Construction d'un corpus parallèle français-comorien en utilisant de la TA français-swahili

Moneim Abdourahamane[1,2] Christian Boitet[1,2] Valérie Bellynck[1,3] Lingxiao Wang[1,2]
Hervé Blanchon[1,2]
(1) LIG, campus, 38041 Grenoble cedex 9, France
(2) UGA, adresse, 38401 Saint Martin-d'Hères, France
(3) G-INP, 47 av. Félix Viallet, 38000 Grenoble, France
prenom.nom@imag.fr

RESUME

Le comorien ou shikomori est une *macro-langue* constituée de 4 dialectes (ngazidja, maore, mwali, ndzuani) très proches entre eux, et assez proches du swahili. Il est très peu doté au niveau des ressources linguistiques informatisées, n'ayant ni corpus, ni dictionnaires, ni outils de correction ou de traduction. Il n'est donc *a priori* pas possible de construire efficacement un corpus parallèle, comme on sait le faire par TA suivie de post-édition (PE) interactive : en français-chinois, 17 mn/page avec Google Translate (GT), 12 mn/page avec le système de TA MosesLIG.fr-zh en utilisant SECTra/iMAG. Nous sommes cependant en train d'y arriver en post-éditant des pré-traductions en swahili produites par GT. Le swahili est utilisé ici non comme langue pivot, mais comme *langue cible auxiliaire*. Nous avons maintenant un corpus de bonne qualité français-ngazidja formé de 34 articles du journal Alwatwan (899 segments, 16224 mots, 65 pages standard). Nous extrayons en parallèle des correspondances lexicales bilingues. La première application sera la lecture active du français pour des locuteurs du comorien ; elle utilisera le dictionnaire et le système de TA dérivés respectivement de la base lexicale et du corpus, en croissance.

ABSTRACT

Building a French-Comorian parallel corpus using French-Swahili MT

Comorian or shikomori is a *macro-language* made of 4 dialects very near one to another (ngazidja, maore, mweli, ndzuani), and quite near to swahili. It is quite under-resourced as far as computerized linguistic resources are concerned, having neither corpora nor dictionaries nor correction or machine translation (MT) tools. It is hence *a priori* not possible to build efficiently a parallel corpus, as we know how to build one using MT followed by online post-editing (PE): for French-Chinese, 17 mn/page with Google Translate (GT), 12 mn/page with the MosesLIG.fr-zh MT system and SECTra/ iMAG. We are however on the way to achieve it by post-editing swahili "pre-translations" produced by GT. Swahili is used here not as a pivot language, but as an *auxiliary target language*. We have now a good quality French-Ngazidja corpus containing 14 articles of the Alwatwan newspaper (899 segments, 16224 words, 65 standard pages). We extract in parallel bilingual lexical correspondences. The first application will be the active reading of French for Comorian speakers; it will use the dictionary and the MT system respectively derived from the lexical database and the growing bilingual corpus.

MOTS-CLES : construction de corpus parallèle, français-comorien, swahili, langue auxiliaire

KEYWORDS: parallel corpus building, French-Comorian, Swahili, auxiliary language

Introduction

Nous nous plaçons dans le cadre des recherches sur l'informatisation de langues et de groupes de langues peu dotées (Berment V. 2004), en visant particulièrement les langues africaines et les langues de la francophonie. L'accroissement du nombre des locuteurs, et de l'usage prévisible des STIC dans ces langues (plus de 300 à informatiser d'ici 30 ans) justifie pleinement la recherche de méthodes efficaces pour construire les ressources puis les outils de TAL nécessaires. Dans le travail décrit ici, il s'agit du comorien ou *shikomori,* qui est en fait une macro-langue[1] regroupant quatre dialectes parlés aux Comores : le *ngazidja*[2] ou grand-comorien (Grande Comore), le *maore* ou mahorais (Mayotte), le *mwali* ou mohélien (Mohéli), et le *ndzuani* ou anjouanais (Anjouan). Or, ce qui est nécessaire la plupart du temps, ce sont des systèmes de bonne qualité. Comme les systèmes de TAL très généralistes (très "couvrants") ne sont pas et ne peuvent pas être de bonne qualité[3], il faut plutôt chercher à construire des systèmes spécialisés à des sous-langages[4], de façon générique et efficace. Avec le Web 2.0, c'est devenu possible, grâce à l'émergence du *développement contributif, participatif ou collaboratif en ligne*.

Dans beaucoup de cas, le couple de langues français-langueX est très peu doté, mais il existe déjà des ressources pour le couple anglais-langueX, et quelques ressources monolingues informatisées et accessibles librement, au moins pour la recherche. C'est par exemple le cas du somali, sur lequel travaille un autre doctorant de notre équipe : même s'il n'y a pas de dictionnaires informatisés accessibles et utilisables pour construire des applications de TAL (ayant quelques milliers d'entrées et les informations morphosyntaxiques de base), il existe des corpus pour le somali[5], et Google Translate (GT) offre un système direct de traduction automatique (TA) anglais-somali, et français-somali par composition (fr-en+en-so). Il est alors possible de développer efficacement un corpus français-somali par post-édition (PE) de résultats de TA spécialisé à un sous-langage (celui des articles d'un quotidien publié en français), et de l'utiliser pour construire un système de TA direct et spécialisé de type Moses à constituants *(phrase-based)*. De plus, il existe un analyseur morphologique en ligne développé par un chercheur finlandais, ce qui permet d'envisager ensuite une version à facteurs *(factor-based)*.

