous ont servi comme corpus parallèle pour déployer un premier système de TA.

La plateforme de post-édition que nous avons utilisée, offre une interface permettant d'exporter un pseudo-document post-édité, avec la possibilité de choisir les informations qu'on souhaite recueillir (système de prétraduction utilisé, temps de post-édition par segment, identifiant du post-éditeur, score attribué à chaque segment, etc.). Outre cette interface, il existe aussi d'autres possibilités pour exporter d'un seul coup plusieurs documents post-édités à l'intérieur d'une mémoire de traductions. Nous avons utilisé cette seconde option pour exporter l'ensemble de nos 100 articles journalistiques depuis la base de données de la mémoire de traductions (Demo2) de SECTra_w/iMAG.

Ensuite, nous avons effectué les prétraitements et un nettoyage des données brutes exportés, car nous avons rencontré un problème d'encodage des caractères, ainsi que la présence de balises HTML et de caractères spéciaux qu'il a fallu supprimer avant de les utiliser.

La boîte à outils Moses disposant également des scripts Perl pour faire la segmentation et le prétraitement nécessaire, nous les avons utilisés pour réaliser ces tâches, avant de faire l'apprentissage de notre système avec Moses.

Ces prétraitements et nettoyages ont quelque peu réduit la taille de nos segments parallèles.

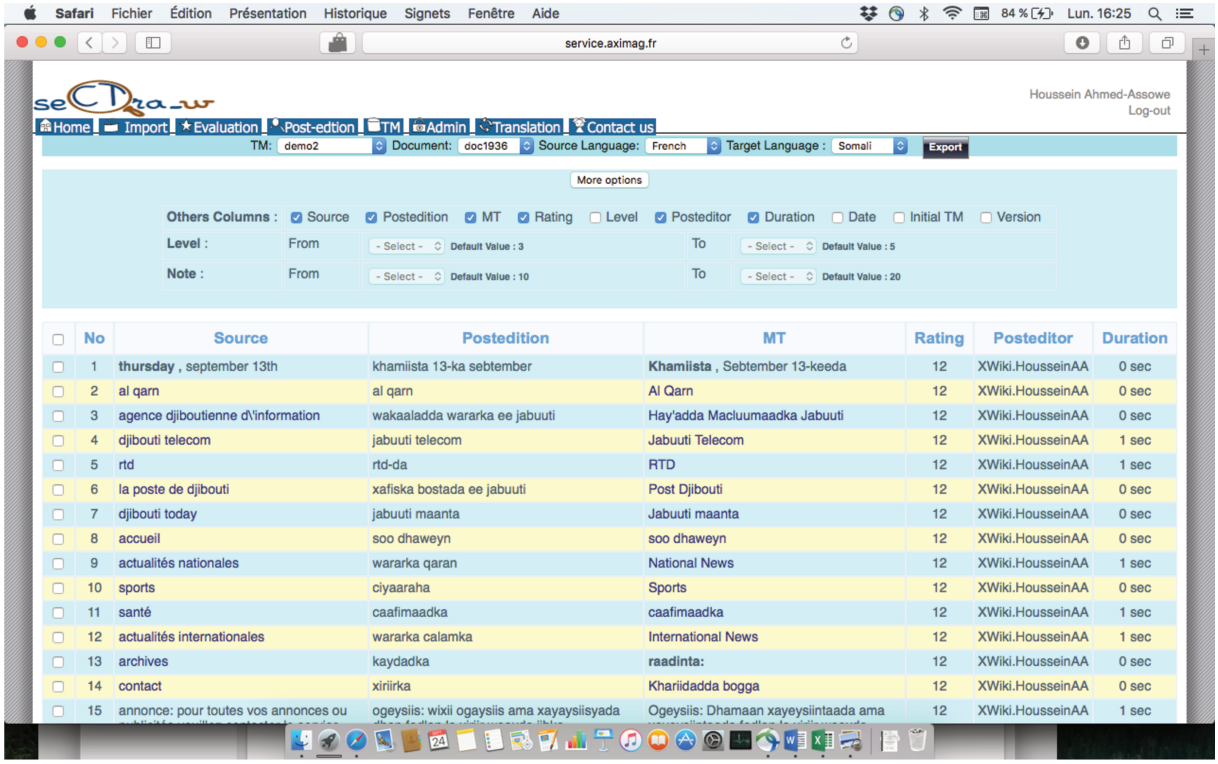

Figure 18 : Interface d'export d'un document post-édité sous SECTra_w/iMAG

III.2.2.1.2 Découpage du corpus d'apprentissage, d'optimisation et d'évaluation

Pour construire notre système de TA sur notre premier corpus, nous l'avons découpé en trois sous-corpus. Le premier sous-corpus représente 85% du corpus post-édité, soit environ 9 069 segments parallèles, et constitue le corpus d'apprentissage.. Le second sous-corpus, appelé corpus d'optimisation ou de développement, représente 10% du corpus total, soit 1 067 segments parallèles. Le dernier sous-corpus, qui contient 643 segments français-somali, est le corpus d'évaluation de notre système de TA.

La taille de ces 3 sous-corpus pourra changer par la suite, quand de nouvelles données bilingues seront ajoutées.

III.2.2.1.3 Construction du système de TA avec un outil complet et libre

Après avoir procédé, au découpage de notre corpus bilingue post-édité en trois sous-corpus, nous avons construit notre système de TA avec l'outil *Moses*. Cet outil permet de construire un rapidement un système de traduction automatique si un corpus parallèle et un grand corpus monolingue sont disponibles pour la paire de langues souhaitée. Il dispose des scripts et programmes informatiques qui permettent de segmenter les deux corpus, et de supprimer les phrases longues et les caractères spéciaux. Cette étape est appelée la phase de prétraitement de l'outil Moses.

Après les opérations de prétraitement, *Moses* peut commencer à fabriquer la table de traductions qu'il utilisera plus tard pour le décodage. Parallèlement à cette seconde phase, nous avons utilisé d'autres outils libres tels que *SRILM*, *IRSTLM* pour construire le modèle du langage de la langue cible qui est dans notre cas le somali.

La construction de la table de traductions avec nos données bilingues de 12 863 (LDJ-fr-so-A) ou 23 568 (LDJ-fr-so-ABC) nous a pris en moyenne 16 à 20h de temps d'exécution.

Le modèle de langage a été construit en 4 heures avec les machines du laboratoire.

Avant d'utiliser le modèle de langage et le modèle de traduction de notre système, nous avons effectué des tâches supplémentaires, notamment l'optimisation de notre corpus de développement à l'aide de l'outil MERT inclus dans Moses, et la sérialisation de nos deux modèles de données.

La dernière étape à effectuer avant d'utiliser notre système de TA est son évaluation avec notre corpus de test.

III.2.2.2 Évaluation comparative des résultats de TA sur des données tests

III.2.2.2.1 Comparaison des scores TER, D_{mix} et du temps de post-édition total (T_{tpe}) des résultats de TA

Pour l'évaluation et l'estimation de la qualité de nos différents systèmes de TA, nous avons décidé de faire d'une part une comparaison des différents scores TER et Dmix entre les prétraductions fournies par chaque système de TA et notre référence à nous qui est constituée par les segments post-édités à partir des résultats de GT.

D'autre part, nous avons comparé le temps de post-édition total que nous avons passé à post-éditer les sorties des systèmes de TA et celui sur les prétraductions de GT pour un petit échantillon. Pour le reste, nous avons comparé le Tpe effectivement passé avec GT et les Tpe estimés à partir des Dmix, pour nos 4 systèmes de TA.

III.2.2.2.2 Comparaison des scores d'évaluation automatique des résultats de TA

Outre l'évaluation que nous effectuerons sur les scores TER et les temps totaux de post-édition de chaque système, nous effectuerons une comparaison entre les différents scores d'évaluation automatique avec les mesures telles que BLEU et NIST.

Comme dans la section précédente, dans cette évaluation nous prendrons les segments post-édités en somali de notre sous-corpus test comme étant traductions de référence.

Ainsi, plus les scores BLEU et NIST seront élevés, meilleures seront les prétraductions réalisées avec les différents systèmes de TA par rapport à GT.

III.2.2.3 Augmentation du corpus par apprentissage incrémental

III.2.2.3.1 Augmentation du corpus post-édité pour l'évaluation de la qualité des différents systèmes

L'une des solutions envisageables pour pallier la petitesse de notre corpus bilingue initial est de poursuivre la post-édition avec les différents systèmes de TA crées à partir de ces données.

Cette solution nous permettra d'augmenter la taille de notre corpus bilingue et par conséquent d'améliorer la qualité de traduction de nos différents systèmes de TA.

Cependant, cette solution impliquera d'avoir une communauté de contributeurs bénévoles permanents qui poursuivront leur tâche de post-édition jusqu'à l'amélioration significative des résultats de la TA.

III.2.2.3.2 L'apprentissage incrémentale pour augmenter les données bilingues

Des travaux et des études récentes comme ceux de [Wang, 2015] ont montré la possibilité d'augmenter et d'améliorer la qualité de traduction des données bilingues post-éditées en utilisant la technique d'apprentissage incrémental.

Elle consiste à augmenter la taille de la table de traductions du système de TA instantanément en y intégrant les post-éditions faites par un ou plusieurs contributeurs sur les sorties des différents systèmes de TA.

III.2.2.4 Première comparaison entre un système de TA statistique et d'un système neuronale

III.2.2.4.1 Analyse de la qualité des résultats de TA d'un système statistique sur une langue peu dotée

Selon [Koehn, 2007] il faut au moins un million de mots traduits dans une paire de langues pour construire un système de TA à l'état de l'art. Mais il est très difficile de trouver une telle quantité de données parallèles pour une paire de langues dont l'une est très faiblement dotée.

À travers les différents systèmes de TA que nous avons construits pour la paire français-somali, nous analyserons la qualité des résultats de TA et montrerons que, malgré la petite taille de nos données bilingues initiales, il est possible de construire un système de TA spécialisé pour une langue peu dotée qui répondrait au besoin d'accès en langue maternelle aux nouvelles journalistiques.

III.2.2.4.2 Analyse de la qualité des résultats de TA d'un système de traduction neuronale pour une langue peu dotée

Depuis 2015, on assiste à un développement considérable des systèmes de traduction neuronale pour les langues bien dotées comme le français, l'anglais, l'espagnol , etc. Cela a été possible du fait qu'il existe une grande quantité de données parallèles pour ces paires de langues, car ces dernières décennies beaucoup de projets de construction de ressources multilingues pour les langues européennes ou celles des autres pays développés ont permis d'accroître leurs données parallèles.

Peu de travaux ont expérimenté la traduction neuronale directe entre deux langues dont l'une est bien dotée et l'autre faiblement dotée.

Nous sommes donc un peu pionnier dans ce domaine. Pour l'instant, nos systèmes de TA neuronale (appris sur les corpus LDJ-fr-so-A et LDJ-fr-so-ABC) donnent des résultats inférieurs à ceux de nos systèmes de TA statistiques développés à partir des 2 mêmes corpus. Notre hypothèse est que cela est dû à la petitesse de notre corpus et que la situation s'inversera quand ils atteindront une taille de 2 à 10 plus grande.

III.2.3 Plan de travail

III.2.3.1 Découverte et premiers travaux sur les systèmes de TA statistique

III.2.3.1.1 Etudes sur la construction des systèmes de TA

Durant nos 3 premiers séjours doctoraux au LIG-GETALP, nous avons réalisé une étude sur la construction des systèmes de TA parallèlement à notre travail de recherche de corpus parallèles somali-anglais. Nous avons été amenés durant cette phase à comprendre et maîtriser les différents composants d'un système de TA ainsi que les outils libres utilisés dans la communauté de TA et au laboratoire pour construire rapidement un système de traduction.

Nous avons découvert durant cette même période, l'intérêt et la construction des modèles de langage avec les outils *SRILM* et *IRSTLM* qui permettent de fabriquer le modèle de la langue cible du système de TA. Pour la construction des corpus bilingues ou des modèles de traduction, nous avons étudié et découvert les outils Giza++, Hunalign etc.

Nous avons appris et compris également qu'en sus du modèle de langage et la table de traductions, il faut choisir un décodeur qui utilise les deux données précédentes pour générer la traduction de la phrase ou segment cible depuis la source. Pour cela nous avons appris

l'utilisation du décodeur le plus utilisé dans le domaine et le plus populaire, qui est inclus dans la boite à outils de TA *Moses*.

III.2.3.1.2 Première expérimentation d'un système de TA avec Moses

Après avoir découvert et bien assimilé les outils et les différentes étapes nécessaires pour construire un système de TA, nous avons effectué notre première expérimentation de TA avec MosesLIG.

Les premières données parallèles que nous avons trouvées sur le Web étaient un corpus bilingue somali-anglais que nous avons extrait de la plateforme OPUS. Il s'agissait d'un corpus d'environ 95 K segments somali-anglais dans le domaine religieux (Tanzil) et d'un corpus en réalité anglais-somali relatif à la localisation de certains logiciels libres (GNOME, UBUNTU).

À partir de ce corpus parallèle, nous avons construit un premier système de TA expérimental somali-anglais avec l'outil Moses. Les détails des données utilisées ainsi que les résultats de notre première expérimentation sont décrits dans les deux tableaux ci-dessous.

Segments	Items
22,98 M	551,9 M

Tableau 10 : Description des données du modèle de langue (ici anglais)

Segments	Items
91,51 K	4,25 M

Tableau 11 : Descriptions des données de la table de traduction (somali-anglais)

Mesure	Score
BLEU	19,35
NIST	5,66
TER	66,36
METEOR	49,77

Tableau 12 : Résultats des scores d'évaluation somali-anglais

III.2.3.1.3 Construction d'un corpus bilingue français-somali spécialisé

L'objectif initial de notre travail de recherche consistait à construire un corpus bilingue pour la TA français-somali. Vu les difficultés que nous avons rencontrées pour constituer un grand corpus parallèle pour la paire français-somali et dans le but de répondre à un besoin précis, nous avons fait le choix de construire un corpus bilingue spécialisé français-somali.

Ce dernier nous servira à mettre en place rapidement un service d'accès en somali du journal *La Nation de Djibouti* et aussi à créer un premier système de TA indépendant pour la traduction des nouvelles journalistiques.

Pour construire ce corpus bilingue, nous avons suivis les différentes étapes de notre stratégie et méthodes de construction du corpus bilingue journalistique français-somali qui sont bien décrites dans la section III.2.1 ci-dessous.

En résumé, voici les 3 étapes que nous avons suivies pour ce travail.

- Choisir et sélectionner les articles du journal à post-éditer,
- Choisir et sélectionner un système de TA pour les prétraductions,
- Utiliser une passerelle web iMAG pour la tâche de post-édition.

III.2.3.1.4 Construction et évaluation de plusieurs systèmes de TA successifs sur les données post-éditées

Malgré la petite taille de nos ressources bilingues, nous avons décidé de construire plusieurs variantes de système de traduction, pour évaluer les améliorations possibles en termes de diminution de temps de post-édition des résultats de TA.

En effet, l'objectif de notre travail n'était pas de construire un système de TA français-somali complet et de bonne qualité, mais d'utiliser nos premières données post-éditées pour poser les premières briques de la construction d'un tel système de TA, déployable pour permettre un accès satisfaisant en somali à *La Nation de Djibouti*.

Des études récentes dans la littérature [Esperança-Rodier E. & al., 2018] et des travaux de sociétés privées actives dans le domaine[42] ont expérimenté divers systèmes de TA avec des approches différentes pour d'une part comparer d'une manière objective la qualité de traduction entre ces différents systèmes en terme score BLEU, NIST et TER mais aussi faire une analyse linguistique ou subjective sur leurs résultats de traduction.

Dans notre cas, il s'agit de construire plusieurs systèmes de TA successifs sur les mêmes données bilingues avec deux outils libres de TA, l'un statistique et l'autre neuronal, afin de faire une comparaison pour déterminer le système qui nous permettra de réduire le temps de post-édition par rapport à GT.

III.2.3.2 Déroulement du travail

III.2.3.2.1 Déroulement effectif

Planning_initial_thèse 24 sept. 2018

Tâches 2

Nom	Date de début	Date de fin
Bibliographie et découverte du domaine de recherche	06/05/13	01/10/13
Rédaction article CEC-TALN	15/05/13	14/06/13
Bibliographie et documentation sur les corpus parallèles	03/03/14	29/08/14
Découverte et premiers travaux sur la TA Statistique	16/03/15	10/09/15
Construction d'un corpus bilingue français-somali spécialisé	12/10/15	31/01/16
Poursuite des post-éditions pour construire le corpus bilingue	01/06/16	19/10/16
Construction du premier systéme de TA français-somali	28/06/17	10/09/17
Construction des systémes de TA hiérarchiques, adaptés et neuronale français-somali	11/09/17	10/11/17
Enquête annotation corpus post-édité Djibouti	25/06/18	15/08/18
Rédaction manuscrit	25/12/17	05/10/18

Figure 19 : Planning des travaux effectifs

[42] https://www.tilde.com/about/news/316

III.2.3.2.2 Gantt final

Planning_initial_thèse 24 sept. 2018

Tâches 2

Nom	Date de début	Date de fin
Bibliographie et découverte du domaine de recherche	06/05/13	01/10/13
Rédaction article CEC-TALN	15/05/13	14/06/13
Bibliographie et documentation sur les corpus parallèles	03/03/14	29/08/14
Découverte et premiers travaux sur la TA Statistique	16/03/15	10/09/15
Construction d'un corpus bilingue français-somali spécialisé	12/10/15	31/01/16
Poursuite des post-éditions pour construire le corpus bilingue	01/06/16	19/10/16
Construction du premier systéme de TA français-somali	28/06/17	10/09/17
Construction des systémes de TA hiérarchiques, adaptés et neuronale français-somali	11/09/17	10/11/17
Enquête annotation corpus post-édité Djibouti	25/06/18	15/08/18
Rédaction manuscrit	25/12/17	05/10/18

Figure 20 : Planning final des travaux de thèse

Conclusion du chapitre III

Dans ce chapitre, nous avons présenté la stratégie que nous avons définie et mise en œuvre pour construire et déployer un système de TA français-somali à partir de notre premier corpus bilingue post-édité.

Pour cela, nous avons présenté également les objectifs pratiques et les questions théoriques auxquels devraient répondre notre travail, concernant le système de TA à construire et les ressources nécessaires pour son déploiement.

Enfin, nous avons présenté en détail la stratégie que nous avons suivie pour construire le corpus bilingue et nos systèmes de TA et pour les évaluer. Ces deux points seront détaillés dans les deux derniers chapitres de ce mémoire.

Chapitre IV Construction et évaluation de 2 corpus pour le somali et le français

Introduction du chapitre IV

Pour développer notre système de TA français-somali, nous avons besoin d'utiliser un corpus bilingue pour cette paire de langues. Étant donné la spécificité de notre besoin, il nous faut construire un corpus bilingue spécialisé pour les nouvelles journalistiques. Pour cela, nous avons utilisé des articles issus du journal *La Nation de Djibouti* en vue de les traduire en somali afin de constituer notre corpus parallèle. En l'absence de traducteurs bilingues, nous avons utilisé la contribution de volontaires bénévoles somalophones connaissant bien le français pour corriger les prétraductions issues de Google Translate.

Ces contributeurs ont post-édité les articles sélectionnés après que nous leur avons mis en place un service d'accès en somali à *La Nation de Djibouti*.

Nous avons utilisé la plate-forme SECTra_w/iMAG pour construire notre corpus bilingue à partir des post-édition faites par nos post-éditeurs.

Les données bilingues recueillies à l'issue de la phase de post-édition ont fait l'objet d'une évaluation avant de les utiliser pour construire un premier système de TA français-somali. L'évaluation de ce corpus post-édité a été effectuée par un groupe d'annotateurs djiboutiens bilingues qui consultent régulièrement la version numérique du journal djiboutien *La Nation de Djibouti*.

Dans la première partie de ce chapitre, nous présenterons la façon dont nous avons mis en place un service d'accès en somali au journal *La Nation de Djibouti,* pour que les locuteurs somalophones de Djibouti et d'ailleurs puissent consulter les nouvelles diffusées dans ce journal dans leur langue maternelle.

La seconde partie de ce chapitre expliquera en détails comment nous avons construit notre premier corpus bilingue en post-éditant à l'aide d'une plate-forme en ligne et libre de droits notre sélection d'articles de ce journal.

La dernière partie de ce chapitre traitera de l'évaluation subjective que nous avons réalisée sur notre corpus bilingue à l'aide d'une enquête que nous avons soumise à des annotateurs bilingues à Djibouti.

IV.1 Mise en place d'un service d'accès en somali du site Web La Nation de Djibouti

IV.1.1 Définition d'une iMAG

D'après [Boitet, Huynh et al., 2010] : « Le concept d'IMAG a été proposé par Ch. Boitet et V. Bellynck en 2006 (Boitet & al. 2008, Boitet & al.2005), et a atteint l'état de prototype en novembre 2008, avec une première démonstration sur le site Web du laboratoire LIG. Il a été adapté au site Web DSR (Digital Silk Road) en avril 2009, puis à 30 autres sites Web. Ces premiers prototypes sont des extensions du système SECTra_w (Huynh & al. 2008) de support en ligne de corpus de traductions.

Une iMAG est une passerelle interactive d'accès multilingue (interactive Multilingual Access Gateway), ressemblant beaucoup à Google Translate, à première vue : on donne une URL (site Web de départ) et une langue d'accès, et on navigue ensuite dans cette langue d'accès. Lorsque le curseur passe sur un segment (le plus souvent une phrase ou un titre), une palette montre le segment source et propose de contribuer en corrigeant le segment cible, en fait en post-éditant un résultat de TA. Avec Google Translate, la page ne change pas après la contribution, et si une autre page contient le même segment, sa traduction est toujours le résultat de TA grossière, pas la version polie post-éditée. La boîte à outils de traduction plus récente Google Translate Toolkit permet de traduire par TA et ensuite de post-éditer en ligne des pages Web complètes tirées de sites tels que Wikipedia, mais de nouveau, les segments corrigés n'apparaissent pas quand on regarde plus tard la page de Wikipedia dans la langue d'accès.

En revanche, une iMAG est dédiée à un site Web élu, ou plutôt au sous-langage élu défini par une ou plusieurs URL et leur contenu textuel. Elle contient une mémoire de traductions (MT) et un dictionnaire spécifique préterminologique (pTD), les deux dédiés au sous-langage élu. Les segments sont prétraduits non pas par un système de TA unique, mais par un ensemble (sélectionnable) de systèmes de TA. Systran et Google sont principalement utilisés aujourd'hui, mais des systèmes spécialisés développés à partir de la MT post-éditée seront également utilisés dans l'avenir.

Les plates-formes contributives puissantes SECTra_w et PIVAX (Nguyen & al. 2007) sont utilisées pour supporter les MT et les pTD. Les pages traduites sont construites avec les meilleurs traductions disponibles des segments source. Pendant la lecture d'une page traduite, il est possible non seulement de contribuer au segment sous le curseur, mais aussi de passer de façon transparente sous l'environnement de post-édition en ligne de SECTra_w, muni d'une aide dictionnairique proactive et de bonnes fonctions de filtrage et de recherche-remplacement, et ensuite de revenir dans le contexte de lecture. »

IV.1.2 Exemple de lecture et de consultation d'un article français de La Nation en somali et post-édition des pré-traductions de Google Translate

IV.1.2.1 Lecture et consultation d'un article français du journal La Nation en somali

Dans cette section, nous allons appliquer la technique iMAG pour la lecture et la consultation d'un article du journal francophone *La Nation de Djibouti* en somali.

Nous avons d'abord choisi un article parmi ceux mis en ligne par l'équipe de rédaction de ce journal[43], ensuite nous l'avons sauvé au format *HTML* avec la commande *GNU/Linux wget* dans un terminal, et enfin nous avons mis le résultat en ligne[44].

Nous y avons accédé à travers l'iMAG correspondante[45].

[43] http://www.lanationdj.com/larabie-saoudite-a-joue-un-role-essentiel-dans-la-normalisation-des-relations-entre-djibouti-et-lerythree-declare-le-president-de-la-republique-ismail-omar-guelleh/

[44] http://somalismt.imag.fr/corporaSo/LaNationdj/2015/

[45] http://service.aximag.fr/xwiki/bin/view/imag/articles-lanationdedjibouti?&depth=2&rurl=translate.google.com&sl=fr&sp=nmt4&tl=so&u=http://somalismt.imag.fr/corporaSo/LaNationdj/2015/27_09_2018_Arabie-Saoudite-role-djibouti-et-erythree-president.htm

Ensuite, nous avons choisi le somali comme langue d'accès et réglé certains paramètres. Une première vue avec les segments français pré-traduits en somali est donnée dans la Figure 21 ci-dessous.

Cette première version, vue à travers l'iMAG, est une assez mauvaise traduction automatique produite par GT en somali, mais elle permet déjà à un lecteur somalophone et non francophone du journal *La Nation de Djibouti* de comprendre de quoi parle l'article.

Figure 21 : Ecran d'une iMAG après TA avec une présentation parallèle (source-cible)

IV.1.2.2 Exemple de post-édition des prétraductions de Google Translate

Nous remarquons qu'il y a des accolades de couleur entourant les segments pré-traduits en somali. C'est une fonctionnalité activable à la demande lorsqu'on consulte pour la première fois une iMAG. Une case à cocher (en haut à droite de la fenêtre) permet de les cacher ou de les afficher.

Dans l'exemple ci-dessus, on voit que tous les segments sont encadrés en couleur rouge ; cela indique que ces segments sont des résultats « bruts » de TA, et n'ont pas encore été post-édités. Par contre, la couleur verte indiquera qu'une post-édition a été effectuée par un post-éditeur connecté, et enfin la couleur orange indiquera les segments dont la post-édition a été faite par un contributeur non connecté (sans doute mais pas nécessairement un contributeur occasionnel).

Pour effectuer la post-édition en qualité de contributeur certifié, ayant accès à l'environnement « avancé » de post-édition de SECTra_w, nous nous sommes connecté avec nos identifiants et avons choisi le mode avancé pour la post-édition.

La Figure 22 ci-dessous montre l'écran de post-édition en mode avancé, avec quelques segments post-édités par nous-même.

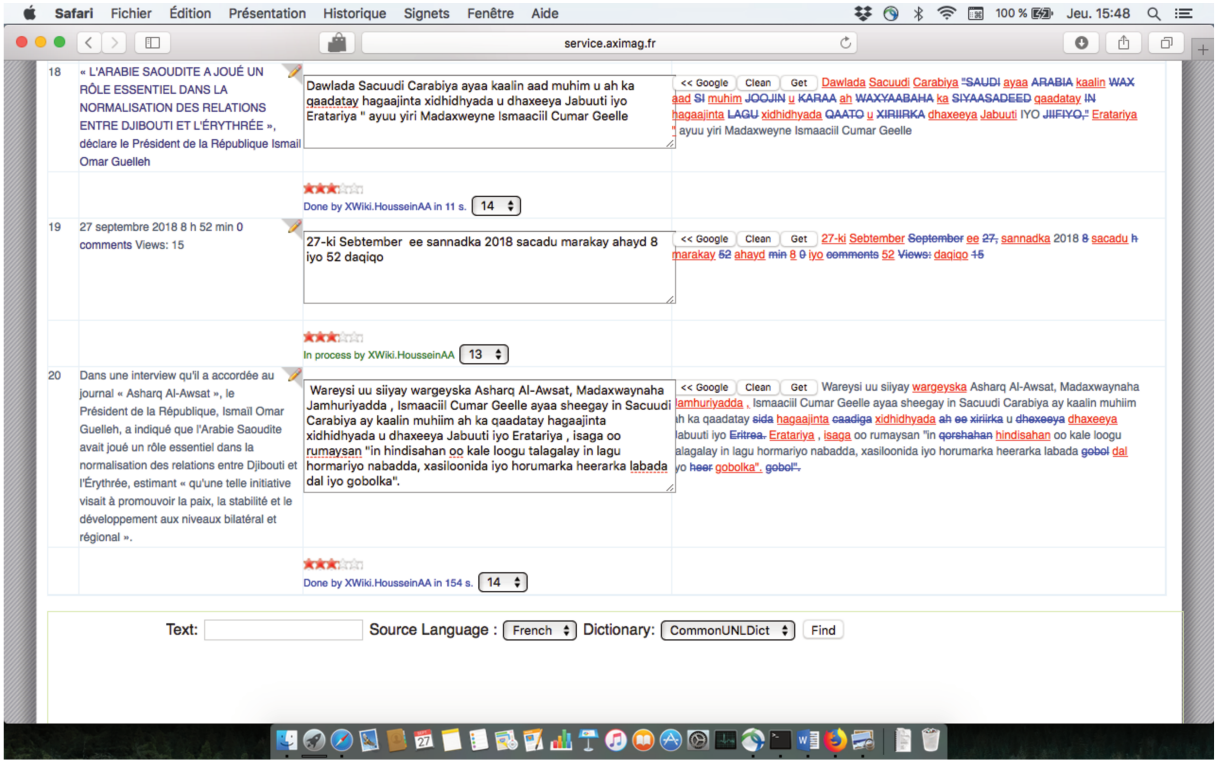

Figure 22 : Écran d'une iMAG après post-édition en mode avancé de quelques segments d'un article de La Nation de Djibouti

Le tableau ci-dessous contient les 3 segments post-édités dans la figure ci-dessus. La dernière colonne du tableau contient les traces de la post-édition, telles qu'affichées (sur demande, en cochant la case « Trace ») dans l'interface de post-édition de SECTra_w/iMAG.

Tous les mots en rouge et soulignés sont ceux ajoutés durant la post-édition, et les mots en bleu et barrés sont ceux qui ont fait l'objet d'une suppression ou d'une substitution.

Segment source	Prétraduction avec GT	Segment post-édité	Trace de la post-édition dans SECTra_w/iMAG
« L'ARABIE SAOUDITE A JOUÉ UN RÔLE ESSENTIEL DANS LA NORMALISATION DES RELATIONS ENTRE DJIBOUTI ET L'ÉRYTHRÉE », déclare le Président de la République Ismail Omar Guelleh	« SAUDI ARABIA WAX SI JOOJIN KARAA WAXYAABAHA SIYAASADEED IN LAGU QAATO XIRIIRKA IYO JIIFIYO, » ayuu yiri Madaxweyne Ismaaciil Cumar Geelle	Dawlada Sacuudi Carabiya ayaa kaalin aad muhim u ah ka qaadatay hagaajinta xidhidhyada u dhaxeeya Jabuuti iyo Eratariya « ayuu yiri Madaxweyne Ismaaciil Cumar Geelle	Dawlada Sacuudi Carabiya ~~« SAUDI~~ ayaa ~~ARABIA~~ kaalin ~~WAX~~ aad ~~SI~~ muhim ~~JOOJIN~~ u ~~KARAA~~ ah ~~WAXYAABAHA~~ ka ~~SIYAASADEED~~ qaadatay ~~IN~~ hagaajinta ~~LAGU~~ xidhidhyada ~~QAATO~~ u ~~XIRIIRKA~~ dhaxeeya Jabuuti IYO ~~JIIFIYO, »~~ Eratariya « ayuu yiri Madaxweyne Ismaaciil Cumar Geelle
27 septembre 2018 8 h 52 min 0 comments Views 15	September 27, 2018 8 h 52 min 0 comments Views : 15	27-ki Sebtember ee sannadka 2018 sacadu marakay ahayd 8 iyo 52	27-ki Sebtember ~~September~~ ee ~~27,~~ sannadka 2018 ~~8~~ sacadu ~~h~~ marakay ~~52~~

Segment source	Prétraduction avec GT	Segment post-édité	Trace de la post-édition dans SECTra_w/iMAG
		daqiqo	ahayd ~~min~~ 8 ~~0~~ iyo ~~comments~~ 52 ~~Views :~~ daqiqo ~~15~~
Dans une interview qu'il a accordée au journal « Asharq Al-Awsat », le Président de la République, Ismaïl Omar Guelleh, a indiqué que l'Arabie Saoudite avait joué un rôle essentiel dans la normalisation des relations entre Djibouti et l'Érythrée, estimant « qu'une telle initiative visait à promouvoir la paix, la stabilité et le développement aux niveaux bilatéral et régional »	Wareysi uu siiyay Asharq Al-Awsat, Madaxwaynaha Ismaaciil Cumar Geelle ayaa sheegay in Sacuudi Carabiya ay kaalin muhiim ah ka qaadatay sida caadiga ah ee xiriirka u dhexeeya Jabuuti iyo Eritrea., oo rumaysan « in qorshahan oo kale loogu talagalay in lagu hormariyo nabadda, xasiloonida iyo horumarka heerarka labada gobol iyo heer gobol ».	Wareysi uu siiyay wargeyska Asharq Al-Awsat, Madaxwaynaha Jamhuriyadda, Ismaaciil Cumar Geelle ayaa sheegay in Sacuudi Carabiya ay kaalin muhiim ah ka qaadatay hagaajinta xidhidhyada u dhaxeeya Jabuuti iyo Eratariya, isaga oo rumaysan « in hindisahan oo kale loogu talagalay in lagu hormariyo nabadda, xasiloonida iyo horumarka heerarka labada dal iyo gobolka ».	Wareysi uu siiyay wargeyska Asharq Al-Awsat, Madaxwaynaha Jamhuriyadda, Ismaaciil Cumar Geelle ayaa sheegay in Sacuudi Carabiya ay kaalin muhiim ah ka qaadatay ~~sida~~ hagaajinta ~~caadiga~~ xidhidhyada ~~ah ee xiriirka~~ u ~~dhexeeya~~ dhaxeeya Jabuuti iyo ~~Eritrea.~~ Eratariya, isaga oo rumaysan « in ~~qorshahan~~ hindisahan oo kale loogu talagalay in lagu hormariyo nabadda, xasiloonida iyo horumarka heerarka labada ~~gobol~~ dal iyo ~~heer~~ gobolka ». ~~gobol ».~~

Figure 23 : Extrait des 3 segments post-édités dans l'iMAG

Dans la figure ci-dessous, nous avons nous-même post-édité les prétraductions de GT afin de mettre en évidence les différents cas de post-édition possibles, que nous avons rencontré durant la phase de construction de corpus bilingue post-édité avec SECTRA_W/IMAG.

Résumons ci-dessous ces différents cas de figure à partir de la post-édition des 3 segments ci-dessus :

- Le premier cas consiste à retraduire totalement le segment prétraduit par GT à cause de sa mauvaise qualité de TA. Dans la figure ci-dessus le segment 1 illustre ce cas.
- Le second cas consiste à effectuer des modifications mineures (insertions, suppressions ou substitutions) sur le segment prétraduit. Dans l'exemple ci-dessus, le segment 3 a fait l'objet de quelques modifications. Par exemple, nous avons ajouté des mots manquants dans la prétraduction de GT, comme Jamhuriyadda ; nous avons remplacé l'expression « sida caadiga ah ee xiriirka » par « hagaajinta xidhidhyada ». Enfin, nous avons ajouté l'expression « labada dal » pour la traduction du mot « bilatéral ».
- Le troisième cas consiste à garder la prétraduction de GT telle qu'elle est, car cela correspond au souhait du post-éditeur et sa qualité de traduction est très bonne par rapport au segment source. Généralement, ce cas est très fréquent pour les segments d'un ou deux mots et que GT a déjà prétraduits et qui se trouvent dans la mémoire de traductions de SECTRA_W/IMAG.

IV.2 Construction d'un corpus bilingue par post-édition avec la plate-forme SECTra_w/iMAG

IV.2.1 La plate-forme SECTRA/IMAG, un outil pour construire des corpus bilingues

Dans la section ci-dessous, nous allons faire une description des travaux antérieurs qui ont utilisé la plateforme SECTRA_W/IMAG pour construire des corpus parallèles spécialisés pour plusieurs paires de langues différentes.

