Boîte à outils TAL pour des langues peu informatisées : le cas du somali

Abdillahi Nimaan[1,2], Pascal Nocera[1], Juan-Manuel Torres-Moreno[1,3]

[1] Laboratoire d'Informatique d'Avignon / UAPV, BP 1228 84911 Avignon, Cedex 9, France

[2] Institut des Sciences et des Nouvelles Technologies / CERD, BP 486 Djibouti, Djibouti

[3] École Polytechnique de Montréal / Département de génie informatique, CP 6079 Succ. Centre-ville - H3C 3A7 Montréal (Québec), Canada

{nimaan.abdillahi, pascal.nocera, juan-manuel.torres}@univ-avignon.fr

Abstract

The use of corpora is a critical phase of systems of natural language processing based on statistical methods. This point is crucial for less equipped and less computerized languages like African languages. This paper aims to present different tools for Somali language and different scenarios to build corpora for this language : A gathered corpora from the web, Syntex, a system of automatic text generation and an automatic translator of Somali.

Résumé

L'utilisation de corpus constitue une phase critique des systèmes de traitement de la langue naturelle basés sur des méthodes statistiques. Ce problème devient crucial pour les langues disposant de peu de ressources électroniques et peu informatisées comme les langues africaines. Cet article décrit un ensemble d'outils pour traiter automatiquement la langue somali, et plusieurs scénarios pour constituer des corpus : collecte de corpus à partir du web, Syntex, un système de synthèse automatique de textes et un traducteur automatique en somali.

Mots-clés : Génération de texte, résumé automatique, TAL statistique, reconnaissance vocale, langue somali.

1. Introduction

Dans la plupart des pays africains, dits de tradition orale, les patrimoines culturels, scientifiques et historiques se transmettent de génération en génération. Ce savoir ancestral accumulé depuis des siècles est aujourd'hui menacé de disparition. Quelles sont les solutions pour éviter que ne sombre dans l'oubli cette importante base de connaissances qui a traversé les âges ? Les organisations internationales (Unesco, 2003) et les pays concernés[1], ont entrepris des programmes de numérisation de leurs archives audio disponibles[2]. Malheureusement, sans outil d'indexation et/ou de transcription automatique, la richesse de ces données reste difficilement accessible. Le développement d'outils de traitement automatique des langues africaines s'avère donc primordial si l'on souhaite sauvegarder et exploiter le patrimoine culturel africain. Nos recherches s'inscrivent dans cette optique. Il s'agit de concevoir des outils d'indexation et de transcription automatiques de la parole pour les langues parlées en Afrique de l'est. Le cas de la langue somali parlé à Djibouti, sera étudié. Bien entendu, le développement de tels outils nécessite des corpus audio manuellement transcrits pour les apprentissages successifs pour la modélisation acoustique et

[1] La plupart des pays africains disposent des archives audio stockées depuis 40 ans dans les stations radio locales.
[2] La république de Djibouti a entrepris un vaste programme de numérisation des archives audios (www.rtd.dj).

des corpus textuels de grande taille pour la modélisation linguistique. Ces corpus textuels sont inexistants dans la plupart de ces pays du fait, précisément, de leur tradition orale. Cet article présente le developpement d'une boîte à outils TAL (Traitement Automatique de la Langue) pour des langues peu dotées en ressources informatiques. En section 2 nous présentons les langues djiboutiennes. En section 3 nous présentons des différents scénarios pour pallier le manque de corpus : la constitution de corpus à partir du web, la synthèse automatique de texte et la traduction automatique. La section 4 présente des outils (racinisation, conjugaison, phonétiseurs,...) et une application au résumé automatique. Des conclusions et perspectives sont présentées en section 5.