Il n'existe rien de ce type pour aucun des dialectes comoriens. Pour équiper le comorien ou même une seule de ses variantes, on ne peut donc pas partir de ressources existantes, de façon opportuniste, et commencer par développer les applications utilisant ces ressources. Il faut d'abord choisir une ou plusieurs applications à réaliser, en fonction de leur utilité potentielle et de leur criticité, et développer les ressources nécessaires. Il semble que, dans le cas des Comores, il faille viser d'abord non pas des systèmes de TA classiques, ou des correcteurs d'orthographe et de grammaire pour le comorien, mais plutôt un outil de *lecture active,* aidant à la compréhension du français, langue d'enseignement et d'administration, dans laquelle tout est publié, alors que rien n'est publié en comorien, mais que lui seul est parlé dans la vie courante. Pour cela, on peut se contenter de TA plus ou moins littérale ou à

---

1 Parmi les exemples connus de macro-langue, on peut citer le yougoslave (croate, serbe, bosniaque et monténégrin), l'hindoustani (hindi et ourdou), et pourquoi pas l'anglais (britannique, américain, australien, et indien). Voir https://fr.wikipedia.org/wiki/Macro-langue.

2 On dit aussi *shingazidja* (code `zdj`), mais le préfixe *shi* ou *ki* signifie simplement "langue" dans les langues bantoues. On dit en français plutôt *swahili* que *kiswahili,* mais *shikomori* (langue des îles) et pas *komori*.

3 Pour la TA, par exemple, on a le « métathéorème » CxAxQ : le produit de la couverture, de l'automaticité et de la qualité (par rapport à la qualité professionnelle) est nécessairement très inférieur à 100%, mais on peut atteindre presque 100% pour deux de ces critères en acceptant que le troisième soit petit. Par exemple, pour un sous-langage représentant peut-être une couverture de 5% ou 10%, comme celui des bulletins météo ou des brèves de sites boursiers, on peut obtenir A=100% et Q=90% ou 95% (18 ou 19/20).

4 Au sens de (Kittredge 1982).

5 www.tshwanedje.com/publications/11ParCor.pdf: *Managing Eleven Parallel Corpora and the Extraction of of data in all official South African languages*. Voir aussi Anneleen Van der Veken et Gilles-Maurice de Schryver (2003. *Les langues africaines sur la Toile. Étude des cas haoussa, somali, lingala et isixhosa*.

constituants, ce qui nécessite le développement d'un corpus français-comorien, d'un dictionnaire français-comorien, et aussi d'un générateur morphologique du comorien.

En l'absence de système de TA, et de tout corpus, bilingue ou monolingue cible, il n'est pas *a priori* pas possible de construire efficacement un corpus parallèle, comme on sait le faire par TA suivie de post-édition (PE) interactive : en français-chinois, nous en sommes à 17 mn/page[6] avec Google Translate (GT), 12 mn/page avec le système de TA MosesLIG.fr-zh en utilisant SECTra/iMAG. Nous y arrivons cependant en post-éditant des pré-traductions en swahili produites par GT. Dans cette nouvelle approche, le swahili est utilisé non pas comme langue pivot, mais comme *langue cible auxiliaire*[7].

Dans la section 1, nous précisons le contexte, les motivations et les objectifs de notre recherche. Dans la section 2, nous illustrons le nouveau processus de construction d'un corpus parallèle par post-édition en ligne de résultats de TA dans une langue voisine mieux dotée, ici le swahili, sur des articles du quotidien Al-Watwan. La section 3 décrit la progression de la construction du corpus et l'évolution de la productivité (de 28 mn/page à 23 mn/page actuellement). Nous commençons à extraire en parallèle des correspondances lexicales bilingues. Enfin, nous concluons sur l'apport actuel et des perspectives.

1 Contexte et motivations

1.1 Le comorien, une macro-langue peu dotée

Les variantes ou dialectes du comorien sont utilisés pour toutes les conversations courantes, par la quasi-totalité de la population des Comores. Il y a cependant de petites minorités qui parlent l'arabe, le malgache ou le gujarâti. Le comorien n'est pas enseigné à l'école, et l'enseignement est en français. Tous les écrits administratifs, politiques, fiscaux, etc. sont en français. Un autre point intéressant est que la majorité de la population est musulmane, et sait lire et prononcer les textes arabes. Cependant, à peine 15% comprennent l'arabe (un peu comme le latin pour les catholiques avant 1965 et Vatican-II). À cause de l'importance linguistique (120 M de locuteurs) et économique du swahili, qui est une sorte de *lingua franca* en Afrique de l'Est (et entre autres la langue la plus parlée en Tanzanie et au Kenya), les Comoriens ayant des contacts avec l'Afrique orientale connaissent assez bien le swahili — bien que https://fr.wikipedia.org/wiki/Langues_aux_Comores ne le mentionne pas.