L'objectif de cette description est de montrer la pertinence de notre stratégie de construction d'un corpus bilingue spécialisée en utilisant un ou plusieurs moteurs de prétraductions pour une post-édition avec la plateforme SECTRA_W/IMAG.

Ces travaux vont de la simple traduction d'une œuvre littéraire anglaise vers le français à la construction d'un corpus bilingue spécialisé pour développer des systèmes de TA français-chinois en passant par la constitution d'un corpus multilingue du chapitre d'un livre scientifique.

IV.2.1.1 Brève description de travaux sur les corpus bilingues construits avec SECTra_w/iMAG

IV.2.1.1.1 Corpus bilingue français-chinois spécialisé

Dans le cadre du projet MACAU mené au sein de l'université de Grenoble à partir de 2012, et dont le but était de fournir un accès multilingue aux cours dispensés à l'université pour les étudiants étrangers, [Kalitvianski et al., 2015] ont construit un corpus aligné d'environ 10 000 segments français-chinois en faisant post-éditer par un groupe d'étudiants chinois des cours d'informatique de niveau licence et master à l'aide de la plate-forme SECTra_w/iMAG.

Les deux figures ci-dessous sont respectivement un exemple du résultat de post-édition d'un chapitre du cours de Complexité Calculatoire et la description des différents thèmes et contenus des cours traduits et post-édités durant ce travail.

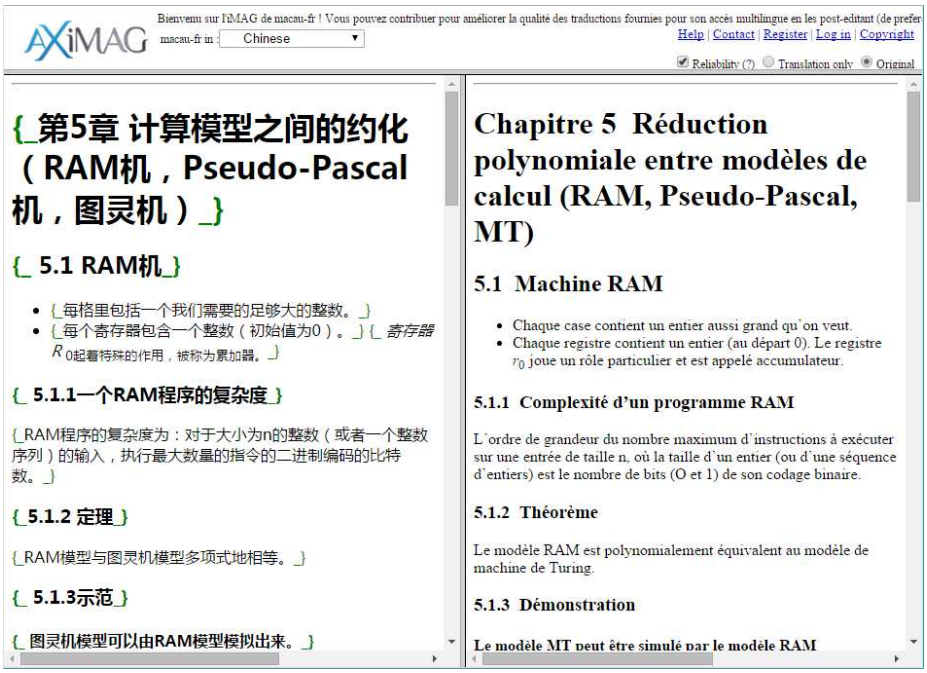

Figure 24 : Chapitre d'un cours de Complexité Calculatoire post-édité avec SECTra_w/iMAG (source : [Kalitvianski et al., 2015])

Subject matter	Content type	Pages (html)	Available Translations
Introduction to Propositional and First-Order Logic	Full book	45	Chinese (full) English (partial) Russian (partial)
Computational Complexity	Lecture notes	13	Chinese (full)
Human-Machine interaction	Teacher lectures	7	Chinese (full)
Formal Languages and Parsing	Teacher lectures, hand-outs	5	Russian (partial)
Modelling of digital systems	Exam paper	2	Chinese (full)
AI and automatic planning	Exam paper	2	Chinese (full)
Introduction to Ergonomics	Student report	1	Chinese (full)

Figure 25 : Description des thèmes et contenus des données du projet MACAU (source : [Kalitvianski et al., 2015])

IV.2.1.1.2 Corpus multilingue spécialisé du chapitre d'un livre

Pour évaluer l'importance de la proximité terminologique entre les langues et dans l'optique de répondre aux besoins des étudiants étrangers pour accéder dans leurs langues aux contenus des documents pédagogiques, une expérience impliquant un groupe de doctorants de 8 langues maternelles[46] différentes a été menée au laboratoire d'informatique de Grenoble à partir du printemps 2013, durant 9 mois, à l'aide la plate-forme SECTra_w/iMAG [Shah R. et al., 2015].

Le contenu du vingt-unième chapitre du livre BEMbook qui traite en langue anglaise du bio-électro-magnétisme, soit environ 18 pages et 4 420 mots, a été choisi pour être traduit et post-édité dans les langues des participants à l'expérience.

Les deux figures ci-dessous présentent respectivement un exemple de post-édition en marathi de quelques segments de ce chapitre, et un graphique sur l'évolution du temps de post-édition par page standard du chapitre 21 du BEMbook sur les 7 langues utilisées finalement durant cette expérience.

[46] Les 8 langues en question étaient : portugais, japonais, russe, espagnol, bengali, hindi, marathi et malayalam.

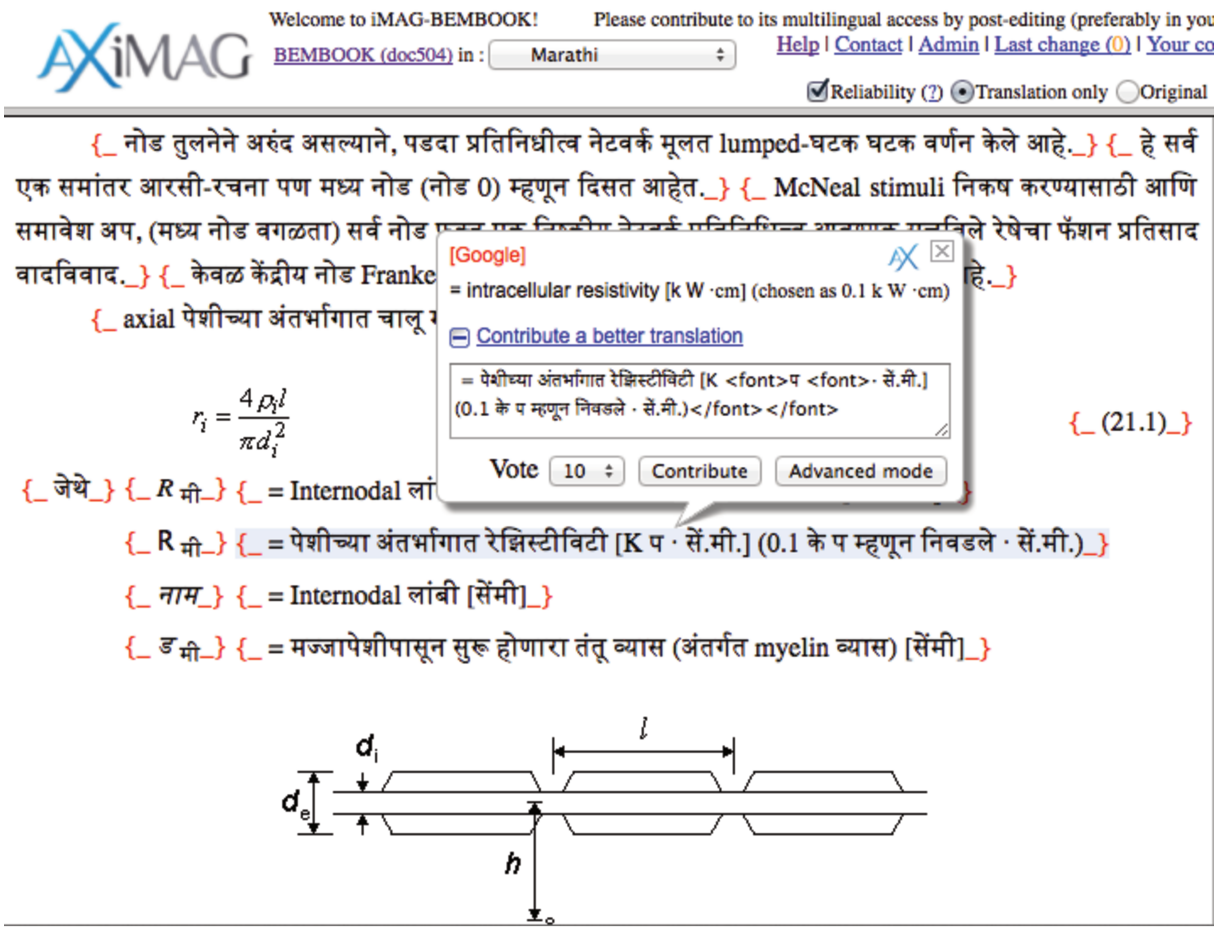

Figure 26 : Écran de post-édition en marathi du chapitre 21 du BEMbook (source : [Shah R. et al., 2015])

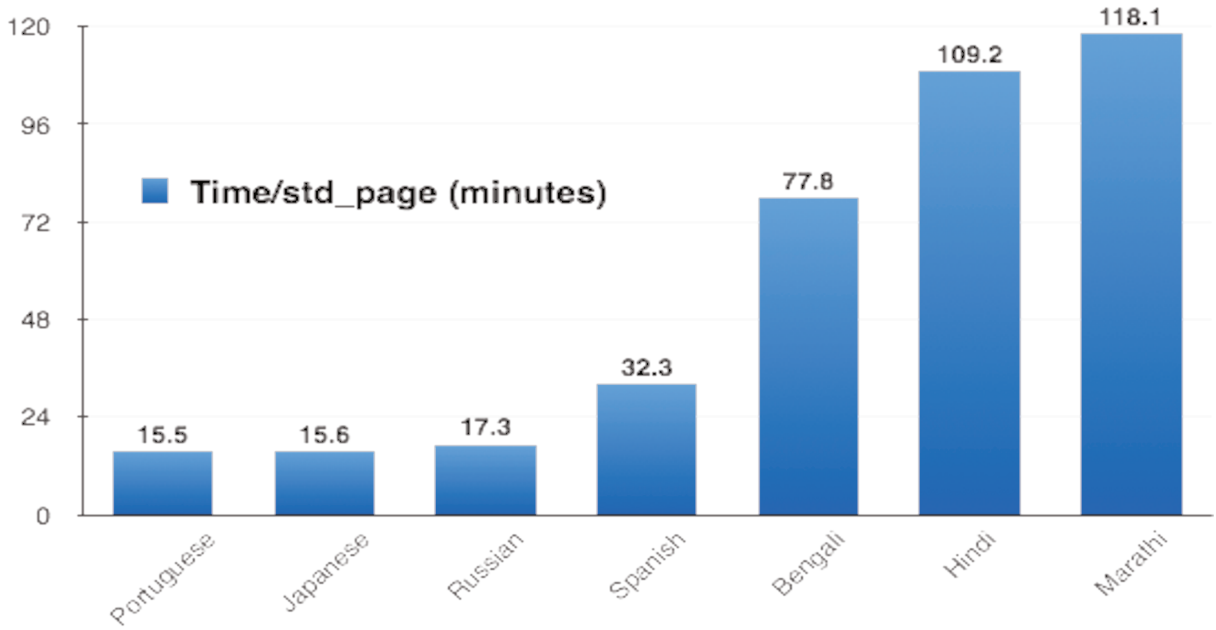

Figure 27 : Graphique de l'évolution du temps de post-édition/page standard des 7 langues (source : [Shah R. et al., 2015])

IV.2.1.1.3 Corpus bilingue anglais-français d'une œuvre littéraire

Une première expérience de traduction automatisée d'une œuvre littéraire utilisant la post-édition pour améliorer la qualité des prétraductions issues du système de TA MosesLIG-fr-en [Besacier & al, 2012] du Laboratoire d'Informatique de Grenoble (LIG) a été menée par [Besacier L., 2014] et plusieurs membres du GETALP, son équipe de recherche.

L'œuvre traduite durant cette expérience est l'essai en anglais intitulé « *The Book of Me* » de l'écrivain américain Richard Powers. Elle est composée de 545 segments et 10 731 mots, et a été divisée en trois blocs de même longueur, traités successivement.

Le tableau ci-dessous résume les nombres de mots en anglais, traduits et post-édités en français durant chaque itération (ou étape).

Itération (nb. seg)	Anglais (nb. mots)	TA français (nb.mots)	PE français (nb.mots
It.1 (184)	3 593	4 295	4 013
It. 2 (185)	3 729	4 593	4 202
It. 3 (176)	3 409	4 429	3 912
Total (545)	10 731	13 317	12 127

Tableau 13 : Corpus source, cible, traduit et corrigé (Source : [Besacier L., 2014])

D'après [Besacier L., 2014] : « Les résultats issus d'un pipeline TA+PE ont été présentés et, pour aller au-delà, les avis d'un panel de lecteurs et d'un traducteur ont été sollicités. Le texte traduit, obtenu après 25h de travail humain, est jugé acceptable par les lecteurs, mais l'avis du traducteur professionnel reste mitigé.

Cette approche suggère une méthodologie de traduction rapide et « low cost », analogue aux traductions de sous-titres de séries TV trouvées sur le Web. Pour l'auteur, c'est la possibilité d'avoir son œuvre traduite dans un plus grand nombre de langues (plusieurs dizaines au lieu d'une poignée – cet essai de R. Powers a d'ailleurs aussi été traduit en roumain avec la même méthodologie).

Mais celui-ci est-il prêt à sacrifier la qualité de traduction (et son contrôle sur celle-ci) au prix d'une diffusion plus large de ses œuvres ? ».

IV.2.2 Construction du premier corpus bilingue français-somali de qualité par post-édition de Google Translate

IV.2.2.1 Difficultés de trouver des corpus bilingues pour les langues peu dotées

Pour entraîner un premier système TA français-somali, nous avons besoin d'un corpus bilingue qui nous permettra de construire les données d'apprentissage de notre système de traduction.

Or, comme dit précédemment, il n'existe pratiquement aucun corpus bilingue français-somali pour le sous-langage sur lequel nous avons choisi d'expérimenter, celui des nouvelles journalistiques.

Au cours de notre recherche, nous avons seulement pu trouver un petit corpus bilingue d'environ 7000 segments, issu du projet OPUS [Jörg Tiedmann, 2012], mais il est dans le domaine religieux pour environ 80%[47], et le reste est dans le domaine de la localisation des

[47] Dans ce domaine, il s'agit de la mise en parallèle de traductions à partir de l'arabe (arabe-français et somali).

logiciels libres[48]. Dans ces conditions, il nous a fallu trouver une méthode efficace pour construire un corpus bilingue pour notre système de TA, formé de vraies (et bonnes) traductions français-somali.

Le problème posé par la construction de corpus parallèles pour les langues peu dotées a été traité dans plusieurs études ces dernières années. Ainsi, [Do D., 2011] a réussi à construire un corpus parallèle de plus de 50 000 segments français-vietnamien à travers l'extraction de corpus comparables récupérés sur le web à partir des dépêches d'agence de presse publiant gratuitement en français et en vietnamien.

Malheureusement, dans notre cas, il a été impossible de trouver un seul corpus comparable français-somali, et encore moins un corpus parallèle français-somali dans le sous-langage des nouvelles journalistiques.

Nous avons adopté une méthode utilisée par plusieurs chercheurs, comme [Wang, 2015] et d'autres[49], qui consiste à créer un corpus bilingue spécialisé à partir des post-éditions des prétraductions fournies par un ou plusieurs systèmes de TA généralistes existants, la post-édition (correction) étant faite manuellement dans un environnement libre et dédié, ce qui permet de gagner beaucoup de temps par rapport à une traduction humaine professionnelle, qui serait de toutes façons bien trop coûteuse pour notre cas.

Cette méthode nous a paru la plus efficace car, depuis août 2013, le somali fait partie des nouvelles langues africaines que Google Translate (GT) propose dans son système de TA en ligne.

IV.2.2.2 Un corpus bilingue spécialisé par post-édition des prétraductions

IV.2.2.2.1 Présentation du journal La Nation de Djibouti

Le journal *La Nation de Djibouti* est le seul organe de presse écrite officielle et francophone en République de Djibouti. Il a été fondé en 1977 juste après que l'ancien Territoire Français des Afars et des Issas (TFAI) est devenu un état indépendant le 27 juin 1977.

C'est le troisième journal d'expression française à Djibouti puisqu'il a été précédé durant la colonisation française du territoire des deux journaux *Djibouti Français* et *Réveil de Djibouti* (1941-1977) [Leroux, R., 1998].

Ce dernier a été créé dans le contexte de la seconde guerre mondiale en janvier 1941, lorsque l'administration française en place à Djibouti a rejoint le camp des alliés et de la résistance. Après le départ de l'administration française, le *Réveil de Djibouti* a été nationalisé et rebaptisé *La Nation de Djibouti,* et mis sous la tutelle du secrétariat général à l'information, et plus récemment du ministère chargé de la communication de la République de Djibouti.

Comme déjà dit plus haut, *La Nation de Djibouti* est l'un deux organes de presse existant à Djibouti, l'autre étant le journal arabophone *Al Qarn*. *La Nation de Djibouti* publie quotidiennement des informations à caractère économique, politique, social et international, destinées aux lecteurs djiboutiens francophone.

La Nation de Djibouti dispose d'un site Web[50] où sont mis en ligne les principaux articles saillants de la version papier du journal. Outre sa fonction de publication journalistique, *La*

[48] Dans ce domaine, il s'agit de la mise en parallèle de traductions à partir de l'anglais (anglais-français et somali)

[49] http://talaf.imag.fr/2016/Actes/ABDOURAHAMANE_ET_AL - Construction d'un corpus parallèle français-comorien en utilisant de la TA français-swahili.pdf

[50] www.lanationdj.com

Nation de Djibouti diffuse également des annonces publicitaires, des offres d'emploi et des appels à participation à des marchés publics. Ce journal est tiré à 3 000 exemplaires chaque jour et est vendu dans les différents kiosques de la capitale à raison de 150 FDJ (soit l'équivalent de 0,74 euros)[51] par exemplaire.

IV.2.2.2.2 Utilisation de SECTra_w/iMAG pour post-éditer les prétraductions de Google

Nous avons utilisé l'outil SECTRA_W/IMAG [Huynh et al., 2008] développé dans notre laboratoire pour construire notre corpus bilingue à partir des post-édition des prétraductions produites par GT. Selon le concepteur de la plate-forme SECTRA_W/IMAG [Huynh et al., 2008], cette dernière est un service web qui vise de façon générale à permettre l'exploitation collaborative sur le web de corpus de traductions multilingues, multi-annotés et multimédia.

Dans cette plate-forme, les post-éditeurs commencent d'abord par charger un ou plusieurs documents qui ont été préalablement mis dans le système par l'administrateur ou l'initiateur de tout projet de post-édition (PE). La plate-forme contient des outils pour segmenter chaque document à post-éditer en un « fichier squelette » contenant les références aux *segments textuels* du document (il s'agit de phrases ou de segments), qu'on désire traduire, et les *segments de présentation* (typiquement, du code html, css, js, etc.). Les segments textuels sont stockés dans une mémoire de traductions (MT), qui est initialisée avec les prétraductions automatiques (en abrégé les TA) produites par le ou les différents systèmes de TA qui ont été choisis pour la tâche de post-édition. Les post-éditions (en abrégé PE) associées à un même segment source (dans une ou plusieurs langues cible) sont aussi stockées dans la MT.

Dans notre cas, nous avons d'abord sélectionné une dizaine d'articles issus du journal en ligne LaNationdj.com[52], et créé un lien web à partir duquel la plate-forme peut récupérer le contenu de chaque article à post-éditer.

La post-édition que nous avons effectuée sur notre corpus de travail est une intervention humaine d'une ou plusieurs personnes maîtrisant parfaitement le français et le somali ; elle a consisté à corriger les prétraductions proposées par Google Translate.

Le mode opératoire le plus utilisé pour cette tâche est le *mode avancé*, visible dans la Figure 28 ci-dessous. Comme on le voit dans la figure, les segments source, la zone de post-édition et la MT, initialisée avec le contenu des prétraductions de GT, sont présentés à l'utilisateur dans un tableau à trois colonnes, avec une ligne par segment source.

Dès que l'utilisateur met son curseur dans la zone de post-édition, son temps de post-édition est incrémenté, jusqu'à ce qu'il finisse ses modifications. Ce temps est mis à jour à la fin de chaque post-édition. On peut voir juste au-dessous de la zone de post-édition le nombre de secondes qu'on a passé à éditer le texte contenu dans la cellule de PE. Ce temps est appelé *temps primaire de post-édition, Tpe_1*.

Le *temps secondaire de post-édition, Tpe_2,* est le reste du temps passé à post-éditer. Le *temps total de post-édition est T_pe = Tpe_1 + Tpe_2*. Tpe_2 est le temps pendant lequel on lit le segment source à post-éditer, on accède à des dictionnaires, et on consulte éventuellement le contenu de la MT pour voir s'il n'y aurait pas de meilleure suggestion que la post-édition ou le résultat de TA avec lequel la cellule de post-édition a été initialisée.

Dans la 3° colonne, présentant les TA (au plus une par système de TA) et les PE (éventuellement plusieurs), on peut afficher la trace des modifications permettant de passer de chaque élément de la MT (TA ou PE) au contenu de la cellule de post-édition, en cliquant sur

[51] https://www.xe.com/fr/currencyconverter/convert/?Amount=150&From=DJF&To=EUR

[52] http://somalismt.imag.fr/corporaSo/LaNationdj/2015/

le bouton « Trace » de l'interface. Les insertions sont en rouge et les suppressions en bleu barré. Il s'agit en fait d'une reconstruction de la trace de l'algorithme de Wagner et Fischer (1974), dans laquelle on remplace N échanges par N suppressions suivies de N insertions.

Figure 28 : Capture d'écran de l'interface de post-édition en mode avancé

Pour notre future évaluation de la qualité de notre corpus bilingue post-édité, nous avons également relevé, à chaque fois qu'on post-éditait un document, le temps total Tpe passé pour effectuer cette PE. En divisant ce temps par le nombre de pages standard (de 250 mots) du document, nous avons pu obtenu le temps de post-édition par page standard.

Au total, nous avons post-édité 100 documents, et produit 10 669 segments bilingues (98 912 mots ou environ 400 pages standard) constituant notre premier corpus bilingue. Ce premier corpus parallèle nous a permis d'effectuer une expérimentation, à l'aide de Moses [Hoang, 2007], d'un premier système TA français-somali.

À partir de ce corpus bilingue de qualité mais relativement petit, nous avons construit un système de TA statistique (en Moses à base de fragments), et un système de TA neuronale (en OpenNMT), les deux très *spécialisés* au sous-langage des nouvelles journalistiques de *La Nation de Djibouti.* Nous avons aussi utilisé un gros corpus monolingue du somali pour construire le modèle de langage nécessaire à la méthode de TA statistique.

Nous avons aussi construit deux systèmes *étendus* (avec Moses et avec OpenNMT) en ajoutant à notre corpus d'apprentissage d'autres corpus parallèles français-somali, qui ne sont pas relatifs au même sous-langage que notre corpus initial, mais ont permis de doubler la taille du corpus, et d'améliorer (très légèrement) les résultats. Cela sera présenté en détail au chapitre V. Avant cela, il nous faut prouver la qualité de notre corpus post-édité, appelé LDJ-fr-so-A (A pour « de très bonne qualité »).

IV.3 Analyse et évaluation de la qualité du corpus post-édité LDJ-fr-so-A

IV.3.1 Caractéristiques du corpus bilingue post-édité

Le corpus bilingue post-édité que nous avons construit contient près de 10 669 segments français-somali. Il a été constitué à partir de la post-édition des prétraductions de 100 articles du journal *La Nation de Djibouti*. Le Tableau 14 ci-dessous présente les caractéristiques et les informations recueillies sur ce corpus.

Nombre de segments (source/cible)	10 669
Nombre de mots (source/cible)	98 912
Nombre de documents post-édités (source/cible)	100
Nombre de pages standard post-éditées	395,6
Temps de post-édition moyenne en minute / page standard	17

Tableau 14 : Caractéristiques du corpus bilingue post-édité

IV.3.2 Analyse du temps de post-édition par page standard du corpus LDJ-fr-so-A

IV.3.2.1 Quelques définitions

Le *temps primaire de post-édition (Tpe_1)* et le *temps secondaire de post-édition (Tpe_2)* ont été définis ci-dessus.

Le *temps total de post-édition* est la somme des deux : *Tpe = Tpe_1 + Tpe_2*.

Le *temps de post-édition par page standard* (*Tpe/p.std)*[53] est le temps moyen de post-édition d'une page standard en langue source (250 mots ou 1400 signes pour le français, 400 caractères pour le chinois ou le japonais).

IV.3.2.2 Evolution du temps de post-édition des segments post-édités

Le tableau ci-dessous montre l'évolution du temps de post-édition total (Tpe) et par page standard (Tpe/p.std) du corpus post-édité.

Numéro du document	Nombre de segments	Temps en mn (Tpe total)	Nombre de mots	Nombre de pages	Temps par page (Tpe/p.std) en minutes
1	91	29	606	2,4	12
2	113	61	1022	4,1	15
3	95	44	770	3,1	14
4	97	45	747	3,0	15
5	98	48	918	3,7	13
6	96	50	777	3,1	16
7	88	16	505	2,0	8
8	88	23	623	2,5	9
9	100	40	859	3,4	12
10	91	24	629	2,5	10

[53] Le *Tpe/p.std* est usuellement exprimé en minutes, et calculé en divisant le temps de post-édition total par le nombre de pages standard de 250 mots du document post-édité. Par exemple, à la ligne 1 du tableau 15 : *Tpe/p.std* = 29/2,4 ≈12 minutes.

Numéro du document	Nombre de segments	Temps en mn (Tpe total)	Nombre de mots	Nombre de pages	Temps par page (Tpe/p.std) en minutes
11	92	31	657	2,6	12
12	89	31	634	2,5	12
13	148	141	1689	6,8	21
14	99	83	842	3,4	25
15	90	25	665	2,7	9
16	93	36	644	2,6	14
17	116	76	1053	4,2	18
18	96	34	808	3,2	11
19	96	36	883	3,5	10
20	109	70	1119	4,5	16
21	97	44	783	3,1	14
22	103	67	1132	4,5	15
23	106	90	1024	4,1	22
24	97	60	740	3,0	20
25	96	31	785	3,1	10
26	102	47	901	3,6	13
27	87	35	564	2,3	16
28	91	51	709	2,8	18
29	90	31	591	2,4	13
30	96	46	660	2,6	17
31	100	57	876	3,5	16
32	93	29	557	2,2	13
33	138	131	1 465	5,9	22
34	114	112	1 196	4,8	23
35	98	46	762	3,0	15
36	108	48	928	3,7	13
37	104	60	745	3,0	20
38	98	47	829	3,3	14
39	96	56	705	2,8	20
40	94	31	747	3,0	10
41	100	47	890	3,6	13
42	89	25	563	2,3	11
43	86	30	633	2,5	12
44	103	72	839	3,4	21
45	100	49	851	3,4	14
46	126	82	1 405	5,6	15
47	92	40	643	2,6	16
48	90	20	615	2,5	8
49	100	34	716	2,9	12
50	107	62	1 036	4,1	15
51	103	52	947	3,8	14
52	112	64	1 309	5,2	12
53	100	76	943	3,8	20
54	101	69	887	3,5	19
55	102	57	939	3,8	15
56	184	189	2 170	8,7	22
57	131	136	1 679	6,7	20

Numéro du document	Nombre de segments	Temps en mn (Tpe total)	Nombre de mots	Nombre de pages	Temps par page (Tpe/p.std) en minutes
58	173	267	2 601	10,4	26
59	94	72	913	3,7	20
60	91	41	686	2,7	15
61	99	69	632	2,5	27
62	106	67	774	3,1	22
63	89	23	589	2,4	10
65	88	21	598	2,4	9
66	113	61	1095	4,4	14
67	91	31	692	2,8	11
68	96	53	689	2,8	19
69	91	39	627	2,5	16
70	91	38	712	2,8	13
71	102	53	928	3,7	14
72	97	52	780	3,1	17
73	107	67	1 153	4,6	15
74	116	127	1 410	5,6	23
75	136	154	1 164	4,7	33
76	111	92	1 140	4,6	20
77	100	48	779	3,1	15
78	116	90	1 249	5,0	18
78	95	60	662	2,6	23
79	90	49	669	2,7	18
80	91	33	704	2,8	12
81	90	30	718	2,9	10
82	95	28	678	2,7	10
83	98	58	688	2,8	21
84	108	84	1 016	4,1	21
85	125	138	1 610	6,4	21
86	156	239	2 069	8,3	29
87	156	180	2 599	10,4	17
88	136	177	1 578	6,3	28
89	115	98	1 102	4,4	22
90	95	42	804	3,2	13
91	96	43	909	3,6	12
92	110	95	999	4,0	24
93	113	67	1 074	4,3	16
94	196	229	3 011	12,0	19
95	128	130	1 524	6,1	21
96	126	123	1 315	5,3	23
97	239	268	3 374	13,5	20
98	100	39	889	3,6	11
99	90	30	654	2,6	11
100	95	37	841	3,4	11
	10 669	6 908	9 8912	395,6	17

Tableau 15 : Évolution du temps total de PE pour 100 articles (400 pages)

IV.3.2.3 Analyse comparative du lien entre les scores TER et du temps de post-édition

Pour évaluer la qualité du corpus bilingue construit à l'aide de la plate-forme SECTRA_W/IMAG et la comparer avec celle des prétraductions produites par GT, nous avons calculé pour chaque document post-édité son nombre de segments, son nombre de mots, le temps passé pour la post-édition de chaque segment, etc.

Cette première évaluation permet de montrer l'évolution du temps de PE et le temps moyen nécessaire pour post-éditer un document. Le but de cette évaluation était d'estimer le temps minimal nécessaire pour post-éditer un document, en faisant l'hypothèse que ce temps pourrait donner des indications sur la qualité de la mémoire de traductions construite à partir des données post-éditées jusqu'à ce jour.

Dès sa version originale, SECTRA_W a permis d'organiser des campagnes d'évaluation de systèmes de TA. D'après le concepteur de la plate-forme [C. Phap Huynh, 2010], pour effectuer une compagne d'évaluation, les organisateurs doivent charger un corpus d'évaluation formé de deux fichiers contenant les segments source et leurs prétraductions à évaluer, et d'un ou plusieurs fichiers de traductions de référence.

Cependant, cette fonctionnalité a été délaissée depuis 2007 au profit d'opérations de traduction de sites Web (par exemple, le site B@bel de l'Unesco) ou de gros documents en ligne (comme EOLSS, Encyclopedia Of Life Support Systems, et les thèses de Lingxiao Wang, Ying Zhang et Ruslan Kalitviansi), ou de sites Web (environ 80 iMAG, la plupart de démonstration), ou encore de construction de corpus parallèles de qualité pour le développement de systèmes de TA empiriques (statistiques ou neuronaux). Nous ne l'avons pas utilisée.

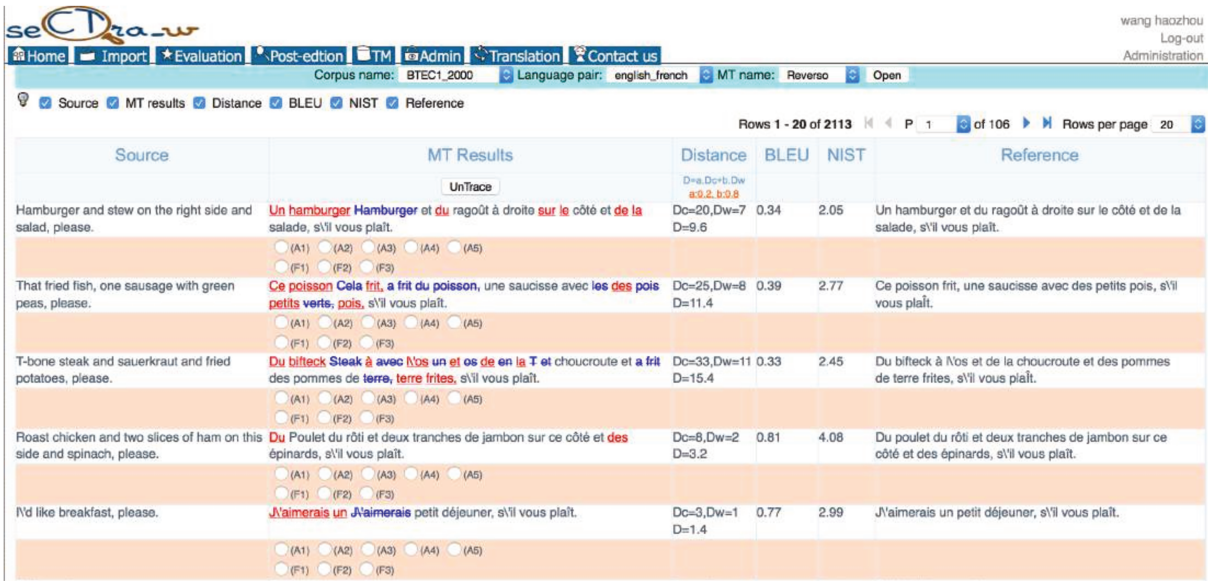

Figure 29 : Capture d'écran de l'interface d'évaluation SECTRA_w

IV.3.3 Auto-notation et évaluation du corpus par des annotateurs bilingues

IV.3.3.1 Protocole d'évaluation

IV.3.3.1.1 Caractéristiques du corpus à annoter

Pour évaluer la qualité de notre corpus post-édité, nous avons réalisé une enquête d'auto-notation sur un article choisi parmi ceux que nous avions déjà post-édités à l'aide de SECTRA_W/IMAG.

Nous avons sélectionné 54 segments de cet article, qui traitait de la question du logement social et décent à Djibouti. Voici les caractéristiques de cet échantillon, qui a fait l'objet d'une évaluation par 9 annotateurs bilingues.

Nombre de segments (source/cible)	54/54
Nombre de mots (source/cible)	431/377

Tableau 16 : Description des segments annotés

IV.3.3.1.2 Profils et types d'annotateurs

Après avoir sélectionné les 54 segments à annoter, nous avons préparé un document décrivant les différentes étapes et tâches que les annotateurs devaient suivre pour réaliser cette enquête. Il fallait que chaque annotateur :

- renseigne son nom complet et son adresse électronique,
- pour chaque segment post-édité, entoure la note attribuée précédemment si elle lui convenait, sinon la biffe et en met une nouvelle,
- éventuellement, corrige la traduction du texte post-édité en l'annotant directement sur le document d'évaluation,
- à la fin de l'évaluation, indique le temps passé (en minutes) à ce travail,
- enfin rende les pages annotées.

Les deux tableaux ci-dessous décrivent les profils des 9 annotateurs bilingues qui ont répondu positivement à notre proposition.

Dénomination	**Profession**	**Niveau d'études**
AIA	Enseignant-chercheur	Universitaire
AWD	Ingénieur	Universitaire
IAH	Enseignant-chercheur	Universitaire
SGD	Enseignant	Universitaire
AMR	Enseignant-chercheur	Universitaire
SAW	Agent Comptable	Universitaire
MAE	Conseiller pédagogique	Universitaire
SDH	Agent Comptable	Universitaire
MAA	Instituteur	Secondaire

Tableau 17 : Profils des annotateurs bilingues

Âge	**Annotateurs**
[30-40]	6
[40-50]	1
[50-60]	2

Tableau 18 : Répartition en âge des annotateurs bilingues

IV.3.3.2 Analyse du résultat d'auto-notation par des juges bilingues

IV.3.3.2.1 Récapitulatif des résultats de notation des juges bilingues sur les 54 segments d'évaluation

Le tableau ci-dessous donne les notes données par les 9 juges bilingues aux 54 segments de l'évaluation.

