

*Adaptation d'un système de recherche
d'information multilingue pour le traitement de
l'arabe*

**LAIB, M., SEMMAR, N.
GARA, F., FLUHR, C.
GREFENSTETTE, G.**

CEA, LIST, France

1. Introduction

Le nombre des documents multilingues mis en ligne va en croissant. Pour pouvoir structurer cette masse de données hétérogènes, les systèmes de recherche d'information s'orientent de plus en plus vers le multilinguisme et tentent de traiter le plus de langues possible, le but étant de pouvoir toutes les traiter un jour. Pour un système de recherche d'information en langage naturel, chaque nouvelle langue représente un défi à relever, plus particulièrement lorsque celle-ci appartient à une famille de langues différentes de celles pour lesquelles le système a été conçu initialement.

Depuis quelques années, le LIC2M (Laboratoire d'Ingénierie de la Connaissance Multilingue Multimédia) a développé un système de recherche d'information multilingue en langage naturel. Ce système traite 9 langues (français, anglais, allemand, espagnole, italien, russe, arabe, chinois et japonais en cours) pour lesquelles

Adaptation d'un système de recherche d'information multilingue pour le traitement de l'arabe

les concepteurs tentent sans cesse d'améliorer les performances en qualité et en temps de traitement.

Dans cet article, nous allons tenter d'expliquer comment le système du LIC2M, développé au départ pour des langues européennes, a été adapté pour traiter une langue sémitique telle que l'arabe. Notre système étant conçu selon un modèle modulaire, il est relativement facile d'y intégrer une nouvelle langue. Les seules difficultés restent celles relatives à la structure de la langue elle-même (structure grammaticale et lexicale) et de sa parenté ou non par rapport aux autres langues déjà traitées par le système.

Nous allons donc, dans un premier temps, expliquer les différentes étapes du traitement linguistique concernant les modules génériques. Nous détaillerons, dans un deuxième temps, le fonctionnement des modules qui ont été conçus spécifiquement pour traiter la structure linguistique propre à l'arabe. Nous expliquerons, par la suite, dans quelle mesure ces modules sont ou seront susceptibles d'être réutilisés pour faciliter l'intégration d'autres langues sémitiques ou non. Nous exposerons enfin l'exemple du projet ALMA européen dans lequel le traitement de l'arabe a été testé (version consultable en ligne).

2. Architecture du système de RI développé au LIC2M

Le système de RI développé au LIC2M est un système modulaire paramétrable pour traiter plusieurs langues. La requête et les documents indexés subissent une analyse linguistique fine qui permet une indexation, une reformulation et une interrogation monolingue et multilingue.