Le shikomori s'est séparé du swahili vers le XI° siècle et il n'y a plus d'intercompréhension orale immédiate, critère utilisé par les linguistes pour dire qu'il s'agit d'une même langue (voir https://fr.wikipedia.org/wiki/Comorien). Les parlers comoriens ne sont pas des dialectes du swahili, mais forment une macro-langue, proche du swahili. C'est cette proximité que nous utilisons pour compenser l'absence de systèmes de TA ayant le shikomori (ou plutôt un de ses dialectes) comme langue cible. Il y a aux Comores une volonté de promotion du comorien depuis 1960 environ[8], et les langues officielles de l'Union des Comores sont le comorien, le français et l'arabe. Incidemment, l'anglais est très peu parlé aux Comores (quelques milliers de personnes seulement, bien moins que de locuteurs du swahili).

[6] Comme dans la profession des traducteurs, nous utilisons comme unités la page standard (250 mots ou 1400 caractères), et le nombre de minutes par page (sans aide informatisée, 60mn/p pour un premier jet et 20mn/p pour une révision donnant une qualité professionnelle). La traduction est usuellement payée au mot (entre 10 et 20 centimes d'euro le mot, typiquement 15, soit 15 €/100 mots).

[7] On réserve le terme de *langage pivot* à une représentation intermédiaire dans un processus de TA. GT utilise un texte anglais comme pivot, ATLAS-II et UNL utilisent des graphes sémantiques et des lexèmes interlingues.

[8] Un lettré comorien, Said Kamar-Eddine (1890-1974), proposa alors une adaptation de l'alphabet arabe pour noter le comorien, l'exposa dans des conférences, et l'appliqua dans le journal Mwanâo qu'il avait fondé.

En quoi le comorien est-il peu doté ? Presque rien n'est publié en comorien, malgré les efforts d'associations de promotion de cette langue, et rien sur le Web de façon régulière, avec un flux permettant de collecter des données de taille suffisante pour développer des applications par des méthodes empiriques. ELRA/ELDA et LDC semblent ne rien avoir non plus concernant le shikomori. Le LACITO offre par contre un conte de 80 phrases en ngazidja[9], avec la version orale, et la version écrite donnant le français en parallèle. Il y en a aussi deux pour le mahorais, le plus long de 270 phrases. Il y a quelques dictionnaires en ligne, mais ce sont plutôt des lexiques de petite taille (moins de 2500 entrées, très peu d'information). D'autre part, il semble qu'il n'existe aucun outil d'anlayse ou de génération morphologique, comme il en existe pour le swahili et même pour le somali.

1.2 Objectifs

1.2.1 Applications visées

Vu la situation aux Comores, l'application la plus utile à développer, d'après les Comoriens consultés, serait la *lecture active du français pour les comorophones*. Il ne s'agit pas de remplacer une page Web par son équivalent en comorien, mais plutôt d'enrichir la page originale par des annotations fixes, ou apparaissant lors du survol par le pointeur, et donnant le ou les équivalents d'un mot ou d'un groupe de mots[10]. Si le système de TA utilisé est assez bon, ce qu'on peut obtenir par spécialisation au sous-langage considéré, on pourra aussi afficher la traduction de toute une phrase en comorien au-dessous du texte original, comme cela est fait par le systèmes Neon du Pr Shi à Xiamen (en-zh et zh-en). Notre premier objectif est de nous concentrer sur le grand-comorien (shingazidja), langue maternelle du premier auteur. Les applications à développer ensuite sont (1) des aides à l'apprentissage du comorien (les 4 dialectes) par les résidents étrangers et par les touristes, (2) un analyseur morphologique en vue de la création de correcteurs, (3) un générateur morphologique, (4) un synthétiseur de parole, et (5) des traducteurs de dialogues finalisés (pour les touristes) tournant sur mobiles et tablettes, comme Jibbigo™ .

Le premier objectif concret est la constitution d'un corpus bilingue fr-zdj de bonne qualité et de taille suffisante pour l'apprentissage d'un système de TA spécialisé à un sous-langage « utile ». C'est l'objet principal de cet article. Le second objectif à court terme est la production d'une base lexicale par acceptions[11] fr-km (pour les 4 dialectes, avec en plus fr, en, sw), au début de petite taille, comme sous-produit du développement du corpus.

Les objectifs à moyen terme sont :

- l'intégration dans un système de lecture active (basé sur les outils de M. Mangeot pour le jp-fr).
- la construction de systèmes de TA fr-km : d'abord Moses de niveau 1 (basé sur un corpus parallèle).
- la construction d'un analyseur morphologique (AM) et d'un générateur morphologique (GM) « génériques » (multi-dialectaux) pour le comorien.

Bien qu'ils soient hors des objectifs de la thèse en cours, mentionnons les objectifs à long terme.