Numéro segment	Note EV1	Note EV2	Note EV3	Note EV4	Note EV5	Note EV6	Note EV7	Note EV8	Note EV9
1	14	20	14	14	14	14	17	14	14
2	15	20	15	15	17	15	13	15	15
3	15	20	15	15	17	15	15	15	5
4	12	20	14	14	15	14	14	16	5
5	14	20	14	14	14	14	14	14	5
6	15	15	12	15	15	13	12	15	10
7	15	12	15	15	15	15	15	15	14
8	14	16	14	14	14	14	17	14	14
9	15	20	11	15	15	13	10	15	5
10	15	18	15	15	15	15	11	15	5
11	15	20	15	15	15	15	18	15	15
12	15	20	15	15	15	13	17	15	15
13	14	20	14	14	14	14	18	15	14
14	14	14	14	14	14	11	17	14	14
15	14	14	14	14	14	14	16	14	14
16	14	14	14	14	14	12	10	14	14
17	14	14	14	14	14	14	17	14	14
18	14	14	14	14	14	14	18	14	5
19	14	14	14	14	14	13	10	14	14
20	16	16	16	16	16	16	16	16	12
21	16	12	15	16	16	16	10	16	12
22	15	15	13	15	15	15	15	15	15
23	15	15	13	15	15	15	15	15	12
24	15	17	15	15	15	15	10	15	15
25	14	14	14	14	14	12	12	17	12
26	14	12	14	14	14	14	10	14	14
27	15	17	15	15	15	11	13	15	15
28	14	14	14	14	14	14	17	14	12
29	13	13	13	13	13	12	13	13	13
30	14	14	13	14	14	14	14	14	14
31	13	13	12	13	13	13	17	13	13
32	14	14	12	14	14	12	10	14	14
33	14	14	14	14	14	14	14	14	14
34	14	14	14	14	14	14	14	14	10
35	14	14	14	14	14	14	17	14	14

Numéro segment	Note EV1	Note EV2	Note EV3	Note EV4	Note EV5	Note EV6	Note EV7	Note EV8	Note EV9
36	13	13	13	13	13	13	13	13	13
37	13	13	13	13	13	13	13	13	13
38	13	14	14	14	14	14	14	14	10
39	14	14	14	14	14	14	14	14	14
40	14	14	14	14	14	14	12	14	14
41	13	13	13	13	13	13	10	13	15
42	14	14	14	14	14	14	17	14	14
43	13	13	11	13	13	13	11	13	13
44	14	14	14	14	14	14	14	14	14
45	14	14	14	14	14	14	17	14	14
46	14	14	14	14	14	14	14	14	14
47	14	14	14	14	14	14	17	14	14
48	14	16	14	14	14	14	17	14	14
49	14	14	12	14	14	14	17	14	15
50	14	18	14	14	14	14	14	14	14
51	14	16	14	14	14	14	16	14	14
52	14	12	14	14	14	14	14	14	14
53	15	19	15	15	15	13	15	15	15
54	15	15	15	15	15	15	15	15	15
Moyenne simple	**14,17**	**15,31**	**13,85**	**14,22**	**14,31**	**13,80**	**14,26**	**14,33**	**12,69**

Tableau 19 : Récapitulatif des notations des juges bilingues (corpus LDJ-fr-so-A)

IV.3.3.2.2 Accord inter-annotateurs et qualité de l'auto-notation

Les premières données que nous avons recueillies auprès des annotateurs se résument comme ci-dessous :

- 54 segments ont été annotés par 9 annotateurs bilingues,
- La classe des notes attribuées par ces mêmes annotateurs est $\{5\} \cup \{10 \ldots 20\}$, soit 12 classes nominales pour cette évaluation.

D'après les résultats précédents, la moyenne des notes des 9 annotateurs est 14,10/20, correspondant à une qualité (ou mention) Bien. De plus, l'accord entre eux est excellent : le coefficient kappa de Fleiss[54] est très bon.

Notons que « le kappa de Fleiss fonctionne pour n'importe quel nombre d'observateurs donnant une classification nominale, à un nombre fixe d'éléments ». Ici, les notes sont des entiers entre 0 et 20, soit en théorie 21 classes. En fait, comme il s'agit de juger des post-éditions de traductions automatiques, les notes ne sont en fait jamais inférieures à 10, sauf la note 5, donnée 6 fois par EV9, ce qui fait 12 classes. On obtient *kappa* ≈ *0,7,* ce qui correspond, avec toutes les précautions d'usage, à une très bonne concordance.

[54] voir https://lemakistatheux.wordpress.com/2013/08/01/le-coefficient-kappa-de-fleiss/, https://fr.wikipedia.org/wiki/Kappa_de_Fleiss, et http://kappa.chez-alice.fr/Kappa_fleiss.htm.

Conclusion du chapitre IV

En partant des résultats de TA français-somali de GT, nous sommes arrivé à construire un très bon corpus de 10 669 segments de 9,5 mots en moyenne (98.912 mots, soit 396 pages standard), constitué de traductions de français en somali. Ce corpus est très homogène, car il a été tiré de certaines rubriques (en ligne) du quotidien « La Nation de Djibouti ». Nous l'avons appelé LDJ-fr-so-A, A signifiant « très bien » ou « très haute qualité ». C'est le tout premier corpus bilingue aligné de traductions français-somali.

Cette qualité a été obtenue en deux temps : d'abord, nous avons nous-même post-édité des « prétraductions » produites par GT (Google Translate) en combinant français-anglais et anglais-somali. D'après les informations disponibles au Tableau 15 et décrivant l'évolution du temps de post-édition avec GT, cela nous a pris environ 17 minutes par page standard, soit, en utilisant la formule de Boitet,[55] une qualité de 13/20. Ensuite, un petit échantillon du corpus a été revu par 9 annotateurs, qui ont donné une note à chaque segment (phrase ou titre), et éventuellement corrigé notre traduction. Quand ils l'ont fait, nous avons ajouté à la note évaluant notre résultat ½ point par correction.

Grâce à ce corpus, qui correspond au « filtrage » d'un corpus usuel de traductions au moins 10 fois plus gros, nous avons pu construire deux systèmes de TA « spécialisés » à notre sous-langage, qui ont donné des résultats bien meilleurs que GT, en termes de BLEU et comparables en termes de temps de PE.

À côté de ce corpus de très grande qualité, les travaux présentés dans ce chapitre nous ont permis de collecter 3 autres corpus de moindre qualité : (1) LDJ-fr-so-A (2194 segments, ≈88 pages), formé de nos premières post-éditions de GT-fr-so de LDJ, (2) OPUS-fr-so (7000 segments) formé par la mise en parallèle de traductions en-fr et en-so, sur des textes techniques ou religieux, et TED-fr-so (≈4000 segments), portant sur des cours/exposés. Dans les 2 derniers cas, il ne s'agit pas d'exemples de traductions. Mais, vu la petite taille de notre « très bon » corpus, nous nous sommes demandé s'il ne serait pas possible d'améliorer les systèmes de TA que nous avons construits en intégrant ces 3 corpus au corpus d'apprentissage.

Cette question est étudiée dans le chapitre suivant, consacré à la construction et à l'évaluation de systèmes de TA pour le couple français-somali.

[55] 18/20 (excellent) si 5mn/p, 16/20 (très bien) si 10 mn/p, 14/20 (bien) si 15 mn/p, 12/20 (assez bien) si 20 mn/p, 10/20 (passable) si 25 mn/p. Au-delà, la TA apporte moins d'aide qu'un système à mémoire de traductions comme Trados™.

Chapitre V Construction et évaluation de deux systèmes de TA statistique et neuronale français-somali

Introduction du chapitre V

Comme nous l'avons rappelé à la fin du chapitre précédent, les traductions français-somali produites par Google Translate (GT) sont estimées à 13,5/20 (assez bien) du point de vue de la qualité d'usage, mesurée par le temps de post-édition moyenne (à 17 mn/p) nécessaire pour obtenir un texte de très bonne qualité. Cela ne veut pas dire que nos post-éditions sont parfaites, mais qu'elles sont fidèles et grammaticales, lisibles en somali comme si les textes avaient été écrits directement en somali. Par contre, quand on met 17 mn/p pour post-éditer des résultats de TA, c'est toujours que les sorties brutes de TA sont en fait inutilisables par des lecteurs ne connaissant pas la langue source, ce qui est le cas pour nos utilisateurs finals envisagés, à savoir des somalophones du continent, connaissant un peu l'anglais et pas du tout le français.

Nous savons d'autre part qu'on peut obtenir des traductions automatiques de bien plus grande qualité en construisant des systèmes de TA « spécialisés à des sous-langages ». Ainsi, [Hajlaoui N., 2014] a démontré qu'on pouvait obtenir une augmentation de 25 points du score BLEU[56] en construisant un système spécialisé au sous-langage du Parlement européen à partir de 50K bisegments, par rapport au gros système généraliste de l'UE appris sur 20M de bisegments (400 fois plus de données).

Nous avons d'abord réévalué les résultats de GT sur notre corpus LDJ-fr-so-A, pour disposer d'un « point de comparaison » (*base line*) sur des données identiques. Nous avons ensuite construit et évalué 4 systèmes de TA, 2 de type « statistique » (Moses à fragments) et 2 de type « neuronal (OpenNMT). Le premier système de chaque type a été construit à partir du corpus LDJ-fr-so-A (10 669 segments), et le second à partir de l'union de tous les corpus français-somali mentionnés au chapitre précédent, soit LDJ-fr-so-ABC (23 863 segments).

V.1 Évaluation de GT-fr-so sur le corpus LDJ-fr-so-A

V.1.1 Matériel et méthode

Nous avons choisi et mis de côté pour toutes nos expériences un « corpus de test », LDJ-fr-so-TEST, de 643 segments, représentant environ 5% du corpus de référence (LDJ-fr-so-A). Nous avons retraduit LDJ-fr-so-TEST avec la version courante de GT, au moment de rédiger ce chapitre.

Cela nous permettra des comparaisons avec les 4 systèmes que nous avons construits.

Nous avons calculé pour GT :

- le score BLEU [Papineni et al., 2002], pour LDJ-fr-so-TEST
- le score TER (translation error rate) [Snover et al., 2006] pour les mêmes corpus
- la distance mPED[57] pour 107 segments extraits du corpus LDJ-fr-so-TEST

[56] Le score BLEU est une similarité, donc dans [0,1]. On le compte en pourcentage, avec 1 point = 1%.

[57] Dm ou mPED (mixed Post-Editing Distance) ou mTER (mixed Translation Error Rate) est la « distance mixte de post-édition » utilisée dans SECTra_w/iMAG. D_m(U, V) = α D_c(U, V) + (1-α) D_w(U,V) où D_c(U, V)

Avant de fournir les résultats de notre expérience sur la TA de GT avec notre corpus d'évaluation, nous allons décrire la démarche que nous avons suivie pour effectuer ce travail.

Notre corpus d'évaluation, dit LDJ-fr-so-TEST, est composé de 643 segments extraits du corpus LDJ-fr-so-A construit sous SECTra_w/iMAG par PE de résultats de GT.

Segments	Items
643	10705

Tableau 20 : Description du corpus TestLDJ-fr-so-A

Nous avons découpé ce corpus de test en une dizaine de sous-corpus d'environ 30 à 50 segments, et avons traduit chacun de ces sous-corpus avec l'interface de traduction de GT. Cette méthode nous a paru adaptée à notre cas, car GT limite à 5000 le nombre de mots qu'on peut traduire d'un seul coup. Par la suite, nous les avons rassemblés et avons calculé les différentes mesures d'évaluation (BLEU, TER, METEOR etc.) sur les sorties de GT, par rapport à notre référence qui se trouve dans la partie complète du corpus LDJ-fr-so-A.

Le tableau ci-dessous récapitule les résultats de ces mesures d'évaluation.

Mesure d'évaluation	Score
BLEU	0,18
METEOR	0,31
TER	0,73

Tableau 21 : Résultats d'évaluation sur LDJ-fr-so-A avec GT

On voit que le score BLEU est très faible (18%), ce qui laisse un bon espoir d'arriver à mieux en construisant des systèmes spécialisés.

Nous avons aussi calculé une autre mesure, objective, le temps total de post-édition Tpe, en post-éditant (sous SECTra_w/iMAG) 107 segments pris dans LDJ-fr-so-TEST.

Nous voyons que la corrélation remarquée par [Wang H., 2015] dans son M2R semble vérifiée : à 1 unité de distance mixte (Dmix) correspond un temps (total) de PE, constant, d'environ 2 secondes pour le français-chinois, et de 2,7 secondes pour le français-somali.

Segment source	Segment initial PE	Traduction avec GT sur LDJ-fr-so-TEST
En cliquant sur l'un de ses liens, vous serez informé en temps réel.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.	diga oo riixaya mid ka mid ah xiriiriyeyaashiisa, waxaa laguu sheegi doonaa waqtiga dhabta ah.
Concernant les opérations de maintien de la paix, et plus spécifiquement les crises en Somalie et en Centrafrique, Washington va augmenter sa participation financière pour équiper les troupes africaines qui seraient déployées sur les terrains.	Ta ku saabsan hawlaha dayactirka nabadda, iyo in si gaar ah dhibaatooyinka Soomaaliya iyo Afrikada Dhexe, Washington waxay kordhin doontaa ka qayb galkeega dhaqaale si lo qalabeeyo ciidamada Afrikaanka ee la geyay dhulka.	Marka la eego hawlaha nabad-ilaalinta, iyo si gaar ah dhibaatooyinka Soomaaliya iyo Jamhuuriyadda Bartamaha Afrika, Washington waxay kordhin doontaa ka qayb qaadashada maaliyadeed si ay u qalabeeyaan ciidamada Afrikaanka ah ee la geeyo dhulka.
Il s'agit en effet de fournir des vivres à une population estimée à 8 millions de personnes touchées par les effets du	Dhab ahaan Waxaa weeye in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.	waa xaqiiqo ah in la siiyo cunto lagu qiyaasay dad lagu qiyaasay 8 milyan oo qof oo saameeya saamaynta isbeddelka cimilada.

est la distance de Levenshtein (en caractères) entre les chaînes (de mots) U = u_1…u_ p et V = w_1…w_q, et D_w(U, V) est leur distance en mots, en comptant comme coût d'échange X(u_i, v_j) = D_c(u_i, v_j).

changement climatique.		
Pour un système d'information sanitaire performant	Sameynta nidaamka macluumaadka caafimaadka ee hufan	si loo helo nidaam macluumaad oo ku habboon caafimaadka
C'est pourquoi les intervenants de l'atelier se sont accordés autour du caractère vital des soins durant la préconception, les périodes prénatale, périnatale et postnatale pour les femmes, les mères et les nouveaux nés.	Tani waa sababta ka hadleyaashii tababarkan ay ku heshiiyeen muhiimada xasaasiga ee daryeelka inta lagu guda jiro abuurista ilmaha, dhalmada ka hor, dhalida iyo muddooyinka umusha loogu tala galay dumarka, hooyooyinka iyo dhallaanka.	taasina waa sababta ka qaybgalayaasha aqoon iswaydaarsiga lagu heshiiyey agagaarka daryeelka muhiimka ah inta lagu guda jiro preconception, dhalmada ka hor, dhalida iyo muddooyinka umusha loogu tala galay dumarka, hooyooyinka iyo dhallaanka.
'Un pays frère et ami vient de perdre des centaines de jeunes" a dit le Président Guelleh qui a condamné cet acte horrible, lâche et cruel.	"Wadan Walaalkeena ico saaxiibkeena ayaa boqolaal dhallinyaro waayey"; ayuu yiri madaxweyne Geelle isago cambaareeyay falkan argagaxa leh, naxariis lahayn iyo fulaynimada ah.	Walaal walaalo ah oo saaxiib ah ayaa ka lumay boqolaal dhalinyaro ah, ayuu yiri Geelle oo cambaareeyay ficilkan naxdinta leh, fuleynimada iyo naxariis darrada ah.
Recherche	Badhis	search
Djibouti Today	Maanta Jibuuti	Jabuuti maanta
La micro finance représente, de part le monde, un moyen de lutte contre la pauvreté en améliorant les conditions de vies des ménages pauvres.	Ammaahda danyarta waa, sida adduunka oo dhan, hab lo la dagaalamo faqriga iyado la hagaajinayo xaaladaha ay ku nool yihiin qoysaska saboolka.	Microfinance waa qalab heer caalami ah oo la dagaallama saboolnimada iyadoo la wanaajinayo xaaladaha nololeed ee qoysaska saboolka ah.
1er Congrès Africain des Transports et de la Logistique (CATL2015) : Moussa Ahmed Hassan aux assises de Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat	Shirkii 1aad ee Afrika ee Gaadiidka iyo Soodhaweynta (CATL2015): Moussa Ahmed Hassan oo ka tirsan Assises de Rabat

Tableau 22 : Exemple de traduction de 10 segments français en somali avec GT

V.1.2 Résultats

Système	Corpus	#segments	#mots	BLEU	TER	mPED/p	Tpe/p
GT	LDJ-TEST	643	9876	0,18	0,73	400 i	18mn

Tableau 23 : Résultat de la TA du corpus test avec GT

Nous obtenons presque le même résultat (18 mn/p) que celui obtenu par L. X. Wang (17 mn/p) au début de son travail sur la construction d'un système français-chinois, quand il a commencé par construire un corpus parallèle français-chinois de qualité. Dans les 2 cas, GT passait par l'anglais, le BLEU était considéré comme mauvais, mais la qualité d'usage, mesurée à partir de temps de PE, était finalement respectable (17 mn/p, la formule de Boitet donnant une note de 13,5/20).

V.2 Systèmes Moses (spécialisé et augmenté)

V.2.1 Matériel et méthode

Pour le système spécialisé, le corpus d'apprentissage est LDJ-fr-so-A, dont on a retiré 5% pour les tests (LDG-fr-so-TEST) et 10% pour le réglage (*tuning*).

Pour le système augmenté, le corpus d'apprentissage est LDJ-fr-so-ABC, dont on a retiré le même corpus de test (LDJ-fr-so-TEST) et 10% pour le réglage (*tuning*). Ces 10% sont formés des mêmes 10% de LDJ-fr-so-A, et de 10% des corpus ajoutés (LDJ-fr-so-B, OPUS-fr-so et TED-fr-so).

Nos 2 systèmes de TA probabiliste ont été réalisés à l'aide de Moses, dans la version à base de fragments (PBMT).

Segment source	Segment initial PE	Traduction avec Moses-LDJ-fr-so-A
En cliquant sur l'un de ses liens, vous serez informé en temps réel.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.	ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.
Concernant les opérations de maintien de la paix, et plus spécifiquement les crises en Somalie et en Centrafrique, Washington va augmenter sa participation financière pour équiper les troupes africaines qui seraient déployées sur les terrains.	Ta ku saabsan hawlaha dayactirka nabadda, iyo in si gaar ah dhibaatooyinka Soomaaliya iyo Afrikada Dhexe, Washington waxay kordhin doontaa ka qayb galkeega dhaqaale si lo qalabeeyo ciidamada Afrikaanka ee la geyay dhulka.	marka la iyo hawlaha ee nabad ilaalinta iyo ka ka sii ay laba ah u samayn ee Soomaaliya iyo Centrafrique, Washington si augmenter ku qaybgalka dhaqaalaha si équiper ay ciidamada Afrika oo ku saabsan terrains in geeyay.
Il s'agit en effet de fournir des vivres à une population estimée à 8 millions de personnes touchées par les effets du changement climatique.	Dhab ahaan Waxaa weeye in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.	tani waa dhab ahaan waxaa ahayd in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.
Pour un système d'information sanitaire performant	Sameynta nidaamka macluumaadka caafimaadka ee hufan	nidaamka macluumaadka caafimaadka ee hufan
C'est pourquoi les intervenants de l'atelier se sont accordés autour du caractère vital des soins durant la préconception, les périodes prénatale, périnatale et postnatale pour les femmes, les mères et les nouveaux nés.	Tani waa sababta ka hadleyaashii tababarkan ay ku heshiiyeen muhiimada xasaasiga ee daryeelka inta lagu guda jiro abuurista ilmaha, dhalmada ka hor, dhalida iyo muddooyinka umusha loogu tala galay dumarka, hooyooyinka iyo dhallaanka.	taasi waa sababta in daneeyayaasha aqoon iswaydaarsiga ku heshiiyeen ee waxay vital daryeelka intii lagu guda jiray préconception, xilliyada ku, asphyxia iyo umusha haweenka, hooyooyinka iyo dhallaanka.
'Un pays frère et ami vient de perdre des centaines de jeunes" a dit le Président Guelleh qui a condamné cet acte horrible, lâche et cruel.	"Wadan Walaalkeena ico saaxiibkeena ayaa boqolaal dhallinyaro waayey"; ayuu yiri madaxweyne Geelle isago cambaareeyay falkan argagaxa leh, naxariis lahayn iyo fulaynimada ah.	"ka dhigo dal frère ami iyo ayaa ku perdre ee centaines dhalinyarada " ayuu yiri madaxweynuhu Geelle kaas oo sameeyay condamné faa "; iideysanaya acte horrible, lâche iyo cruel.
Recherche	Badhis	Badhis
Djibouti Today	Maanta Jibuuti	maanta Jibuuti

La micro finance représente, de part le monde, un moyen de lutte contre la pauvreté en améliorant les conditions de vies des ménages pauvres.	Ammaahda danyarta waa, sida adduunka oo dhan, hab lo la dagaalamo faqriga iyado la hagaajinayo xaaladaha ay ku nool yihiin qoysaska saboolka.	ammaahda danyarta meel,, ee dunida, hab ee ka dhanka ah saboolnimada ku améliorant xaaladaha vies qoysaska saboolka ah.
1er Congrès Africain des Transports et de la Logistique (CATL2015) : Moussa Ahmed Hassan aux assises de Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat	kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat

Tableau 24 : Exemple de traduction de 10 segments français avec MosesLIG-LDJ-fr-so-A

Segment source	Segment initial PE	Traduction avec Moses-LDJ-fr-so-ABC
En cliquant sur l'un de ses liens, vous serez informé en temps réel.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.
Concernant les opérations de maintien de la paix, et plus spécifiquement les crises en Somalie et en Centrafrique, Washington va augmenter sa participation financière pour équiper les troupes africaines qui seraient déployées sur les terrains.	Ta ku saabsan hawlaha dayactirka nabadda, iyo in si gaar ah dhibaatooyinka Soomaaliya iyo Afrikada Dhexe, Washington waxay kordhin doontaa ka qayb galkeega dhaqaale si lo qalabeeyo ciidamada Afrikaanka ee la geyay dhulka.	Marka la eego ah iyo ee nabad ilaalinta ee iyo ka iyo in ka badan wuxu caawin doona kordhinta helitaanka waxbarasho Centrafrique Soomaaliya ' oo, Washington ahayn ka dhaqaale ku ka équiper Afrika oo u troupes iyo ku saabsan. terrains déployées way ku
Il s'agit en effet de fournir des vivres à une population estimée à 8 millions de personnes touchées par les effets du changement climatique.	Dhab ahaan Waxaa weeye in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.	Tani waa xaqiiqdii in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.
Pour un système d'information sanitaire performant	Sameynta nidaamka macluumaadka caafimaadka ee hufan	Nidaamka macluumaadka caafimaadka ee hufan
C'est pourquoi les intervenants de l'atelier se sont accordés autour du caractère vital des soins durant la préconception, les périodes prénatale, périnatale et postnatale pour les femmes, les mères et les nouveaux nés.	Tani waa sababta ka hadleyaashii tababarkan ay ku heshiiyeen muhiimada xasaasiga ee daryeelka inta lagu guda jiro abuurista ilmaha, dhalmada ka hor, dhalida iyo muddooyinka umusha loogu tala galay dumarka, hooyooyinka iyo dhallaanka.	Taasi waa sababta oo ee iyo baahiyaha aqoon iswaydaarsiga vital waa dhinaca bulshada ee dalka ku heshiiyeen ee daryeelka ee wax lagu préconception lagu prénatale,, haweenka iyo u postnatale périnatale, iyo dhimashada dhallaanka. oo hooyooyinka
'Un pays frère et ami vient de perdre des centaines de jeunes" a dit le Président Guelleh qui a condamné cet acte horrible, lâche et cruel.	"Wadan Walaalkeena ico saaxiibkeena ayaa boqolaal dhallinyaro waayey"; ayuu yiri madaxweyne Geelle isago cambaareeyay falkan argagaxa leh, naxariis lahayn iyo fulaynimada ah.	wuxuuna iyo "Markaa wadan ami perdre ka "; dhallinyarada boqolaal ayuu yiri Geelle kaas oo sameeyay Madaxweynaha condamné, iyo. lâche, cruel horrible
Recherche	Badhis	Badhis
Djibouti Today	Maanta Jibuuti	Maanta Jibuuti

La micro finance représente, de part le monde, un moyen de lutte contre la pauvreté en améliorant les conditions de vies des ménages pauvres.	Ammaahda danyarta waa, sida adduunka oo dhan, hab lo la dagaalamo faqriga iyado la hagaajinayo xaaladaha ay ku nool yihiin qoysaska saboolka.	Ammaahda danyarta ka tahay, adduunka ah ah ee ka dhanka ah saboolnimada améliorant ka xaaladaha vies qoysaska saboolka ah.
1er Congrès Africain des Transports et de la Logistique (CATL2015) : Moussa Ahmed Hassan aux assises de Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat

Tableau 25 : Exemple de traduction de 10 segments français avec MosesLIG-LDJ-fr-so-ABC

V.2.1.1 Architecture d'un système de TA statistique construit avec Moses

Un système de TA probabiliste[58] « à fragments » construit avec la boîte à outils Moses [Koehn et al., 2007] utilise plusieurs éléments dépendant de la paire de langues et du genre de textes à traduire :

- un modèle de traduction de segments, appelé également « table de traduction », construit à partir d'un corpus d'apprentissage formé d'exemples de traductions,
- un modèle de langage de la langue cible dont l'objectif est de s'assurer de la grammaticalité des hypothèses de traduction proposées par le système PBMT,
- un modèle de ré-ordonnancement qui permet de gérer les alignements des segments de traduction.

La Figure 30 ci-dessous présente l'architecture générale d'un décodeur PBMT classique. Le « décodeur » est un algorithme invariant, indépendant des langues et des textes à traiter, qui calcule une liste ordonnée d'hypothèses de traduction par « optimisation combinatoire ». Il utilise une recherche en faisceau (*beam search*) dans l'espace de recherche, lui-même organisé dans une structure factorisante complexe (treillis, arbres et files avec priorités).

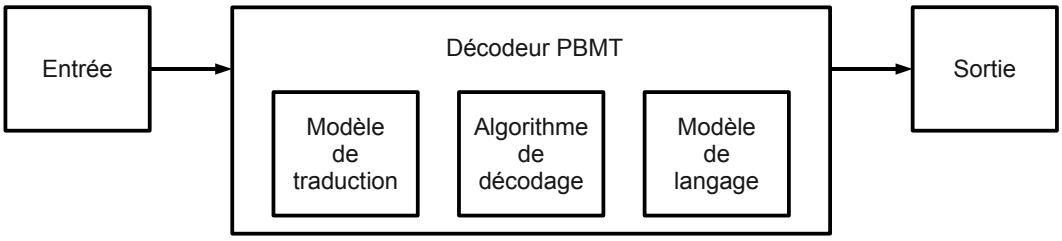

Figure 30 : Architecture d'un décodeur PBMT classique

Durant la phase de décodage, le décodeur sélectionne les segments de la langue cible les plus probables d'une phrase source à traduire en fonction des différents scores associés à chacun de ce modèle.

Pour minimiser les erreurs de traduction, le décodeur contient des outils pour effectuer l'optimisation ; le corpus dédié à cette tâche est généralement appelé corpus de développement ou corpus d'optimisation. Cette phase d'optimisation s'effectue selon la méthode proposée par [Och, 2003] et est appelée MERT (Minimum Error-Rate Training).

[58] Le fonctionnement d'un système « SMT » (statistical MT) est en fait probabiliste, les probabilités ayant été estimées par des calculs statistiques préalables effectués durant la phase d'apprentissage.

V.2.1.2 Développement du système de TA probabiliste

Nous avons tout d'abord calculé les alignements des mots de notre corpus en utilisant GIZA++ [Och, 2003], qui implémente les algorithmes des différents modèles IBM (1 à 5) [Brown et al., 1993] et le modèle HMM [Vogel et al., 1996]. Ces alignements servent à construire la table de traductions. Le modèle de ré-ordonnancement est lui aussi construit et celui-ci contient les informations sur les positions de mots des phrases ou segments qui ont été précédemment traduits.

Enfin le modèle de langage est construit à l'aide de l'outil IRSTLM [Federico et al., 2008].

La construction de notre premier système de TA français-somali avec Moses nous a pris environ 16 heures, de la phase d'entraînement jusqu'à l'évaluation des scores BLEU [Papineni et al., 2002].

Les différents tableaux ci-dessous résument les données utilisées pour construire nos deux systèmes de TA à base de fragments, l'un spécialisé et l'autre augmenté, avec la boite à outils libre Moses, version MosesLIG.

Type corpus	Segments	Items source	Items cible
Apprentissage	10 933	181 377	164 764
Validation	1 286	21 252	19 540
Evaluation	643	10 705	9 826

Tableau 26 : Description des données du système MosesLIG-LDJ-fr-so-A

Type corpus	Segments	Items source	Items cible
Apprentissage	20 920	360 545	299 952
Validation	2 395	27 890	24 927
Evaluation	643	10 705	9 826

Tableau 27 : Description des données du système MosesLIG-LDJ-fr-so-ABC

V.2.2 Résultats

Mesure évaluation	Score
BLEU	0,32
TER	1,15

Tableau 28 : Scores BLEU et TER du système MosesLIG-LDJ-fr-so-A

Mesure évaluation	Score
BLEU	0,34
TER	1,01

Tableau 29 : Scores BLEU et TER du système MosesLIG-LDJ-fr-so-ABC

V.2.2.1 Récapitulatif et commentaires des résultats de l'évaluation des deux systèmes de TA Moses à base de fragments français-somali

Système	Corpus	#segments	#mots	BLEU	TER	mPED/p	Tpe/p
Moses	LDJ-A	12 863	213 336	0,32	1,15	422i	19mn
Moses	LDJ-ABC	23958	334 705	0,34	1,01	377	17mn

Tableau 30 : Résultats des systèmes de TA probabiliste à base de fragments

Les résultats de nos deux expériences de traduction avec les deux systèmes de TA statistique avec Moses sont encourageants et affichent de bons scores BLEU et de Tpe/p. Le score BLEU du système de TA MOSES augmenté a presque doublé par rapport à celui de GT.

Les résultats de nos calculs du temps de post-édition par page standard du corpus test sur les deux systèmes de TA statistique MosesLIG-LDJ-fr-so-A et MosesLIG-LDJ-fr-so-ABC sont estimés respectivement à 17 et 19 minutes par page standard.

V.3 Systèmes OPENNMT (spécialisé et augmenté)

V.3.1 Matériel et méthode

Nous avons utilisé les mêmes ressources, et avons adapté la méthode au cas d'un système de TA « neuronale », pour lequel les composants sont différents. En particulier, il n'y a pas de « modèle de langue ».

Pour le système spécialisé, le corpus d'apprentissage est donc toujours LDJ-fr-so-A, dont on a retiré 5% pour les tests (LDJ-fr-so-TEST) et 10% pour le réglage (*tuning*).

Pour le système augmenté, le corpus d'apprentissage est LDJ-fr-so-ABC, dont on a retiré le même corpus de test (LDJ-fr-so-TEST) et 10% pour le réglage (*tuning*). Ces 10% sont formés des mêmes 10% de LDJ-fr-so-A, et de 10% des corpus ajoutés (LDJ-fr-so-B, OPUS-fr-so et TED-fr-so).

Les deux tableaux ci-dessous contiennent la traduction de 10 segments du journal *La Nation de Djibouti*, traduits par avec nos deux systèmes de TA neuronale construits avec OPENNMT.

Comme on le voit dans le Tableau 35 et le Tableau 36, les résultats de TA neuronale de type OPENNMT sont moins bons que ceux obtenus avec la TA statistique de type MOSES.