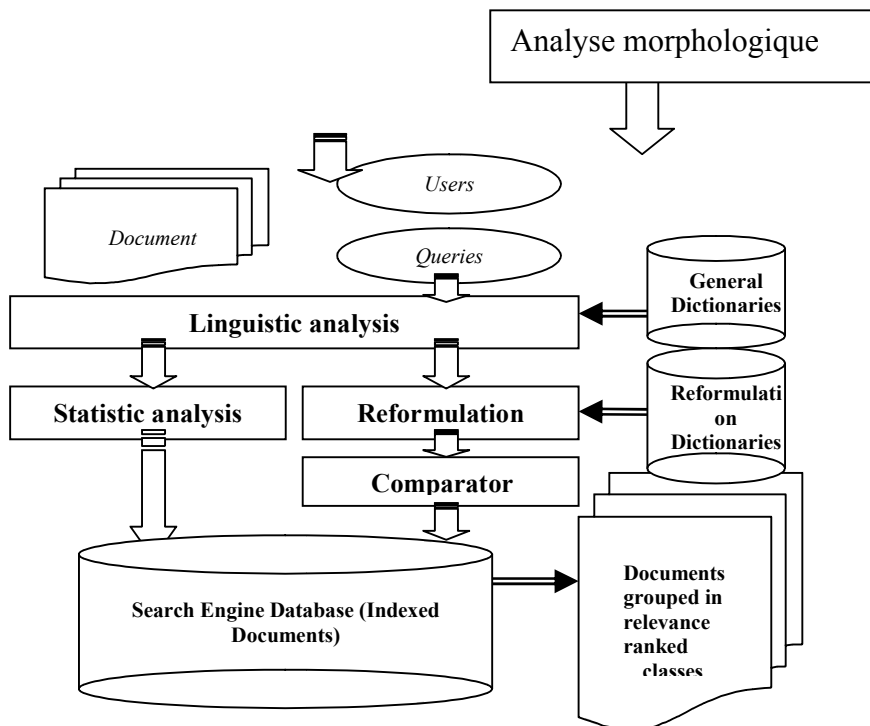
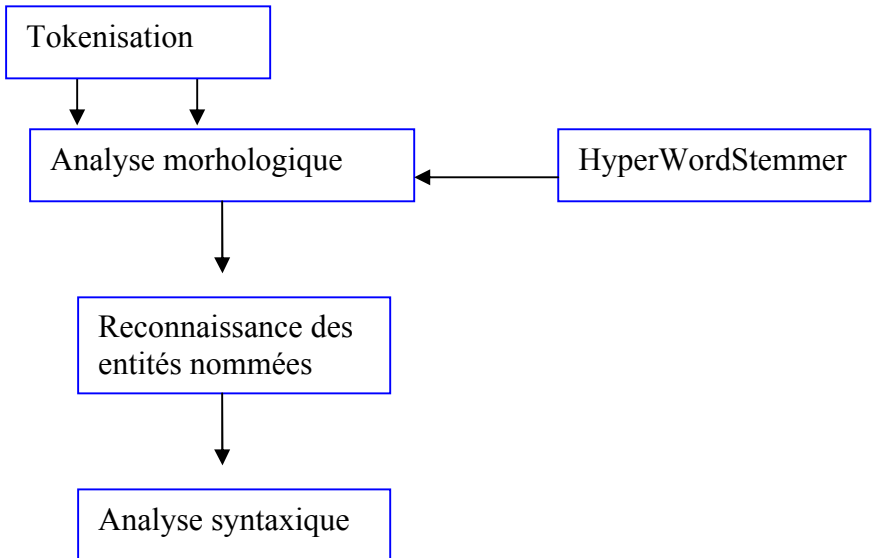


Figure 1: Architecture du système de recherche d'informaion crosslingue du LIC2M

3. L'analyse linguistique :

Il s'agit d'une étape fondamentale du système. Les modules qui permettent de faire le traitement linguistique ont tous été développés au départ pour analyser des langues européennes telles que le français et l'anglais. Pour permettre au système de traiter une langue comme l'arabe, nous avons développé un HperWordStemmer capable de traiter les mots agglutinés de l'arabe.

«لإستعمالهم للعقل الإلكتروني في الجزائر»



3.1 L'analyse morphologique

Elle se déroule selon les étapes suivantes :

- La tokenisation permet de séparer les textes en entrée en mots (unités minimale traiter par la suite).
- La recherche dans le dictionnaire 'full form' tente de récupérer les informations linguistiques (lorsqu'elles existent) pour chaque mot repéré dans le texte.
- La recherche des alternatives orthographiques sert à récupérer les accentuations des mots mal orthographiés et les voyelles pour les mots en arabe.
- La reconnaissance des expressions idiomatiques

3.2 L'analyse syntaxique qui consiste à reconnaître les relations de dépendance entre les mots et entre les syntagmes.

3.3 La reconnaître des entités nommées pour les noms de lieux, de personnes, de mesures, ...

3.4 La recherche d'information: une fois l'indexe des documents constitué, la requête est reformulée, ce qui permet une interrogation monolingue et crosslingue des documents.

Notre système a été développé au départ pour des langues européennes telles que le français, l'anglais, l'allemand, l'espagnol, ... L'ajout de l'arabe a nécessité un certain nombre d'adaptations.

4.L'adaptation du système pour le traitement de l'arabe

Les changements apportés au système ont concerné essentiellement deux parties : les ressources lexicales et l'HyperWordStemmer.

4.1 Les ressources lexicales

- Les entrées lexicales dans le dictionnaire arabe sont toutes voyellées et à chaque entrée voyellée correspond une entrée non voyellée

- Le système est doté de dictionnaires de proclitiques et d'enclitiques.