2. Présentation sommaire des langues djiboutiennes

La république de Djibouti se situe à l'est du continent africain, à l'entrée de la mer rouge et appartient à la Corne de l'Afrique. Elle est bordée par l'Érythrée au nord, l'Éthiopie à l'ouest et la Somalie au sud. Quatre langues sont parlées à Djibouti, dont deux officielles (français et arabe) et deux autochtones (somali et afar). Les langues afar et somali, sont non seulement parlées à Djibouti mais également dans plusieurs autres pays de la région : l'Éthiopie, l'Érythrée et la Somalie. Ces deux langues appartiennent à la même famille linguistique afro-asiatique, sous-branche couchitique-est (SIL, 2004). Toutes les deux sont écrites en caractères latins. Nous nous intéresserons par la suite à la langue somali. L'alphabet somali se compose de 22 consonnes et de 10 voyelles (5 longues et 5 courtes). La structure phonétique (Saeed, 1999) est présentée dans le tableau 1. Les différentes variantes de cette langue sont le somali-somali (souvent appelé somali), le dabarre, le garre, le jiiddu, le maay et le tunni. Les composantes somali-somali (80%) et somali-maay (17.8%) sont les plus répandues. Nos travaux actuels portent sur la variante somali-somali (parlé par une population d'environ 11 millions de personnes[3]), qui est aussi celle parlée à Djibouti. C'est également une langue tonale avec deux à trois tons lexicaux (Le-Gac, 2001 ; Hyman, 1981). La transcription actuelle ne prend pas en compte ces tons. D'où l'existence de quelques homographes-hétérophones. Exemples : *beer* (le foie) et beer (le champ) ; *dameer* (l'âne) et *dameer* (l'ânesse) ; *gool* (dromadaire) et *gool* (lionne). Une autre particularité de cette langue est la composition en racines (sous-mots) de la plupart de ses mots (Bendjaballah, 1998). Ces racines sont des unités monosylabiques le plus souvent de formes : CVC, CV, VC (C=Consonne, V=Voyelle).

	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 1 : structure phonétique des consonnes de la langue somali.

[3]www.ethnologue.com

Par exemple, les mots *birlab* (aimant) ou *shinbir* (oiseau) sont tous les deux composés de deux radicaux de type CVC. Plusieurs dictionnaires existent, notamment des dictionnaires somali-français et français-somali[4], somali-italien, somali-anglais, etc. Depuis 2002, sont réalisés à Djibouti, d'importants efforts de promotion, de recherches linguistiques[5], de standardisation (Erey-bixin, 2003), et surtout la création d'un dictionnaire illustré (Carab, 2004) somali-somali de 40 000 entrées.

3. Constitution de corpus

À partir du web. L'utilisation du web pour constituer des corpus a été étudiée par de nombreuses équipes (Ghani *et al.*, 2000 ; Vaufreydaz *et al.*, 1999). Pour notre part, nous avons constitué automatiquement un corpus de ≈ 3 millions de mots bruts à partir de l'aspiration de sites Internet (sites de journaux, associations culturelles, etc.). Le tableau 2 montre la distribution du corpus constitué. Ce corpus est composé de 121K de mots différents (avec les formes fléchies), de 4.4K de racines différentes et de 36 phonèmes différents. Ce corpus nous servira de référence. Le programme de racinisation développé à cet effet est présenté à la section 4. La liste des phonèmes est la suivante : *null, pause, a, aa, b, c, d, dh, e, ee, f, g, h, i, ii, j, k, kh, l, m, n, o, oo, q, r, s, sh, t, u, uu, w, x, y, '*, (les phonèmes *t* et *k* sont décomposés en leurs parties occlusives et explosives). La figure 1a montre une étude de la distribution de Zipf des mots du corpus de référence somali. Le corpus Hansard (français-anglais) des débats du parlement Canadien, a été aussi étudié pour comparaison. La constante *K* (fréquence ∗ rang = *K*) du somali est plus faible que celle du français ou de l'anglais, donc en principe plus facile à traiter informatiquement. La figure 1b montre l'importance des mots-type dans la langue somali par rapport au total de mots bruts. Ce nombre est bien plus important qu'en anglais ou français. La complexité a été calculée comme nombre mots / nombre mots-type.

Phrases	Mots	Mots distincts	Racines	Racines distinctes	Phonèmes	Phonèmes distincts
84.7k	2.82M	121k	6M	4.4k	14M	36

Tableau 2 : Caractéristiques du corpus de texte somali issu du web.

Nous avons appris un modèle de langage trigramme sur ce corpus, pour des tests de reconnaissance automatique de la parole (RAP). Un lexique des 20K mots les plus fréquents en a été extrait et a été phonetisé avec l'outil Somphon[6]. Nous avons utilisé les outils de Carnegie Mellon University (Rosenfeld, 1995) pour l'apprentissage du modèle de langage. Ce modèle est composé de 726K bigrammes et de 1.75M trigrammes. La perplexité $PP = 2^H$, $H = -Pn \log p(x)/n$ du corpus de test est *52.16* avec un taux de mots hors vocabulaire de 6.74%. Les premiers résultats de RAP avec le moteur Speeral du LIA (Nocera *et al.*, 2002) ont donné des résultats encourageants de 20% de taux d'erreur mot.

