

Une application de recherche par racines sur le Web développée sur la plate-forme PHARAS

Mourad LOUKAM* , Amar BALLA ** et Mohamed Tayeb LASKRI***

* Département d'informatique, Faculté des sciences et sciences de l'ingénieur, Université Hassiba Benbouali , Chlef, Algérie. Email: loukam@hotmail.com

** Ecole supérieure d'informatique (ex-INI), Oued Smar, Alger, Algérie. Email: a_ball@ini.dz

*** Laboratoire de recherche en informatique, Université Badji Mokhtar, Annaba, Algérie. Email: laskri@univ-annaba.org

Résumé— Dans ce papier, nous présentons une application de recherche par racines sur le Web. Il s'agit d'un outil composé d'un analyseur morpho-lexical de l'arabe standard, d'un générateur de formes fléchies et formes dérivées et d'un module de recherche sur internet.

L'objectif de ce travail est de permettre une recherche avancée sur le web essentiellement dominée actuellement par le concept de « mot clé ». Notre outil opère une expansion de la requête de recherche en faisant d'abord une extraction de la racine du mot clé et ensuite en ajoutant à la requête toutes les formes dérivées et fléchies du mot clé à rechercher.

Mots clés — Analyse morphologique, Arabe standard, Internet, Racine, TALN.

I. INTRODUCTION

La plupart des moteurs sur le web classique se basent sur des techniques de « matching » opérant sur des mots clés. C'est-à-dire qu'ils ne font que rechercher un mot clé, représenté par une chaîne de caractères, dans les pages d'index du web. Par exemple, en soumettant le mot clé « العربية » à l'un de ces moteurs, celui-ci repérera uniquement les pages web contenant exactement cette chaîne de caractères, mais il ignorera les autres pages contenant les autres déclinaisons possibles du mot, et qui peuvent être intéressantes pour la personne ayant lancé la recherche, comme : العرب, العربي, الأعراب, التعريب... etc.

L'objectif de notre travail était justement d'apporter cette nouvelle possibilité à la recherche sur le web. En clair, le système développé opère une extension de la requête initiale en ajoutant les formes fléchies et dérivées (figure1).

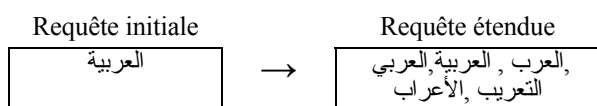


Figure 1. Objectif : expansion de la requête en ajoutant les formes fléchies et dérivées

Pour développer ce système de recherche basée sur la racinisation et utilisation la flexion et la dérivation, nous avons eu recours à la plate-forme PHARAS (Plate-forme d'analyse pour l'arabe standard, basée sur le formalisme HPSG) [1]. Cette plate-forme, actuellement en validation, est constituée d'une chaîne de modules d'analyse,

morpho-lexical et syntaxique notamment, reprenant les concepts de base du formalisme HPSG (structures de traits, principes, règles, ...). Pour développer notre application de recherche par racines, il fallait travailler principalement sur les outils suivants :

Un analyseur morpho-lexical qui doit reconstituer la racine du mot clé introduit.

Un générateur qui doit opérer une expansion de la requête en y ajoutant les formes fléchies et dérivées du mot.

Un module de recherche proprement dite sur Internet

II. TRAVAUX CONNEXES

Plusieurs travaux ont été faits dans le domaine de la morphologie de l'arabe, particulièrement depuis la fin des années 1990. Les objectifs étaient différents et les méthodes proposées aussi. Nous citons, notamment : l'utilisation de la théorie des automates à états finis [2] et les réseaux de transitions augmentés ATN [3]. Des méthodes hybrides, alliant à la fois une représentation « classique » des unités lexicales (dictionnaires, lexiques, ...etc) avec des méthodes de représentation des connaissances ont vu le jour. Nous citons particulièrement l'analyseur développé par Tim Buckwalter pour le compte du Linguistic Data Consortium (LDC) [4], l'outil développé par Xerox permettant l'analyse en ligne de mots en arabe [5][6] et l'outil développé par Systran [7][8] réalisé dans le but d'une traduction anglais/arabe.