Segment source	Segment initial PE (post-édité)	Traduction avec OpenNMT/LDJ-fr-so-A
En cliquant sur l'un de ses liens, vous serez informé en temps réel.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.	ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.
Concernant les opérations de maintien de la paix, et plus spécifiquement les crises en Somalie et en Centrafrique, Washington va augmenter sa participation financière pour équiper les troupes africaines qui seraient déployées sur les terrains.	Ta ku saabsan hawlaha dayactirka nabadda, iyo in si gaar ah dhibaatooyinka Soomaaliya iyo Afrikada Dhexe, Washington waxay kordhin doontaa ka qayb galkeega dhaqaale si lo qalabeeyo ciidamada Afrikaanka ee la geyay dhulka.	marka la iyo hawlaha ee nabad ilaalinta iyo ka ka sii ay laba ah u samayn ee Soomaaliya iyo Centrafrique, Washington si augmenter ku qaybgalka dhaqaalaha si équiper ay ciidamada Afrika oo ku saabsan terrains in geeyay.
Il s'agit en effet de fournir des vivres à une population estimée à 8 millions de personnes touchées par les effets du changement climatique.	Dhab ahaan Waxaa weeye in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.	tani waa dhab ahaan waxaa ahayd in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.
Pour un système d'information sanitaire performant	Sameynta nidaamka macluumaadka caafimaadka ee hufan	nidaamka macluumaadka caafimaadka ee hufan
C'est pourquoi les intervenants de l'atelier se sont accordés autour du caractère vital des soins durant la préconception,	Tani waa sababta ka hadleyaashii tababarkan ay ku heshiiyeen muhiimada xasaasiga ee daryeelka inta lagu guda jiro	taasi waa sababta in daneeyayaasha aqoon iswaydaarsiga ku heshiiyeen ee waxay vital daryeelka intii lagu

les périodes prénatale, périnatale et postnatale pour les femmes, les mères et les nouveaux nés.	abuurista ilmaha, dhalmada ka hor, dhalida iyo muddooyinka umusha loogu tala galay dumarka, hooyooyinka iyo dhallaanka.	guda jiray préconception, xilliyada ku, asphyxia iyo umusha haweenka, hooyooyinka iyo dhallaanka.
'Un pays frère et ami vient de perdre des centaines de jeunes" a dit le Président Guelleh qui a condamné cet acte horrible, lâche et cruel.	"Wadan Walaalkeena ico saaxiibkeena ayaa boqolaal dhallinyaro waayey"; ayuu yiri madaxweyne Geelle isago cambaareeyay falkan argagaxa leh, naxariis lahayn iyo fulaynimada ah.	"ka dhigo dal frère ami iyo ayaa ku perdre ee centaines dhalinyarada " ayuu yiri madaxweynuhu Geelle kaas oo sameeyay condamné faa "; iideysanaya acte horrible, lâche iyo cruel.
Recherche	Badhis	Badhis
Djibouti Today	Maanta Jibuuti	maanta Jibuuti
La micro finance représente, de part le monde, un moyen de lutte contre la pauvreté en améliorant les conditions de vies des ménages pauvres.	Ammaahda danyarta waa, sida adduunka oo dhan, hab lo la dagaalamo faqriga iyado la hagaajinayo xaaladaha ay ku nool yihiin qoysaska saboolka.	ammaahda danyarta meel,, ee dunida, hab ee ka dhanka ah saboolnimada ku améliorant xaaladaha vies qoysaska saboolka ah.
1er Congrès Africain des Transports et de la Logistique (CATL2015) : Moussa Ahmed Hassan aux assises de Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat	kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat

Tableau 31 : Exemple de traduction de 10 segments français avec le système OpenNMT/LDJ-fr-so-A

Segment source	Segment initial PE	Traduction avec Moses-LDJ-fr-so-A
En cliquant sur l'un de ses liens, vous serez informé en temps réel.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.	Ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.
Concernant les opérations de maintien de la paix, et plus spécifiquement les crises en Somalie et en Centrafrique, Washington va augmenter sa participation financière pour équiper les troupes africaines qui seraient déployées sur les terrains.	Ta ku saabsan hawlaha dayactirka nabadda, iyo in si gaar ah dhibaatooyinka Soomaaliya iyo Afrikada Dhexe, Washington waxay kordhin doontaa ka qayb galkeega dhaqaale si lo qalabeeyo ciidamada Afrikaanka ee la geyay dhulka.	Marka la eego ah iyo ee nabad ilaalinta ee iyo ka iyo in ka badan wuxu caawin doona kordhinta helitaanka waxbarasho Centrafrique Soomaaliya ' oo, Washington ahayn ka dhaqaale ku ka équiper Afrika oo u troupes iyo ku saabsan. terrains déployées way ku
Il s'agit en effet de fournir des vivres à une population estimée à 8 millions de personnes touchées par les effets du changement climatique.	Dhab ahaan Waxaa weeye in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.	Tani waa xaqiiqdii in cunto la siiyo dad lagu qiyaasay 8 milyan oo qof oo ay saameeyeen isbedelka cimilada.
Pour un système d'information sanitaire performant	Sameynta nidaamka macluumaadka caafimaadka ee hufan	Nidaamka macluumaadka caafimaadka ee hufan
C'est pourquoi les intervenants de l'atelier se sont accordés autour du caractère vital des soins durant la préconception, les périodes prénatale, périnatale et postnatale pour les femmes,	Tani waa sababta ka hadleyaashii tababarkan ay ku heshiiyeen muhiimada xasaasiga ee daryeelka inta lagu guda jiro abuurista ilmaha, dhalmada ka hor, dhalida iyo muddooyinka	Taasi waa sababta oo ee iyo baahiyaha aqoon iswaydaarsiga vital waa dhinaca bulshada ee dalka ku heshiiyeen ee daryeelka ee wax lagu préconception lagu prénatale,, haweenka iyo u

les mères et les nouveaux nés.	umusha loogu tala galay dumarka, hooyooyinka iyo dhallaanka.	postnatale périnatale, iyo dhimashada dhallaanka. oo hooyooyinka
'Un pays frère et ami vient de perdre des centaines de jeunes" a dit le Président Guelleh qui a condamné cet acte horrible, lâche et cruel.	"Wadan Walaalkeena ico saaxiibkeena ayaa boqolaal dhallinyaro waayey"; ayuu yiri madaxweyne Geelle isago cambaareeyay falkan argagaxa leh, naxariis lahayn iyo fulaynimada ah.	wuxuuna iyo "Markaa wadan ami perdre ka "; dhallinyarada boqolaal ayuu yiri Geelle kaas oo sameeyay Madaxweynaha condamné, iyo. lâche, cruel horrible
Recherche	Badhis	Badhis
Djibouti Today	Maanta Jibuuti	Maanta Jibuuti
La micro finance représente, de part le monde, un moyen de lutte contre la pauvreté en améliorant les conditions de vies des ménages pauvres.	Ammaahda danyarta waa, sida adduunka oo dhan, hab lo la dagaalamo faqriga iyado la hagaajinayo xaaladaha ay ku nool yihiin qoysaska saboolka.	Ammaahda danyarta ka tahay, adduunka ah ah ee ka dhanka ah saboolnimada améliorant ka xaaladaha vies qoysaska saboolka ah.
1er Congrès Africain des Transports et de la Logistique (CATL2015) : Moussa Ahmed Hassan aux assises de Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat	Kulankii Kowaad ee Afrikanka ee gaadiidka iyo xamuulka (CATL2015) : Musa Axmad Xassan oo tagay fadhiga Rabat

Tableau 32 : Exemple de traduction de 10 segments français avec le système OpenNMT/LDJ-fr-so-ABC

Nos deux systèmes de TA neuronale ont été réalisés avec l'outil OPENNMT, qui fournit une boîte à outils ouverte.

Les caractéristiques des données utilisées pour construire ces deux systèmes de TA neuronale, spécialisé et augmenté, sont résumées dans le Tableau 33 et le Tableau 34 ci-dessous :

Type corpus	Segments	Items source	Items cible
Apprentissage	10 933	181 377	164 764
Validation	1 286	21 252	19 540
Evaluation	643	10 705	9 826

Tableau 33 : Description des données du système OpenNMT/LDJ-fr-so-A

Type corpus	Segments	Items source	Items cible
Apprentissage	20 920	360 545	299 952
Validation	2 395	27 890	24 927
Evaluation	643	10 705	9 826

Tableau 34 : Description des données du système OpenNMT/LDJ-fr-so-ABC

V.3.1.1 Architecture d'un système de TA neuronale avec OPENNMT

Le second système de TA construit à partir des données bilingues initiales est un modèle de TA neuronale. Il utilise l'une des dernières technologies du domaine, issue de l'apprentissage profond, pour améliorer la qualité de traduction d'un système de TA.

Un système de TA neuronal est basé sur l'approche d'apprentissage séquence à séquence avec le mécanisme d'attention [Sutskever et al., 2014, Bahdanau et al., 2015].

Comme le montre la Figure 31, un système de TA neuronale contient généralement les trois composants suivants :

- **Un encodeur**. Le rôle de l'encodeur est de transformer un segment source en une liste de vecteurs avec un vecteur par symbole d'entrée,
- **Un décodeur**. Étant donné la liste des vecteurs, le décodeur produit un symbole à chaque fois jusqu'à ce qu'il rencontre le symbole spécial de fin de segment (EOS),
- **Le mécanisme d'attention**. Le module d'attention permet de faire la connexion entre l'encodeur et le décodeur, en mettant l'attention sur certaines parties du segment source durant le processus de traduction.

La Figure 31et la Figure 32 ci-dessous montrent respectivement l'architecture globale du système de TA neuronale de Google et donnent une vue schématique d'un système de TA neuronale avec ses différentes composantes.

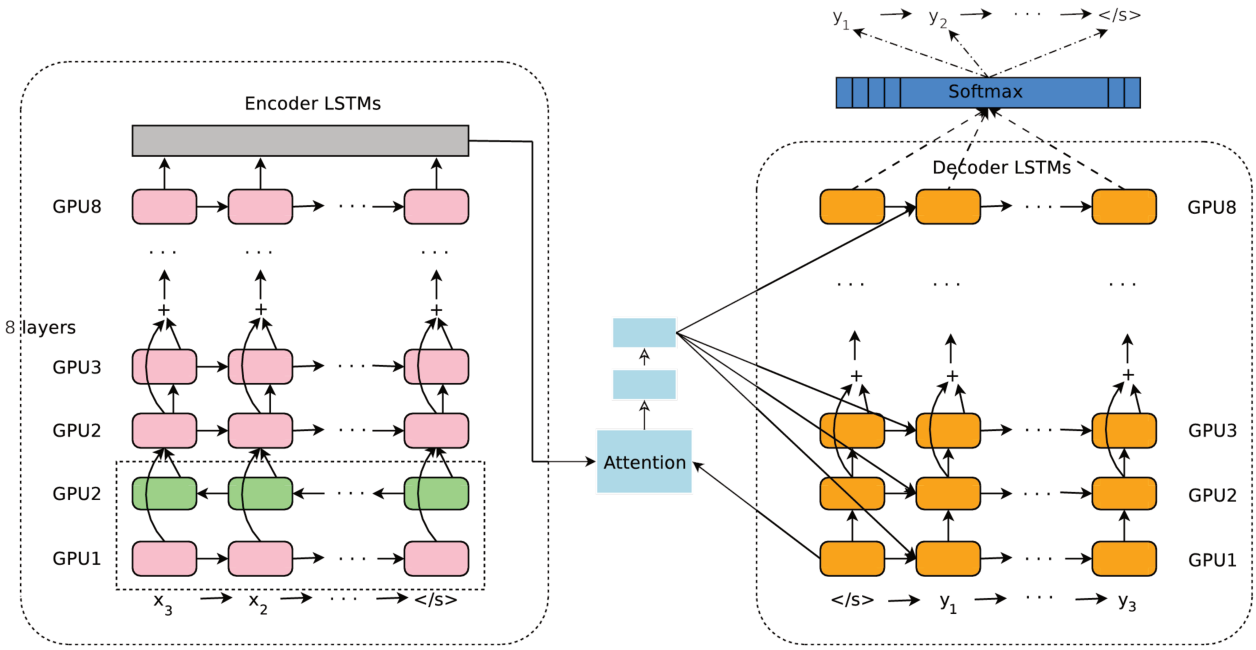

Figure 31 : Architecture du modèle de TA neuronale GNMT (Google's Neural Machine Translation)[59]

[59] A gauche se trouve l'encodeur, à droite le décodeur, et au centre le module d'attention.

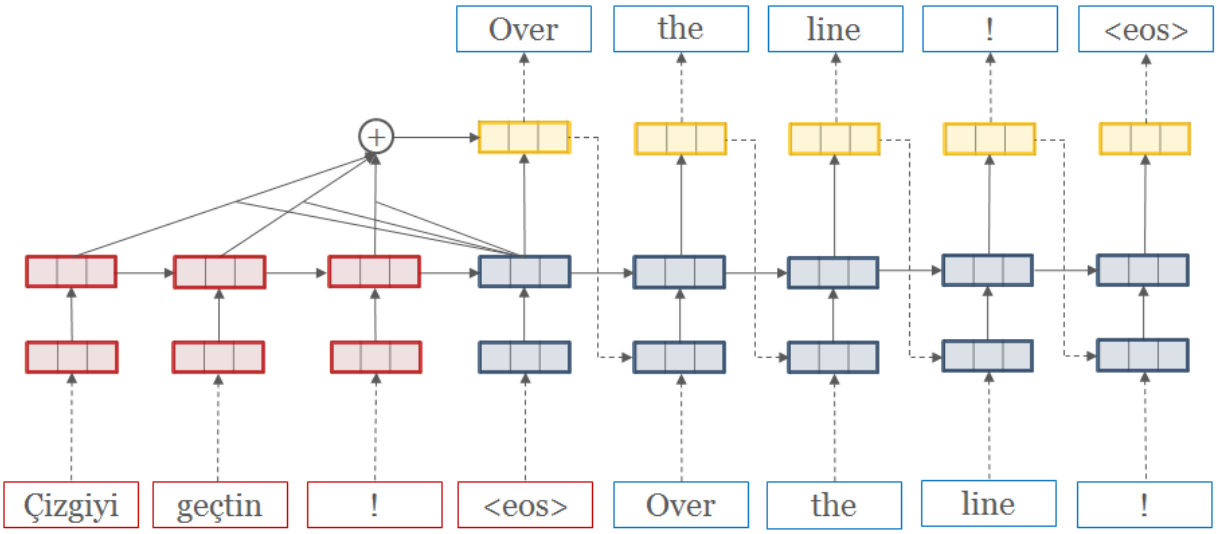

Figure 32 : Vue schématique du système de TA neuronale OpenNMT (source : [Klein et al., 2017])

V.3.1.2 Développement de deux systèmes de TA neuronale

Pour nos deux expériences de TA neuronale, spécialisé et augmenté, nous avons, comme annoncé plus haut, utilisé OPENNMT [Klein & al., 2017]. Cette boite à outils contient tous les outils nécessaires pour construire un système de TA neuronal complet et libre de droits. De plus, elle est très bien documentée et facile d'utilisation.

Le système OPENNMT est le successeur du modèle d'apprentissage profond de type séquence à séquence destiné à la traduction automatique.

Les principales fonctionnalités d'OPENNMT (apprentissage, traduction et inférence) sont implémentées dans le cadre mathématique appelé LUA/TORCH.

Contrairement aux systèmes de TA probabiliste à base de fragments, les systèmes de TA neuronale comme OPENNMT utilisent seulement un grand corpus bilingue d'une paire de langues à traduire. Cette possibilité qu'offrent les systèmes de TA neuronale trouve tout son intérêt lorsqu'il s'agit de développer des systèmes de TA pour des langues faiblement dotées, qui ne disposent pas des très grands volumes de données monolingues nécessaires pour obtenir de bons modèles de langage.

Dans notre cas, pour faire l'apprentissage du modèle de traduction neuronale de nos deux systèmes (OpenNMT-LDJ-fr-so-A et OpenNMT-LDJ-fr-so-ABC), nous avons utilisé un encodeur-décodeur fondé sur un modèle bi-LSTM à 2 couches cachées de 500 neurones, avec des représentations vectorielles (word embeddings, ou *plongements de mots*) de taille 500.

Nous avons utilisé également un modèle d'attention standard. Enfin, nous avons appliqué un « dropout » naïf (avec un taux de 0,3) durant l'entraînement sur les connexions entre les couches LSTM dans l'encodeur-décodeur.

Le Tableau 35 et le Tableau 36 ci-dessous contiennent les meilleurs scores des mesures d'évaluation BLEU et TER des résultats de la TA sur le corpus d'évaluation LDJ-TEST.

Ces deux scores ont été atteints à la 13ème itération pour le système de TA neuronale spécialisé et à la 25ème itération pour le système de TA neuronale augmenté, construit à partir du corpus LDJ-fr-so-ABC.

N° Itération (epoch)	Mesures Evaluation	Scores
13	BLEU	0,11
	TER	1,03

Tableau 35 : Meilleurs scores BLEU et TER des 13 itérations du système OpenNMT/LDJ-fr-so-A

N° Itération (epoch)	Mesures Evaluation	Scores
25	BLEU	0,17
	TER	0,83

Tableau 36 : Meilleurs scores BLEU et TER des 25 itérations du système OpenNMT/LDJ-fr-so-ABC

V.3.2 Résultats

V.3.2.1 Récapitulatif et commentaires des résultats de l'évaluation des deux systèmes de TA neuronale français-somali

Système	Corpus	#segments	#mots	BLEU	TER	mPED/p	Tpe/p
OpenNMT	LDJ-A	12 863	194 135	0,11	1,03	555i	25mn
OpenNMT	LDJ-ABC	23 958	334 705	0,17	0,83	511i	23mn

Tableau 37 : Résultats des systèmes de TA neuronale

Les résultats de nos deux expérimentations de données bilingues post-éditées avec un système de TA neuronale montrent une légère amélioration de 6% du score BLEU, et une baisse du score TER de 20% entre le système spécialisé et le système augmenté.

Cette situation s'explique par l'augmentation de la taille de nos deux données bilingues d'apprentissage : en effet, aux 12 863 segments spécialisés, nous avons ajouté 11 095 segments « hors domaine » (soit environ 110% de plus) dans le système augmenté, par rapport à nos données bilingues initiales.

En dépit du fait que ces nouvelles données n'étaient pas des traductions français-somali, mais des traductions de segments anglais vers le français et le somali, ensuite alignées, et qu'elles n'étaient pas du même sous-langage (journalistique), l'ajout de ces données supplémentaires a permis d'augmenter les deux scores de nos mesures d'évaluation objective pour la TA neuronale français-somali.

Toutefois, les deux systèmes de TA neuronale obtiennent les deux plus mauvais scores en terme du temps de post-édition, par rapport aux systèmes de TA statistiques construits avec MOSES.

Si nous voulons utiliser nos différents systèmes de TA spécialisée pour poursuivre la tâche de construction d'un vraiment grand corpus bilingue français-somali, de haute qualité, nous utiliserons sans doute les systèmes construits avec MOSES, car ils permettront à nos post-éditeurs de traduire mieux qu'avec OPENNMT, et surtout en moins de temps.

Conclusion du chapitre V

Pour conclure, regroupons les résultats précédents dans un même tableau. Les meilleurs résultats obtenus en termes de mesure d'évaluation (BLEU) et le temps de post-édition par page standard (Tpe/p) sont coloré en rouge dans le Tableau 38 ci-dessous.

Système	Corpus	#segments	#mots	BLEU	TER	mPED/p	Tpe/p
Moses	LDJ-fr-so-ABC	23 958	299 952	**0,34**	1,01	377i	**17mn**
Moses	LDJ-fr-so-A	12 863	164 764	**0,32**	1,15	422i	**19mn**
OpenNMT	LDJ-fr-so-ABC	23 958	299 952	0,17	0,83	555i	25mn
GT	LDJ-TEST	643	9876	0,18	0,73	400i	**18mn**
OpenNMT	LDJ-fr-so-A	12 863	164 764	0,11	1,03	511i	23mn

Tableau 38 : Récapitulatif global des résultats des différents systèmes de TA français-somali

Nous avons donc pu construire à partir de notre très bon corpus LDJ-fr-so-A deux systèmes de TA de type Moses qui sont objectivement meilleurs que GT. Cela confirme le « métathéorème » CAQ : « en tout automatique (automaticité A = 100%), il est impossible d'avoir à la fois une bonne couverture (C) et une bonne qualité (Q). Métaphoriquement[60] , si deux des termes sont proches de 90%, le troisième sera proche de 10%. »

En particulier, un système de TA statistique construit à partir d'un très bon corpus de relativement petite taille (400 pages, 100 000 mots) s'avère d'emblée meilleur que le système GT, très généraliste. Pour un système neuronal, il faut un corpus 2 fois plus grand pour dépasser significativement GT en termes de BLEU.

Nous confirmons aussi que la tentative d'augmenter un très bon corpus de référence dans un corpus plus gros dans le cadre de développement d'un système de TA statistique pour une langue peu dotée, n'améliore que très peu les scores des mesures d'évaluation. Des expériences passées ont d'ailleurs montré que, si on fait croître le « bon corpus », l'ajout d'autres données « hors sous-langage » fait plutôt diminuer les scores.

Quoi qu'il en soit, dans notre cas, on se retrouve avec un système un peu meilleur que GT, et surtout possiblement utilisable pour faire grossir le « bon corpus » LDJ-fr-so-A par post-édition collaborative bénévole et ainsi, comme pour le français-chinois, on peut raisonnablement espérer arriver à un temps total de post-édition Tpe de 10-12 mn/page, bien meilleur que GT (19 mn/p).

Dans le meilleur des cas, nous avons construit un système qui demande environ 17 minutes de PE par page standard, un tout petit peu moins qu'avec GT par rapport à nos systèmes de TA construits sur des données spécialisés, ce que nous évaluons à 13,5/20 environ.

Cependant, un lecteur somalophone mais non francophone ne peut pas encore comprendre avec précision (et fiabilité) un article du journal *La Nation de Djibouti* traduit par un système de TA. Il faudrait atteindre une bien meilleure qualité. Cela semble possible, à terme, car les lecteurs djiboutiens du journal, qui sont somalophones et francophones, pourraient facilement et surtout rapidement contribuer, chacun de façon occasionnelle et opportuniste, à l'amélioration continuelle du système, et donc à l'augmentation immédiate de la qualité (informationnelle et linguistique) de la version somalophone, en temps réel — car on n'a pas envie de lire le journal d'hier, et encore moins d'avant-hier !

[60] On peut mesurer la qualité d'usage à partir du temps de postédition, mais il existe seulement des pistes pour mesure la « couverture ». Que représente un « sous-langage » donné (construit ou observé) par rapport à « toute la langue », si même ce terme a un sens ?

Conclusions et perspectives

Dans le cadre de cette thèse, nous avons abordé la question de l'informatisation des langues peu dotées de l'espace francophone en Afrique, et plus particulièrement celle du somali, langue très faiblement dotée parlée en République de Djibouti et dans 3 autres pays de la Corne d'Afrique.

La question principale que nous avons traitée dans cette thèse était de trouver et prouver comment on peut répondre à un besoin réel et précis pour les locuteurs somalophones de Djibouti et du continent, à savoir, accéder en somali aux articles en français publiés quotidiennement sur le site web du journal *La Nation de Djibouti*.

La première partie de notre travail a consisté à étudier et présenter l'état de l'art de l'informatisation du somali. Pour cela, nous avons recensé l'ensemble des ressources, outils de base, applications et services créés ou construis pour la langue somalie à ce jour. Cet état de l'art nous a permis d'avoir un aperçu global sur les besoins d'informatisation de cette langue, et de proposer et appliquer une méthodologie et une nouvelle stratégie pour construire les ressources langagières nécessaires au déploiement d'un système de TA indépendant et spécialisé.

La stratégie que nous avons utilisée a permis de construire un tout premier corpus bilingue français-somali de 98 912 mots (environ 400 pages standard), réalisé à partir des post-éditions des pré-traductions de *Google Translate* (GT) sous SECTra_w/iMAG, et de très bonne qualité.

Pour évaluer subjectivement la qualité de ce corpus bilingue, nous avons réalisé une enquête d'évaluation et d'auto-notation sur un échantillon de segments bilingues. Les résultats de cette enquête sont très encourageants, et indiquent la bonne qualité linguistique du corpus post-édité.

La seconde partie de notre travail a été consacrée à la construction de deux systèmes de TA français-somali basés sur LDJ-fr-so-A et sur d'autres données bilingues constituées à partir du web, LDJ-fr-so-B et LDJ-fr-so-C. Le système étendu a été construit à partir de leur union, appelée LDJ-fr-so-ABC. Nous avons utilisé deux outils libres de droits dont l'un utilise l'approche de TA probabiliste (*Moses*) et l'autre l'approche de la TA neuronale (*OpenNMT*).

Les évaluations faites sur les résultats de TA de ces deux systèmes ont montré clairement que la construction d'un ou plusieurs systèmes de TA spécialisés à partir d'un très bon corpus relativement de petite taille produisait des performances meilleures que celles de GT en termes de mesure BLEU et de temps de post-édition.

Les perspectives de notre recherche sont multiples. D'abord, nous souhaitons poursuivre la construction du corpus bilingue post-édité LDJ-fr-so-A, en fédérant une communauté de post-éditeurs bénévoles, afin d'augmenter sa taille, tout en maintenant sa qualité, et d'améliorer notre système de TA incrémentalement. Nous désirons expérimenter par la suite d'autres approches de TA plus « vectorielles », avec nos données annotées morpho-syntaxiquement.

Nous prévoyons également de mettre à la disposition de la communauté l'ensemble des ressources construites dans le cadre de cette thèse.

Enfin, nous comptons définir un cadre approprié pour valoriser notre expertise en TA français-somali, et adapter notre méthode pour développer d'autres systèmes de TA entre le français et au moins une langue peu dotée de l'espace francophone, en commençant par l'afar, langue couchitique peu dotée de la Corne de l'Afrique.

Bibliographie

1. [Abdillahi N. et al., 2007] Abdillahi, Nimaan & Nocera, Pascal & Torres-Moreno, Juan-Manuel. (2018). Boîte à outils TAL pour des langues peu informatisées : le cas du somali.

2. [Abdullahi, 1996] Diriye Abdullahi, Mohamed. *Parlons somali*. Paris: L'Harmattan, 1996.

3. [Assowe, 2011] Assowe, Houssein Ahmed. *Étude linguistique et approches de l'étiquetage morphosyntaxique du somali*. Mémoire de Master 2. Université Michel de Montaigne, Bordeaux 3, 2011.

4. [Avramidis, 2012] Avramidis, E., 2012. *Comparative Quality Estimation: Automatic Sentence-Level Ranking of Multiple Machine Translation Outputs*. *Proceedings of 24th International Conference on Computational Linguistics*, (December 2012), pp.115–132. Available at: http://www.aclweb.org/anthology/C12-1008.

5. [Banerje & Lavie 2005] Satanjeev Banerjee and Alon Lavie. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In Proc. of ACL WIEEMMTS, 2005.

6. [Barillot, X., 2002] Barillot, X. (2002). Morphophonologie gabaritique et information consonantique en somali et dans les langues est-couchitiques. Thèse de Doctorat. Université Paris 7.

7. [Berment, 2004] Vincent Berment. *Méthodes pour informatiser des langues et des groupes de langues peu dotées.* Thèse de doctorat, Université Joseph Fourier, Grenoble 1, 2004.

8. [Besacier L., 2014] Besacier, L. *Traduction automatisée d'une œuvre littéraire : une étude pilote.* Proc. Traitement Automatique du Langage Naturel (TALN). Marseille, France, 2014.

9. [Besacier & al., 2012] Besacier L., Lecouteux B., Azouzi M. & Luong Ngoc Q. (2012). *The LIG English to French Machine Translation System for IWSLT*. In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT).

10. [Blanchon & Boitet, 2007] Blanchon, H., & Boitet, C. *Pour l'évaluation des systèmes de TA par des méthodes externes fondées sur la tâche*. Journée sur la Traduction Automatique organisée par l'ATALA. Paris. 1er décembre 2007. Jeu de 27 transparents en complément de l'article TAL vol. 48(1), 2007.

11. [Boitet, 2007] Christian Boitet. *Corpus pour la TA : types, tailles et problèmes associés, selon leur usage et le type de système*. Revue française de linguistique appliquée 2007/1 (Vol. XII), p. 25-38, 2007.

12. [Boitet, Bellynck et al., 2008] Boitet, C., Bellynck, V., Mangeot, M. and Ramisch, C. *Towards Higher Quality Internal and External Multilingualization of Web Sites.* Proc. ONII-08 (Summer Workshop on Ontology, NLP, Personalization and IE/IR) IITB, Mumbai, Inde. 8 p., 2008.

13. [Boitet, Huynh et al., 2010] Boitet, C., Huynh, C.-P., Nguyen, H.-T. and Bellynck, V. The iMAG concept: multilingual access gateway to an elected Web site with incremental quality increase through collaborative post-edition of MT pretranslations. Proc. Traitement Automatique du Langage Naturel (TALN). Montréal, Canada, 2010.

14. [B.W. Andrzejewski, 1974] B. W. Andrzejewski. *The Introduction of a National Orthography for Somali*. African Language Studies, 15, 1974, pp. 199-203.

15. [Breton, 2003] Breton R.J.L., Mazoyer K. *Atlas des langues du monde: une pluralité fragile.* Paris: Éditions Autrement, 2003.

16. [Brown, Pietra et al., 1993] Brown, P. F., Pietra, V. J. D., S. a. D. and Mercer, R. L. *The mathematics of statistical machine translation: Parameter estimation.* Computational Linguistics 19(2): pp. 263-311, 1993.

17. [C. Quadir & N. Awde, 1999] *Somali-English/English-Somali Dictionary & Phrasebook* (Hippocrene Dictionary & Phrasebook). Hippocrene Books. 1999.

18. [Cabdalla, 1999] Cabdalla C. Mansuur, Puglielli A. *Barashada Naxwaha af Soomaaliga: a Somali school grammar.* London, England: HAAN Associates, 1999. 304 p.

19. [Callison-Burch et al., 2006] Callison-Burch, C., Osborne, M. and Koehn, P. *Re-evaluating the Role of BLEU in Machine Translation Research.* Proc. EACL-2006, pp. 249-256, 2006.

20. [Carbonell et al., 2006] Carbonell, Jaime G.; Lavie, Alon; Levin, Lori; and Black, Alan. *Language Technologies for Humanitarian Aid*. 2006. Institute for Software Research. Paper 394.

21. [Cettolo et al., 2012] Mauro Cettolo, Christian Girardi, and Marcello Federico.*WIT*^{3}*: Web inventory of transcribed and translated talks.* In Proceedings of EAMT-2012.

22. [Chiang, 2007] David Chiang. *Hierarchical phrase-based translation.* Computational Linguistics, 33(2):201–228, 2007.

23. [Do D., 2011] Extraction de corpus parallèles pour la traduction automatique depuis et vers une langue peu dotée. Thèse, Université de Grenoble. 2011.

24. [Doddington, 2002] George Doddington. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurence Statistics.* In Human Language Technology: Notebook Proceedings, pp. 128–132, 2002.

25. [Enguehard & Mbodj, 2004] Chantal Enguehard, Cherif Mbodj. *Des correcteurs orthographiques pour les langues africaines.* Bulletin de linguistique appliquée et générale (BULAG), 2004, pp. 51-68.

26. [Enguehard, 2005] Enguehard, C. (2005). Des correcteurs orthographiques pour collecter et diffuser les connaissances linguistiques en Afrique subsahararienne. 27th Internationalization and Unicode Conference, atelier « Unicode and Language Support in Francophone Africa », Avril 2005, Berlin, Allemagne. 2005.

27. [Ethnologue, 2017] *« Ethnologue: Languages of the World ».* https://www.ethnologue.com/. Consulté le 11 novembre 2017.

28. [Espla-gomis et al., 2009] Miquel Esplà-Gomis. *Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites.* In Proceeding of MT Summit XII, Ottawa, Canada. Association for Machine Translation in the Americas, ed., 2009.

29. [Falaise, A. et al., 2011] Falaise, A., Tutin, A., & Kraif, O. (2011). *Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques*. Traitement Automatique des Langues Naturelles, 88.

30. [Farah, A.G., 2008] Dictionnaire français - somali = Qaamuus faransiis - af soomaali. Paris: L'Harmattan, 2008.

31. [Farah, A.G., 2009] Dictionnaire somali-français = Qaamuus af soomaali-faransiis. Paris, Montréal: L'Harmattan, 2009.

32. [Federico et al., 2008] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. *IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models.* In Proceedings of Interspeech, pp. 1618–1621, Melbourne, Australia, 2008.

33. [Gamon et al., 2005] Gamon, M., Aue, A. & Smets, M., 2005. *Sentence-level MT evaluation without reference translations: Beyond language modeling*. Proceedings of the European Association for Machine Translation, pp.103–111. Available at: http://www.mt-archive.info/EAMT-2005-Gamon.pdf.

34. [Gupta et al., 2013] Gupta, R., Joshi, N. & Mathur, I., 2013. *Quality Estimation of English-Hindi Outputs using Naïve Bayes Classifier*. *arXiv preprint*, pp.3–6.

35. [Hajlaoui N., 2014] Hajlaoui, N., (2014). SMT for restricted sublanguage in CAT tool context at the European Parliament. ASLING, 2014.

36. [H.A Assowe, 2013] Ahmed Assowe, Houssein. (2013). Approche pour un premier étiqueteur morphosyntaxique d'une langue très peu dotée: le cas du somali. CEC-TAL'13. 23-27 Septembre, Montréal, 2013.

37. [H. Davis et al., 1952] H. Davis, K. *Automatic Recognition of Spoken Digits*. The Journal of the Acoustical Society of America. 1952.

38. [Hajlaoui & Boitet, 2008] Hajlaoui, N. and Boitet, C. (2008). *TA statistique à petits corpus pour des petits sous-langages.* Proc. TOTh-2008 Conférence sur Terminologie & Ontologie : Théories et Applications. 20 p., 2008.

39. [Huynh & al., 2008] Huynh, C.-P., C. BOITET et H. BLANCHON (2008). SECTra_w: an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. Proc. LREC-08, Marrakech, 8 p.

40. [C. Phab Huynh, 2010]: Huynh, C.-P. Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimédia. Thèse de doctorat, Université Joseph Fourier, 2010.

41. [Kalitvianski, Boitet et al., 2012] Kalitvianski, R., Boitet, C. and Bellynck, V. *Collaborative Computer-Assisted Translation Applied to Pedagogical Documents and Literary Works.* COLING (Demos): 255–260, 2012.

42. [Kalitvianski et al., 2015] Ruslan Kalitvianski, Valérie Bellynck, Christian Boitet. *Multilingual Access to Educational Material Through Contributive Post-editing of MT Pre-translations by Foreign Students.* The 14th International Conference on Web-based Learning, Jan 2017, Guangzhou, China. Advances in Web-Based Learning, ICWL 2015.

43. [Kalitvianski et al., 2016] Ruslan Kalitvianski, Lingxiao Wang, Valérie Bellynck, Christian Boitet. *An Aligned French-Chinese corpus of 10K segments from university educational material.* Proc. 3rd Workshop on Natural Language Processing Techniques for Educational Applications, Dec. 2016, Osaka, Japan. 2016.

44. [Klein & al., 2017] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A.-M. *OpenNMT: Open-Source Toolkit for Neural Machine Translation*, 2017. 67-72. 10.18653/v1/P17-4012.

45. [Koehn et al., 2007] Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. *Moses: Open source toolkit for statistical machine translation.* In Proc. ACL'07, pp. 177–180, 2007.

46. [Kukich K., 1992] Kukich K. 1992. "*Techniques for automatically correcting words in text*". Computing Surveys, 24(4): 377–439.

47. [Lecarme, 2002] Lecarme, Jacqueline, 2002. "Gender 'Polarity': Theoretical Aspects of Somali Nominal Morphology". In: Paul Boucher (ed.), Many Morphologies, Somerville, Mass.: Cascadilla Press, pp. 109–141.

48. [Leroux, R., 1998] Leroux R., Le Réveil de Djibouti 1968-1977. Simple outil de propagande ou véritable reflet d'une société ?, Paris, L'Harmattan, 1998.

49. [Levenshtein, 1966] Levenshtein, 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet physics doklady, 10, pp.707–710.

50. [Mangeot et al., 2011] Mathieu Mangeot, Chantal Enguehard. *Informatisation de dictionnaires langues africaines-français.* Actes des journées LTT 2011, Villetaneuse, France. 11 p., 2011.

51. [Morin, 1986] Didier Morin. *Le parcours solitaire de la Somalie*. Politique africaine, no 23. Paris, Karthala, septembre 1986, p. 57-66.

52. [Nasredine et al., 2016] Nasredine Semmar, Othman Zennaki, Mariama Laib. *Etude de l'impact d'un lexique bilingue spécialisé sur la performance d'un moteur de traduction à base d'exemples*. Proc. Traitement Automatique du Langage Naturel (TALN). Paris, France, 2016.

53. [Navigli, R. & Ponzetto, S. P., 2010] Navigli, R. and Ponzetto, S. P. (2010). *BabelNet: Building a Very Large Multilingual Semantic Network*. Proc. Artificial Intelligence, Elsevier, pp. 217-250.

54. [Negri et al., 2012] Negri, M. et al., 2012. *Match without a Referee: Evaluating MT Adequacy without Reference Translations*. Proceedings of the 7th Workshop on Statistical Machine Translation, pp.171–180.

55. [Nimaan, 2007] Sauvegarde du patrimoine oral africain : conception de système de transcription automatique de langues peu dotées pour l'indexation des archives audio. Thèse de doctorat, Université d'Avignon et des Pays du Vaucluse, Avignon, France, juillet 2007.

56. [Och F., J., 2002] Och, F. J. (2002). *Statistical machine translation: from single-word models to alignment templates*. Thèse de doctorat, Bibliothek der RWTH Aachen.

57. [Och, 2003] Och, F.J. 2003. *Minimum error rate training in statistical machine translation.* In Proc. ACL'03, volume 1, pp. 160–167, Sapporo, Japan, 2003.

58. [Osborn, 2011]: Osborn, D. (2011). *Les langues africaines à l'ère du numérique*. Laval, Canada: Presses de l'Université Laval.

59. [Paice & Chris D., 1990] Paice Chris D. "*Another stemmer*". ACM SIGIR Forum, Volume 24, No. 3. 1990, 56-61.

60. [Papineni et al., 2002] Papineni, K., S. Roukos, T. Ward, and W.J. Zhu. *BLEU: A method for automatic evaluation of machine translation.* In Proc. ACL'02, pages 311–318, Philadelphia, USA, 2002.

61. [Pineau & Boitet, 2004] Pineau, M. & Boitet, C., 2004. *Comparaison de résultats de traduction (automatique ou non)*. Université Joseph Fourier. Available at: https://www.ujf-grenoble.fr.

62. [Porter, 1980] M. F. Porter. *An Algorithm for Suffix Stripping.* Program, 14 (3), 30–137, 1980.

