Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014

Mathieu Mangeot[1], Fatiha Sadat[2]
(1) GETALP-LIG, 41 rue des mathématiques, 38041 Grenoble Cedex 9
(2) UQÀM, 201 av du Président Kennedy, Montreal, QC, Canada
Mathieu.Mangeot@imag.fr, Sadat.Fatiha@uqam.ca

Préface

1 Motivations et objectifs

Dans la suite du premier atelier TALAf qui s'est tenu le 8 juin 2012 à Grenoble, lors de la conférence JEP-TALN-RECITAL 2012 (voir les actes : http://aclweb.org/anthology//W/W12/#1300), nous proposons une nouvelle édition de cet atelier lors de la conférence TALN 2014 le premier juillet à Marseille. Nous accueillons les travaux menés sur toutes les langues peu dotées d'Afrique y compris l'arabe dialectal de l'Afrique du nord (maghrébin).

Les recherches en traitement automatique des langues africaines sont actuellement à l'orée de développements majeurs. Les efforts de reconnaissance des langues nationales et de standardisation des différents alphabets commencent à porter leurs fruits. Au Niger, par exemple, les alphabets des langues fulfulde, haussa, kanouri, songhai-zarma et tamajaq ont été définis par des arrêtés du gouvernement en 1999. Par ailleurs, un certain nombre de collègues formés dans les pays du Nord reviennent dans leur pays avec la volonté de continuer leur travail sur les langues locales. Il y a également des diasporas disposant de moyens technologiques leur permettant de contribuer directement en ligne et de manière bénévole.

Pour autant, les langues nationales de la plupart des pays d'Afrique sont peu dotées (langues-π) : les ressources électroniques disponibles sont rares, mal distribuées, voire inexistantes. Seules sont accessibles les fonctions d'édition et d'impression rendant l'exploitation de ces langues difficile. Au moment où il est question de les introduire dans le système éducatif, de créer des normes d'écriture standardisées et stabilisées et surtout de développer leur usage à l'écrit et à l'oral dans l'administration et la vie quotidienne, un développement de ces langues s'impose comme une nécessité vitale.

Développer le traitement automatique de langues africaines nécessite l'élaboration de ressources qui seront les fondements à partir desquels des traitements plus élaborés peuvent être construits. Il apparaît indispensable de constituer en premier lieu des corpus écrits et oraux annotés aussi larges que possibles. À partir de tels corpus, il est possible d'extraire des exemples pour aider à la constitution de dictionnaires ou de mettre au point des modèles de langage pour la reconnaissance vocale. Toutefois, la constitution de tels corpus reste une entreprise délicate dans le contexte de langues peu dotée car les transcriptions souffrent du manque de standardisation de la langue et l'enrichissement de corpus reste très onéreux.

Le développement d'applications à base de traitement de l'oral peut être considéré comme prioritaire dans des régions de tradition orale. De plus, l'usage de téléphones mobiles, très répandu, permet d'imaginer un déploiement rapide de ces applications.

Les dictionnaires sont également nécessaires pour construire les outils de base tels les correcteurs orthographiques (qui peuvent servir à leur tour pour corriger les corpus écrits) ou encore pour l'aide à la transcription de corpus oraux. Il existe parfois des dictionnaires bilingues couplant la langue officielle et une langue nationale. Par exemple, au Mali, le père Charles Bailleul est l'auteur d'un dictionnaire bambara-français ; au Niger, le projet éducatif SOUTÉBA a créé cinq dictionnaires bilingues destinés aux enfants de primaire. Mais ceux-ci existent uniquement en version papier ou sous forme de fichiers d'éditeurs de texte (format.doc). Informatiser ces dictionnaires pour les rendre utilisables par des outils de traitement automatique nécessite, dans un premier temps, d'ajouter des informations manquantes : prononciation, règles de flexion morphologiques et flexionnelles, exemples et traductions tirés de corpus, etc.Il s'agit dans un premier temps de les informatiser (les transformer dans un format utilisable par des outils de traitement automatique) et de les compléter avec des informations manquantes : prononciation, exemples et traductions tirés de corpus, etc. Des astuces peuvent parfois être inventées pour pallier le manque de ressources. Par exemple, s'il n'existe pas de corpus oraux avec transcriptions, il est possible de constituer un corpus oral de textes lus.