La plupart de ces travaux utilisent, plus ou moins, les caractéristiques importantes qui font la remarquable richesse du lexique de l'arabe, à savoir les notions de racine et de schème. Selon notre point de vue, le travail qui a le plus tiré profit de l'organisation du lexique arabe en racine/schème est celui qui a été réalisé sous l'égide de l'ALECSO qui a consisté à mettre en place un système de génération des formes fléchies et dérivées [9]. Les auteurs de ce travail ont répertorié la plupart des types de racines de l'arabe, il en ressort que les racines trilitères sont majoritaires (plus de 60%) (voir table 1). Les auteurs affirment que leur système est capable de générer les formes fléchies de plus de 18.000 verbes et structures dérivables.

TABLE I.
FREQUENCE DES TYPES DE RACINES DE LA LANGUE ARABE

Racine	Nombre	Taux
Deux éléments	115	01.33 %
Trois éléments	7198	63.43%
Quatre éléments	2739	32.95%
Cinq éléments	295	02.29%

Malgré son importance, le travail réalisé [9] ne constitue en fait qu'une phase, celle de la génération, de notre application de recherche par racines. En effet, dans notre système, le traitement commence d'abord par une phase d'analyse, celle de la racinisation, pour extraire la racine d'un mot quelconque entré, ensuite générer toutes ses formes fléchies et dérivées.

DIINAR (Dictionnaire INformatisé de l'ARabe) [10] est un autre travail qui mérite d'être cité parmi les travaux connexes. Cependant il s'agit plutôt d'une base de données de ressources lexicales que d'un logiciel d'analyse. Cette base contient près de 120.000 lemmes différents (verbes, adjectifs, noms, ...etc).

III. ARCHITECTURE GENERALE DE L'APPLICATION

La figure suivante résume l'architecture globale de notre application de recherche.

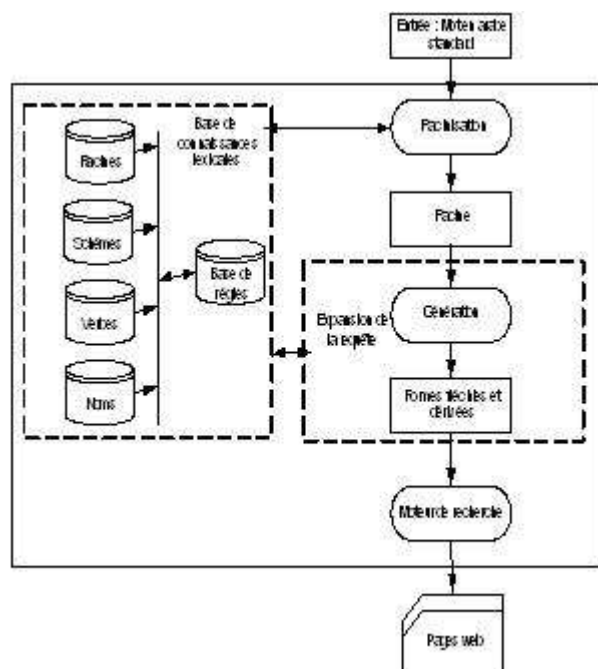


Figure 2. Architecture générale de l'application de recherche

A travers ce schéma, on peut comprendre que sommairement notre système reçoit en entrée un mot écrit en arabe standard, extrait sa racine, produit ses formes fléchies et dérivées à partir de la racine, fait une expansion de la requête et lance la recherche sur internet.

Le système s'articule autour d'un système expert constitué d'une base de connaissances morphologiques et d'un moteur d'inférence. Ce principe est repris de nos travaux précédents [1][11].

Dans la section suivante on fera une description succincte de chacune des étapes de fonctionnement du système.

Entrée du système : Correspond à un mot écrit en Arabe standard non voyellé.

Le module d'analyse (Racinisation): on applique une lemmatisation légère qui consiste à déceler si des préfixes ou suffixes ont été ajoutés au mot. Pour extraire la racine d'un mot, il faut connaître le schéma par lequel il a été dérivé et supprimer les éléments flexionnels (préfixes, suffixes).

Le module de génération : se base sur le système dérivationnel très riche de l'arabe standard. Les éléments et les opérations de base qui interviennent dans la formation des mots sont : les racines, les préfixes qui n'intervient que dans la conjugaison des verbes à l'accompli, l'inaccompli et à l'impératif, les suffixes qui interviennent dans la conjugaison des verbes (inaccompli, accompli, impératif) et dans la déclinaison des noms et enfin les schémas.

Le générateur morphologique réalise deux fonctions principales: la conjugaison (pour les verbes) et la dérivation (pour les noms).

Le conjugaison (flexion verbale), réalisée par un module, a pour rôle de conjuguer tous les verbes appartenant aux différentes classes (verbes sains الأفعال الصحيحة, verbes faibles الأفعال المعتلة, ...). La génération se fait à partir des racines, schémas et en utilisant les règles de flexion spécifiques à chaque verbe.