63. [Potet, 2013] Marion Potet. Vers l'intégration de post-éditions d'utilisateurs pour améliorer les systèmes de traduction automatique probabilistes. Thèse de doctorat, Université de Grenoble, 2013.

64. [Pouliquen et al., 2013] Pouliquen, B., Elizalde, C., Junczys-Dowmunt, M., Mazenc, C. and Garcia-Verdugo, J. (2013). *Large-scale multiple language translation accelerator at the United Nations*. Proc. MT Summit. Nice, France.

65. [Esperança-Rodier E. & al., 2018] Esperança-Rodier, E., Becker, N. (2018). *Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs*. Proc. Plate-Forme Intelligence Artificielle. Nancy, France.

66. [Rubino, 2012] R. Rubino, S. Huet, F. Lefèvre, and G. Linarès. *Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique*. Proc. Conférence en Traitement Automatique des Langues Naturelles (TALN-2012), pp. 527-534, Grenoble, 2012.

67. [Saeed, 1999] *Somali (London Oriental and African Language)*. Johns Benjamin Publishing Company, ISBN 10: 9027238103, 1999.

68. [Saeed,1984] Saeed, John I. *The syntax of focus & topic in Somali*. Cushitic language studies, Kuschitische Sprachstudien, Helmut Buske Verlag, Hamburg, 1984.

69. [Scannel, K., 2007] Kevin Scannel, *The Crúbadán Project: Corpus building for under-resourced languages*, in "Building and Exploring Web Corpora": Proceedings of the 3rd Web as Corpus Workshop, Louvain-la-Neuve, Cahiers du Cental **4** (2007), pp. 5-15.

70. [Shah R. et al., 2015] Shah, Ritesh, Christian Boitet, Pushpak Bhatacharyya, Mithun Padmakumar, Leonardo Zilio, Ruslan Kalitvianski, Mohammad Nasiruddin, Mutsuko Tomokiyo & Sandra Castellanos Paez. *Post-editing a chapter of a specialized textbook into 7 languages: importance of terminological proximity with English for productivity.* In Proceedings of the 12th International Conference on Natural Language Processing (ICNLP-2015).

71. [Snover et al., 2006] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. *A Study of Translation Edit Rate with Targeted Human Annotation.* Proceedings of Association for Machine Translation in the Americas, 2006.

72. [Steinberger et al. 2006]: Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, Dániel Varga (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

73. [Tiedemann, 2012] Jörg Tiedemann. *Parallel data, tools and interfaces in OPUS.* In Proc. LREC-2012, Istanbul, Turkey.

74. [Tomokiyo, M. et al., 2000] Tomokiyo, M., Mangeot, M. and Planas, E. (2000). *Papillon : a Project of Lexical Database for English, French and Japanese, using Interlingual Links*. Proc. Journées Science et Technologie (JST-2000), Tokyo, 3 p.

75. [Van der Verken et al., 2003] Van Der Veken, A., de Schryver, G.-M., « *Les langues africaines sur la Toile: étude des cas haoussa, somali, lingala et isixhosa* ». Les cahiers du RIFAL n°23, pp.33-45, novembre 2003.

76. [Vogel et al.1996] Stephan Vogel, Hermann Ney, and Christoph Tillmann. *HMM-based Word Alignment in Statistical Translation.* COLING, pp. 836–841, Copenhagen, August 1996.

77. [Wagner & Fischer, 1974] Wagner, R.A. & Fischer, M.J., 1974. *The String-to-String Correction Problem*. Journal of the ACM, 21(1), pp.168–173.

78. [Wang, 2015]: Wang, H. (2015). Évaluation comparative de la qualité d'usage de plusieurs systèmes de TA français-chinois en fonction de la tâche de post-édition. Mémoire de M2R, Grenoble, Université Grenoble Alpes.

79. [Wang L. X, 2015] Lingxiao Wang. Outils et environnements pour l'amélioration incrémentale, la post-édition contributive et l'évaluation continue de systèmes de TA. Application à la TA français-chinois. Thèse de doctorat, Université Grenoble Alpes, 2015.

80. [Wang L. & C. Boitet, 2013] L. Wang, C. Boitet. Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost. Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, Sept. 2, 2013, p. 103–110.

81. [Zhang, Y. & Mangeot, M., 2013] Zhang, Y. and Mangeot, M. (2013). *Bases lexicales multilingues : traitement des acronymes*. Proc. JCLTT 2013, Bruxelles, 16 p.

Chapitre VI Annexes

Voici en annexe des extraits de la TA et de la post-édition de notre corpus de test, traduit automatiquement par Google Translate (GT), et par nos 4 systèmes de TA. Il y a 105 segments (1824 mots au total), allant du très court au très long (souvent à cause de segmentations incorrectes).

Après ces extraits, on trouvera une monographie de l'auteur sur la morphosyntaxe du somali.

VI.1 Extraits de la traduction des 643 segments de notre corpus de test corpus_fr-so_LDJ-TEST avec Google Translate

Numéro du segment	Source	Segment PE_initial	Segment traduit avec GT
1	© Copyright 2017 — La Nation.	xuquqda la dhowray 2017- La Nation	© Copyright 2017 - Waddanka.
2	plateforme du Projet " E Campus " de l'Université de Djibouti : une vocation régionale 2281 views	hagaajinta mashruuca " E-CampUs " ee jamacadda jabuuti : Ujedoyin heer gobol ah	Mashruuca "E Campus" ee Jaamacadda Jabuuti: xirfad goboleed 2281 views
3	il s'est ensuite entretenu avec M. Wang, secrétaire-général Adjoint de l'UIT, candidat à la succession du Dr Touré.	ka dibna wuxuu la kulmay M. Wang, ku xigeenka xoghayaha guud ee ITU-ta, ahna musharax in uu badalo rabba Dr Touré.	Kadib wuxuu la hadlay Mr Wang, Ku-xigeenka Xoghayaha Guud ee ITU, Musharax u tartamidda Dr Touré.
4	ce qui permettrait de dépasser les petites activités de commerce peu rentables et ayant une faible valeur ajoutée.	taas waxay suurta galin in laga gudbo hawlaha ganacsi ee yar yar ee waxka tar leh qiimaha ku daray hooseeyo.	Tani waxay suurtogal ka dhigeysaa in ay ka baxaan ganacsiyada yar yar, kuwa aan faa'iido doonka ahayn iyo kuwa hooseeya.
5	ce qui signifie pratiquement la disponibilité d'équipements adéquats, de protocoles claires et d'un personnel parfaitement formé et préparé.	tani waxay ka dhigan tahay ficil ahaan helitaanka qalab ku filan, hab-cad iyo shaqaalaha oo si fiican u tababaran oo diyaarna ah.	taas oo micnaheedu yahay helitaanka qalab ku filan, qawaaniin cadcad iyo shaqaale tababaran oo diyaar ah.
6	la fondation « roi Salman » apporte un appui significatif au ministère de la Santé	aasaaska " Boqor Salman " ayaa waxay siisay taageero weyn Wasaaradda Caafimaadka	King Salman Foundation waxay bixisaa taageero macne leh oo wasaarada caafimaadka
7	Thursday, April 30th	Khamiis 30-ka Abriil	Khamiis, 30-ka April
8	catégories	qaybo	qaybaha
9	Publicité	Xayaysiis	advertising
10	sa croissance a été spectaculaire : il compte désormais 150 États contractants, avec près de 220 ' 000 demandes déposées en 2015, et continue de croître à un rythme soutenu.	Kobaceega ayaa ahaa mid cajiib ah : hadda wuxu leeyahay 150 dawladood oo xubin ka ah, iyada oo ku dhowaad 220,000 oo dalabyo la soo gudbiyey 2015, kaas oo sii waday inay koraan si xawli leh.	kobaca ayaa cajiib ah, waxa ay haatan u leedahay 150 iibsanaya Mareykanka, iyada oo ku dhowaad 220 '000 codsiyada gudbiyay ee 2015, oo ku sii socda si ay u koraan at xawaaraha joogto ah.

11	Voeux du Président de la République pour l’année 2016	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyo ka socda Madaxweynaha Jamhuuriyadda ee 2016
12	Name (required)	magac (waajib)	Magaca (loo baahan yahay)
13	le 23 mars 1975, un commando du Front de la Libération de la Côte Française des Somalis prend en otage l’ambassadeur de France à Mogadiscio, M. Jean Guery.	23-kii Maarso sannadkii 1975, koox ka taliye ka tirsan Jabhadda Xoreynta ee Xeebaha Soomaaliya ee Faransiiska ayaa safiirkii Faransiiska ee Jean-Guyane, oo ahaa afduub loo haystay.	23-kii Maarso, 1975-kii, taliye ku-xigeenka Jabhadda Xoreynta ee Xeebaha Soomaaliya ee Faransiisku waxay qabsadeen Safiirkii Faransiiska ee Muqdisho, Mudane Jean Guery.
14	Follow Me	I soo raac	Raac aniga
15	E-mail :	E-mail :	E-mail:
16	Sports	Sbortiga	Sports
17	en attendant, le pays importe actuellement 65 % d’énergie hydroélectrique via la ligne d' interconnexion électrique avec l' Ethiopie voisine.	inta taas la sugayo, dalka hadda wuxu dibadda ka keensadaa 65 % korontada biyaha iyada uu kenno khadka isku xidhan ee korontada ee dalka deriska ee Ethiopia.	isla mar ahaantaana waddanku waxa uu hadda soo dejiyey 65% awoodda korantada iyada oo loo marayo khadka xidhiidhinta korontada ee Itoobiya la deriska ah.
18	© Copyright 2017 — La Nation.	xuquqda la dhowray 2017-La Nation	© Copyright 2017 - Waddanka.
19	il s' agit là d' une initiative de Djibouti destinée à jeter les bases des « Jeux de l' IGAD » en en élaborant la charte.	tani waa hindise Jabuuti ka yimid si ay u sameyso asaaska qorshaha " Ciyaraaha IGAD " iso si lo dhigo dastuurka.	waa qorshe Jabuuti loogu talagalay in la aasaaso aasaaskii loogu talagalay "IGAD Ciyaaraha" iyada oo loo marayo axdiga.
20	recherche	Badhis	search
21	ministère des affaires musulmanes / FNUAP : seconde rencontre des leaders religieux sur les MGF	Wasaaradda Arrimaha Islaamka / FNUAP : shirkii labaad ee Hogaamiyayaasha diinta ee ku saabsan Gudniinka Fircooniga ah	Wasaaradda Arrimaha Muslinka / UNFPA: Shirkii Labaad ee Hogaamiyaasha Diinta ee FGM
22	ce jour là, les jeunes rotariens n' avaient pas manqué de remercier très chaleureusement, les responsables du Centre Mère et Enfant et plus particulièrement, la Première dame du pays qui s' est très impliquée en faveur du bien-être des enfants ainsi que dans celui des enfants vivants avec des handicaps comme les sourds et malentendants.	maalintaas, dhallinyarada Rotarianka kama ay moogin inay ugu mahad celiyaan si aad u diiran, saraakiisha Xarunta Hooyada iyo Carruurta iyo gaar ahaan marwadda kobaad taas oo aad ugu lug lahayd fayo-qabka carruurta iyo ka carruurta naafada ah sida kuwa dhagaha ama maqalka la.	maalintaas, dhallinyarada Rotarians n 'weli ma ay ku fashilantay in ay si daacad ah ugu mahad naqayaa madaxda Xarunta Hooyada iyo Ilmaha iyo gaar ahaan, Marwada Koowaad ee dalka s ah' waa mid aad u lug ah ee fayo-qabka ah ee carruurta iyo kuwa carruurta naafada ah sida dhegoolaha iyo maqalka adag.

23	enfin, les hommes d' affaires des deux continents ont décidé de soumettre les conclusions de leurs travaux de deux jours à l' aval des chefs d' Etat et de gouvernements d' Afrique et de l' Union Européenne qui participent au 4ème sommet afro-européen.	ugu dambeyntiina, ganacsatada ka kala yimid labada qaaradood waxay go ' aansadeen in ay soo gudbiyaan gabagabadii shaqadoodii labada cisho oo ay taageereen hoggaamiyayaasha dawladaha iyo xukumadaha Afrika iyo Midowga Yurub ee ka qaybqaatay shirkii 4aad ee Afrika-Yurub.	Ugu dambeyntii, ka ganacsatada oo ka mid ah laba qaaradood ay go'aansadeen in ay soo gudbiyaan natiijada shaqadooda laba maalmood geysto madaxda State iyo Dowladda ee Africa iyo Midowga Yurub ee ka qeyb 4aad Summit Afro- Midowga Yurub.
24	celles-ci prennent généralement contact avec les dirigeants d' Hormar.	waxay guud ahaan la xiriiraan hogaamiyaasha Hormar.	kuwani waxay badanaa la xiriiraan hogaamiyaasha Hormar.
25	Tweet	Tweet	Tweet
26	s' exprimant à son tour, le ministre de l' Habitat, de l' Urbanisme et l' Environnement, Mohamed Moussa Ibrahim, a indiqué que les changements climatiques constituent l' un des défis majeurs du 21e siècle et qui ont des impacts significatifs sur notre continent.	isagoo ka hadlayay ayaa isgauna, Wasiirka Guryeynta, Qorshaynta Magaalada iyo Degaanka, Qorshaha iyo Deegaanka, Mohamed Moussa Ibrahim, ayaa sheegay in isbeddelka cimilada waa mid ka mid ah caqabadaha ugu waaweyn ee qarniga 21aad oo ay saamayn weyn ku qaaradda.	lagu muujiyay markeeda, Wasiirka Housing, Qorshaha Magaalada iyo Deegaanka, Mohamed Moussa Ibrahim, ayaa sheegay in isbedelka cimilada waa mid ka mid ah caqabadaha ugu waaweyn ee qarniga 21aad oo ay leeyihiin saamayn weyn ku qaaradeena.
27	iftar collectif à Yoboki	Afurka wadajirka ee Yoboki	Dhaqdhaqaaqa dhexdhexaad ah ee Yoboki
28	la parole à … Ali Guelleh Aboubaker Ministre auprès de la Présidence, chargé de l' Investissement	hadalkii … Cali Geelle Abubaker Wasiirka Madaxtooyadda mas ' uul ka ah malgeeshiga	Ergeyga... Cali Geelle Wasiirka Warfaafinta ee Madaxtooyada, oo mas'uul ka ah Maalgashiga
29	6 août 2017 9 h 01 min 0 comments Views : 9	6 agusto 2017	6 Ogosto 2017 9 h 01 min 0 comments Views: 9

30	cette assistance technique qui mobilise trois experts consultants a pour objectif d' appuyer le ministère dans la mise en œuvre de sa feuille de route par le biais d' une assistance technique avisée à la tenue des assises sur la décentralisation que celui-ci compte organiser, à l' élaboration de deux documents, l' un portant sur un avant-projet de « code des Collectivités Territoriales en République de Djibouti » et l' autre sur une stratégie visant le transfert de ressources vers les Conseils Régionaux, à partir du budget de l' Etat, de la fiscalité locale et de l' aide des bailleurs de fonds.	Gargaarkani farsamo oo abaabulay saddex taliyayaal ayaa looga golleeyahay in lagu taageero Wasaaradda hirgelinta tusmo hawleedkeega iyada loga gargaarayo xagga farsamadda iyado la sameyn fadhi guud oo ku sabsan mamul balaadhinta oo lagu tala-jiro in la abaabulo, sameynta laba dokumenti, mid ka mid ah oo ku saabsan xeer-qabyo ah " Xeerka Dawlada Hoose ee Jamhuuriyadda Jabuuti " iyo mid kale oo ku saabsan istaraatijiyad ku wareejista khayraadka Golayaasha Gobolka, ee miisaaniyadda dawlada, canshuurta degaanka iyo gargaarka deeq bixiyayaasha.	gargaar farsamo oo abaabushey saddex taliyayaal looga golleeyahay in lagu taageero Wasaaradda hirgelinta ay roadmap iyada oo gargaar ah xog farsamo assizes ku saabsan baahinta in ay tala-jira in uu abaabulo, in diyaarinta laba dokumenti, mid ka mid ah oo ku saabsan qabyo "code of Dawlada Hoose ee Jamhuuriyadda Jabuuti" a iyo kuwa kale oo ku saabsan istaraatijiyad for kala iibsiga khayraadka si Golayaasha Gobolka, miisaaniyadda Canshuurta dawlad-goboleedka, canshuuraha iyo deeq-bixiyeyaasha.
31	Apres un mois passé à Djibouti, je m’émerveille toujours autant de l' accueil des gens ici.	kadib markii aan bil joogay Jabuuti, wali waxaan la yaabanahay soo dhaweeyeenta dadka halkan.	Kadib markii ay ku qaadatay Jabuuti, weli waan la yaabay dadka soo dhaweynta ah.
32	elle reste un puissant levier grâce auquel chaque épargnant peut surmonter certaines difficultés lors d' une crise financière.	wuxu noqdaa kabaal awood leh oo qof kasta oo keydsadaa kaga gudbi karo dhibaatooyinka qaar ka mid ah intii lagu jiro xiisadda dhaqaale.	waxa ay ahaanaysaa awood xoog leh iyada oo ilaaliyeyaashu ka adkaan karaan dhibaatooyinka qaarkood inta lagu jiro xiisad dhaqaale.
33	et les sujets ne manquent pas.	Maadooyinkaasina maaha kuwo maqaan.	iyo maadooyinka aan haysan.
34	Pis, certains ménages sont surendettés indépendamment de leurs revenus.	Kaba si daran, qoysaska qaarkood way xog bay u daymaysan yihiin iyado loo eegin dakhligooga.	Waxaa sii xumaanaya, qoysaska qaar ayaa ka baxsan dakhliga iyada oo aan loo eegin dakhligooda.
35	un atelier sur le leadership et le développement des compétences de gestion s' est ouvert hier au sein de l' institut national d' administration publique (INAP).	aqoon iswaydarsi ku saabsan hogaaminta iyo horumarinta xirfadaha maamulka ayaa ka furmay shalay Machadka Qaranka ee Maamulka Guud (INAP).	seminar ku saabsan hoggaanka iyo horumarinta xirfadaha maamulka s 'furay shalay ee Machadka Qaranka ee Maamulka Guud (INAP).
36	et la France et ses entreprises ont des atouts, qu' il nous faut désormais davantage valoriser », a-t-il affirmé.	Faransiiskuna iyo shirkadihisu waxay leeyihiin faa ' iidoyin, ku habon in hadeed la qiimeyo ", ayuu yidhi.	iyo France iyo shirkadaha ay leeyihiin awooda, in aan waa in qiimaha hadda ka badan, "ayuu yiri.
37	Voeux du Président de la République pour l' année 2016	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyo ka socda Madaxweynaha Jamhuuriyadda ee sanadka 2016

38	cette importante conception de partage des tâches dans le système du PCT favorise la coopération internationale et la collaboration entre les offices délivrant des brevets dans de nombreux pays, une assistance leur étant offerte dans le cadre de leurs efforts visant à traiter les demandes de brevets de façon rationnelle et efficace.	Qodobkan muhiimka ah ee wadaagista shaqada ee nidaamka PCT-du wuxuu kor u qaadaa iskaashiga caalamiga ah iyo iskaashiga xafiisyada patentyada bixiya ee wadamo badan, iyaga oo siinaya caawimo dadaalkooda lagu xalinayo codsiyada patentka si aqoon iyo waxtar leh.	sharing this muhiim ah oo hawlaha design nidaamka PCT ku kor iskaashiga caalamiga ah iyo iskaashiga u dhaxeeya xafiisyo bixinta shatiyada dalal badan, taageerada ay qabka bixiyeen qayb ka ah dadaallada ay si wax looga qabto codsiyada patent si caqli iyo waxtar leh.
39	Recevez les Dernières Informations Par Email - Abonnez-Vous Gratuitement !	Hel Macluumaadyadi ugu dambeyay dhinaca Emailka - Sajilad bilash ah !	Hel Qaybta Macluumaadka Ugu Dambeeya Email ahaan - Soo-gali Free!
40	la Parole à … Cheick Abdourahman Mohamed Ali alias Abdourahman Chamsudin Président du haut comité islamique de la Fatwa et secrétaire général du réseau des religieux de la région SHAMKAT	wareysiga … Shiikh Cabdiraxmaan Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin Madaxweynaha Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada dadka diinta ee gobolka SHAMKAT	Ereyga... alias Cheick Mohamed Ali Abdurahman Abdurahman Chamsudin Madaxweynaha Guddiga Islamic Sare ee Fatwa oo ah xoghayaha guud ee shabakada diinta ee SHAMKAT gobolka
41	Électrification rurale – Une nouvelle donne « Mern dans Électrification rurale – Une nouvelle donne	koronto geynta baadiya - arrin cusub	Qalabaynta Guryaha - Mashruuca Cusub "Mern ee Farsamooyinka Dhexdhexaadinta - Heshiiska Cusub
42	il s' agit de Safia Ali Said, bibliothécaire, récompensée dans la série « Innovation Administrative ».Dans le domaine de l' innovation pédagogique, Mme Hibo Moumin a été récompensée, ainsi que M. Abdoulkader, M. Abdallah Abass et M. Koffi.	waxa ka mid aha Safia Cali Saciid, oo ka shaqaysa maktabadda iyado lagu abaalmariyay qaybta " Hal-abuurnimada Maamulka ".Qaybta hal-abuurnimo waxbarasho, M. Hibo Mumin ayaa lagu abaal mariyay, iyo Md. Cabdulkader, Cabdala Cabaas iyo Md. Kofi.	waa Safiya Cali Siciid, oo ah laybareeriye, oo lagu abaalmariyay "Innovation Innovation".Dawladda Hibo Moumin ayaa lagu abaalmariyey, sidoo kale Mr. Abdoulkader, Abdallah Abass iyo Mr. Koffi.
43	Publicité	Xayaysiis	advertising
44	Commentaires récents	sharaxadi ugu dambeysay	Faallooyinka dhawaan
45	Voeux du Président de la République pour l' année 2016	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyo ka socda Madaxweynaha Jamhuuriyadda ee sanadka 2016
46	ne ratez plus les mises à jour de la Nation	ha ka maqnaan Wararka cusub ee jariidada Qaranka	ha seegin taariikhda qaranka
47	celles-ci sont aujourd' hui au nombre de 66, dont 58 en vigueur, 8 ayant été dénoncées par les autorités djiboutiennes.	kuwaas maanta tiradoodu waxay tahay 66, oo 58 ka mid ah ayaa dhaqan galay, 8 ka mid ah waxaa lagu eedeeyay mas ' uuliyiinta Jabuuti.	maanta waxaa jira 66 ka mid ah, oo 58 ka mid ah ay haystaan, 8-na waxaa lagu eedeeyay mas'uuliyiinta Jabuuti.
48	catégories	qaybo	qaybaha
49	Recevez Notre Newsletter	Hel Newsletterkeena	Hel wargeyskayaga

50	le point avec … Kaltoum Hassan Omar représentante d' une association d'engueilla	kulankii … Kaltum Xasan Cumar oo wakiil ka ah urur degmada engueilla	dhibicda... Kaltoum Xasan Cumar wakiil ka mid ah ururada nacaybka
51	la rénovation en mars 2017 de l' unité de néonatologie à l' hôpital Cheiko de Balbala répond à l' objectif national de réduire la morbi-mortalité néonatale.	dib-u-habeynta bishii maarso ee sannadka 2017 ee qaybta dhakhtarka dhallanka ee Isbitaalka Cheiko ee Balbala ayaa ka jawaabay hadafka qaran ee yaraynta xanuunada iyo dhimashada dhallaanka.	dib-u-habaynta Qeybta Cudurka 'Neonatology Unit' ee Isbitaalka Cheiko ee Balbala bishii Maarso 2017 ayaa la kulmay hadafka qaran ee dhimista dhalmada iyo dhimashada dhalmada.
52	de notre côté, nous bénéficions de points de vente idéalement positionnés, notamment à Djibouti et désormais aussi en province, d' un dépôt en propre pour l' aviation, et d' un contrat exclusif de distribution de Lubrifiants à la marque Total, réel expert dans le domaine des Lubrifiants.	dhinaceena, waxaynu ka faa ' iideysanaa goobaha iibka ku haboon, gaar ahaan Jabuuti iyo hadda gobolada sidoo kale, kaydka dhigashada ee u gaarka ah saliidaha, iyo qandaraas gaar ah ee waxaan leenahay boosaska boosaska, gaar ahaan Jabuuti iyo hadda gobollada, deebaajiga gaarka loo leeyahay ee duulimaadka, iyo qandaraas gaar ah oo loogu talagalay Lubrifiants ilaa calaamadda Wadarta khabiirka dhabta ah gudaha bakhaarada.	Dhinaceena, waxan ka faa'iideysan karnaa iibsashada goobaha laga iibsado, gaar ahaan Jabuuti iyo sidoo kale sidoo kale gobollada, iyada oo kaydka duulimaadka leh, iyo qandaraas gaar ah oo loogu talagalay qeybta Total, khabiir dhab ah. gudaha bakhaarada.
53	E-Campus – L' Université de Djibouti se dote d' une plateforme numérique 28257 views	E-Campus - Jamaacadda Jabuti waxay yeelatay boostejo kombuyutareysan	E-Campus - Jaamacadda Jabuuti waxay leedahay muuqaal dijitaal ah oo dhan 28257 views
54	Publicité& Annonces	Xayaysin & Qoritaanka	Raadinta iyo Digniinta
55	en cliquant sur l' un de ses liens, vous serez informé en temps réel.	ku riixida mid ka mid ah xiriiryadan, waxaa lagugu soo wargelinayaa si joogta ah.	adiga oo riixaya mid ka mid ah xiriiriyeyaashiisa, waxaa laguu sheegi doonaa waqtiga dhabta ah.
56	Economie	dhaqaalaha	dhaqaalaha
57	“ Une nécessaire vulgarisation des mesures multidimensionnelles de la pauvreté ”	" qalab lagama maarmaanka u ah tallaabooyinka kala duwan ee saboolnimada "	Qalab lagama maarmaanka u ah tallaabooyin fara badan oo saboolnimo ah
58	interrogé sur le contexte sécuritaire, le maire a précisé que la question relevait de l' intérêt national.	isago wax laga weydiiyay arrimaha ammaanka, duqa magaalada ayaa sheegay in arrintu ay ahayd danta qaranka.	Markii la weydiiyay xaaladda ammaanka, ayaa duqa ayaa sheegay in arrintu ahayd mid danta qaranka ah.

59	ces matériels sont essentiellement constitués de médicaments essentiels pour la santé de la mère et de l' enfant, d' équipements de protection, de stérilisateurs, de générateurs électriques, de tentes, de couvertures ou encore d' outils nécessaires aux services d' urgence et de réanimation.	qalabkani waxay ka kooban yahay daawooyin muhiim u ah caafimaadka hooyooyinka iyo carruurta, qalabka ilaalinta, sterilizers, koronto-dhaliyaha korontada, teendhooyinka, bustayaal, qalabyada adeegyada gurmadka iyo Dib-u-furista.	maadooyinkaani waxay ka kooban yihiin alaabooyin daaweyn oo muhiim u ah caafimaadka hooyada iyo ilmaha, qalabka ilaalinta, sterilizers, koronto-dhaliyaha korontada, teendhooyinka, busteyaalka ama qalabyada lagama maarmaanka u ah adeegyada gurmadka. dib u soo kabashada.
60	le projet cible les jeunes déscolarisés en chômage dans la ville d' Arta.	mashruucan waxa uu beegsanayaa dhalinyarada iskoolada ka baxday ee shaqo la ' aanta hayso ee magaalada Carta.	mashruucani wuxuu ka soo jeedaa dhallinyarada shaqo la'aanta ah ee magaalada Arta.
61	de revoir, non seulement ses actes et sa représentation du monde mais aussi, l' état d' esprit qui teinte ses démarches et l' image qu' il nous donne de lui.	in dib loo eego, ma aha oo kaliya falalkiisa iyo mataladiisa adduunka laakiin sidoo kale, xaaladda maskaxda taas oo xumeynaysa talaabooyinkiisa iyo sawirka uu iska bixinayo.	inuu dib u eego, ma aha oo kaliya ficilkiisa iyo matalaadiisa adduunka, laakiin sidoo kale, gobolka maskaxda kuwaas oo midabkiisa iyo muuqaalkiisa uu na siinayo.
62	il a par ailleurs mis l' accent sur les relations privilégiées qu' entretient notre pays avec cette agence onusienne.	waxa kale oo uu carabka ku adkeeyay xiriirka mudnaanta leh ee dalkeenna uu la leeyahay hay ' addan Qaramada Midoobay.	wuxuu sidoo kale xoogga saaray xiriirka khaaska ah ee wadankeena uu la leeyahay hay'adan UN.
63	Leave a Reply	ka tag jawaabtada	Ka tag hadal
64	Please try it manualy.	fadlan isku day dhinaca gacanta.	Fadlan tijaabi manualy.
65	VIH-Sida : Prudence est mère de sûreté	HIV-AIDS-ka : maqnaanshaha waa hooyada amniga	HIV-AIDS: Maqnaanshaha waa hooyada amniga
66	investissement étranger : Touchroad signe un mémorandum d' entente avec Djibouti 2 comments	maalgashiga Dibadda : Touchroad o la saxiixday diwan is-fahamin dawlada Jabuuti	maalgashiga shisheeye: Touchroad ayaa saxiixday heshiis saxeexa Jabuuti 2 faallooyin
67	l' ambassadeur américain Tom Kelly a été reçu hier par le ministre de la Communication chargé des postes et des Télécommunications, Abdi Youssouf Sougueh.	Safiirka Maraykanka Tom Kelly ayaa qabilay shalay Wasiirka Isgaarsiinta, qaabilsan Boostada iyo Isgaarsiinta, Cabdi Yusuf Suge.	Danjiraha Maraykanka Tom Kelly ayaa shalay la soo gaadhay wasiirka warfaafinta ee maamulka iyo Isgaadhsiinta, Cabdi Youssouf Sougueh.
68	Voeux du Président de la République pour l' année 2016	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyo ka socda Madaxweynaha Jamhuuriyadda ee sanadka 2016

69	le projet, mis en œuvre par l' ONG " Paix et Lait en partenariat avec l' association " Action Plus " d' Obock, vise à contribuer au développement national, notamment par l' amélioration de l' accès à l' eau potable des communautés et des ménages vulnérables de Balbala et d' Obock.	mashruucu waxa, fuliyey hay ' adda " Nabad iyo Canno " iyaga oo shuraako la ah ururka ' ee NGO ' Action More ' ' Obokh, ujeedadeedu tahay in ay ka qayb qaataan horumarinta qaranka, gaar ahaan iyadoo la sii wanaajinayo helidda biyaha ee bulshada iyo qoysaska nugul Balbala iyo Obokh la cabbo.	mashruuca, oo ay hirgelisay NGO "Peace and Coke" oo iskaashi la leh ururka "Action Plus" Obock, ujeedadeedu tahay in ay gacan ka geysato horumarinta qaranka, oo ay ku jirto wanaajinta helitaanka biyaha la cabbo bulshooyinka iyo qoysaska nugul ee Balbala iyo Obock.
70	de notre côté, nous bénéficions de points de vente idéalement positionnés, notamment à Djibouti et désormais aussi en province, d' un dépôt en propre pour l' aviation, et d' un contrat exclusif de distribution de Lubrifiants à la marque Total, réel expert dans le domaine des Lubrifiants.	dhinaceena, waxaynu ka faa ' iideysanaawaxaan leenahay boosaska boosaska, gaar ahaan Jabuuti iyo hadda gobollada, deebaajiga gaarka loo leeyahay ee duulimaadka, iyo qandaraas gaar ah oo loogu talagalay Lubrifiants ilaa calaamadda Wadarta khabiirka dhabta ah gudaha bakhaarada.	Dhinaceena, waxan ka faa'iideysan karnaa iibsashada goobaha laga iibsado, gaar ahaan Jabuuti iyo sidoo kale sidoo kale gobollada, iyada oo kaydka duulimaadka leh, iyo qandaraas gaar ah oo loogu talagalay qeybta Total, khabiir dhab ah. gudaha bakhaarada.
71	le ministre Aramis leur a exprimé sa reconnaissance pour leurs apports, individuels et collectifs, à la réalisation des activités de son département.	Wasiir Aramis ayaa wuxuu muujiyay sida uu ugu mahad naqay kaalintooda, shaqsiyaadka iyo wadajirka, si loo fuliyo hawlaha waaxda.	Wasiirka Aramis wuxuu muujiyay sida uu ugu mahad naqay tabarucdooda, shakhsiyan iyo mid wadajir ah, si loo xaqiijiyo hawlaha waaxda.
72	Voeux du Président de la République pour l' année 2016	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyo ka socda Madaxweynaha Jamhuuriyadda ee sanadka 2016
73	nuit du destin à Haramous 2010 views	habeenka laylatul qadriga ee Haramous	Habeenkii hore ee Haramous 2010 views
74	5 juillet 2016 9 h 53 min 0 comments Views : 12	5-ta Julay 2016	July 5, 2016 9 h 53 min 0 comments Views: 12
75	santé des nouveau-nés : le point avec Dr Timiro Aden, néonatologue à l' hôpital « Cheikho » de balbala	caafimaadka dhalaanka cusub : wareysiga Dr. Timiro Aadan, oo ah dhakhtarka dhalaanka ee cusbitaalka " Sheikho " ee Balbala	caafimaadka dhalaanka cusub: dhibicda Dr. Timiro Aden, dhakhtarka neonatoole ee isbitaalka "Cheikho" ee balbala

76	présent dans une trentaine de pays, le développement du groupe s' accélère ces dernières années, notamment sur le continent africain, là ou les majors classiques ont tendance à se faire plus rares … Il se fait donc naturellement un jeu de vases communicants qui nous va très bien et qui nous permet de démontrer la capacité du groupe à opérer de manière efficace, en parfait partenariat avec les acteurs locaux.	iyadoo joogtaa soddon waddan, horumarinta kooxda ayaa dardar galay sanadihii la soo dhaafay, gaar ahaan qaarada Afrika, halkaas oo shirkadaha ugu muhiimsan iyo kuwa caadiga ah ay noqdeen kuwo liita … Sidaa darteed dabiici ahaan, waxaa jira ciyaarta xirmoyinka isgaarsiiya oo noo habboonaoo noo suuro gelin kara inaan muujinno awooda kooxda ee ah inay si firfircoon uga shaqeeyaan, si wada jir ah ula wadaagaan ciyaartoyda maxalliga ah.	oo ay ku nool yihiin soddon waddan, horumarinta kooxdu waxay dardargelisay sanadihii ugu dambeeyay, gaar ahaan qaarada Afrika, halkaasoo caan ka mid ah kuwa caanka ah ay aad u yaryihiin... Sidaas awgeed, dabiici ahaan waa xirmooyin maraakiib ah oo noo yimaada si aad u wanaagsan oo noo ogolaanaya inaan muujinno awooda kooxda in ay si firfircoon uga shaqeeyaan, iskaashi buuxa oo la leh jilayaasha maxalliga ah.
77	elle contribuera ainsi à l' indépendance hydrique du pays et assurera la production d' une eau traitée localement pour les Djiboutiens.	waxay gacan ka geysan doontaa madax-bannaanida biyaha ee waddanka iyo hubinta soo-saarka biyaha la macaaneyay ee maxaliga ee loogu talagalay biyaha reer Jabuuti.	taasi waxay gacan ka geysan doontaa xornimada biyaha ee waddanka iyo hubinta wax soo saarka biyaha laga daaweeyo ee Jabuuti.
78	© Copyright 2017 — La Nation.	xuquqda la dhowray 2017-La Nation	© Copyright 2017 - Waddanka.
79	le directeur de la société djiboutienne des chemins de fer, M. Mohamoud Robleh Dabar, le directeur de la société éthiopienne des chemins de fer « ERC », M. Getachew Bedru et le directeur de la société chinoise CCECC en charge de la construction de la nouvelle ligne ferroviaire, M. Ding Zhaojun ont pris part à la cérémonie.	agaasimaha shirkadda tareenka Jabuuti, Mr. Roble Geelle Moxamoud Agaasimihii shirkadda tareenka Itoobiya " ERC " Mr. Getachew Bedru iyo agaasimaha shirkadda Shiinaha CCECC ee mas ' uul ka ah dhismaha tareenka cusub, Mr. Ding Zhaojun ayaa ka qayb qaatay xaflada.	Agaasimaha shirkadda Jabuuti, Mohamoud Rooble Dabar, agaasimaha shirkadda tareenada ee Itoobiya "ERC", Mr. Getachew Bedru iyo agaasimaha shirkadda Shiinaha ee CCECC ee mas'uul ka ah dhismaha khadka cusub ee tareenka, Mr. Ding Zhaojun ayaa ka qayb galay xafladda.
80	aussi, le sens de cette chanson se trouve en symbiose avec le message de Martin luter King qui disait en s' adressant aux noirs américains : ″ si un homme à plus trois siècle de retard sur les autres, il devra faire un tour de force impossible pour rattraper son retard ″.	sidoo kale, macnaha heestani waxay la mid tahay sida fariintii Martin luter King kaas oo lahaa isaga oo lahadlayaa maraykanka madow : " haddii nin saddex qarniyo uu ka dambeeyo kuwa kale, waxay tahay in uu sameeyaa safar xoog ah si u so qabto dibudhiciisa. "	Sidoo kale, macnaha sheekadani waxay ku dhex jirtaa fariinta Martin Luter King oo yiri isagoo la hadlaya dadka madow ee Maraykanka ah: "Haddii nin ka badan seddex qarniyadood oo ka dambeeya kuwa kale, waa inuu sameeyaa dalxiis aan suurtagal ahayn si ay u qabsadaan ".

