Demande de financement pour le projet DiLAF

Contexte

Les attentats sévissent en Europe, mais aussi en Afrique, notamment au Mali, au Niger, ou encore au Burkina Faso. Ils déstabilisent des États fragiles dont le système scolaire reste largement insuffisant. Ainsi, le taux d'analphabétisme est important, jusqu'à 80% dans la zone sahélienne. Le faible niveau d'éducation de la population est un des freins au développement économique, social et politique de ces pays. Cette situation peut conduire une partie de la populations à migrer ou à se radicaliser.

De plus, ces pays font face à des difficultés d'ordre linguistique qu'il est difficile de surmonter[1]: la **langue officielle** est le français. C'est aussi la langue d'enseignement. Mais cette langue n'est pas la langue maternelle d'une grande part de la population (environ 90 %) et reste peu utilisée en dehors des échanges avec les autorités. Les enfants apprennent donc à lire dans une langue qu'ils ne comprennent pas et dont ils devront en sus acquérir le vocabulaire, les règles de grammaire et de bon usage.

Des expériences d'écoles bilingues/multilingues ont vu le jour depuis plusieurs dizaines d'années. Les enfants y apprennent à lire dans leur langue maternelle ; le français est introduit de manière graduelle au cours de l'éducation primaire. De bons résultats sont observés[2], cette stratégie contribue donc à améliorer le niveau d'éducation de la population. En particulier, "*Les filles auraient de meilleurs résultats si elles étaient scolarisées dans des idiomes qu'elles maîtrisent, les langues nationales.*"[3]

Mais ces langues sont "peu dotées" : les ressources linguistiques les concernant sont peu diffusées ou inaccessibles. Ainsi même les professeurs ne disposent que de petits manuels rédigés dans ces langues. Ils sont privés des ressources à la fois élémentaires et fondamentales que sont les dictionnaires et les grammaires. Évidemment la population n'a pas non plus accès à de ces ressources, ce qui entrave le développement des échanges techniques, scientifiques, culturels et économiques.

De plus, ces langues sont également "peu dotées" en ce qui concerne le Traitement Automatique des Langues (TALN) : correction orthographique, traduction automatique, synthèse de la parole, dictée automatique, etc. Alors que nombre de ces applications seraient très utiles dans une société où le taux d'analphabétisme est important (autour de 80 %).

Historique et résultats préliminaires

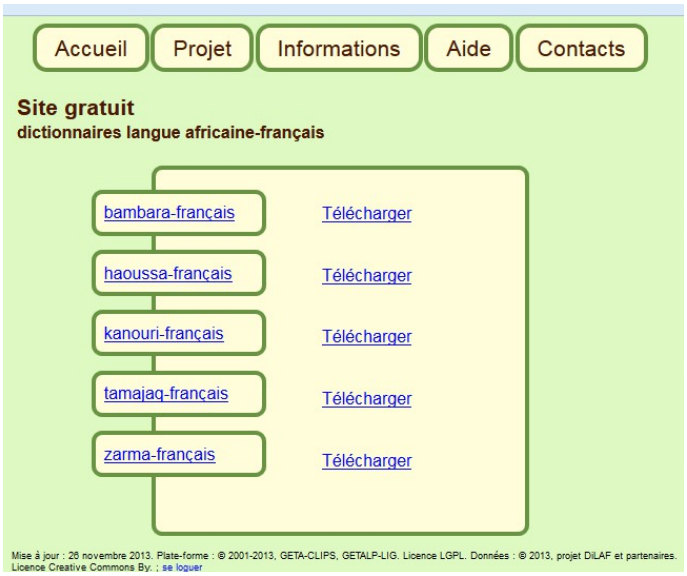
DiLAF est un projet de recherche de mise ligne de dictionnaires de langues peu dotées.

Les dictionnaires peuvent être consultés et téléchargés gratuitement sur le site web www.dilaf.org

À notre connaissance il n'existe aucun site web offrant l'accès gratuit à des dictionnaires de langues africaines de qualité académique.

La phase initiale du projet a été financée par l'Organisation internationale de la Francophonie.

1 et qui s'ajoutent à une situation socio-économique affligeante.
2 par l'Académie Africaine des Langues (ACALAN) ou l'Organisation Internationale de la Francophonie
3 Plan Décennal du Développement de l'Éducation (PDDE) au Niger



Page d'accueil

Cinq dictionnaires ont été mis en ligne depuis 2013. Ils ont été téléchargés 370 fois, en majorité par des personnes africaines, surtout sahéliennes. Il s'agit surtout de lycéens, commerçants, enseignants, proviseurs, rédacteurs de manuels, linguistes, évangélistes, agents de service public (planification, statistiques, formation professionnelle), chercheurs, techniciens (informatique, pisciculture), journalistes, chômeurs, etc.[4]

Le projet semble donc combler un besoin au sein des populations africaines, comme nous en avions fait l'hypothèse.