[4]http://www.dictionaric.com
[5]Création d'un institut des langues : http://www.cerd.dj
[6]Outil de phonétisation des textes en somali développé au LIA.

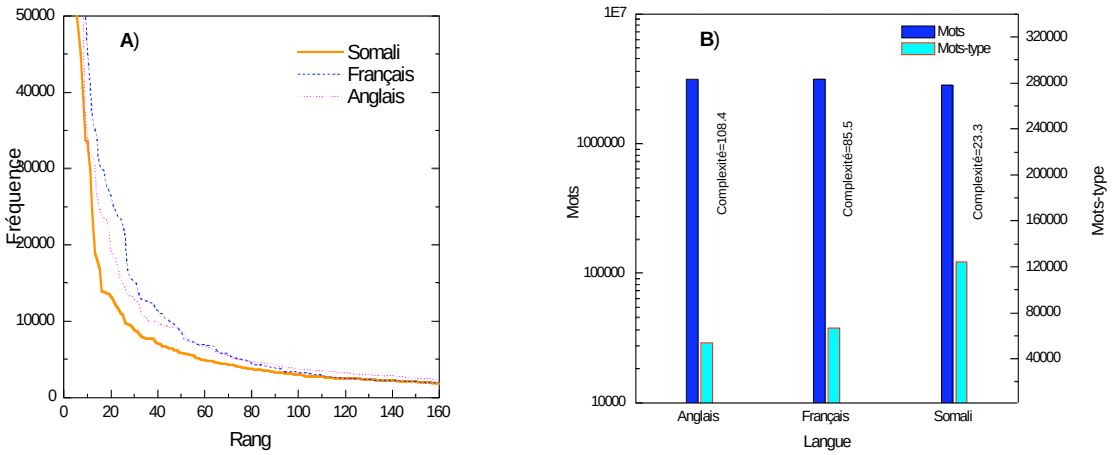

Figure 1. Corpus constitués à partir du web.

Synthèse de texte. À partir d'un corpus amorce de petite taille, nous avons généré un texte aléatoire afin de constituer un corpus d'une taille plus importante. Nous pensons que l'utilisation de *n*-grammes permet de générer un texte avec les propriétés de la langue d'origine. Il ne s'agit pas de générer un texte avec une sémantique, mais uniquement un texte aléatoire lexicalement correct. Pour la RAP il est impératif de générer, en particulier, des nouveaux bigrammes existants dans la langue. Nous avons construit Syntex, un générateur de texte de type Markovien. Nous allons expliquer l'algorithme avec des 3-grammes. À partir d'un unigramme du corpus amorce, nous retrouvons ses probables bigrammes. Nous gardons un bigramme (au hasard) et nous retrouvons leurs probables trigrammes. Nous prenons ensuite le dernier mot du trigramme comme point de départ et nous itérons le processus jusqu'à obtenir le nombre de paragraphes désiré. Il faut contrôler le nombre moyen de mots par phrase. Ce nombre a été fixé à 35 dans cette expérience pour être proche du nombre moyen de mots par phrase (33) du corpus de référence. En pratique, ce processus peut être itéré en utilisant des valeurs de *n* allant de 3 à 10. Selon la taille du corpus de départ de *Pi* phrases, nous avons généré des corpus synthétiques de taille *Ps* phrases. Nous avons utilisé deux tailles d'amorces : un petit corpus avec *Pi* = 15 et un autre avec *Pi = 1000* phrases, nous avons ainsi généré des corpus synthétiques (1 et 2) d'environ 3 millions de mots. Nous avons calculé la perplexité des corpus de référence et synthétiques. Ainsi, nous avons une perplexité du corpus somali issu du web *PP = 52.16* et une perplexité du corpus synthétique 1 de *PP = 46.7* et du corpus synthétique 2, *PP = 54.7*. Sur la figure 3a nous montrons également les courbes de Zipf pour le texte synthétique. On voit que le corpus généré avec un faible nombre de phrases de départ se place plus loin de la courbe du corpus de référence. Par contre, la courbe de 1000 phrases de départ colle assez bien à l'allure de la courbe du corpus de référence issu de web. Sur la figure 3b nous montrons l'erreur quadratique moyenne entre le corpus somali de référence et les corpus synthétiques : on voit que l'erreur est inférieure pour le corpus avec *Pi* plus grand. Ceci confirme notre hypothèse que des corpus avec des propriétés assez proches de celles du corpus de référence peuvent être générés automatiquement. Bien entendu ces propriétés ne suffisent pas pour la RAP et les premiers tests avec des modèles de langage appris sur ces textes synthétiques ne sont pas concluants. Ces résultats étaient tout à fait prévisibles, compte tenu de l'absence de nouveaux bigrammes ou de nouvelles informations dans le texte généré. Syntex a généré un texte aléatoire avec des