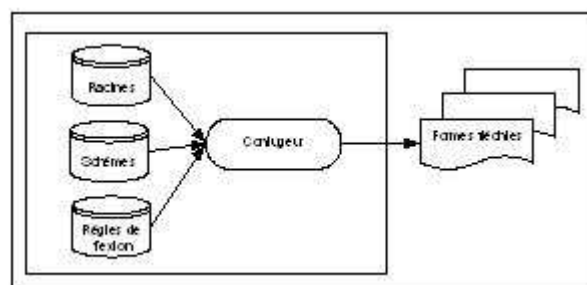


Figure 3. Fonctionnement du conjugueur (générateur de formes fléchies)

A titre d'exemple, la table suivante illustre la conjugaison du verbe sain "فعل" à l'accompli, à l'inaccompli et à l'impératif produite par le générateur.

TABLE II.
EXEMPLE DE GENERATION DE FORMES FLECHIES : VERBE SAIN فعل

		accompli	inaccompli	impératif
1ère pers.	Sing	فعلت	أفعل	
	Plur	فعلنا	تفعل	
2ème pers.	Sing/M.	فعلت	تفعل	افعل
	Sing/F.	فعلت	تفعلين	افعلي
	Duel/M.	فعلتما	تفعلان	افعلا
	Duel/F.	فعلتما	تفعلان	افعلا
	Plur/M.	فعلتم	تفعلون	افعلوا
	Plur/F.	فعلتن	تفعلن	افعلن
3ème pers.	Sing/M.	فعل	يفعل	
	Sing/F.	فعلت	تفعل	
	Duel/M.	فعلا	يفعلان	
	Duel/F.	فعلتا	تفعلان	
	Plur/M.	فعلوا	يفعلون	
	Plur/F.	فعلن	يفعلن	

A l'état actuel, les différentes formes de verbes traités dans notre système sont:

TABLE III.
CLASSES DE VERBES COUVERTS

Classe	Sous-classe
verbes sains (الأفعال الصحيحة)	verbes normaux (الأفعال السالمة).
	verbes sourds (الأفعال المضاعفة).
	verbes hamzes (الأفعال المهموزة).
verbes faibles (الأفعال المعتلة)	verbes assimilés (الأفعال المثالية)
	verbes concaves (الأفعال الجوفاء)
	verbes manquants ou défectueux (الأفعال الناقصة).
	verbes dits المفروق اللفيف
	verbes dits المقرون اللفيف

La classe des verbes sains (الأفعال الصحيحة) est formée de trois sous classes

les verbes normaux (الأفعال السالمة). Ce sont des verbes dans lesquels la Hamza ne constitue pas une lettre radicale et deux lettres radicales identiques ne peuvent se suivre, par Exemple : (كتب).

les verbes sourds (الأفعال المضاعفة). Ce sont des verbes dans lesquels l'une des lettres radicales est doublée. Par Exemple: (عد).

les verbes hamzes (الأفعال المهموزة). Ce sont des verbes où la Hamza constitue l'une des lettres radicales. Par Exemple : (قرأ).

La classe des verbes faibles (الأفعال المعتلة) regroupe cinq sous classes qui sont :

les verbes assimilés (الأفعال المثالية). Ce sont des verbes dans lesquels la première lettre radicale est faible (ا، و، ي). Ils ont appelés ainsi parce qu'ils sont assimilés au verbe sain et se conjuguent de la même manière qu'eux à l'accompli actif et passif. Par Exemple : (ورث), assimilé waw) et (ينع, assimilé ya).

les verbes concaves (الأفعال الجوفاء). Le verbe concave (creux) est un verbe dont le 2ème radical est (و) ou (ي), il est ainsi appelé parce que la lettre faible se trouve au milieu. Dans ce type de verbe (و) et (ي) se changent en alif (ا) à la 3ème personne du masculin (singulier, duel ou pluriel) et à la 3ème personne du féminin (singulier et duel) à l'accompli. Exemple : باع ، قال ، جوف

les verbes manquants ou défectueux (الأفعال الناقصة). Un verbe défectueux est un verbe dans lequel la dernière lettre radicale est faible, la lettre «ا» ou «ي», est à l'origine «ي» ou «و». Tel que : خشى سعى، دعا.(رمى) :

les verbes dits المفروق اللفيف. Si la première (ف) et la troisième (ل) lettre radicale sont faibles, on dit qu'il est المفروق اللفيف, par exemple : وشي وقي .

les verbes dits المقرون اللفيف. Si la première (ف) et la deuxième (ع) ou la deuxième (ع) et troisième (ل) lettre radicale sont faibles, on dit qu'il est المقرون اللفيف, par exemple : طوى، أوى، عوى .