81	après les élections législatives de 2013 qui avaient débouché sur une crise post- électorale, Djibouti vit de nouveau une échéance électorale qui est importante à plus d' un titre.	ka dib doorashadii sharci-dajinta ee 2013-kii taas oo horseeday xiisad doorashada kadib, haddana Jabuuti mar kale waxay markhaati u ah xilliga kama dambaysta ah ee doorashada oo muhiim u ah siyaabo badan.	ka dib doorashadii sharci-dajinta ee 2013-kii oo horseeday xiisad dib-u-dhac doorasho, haddana Jabuuti mar kale waxay soo wajahday xilligii kama dambaysta ahaa ee doorashada oo muhiim u ah wax ka badan hal sabab.
82	Historien, Archéologue et patrimoines à l' Université de Djibouti	Taariikhyahan, Jaamacadda Jabuuti	Historian, Archaeologist iyo Heritage ee Jaamacadda Jabuuti
83	Publicité& Annonces	xayeysiis & Ogeysiis	Raadinta iyo Digniinta
84	le président de la République, M. Ismaïl Omar Guelleh, a parrainé jeudi dernier l' inauguration du nouvel hôpital militaire, dénommé « Omar Hassan El Béchir et érigé aux abords de la route d' Arta dans la capitale.	Madaxweynaha Jamhuuriyadda, Mr. Ismaaciil Cumar Geelle, ayaa gudominayay khamiisti hore xaflada caleema saarka isbitaalka cusub ee ciidamada, oo logu magac daray " Cumar Xasan Al Bashir " kaas oo laga taagay meel u dhow wadada Carta ee casimadda.	Madaxweynaha Jamhuuriyadda, Mudane Ismaaciil Cumar Geelle, Khamiistii la soo dhaafay ayaa maal galiyay xafladdii cisbitaalka cusub ee loo yaqaan "Omar Hassan El Bashir" oo lagu dhisay waddada Carta.
85	Sports	Sbortiga	Sports
86	les visiteurs n' étaient pas venus les mains vides.	kooxda boqdayaasha lama ay iman gacamo madhan.	booqdayaashu ma iman wax madhan.
87	bref, le présent texte de loi vient compléter l' arsenal juridique existant en matière de lutte contre la traite humaine.	marka la soo koobo, sharcigan ayaa dhameystirayaa sharcigi si jiray ee dagaalka ka dhanka ah ka ganacsiga dadka.	Muddada gaaban, sharcigan ayaa sii kordhinaya suxufiyiinta sharci ee jira ee la dagaallanka ka ganacsiga dadka.
88	se souvenir de moi	igu soo xusuuso	Xusuuso aniga
89	Institut djiboutien d' études diplomatiques / Institutions belges de formations supérieures : une mise en réseau 3096 views	Machadka Jabuuti ee waxbarashada diblomaasiyadeed / Hay ' adaha Belgiumka ee tacliinta sare, Isku shabakeynta	Machadka Jabuuti of Studies diblomaasiyadeed / hay'adaha Belgian tacliinta sare: xirka 3096 views
90	Feed Rss	Faallooyinka aqristaha	Feed Rss
91	M. Hassan Mohamed Kamil a précisé que ces structures sont des outils d' intégration sociale et constituent de surcroit, de véritables plateformes de concertation et d' initiation à la gestion participative en faveur des populations en général.	Md. Xasan Maxamad Kamil ayaa cadeeyay in dhismayaasha ay yihiin kuwa loogu talagalay is-dhexgalka bulshada iyo waliba in ay noqdan, dhufto dhab ah oo wada tashi iyo is barasho iyo maamulka ka qaybqaadashada loogu talagalay dadweynaha guud.	Mohamed Hassan Kamil ayaa sheegay in dhismayaasha kuwanu waa qalab loogu talagalay is dhexgalka bulshada iyo marka lagu daro, dhufto ee wada tashi dhab ah oo sal-dhigida in maamulka ka qaybqaadashada dadweynaha guud.

92	la communauté Djiboutienne et Somalienne installée au Minnesotta organise une grande réception en l' honneur de la délégation officielle conduite par le président Guelleh.	Jaaliyadda Jabuuti iyo Somaliya ee ku nool Minnesotta waxay qabanqaabinsay soo dhaweyn ballaaran oo loogu martiqaadayo wafdiga rasmiga ah ee uu hogaaminayo Madaxweyne Geelle.	bulshada Jabuuti iyo Soomaali lagu rakibay Minnesotta qabanqaabisaa soo dhoweynta ee sharaf ergada rasmiga ah oo uu hoggaaminayo Madaxweyne Geelle.
93	l' oraison funèbre devait être dite par le colonel Wahib Hassan Kalinleh, qui est à la tête de l' Armée de l' air.	Qudbada aaska waxay ahayd inuu yidhaahdo Kornayl Wahib Hasan Qalinle, oo ah madaxa Ciidamada Cirka.	qudbad aaska lahaa in loogu yeedho by Colonel Hassan Wahib Kalinleh, kan madaxa ah oo ah Air ciidamada.
94	la Nation : - Le titre de votre livre s' intitule ” L' ambivalente libéralisation du droit du travail en République de Djibouti “, pourquoi le choix de ce titre ?	Jaridada Qaranka : - Magaca buuggaaga ayaa xaq u leh " Xuriyeynta Sharciga ah ee Sharciga Shaqada ee Jamhuuriyadda Jabuuti ", maxay tahay sababta doorashada ee magacan ?	Nation: - Cinwaanka aad buug ee xaq: "libaraaliyadda The jirio sharciga shaqada ee Jamhuuriyadda Jabuuti", maxaad u dooran title this?
95	Actualités Nationales	Wararka qaran	National News
96	dans ses premiers mots, il a rendu un vibrant hommage au parrain de la cérémonie et à tous les enseignants de l' université qui se sont dévoués pour les aider à réussir dans leur cursus universitaire.	Erayadiisi ugu horeysay, wuxu ka go ' dayna uu siiyo la ' yihiin ee xafladda iyo dhamaan macalimiinta jaamacadda kuwaas oo naftooda u go ' ay si ay uga caawiso inuu guulaysto in jaamacadda ay ku.	in erayada ugu horeysay, ka go'dayna uu siiyo la'yihiin ee xaflada oo dhan macalimiinta ee Jaamacadda kuwaas oo naftooda u go'ay si ay uga caawiso inuu guulaysto in jaamacadda ay.
97	recherche	Badhis	search
98	Actualités Nationales	Wararka qaran	National News
99	gendarmerie Nationale / Armée Chinoise : des prises de contact de haut niveau	jeendaranka qaranka / Ciddanka shiinaha : xidhiidhada heerka sare	Jandarma Qaran / Ciidan Shiineys ah: Xiriir heer sare ah
100	sans doute croient-elles au retour de l' embellie de 2014. a raison selon Mohamed Osman Allaleh qui table sur les énormes besoins en pierres taillées des futurs chantiers de pavage des rues et ruelles de la capitale et de sa banlieue.	shaki kuma jiro inay aaminsan yihiin soo noqoshada horumarinta 2014. Maxamed Cusmaan Allaaleh uu ku tiirsan yahay baahida weyn ee dhagxaan la jarjaray oo ah goobaha mustaqbalka fog ee waddooyinka iyo magaalooyinka iyo agagaarkeeda.	shaki kuma jiro inay aaminsan yihiin soo laabashada 2014. waa xaq by Mohamed Osman Allaleh in kahadashaa baahida weyn ee xaareysa dhagaxyo jaray goobaha mustaqbalka ee jidadka iyo surimmada caasimadda iyo agagaarkeedii.
101	Archives	kaydadka	archives
102	© Copyright 2017 — La Nation.	xuquqda la dhowray 2017-La Nation	© Copyright 2017 - Waddanka.
103	Publicité& Annonces	xayeysiiska & Ogeysiisyo	Raadinta iyo Digniinta
104	CNIPLCC / MCPT : la commission nationale de lutte contre la corruption imprime ses marques	CNIPLCC / MCPT : Guddiga Qaranka ee ka dhanka ah musuqmaasuqa ayaa dabaacay sumaddiisi	CNIPLCC / MCPT: komishanka qaranka ka dhanka ah musuqmaasuqa qora magac ay
105	Actualités Nationales	Wararka qaran	National News

VI.2 Annotation de quelques évaluateurs des 54 segments post-édités

Notes et corrections de l'évaluateur 7

Remarque : L'évaluateur 7 a modifié 21 segments parmi les 54 pots-édités qu'il a évalués, ce qui fait environ 40% de des segments de l'échantillon. Les segments dans le tableau ci-dessous colorés en rouge sont ceux qui ont été corrigés et modifiés par l'évaluateur. Les modifications vont de la simple modification d'un ou plusieurs mots du segment jusqu'à son remplacement total.

#segment	segment initial PE	PE EV 7	Note EV7
1	Axad, 6da agusto	Axad, 6da agusto	17
2	Al Qarn	Al Qarn	13
3	Wakaaladda Wararka ee Jabuuti	Wakaaladda warfafinta ee Jabuuti	15
4	Jabuuti Telecom	Jabuuti Telecom	14
5	RTD-da	RTD-da	14
6	Xafiska bostada ee Jabuuti	Xafiska bosaha ee Jabuuti	12
7	Maanta Jabuuti	Maanta Jabuuti	15
8	Soo dhaweyn	Soo dhaweyn	17
9	Wararka qaran	Wararka dalka	10
10	Sbortiga	Ciyaaraha	11
11	Dhaqaalaha	Dhaqaalaha	18
12	Caafimaadka	Caafimaadka	17
13	Wararka Calamka	Wararka Calamka	18
14	Kaydadka	Kaydadka	17
15	Xidhidh	Xidhidh	16
16	Ogeysiis: Wixii ogaysiis ama xayaysiisyada dhan fadlan la xiriir waaxda iibka.	Ogeysiis: wixii ogaysiis ama xayaysiis ah fadlan la xidhiidh waaxda iibka.	10
17	Email: nationdjib@gmail.com / servicecommercial.nation@gmail.com	Email: nationdjib@gmail.com / servicecommercial.nation@gmail.com	17
18	Aasaaska IOG : Guryo cusub oo la siiyay dhibanayaashi 18ka Luulyo	Aasaaska IOG : Guryo cusub oo la siiyay dhibanayaashi 18ka Luulyo	18
19	30 Luulyo 2017	30 Luulyo 2017 sacadda sideedii subax iyo saddexiyo toban daqiqo.	10
20	Khamiistii hore PK 13, mid kasta oo ka mid ah 200 ee qoysaska dhibanayaasha ee 18ka Luulyo ayaa helay guri cusub.	Khamiistii hore goobta PK13, qof kasta oo ka mid ahaa 200 ee qoysaska dhibanayaasha ee 18ka Luulyo ayaa helay guri cusub.	16
21	Guryahani uu siiyay madaxweynuhu dhibanayaasha Balbala waxay yihiin guulaha ugu horeyay ee aasaaska IOG aa loo aasasay guryaha.	Guryihii Madaxweynuhu u deeqay dhibanayaasha balbala waxay yihiin quulaha ugu horeyay ee aasaaaska IOG ee loo magacaabay guryaha	10
22	Waxay ka kooban yihiin laba qol, musqusha iyo qolka karinta , guryahanna waa la fidsan kara ilayo F4 daaradiisa weyn awgeed.	Iyaga oo ka kooban laba qol, musqusha iyo qolka karinta, guryahaasi waa la baladhin kara ilayo F4 daaradiisa weyn awgeed.	15
23	Waayo, dhibbanayaasha "Afar-mitir," maalinta 27 Luulyo waxay ahayd maalin lagu qoro dhagax cad.	Waayo, dhibanayaasha lagu magacaabay "Afarta mitir" malinta 27 luuliyo waxay ahayd maalin tariikh gal ah oo lagu qoro dhagax cad.	15

#segment	segment initial PE	PE EV 7	Note EV7
24	Dareenku wuxu ahaa mid ka muuqda ka-faa'iideystayaasha; kuwaas oo, ka dib markii ay ku noolayeen sanado guryahooda cooshadaha ah, wax kasta uu ka baabbi'iyey dabkii 18 Luulyo.	Dareenku wuxu ahaa mid ka muuqda wajiga kuwi ka faa'iideystay, kuwaas oo mudda badan ku noola guryahoodi cooshada aha, oo dabki 18 luliyo wax kasta ka baabiiyay.	10
25	Waa kuwan hadda la dejiyay guryo ku istaahilo magaca.	Waa kuwan iyaga oo la dajiyey guryo u qalma.	12
26	Magaaladan yar , ay dhistay aasaaska uu abuuray dhawaan madaxa dawladda waxay leedahay barxad lagu ciyaaro ee carruurta, dugsi hoose iyo saldhiga booliska.	Xafadan yar ee ay dhistay aasaaska uu dhawaan abuuray madaxweynaha xukuumadu, waxay leedahay goobta ciyaaraha ee caruurta, dugsi hoose iyo saldhiga booliska.	10
27	Iyago leh ilmo tiiraanyo leh, qaar ka mid ah hooyooyinka, kuwaas oo hadaladoodi ay ka soo ururiyeen asxaabteenna telefishanka, ayaa ugu mahad celiyay Madaxweynaha deeqsinimada weyn ee uu ku maareeyay arrintooda tan iyo maalintii ugu horeysay.	Iyago tiiraanyo leh, oo ilmo ku joogto, qaar ka mid ah hooyooyinka, kuwaas oo hadaladoodi ay ka soo ururiyeen asxaabteenna telefishanka, ayaa ugu mahad celiyay Madaxweynaha deeqsinimada weyn ee uu ku maareeyay arrintooda tan iyo maalintii ugu horeysay.	13
28	Tweet	Tweet	17
29	Ka tag jawaabtada	Dhaaf jawaabta	13
30	Halkaan ku rix sidad u joojisid jawabta	Halkan taabo sidad u joojisid jawabta	14
31	Magac (loo baahan yahay)	Magac (loo baahan yahay)	17
32	Mail (aan la daabacayn) (loo baahan yahay)	Mail (aan la baafin karin)(loo baahan yahay)	10
33	Websit-ka	Websit-ka	14
34	Xidhidh isku noqnoqda	Xidhiidhka isku noqnoqda	14
35	Fidiyoyada ugu so horeya	Fidiyoyada ugu so horeya	17
36	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP 8-di janayo 2016	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Heshiiska UMP 8-di janayo 2016	13
37	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Gobaabinta Madaxweynaha Jamhuuriyadda ee sannadka 2016	13
38	Ma karaan socosho.	Ma karaan socosho.	14
39	Fadlan isku day dhinaca gacanta.	Fadlan isku day dhinaca gacanta	14
40	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP-da	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee heshiska UMP-da	12
41	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Gobaabinta Madaxweynaha Jamhuuriyadda ee sannadka 2016	10
42	Xayaysiis	Xayaysiis	17
43	Raac warka Fasebuga	La soco warka Fasebuga	11
44	Hel Wargeyskeena	Hel Wargeyskeena	14
45	Hel Macluumaadyadi ugu dambeyay dhinaca Emailka - Sajilad bilash ah!	Hel Macluumaadyadi ugu dambeyay dhinaca Emailka - Sajilad bilash ah!	17
46	Iimeelka:	Iimeelka	14
47	Qaybo	Qaybo	17
48	Qaybo	Qaybo	17
49	Badhis	Baadhitaan	17
50	10-ki Maqaallo ee ugu dambeeyay	10-ki Maqallo ee ugu dambeeyay	14
51	Wareysiga … Shiikh Cabdiraxmaan	Wareysiga … Shiikh Cabdiraxmaan	16

#segment	segment initial PE	PE EV 7	Note EV7
	Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin Madaxweynaha Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada dadka diinta ee gobolka SHAMKAT	Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin Madaxweynaha Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada dadka diinta ee gobolka SHAMKAT	
52	WAIDHW/Guddiga Sare ee Fatwada: Laba maalmood si lo so saro talooyin ka dhanka ah MGF-ta	WAIDHW/Guddiga Sare ee Fatwada: Laba maalmood si lo so saro talooyin ku saabsan dhanka ah MGF-ta	14
53	Doorashada madaxtooyada ee Ruwanda: Madaxa dawalad ayaa fariin hambalyo u diray Madaxweyne Kagame	Doorashada madaxtooyada ee Ruwanda: Madaxweynaha dalka ayaa fariin hambalyo u diray Madaxweyne Kagame	15
54	FTX 2017: EASF-ta ayaa sameysatay fadhiyadii diyaarinta ee Addis Ababa	FTX 2017: EASF-ta ayaa sameysatay fadhiyadii diyaarinta ee Addis Ababa	15

Résultat de l'évaluateur 8

Remarque : L'évaluateur 8 a modifié 23 segments parmi les 54 pots-édités qu'il a évalués, ce qui fait environ un peu plus de 42% des segments de l'échantillon. Les segments dans le tableau ci-dessous colorés en rouge sont ceux qui ont été corrigés et modifiés par l'évaluateur. Les modifications vont de la simple modification d'un ou plusieurs mots du segment jusqu'à son remplacement total.

#segment	segment initial PE	PE EV 8	Note EV8
1	Axad, 6da agusto	Axad, 6da ogoss	14
2	Al Qarn	Al Qarn	15
3	Wakaaladda Wararka ee Jabuuti	Wakaaladda Wararka ee Jabuuti	15
4	Jabuuti Telecom	Jabuuti Telecom	16
5	RTD-da	RTD-da	14
6	Xafiska bostada ee Jabuuti	Xafiska bostada ee Jabuuti	15
7	Maanta Jabuuti	Maanta iyo Jabuuti	15
8	Soo dhaweyn	Soo dhaweyn	14
9	Wararka qaran	Wayaha Qaran	15
10	Sbortiga	Sbortiga	15
11	Dhaqaalaha	Dhaqaalaha	15
12	Caafimaadka	Caafimaadka	15
13	Wararka Calamka	Wayaha Calamka	15
14	Kaydadka	Kaydadka	14
15	Xidhidh	Xidhidh	14
16	Ogeysiis: Wixii ogaysiis ama xayaysiisyada dhan fadlan la xiriir waaxda iibka.	Ogaysiis : Ogaysiis ama iidheh dhamantod fadlan laso xidhidha waaxda bayac mushtarka.	14
17	Email: nationdjib@gmail.com / servicecommercial.nation@gmail.com	Email: nationdjib@gmail.com / servicecommercial.nation@gmail.com	14
18	Aasaaska IOG : Guryo cusub oo la siiyay dhibanayaashi 18ka Luulyo	Hay'adda IOG : Guryo cusub oo la siiyay barakacayasha 18kii Luulyo	14
19	30 Luulyo 2017	30 Luulyo 2017 8h 13 midhid.	14

20	Khamiistii hore PK 13, mid kasta oo ka mid ah 200 ee qoysaska dhibanayaasha ee 18ka Luulyo ayaa helay guri cusub.	Khamiistii hore PK 13, mid kasta oo ka mid ah 200 ee qoysaska dhibanayaasha ee 18ka Luulyo ayaa helay guri cusub.	16
21	Guryahani uu siiyay madaxweynuhu dhibanayaasha Balbala waxay yihiin guulaha ugu horeyay ee aasaaska IOG aa loo aasasay guryaha.	Guryahanii uu gudonsiyay madaxweynaha dalka dhibanayashi balbala ayaa ka dhan ah waxqabadki u horeyey ee hay'adda IOG ee guriyaynta.	16
22	Waxay ka kooban yihiin laba qol, musqusha iyo qolka karinta , guryahanna waa la fidsan kara ilayo F4 daaradiisa weyn awgeed.	Waxay ka kooban yihiin laba qol, musqusha iyo qolka karinta , guryahanna waa la fidsan kara ilayo F4 daaradiisa weyn awgeed.	15
23	Waayo, dhibbanayaasha "Afar-mitir," maalinta 27 Luulyo waxay ahayd maalin lagu qoro dhagax cad.	Waayo, dhibbanayaasha "Afar-mitir," maalinta 27 Luulyo waxay ahayd maalin lagu qoro dhagax cad.	15
24	Dareenku wuxu ahaa mid ka muuqda ka-faa'iideystayaasha; kuwaas oo, ka dib markii ay ku noolayeen sanado guryahooda cooshadaha ah, wax kasta uu ka baabbi'iyey dabkii 18 Luulyo.	Shucurta oo laga dareemayay kafaa'iideystayasha qarkod, ayako oo kadib marki ay ku noolayeen sanado dheer guryo coshadaha ku xolobelay dabkii 18 Luulyo.	15
25	Waa kuwan hadda la dejiyay guryo ku istaahilo magaca.	Waa kuwan hadda la dejiyay guryo ku istaahilo magaca.	17
26	Magaaladan yar , ay dhistay aasaaska uu abuuray dhawaan madaxa dawladda waxay leedahay barxad lagu ciyaaro ee carruurta, dugsi hoose iyo saldhiga booliska.	Xafadan yar, ay dhistay hay'adda uu abuuray dhawan Madaxweynaha Dalka aya ka koban xaruun ciyareed ee carruurta, dugsi hoose iyo xaruun booliss.	14
27	Iyago leh ilmo tiiraanyo leh, qaar ka mid ah hooyooyinka, kuwaas oo hadaladoodi ay ka soo ururiyeen asxaabteenna telefishanka, ayaa ugu mahad celiyay Madaxweynaha deeqsinimada weyn ee uu ku maareeyay arrintooda tan iyo maalintii ugu horeysay.	Ayago la ilmaynaya farxad, hooyooyinka qarkood oo ay waraysten walalehenka tvga, waxay uu cabireen madaxweynaha mahadnaq, sida uu ugu hawlayahay arimahoga min malinti uu horeysay iyo ay ku dhehantahay deqsinimo weyn.	15
28	Tweet	Tweet	14
29	Ka tag jawaabtada	Ka tag jawaabtada	13
30	Halkaan ku rix sidad u joojisid jawabta	Haalka ku rix sidad uu babi'isi jawabta.	14
31	Magac (loo baahan yahay)	Magac (loo baahan yahay)	13
32	Mail (aan la daabacayn) (loo baahan yahay)	Mail(aan lafafinayn) (la rabo)	14
33	Websit-ka	Websit-ka	14
34	Xidhidh isku noqnoqda	Xidhid tidcan	14
35	Fidiyoyada ugu so horeya	Muqal shidane	14
36	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP 8-di janayo 2016	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP 8-di janayo 2016	13
37	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Bogadinta madaxweynaha jamhuriyada ee sanadka 2016.	13
38	Ma karaan socosho.	Ma karaan socosho.	14
39	Fadlan isku day dhinaca gacanta.	Fadlan ku day gacanta	14
40	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP-da	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee heshiska UMP-da	14
41	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Bogadinta madaxweynaha jamhuriyada ee sanadka 2016.	13
42	Xayaysiis	Iidheh	14
43	Raac warka Fasebuga	Kala soco wayaha faysbuga.	13

44	Hel Wargeyskeena	Hel Wargeyskeena	14
45	Hel Macluumaadyadi ugu dambeyay dhinaca Emailka - Sajilad bilash ah!	Hel Macluumaadyadi ugu dambeyay dhinaca Emailka - Sajilad bilash ah!	14
46	Iimeelka:	Iimeelka	14
47	Qaybo	Qaybo	14
48	Qaybo	Qaybo	14
49	Badhis	Badhitan	14
50	10-ki Maqaallo ee ugu dambeeyay	10-ki Maqallo ee ugu dambeeyay	14
51	Wareysiga … Shiikh Cabdiraxmaan Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin Madaxweynaha Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada dadka diinta ee gobolka SHAMKAT	Hadalka …. Shiikh Cabdiraxmaan Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin madaxa Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada culuma'udiinta gobolka SHAMKAT	14
52	WAIDHW/Guddiga Sare ee Fatwada: Laba maalmood si lo so saro talooyin ka dhanka ah MGF-ta	WAIDHW/Guddiga Sare ee Fatwada: Laba maalmood si lo so saro talooyin ka dhanka ah MGF-ta	14
53	Doorashada madaxtooyada ee Ruwanda: Madaxa dawalad ayaa fariin hambalyo u diray Madaxweyne Kagame	Doorashada madaxtooyada ee Ruwanda: Madaxa dawalad ayaa fariin hambalyo u diray Madaxweyne Kagame	15
54	FTX 2017: EASF-ta ayaa sameysatay fadhiyadii diyaarinta ee Addis Ababa	EASF oo qabsatay fadhiyadii diyarinta ee Adis Abeba	15

Notes et corrections de l'évaluateur 9

Remarque : L'évaluateur 9 a modifié 30 segments parmi les 54 pots-édités qu'il a évalués, ce qui fait environ 55% du total de l'échantillon. Les segments dans le tableau ci-dessous colorés en rouge sont ceux qui ont été corrigés et modifiés par l'évaluateur. Les modifications vont de la simple modification d'un ou plusieurs mots du segment jusqu'à son remplacement total. Sur les 6 segments auquel il attribué la note 5, il en a modifié et corrigé 5.

#segment	segment initial PE	PE EV 9	Note EV9
1	Axad, 6da agusto	Axad, 6da agusto	14
2	Al Qarn	Al Qarn	15
3	Wakaaladda Wararka ee Jabuuti	Wakaaladda Wararka ee Jabuuti	5
4	Jabuuti Telecom	War isgadhsinta Jabuuti	5
5	RTD-da	Idaacada Fogaalaraga ee Jabuuti	5
6	Xafiska bostada ee Jabuuti	Xafiska dhambal gudbinta Jabuuti	10
7	Maanta Jabuuti	Maanta iyo Jabuuti	14
8	Soo dhaweyn	Xashada soo dhaweynta	14
9	Wararka qaran	Wararka Dalka	5
10	Sbortiga	Ciyaaraha	5
11	Dhaqaalaha	Dhaqaalaha	15
12	Caafimaadka	Caafimaadka	15
13	Wararka Calamka	Wararka Dunida	14
14	Kaydadka	Kaydadka	14
15	Xidhidh	Xidhidh	14

16	Ogeysiis: Wixii ogaysiis ama xayaysiisyada dhan fadlan la xiriir waaxda iibka.	Ogaysiis : wixii ogaysis ah ama xayaysis ah Fadlan la xidhiidh waaxda bayac mushtarka	14
17	Email: nationdjib@gmail.com / servicecommercial.nation@gmail.com	Email: nationdjib@gmail.com / servicecommercial.nation@gmail.com	14
18	Aasaaska IOG : Guryo cusub oo la siiyay dhibanayaashi 18ka Luulyo	Aasaaska IOG : Guryo cusub oo la siiyay barakacayasha 18kii Luulyo	5
19	30 Luulyo 2017	30 Luulyo 2017	14
20	Khamiistii hore PK 13, mid kasta oo ka mid ah 200 ee qoysaska dhibanayaasha ee 18ka Luulyo ayaa helay guri cusub.	Khamistii inasoo dhaftay PK 13 ayaa la siiyey mid kasta oo ka mid ah 200 ee qoysaska barakacayasha 18kii Luulyo guri cusub.	12
21	Guryahani uu siiyay madaxweynuhu dhibanayaasha Balbala waxay yihiin guulaha ugu horeyay ee aasaaska IOG aa loo aasasay guryaha.	Guryahani waxay yihiin kuwii ugu horeyay uu siiyay madaxweynuhu barakacayasha Balbala ee aasaaska IOG oo loo aasasay qorshaha guryenta.	12
22	Waxay ka kooban yihiin laba qol, musqusha iyo qolka karinta , guryahanna waa la fidsan kara ilayo F4 daaradiisa weyn awgeed.	Waxay ka kooban yihiin laba qol, musqusha iyo qolka karinta , guryahanna waa la fidsan kara ilayo F4 daaradiisa weyn awgeed.	15
23	Waayo, dhibbanayaasha "Afar-mitir," maalinta 27 Luulyo waxay ahayd maalin lagu qoro dhagax cad.	Waayo, barakacayasha "Afar-mitir", malinta 27 Luulya waxay ahayd maalin lama ilobaan ah.	12
24	Dareenku wuxu ahaa mid ka muuqda ka-faa'iideystayaasha; kuwaas oo, ka dib markii ay ku noolayeen sanado guryahooda cooshadaha ah, wax kasta uu ka baabbi'iyey dabkii 18 Luulyo.	Dareenku wuxu ahaa mid ka muuqda ka-faa'iideystayaasha; kuwaas oo, ka dib markii ay ku noolayeen sanado guryahooda cooshadaha ah, wax kasta uu ka baabbi'iyey dabkii 18 Luulyo.	15
25	Waa kuwan hadda la dejiyay guryo ku istaahilo magaca.	Waa kuwan hadda la dejiyay guryo ku istaahilo magaca.	12
26	Magaaladan yar , ay dhistay aasaaska uu abuuray dhawaan madaxa dawladda waxay leedahay barxad lagu ciyaaro ee carruurta, dugsi hoose iyo saldhiga booliska.	Magaaladan yar , ay dhistay aasaaska uu abuuray dhawaan madaxa xukumada waxay leedahay barxad carurtu ku ciyarto, dugsi hoose iyo saldhiga askarta nabad sugiida.	14
27	Iyago leh ilmo tiiraanyo leh, qaar ka mid ah hooyooyinka, kuwaas oo hadaladoodi ay ka soo ururiyeen asxaabteenna telefishanka, ayaa ugu mahad celiyay Madaxweynaha deeqsinimada weyn ee uu ku maareeyay arrintooda tan iyo maalintii ugu horeysay.	Iyago leh ilmo tiiraanyo leh, qaar ka mid ah hooyooyinka, kuwaas oo hadaladoodi ay ka soo ururiyeen asxaabteenna telefishanka, ayaa ugu mahad celiyay Madaxweynaha deeqsinimada weyn ee uu ku maareeyay arrintooda tan iyo maalintii ugu horeysay.	15
28	Tweet	Tweet	12
29	Ka tag jawaabtada	Ka gudub	13
30	Halkaan ku rix sidad u joojisid jawabta	Halkan guji si ad uga noqotid jawabta	14
31	Magac (loo baahan yahay)	Magacaga sheeg	13
32	Mail (aan la daabacayn) (loo baahan yahay)	Mail(qarsoonan dona sheeg) (loo baahan yahay)	14
33	Websit-ka	Websit-ka	14
34	Xidhidh isku noqnoqda	Xidhidh isku noqnoqda	10
35	Fidiyoyada ugu so horeya	Fidiyoyasha labogey	14
36	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP 8-di janayo 2016	Khudbadi Madaxweynaha Jamhuuriyadda Ka jediyey Shirka UMP 8-di janayo 2016	13
37	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyada Madaxweynaha Jamhuuriyadda ee sannadka 2016	13

38	Ma karaan socosho.	Ma karaan socosho	10
39	Fadlan isku day dhinaca gacanta.	Fadlan gacanta ku day	14
40	Khudbadi Madaxweynaha Jamhuuriyadda Jabuuti ee Shirka UMP-da	Khudbadi Madaxweynaha Jamhuuriyadda Ka jediyey Shirka UMP	14
41	Rabbintanka Madaxweynaha Jamhuuriyadda ee sannadka 2016	Hambalyada Madaxweynaha Jamhuuriyadda ee sannadka 2016	15
42	Xayaysiis	Idhe	14
43	Raac warka Fasebuga	Kala soco warka alada Facebookga	13
44	Hel Wargeyskeena	Naga gudon wargeyskayaga	14
45	Hel Macluumaadyadi ugu dambeyay dhinaca Emailka - Sajilad bilash ah!	Naga gudon Maclumaadyadi ugu dambeeyay dhinaca Emailka- Sajilad bilash ah!	14
46	Iimeelka:	Iimeelka	14
47	Qaybo	Qaybo	14
48	Qaybo	Qaybo	14
49	Badhis	Dondonis	15
50	10-ki Maqaallo ee ugu dambeeyay	10-ki Maqallo ee ugu dambeeyay	14
51	Wareysiga … Shiikh Cabdiraxmaan Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin Madaxweynaha Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada dadka diinta ee gobolka SHAMKAT	Wareysiga … Shiikh Cabdiraxmaan Maxamed Cali oo lo yaqaan Cabdurahman Shamsudin Madaxweynaha Guddiga Sare Islamiga ee Fatwada ahna xoghayaha guud ee shabakada dadka diinta ee gobolka SHAMKAT	14
52	WAIDHW/Guddiga Sare ee Fatwada: Laba maalmood si lo so saro talooyin ka dhanka ah MGF-ta	WAIDHW/Guddiga Sare ee Fatwada: Laba maalmood si lo so saro talooyin ka dhanka ah MGF-ta	14
53	Doorashada madaxtooyada ee Ruwanda: Madaxa dawalad ayaa fariin hambalyo u diray Madaxweyne Kagame	Doorashada madaxtooyada ee Ruwanda: Madaxa dawalad ayaa dhambal u diray Madaxweyne Kagame	15
54	FTX 2017: EASF-ta ayaa sameysatay fadhiyadii diyaarinta ee Addis Ababa	FTX 2017: EASF-ta ayaa sameysatay fadhiyadii diyaarinta ee Addis Ababa	15

VI.3 Morphosyntaxe du somali

Introduction

Cette monographie a été réalisée dans le cadre d'un travail de recherche entrepris dans un premier temps à l'Université de Bordeaux, portant sur une étude linguistique des problèmes et des approches relatifs à l'étiquetage morphosyntaxique du somali, puis enrichi et poursuivi dans le cadre de cette thèse.

Le but de ce document est de fournir les bases nécessaires à la compréhension de la morphosyntaxe du somali, en vue de construire des outils et applications de traitement automatique du langage naturel pour cette langue, tels qu'un analyseur et un générateur morphologiques, un étiqueteur morphosyntaxique, ou un correcteur grammatical.

Nous présenterons d'abord une description générale du somali (origine, typologie, aire de diffusion, dialectes), puis les classes lexicales productives (noms, verbes, adjectifs, adverbes), les classes « grammaticales » (non productives, à savoir les pronoms, les appositions verbales, les déterminants, les marqueurs de types de phrase, les conjonctions, et les interjections ou

phatiques). Nous passons ensuite à la morphologie flexionnelle et dérivationnelle. Nous terminons par une introduction à la syntaxe des phrases affirmatives simples et de leur noyau verbal, potentiellement assez complexe.

