Démonstration de *Kawâkib*, outil permettant d'assurer le feedback entre grammaire et corpus arabe pour l'élaboration d'un modèle théorique

André Jaccarini[1], Christian Gaubert[2]
(1) MMSH, CNRS, 5 rue du Château de l'Horloge, 13094 Aix-en-Provence
(2) IFAO, 37 rue Cheikh Aly Yousef, Qasr al Ayni, Le Caire, Egypte
jaccarini@mmsh.univ-aix.fr, cgaubert@ifao.egnet.net

Résumé. Kawâkib est un outil assurant le feedback entre corpus arabe et grammaire. Ce logiciel interactif en ligne démontre le bien fondé de la méthode de variation des grammaires arabes pour l'obtention de l'algorithme optimal tant au niveau de l'analyse morphologique, cruciale étant donnée la structure du système sémitique, que syntaxique ou dans le domaine de la recherche de critères pertinents et discriminants pour le filtrage des textes.

Abstract. Kawâkib is a tool allowing feedback between arabic corpus and grammar. As far as methodology is concerned, this interactive online software implements and illustrates the grammar variation method that aims to determine the optimal algorithm, either for morphology – which is essential in semitic languages - or for syntax. The software also permits the search for criteria for text filtering.

Mots-clés : arabe, automates, analyseurs, opérateurs linguistiques, mots-outils, filtrage de corpus.
Keywords: arabic, automata, parsers, lingistic operators, tool words, corpus filtering.

Kawâkib est un outil assurant le feedback entre corpus arabe et grammaire. Sur le plan méthodologique, ce logiciel interactif en ligne démontre le bien fondé de la méthode de variation des grammaires arabes en vue de l'obtention de l'algorithme optimal tant au niveau de l'analyse morphologique, cruciale étant donnée la structure du système sémitique, que syntaxique ou dans le domaine de la recherche de critères pertinents et discriminants pour le filtrage des textes. Kawâkib contient :

- 1 - une bibliothèque d'automates exprimant des règles morphosyntaxiques et des opérateurs de détection des relations discursives
- 2 - des fonctions de "radiographie linguisitique" attribuant des valeurs numériques à des textes à tout venant en vue du filtrage

Ce logiciel doit permettre au chercheur de spécifier lui-même sa grammaire ou de modifier les grands schèmes d'automates proposés par le système, de les instancier, de les mettre en œuvre pour les modifier ensuite rétroactivement en fonction des résultats.

Cette démonstration a pour objectif de montrer :

- a - la richesse des ressources linguistiques
- b - le souplesse de leur utilisation
- c - la fécondité du feedback entre modèle théorique et implémentation

Les fonctionnalités linguistiques immédiatement accessibles à l'utilisateur sont les suivantes :

- Analyses morphologiques des noms et verbes (trilitères et quadrilitères) par automates à états finis et transducteurs, avec ressources linguistiques minimales
- Mise en évidence des ambiguïtés et de leur hiérarchie
- Analyse/recherche des opérateurs tokens (associés aux mots-outils) organisés en 24 catégories
- Statistiques pour ces tokens et leurs catégories
- Repérage et stockage de combinaisons et séries de tokens
- Recherche de mots dans un texte et analyse du contexte
- Recherche de racines ou motif de racines (R1 X R3 par exemple)
- Recherche des racines les plus fréquentes et calcul du seuil de couverture
- Recherche de répétition

- Recherche d'indice de temporalité
- Recherche de discrours raportés
- Recherche de différentes relations discursives liée aux token
- Recherche de motifs combinant mots, tokens, opérateurs logiques par expressions régulières et automates
- Outils pour la synthèse de fragments de grammaires à partir de ces résultats fournis par nos parseurs de segments de phrases.

Ces fonction peuvent être appliquées à des textes tout venants ou importés dans le corpus du logiciel, qui accepte les documents arabes de tout format (doc, docs, rtf, pdf, etc.) pour en extraire le texte.

La démonstration repose sur le déroulé de scénarios d'interaction grammaire corpus.

Kawâkib, analyse du contexte du token *mā*