- Construction d'un système Moses factorisé, puis d'un système expert écrit en Ariane-G5 ou Ariane-Y et passant par le pivot anglo-sémantique UNL.
- Extension aux 4 dialectes du shikomori.
- Construction d'un système de TA de dialogues oraux de type Jibbigo.

[9] http://lacito.vjf.cnrs.fr/pangloss/tools/list_rsc.php?lg=Ngazidja

[10] C'est de la traduction *pidgin,* terme introduit en 1970 par B. Harris, translatologue, directeur de TAUM, UdM.

[11] comme PIVAX (Nguyen H.T. 2009, Zhang Y. 2016), dans laquelle les sens de mots ou *lexies* de chaque langue sont reliées aux lexies synonymes de la même langue ou d'autres langues par des *axies* (acceptions interlingues).

2 Méthodologie de la construction d'un corpus fr-zdj

2.1 État de l'art de la production de corpus parallèles

La méthode la plus utilisée est la récupération ou l'extraction de gros corpus existants, suivie de filtrage. C'est ce qui est fait à la CEE à Luxembourg, où on a appris un système généraliste sur 20 M segments (phrases), et aussi au WIPO et à l'ONU (à Genève). En partant de mémoires de traductions contenant 11 ans de travail des traducteurs professionnels de l'ONU, Bruno Pouliquen a ainsi développé à leur demande un *accélérateur de traduction* dont ils sont très satisfaits (Pouliquen B. 2013). En l'absence de corpus parallèles, on arrive dans certaines situations à en construire à partir de corpus comparables. C'est ce qu'a fait DO Diep pour le vietnamien-français dans sa thèse, en partant des dépêches d'agences de presse publiant en français et en vietnamien au Vietnam (Do D. 2011).
Aucune de ces méthodes ne convient dans notre cas. De ce point de vue, on ne peut pas le faire non plus pour un couple peu doté de langues très bien dotées, comme le français-chinois. Ainsi, WANG Lingxiao indique dans sa thèse (Wang L. X. 2015) qu'il a dû créer un corpus de 9000 segments fr-zh pour construire un système prototype pour un industriel (EDF ou Renault), car il ne disposait d'aucune mémoire de traductions et d'aucun corpus bilingue comparable, et car l'utilisation du corpus MultiUN, non ciblé, donnait un système de trop basse qualité pour être utilisable. Il a ensuite développé en parallèle son propre système MosesLIG-fr-zh.
La méthode développée ces dernières années dans notre laboratoire consiste à construire un système de TA spécialisé, appris à partir des pré-traductions produites par un ou plusieurs systèmes de TA généralistes existants, si possible gratuits[12], et post-éditées dans un environnement dédié, permettant de gagner beaucoup de temps par rapport à la traduction humaine, comme c'est le cas avec l'environnement SECTra/iMAG. La construction par Houssein Assowe d'un corpus français-somali visant à développer un système de TA fr-so est en cours, sur ce même principe et avec le même environnement.

2.2 Nouvelle approche proposée

Pour construire un corpus français-ngazidja, nous profitons de la proximité du swahili et de l'existence d'un service de TA gratuit en ligne offert par Google (GT-fr-sw). La mise en œuvre de cette approche a nécessité une adaptation de l'outil SECTra/iMAG par le 4° auteur. La traduction vers le ngazidja est initialisée par l'anglais au niveau de la page Web (car la mémoire de traductions est initialisée au swahili, les cellules devant contenir les segments en comorien étant vides au début).

Au niveau de l'interface de post-édition spécialisée de SECTra, on utilise une vue montrant le swahili comme *langue intermédiaire*. C'est un artefact dû au fait qu'on utilise une présentation faite pour le cas d'une TA *composée*, passant par un *pivot textuel*, comme l'anglais pour le français-swahili. Dans notre nouvelle organisation, le swahili est une *langue cible auxiliaire*. Dans le futur, SECTra sera modifié pour mieux tenir compte de cette nouvelle façon de l'utiliser.

2.3 Choix du corpus source

Les critères de choix du corpus source sont les suivants.

- Accessibilité sur le Web, gratuité, et croissance régulière.
- Utilité potentielle du système de TA dérivé (sachant qu'on l'utilisera pour la lecture active).
- Synergie avec le travail similaire en cours sur le français-somali.

Les possibilités trouvées sont les quotidiens *Al-Watwan* et *La Gazette des Comores*. Le premier a été choisi, car il y a une version parallèle publiée en arabe, d'où d'autres ouvertures possibles dans le futur.

[12] Une licence *desktop* d'un système commercial n'est pas très onéreuse et l'utiliser peut être une bonne option.