VI.3.1 Description du somali : une langue africaine peu dotée

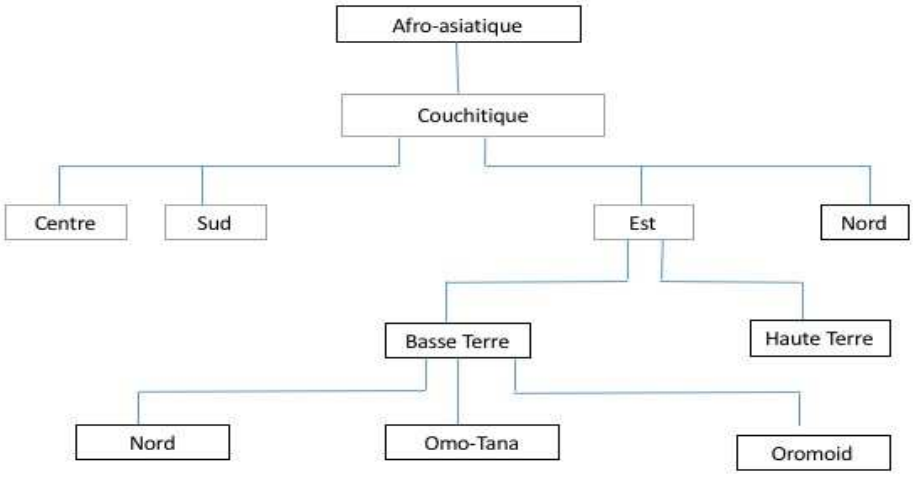

Figure 33 : Famille des langues afro-asiatiques couchitiques

VI.3.1.1 Origine et typologie

La langue somalie fait partie de la branche des langues couchitiques, et appartient à une sous-division de la grande famille des langues dites afro-asiatiques ou chamito-sémitiques avec l'omotique, le berbère, le sémitique et l'égyptien ancien. La sous-branche des langues couchitiques est composée d'une trentaine de langues, et la langue somalie est la deuxième langue couchitique en nombre de locuteurs après l'oromo[61]. Elle appartient également à la sous branche est-couchitique ou au LEC (Lowlands East Cushitic) avec le rendille et le boni.

L'aire géographique de la langue somalie s'étend entre le sud de Djibouti, le sud-est de l'Éthiopie, la Somalie, et le nord-est du Kenya, comme le montre la figure suivante.

[61] L'oromo est la première langue parlée en Éthiopie avec au moins 40 millions de locuteurs.

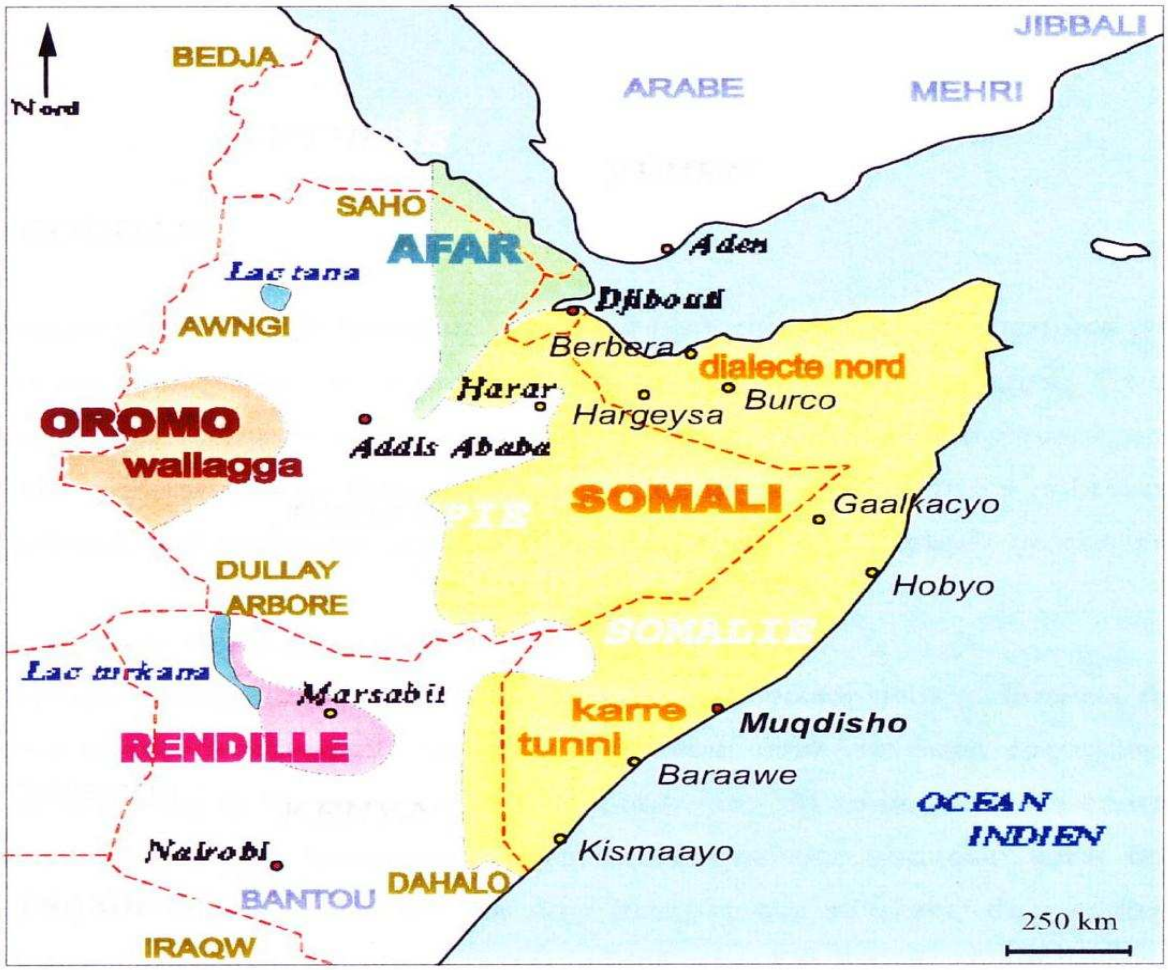

Figure 34 : Aire géographique du somali[62]

La position de la langue somalie dans les sous-branches des langues couchitiques et des familles afro-asiatiques est donnée dans la Figure 35.

Contrairement à d'autres langues africaines, et bien qu'elle ait trois variantes ou dialectes, la langue somalie jouit d'une relative homogénéité. Il y a le *dialecte du nord* (ou somali standard), le dialecte *benadiri* (ou somali littoral) et enfin le *may*.

Le dialecte du nord de la Somalie est le mieux développé dans la péninsule somalienne, grâce à l'immigration de certains clans ou tribus du nord de la Somalie vers le sud et le centre. Il est communément parlé et compris à la fois dans le sud, le centre et le nord de la Somalie, dans le nord-est du Kenya, le sud-ouest éthiopien, et enfin dans le sud de la République de Djibouti.

[62] Source : [Barillot, X., 2002]

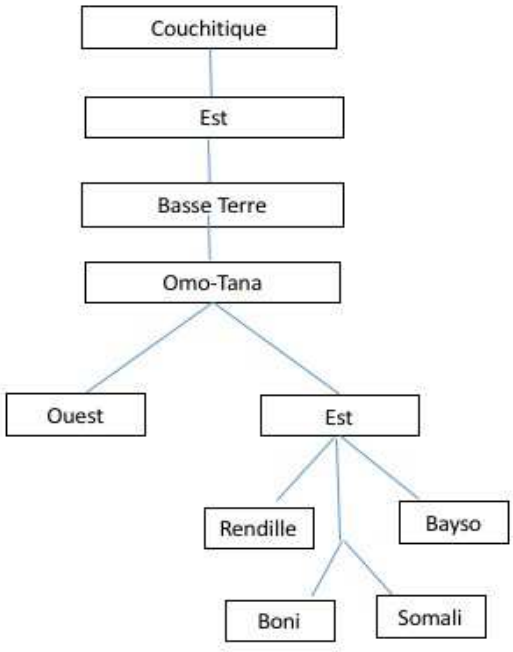

Figure 35 : Schéma des langues couchitiques de l'est

Le dialecte officiel qui a été choisi pour l'adoption de l'écrit du somali est celui du nord de la Somalie. Il est considéré comme étant le somali standard par toutes les communautés somalies de par le monde [Saeed 1999], [Abdullahi 1995]. Voici, quelques détails supplémentaires sur les trois principaux dialectes du somali repèrés par les études linguistiques [Saeed 1999], [Abdullahi, 1995].

- Le *may* est la langue des rahanweins, l'une des grandes tribus somalies habitant dans le sud-ouest du pays. La ville de Baidoa (Baydhaba) se trouve à 250 km de Mogadiscio et est le chef-lieu de ce dialecte. La plupart des locuteurs de ce dialecte habitent dans cette ville et ses alentours. C'est un dialecte incompréhensible pour les Somalis du nord et de Djibouti.
- Le *benadiri* est le dialecte des habitants de Mogadiscio ainsi que de la région côtière dite Benaadir, allant de la ville de Marka à Mogadiscio. On appelle aussi ce dialecte le *somali du littoral.*
- Le *somali du nord* est le dialecte parlé dans le centre et dans le nord-est (Puntland), dans le nord-ouest (Somaliland) ainsi qu'en république de Djibouti. Il fait office de langue commune et officielle de tous les Somalis.

VI.3.1.2 Typologie syntaxique du somali

La langue somalie est une langue de type SOV (Sujet-Objet-Verbe) c'est-à-dire que la structure syntaxique des phrases affirmatives simples de cette langue est généralement composée du sujet, suivi de l'objet de la phrase, et enfin du verbe (ou du noyau verbal, cf. infra).

Du point de vue morphologique, le somali appartient aux langues dites concaténatives.

D'autre part, c'est une langue à accent tonal, et le ton joue un rôle important dans sa morphosyntaxe comme on le verra plus loin.

Le nombre de locuteurs du somali est estimé à 16,492 millions [Ethnologue, 2017]. Ces locuteurs habitent majoritairement dans quatre pays de la Corne de l'Afrique : Djibouti, Somalie, Éthiopie et Kenya (voir Figure 35).

Comme indiqué plus haut l'alphabet officiel de la langue somalie est l'alphabet latin avec l'ajout de digraphes. Il y a 10 voyelles dont 5 courtes et 5 longues, ainsi que 21 consonnes. Il n'y a pas d'organisme officiel de régulation et de standardisation de l'orthographe en somali, d'où la difficulté de choisir une orthographe fixe pour cette langue. Dans le cadre de ce travail, nous avons adopté l'orthographe officielle de la langue somalie qui est celle utilisée en République de Somalie.

VI.3.1.3 Structure phonologique et phonétique du somali

Comme dans toutes les langues naturelles une syllabe est composée en somali d'un ou plusieurs phonèmes et chaque mot de la langue somalie contient au moins une syllabe. La structure syllabique du somali est simple et elle est constituée comme suit :

- **v** : Une seule voyelle qu'elle soit brève ou longue ou bien une diphtongue.
  Exemple : *ú* (à), *oo* (et), *èy* (chien)
- **cv** : Une consonne suivie d'une voyelle (brève ou longue) ou d'une diphtongue.
  Exemple : *kú* (Dans), *síi* (Donner), *cáy* (insulte).
- **vc** : Une voyelle (brève ou longue) suivi d'une consonne.
  Exemple : *ul* (bâton), *il* (œil).
- **cvc** : Une consonne suivie d'une voyelle (brève ou longue), suivie d'une consonne.
  Exemple : *nin* (garçon).

Les quatre structures ci-dessus avec des voyelles brèves forment la structure syllabique de base du somali. Cependant, il faut ajouter 4 autres formes de syllabes, obtenues en remplaçant chaque voyelle brève par une voyelle longue. Au total, il existe donc 8 structures syllabiques en somali.

VI.3.2 Les catégories grammaticales du somali

La langue somalie est une langue à morphologie concaténative et agglutinante. Présenter les différentes catégories grammaticales de cette langue revient à effectuer une classification basée sur la morphologie et la distribution des mots. Il existe plusieurs travaux qui ont été réalisés sur la grammaire, la syntaxe et la morphologie du somali : [Saeed, 1984], [Abdullahi, 1995], [Abdalla, 1999].

Notre classification et présentation des catégories grammaticales ci-dessous est inspirée d'une étude précédente quasi-complète et détaillée que nous avons faite sur la classification des mots du somali [Assowe, 2011].

Selon les études et travaux cités, la langue somalie dispose de deux types de catégories grammaticales : les classes lexicales et les classes grammaticales.

Dans la suite, nous présentons succinctement ces différentes catégories grammaticales du somali.

VI.3.2.1 Les classes lexicales du somali

VI.3.2.1.1 Les noms

Les unités linguistiques qui peuvent être catégorisés comme étant des noms en somali occupent les fonctions de sujet ou de complément au sein de la phrase. D'un point de vue

morphologique, les noms ont un genre inhérent grâce aux morphèmes déterminants qui leur sont obligatoirement affixés, ou à travers leur accent tonique. En somali, il existe deux genres, le « masculin » et le « féminin ». Les noms ont également le trait grammatical du *nombre*, qui distingue les noms au singulier des noms au pluriel à l'aide de morphèmes flexionnels.

Il existe quatre sous-catégories de noms en somali : noms communs (NC), noms propres (NP), nombres cardinaux (Num) et enfin les pronoms indépendants.

VI.3.2.1.2 Les verbes

Les verbes peuvent être « de base », ou se former à partir d'autres classes de mots selon certaines règles morphologiques.

En sémantique, le verbe dénote un procès qui peut être un état, une action ou un événement qui pour se réaliser implique le temps (présent, passé et futur).

En somali, le verbe constitue le noyau central de la phrase. Cependant, l'existence de phrases sans verbe est également attestée dans cette langue grâce entre autres au rôle de prédicat joué dans certaines phrases par le thématiseur[63] waa.

D'un point de vue morphologique, le verbe est identifiable grâce aux morphèmes flexionnels de temps, d'aspect, de personne et du monde qui lui sont suffixés.

On distingue 3 types de verbe en somali en fonction des différentes formes qu'ont les désinences verbales : verbes à flexion suffixale, verbes à préfixes, et verbes à flexion préfixale (*yahay* (être)).

VI.3.2.1.3 Les adjectifs

Il y a un petit nombre d'adjectifs « de base » (têtes de familles dérivationnelles). Les autres sont formés à partir de noms ou de verbes.

Les adjectifs modifient les noms et occupent la position post-nominale ou la position de prédicat en tant que complément de la copule « *yahay* ».

Ils s'accordent seulement en nombre avec les noms qu'ils modifient, et peuvent avoir une forme plurielle par reduplication.

Ils se distinguent des noms et des verbes par le fait qu'ils n'ont pas de morphème déterminant et qu'ils n'ont ni genre ni nombre inhérent.

VI.3.2.1.4 Les adverbes

Les adverbes sont une catégorie grammaticale difficile à identifier en somali, car ils ne peuvent être considérés ni comme une classe lexicale (productive) ni comme une classe grammaticale (mots outils, pronoms, etc., formant des classes non productives). Certains adverbes se forment à partir d'autres mots (noms et adjectifs) tandis que d'autres peuvent seulement jouer le rôle d'adverbe dans la phrase.

---

[63] [suite à une discussion avec Ch. Boitet] Nous préférons le terme « thématiseur » ou « marqueur de thématisation » au terme « marqueur de focus », parfois utilisé en imitation de l'anglais, mais qui est faux. En effet, quand on a une opposition « topic / focus » (ou « tema / rema » dans l'École de Prague, ou « thème / rhème » dans l'analyse statutaire de J. M. Zemb), la thématisation (ou mise en relief, ou emphase) porte bien sur un élément thématique (ce dont on parle, une partie du cadre discursif). Le « focus » ou « foyer » (deux termes incorrects en français), c'est le rhème, ou prédicat, ce qu'on dit du thème. Ainsi, on peut dire « c'est Paul qui accompagnera Pierre demain », ou « c'est demain que Paul accompagnera Pierre », mais pas ou très difficilement « c'est accompagner que fera Paul pour Pierre demain ».

D'un point de vue morphologique, les adverbes se répartissent en adverbes simples, adverbes nominaux, adverbes propositionnels et adverbes compléments.

Selon la position qu'ils occupent dans la phrase, les adverbes peuvent être post-nominaux, postverbaux ou adjectivaux. Ils sont invariables en genre et en nombre, mais peuvent avoir le trait grammatical du cas lorsqu'ils jouent le rôle d'un génitif au sein d'un syntagme nominal.

VI.3.2.2 Les classes grammaticales du somali

VI.3.2.2.1 Les pronoms clitiques

En somali, il existe deux types de pronoms clitiques, les pronoms clitiques sujets et les pronoms clitiques objets. Ils constituent une classe grammaticale et occupent toujours la position préverbale au sein du groupe verbal et peuvent le cas échéant devenir sujet du verbe de la phrase. Dans ce dernier cas, ils subissent une coalescence avec le thématiseur.

a. ***Les pronoms clitiques sujets***

Les pronoms clitiques sujets du somali précèdent toujours le verbe et sont le sujet de la phrase. Leurs équivalents en français sont les pronoms personnels sujets (je, tu, il/elle, nous, vous, ils/elles).

Leurs formes peuvent changer, surtout lorsqu'ils font l'objet d'une coalescence avec les marqueurs de phrases ou avec des thématiseurs comme ***waa***, ***baa*** et ***ayaa***.

b. ***Les pronoms clitiques objets***

Les pronoms clitiques objets assument les fonctions de complément d'objet direct ou indirect du verbe. Il existe deux séries de pronoms clitiques objets (voir Tableau 39).

I	Ku	Na	Ina	Idin

Tableau 39 : Pronoms clitiques objet série 1 et 2

VI.3.2.2.2 Les pronoms indépendants

Les pronoms indépendants ou pronoms emphatiques sont une sous-catégorie de la classe des noms. Ils assument la fonction de sujet ou d'objet dans la phrase, et occupent la position de tête dans les propositions relatives. Ils peuvent être mis en relief par des thématiseurs nominaux tels que ***baa*** et ***ayaa***.

D'un point de vue morphologique, ils ont un genre inhérent, et un cas. Ils peuvent avoir des morphèmes déterminants et avoir une existence indépendante dans la phrase, contrairement aux pronoms clitiques, surtout lorsqu'ils précèdent une conjonction ou des thématiseurs. Le Tableau 40 ci-dessous récapitule les pronoms indépendants du somali :

Pronoms personnels + Ka	**Pronoms emphatiques**
Ani+ga	Aniga (Moi)
Adi+ga	Adiga (Toi)
U+ga	Isaga (Lui)
Ay+da	Iyada (Elle)
Aynu+da	Inaga (Nous inclusif)
Anu+ga	Anaga (Nous exclusif)
Ad+ka	Idinka (Vous)
Ay+ka	Iyaga (Ils/Elles)

Tableau 40 : Les pronoms indépendants somali

VI.3.2.2.3 Les appositions verbales

Les appositions verbales (ou prépositions locatives) servent à indiquer le lieu de l'action du verbe. Elles précèdent souvent ce dernier et ne suivent pas systématiquement le nom auquel elles se rapportent.

La présence des appositions verbales dans une phrase peut changer le sens d'un verbe, par exemple, dans le cas d'une apposition directionnelle ou locative. Il existe 4 appositions verbales en somali : ***u*** (à, pour), ***ku*** (dans, dedans), ***ka*** (de, vers l'extérieur), ***la*** (avec, en compagnie de).

VI.3.2.2.4 Les déterminants

Les déterminants s'affixent aux noms (communs ou propres) sous forme de morphèmes grammaticaux. Ils ne peuvent avoir une existence indépendante dans la phrase, à l'exception des déterminants possessifs, des démonstratifs et des interrogatifs. Les articles définis s'affixent seulement aux noms communs.

Il existe quatre types de déterminants en somali : les articles définis, les déterminants possessifs, les déterminants démonstratifs et les déterminants interrogatifs. Leurs formes masculines commencent par la consonne /***k***/ et leurs formes féminines par /***t***/.

Durant leur affixation avec les noms, les déterminants peuvent subir un changement phonétique dû au phénomène dit « *sandhi* » attesté en somali. Dans certaines phrases, on remarque aussi l'existence de déterminant sous forme de pronoms verbaux, surtout lorsqu'ils sont en position préverbale.

Les articles définis indiquent la position du substantif dans le temps et l'espace. Si le référent du substantif est distant, l'article défini prend les formes suivantes : /k***ii***/ et /t***ii***/.

Par contre, les articles définis /ka/ et /ta/ ne sont pas marqués spatialement et sont utilisés surtout lorsque le substantif est proche et se situe dans le futur ou au présent.

Le tableau ci-dessous récapitule les articles définis en somali.

Articles définis non éloignés	**Articles définis éloignés**
Ka (Masculin)	*Kii* (Masculin)
Ta (Féminin)	*Tii* (Féminin)

Tableau 41 : Les articles définis en somali

Les déterminants démonstratifs s'affixent aux noms propres et communs et peuvent parfois exister d'une manière indépendante dans la phrase.

Leur rôle dans la phrase est d'être le complément ou l'objet. Ils sont toujours suivis par un verbe et son pronom clitique, fusionné avec un thématiseur.

Enfin, les déterminants démonstratifs jouent parfois le rôle de prédicat dans les phrases déclaratives sans verbe.

Le tableau ci-dessous récapitule les articles définis en somali :

Démonstratifs masculins	**Démonstratifs féminins**
Kan (proche)	Tan (proche)
Kaa, Kaas (éloigné)	Taa, Taas (éloigné)
Keer (pas très éloigné)	Teer (pas très éloigné)

Tableau 42 : Les articles démonstratifs du somali

Les déterminants interrogatifs portent aussi sur les substantifs et ont deux formes : le masculin avec **kee** et le féminin avec **tee**. La fusion des interrogatifs avec certains substantifs forme des adverbes interrogatifs. Le tableau ci-dessous récapitule ces interrogatifs.

Si (manière) + tee	Sidee (Comment, de quelle manière)
Hal (lieu) + kee	Halkee (Ou, quel endroit)
Xag (côté) + kee	Xaggee (Ou, de quel côté)
In (quantité indeterminée)+kee	Intee (Combien)

Tableau 43 : Les articles (adverbes) interrogatifs du somali

Les déterminants possessifs du somali s'accordent en genre et peuvent s'affixer avec des substantifs. Le substantif pluriel reçoit le même possessif que son singulier. Le tableau ci-dessous récapitule les déterminants possessifs du somali en fonction du genre.

Masculin	**Féminin**	**Sens**
Kayga	Tayda	Mien/Mienne
Kaaga	Taada	Tien/Tienne
Kiisa	Tiisa	Sien/Sienne (à lui)
Keeda	Teeda	Sien/teeda (à elle)
Kaayaga	Taayada	Le/la Notre (exclusif)
Keenna	Teenna	Le/la Notre (inclusif)
Kiinna	Tiinna	Votre
Kooda	Tooda	Leur

Tableau 44 : Les déterminants possessifs du somali

VI.3.2.2.5 Les marqueurs de types de phrase

Les marqueurs de phrase en somali sont des particules qui permettent d'indiquer le type de la phrase, sa fonction rhétorique ou pragmatique et parfois même le temps (conjugaison). Seules les phrases impératives n'ont pas de marqueurs de type de phrase. Le rôle de ces marqueurs ne se limite pas seulement à indiquer le type de la phrase ; ils mettent aussi en relief tout ou partie de la phrase, comme le sujet, l'objet et le verbe, ou bien des groupes nominaux.

Il existe en tout 5 marqueurs de phrase pour le somali ; ils sont récapitulés dans le tableau ci-dessous.

Marqueurs	**Type de phrase**
Waa	Déclaratives
Ma	Interrogatives positives
Wa	Conditionnelles positives
Hà	Optatives positives
Show	Potentielles positives

Tableau 45 : Marqueurs de type phrase du somali

VI.3.2.2.6 Les conjonctions

Les conjonctions sont des mots invariables qui servent à marquer la liaison entre les unités linguistiques ou constituants d'une phrase (syntagmes). Il existe deux types de conjonction en somali : les conjonctions de coordination et les conjonctions de subordination.

La coordination ou la simple conjonction s'effectue à l'aide de la particule ***iyo*** qui permet de relier des substantifs ou syntagmes nominaux.

Pour assurer la conjonction entre deux propositions complétives, on utilise la particule ***oo***.

La particule ***na*** permet de lier deux propositions, et son équivalent français est « puis ».

Enfin, la conjonction de deux propositions relatives se fait à l'aide des conjonctions suivantes : ***oo***, ***ee***, ***haddana***, ***hase*** et ***yeeshe***.

Enfin, la conjonction disjonctive (ou) est réalisée à l'aide des particules ***ama*** et ***mise,*** et la conjonction de subordination est la particule ***in***.

VI.3.2.2.7 Les thématiseurs

Les thématiseurs du somali mettent en évidence ou focalisent certaines parties de la phrase comme le syntagme nominal (ou substantif) ou le noyau verbal (ou le verbe).

Leur rôle consiste à indiquer l'information la plus importante ou saillante de la phrase. Ces thématiseurs sont considérés comme des fonctions du discours ou fonctions pragmatiques. Leurs équivalents en français sont les clivées telles que 'C'est ….. X qui/que…'.

Le tableau ci-dessous récapitule les différents thématiseurs du somali.

Marqueurs	**Thématisation (ou emphase)**
Baa/Ayaa	Syntagme nominal, substantif ou noms
Waxa/Waxaa	Syntagme verbal, verbe

Tableau 46 : Thématiseurs du somali

VI.3.2.2.8 Les interjections

Les interjections (ou « phatiques ») constituent une classe grammaticale. Elles sont utilisées souvent dans le discours oral, pour attirer l'attention de l'interlocuteur, exprimer ses émotions ou son état d'esprit (ses affects).

Les interjections prennent des formes différentes selon qu'on s'adresse à un interlocuteur masculin ou féminin, ou qu'on s'adresse ou parle à des animaux. Toutefois, elles ne portent aucun trait grammatical et ne jouent aucun rôle dans la structure morphologique des phrases dans lesquelles elles apparaissent.

Elles sont utilisées dans le cadre de discussions ou de conversations, ou pour répondre à des questions ou à quelqu'un ou dans un discours poétique. Le tableau ci-dessous récapitule les interjections les plus fréquemment rencontrés dans du texte ou discours oral somali.

Interjections	**Sens**
Hàa	Oui
May	Non
Màya	Non
Hée	Oui !
Hàye	Et alors !
Hàayye	D'accord
Hooyàale	Métrique poétique
Hée dhé	Dis-moi

Tableau 47 : Les interjections du somali

VI.3.3 La morphologie flexionnelle et dérivationnelle du somali

La morphologie est une branche de la linguistique qui traite de la structure interne des mots. L'approche qui domine le champ linguistique dans ce domaine depuis les années 50 est l'approche dite structuraliste. Selon cette approche, les mots d'une langue sont formés de

morphèmes ou lexèmes qui sont les unités minimales (c'est à dire indécomposables) porteuses de sens.

Un *morphème* est une unité abstraite qui peut avoir plusieurs formes graphiques et phoniques appelées *morphes*. Deux morphes différents sont appelés *allomorphes* lorsqu'ils correspondent au même morphème.

On distingue traditionnellement la morphologie dérivationnelle, qui étudie la relation entre lexèmes qui peuvent être considérés comme appartenant à une même famille ou classe de mots, et la morphologie flexionnelle, qui étudie la relation entre les différentes formes d'un même lexème.

En somali, ces deux sous-branches de la morphologie fournissent chacune des indications sur les différentes classes lexicales et par conséquent leur étude théorique et leur compréhension sont nécessaires pour mener à bien les opérations de lemmatisation et d'étiquetage lexical.

Dans les sections ci-dessous, nous parlerons dans un premier temps de la morphologie flexionnelle et dans un second temps de la morphologie dérivationnelle. Ainsi mettrons-nous l'accent tout au long de cette partie sur leurs rôles dans la morphosyntaxe des mots de cette langue.

VI.3.3.1 Les morphèmes flexionnels

VI.3.3.1.1 Propriétés des morphèmes flexionnels

Les morphèmes flexionnels du somali sont des morphèmes grammaticaux qui ont leurs propres paradigmes ; ils se distinguent des morphèmes dérivationnels par leurs propriétés sémantiques et combinatoires.

Les trois propriétés ci-dessous permettent de définir les morphèmes flexionnels.

- Les morphèmes flexionnels ne permettent pas de créer de nouveaux mots. Par exemple, les mots *nin* (homme), *niman* (hommes), *nimanyahaw* Ô hommes) sont quatre formes différentes du même lexème *NIN*.
- L'affixation ou l'adjonction d'un morphème flexionnel à une catégorie lexicale ou grammaticale suit certaines règles qu'on peut qualifier de combinatoires. Il existe différents morphèmes du pluriel pour les noms, les adjectifs et les verbes.
- Les morphèmes flexionnels ne modifient pas le sens du lexème auquel ils s'appliquent. La différence sémantique et graphique entre un radical et sa forme fléchie se situe seulement au niveau de l'information grammaticale supplémentaire apportée par le morphème flexionnel.
  Par exemple, la différence sémantique entre *naag* et *naago* ou *yar* et *yaryar* réside dans les morphèmes du pluriel é /-*o*/ et /-*yar*/, respectivement pour le nom commun *naag* et pour l'adjectif *yar*. Ces morphèmes indiquent le changement du trait grammatical nombre du nom et de l'adjectif qui est maintenant au pluriel.

Les morphèmes flexionnels réalisent les variables ou traits grammaticaux morphologiques sur les classes lexicales. Ce sont des indices formels qui permettent d'indiquer les rapports que peut entretenir la base d'un mot avec le reste de l'énoncé.

Par exemple, la variable nombre est réalisée par des morphèmes flexionnels différents, selon qu'elle s'applique à un nom, un verbe ou à un adjectif.

La morphologie non verbale peut être définie négativement : elle regroupe tous les phénomènes de variation grammaticale qui peuvent affecter les classes lexicales non verbales.

Dans la section suivante, nous parlerons des différentes variables et de morphèmes flexionnels associés, qui s'appliquent aux classes nominales.

VI.3.3.1.2 Les variables de la flexion non verbale

a. ***Le genre***

Le genre est une variable qui concerne en somali les classes lexicales du nom et de l'adjectif, et les classes grammaticales des déterminants et des pronoms.

C'est une catégorie grammaticale linguistique, distincte du genre naturel (sexe). Ce dernier est une catégorie extralinguistique qui classe les référents humains ou les espèces animales en mâle et femelle. Ce classement étant arbitraire et naturel, il ne peut avoir une incidence sur les catégories grammaticales des mots.

Les valeurs que peut prendre le genre sont seulement le masculin ou le singulier ; il n'existe pas de genre neutre en somali. Le genre en somali est souvent arbitraire pour les mots de base.

Il existe également le phénomène de *polarité* attesté dans d'autres langues couchitiques. La polarité concerne surtout les noms, qui peuvent changer arbitrairement de genre en passant du singulier au pluriel.

Ce changement de genre s'effectue soit par la forme des morphèmes déterminants suffixés aux noms, soit par l'accent tonique, soit par la présence de certains morphèmes de pluralité qui permettent d'indiquer le genre du nom.

Les propriétés ci-dessous permettent d'indiquer le genre des classes de mots du somali.

- Pour la classe du nom, qui regroupe les noms communs, les noms propres et les numéraux, la forme orthographique du morphème déterminant permet d'indiquer leur genre, lorsqu'ils sont définis.
- Dans le cas d'un nom indéfini, l'accent tonique ainsi que l'accord grammatical avec le verbe permet de déterminer le genre.

b. ***Le nombre***

Le nombre est une catégorie grammaticale qui concerne les classes lexicales des noms et adjectifs. Il apporte une information grammaticale supplémentaire sur le lexique et permet de distinguer les mots singuliers des mots pluriels.

Son rôle ne se limite pas seulement à distinguer les éléments uniques ou isolés des éléments pluriels ; il permet aussi d'opérer une sous-catégorisation à l'intérieur de la classe des noms, entre les noms comptables et les noms non comptables.

Un nom comptable doit obligatoirement avoir un morphème de pluralité qui peut prendre différentes formes. Cela permet définir ce que les linguistes comme [Saeed, 1999] ont appelé « déclinaisons nominales ».

La sous-catégorie des noms non comptables regroupe les noms de masse et les noms collectifs. Les noms de masse n'affichent pas normalement une distinction entre un singulier et un pluriel mais se subdivisent entre des noms ne pouvant s'accorder qu'avec des verbes conjugués à la 3ème personne du pluriel, et des noms ne s'accordant qu'avec des verbes aux personnes du singulier. Ainsi, les noms de masse n'ont pas un genre inhérent.

Les noms collectifs n'ont pas d'information morphologique indiquant leur nombre, mais ils peuvent former des syntagmes nominaux avec des numéraux dont ils constituent la tête dominante dans une proposition relative. Contrairement aux noms de masse, ils ne

s'accordent pas avec les verbes conjugués lorsque leur syntagme occupe la position de sujet dans la phrase.

Enfin les noms collectifs ou de masse qui finissent par la voyelle /-o/, qui est un morphème de pluralité, s'accordent en pluriel avec les verbes dont ils sont sujets.

Ainsi, seuls les noms communs comptables peuvent avoir un paradigme flexionnel de pluralité. Ce paradigme forme l'ensemble des déclinaisons possibles d'un nom et sera traité dans la section ci-dessous.

En effet, il existe différentes façons pour la catégorie grammaticale du nombre de manifester l'opposition entre les noms comptables pluriels ou singuliers. La forme au singulier des noms comptables étant la forme de base de ces mots, nous traiterons dans les sections suivantes les différentes manières de former le pluriel des noms comptables en somali.

VI.3.3.1.3 Formation du pluriel des noms comptables

a. ***Pluriels prosodiques***

Ces pluriels sont appelés pluriels prosodiques parce qu'ils ne contiennent pas de suffixe de pluralité, et que la formation du pluriel s'effectue au niveau du changement de ton et par l'accentuation de certaines voyelles.

Tous les noms masculins qui ont un ton haut pénultième deviennent féminins au pluriel. Le changement du genre d'un mot du masculin au féminin se manifeste à l'aide de l'accord de ces mots avec l'article définies **Ka** (masculin) qui se transforme en **ta** au féminin pluriel.

Le Tableau 48 donne quelques exemples de noms somalis avec leurs pluriels prosodiques.

Noms	Articles définis	Genre	Noms	Articles défnis	Genre	Mots en français
Dibi	(-ga)	M	Dibi	(-da)	F	Boeufs
Ey	(-ga)	M	Ey	(-da)	F	Chiens
Arday	(-ga)	M	Arday	(-da)	F	Étudiants
Soomali	(-ga-	M	Soomaali	(-da)	F	Somaliens
Madax	(-a)	M	Madax	(-da)	F	Têtes (Dirigeants)
Tuug	(-a)	M	Tuug	(-ta)	F	Voleurs

Tableau 48 : Exemples de pluriels prosodiques

D'après [Lecarme, 2002], les noms pluriels de cette classe deviennent féminins lorsqu'ils sont associés à des articles définis ou des pronoms, une propriété partagée avec la langue arabe.

[Lecarme, 2002] réfute également l'idée de certains auteurs qui assignaient une classe sémantique à cette catégorie de pluriels.

Pour elle, les pluriels prosodiques sont simplement des pluriels avec zéro suffixation, du fait qu'il n'y a pas de suffixe de pluralité dans leurs formes au pluriel.

b. ***Pluriels internes avec le morphème /-aC/***

Certains noms comptables du somali ont des pluriels internes lorsqu'ils se terminent par le morphème /**a**/. Ainsi, tous les noms masculins monosyllabiques qui se terminent par un a accentué voient leurs dernières consonnes doubler après le morphème a, et cela avec copie de la dernière consonne du radical.