4.2 L'HyperWordStemmer

Ce module a été développé pour traiter le phénomène de l'agglutination. Chaque mot non trouvé dans le dictionnaire des formes subit un traitement pour séparer les enclitiques et proclitiques pour récupérer la forme originale du mot. Lorsque cette forme a été altérée par les ajouts des proclitiques et des enclitiques, le système est capable de récupérer la forme d'origine.

ب + كرت + هم \ → بكرتهم (avec leur ballon)
ت → ة
كرت → كرة
هم + هوى + و → وهواهم
هما + في → فيهما

Adaptation d'un système de recherche d'information multilingue pour le traitement de l'arabe

5. Application : le prototype ALMA

URL : <http://alma.oieau.fr/alma>

Ce prototype a été développé lors du projet ALMA en 2004, il est consultable en ligne. La base est constituée de 50 documents en relation avec le traitement et la gestion des ressources en eau ...

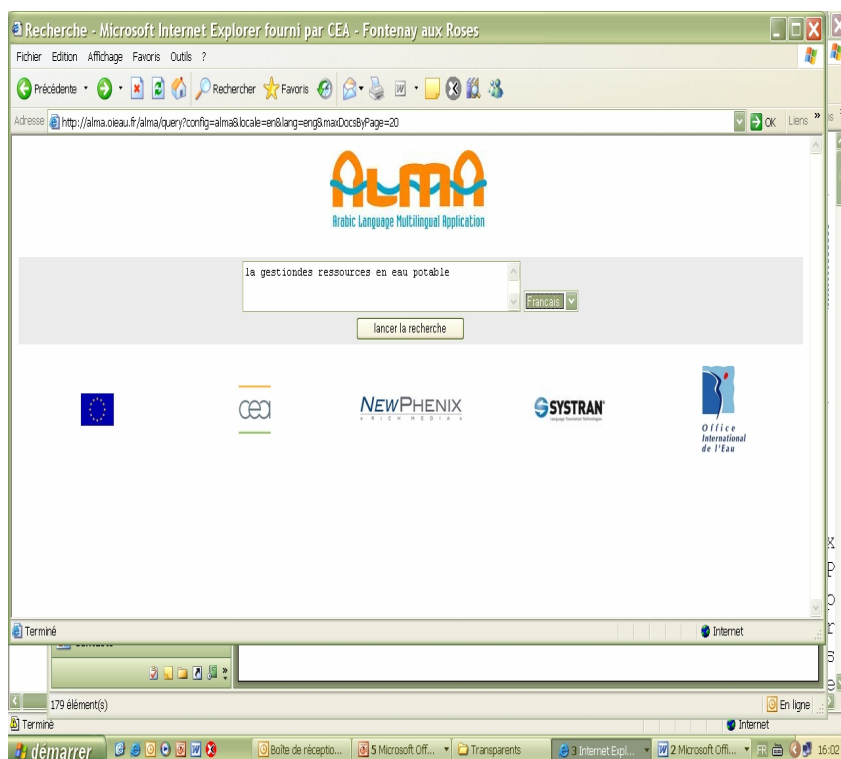


Figure 2: Requête en langage naturelle

Adaptation d'un système de recherche d'information multilingue pour le traitement de l'arabe