nouvelles phrases, mais il n'a pas généré de nouveaux mots ni de nouveaux *n*-grammes. Pour donner plus de richesse à la synthèse, nous avons envisagé trois stratégies : l'introduction aléatoire de nouveaux mots à partir d'un lexique ; la génération des nouvelles formes fléchies avec un conjugateur de verbes (cf. section 4) et un mélange des deux (voir figure 2). Des tests sont en cours.

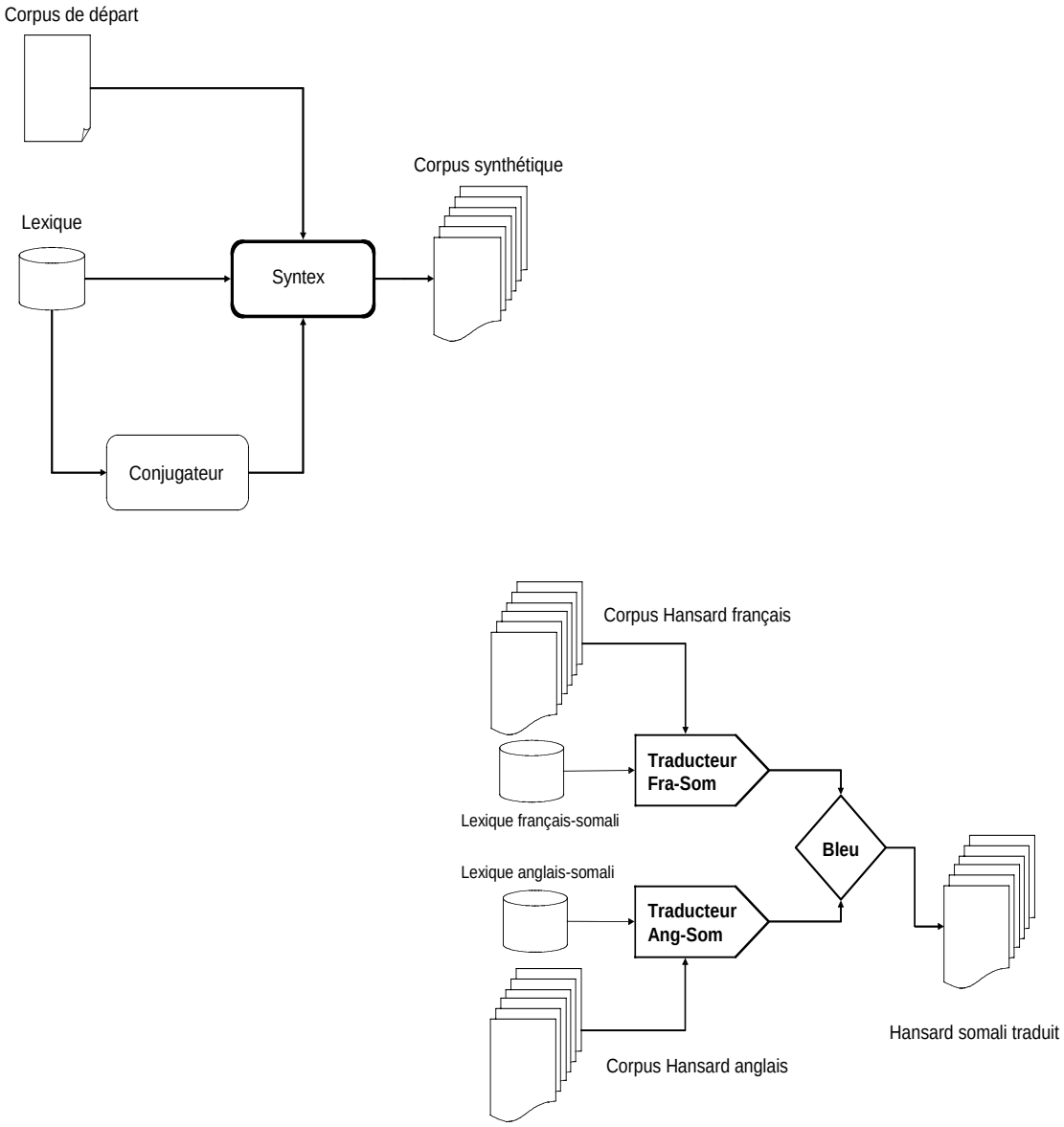

Figure 2. Systèmes pour la génération de corpus synthétiques et la traduction automatique.

Corpus issus de la traduction automatique. Nous avons effectué la traduction directe et alignée. En traduction directe, nous avons généré une traduction simple français-somali à l'aide d'un modèle unigramme et d'un lexique de base. Nous avons lemmatisé le corpus Hansard français à la volée afin d'avoir des formes normalisées des mots. Ceci donne lieu à une traduction somali lemmatisée qui est cependant moins gênante qu'une traduction française lemmatisée, car un traitement de surface permet de retrouver une traduction acceptable. Par exemple la phrase : « j'avais six ans » serait lemmatisée comme « je avoir six