Enfin, il y a lieu de prendre en compte les contraintes socio-économiques s'exerçant sur la population des locuteurs : les ressources économiques sont limitées, les ressources humaines qualifiées sont rares, les recherches sont sporadiques et isolées, les résultats confidentiels et parcellaires. Il est donc nécessaire de définir des méthodologies économes en coût d'achat de logiciels et en temps de travail qualifié visant à produire des résultats pérennes, partagés et faciles à enrichir. La constitution de ressources linguistiques de manière générale, et plus encore pour les langues africaines devrait donc respecter un certain nombre de principes : utilisation d'outils en source ouverte, définition et utilisation de standards (ISO, Unicode), transfert de connaissances entre les collègues des pays du Nord et du Sud, disponibilité des ressources sous licence ouverte (Creative Commons), etc.

Cet atelier a pour but d'effectuer un état des lieux des travaux de constitution de ressources linguistiques de base (dictionnaires, corpus oraux et écrits), de mettre au point des méthodologies simples et économes d'élaboration de ressource, d'échanger sur les techniques permettant de se passer de certaines ressources inexistantes et de fixer un certain nombre de principes pour les futurs travaux dans le domaine.

Les ateliers TALAf sont soutenus par l'association LTT (Lexicologie Terminologie Traduction).

2 Présentation des articles

L'atelier a reçu treize soumissions. Onze articles ont été rédigés en français et deux en anglais. Pour mémoire, l'atelier TALAf avait reçu 12 soumissions.

Parmi ces articles, six ont été acceptés en première lecture, et quatre acceptés après révision. Tous les articles portent sur l'écrit, ce qui s'explique par le thème de la conférence principale TALN.

La diversité linguistique est présente puisque huit langues figurent dans les articles acceptés : amazighe (kabyle), bambara, maninka, haoussa, ikota, mwan, yambetta, wolof.

La plupart des travaux portent sur le bambara (3) et le wolof (2).

Les auteurs se répartissent entre cinq pays : Cameroun (2), France (6), Niger (2), Russie (2), Sénégal (3)

Les articles acceptés se regroupent autour de trois thèmes principaux :

2.1 Corpus

– Valentin Vydrin : *Projet des corpus écrits des langues manding : le bambara, le maninka*.

Cet article traite d'un projet de construction de corpus écrits pour le bambara et le maninka. Dans le futur, il est envisagé une extension à l'oral. Ces langues sont parlées majoritairement au Mali, Guinée, Sénégal, Côte_d'Ivoire et Burkina_Faso.

– Kirill Maslinsky : *Daba: a model and tools for Manding corpora*.

Cet article, rédigé en anglais, traite du même projet que l'article précédent et l'aborde sous l'angle des outils utilisés.

– Mahfoud Mahtout : *Méthodologie pour la structuration semi-automatique du corpus dans une perspective de traitement automatique des langues : le cas du dictionnaire français-kabyle*.

Cet article traite de l'élaboration d'un dictionnaire français-kabyle. Le kabyle, est une langue berbère parlée principalement en Kabylie, région d'Algérie. Le nombre de locuteurs est évalué entre 5 et 6 millions.

2.2 Morphologie et orthographe

– Brunelle Magnana Ekoukou : *PFM : pour une implémentation de la morphologie de l'ikota dans XMG*.

Dans cet article, le formalisme XMG (eXtensible MetaGrammar) est utilisé pour décrire les variations morphologiques de l'ikota, langue bantoue parlé au Gabon. Le nombre de locuteurs natifs est évalué à 43 000.

– Jean-Jacques Méric : *Un vérificateur orthographique pour la langue bambara*.