3 Progression et évaluation de la construction du corpus

3.1 Illustration du processus sur un exemple

Nous avons commencé par mettre en place une iMAG (passerelle interactive d'accès multilingue, ou *interactive Multilingual Access Gateway*) pour le site Web `alwatwan.net`. Mais il y a des pages de même nom qui changent d'un jour à l'autre, et des pages qu'on n'a pas le temps de post-éditer dans la journée, et qui changent d'url le lendemain. Bien qu'elles soient toujours accessibles dans les archives du site Web, elles sont considérées comme différentes par notre logiciel SECTra/iMAG. C'est pourquoi, comme pour le somali, nous avons mis en place un site Web (comorien.imag.fr) sur lequel nous copions les pages que nous voulons traduire pour ensuite mettre les « bons » bi-segments[13] dans le corpus d'apprentissage utilisé pour construire notre système de TA, spécialisé à ce sous-langage.

Concrètement, l'iMAG pointe maintenant vers ce site, et nous utilisons la commande `wget` d'Unix pour y mettre les fichiers à traduire. Pour l'instant, nous utilisons indirectement le segmenteur de GT, et comptons bientôt le remplacer par notre propre segmenteur `SegDoc` (Kalitvianski 2014).

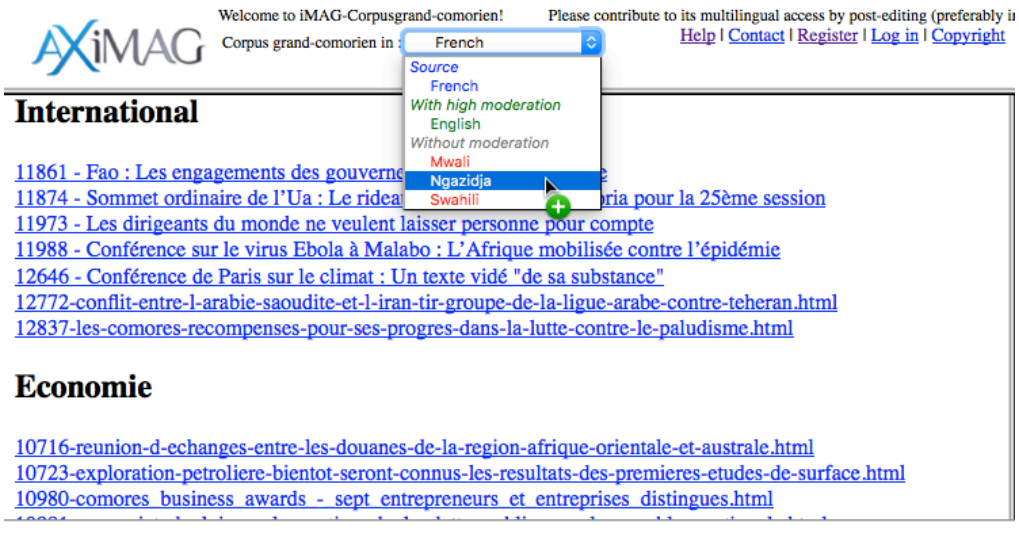

FIGURE 1 : accès au corpus source destiné à être traduit

{_ **PRODUCTION.**_} {_ **The donation worth 125 million Swiss francs was approved by the Minister of Production, Environment, Energy, Industry and Handicrafts.**_}	**PRODUCTION. Le don d'une valeur de 125 millions de francs a été réceptionné par la ministre de la Production, de l'environnement, de l'énergie, de l'industrie et de l'artisanat.**

FIGURE 2 : vue initiale parallèle

[13] On les définit par une condition sur leur niveau de fiabilité et sur leur score de qualité (évalué par le post-éditeur et éventuellement modifié par un réviseur ou un modérateur). Par exemple, *** (bilingue) et ≥13/20 ou **** (traducteur professionnel) et ≥11/20.

Le post-éditeur (premier auteur) choisit la langue cible ngazidja, puis un article (appelé *document* dans SECTra/iMAG), qui apparaît en anglais (et français en regard si on coche la case `Original`). Voir la FIGURE 1. Il arrive à la page Web, reconstituée en anglais. Voir la FIGURE 2. Il accède ensuite à l'interface avancée de PE, qui dans ce cas montre comme TA l'anglais, et comme TM *(Translation Memory)* ce qui a déjà été fait (en comorien, voir FIGURE 3). Pour effectuer la PE à partir du swahili, il utilise la fonction `Postedit Corpus` de SECTra, qui montre le swahili dans la colonne normalement prévue pour une langue pivot de TA. Voir la FIGURE 4.

SECTra mesure le *temps primaire de post-édition*, `Tpe_1`, c'est à dire le temps passé à modifier la cellule de post-édition contenant un segment prétraduit. Dans notre cas, le post-éditeur commence par y copier la traduction vers le swahili, puis modifie le texte (relativement peu) pour obtenir du ngazidja. `Tpe_1` est inférieur au *temps total de post-édition,* `Tpe_tot`, car il faut y ajouter le *temps secondaire de post-édition,* `Tpe_2`, qui correspond au temps passé à lire le texte source avant de cliquer dans la zone de post-édition, et surtout à chercher des équivalents en comorien à des mots ou à des termes français. Nous mesurons le temps total `Tpe_tot` au niveau d'une session, en notant l'heure de début de la PE du premier segment traité, et l'heure de fin du dernier. On ne peut s'astreindre à noter cela pour chaque segment que pour quelques dizaines de segments, car c'est très fastidieux. D'autre part, cela brise le rythme du travail de PE et le ralentit considérablement. Mais, si on le fait par exemple pour les *pages logiques* présentées par SECTra[14], cela n'induit presque pas de perte de temps, puisque de toutes façons il faut aller en haut de page sélectionner la page logique suivante.