Pour chaque modèle de verbe, des règles morphologiques appropriées ont été conçues. Ces règles ont été reconstituées à partir des manuels de la morphologie de l'arabe standard .

La génération concerne aussi la dérivation. Il s'agit de générer automatiquement toutes les formes dérivées verbales et nominales à partir de la racine (figure ...). L'opération utilisée est l'instanciation de la racine aux différents schémas, en utilisant les règles morphologiques de la dérivation..

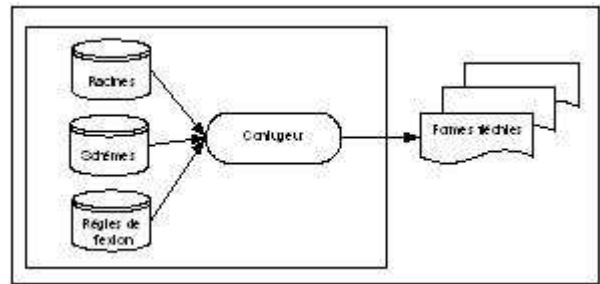


Figure 4. Fonctionnement du dérivateur (générateur de formes dérivées)

Exemple: Voici un exemple de dérivation du verbe كتب, qui donne toutes les dérivations possibles, après l'instanciation de la racine كتب aux différents schémas.

TABLE IV.
EXEMPLE DE DERIVATION A PARTIR DU VERBE كتب

schème	فعل	أفعل	فعال	فعليل	مفعلة
Formes dérivées	كتب	أكتب	كتاب	كتيب	مكتبة

IV. TESTS ET EVALUATION

Dans cette section, nous donnons un exemple réalisé par notre application de recherche par racines. Soit en effet à rechercher sur internet le mot (verbe trilitère) 'تسمعون'.

L'analyse aboutira aux résultats suivants, dont notamment la racine سمع et le schème فعل.

Résultats de l'analyse du mot 'تسمعون'

Racine	schème	Type	Temps	mot
سمع	فعل	سالم	المضارع	تسمعون

La génération donnera toutes les déclinaisons possibles de la racine trouvée. Elles sont contenues dans le tableau suivant :

TABLE V.
RESULTATS : FORMES FLECHIES

الأمر	المضارع	الماضي	الضمائر
	أسمع	سمعت	أنا
	نسمع	سمعنا	نحن
اسمع	تسمع	سمعت	أنت
اسمعي	تسمعي	سمعت	أنت
اسمعا	تسمعان	سمعتما	انتما
اسمعوا	تسمعون	سمعتم	انتم
اسمعن	تسمعن	سمعتن	انتن
	يسمع	سمع	هو
	تسمع	سمعت	هي
	يسمعان	سمعا	هما للمذكر
	تسمعان	سمعتا	هما للمؤنث
	يسمعون	سمعوا	هم
	تسمعن	سمعن	هن

TABLE VI.
RESULTATS : FORMES DERIVEES

اسم الفاعل	الكلمة	مذكر	مؤنث
الرفع	مفرد	سماع	سماعة
	مثنى	سماعان	سماعتان
	جمع	سماعون	سماعات
ال نصب	مفرد	سماعا	سماعة
	مثنى	سماعين	سماعتين
	جمع	سماعين	سماعات
الجر	مفرد	سماع	سماعة
	مثنى	سماعين	سماعتين
	جمع	سماعين	سماعات

Enfin le lancement de la recherche sur Internet . Le script de recherche est généré automatiquement par l'application. En utilisant le moteur de recherche Google , par exemple, le résultat de la page générée est ouvert automatiquement sur une page du navigateur web (figure 5).