Méthodologie

Produire *ex nihilo* un dictionnaire électronique pour une langue représente un effort de plusieurs années. Or des dictionnaires éditoriaux (destinés à être distribués sous forme imprimée) existent.

Pour ce projet nous avons développé une stratégie originale et à moindre coût : il s'agit de convertir des dictionnaires éditoriaux déjà existants (rédigés avec des éditeurs de texte usuels) en dictionnaires électroniques. Les dictionnaires sont mis en ligne avec l'accord des auteurs. Ils sont alors protégés par une licence Creative Commons CC BY-NC-SA 2.0 . Celle-ci autorise la consultation et la réutilisation non commerciale des dictionnaires.

Les dictionnaires électroniques produits respectent le standard international ISO/TC 37/SC 4 N130, ce qui permet, à la fois, l'**affichage en ligne** des dictionnaires et l'**utilisation par des applications de TALN**.

La méthodologie de conversion a été définie, rédigée et testée lors de la phase initiale du projet. Elle comprend sept étapes : 1-prétraitements ; 2-conversion vers Unicode ; 3-choix des noms des éléments XML et structuration ; 4-balisage à l'aide d'expressions rationnelles ; 5-contrôle des listes fermées ; 6-licence ; 7-documentation.

Nous avons établi qu'un dictionnaire peut être converti en un à deux mois[5] de travail par une équipe alliant un informaticien spécialiste du TALN et un linguiste ou lexicographe de la langue du dictionnaire.

Voici les principales publications sur le projet DiLAF :

4 Malheureusement le nombre de consultations en ligne n'a pas été comptabilisé.

5 Cette durée peut varier en fonction du format du dictionnaire récupéré et de la maîtrise de la méthodologie par l'équipe.

Enguehard, C. Mangeot, M. DILAF : des dictionnaires africains en ligne et une méthodologie. Actes de l'atelier Traitement automatique des langues africaines "collecte et organisation de ressources linguistiques". Dakar. Sénégal. 22 novembre 2014.

Enguehard, C. Mangeot, M. LMF for a selection of African Languages. Chapter 7. in Gil Francopoulo (dir.), LMF: Lexical Markup Framework, theory and practice. Ed., Hermès science. Paris, France. 17 p. 2013.

Mangeot, M. Enguehard, C. Des dictionnaires éditoriaux aux représentations XML standardisées. Chapter 8. in Gala, Núria et Michael Zock (dir.), Ressources Lexicales: Contenu, construction, utilisation, évaluation. xii, 364 pp. (pp. 255-290). 2013.

Objectifs

Le projet poursuit plusieurs objectifs à long terme :
— lutter contre l'analphabétisme et réduire le taux d'échec scolaire,
— faire bénéficier les langues nationales des outils de Traitement Automatique des Langues Naturelles (TALN).

À moyen terme :
— faciliter l'accès aux dictionnaires, favoriser l'expression d'écrits en langue nationale, principalement chez les jeunes,
— encourager la production de pages web bilingues et dans les langues nationales,

À court terme :
— **ajout d'un dictionnaire fulfulde-français** sur le site web DiLAF. Nous avons recueilli les fichiers sources de la version éditoriale et l'autorisation des auteurs ;
— **enrichissement de cinq dictionnaires** (haoussa, kanouri, tamajaq, wolof et zarma), notamment la traduction en français des exemples d'usage et, si possible des définitions. Il s'agit de constituer ainsi des corpus bilingues qui seront libres d'usage pour la recherche ;
— **recherche de nouveaux dictionnaires à mettre en ligne** (récupération des fichiers sources et autorisation des auteurs).

Organisation

Le projet est géré par l'Université de Nantes via le laboratoire d'Informatique de Nantes-Atlantique (LINA). Ce dernier se chargera d'organiser les missions des collaborateurs africains en son sein (accès à un bureau, à un ordinateur, aux ressources bibliographiques, etc.).

Pour chaque dictionnaire, deux missions de trois semaines chacune sont nécessaires : une première mission permettra d'avancer le travail et de collecter les questions linguistiques ou lexicographiques qui ne peuvent être tranchées sur place. Les collaborateurs seront chargés de consulter des collègues afin d'apporter des réponses à ces questions.