Le bambara est parlé majoritairement au Mali. Le nombre de locuteur varie entre 10 et 13 millions selon les estimations.

- Lawaly Salifou & Harouna Naroua : É*tude et conception d'un correcteur orthographique pour la langue haoussa*.

Cet article traite de la conception d'un correcteur orthographique programmé en Java selon des technique standard pour la langue haoussa. Le haoussa est parlé principalement au Niger, au Nigeria et au Tchad. nombre de locuteurs est évalué entre 40 et 50 millions.

2.3 Lexique

- Manifi Abouh Maxime, Yves Julien & Sadembouo Etienne : *De la dénomination des concepts techniques dans l'élaboration d'un lexique thématique agricole bilingue français-yambetta*.

Le yambetta (encore appelé nigi) est une langue bantu du Mbam qui a pour code 520 dans l'Atlas linguistique du Cameroun (ALCAM). C'est une langue parlée par une minorité de 3700 personnes (Gordon, R.,and Grimes, B., 2005) dans la région de savane arborée située entre Bafia et Ndikiniméki, dans la vallée du Mbam au Cameroun.

- Mouhamadou Khoule, El Hadji Mamadou Nguer & Mouhamaad Ndiankho Thiam : *Vers la mise en place d'un lexique basé sur LMF pour la langue Wolof*.

LMF : Lexical Markup Framework est un standard ISO pour la représentation de lexiques. Le Wolof, parlé principalement au Sénégal, est la langue véhiculaire de ce pays. Le nombre de locuteurs est évalué à environ 11 millions.

- Elena Perekhvalskaya : *The Mwan language: dictionary and corpus of texts*.

Cet article, rédigé en anglais, décrit la construction de ressources écrite pour le monan (Mwan en anglais). Cette langue mandée est parlée en Côte d'Ivoire. Le nombre de locuteurs est estimé à 17 000.

- Abibatou Diagne : *De quelques problèmes de traduction des adjectifs relationnels du français vers le wolof : étude sur corpus de terminologie commerciale*.

Cet article traite également de la langue wolof, parlée essentiellement au Sénégal.

3 Comité de programme

Laurent Besacier (LIG, Grenoble, France)

Philippe Bretier (Voxygen, Pleumeur-Bodou, France)

Khalid Choukri (ELDA, Paris, France)

Mame Thierno Cissé (ARCIV, Université Cheikh Anta Diop, Dakar, Sénégal)

Denys Duchier (Université d'Orléans, Orléans, France)

Chantal Enguehard (LINA, Nantes, France)

Gil Francopoulo (Tagmatica, Paris, France)

Mathieu Mangeot (LIG, Grenoble, France)

Chérif Mbodj, (Centre de Linguistique Appliquée de Dakar, Sénégal)

Kamal Naït-Zerrad (INALCO, Paris, France)

Pascal Nocera, (Université d'Avignon, France)

Francois Pellegrino, (DDL, Lyon, France)

Fatiha Sadat (UQAM, Montréal, Canada)

Mamadou Lamine Sanogo (INSS, Ouagadougou, Burkina-Faso)

Emmanuel Schang (Université d'Orléans, Orléans, France)

Gilles Sérasset (LIG, Grenoble, France)

Valentin Vydrin (LLACAN-INALCO, Paris, France)

4 Conclusion

Cette deuxième édition montre l'intérêt d'un atelier francophone sur le traitement automatique des langues africaines. Le TAL en Afrique est en train de prendre son essor. Les travaux restent encore éparpillés mais cet atelier et de la liste de discussion par courriel talaf@imag.fr permet de construire et de structurer la communauté qui se met en place actuellement. Les savoirs et savoirs-faire doivent également être capitalisés pour resservir pour d'autres langues et d'autres contextes.

Le prochain atelier TALAf est prévu pour 2016, conjointement avec la conférence JEP-TALN-RÉCITAL. Ce sera certainement l'occasion de recueillir des soumissions portant sur l'oral.