Figure 5. Exemple de recherche avec expansion de la requête du mot : تسمعون

Examinons un autre exemple, il s'agit du mot à racine quadrilitère : مدرج

L'analyse aboutira à la racine مدرج et le schème فعل

Résultats de l'analyse du mot مدرج

Racine	schème	Type	mot
مدرج	فعل	اسم فاعل	مدرج

La dérivation aboutira aux déclinaisons suivantes :

الأمر	المضارع	الماضي	الضمائر
	أدرج	درجت	أنا
	ندرج	درجنا	نحن
درج	تدرج	درجت	أنت
درجي	تدرجي	درجت	أنت
درجا	تدرجان	درجتما	انتما
درجوا	تدرجون	درجتهم	انتم
درجن	تدرجن	درجتن	انتن
	يدرج	درج	هو
	تدرج	درجت	هي
	يدرجان	درجا	هما للمذكر
	تدرجان	درجتا	هما للمؤنث
	يدرجون	درجوا	هم
	تدرجن	درجن	هن

Noms dérivés :

اسم الفاعل	الكلمة	مذكر	مؤنث
الرفع	مفرد	مدرج	مدرجة
	مثنى	مدرجان	مدرجتان
	جمع	مدرجون	مدرجات
النصب	مفرد	مدرج	مدرجة
	مثنى	مدرجين	مدرجتين
	جمع	مدرجين	مدرجات
الجر	مفرد	مدرج	مدرجة
	مثنى	مدرجين	مدرجتين
	جمع	مدرجين	مدرجات

Enfin recherche sur internet donne le résultat suivant.



Figure 6. Exemple de recherche avec expansion de la requête du mot :
مدرج

V. CONCLUSION ET PERSPECTIVES

Nous avons présenté dans cet article une application d'analyse morpho-lexicale pour l'arabe standard développée dans l'optique d'une recherche par racines sur Internet. L'application reçoit en entrée un mot clé en arabe standard, et doit lancer sur le web la recherche de toutes les déclinaisons possibles (formes fléchies et dérivées) de ce mot.

La recherche sur le web n'est qu'une application possible de l'outil de racinisation que comporte cette application. En fait, la racinisation est un élément central et incontournable dans toute application de TALN s'intéressant d'une manière ou d'une autre au lexique de l'arabe standard, qui est fondé sur les concepts de racine et schème.

Plusieurs perspectives s'offrent à notre travail, nous pouvons citer entre autres :

Compléter le travail de l'analyseur en augmentant la couverture des phénomènes morphologiques à étudier.

Traitement des mots clés ayant un statut spécifique : mot composé, nom propre, ...etc.

REFERENCES

- [1] M.Loukam , « PHARAS : Une plateforme d'analyse basée sur le formalisme HPSG pour l'arabe standard », Actes du premier séminaire sur le langage naturel et l'intelligence artificielle LANIA'2007, 20/21 Novembre 2007, Chlef/Algérie, p 31-40.
- [2] A.Freeman , Brill's POS tagger and a morphology parser for Arabic. In ACL'2001 Workshop on Arabic language processing, 6 Juillet 2001, Toulouse, France.
- [3] E.Othmane, K.Shaalan and A.Rafea, A Chart Parser for Analyzing Modern Standard Arabic, Machine Translation for Semitic Languages: Issues and Approaches, September 23, 2003 New Orleans, Louisiana, U.S.A.
- [4] T.Buckwalter, Buckwalter Arabic Morphological Analyzer, Linguistic Data Consortium, Philadelphia, 2002.
- [5] K.Beasley. Arabic finite-State Morphological Analysis & Generation, Proceedings of COLING'96, August 5-9, 1996, Copenhagen. pp.89-94
- [6] K.Beasley. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001, Rapport de recherche, Xerox Research Centre Europe.
- [7] J.Dichy and A.Farghali, Roots & Patterns vs. Stems plus Grammar-Lexis specifications: on what basis should a multilingual lexical database centred on Arabic be built? , Machine Translation for Semitic Languages: Issues and Approaches, September 23, 2003 New Orleans, Louisiana, U.S.A.
- [8] A.Farghali , I.Senellart., Intuitive Coding of the Arabic Lexicon, Machine Translation for Semitic Languages: Issues and Approaches, September 23, 2003 New Orleans, Louisiana, U.S.A.
- [9] A.Attar., M.Bawab and O.Al Dakkak , Arabic Lexical Database , Damascus University, Syria, 2007.
- [10] J.Dichy and M.Hassoun, The DIINAR.1-« معالي » Arabic Lexical Resource, an outline of contents and methodology. In The ELRA Newsletter, Vol. 10, n°2, April-June 2005 : 5-10.
- [11] M.Loukam, A.Abbache and M.T.Laskri, « Un analyseur morpho-lexical à base de système expert en vue d'une analyse en HPSG », Actes de la conférence Internationale sur le traitement automatique de la langue arabe CITALA'07, 18/19 Juin 2007, Rabat/Maroc, p 159-166.