Une seconde mission permettra de prendre en compte les réponses aux questions, d'ajouter éventuellement de nouvelles entrées aux dictionnaires et d'achever la révision.
Du fait des problèmes de sécurité dans la zone sahélienne, les recherches se dérouleront au LINA, à Nantes.

Chantal Enguehard, enseignante-chercheuse en informatique (membre de l'UMR LINA) et spécialiste en TALN, assure la coordination et la direction scientifique du projet depuis ses débuts, en collaboration avec Mathieu Mangeot (enseignant-chercheur à l'Université de Grenoble).

Les collaborateurs africains travaillant sur ce projet sont des professionnels : chercheurs en

informatique ou en linguistique, lexicologues. Ils appartiennent à des institutions spécialisées dans les langues nationales. Il s'agit de l'INDRAP (Institut National de Documentation de Recherche et d'Animation pédagogiques) et de la DRELN (Direction de la Recherche et de l'Equipement des Langues Nationales) au Niger, et du CLAD (Centre de Linguistique Appliquée de Dakar), au Sénégal.

Demande de financement

Nous sollicitons le CNRS pour financer le projet quant à ses objectifs à court terme : ajout d'un dictionnaire fulfulde-français et enrichissement de cinq dictionnaires. Ces langues sont parlées par plusieurs dizaines de millions de personnes[6].

Coûts unitaires

Coût d'une mission d'un collaborateur africain en France :
1400 euros de transport
+ frais de séjour (22 jours à 110 euros/ jour)
= 3820 euros

Coût d'une mission de Mathieu Mangeot à Nantes :
550 euros de transport
+ frais de séjours (5 jours à 110 euros/ jour)
= 1100 euro

Coût d'une mission pour publication
500 euros de transport
200 euros d'inscription
+ frais de séjours (5 jours à 110 euros/ jour)
= 1250 euros

Budget total

	Coût unitaire	Nombre	Coût	Financement
Mission de Mathieu Mangeot à Nantes	1 050	1	1 100	CNRS
Mission pour publication	1 250	1	1 250	CNRS
Missions de collaborateurs africains en France	3 820	10	38 200	CNRS
Stage	600	5	3000	CNRS
Frais de gestion	20%	1	8 710	CNRS
Frais co-encadrement par M. Mangeot	100 998	0,5	4 208	U. Grenoble
Frais coordination par C. Enguehard	100 998	4	33 666	U. Nantes
Total			92634	

Aussi nous sollicitons du CNRS un financement de **52 260 euros** (soit 55% du budget total).

6 Baromètre Calvet des langues du monde

Dimension humaine et éthique

Ce projet s'inscrit dans une démarche éthique de recherche en TALN sur les langues peu dotées[7] puisqu'il s'agit de réaliser des recherches sur des langues qui, non seulement sont peu dotées, mais aussi pour lesquelles, *en sus*, il y a peu de recherches et peu de moyens. Il faut ajouter que, localement, le statut hégémonique du français par rapport aux langues nationales prolonge le sentiment d'injustice lié à la période coloniale dont le souvenir reste vivace. Investir en recherche dans ces langues au même niveau que le français permettrait aux locuteurs de se sentir revalorisés quant à leur culture, et d'accroître le sentiment de confiance et de fierté, qualités nécessaires au développement endogène toute nation.

Il s'agit aussi de prolonger des collaborations scientifiques mises en œuvre depuis une vingtaine d'années et qui ont été brutalement interrompues suite à l'enlèvement et l'exécution de deux français par Al-Qaïda à Niamey en janvier 2011.

Outre les objectifs scientifiques déjà annoncés, ce projet sera l'occasion de renforcer les liens scientifiques entre nos pays et d'effectuer des transferts de connaissances. Il permettra aussi aux collaborateurs africains d'effectuer un séjour dans un laboratoire en France et d'avoir accès aux bibliothèques universitaires, rompant ainsi l'isolement dans lequel ils se trouvent.

Contact

Chantal Enguehard
LINA - UMR CNRS 6241
2, rue de la Houssinière
BP 92208
44322 Nantes Cedex 03

téléphone : 02 51 12 58 55
télécopie : 02 51 12 58 12

chantal.enguehard@univ-nantes.fr
http://www.sciences.univ-nantes.fr/info/perso/permanents/enguehard/

7 Enguehard, C. Mangeot, M. Favorisons la diversité linguistique en TAL. Journée d'étude de l'ATALA. "Ethique et Traitement Automatique des Langues". 22 novembre 2014. Paris.