Noms	Articles définis	Genre	Noms	Articles définis	Genre	Mots en français
Af	(-ka)	M	Afaf		M	Bouches
Nin	(-ka)	M	Niman	(-ka)	M	Garçons
War	(-ka)	M	Warar	(-ka)	M	Informations
Roob	(-ka-	M	Roobaab	(-ka)	M	Pluies
Tuug	(-ga)	M	Tuugag	(-ga)	M	Voleurs

Tableau 49 : Exemples de pluriels internes avec le morphème /-aC/

Dans ces pluriels internes, les noms singuliers masculins restent toujours masculins au pluriel, d'où l'absence de polarité dans les pluriels internes avec le morphème /**–a**/.

c. ***Pluriels internes avec le morphème –Co***

Lorsque les noms comptables singuliers sont constitués de plus d'une syllabe et qu'ils se terminent par une consonne, cette dernière est copiée dans les pluriels de ces mots, et on remplace la dernière voyelle a du mot par o.

En cas d'impossibilité de copier la dernière consonne, c'est-à-dire lorsque le mot se termine par une consonne gutturale (/**x**/, /**c**/, /**j**/, /**q**/, /**s**/, /**g**/) ainsi que la voyelle /**i**/, on ajoute au radical du mot le suffixe de pluralité /**-yo**/ pour former son pluriel.

Noms	Articles définis	Genre	Noms	Articles définis	Genre	Mots en français
Inan	(-ka)	M	Inammo	(-a-da)	F	Garçons
Doofar	(-ka)	M	Doofaarro	(-a-da)	F	Singes
Qalin	(-ka)	M	Qalimmo	(-a-da)	F	Stylos
Dagaal	(-ka-	M	Dagaallo	(-a-da)	F	Guerres
Dhaagax	(-a)	M	Dhaagaxyo	(-a-da)	F	Cailloux

Tableau 50 : Pluriels internes avec le morphème -Co

Comme on le voit dans les exemples ci-dessus, les noms polysyllabiques masculins changent de genre dans leurs formes plurielles, d'où l'existence de la polarité dans ces pluriels.

d. ***Pluriels internes avec le morphème –o (M)***

Les mots en somali deviennent pluriels après l'ajout de la voyelle /**-o**/ à la fin du mot comme suffixe de pluralité.

Dans ces pluriels, tous les mots féminins singuliers ont une polarité de genre dans leurs pluriels, tandis que les mots masculins singuliers restent toujours masculins au pluriel.

e. ***Pluriels avec –yaal (F ou M) et –oyin (M)***

Les morphèmes de pluralité /**–yaal**/ et /**–oyin**/ peuvent être suffixés à des mots masculins se terminant par une voyelle pour former leurs pluriels.

Seuls les mots qui forment leurs pluriels avec le morphème de pluralité /**-oyin**/ ont une polarité dans leur genre tandis que les pluriels avec le morphème /**-yaal**/ n'ont pas de polarité dans leur genre.

f. ***Le cas en somali***

Le cas est une catégorie grammaticale permettant d'indiquer la fonction syntaxique d'un mot dans la proposition ou la phrase. Il concerne surtout la classe lexicale du nom, et plus particulièrement ses deux sous-classes du nom commun et du noms propre.

Le cas d'un nom s'exprime soit par l'adjonction d'un morphème de cas à la fin du nom, soit par la présence d'un accent tonal.

Le cas d'un nom en somali dépend surtout du rôle que ce nom joue dans la phrase ou dans la proposition subordonnée. Les noms féminins singuliers ont un accent tonal haut (AP1) au cas absolutif, tandis que les noms masculins singuliers ont un accent tonal bas (AP2), ainsi que tous les noms pluriels, à l'exception des noms masculins de la déclinaison 2B et 7 qui ont un accent tonal haut AP1 au singulier, tandis que les noms de la déclinaison 6 ont un accent tonal AP2 au pluriel.

Le changement de l'accent tonique est réalisé souvent lors du passage d'un nom du cas absolutif aux autres cas.

Le tableau ci-dessous indique le passage d'un cas à la dernière more du mot.

ABS→	**Nominatif (NOM)**	**Génitif (GEN)**	**Vocatif (VOC)**
	AP3	AP1	AP4

Tableau 51 : Les cas du somali

Nom	**Cas**	**Nom**	**Cas**	**Nom**	**Cas**	**Nom**	**Cas**
Cáli	ABS	Cali	NOM	Calí	GEN	Calí	VOC
bisád	ABS	bisadi	AP3	bisád	AP1		
carrúur	ABS	carruuri	NOM	carrúur	GEN		

Tableau 52: Cas des noms somalis

Le cas des noms peut prendre quatre valeurs : absolutif, nominatif, génitif et vocatif.

f.i **Le cas absolutif**

C'est le cas par défaut de tous les noms communs en somali, c'est-à-dire lorsque le nom n'est ni un sujet ni un nom vocatif ou génitif. Souvent l'objet direct ou indirect des phrases en somali est à ce cas. Les noms associés avec des prépositions sont également à ce cas, ainsi que les noms communs de la déclinaison 1,2 et 3. C'est également la forme de citation ou le lemme d'un nom en somali.

mindí (couteau)

naág (femme)

babuúr (voiture)

f.ii **Le cas nominatif**

Ce cas est utilisé lorsque le nom est le sujet de la phrase.

Le nom est sujet lorsqu'il est indéfini ou n'a pas de suffixe déterminant (article) ni un autre suffixe, adjectif ou non génitif. La forme du cas sujet est marquée par l'absence d'accent tonal sur les voyelles (longues ou brèves) du nom.

Pour marquer ce cas, on ajoute un /-i/ aux noms féminins finissant par une consonne.

mindi couteau (sujet)

naagi femme (sujet)

baabuur voiture (sujet)

f.iii **Le cas génitif**

Le génitif est utilisé en somali pour indiquer la possession, c'est-à-dire pour indiquer qui possède tel objet ou bien encore le lien parental entre un parent et son enfant.

Ce cas est marqué phonologiquement par l'accentuation de la pénultième du nom du possesseur et l'ajout du suffixe /**-i**/ au nom possédé.

Búuggií Maxaméd (Buug) « Le livre de Mohamed (Mohamed's book) ».

Gabadhií Cali (Gabadh) « La fille d'Ali »

Par ailleurs, les noms qui sont féminins au singulier et qui ne se terminent pas par la voyelle /-**i**/ peuvent former leur génitif par l'ajout du suffixe génitif /-**eéd**/ pour les noms finissant par une consonne et /-**yeéd**/ pour se terminent par la voyelle /-**i**/.

Xanúun lugeéd [luug- eéd] -> Maladie de pied

Af shimbireéd [Shimbir-**eéd**] -> Langue des oiseaux.

Les noms féminins au singulier et qui forment leurs pluriels avec le morphème de pluralité /-***o***/ deviennent des noms génitifs par l'ajout du suffixe agentif /-***ó***/.

Exemple : mídabka shimbiroód [Shimbir-**o-ód**] La couleur des oiseaux

Les noms des animaux domestiques peuvent former leurs génitifs par l'ajout du suffixe génitif /**-aád**/ à la place de /**-eéd**/ ou /**-oód**/.

caanó lo**'aád** (lo') -> Lait de vache

caanó riy**aád (riyo, rí')** -> Lait de chèvre.

f.iv **Le cas vocatif**

On ajoute des suffixes vocatifs aux noms lorsqu'on s'adresse directement à quelqu'un pour le héler, attirer son attention, lui crier secours ou bien pour obtenir l'aide d'un personnage saint. On utilise le cas vocatif à la fois avec des noms communs et avec des noms propres.

Il existe six suffixes vocatifs, dont un pour le masculin singulier et les féminins pluriels des noms communs : /**-yahaw**/, un autre pour les noms communs féminins singuliers : /**-yahay**/, un autre pour les noms propres masculins (ou mâles) : /**-ow**/ et enfin trois suffixes vocatifs pour les noms propres féminins (ou femelle) : /**-eey**/, /**-ay**/, /**-oy**/.

Nin**yahaw** (nin) « Oh homme ! »

Gabdhayahaw (Gabdh**-o-yahaw**) « Oh les filles »

Naag**yahay** (naag) « Oh femme ! »

Cheekh Cabdulqaadir**ow !** (Cadulqaadir) « De cheikh Abdoulkader (prénom d'un saint irakien)

Ruunn**eey**! (Ruun) « Oh verité! » Caashaay! (Caasha) « Oh Aicha ! »

Cibaadooy! (Cibaado) « Oh Ibado:

VI.3.3.1.4 Les variables de la flexion verbale

Lors de la conjugaison d'un verbe en somali, le mot-forme obtenu celui-ci est composé d'une part d'un morphème lexical appelé radical, et d'autre part d'une désinence verbale qui elle-même se décompose en morphème de mode, d'aspect, de temps et de personne.

Schématiquement, un verbe conjugué en somali se compose de la façon suivante :

[Radical+ Affixes+ Accord + Flexion], où **:**

- **Radical** est la forme de base du verbe, c'est-à-dire sa forme à la 2ème personne du singulier de l'indicatif ou de l'impératif,
- **Affixes** contient les affixes ajoutés au radical pour dériver d'autres verbes,
- **Accord** contient des informations sur le genre et le nombre de la conjugaison
- **Flexion** contient les informations sur le temps, le mode et l'aspect de la conjugaison sous forme de suffixe.

Dans les sections ci-dessous, nous allons présenter les différents morphèmes qui s'ajoutent aux verbes après le radical et les affixes lexicaux. Il s'agit des morphèmes de mode, d'aspect, de temps et de personne.

a. ***Le mode***

Le mode est une catégorie grammaticale de la morphologie des verbes du somali. Il décrit la façon dont le verbe exprime le fait, l'état ou le procès. Ce dernier peut être affirmatif, réel, éventuel, un ordre, un conseil ou même un souhait etc. Les différents types de mode verbal du somali sont : le déclaratif, l'impératif, l'interrogatif, le conditionnel, l'optatif et le potentiel.

Contrairement à d'autres langues, le type de mode verbal ne se manifeste pas systématiquement par des désinences verbales (terminaisons). En effet, certaines catégories ou particules grammaticales comme les marqueurs de phrase ou les particules interrogatives peuvent indiquer le mode.

a.i **Le mode déclaratif**

Le mode déclaratif est l'un des deux modes des verbes du somali qui n'a pas des désinences verbales. Les marqueurs de phrases déclaratifs qui précèdent le verbe conjugué permettent d'indiquer ce mode.

a.ii **Le mode impératif**

Le mode impératif est utilisé pour donner des ordres ou des injonctions. C'est un mode personnel et il ne comporte que trois personnes : première personne du singulier, deuxième personne du singulier et enfin deuxième personne du pluriel.

Pour les formes verbales de base (ou première conjugaison) et dérivées, le morphème du mode impératif est nul à la deuxième personne du singulier. Cette forme de l'impératif est considérée comme étant la forme de citation ou de l'infinitif des verbes en somali (Saeed, 1999 & Abdullahi, 1995).

Pour la deuxième personne du pluriel, le suffixe « a » s'ajoute comme suffixe à la forme infinitive du verbe de base, le suffixe « ya » pour les verbes dérivés causatifs et factitifs, le suffixe « ada » pour les verbes d'expérience, et enfin le suffixe « a » pour les verbes autobenéfactifs .

La forme du morphème du mode impératif change lorsque le verbe est précédé de particules négatives.

a.iii **Le mode conditionnel**

Le mode conditionnel exprime une situation hypothétique dans le présent ou le passé. Il est utilisé pour un acte, un événement ou une action qui ne s'est pas réalisée mais que le locuteur aimerait réaliser. Ce mode n'a pas de morphème grammatical, car la conjugaison d'un verbe

au conditionnel se réalise sous une forme composée avec l'utilisation de l'adjectif « ***leh*** » (acte de posséder ou d'avoir quelque chose) auquel on ajoute la copule « ***yahay*** » dans sa forme du passé. Seuls les verbes de la première, deuxième et troisième conjugaison peuvent se conjuguer au mode conditionnel.

a.iv **Le mode optatif**

Le mode optatif désigne un mode irréel et il est surtout utilisé pour exprimer un souhait, un espoir ou un conseil. Il se caractérise par la présence de la particule « ha » devant le verbe qui est conjugué à la forme du subjonctif. Comme le mode conditionnel, ce mode ne concerne que les trois premières conjugaisons verbales du somali.

a.v **Le mode potentiel**

Le mode potentiel est lui aussi un mode irréel ; il désigne la possibilité, la praticabilité de réaliser un acte, un événement ou un procès. Il se caractérise aussi par la présence devant le verbe de la particule de potentialité « show ». Il n'y a pas de forme négative pour ce mode.

b. ***Le temps***

Le temps est une catégorie grammaticale permettant de localiser l'événement, l'état, l'action ou le procès porté par le verbe sur un axe linéaire à trois régions qui sont le présent, le passé et le futur.

Le présent et le passé se réalisent morphologiquement au moyen de morphèmes de temps, alors que le futur prend une forme composée avec la présence de l'auxiliaire « doon » (vouloir) qui spécifie le futur. Les morphèmes de temps font partie du paradigme flexionnel de la conjugaison des verbes. Ils peuvent changer leurs formes selon que la forme conjugale du verbe est précédée de particule de négation ou que le verbe se trouve dans la proposition principale ou subordonnée dans le cadre d'une phrase complexe.

Le présent est spécifié par la présence du morphème flexionnel de temps « aa » placé juste après le radical, et le passé par « ay ».

Le verbe auxiliaire du futur contient le morphème flexionnel du temps présent et le verbe principal se termine par le suffixe « i ». Cette dernière opération suit certaines règles phonologiques telles que l'assimilation et conduit parfois à modifier la forme du verbe.

c. ***L'aspect***

L'aspect est une catégorie grammaticale de la flexion verbale qui permet de spécifier la façon dont le procès, l'état ou l'action exprimée par le verbe s'est déroulé. Il est marqué morphologiquement par la présence ou non d'un affixe d'aspect. Ce dernier permet surtout de distinguer l'aspect accompli du progressif ou inaccompli.

Il existe trois aspects en somali.

- **Accompli.** Cet aspect indique un état, une action ou un procès qui s'est achevé et qui n'a eu lieu qu'une seule fois. Il concerne à la fois le présent et le passé.
- **Progressif (ou inaccompli)**. Cet aspect indique un état, un procès ou une action qui est en progression ou en train de se réaliser. On remarque cet aspect par la présence du morphème d'aspect ***ay*** entre le verbe et le morphème de temps.
- **Habituel**. Cet aspect indique un état, un procès ou une action qui est habituel ou qui se répète à la fois au présent et au passé. On reconnait cet aspect à la présence de l'auxiliaire « jir » après un verbe conjugué à cet aspect au passé, et à l'adjonction des morphèmes de temps au verbe auxiliaire.

d. ***La personne***

Le morphème flexionnel de personne est une catégorie grammaticale permettant d'indiquer le genre ou le nombre du verbe conjugué. Cette information spécifie surtout l'accord du verbe avec son sujet. Il existe cinq morphèmes de personne en somali, qui sont les suivants.

- La deuxième personne du singulier et la troisième personne du féminin ont le même morphème de personne qui est « **t** ».
- La première personne du pluriel (inclusif ou exclusif) a pour morphème de personne « **n** ».
- La deuxième personne du pluriel a pour morphème de personne « **t** ».
- Enfin, il n'existe pas de morphème de personne pour la première personne du singulier et pour la troisième personne du singulier masculin.

VI.3.3.2 Les morphèmes dérivationnels

VI.3.3.2.1 Propriété des morphèmes dérivationnels

Les morphèmes dérivationnels du somali sont des morphèmes grammaticaux non autonomes qui prennent souvent la forme d'un suffixe. Leur rôle consiste plutôt à créer de nouvelles unités linguistiques ou mots à partir des unités de base susceptibles de recevoir des morphèmes dérivationnels, comme les classes lexicales des noms, verbes et adjectifs.

Ainsi, en somali, les morphèmes dérivationnels relèvent à la fois du domaine de la morphosyntaxe et de la morpholexicologie.

Les propriétés ci-dessous permettent de distinguer les morphèmes dérivationnels des morphèmes flexionnels :

- Les nouvelles unités linguistiques créées à l'aide des morphèmes dérivationnels changent de catégorie grammaticale ou de classe de mots.
- L'opération de dérivation est régie par des critères et des restrictions sémantiques. Ainsi, chaque morphème dérivationnel du somali opère sur une seule classe lexicale et produit une nouvelle unité linguistique avec une signification précise et une nouvelle catégorie grammaticale.
- Toutes les unités linguistiques formées avec des morphèmes dérivationnels peuvent recevoir des informations supplémentaires sous forme de morphèmes flexionnels, en fonction de leur nouvelle catégorie grammaticale.

VI.3.3.2.2 La dérivation non verbale

a. ***Les noms dérivés***

a.i **Dérivation d'un nom à partir d'un verbe**

Il existe plusieurs suffixes verbaux qui permettent de créer de nouveaux noms à partir des verbes. Généralement, les nouveaux noms créés à l'aide de ces suffixes verbaux jouent le rôle de noms d'agent ou de gérondifs. Chaque groupe ou catégorie de verbe en langue somali a son propre suffixe verbal pour dériver de nouveaux noms.

Le suffixe verbal /**-id**/ ou /**-is**/ est réservé pour la dérivation nominale à partir des verbes de base ou de la première conjugaison, le suffixe /**-a**/ est réservé pour dériver des noms à partir des verbes autobénéfactifs, le suffixe /***-in***/ à partir des verbes causatifs. Au total, il existe selon [Barillot, X., 2002] une quarantaine de suffixes verbaux en somali qui permettent de dériver des noms à partir d'un verbe, autrement dit des noms déverbaux.

a.ii **Dérivation d'un nom à partir des suffixes nominaux**

Il y a plusieurs types de noms formés par l'ajout de suffixes nominaux à des verbes ou des adjectifs. Les noms verbaux féminins abstraits sont formés par la suffixation du suffixe /***-id***/ avec des verbes de base. Pour les verbes factitifs et causatifs, on utilise le suffixe /***-n***/ et enfin le suffixe /***-asho***/ pour les verbes médio-passifs.

Il existe trois suffixes nominaux (/**-nimó/**, / **-tooyo**/ et /**-ád**/) qui permettent de créer des noms de qualité à partir des noms communs par suffixation.

Pour dériver un nouveau nom, le suffixe nominal /**-nimó/** peut s'ajouter à une racine nominale **(a)**, à un nom lui-même dérivé **(b)** ou bien à une forme composée **(c)**. Ce suffixe porte toujours l'accent tonal et **le nom dérivé avec ce suffixe est toujours féminin et n'a pas de forme plurielle**.

- Racine nominale + /-nimó/

  Carruur (Enfants) + /**-nimó/**→ Carrur**nimó** (Enfance)

  Gabar (Vierge) + /**-nió/** →Gabar**nimó** (Vierginité)

- Nom derivé + /-nimó/

  Baré **[**Bar+**é] (**Enseignant**) + /-nimó/** → Baré**nimó** (Enseignement)

  Duuliye [duul+**i**+**é**] (Pilote) + **/-nimó/** → Duuliye**nimó (**Pilotage**)**

- Nom composé

  Afmiinshaar [Af (bouche) +miinshar (troueuse)] (Débateur éloquent) + /**nimó/** → Afmiinshaar**nimó (**Le fait d'être un débateur éloquent**)**

  Dhiigyacab [Dhiig (sang) + acab (boire)] (Sanguinaire) + /**nimó/** → Dhiigayacab**nimó** (Le fait d'être sanguinaire)

Le suffixe nominal /**-tooyó/** est lui aussi porteur d'un accent tonal ; **le nom dérivé avec ce suffixe est toujours au féminin singulier** et à la différence du suffixe /**nimó/,** ce suffixe ne s'affixe qu'aux formes radicales des noms (a) et des verbes (b).

- Formes radicales nominales avec /-tooyó/

  Gacal (Belle famille) + /**-tooyó/** → Gacal**tooyó (**Lien de belle famille**)**

  Madax (Tête/Responsable/Chef**) + /-tooyó/** → Madax**tooyó** (Présidence**)**

  Boqor (Roi) **+ /-tooyó/** → Boqor**tooyó** (Monarchie/Royaume)

- Formes radicales verbales avec /-tooyó/ (Seulement 2 verbes)

  Xad (Voler) + **/-tooyó/** → Xad**tooyó** (Vol)

  Qad (Jeûner) + **/-tooyó/** → Qad**tooyó** (Jeûne)

Le suffixe nominale /**-ád**/ a lui aussi un accent tonal et permet de dériver un nom à partir d'un nom ou d'un adjectif.

Si le nom auquel il s'affixe se termine par la voyelle /**-i**/, un /**-y-**/ est inséré entre le nom et le suffixe (/-i-y- **ád**). **Le nom dérivé avec ce suffixe est toujours féminin au singulier**.

Burjuwaási (Bourgeoisie) + /**- ád /** → Burjuwaasiy**ád** (Bourgeoisie ou le fait d'être bourgeois)

Dammiin (Garant) + /**- ád /** → Dammiin**ád**/Dammaan **ád** (Garantie)

Xanaq (Faute) + /**- ád /** → Xanaq**ád** (Faute)

a.iii **Dérivation d'un nom à partir d'un nom**

La dérivation d'un nom à partir d'un autre nom permet en langue somalie surtout de dériver des noms féminins à partir de noms masculins à l'aide du morphème de féminin /**-ad #**/ ou des noms pluriels à partir de noms masculins, grâce à un changement de l'accent tonal sur les voyelles du mot, et sans l'ajout d'un suffixe dans ce dernier cas.

Contrairement aux mots dérivés formés à partir d'autres classes lexicales traitées dans la section ci-dessus, les nouveaux noms ici dérivés sont seulement formés à partir de noms.

Le suffixe féminin /**-ad#/** est un emprunt à l'arabe, il s'agit de la forme somalisée du morphème de genre de l'arabe /**-at#**/.

- Avec le morphème /**-ad**/ :

  Boqor (Roi) + /**-ad**/ → Boqor**ad [**En arabe Malik (Roi) → Malika (t)]

  Arday (Etudiant) + /**-ad**/→ Arday**ad** (Etudiante)

  Curyaan (Handicapé) + /**-ad**/→ Curyaan**ad** (Handicapée)

  Daayéer (Singe) + /**-ad**/ → Daayeer**ad** (Guenon)

- Avec le changement de l'accent tonal du somali, on peut changer le genre ou le nombre d'un nom en somali.

  Dam**é**er- (Ane) → Dame**é**r (Anesse)

  Inan (garçon) → inán (Fille)

  Ey (chien)→eý (chiens)

b. ***Les adjectifs dérivés***

Les adjectifs de base ne sont pas très nombreux en somali ; la plupart des adjectifs sont dérivés à partir de verbes grâce à l'affixe statique /**-an**/ et une autre partie dérive à partir des noms. Les adjectifs dérivés à partir d'un verbe causatif par exemple décrivent l'état achevé par la cause. On ajoute un allomorphe /**-s**/ pour dériver un adjectif d'un verbe causatif. Leurs équivalents en français sont les participes passés à valeur d'adjectif.

kári (Cuis) + /-an/ → karsán (déjà cuit)

Búuxi (remplis) + /-an/ → Búuxsan (plein ou déjà remplis)

Les mêmes effets sémantiques et phonologiques se produisent aussi lorsqu'on dérive un adjectif d'un verbe factitif.

afée [af+/-ee/] « Faites lui une bouche ou sortie/aiguiser » + /-an/ → afaysan « aiguisé »

Sumée « empoisonner » + /**-an**/→ sumaysán « Empoisonné »

Avec les verbes inchoatifs, cet affixe forme des adjectifs dérivés décrivant l'effet du verbe sur l'objet.

àamus + /**-an**/ « silence » → aamusàn « Silencieux »

Engeg + /**-an**/ « sec » → éngegàn « Devenir sec/à sec »

VI.3.3.2.3 La dérivation verbale

La dérivation verbale est un processus morphologique permettant de dériver de nouveaux verbes à partir des verbes de base, des noms ou des adjectifs. La dérivation verbale permet de changer la catégorie grammaticale du mot lorsque la forme de base n'est pas un verbe. Les différents groupes ou catégories de verbes en somali sont formés également avec la dérivation verbale. Ainsi un verbe de base ou de conjugaison 1 peut devenir de conjugaison 2, 3 et 4 selon la forme du suffixe auquel on l'affixe.

La dérivation permet aussi la détermination de la catégorie syntaxique du verbe ; c'est grâce à la dérivation que les verbes intransitifs, statifs et factitifs se forment en somali.

a. ***Les verbes causatifs***

Il existe deux types d'affixes causatifs /**-is**/ et /**-i**/ qui permettent de dériver des verbes causatifs à partir de verbes de base ou de verbes inchoatifs. Le sens d'un verbe causatif est d'exprimer l'idée que quelqu'un a causé ou provoqué l'acte de faire quelque chose. Les affixes causatifs ajoutent ainsi une information supplémentaire sur la cause de l'action.

Au niveau de la conjugaison, les verbes causatifs font partie du deuxième groupe des verbes. Dans la conjugaison d'un causatif, on ajoute la consonne /-y/ devant un morphème flexionnel commençant par une voyelle. Les verbes causatifs deviennent transitifs après la dérivation.

Exemple de verbes causatifs

Búux (Intr) « être plein » → búuxi (Tr) « Remplir »

Kár (Intr) « chauffe » → kári (Tr) « Cuisiner »

b. ***Les verbes inchoatifs***

Les verbes inchoatifs sont des verbes intransitifs formés par l'ajout de morphèmes verbaux à des formes qui ne sont pas verbales comme les noms communs masculins ou féminins.

Il existe deux suffixes qui permettent de créer ce type de verbe à partir d'un nom (**-ow**) ou à partir d'un adjectif (**-aan**). Le verbe obtenu prend l'accent tonal du nom après la suffixation.

Le sens des verbes créés par cette dérivation est généralement « devient/devenir N/Adj » pour N le sens du nom de base et Adj le sens de l'adjectif. Ils appartiennent au quatrième groupe des verbes dans la classification des groupes verbaux en somali.

Suffixe verbal /**-ow**/ avec des noms

Báraf (Glace/Neige) + /**-ow**/ → baraf**ów** (Devient neige, glace ou avoir froid)

Biyó (Eau) + /**-ow**/ → Biy**ów** (Devenir liquide, avoir de l'eau sur soi même)

Gáal (Infidéle) + /**-ow**/ →gaal**ów** (Devenir infidèle)

Magáalo (Ville, centre urbain) + /**-ow**/ → Magaal**ów** (Devenir une ville, un citadin)

Nácas (Imbécile) + /**-ow**/ → Nacas**ów** (Devenir imbécile)

Túug (Voleur) + /**-ow**/ → Tuug**ów** (Devenir voleur)

Adág (Dur/Solide) + /**-aan**/ → adk**ów** (Devenir dur/solide)

Dhow (Proche) + /**-aan/** → Dhow**ów** (s'approcher)

Kulúl (Chaud) + /**-aan**/ → kulul**ów** (Devenir chaud, chauffer)

Yár (Petit) + /**-aan**/ → yar**ów** (Devenir petit)

Ce suffixe a deux allomorphes qui sont surtout utilisés pendant la conjugaison des verbes inchoatifs.

Ainsi, /***-ow***/ devient /***-oob***/ lorsqu'on ajoute au verbe un morphème flexionnel de temps commençant par une voyelle, mais reste le même devant un morphème flexionnel commençant par une consonne ou vide (**ø**).

Wúu [***Wáa + uu***] (Il) baraf-oob-ay [***báraf + ow + ay***] (a gelé). Il a gelé

Wáy [***Wáa + ay***] (Elle) baraf-ow-day [***báraf + ow + t-ay***] (a gelée). Elle a gelé.

Le suffixe verbal /**-ow**/ subit lui aussi des changements lorsqu'il s'affixe avec des adjectifs. Ainsi, si l'adjectif de base se termine par la consonne /g/, on remplace cette consonne par /k/.

Il y a une assimilation qui se réalise lorsqu'on ajoute le suffixe verbal /-ow/ à un adjectif finissant par /k/.

Suffixe verbal /**-ow**/ avec des adjectifs de base ou dérivés.

adag (dur|solide) + /***-ow*** → adkow

gaabán (court) + /**-ow**/ → gaabn**ów** (Devenir dur/solide)

sahlán (Facile, simple) + /**-ow/** → Sahlan**ów** (devenir facile/simple)

qaalisán (Cher) + /**-ow**/ → qaalisan**ów** (Devenir cher)

wanaagsán (Bien, bon) + /**-ow**/ → Wanaags**ów** (Devenir quelqu'un de bien ou bon)

Dans la conjugaison des verbes inchoatifs, le suffixe /**-aw**/ est seulement utilisé pour la forme impérative, le suffixe /**-aan**/ pour la forme infinitive des verbes ou la forme de citation, les conjugaisons avec auxiliaire et le progressif, tandis que le suffixe /–**aad/** est utilisé dans les autres cas.

adkaw [**adag+ow]** Deviens solide|dur

waan [***Waa + aan***] Je adkaanayaa [***adkaan+ayaa***] suis en train de devenir solide|dur

waad [***Wáa+ aad***] Tu wanaagsanaatay [***wanaagsanaad + tay***] es devenu bien.

way [***Waa + ay***] Elle adkaatay [***adkaad +tay***] est devenue solide.

c. ***Les verbes médio-passifs***

Les verbes médio-passifs se forment à l'aide de l'affixe medio-passif /**-at**/ à partir d'un verbe autre qu'un verbe de base, c'est-à-dire lui-même dérivé. Certains considèrent ces verbes comme des verbes autobénéfactif comme (B. W.Andrzejewski, 1974 et Puglielli, 1984). Généralement on utilise ces verbes pour dire à quelqu'un de faire un acte pour son propre bénéfice d'où l'autobénéfactif. Dans la conjugaison, l'affixe /**-at**/ devient /**-ad**/ devant un verbe finissant par une voyelle. Ils dérivent souvent des verbes causatifs.

diiri « Chauffe » → diirsó « Chauffe pour toi-même »

Kàri « cuit » → karsó « Cuit pour toi-même »

d. ***Les verbes d'expérience***

Les verbes d'expérience sont des verbes créés à partir de la suffixation d'un nom féminin avec le morphème /**-ood**/. Dans la conjugaison, le suffixe /***-ood***/ devient /***-oon***/ pour la forme infinitive et pour les temps continus (passé, présent). La forme réduite de cet affixe /***-oo***/ sert parfois comme morphème flexionnel du mode impératif. Ce sont des verbes statiques et c'est pour cela qu'on les conjugue souvent dans les temps continus. Les noms de base de ces verbes ont un ton haut pénultième et il y a une alternation tonale après la dérivation du verbe, car l'accent tonal se déplace sur la more.

cádho (N F) « colère » → cadhóod « le fait d'être fâchée »

dháxan (N F) « froid, humidité » → dhaxmóod « Le fait d'avoir froid »

gájoo (N F) « une faim » → gajóod « Le fait d'avoir faim »

e. ***Les verbes factitifs***

Les verbes factitifs expriment l'idée de « faire faire l'action » et ils sont dérivés par suffixation à partir d' un verbe ou d'un adjectif à l'aide du morphème factitif /-ays/.

La forme infinitive de ces verbes contient le morphème /**-ee**/ et leurs formes impératives le morphème /**-ayn|eyn**/ et enfin le morphème /**-ays**/ pour le reste.

Les verbes factitifs dérivés à partir d'un nom sont des verbes transitifs. Ils font partie du troisième groupe de verbes du somali.

bír (N M) « fer » → birée «Le fait de faire en fer »

Dhár (N F) → Dharée « Le fait de faire habiller »

Lorsqu'ils sont dérivés à partir d'un adjectif, ils signifient l'idée de réaliser la qualité ou l'attribut exprimé par l'adjectif. Le morphème /**-ee**/ permet de créer les verbes factitifs par suffixation.

Ils appartiennent également au troisième groupe des verbes de conjugaison.

adág (Adj) « dur, solide »→ adkée « L'acte de rendre dur ou solide quelque chose »

Cád (Adj) « Blanc » → caddée « L'acte de rendre blanc ou propre quelque chose »

Les verbes factitifs dérivés à partir d'un adjectif d'attribut de lieu deviennent des verbes intransitifs et ils expriment l'idée d'être dans une position donnée dans l'espace.

Dambé (Adj) « derrière » → Dambée « le fait d'être ou se mettre derrière »

Saré (Adj) « au-dessus » → Sarée « le fait d'être ou se mettre au-dessus ».

Ils se conjuguent de la même façon que les verbes du 3ème groupe.

VI.3.4 La syntaxe du somali

Dans cette section nous allons brièvement décrire la structure syntaxique des phrases simples du somali.

En effet le regroupement des unités linguistiques appelés syntagmes ou groupe peuvent influencer la morphosyntaxe des mots d'une phrase, en fonction de la position et de la fonction qu'ils ont dans le syntagme.

Ils peuvent subir des modifications phonologiques ou morphologiques. Les deux syntagmes qui nous intéressent ici et qu'on retrouve souvent dans les phrases somalies sont : le syntagme nominal et le syntagme verbal. Décrivons-les brièvement.

VI.3.4.1 Le syntagme verbal

Le syntagme verbal simple est constitué d'un ou plusieurs substantifs et d'un verbe qui occupe la position dominante ou centrale dans le syntagme. C'est un syntagme à tête finale puisque le verbe se trouve en dernière position. Il est composé de plusieurs éléments qui sont liés au verbe sous forme de clitiques.

La structure basique d'un syntagme verbal en somali est la suivante.

[Pro Suj ProObj1 Adpo CliAdv1, CliAdv2, ProObj2 VM ou VBAUX VBINF]

- ***ProSuj*** est un pronom clitique sujet et c'est lui qui occupe la première position dans le syntagme,
- ***ProObj*** est un pronom objet clitique de série 1,
- ***Adpo*** est une apposition verbale,
- ***CliAdv1*** est un clitique adverbial souvent allatif ou ablatif,
- ***ClidAdv2*** est un adverbe nominal comme ***wada*** ou ***kala***
- ***ProObje2*** est un pronom objet de seconde série,
- ***VBM***, ***VBAUX*** ou ***VBINF*** sont respectivement un verbe principal, un verbe auxiliaire, ou un verbe infinitif.

Un syntagme verbal ne peut en général pas former tout seul une phrase complète ; il lui faut lui ajouter des marqueurs de type de phrase (***MPh***) et des thématiseurs (***MFocus***). Enfin, le pronom sujet clitique ***ProSuj*** peut jouer le rôle de sujet ou d'objet dans le syntagme verbal, et ce dernier peut contenir seulement un verbe à l'infinitif.

VI.3.4.2 Le syntagme nominal

Le syntagme nominal en somali est composé d'un substantif (nom) suivi d'autres mots dépendant de lui comme les déterminants, les propositions modificatrices et les adjectifs.

Les déterminants sont suffixés au nom tandis que les deux autres sont de dépendants qui peuvent suivre le nom en tant que groupe ou entité morphologique indépendante.

Plusieurs syntagmes nominaux peuvent se combiner pour former un syntagme complexe à l'aide de la conjonction de coordination i***yo*** avec plusieurs têtes dominantes. Les dépendants du substantif dans un syntagme nominal sont les articles définis, les démonstratifs, les interrogatifs et les possessifs.

En cas de présence de plusieurs déterminants dans un même syntagme, c'est le possessif qui occupe la première position, juste après le substantif.

Le schéma ci-dessous représente les différents constituants ou mots composant un syntagme nominal :

Nom + possessif + {article défini démonstratif interrogatif}