an » puis traduite comme « aan wax lix sanad », et le traitement de surface donnerait « waxaan lix sanadood », qui est plus proche de la véritable traduction « waxaan jiray lix sanadood ». Pour la traduction alignée, nous avons utilisé le corpus Hansard bilingue. Nous avons réalisé une traduction directe français-somali et une autre anglais-somali. Nous avons gardé la meilleure traduction en terme de score Bleu, mesure qui comptabilise de façon pondérée le nombre de *n*-grammes qu'une traduction automatique partage avec une traduction de référence. Elle prend des valeurs entre [0,1], où 1 est synonyme d'une bonne traduction. (Papineni *et al.*, 2002) rapporte que cette mesure utilisée dans les évaluations NIST (cf. www.nist.gov), est corrélée à des jugements produits par des humains. Le système de traduction automatique est montré sur la figure 2, à droite.

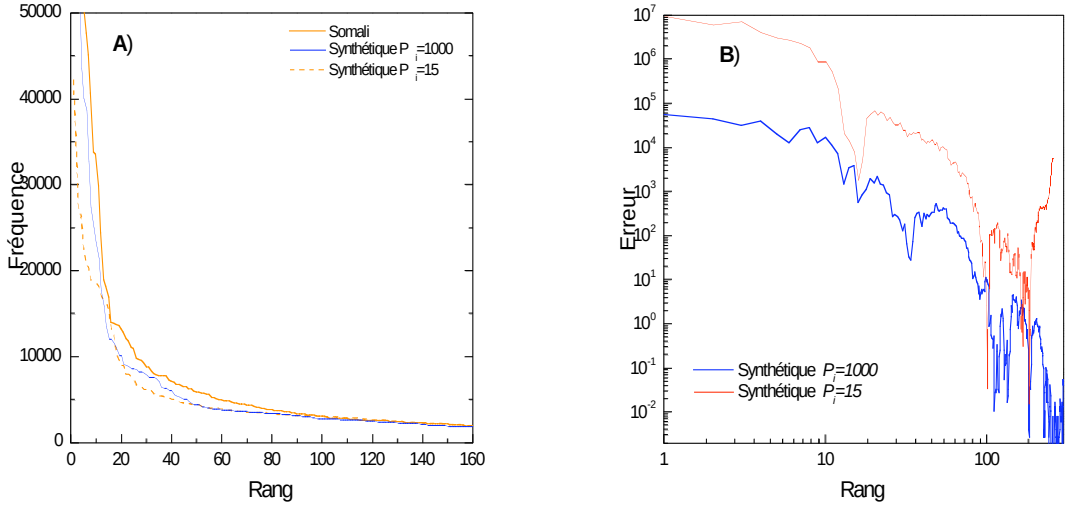

Figure 3. A) Loi de Zipf du corpus du web (gros trait) et corpus synthétiques, Pi=15 (pointillé) et Pi=1000 (trait continu fin) phrases de départ. B) Erreur quadratique moyenne entre le corpus somali issu du web et ceux synthétiques.