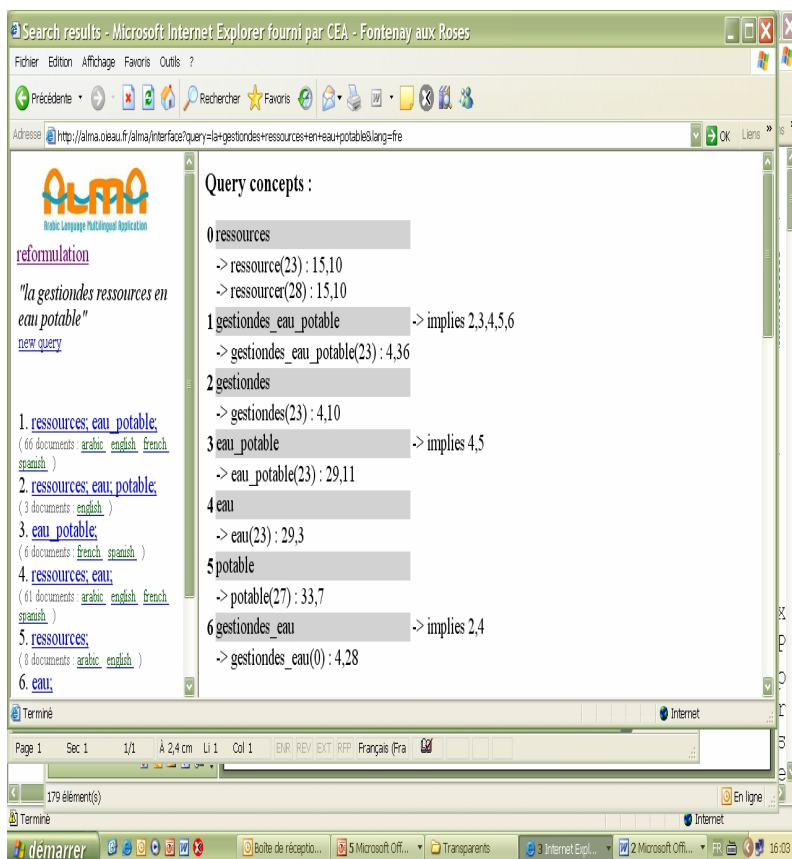


Figure 2: mots composés et reformulation de la question

Adaptation d'un système de recherche d'information multilingue pour le traitement de l'arabe

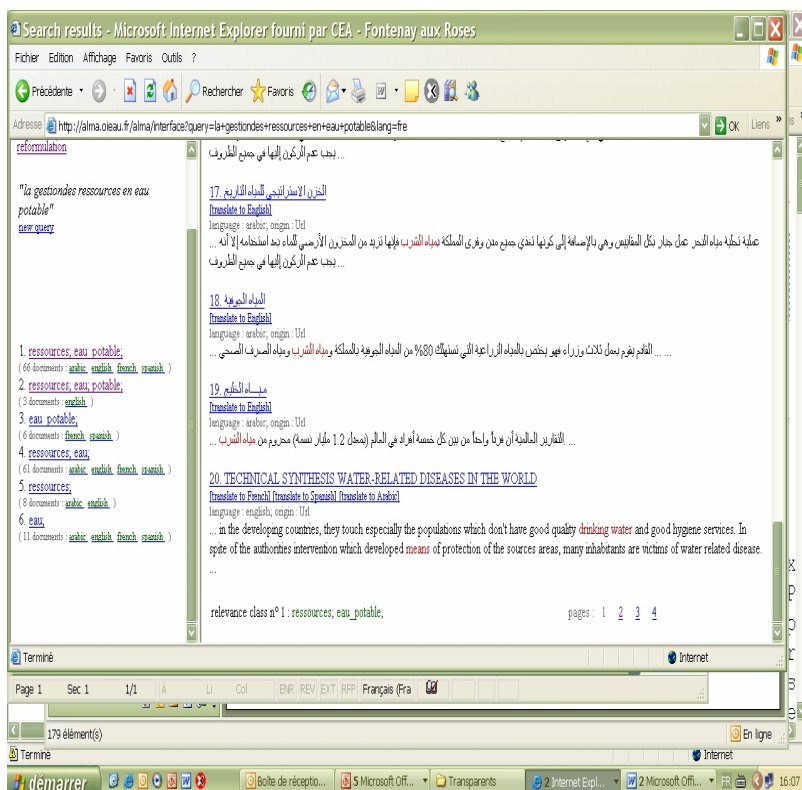


Figure 2: documents pertinents retournés en réponse

Adaptation d'un système de recherche d'information multilingue pour le traitement de l'arabe



Figure 2: visualisation du résultat dans un document pertinent

6. Conclusion et perspectives

- Réutilisabilité: appliquer la méthode de segmentation de l'arabe sur d'autres langues (espagnol, italien, hongrois, hébreu,...)
- Améliorer les ressources linguistiques de l'arabe : dictionnaire de reformulation monolingue pour augmenter le rappel
- Ajouter un traitement sémantique pour améliorer la précision