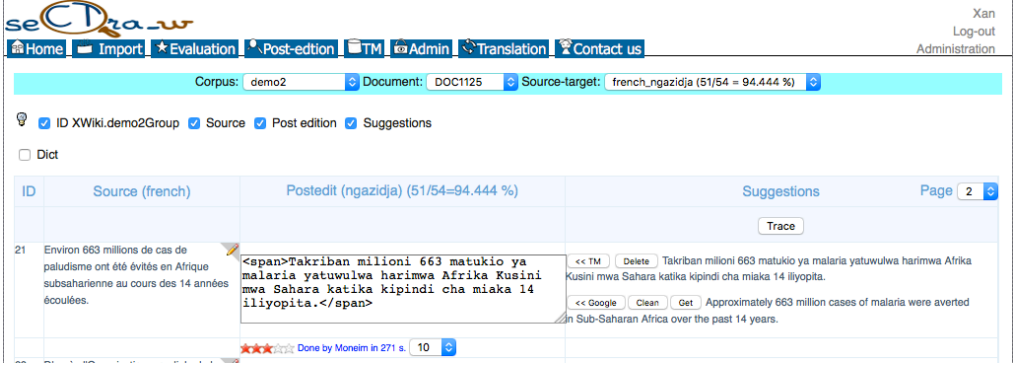

FIGURE 3 : vue classique de PE, avec sortie intermédiaire en anglais

Enfin, nous utilisons la fonction offerte par SECTra d'auto-évaluation de la qualité, segment par segment. Depuis sa conception en 2007 (Huynh C. P. & al. 2008, 2010), SECTra associe à chaque segment, pour chaque langue cible, un niveau de fiabilité (origine de la traduction, de 1 à 5 étoiles[15]), et un score de qualité (de 0/20 à 20/20), initialisé selon le profil du post-éditeur, et modifiable par lui-même, puis par un *modérateur* s'il y en a un.

3.2 Quantité et évolution de la productivité

Au début, nous en étions à environ 33 mn/page, ce qui est à peine mieux que si l'on fait de la traduction humaine classique avec une mémoire de traductions (avec Trados™ par exemple). Au fur et à mesure que le post-éditeur maîtrise mieux l'outil, et cherche moins d'équivalents lexicaux en comorien, à cause bien sûr de la répétition des mêmes mots et termes dans les articles des mêmes rubriques, le temps total

[14] Une page logique est constituée d'un nombre paramétrable de segments. Par défaut, elle en contient 20, ce qui correspond à peu près à une page physique si les segments (phrases ou titres) ont en moyenne 12,5 mots.

[15] * : TA mot à mot, ** : sortie brute de TA, *** : PE par un locuteur de la langue cible, **** : PE par un traducteur professionnel, ***** : traduction par un traducteur certifié par le site source.

`Tpe_tot` diminue. Au bout de 6750 mots (27 pages), il était de 21 mn/page, pour le dernier document traité (1,5 pages). C'est environ 25% de plus que si nous avions un système GT indirect mais allant vers le comorien (en supposant que cela nous prendrait 17 mn/p comme pour le fr-zh).

FIGURE 4 : interface avancée de PE utilisant le swahili comme *langue cible auxiliaire*

N° d'entrée	Corpus	Doc	Nombre de segments	Temps en mn	Nombre de mots	Nombre de pages	Temps par page	Segments /page	Cumul de mots	Moyenne en mn/p sur les 5 derniers documents	Cumul de pages
1	iMAG-demo2	Doc1114	15	52,2	400	1,60	32,63	9,375	400,00		1,6
2	iMAG-demo2	Doc1115	25	58,47	581	1,92	30,45	13,02083	981,00		3,924
3	iMAG-demo2	Doc1116	22	48,5	396	1,58	30,70	13,92405	1377,00		5,508
4	iMAG-demo2	Doc1117	21	55,3	576	2,30	24,04	9,130435	1953,00		7,812
5	iMAG-demo2	Doc1118	29	53,07	456	1,82	29,16	15,93407	2409,00	27,8	9,636
6	iMAG-demo2	Doc1121	22	48,15	367	1,47	32,76	14,96599	2776,00	27,7	11,104
7	iMAG-demo2	Doc1125	54	85	1058	4,23	20,09	12,76596	3834,00	25,4	15,336
11	…										
30	iMAG-demo2	Doc1156	22	40	427	1,7	23,53	12,94	14286	19,99	57,1
31	iMAG-demo2	Doc1157	20	30	255	1	30,00	1,54	14541	21,88	58,2
32	iMAG-demo2	Doc1158	28	48	361	1,4	34,29	20,00	14902	23,86	59,6
33	iMAG-demo2	Doc1159	33	46	693	2,8	16,43	2,54	15595	22,71	62,4
34	iMAG-demo2	Doc1160	31	43	629	2,5	17,20	12,40	16224	21,88	64,9
Totaux		34	899	1502,79	16224	64,62	23,16				

FIGURE 5 : évolution du temps de PE et moyenne atteinte après environ 16224 mots (65 pages standard, 34 documents, 899 segments) en moyenne sur les 5 derniers documents.