Nous montrons ci-bas, un exemple de traduction automatique d'un fragment du Petit Prince, de A. de Saint-Exupéry :

Le petit prince. Lorsque j'avais six ans j'ai vu, une fois, une magnifique image, dans un livre sur la Forêt Vierge qui s'appelait « Histoires Vécues ». Ça représentait un serpent boa qui avalait un fauve. Voilà la copie du dessin. On disait dans le livre : « Les serpents boas avalent leur proie tout entière, sans la mâcher. Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion ». J'ai alors beaucoup réfléchi sur les aventures de la jungle et, à mon tour, j'ai réussi, avec un crayon de couleur, à tracer mon premier dessin.

Yarkii ina boqor. *Markii waxaan lix sanadood waxaan arkid, kow jeer, kow qurux badan sawir, ku kow buug dul saaran ka kayn hawdka yaad is wicid sheeko noolow. Taasina matalaad kow mas jabiso yaad liqid kow bahal. Wakaa ka nuqlid ee ka sawirkii. Anaga dhehitaan ku ka buug ka mas jabiso liqid kiisa ugaadh gabigeeda dhan, la aan ka calajin. Marka xiga aan ma awood is jeclaysii silig iyo aan seexasho inta ka lix bil ee kiisa dheefshiid. Aan leeyahay markaas in badan fikir dul saaran ka mu*

aamaraad ee ka howd iyo, u kay dalxiis, waxaan guuleyso, maca kow laabbis ee midab, u jeex kay kowaad sawirkii. kay sawirkii kowaad.

4. Outils pour le TAL

Lemmatisation. La lemmatisation consiste à trouver le lemme des formes fléchies et à ramener les mots pluriels et/ou féminins au masculin singulier, ainsi, *chanter*, *chantaient* et *chanteur* sont ramenés à la même forme *chant.* Ce processus permet d'amoindrir la difficulté dimensionnelle qui pose de très sérieux problèmes lors de la représentation vectorielle de textes. Ainsi, la lemmatisation permet de diminuer le nombre de termes ce qui est essentiel pour des algorithmes comme les systèmes de résumé automatique. Nous nous sommes inspiré de l'algorithme de (Porter, 1980) pour développer un algorithme de lemmatisation du somali.

Racinisation. Elle consiste à décomposer en racines-syllabes de formes CVC, CV ou VC. La langue somali se prête assez aisément (Bendjaballah, 1998) à cet exercice et nous avons pu mettre en place un algorithme d'extraction de ces sous-mots. Il s'agit de parcourir le texte de gauche à droite et d'extraire les racines dans l'ordre CVC, CV, VC et V.

Conjugateur. Il réalise une tâche opposée à celle de la lemmatisation : à partir d'un lemme ou d'un verbe à l'infinitif, il fait des expansions (à base de règles) afin d'obtenir les formes fléchies selon un temps et une personne. Encore une fois, le somali étant une langue très régulière concernant les verbes, se prête bien à cette tâche (Guure, 1999 ; Carab, 2004). Le temps futur est généré à partir du verbe + *doon.* Des tests avec Syntex et la mise en place d'un lexique enrichi sont actuellement en cours.

Transducteurs : chiffres, téléphones, dates et abréviations. Plusieurs transducteurs ont été développés afin de normaliser les chiffres et les abréviations des différents corpus. Pour les chiffres, un filtre détermine leur classe : numéro de téléphone, date, heure ou valeur numérique. Les exemples ci-dessous donnent une illustration des transformations effectuées :

- 00-253-343098 (tél.) → *eber eber laba shan sadex sadex afar sadex eber sagaal sideed*
- 14/10/2005 (date) → *afar iyo tobankii bishii tobnaad laba kun iyo shan*
- 14 :30 (heure) → *labadii iyo badhkii*
- 25% (pourcentage) → *boqolkiiba shan iyo labaatan*
- 452548 (nombre) → *afar boqol laba iyo konton kun shan boqol sideed iyo afartan*

Les abréviations tels que *Md* ou *Mud* (*Mudane* / Monsieur), *M* (*Marwo* / Madame), Dr (*Doktor* / Docteur), *Col* (Colonel), *Muj* (*Mujaahid* / combattant), *C/rasaaq* (*Cabdulrasaaq*), *Sh* (*Sheekh* ou *Shilin* / Cheick ou Shilling) ont été également prises en compte.

Phonétiseur Somphon. Conformément au phonétiseur Lia_Phone développé au LIA (Bechet, 2001), notre phonétiseur somali prend en entrée un texte ou un lexique et il génère la représentation phonétique correspondante. Le somali dispose de peu de lettres qui ne se prononcent pas et les liaisons n'existent pas, ce qui évite des nombreuses ambiguïtés. Ceci est un véritable atout pour sa phonétisation. Le codage des 36 phonèmes est effectué avec 2 caractères. Par exemple le phonème « b » est codé par bb, et le « oo » (o long) est codé par oo, tandis que le phonème « o » (o court) est codé par oh. Les caractères transcrits avec deux lettres sont codés avec ces deux caractères (« dh » → dh ; « kh » → kh, etc.), par exemple : (voiture) *baabuur* → bb aa bb uu rr ; (Djibouti) *jibuuti* → jj ih bb uu tt ih ; (Mexico) *Meksiko* → mm eh Okk Bkk ss ii Okk Bkk oh.

Résumé automatique par *Cortex*. Développé au LIA, Cortex génère des résumés de texte par extraction de phrases (Torres-Moreno *et al.*, 2002), méthode qui s'avère performante dans cette tâche (Saggion and Lapalme, 2000 ; Morris *et al.*, 1999). Il est basé sur des algorithmes numériques et reste très indépendant de la langue. Nous avons utilisé nos algorithmes de racinisation et transduction couplés avec le système Cortex. Nous présentons en exemple, le résumé du chapitre I du Petit Prince, effectué par un enseignant de français, ainsi que les résumés automatiques en français et somali générés par Cortex. On peut remarquer la qualité des phrases retenues.

Résumé humain : Un garçon, à l'age de six ans, lit le livre « Histoires Vécues » qui a dit « les serpents avalent leur proie tout entière, sans la mâcher ». Il a fait le dessin d'un serpent qui a avalé un éléphant, mais il ressemble un chapeau. Il a dessiné l'intérieur du serpent boa. Il avait été découragé par l'insuccès de ses dessins. Les grandes personnes ne comprennent jamais rien toutes seules. Il a choisi un métier comme pilote des avions. Il ne parlait ni de serpents boas, ni de forets vierges, ni d'étoiles aux grandes personnes.

Résumé automatique : Lorsque j'avais six ans j'ai vu une image dans un livre sur la Forêt Vierge s'appelant Histoires Vécues. Ça représentait un serpent boa avalant un fauve. J'ai réfléchi sur les aventures de la jungle et j'ai réussi avec un crayon de couleur à tracer mon premier dessin. J'ai montré mon chef-d'oeuvre aux adultes et je leur ai demandé si mon dessin leur faisait peur. Mon dessin ne représentait pas un chapeau. J'ai dessiné l'intérieur du serpent boa afin que les adultes puissent comprendre. Les adultes m'ont conseillé d'abandonner les dessins de serpents boas ouverts ou fermés et de m'intéresser à la géographie à l'histoire au calcul et à la grammaire. J'ai abandonné à l'âge de six ans une carrière de peinture. Et je ne lui parlais ni de serpents boas ni de forêts vierges ni d'étoiles.

Résumé automatique (somali) : *Markii aan jiray lix sanadood waxaan arkay maalin sawir qurux badan oo ku yaala buug ka hadlaya kaynta hawdka oo magaciisu ahaa Taariikh Nololeed. Markaan sawirkaa arkay aad ayaan uga fikiray arimaha ka dhaca kaynta isla markiina waxaan ku guuleystay aniga oo isticmaalay qalin nashqad leh inaan soo saaro sawirkaygii ugu horeeyay. Waxaan tusay farshaxankaygii dadkii iga weyna oo waxaan weydiiyay inay ka cabsanayaan sawirkayga. Waxaan sawirkii aad ugu muujiyay gudaha jabisada oo furan si dadka waaweyni u fahmaan. Waxay markaa dadkii waaweynaayi igu dardaareen inaan iskaga hadho sawiradan jabisooyinka furan ama xidhan oo waxay igula taliyeen inaan isku taxaluujiyo jiqoraafiga taariikhda xisaabta iyo naxwaha. Dadka waaweyni waxba ma fahmaan keligood arintaasina way daalisa ubadka oo markasta macno bixiya. Markaan la kulmo qof weyn oo garaad leh waxaan ku tijaabin jiray sawirkaygii kowaad oo aan haystay. Aniguna Waan iska dhaafi jiray oo kama aan hadli jirin jabisooyinka iyo kaynta hawdka iyo xidigaha.*