3.3 Qualité

Il est prématuré pour l'instant d'organiser une évaluation externe. Nous comptons le faire dès que nous aurons pu motiver des contributeurs comorophones à participer à la construction de ce corpus. Nous pourrons alors faire de l'évaluation croisée, en notant les scores initiaux d'auto-évaluation, puis les scores corrigés par les évaluateurs.

Une autre façon d'évaluer la qualité de la PE initiale sera de demander aux évaluateurs de fonctionner en réviseurs, et de mesurer le temps qu'ils passeront pour arriver à une qualité jugée parfaite. Étant donné qu'il n'y a pas de traducteurs professionnels pour le comorien, et *a fortiori* pas d'écoles de traduction

professionnelle, il ne sera pas possible de prendre comme juges des professeurs de traduction ou des traducteurs professionnels.

Enfin, on pourra aussi évaluer la qualité de ce corpus indirectement, en évaluant la qualité du système de TA (et de ses versions successives suivant sa croissance) qui en sera tiré. Pour l'instant, nous faisons de l'auto-évaluation, et voyons qu'elle est assez bonne à bonne en général, mais peut aussi être mauvaise, la plupart du temps quand le post-éditeur n'a pas trouvé d'équivalent en ngazidja pour un terme français.

Conclusion et perspectives

Nous avons présenté une nouvelle approche permettant de construire un corpus parallèle de bonne qualité (autoévaluée sur chaque segment) pour un couple de langues dont l'une est très peu dotée, n'ayant ni système de TA, ni corpus monolingue ou bilingue (parallèle ou comparable), ni dictionnaires informatisés accessibles, ni outils d'aucune sorte comme des correcteurs et des modules morphologiques (générateur, analyseur, lemmatiseur, racineur, segmenteur, transcripteur…). Cette méthode consiste à post-éditer en ligne des prétraductions dans une langue voisine. C'est le cas du shikomori, pour lequel nous utilisons le swahili comme *langue cible auxiliaire*.

Au moment de la rédaction, nous avons un corpus de bonne qualité français-ngazidja, accessible via l'iMAG http://service.aximag.fr/xwiki/bin/view/imag/Corpusgrand-comorien, formé de 34 articles du journal Alwatwan (899 segments, 16224 mots, 65 pages standard), et sommes descendus de 28 mn/p au début à 23 mn/p maintenant, en moyenne sur les 5 derniers documents. Nous commençons à extraire en parallèle des correspondances lexicales bilingues. Il y a deux perspectives. (1) Nous allons bientôt construire une première version MosesLIG-fr-zdj et espérons obtenir une productivité de l'ordre de 15 mn/p, en l'utilisant pour produire les prétraductions. (2) La première application visée est la lecture active du français pour des locuteurs du comorien ; elle utilisera le dictionnaire et le système de TA dérivés respectivement de la base lexicale et du corpus, en croissance, et les techniques développées récemment par M. Mangeot pour le japonais.

Remerciements

Nos remerciements vont à la BID (banque islamique de développement), qui finance la bourse de thèse du premier auteur, au professeur Ahmed Chamanga, qui nous a communiqué son lexique sous forme électronique et nous aide de ses conseils, ainsi qu'à Mathieu Mangeot qui met à notre disposition ses outils pour la construction d'une base lexicale et d'une fonction de lecture active.

Références

AHMED ALI S. (2012) *Le "shikomori", un patrimoine linguistique à l'agonie !* (Blog urentre sur le shikomori), Moroni, http://urentre.over-blog.com/article-le-shikomori-un-patrimoine-linguistique-a-l-agonie-par-salim-ahmed-ali-enseignant-de-physique-101474843.html

AHMED-CHAMANGA M. (2008) Introduction à la grammaire structurale du comorien : le shingazidja. Y. M. publications, ed., *Introduction à la grammaire structurale du comorien,* Paris, 2 vol.,

BERMENT V. (2004) Méthodes pour informatiser les langues et les groupes de langues peu dotées. *Thèse,* UJF, 250 p.

BEY Y., KAGEURA K. ET BOITET C. (2006) *Data Management in QRLex, an Online Aid System for Volunteer Translators. IJCLCLP (International Journal of Computational Linguistics and Chinese Language Processing),* **11**/4, pp. 349—376.

BEY Y., KAGEURA K. ET BOITET C. (2007) *BEYTrans: A Free Online Collaborative Wiki-Based CAT Environment Designed for Online Translation Communities*. Proc. PACLIC-21 (Pacific Asia Conference on Language, Information and Computation), Seoul, 8 p.