5. Conclusion, discussion et perspectives

Nous avons développé plusieurs outils de traitement automatique pour la langue somali parlée à Djibouti et en Afrique de l'est. La plupart de ces outils (transducteurs, lemmatiseur, raciniseur, conjugateur et phonétiseur) sont conçus pour répondre à l'objectif premier de nos travaux : la sauvegarde du patrimoine culturel oral par le développement d'outils informatiques d'indexation et de transcription automatique. Certains outils peuvent être utilisés pour développer d'autres systèmes comme ceux du *text-to-speech*, qui permettraient une meilleure utilisation des nouvelles technologies par des populations analphabètes. Nous

avons aussi développé Syntex, un générateur markovien de textes aléatoire performant. Il permet de générer de vastes corpus, indépendamment de la langue. L'introduction de nouveaux mots se fait au moyen d'un conjugateur de verbes. La tournure à la voix passive et un lexique que l'on pourra enrichir sont en cours. Nos tests ont montré que notre algorithme est capable de synthétiser des corpus avec des propriétés semblables à celles des véritables corpus. Le système de traduction automatique a fourni des résultats mitigés. Si l'on considère la taille des lexiques utilisés, le système aligné a cependant produit des traductions de meilleure qualité. Des tests avec l'utilisation des quadrigrammes les plus probables ainsi que l'introduction de nouveaux mots passant par la conjugaison de verbes (accord de la phrase) sont en cours. D'un autre côté, nous travaillons actuellement sur une indexation automatique de bases de données audio, en utilisant les sous-mots, « racines » de la langue somali, pour pallier l'insuffisance des corpus de départ. Nos outils TAL seront disponibles sur le site du LIA à l'adresse www.lia.univ-avignon.fr.

Références

Bechet F. (2001). Lia_phone : Un système complet de phonétiseur de textes. *TAL*, 2(1) : 47–67.

Bendjaballah S. (1998). La palatisation en somali. *Linguistique Africaine*, 21 : 5-52.

Carab S. H. (2004). *Ileeye 2004. Qaamuus. Ereykoobe*. Machadka Afafka ee Jabuuti.

Erey-bixin (2003). *Ereyyada wararka ee war-baahinta*. Machadka Afafka ee Jabuuti.

Ghani R., Jones R., and Mladenic D. (2000). *Mining the web to create minority language corpora*. Berlin.

Guure F. C. (1999). *Dictionnaire somali-français*. L'Harmattan, France.

Hyman L. (1981). Tonal accent in somali. *Studies in African linguistics*, (12) : 169–203.

Le-Gac D. (2001). *Structure prosodique de la focalisation : cas du somali et du français*.

Morris A., Kasper G., and Adams D. (1999). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1) : 17–35.

Nocera P., Linares G., Massonie D., and Lefort L. (2002). *Phoneme lattice based A* search algorithm for speech recognition*, TSD.

Papineni K., Roukos S.,Ward T., and ZhuW. (2002). Bleu : a method for automatic evaluation of machine translation. In *40th Annual Meeting of the ACL*, Philadelphia, Pennsylvania, USA : 311–318

Porter M. (1980). An algorithm for suffix stripping. *Program*, 3(14) : 130–137.

Rosenfeld R. (1995). The CMU statistical language modeling toolkit, and its use. In *ARPA Spoken Language Technology Workshop*, Austin, Texas, USA.

Saeed J. (1999). *Somali (London Oriental and African Language 10)*. Johns Benjamins Publishing Co., Amsterdam/Philadelphia.

Saggion H. and Lapalme G. (2000). Concept identification and presentation in the context of technical text summarization. In *Automatic Summarization Workshop*, pp 1–10, ANLP/NAACL, Seattle.

SIL I (2004). *Ethnologue : Language of the World*. 14th edition, USA.

Torres-Moreno J.-M., Velazquez-Morales P., et Meunier J.-G. (2002). Condensés de textes par des méthodes numériques. In *JADT*, IRISA/INRIA France : 723–734.

Unesco (2003). Convention pour la sauvegarde du patrimoine culturel immatériel. *www.unesco.org*.

Vaufreydaz D., Akbar M., and Roullard J. (1999). Asru'99. In *Internet documents : a rich source for spoken language modelling*, pp 177–280, Keystone, Colorado, USA.