BOITET C. (2013) *Infrastructures in CLIR for patents: what we (think we) have and what we want*. Proc. RANLP, MUMIA WG-2 meeting, Hissar, Bulgaria, 4 p. (slides)

BOITET C. (2013) *Infrastructures for multilingual IR and MT. Proc. MUMIA WG-2 meeting,* Tallinn, Estonia, 24 p. (slides)

BOITET C. (2013) *Emerging trends towards a multilingual semantic web*. Proc. LICIA seminar, UFRGS, Porto Alegre, 16 p. (slides)

BOITET C., PHAP H. C., NGUYEN H. T. ET BELLYNCK V. (2010) *The iMAG concept: multilingual access gateway to an elected Web site with incremental quality increase through collaborative post-edition of MT pretranslations*. Proc. TALN-2010, UdM, Montréal, 8 p. (bilingual demonstration presentation)

CHEN Y., WANG L., BOITET C. ET SHI X. (2014) *On-going Cooperative Research towards Developing Economy-Oriented Chinese-French SMT Systems with a New SMT Framework*. Proc. TALN-2014, Marseille, France, 6 p.

DAOUD M., BOITET C., KAGEURA K. ET MANGEOT M. (2010) *Indirect Contribution to Multilingual Lexical Resources through Online Cultural Activities*. Proc. NLP'KE-10 (Natural Language Processing & Knowledge Engineering), CNU (Central Normal University), Beijing, 9 p.

DIEP DO T. N. (2011) *Extraction de corpus parallèles pour la traduction automatique depuis et vers une langue peu dotée*. Thèse, UdG.

GUILBAUD J.-P., BOITET C. ET BERMENT V. (2013) *Un analyseur morphologique étendu de l'allemand traitant les formes verbales à particule séparée. An extended morphological analyzer of German handling verbal forms with separated separable particles*. Proc. TALN-2013, Les Sables d'Olonne, 10 p.

HADRICH BELGUITH L. S., RAHMA; SADAT, FATIHA (2012) *L'extraction de phrases en relation de traduction dans Wikipédia*. Sfax (Tunisie), Montréal (Canada), 12 p.

NEGRE X. (2002-2016) *Dictionnaire comorien*. LEXILOGOS, mots et merveilles d'ici et d'ailleurs http://www.lexilogos.com/comorien_dictionnaire.htm.

WANG L. (2013) *Contributive online production of HQ parallel corpora and task-based evaluation of MT systems can be obtained through iMAGs with no added cost*. Proc. RANLP-2013, Workshop on Adaptation of language resources and tools for closely related languages and language variants, Nice, France, 8 p. (with a demo)

WANG L., AZOUZI M., BOITET C. ET BESACIER L. (2013) *T7o — Intergiciel de génération et d'exploitation de systèmes de TA probabilistes basés sur Moses et spécialisés à des sous-langages*. Livrable L3.5-v2, projet ANR Traouiero, 24/1/2013.

WANG L. ET BOITET C. (2013) *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*. Proc. Machine Translation Summit XIV, 2nd Workshop on Post-Editing Technologies and Practice (WPTP-2), Nice, France, 8 p.

WANG L. ET BOITET C. (2014) *Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems*. March 2014, MUMIA COST project.

Ce résumé en ngazidja a été ajouté à la demande du comité de lecture de TALAf. Il a été produit par la même méthode que celle présentée dans l'article pour créer un corpus français→comorien.

Uhandziha wa mwili yilolanao kifaransa-kikomori ha urumia tafsiri ya otomatik kifaransa-kiswahili

MUKHTASAR

Shikomori yo lugha yenya lahaja jumla nne (njazidja maore, mwali, ndzuani) karibu hakuu na kila mmoja, na karibu Kiswahili utosha. Nguio maskini hakuu ha maalumati ya lugha kompyuta, ketsina mwili wala kamusi, wala zana za usahihisha au tafsiri. Ha hiyo kena humkinichiha ufanisi uhanziha mwili ufanihawo, hama anajulikana ha tarjama otomatik (TA) ikifuatiwa na baada ya edition (PE) interactive: kichina-kifaransa, 17 min / ukurasa na Google Translate (GT), 12 min / ukurasa na mfumo TA MosesLIG.fr-zh urumia SECTra/iMAG. Hata hivyo, sisi ni kujaribu ufanya hivyo harimwa post edition kabla ya tafsiri ha kiswahili zinazozalishwa na GT. Shiswahili shirumichiha pvanu hama lugha pivot, lakini hama lugha msaidizi. Ngasina avasa mwili wa bora kifaransa-kingazidja yapanguiha ha 34 makala ya gazeti Alwatwan (899 makundi, 16224 makalima, 65 safha standard). Ngari towawo ha sambamba mawasulano ya hati zalugha bili. Ye appli yahandani yadumaniwa ndo usoma wa wasemaji wa kifaransa haajli ya waronguwa wa shikomori; nguedjorumia ye kamusi na mfumo TA inayotokana mtiririko utoka ya “base lexicale” na uhua.

MAKALIMA-YAUBUA:
uhanziha wa mwili yifanishihao, kifaransa-kikomori, kiswahili, lugha msaidizi