



N° d'ordre : .....

**UNIVERSITE DE MOHAMED BOUDIAF - M'SILA**  
**FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE**

**Département d'Informatique**

**MEMOIRE de fin d'étude**

**Présenté pour l'obtention du diplôme de MASTER**

**Domaine : Mathématiques et Informatique**

**Filière : Informatique**

**Spécialité : Systèmes d'Informations Avancés**

**Par: Baâli Meriem**

**SUJET**

**Utilisation de la technique des n-grammes Dans  
l'extraction des racines en langue arabe.**

**Soutenu publiquement le : / /2015 devant le jury composé de :**

.....

**Mr .KADRI Said**

.....

.....

**Université de M'sila**

**Université de M'sila**

**Université de M'sila**

**Université de M'sila**

**Président**

**Rapporteur**

**Examineur**

**Examineur**

**Promotion : 2014 /2015**

# Table des matières

|                            |    |
|----------------------------|----|
| INTRODUCTION GENERALE..... | 01 |
|----------------------------|----|

## CHAPITRE I : Les caractéristiques de la langue arabe

|                                                       |    |
|-------------------------------------------------------|----|
| I.1.Introduction.....                                 | 04 |
| I.2. Particularité de la langue arabe .....           | 04 |
| I.2.1. La structure morphologique d'un mot arabe..... | 05 |
| I.2.2. Catégories des mots.....                       | 06 |
| 1) Le verbe.....                                      | 06 |
| 2) Le nom.....                                        | 08 |
| 3) Particule .....                                    | 10 |
| I.2.3. Morphologie arabe .....                        | 10 |
| 1) Le schème.....                                     | 10 |
| 2) La racine .....                                    | 11 |
| a. L'importance de l'extraction de la racine.....     | 11 |
| b. Domaine d'application.....                         | 12 |
| 3) Les affixes.....                                   | 12 |
| 4) Le stem.....                                       | 12 |
| 5) Les mots dérivés .....                             | 13 |
| 6) Les mots outils.....                               | 13 |
| 7) Les diacritiques.....                              | 13 |
| 8) La šhadda.....                                     | 13 |
| 9) Le tanwin.....                                     | 13 |
| I.3.Conclusion .....                                  | 14 |

## CHAPITRE II : Méthodes d'extraction de la racine en langue arabe

|                                                                |    |
|----------------------------------------------------------------|----|
| II.1.Introduction.....                                         | 16 |
| II.2. Processus d'extraction de la racine d'un mot arabe ..... | 16 |
| II.3. Traitements communs .....                                | 17 |

|                                                                                                             |    |
|-------------------------------------------------------------------------------------------------------------|----|
| II.3.1. Normalisation.....                                                                                  | 17 |
| II.4. Les méthodes d'extraction de la racine d'un mot arabe .....                                           | 17 |
| II.4.1. Schéma général de la classification.....                                                            | 17 |
| II.4.1.1. Classe de l'analyse morphologique.....                                                            | 19 |
| II.4.1.1.1. Suppression des affixes .....                                                                   | 19 |
| II.4.1.1.1.1. Suppression des affixes avec dictionnaire.....                                                | 19 |
| II.4.1.1.1.2. Suppression des affixes sans dictionnaire .....                                               | 20 |
| a) Lemmatiseur léger .....                                                                                  | 20 |
| b) Lemmatiseur du Saint Qur'an.....                                                                         | 21 |
| c) Lemmatiseur linguistique .....                                                                           | 21 |
| II.4.1.1.2. Reconnaissance des modèles et traitement des affixes.....                                       | 22 |
| II.4.1.1.2.1. Reconnaissance des modèles et traitement des affixes<br>sans dictionnaire.....                | 22 |
| a) Lemmatiseur computationnel.....                                                                          | 22 |
| b) Lemmatiseur sans dictionnaire.....                                                                       | 23 |
| c) Lemmatiseur des pluriels irréguliers sans dictionnaire.....                                              | 24 |
| d) Lemmatiseur léger transcrit.....                                                                         | 24 |
| II.4.1.1.2.2. Reconnaissance des modèles et traitement des affixes<br>avec Dictionnaire.....                | 27 |
| a) Lemmatiseur de Khoja.....                                                                                | 27 |
| b) Lemmatiseur des pluriels irréguliers avec dictionnaire.....                                              | 27 |
| II.4.1.1.2.3. Reconnaissance des modèles et traitement des affixes avec<br>règles et sans dictionnaire..... | 28 |
| II.4.1.1.2.4. Reconnaissance des modèles et traitement des affixes avec<br>règles et dictionnaire .....     | 28 |
| a) Lemmatiseur Xerox .....                                                                                  | 28 |
| b) Lemmatiseur par génération systématique .....                                                            | 29 |
| II.4.1.1.3. La traduction.....                                                                              | 30 |
| II.4.1.2. Classe de l'analyse statistique.....                                                              | 31 |
| II.4.1.2.1 Calcul des coefficients de la ressemblance.....                                                  | 31 |
| II.4.1.2.2. Calcul des coefficients de la dissemblance.....                                                 | 32 |
| II.5. Conclusion.....                                                                                       | 33 |

## CHAPITRE III : Réalisation et expérimentation

|                                                                  |           |
|------------------------------------------------------------------|-----------|
| III.1. Introduction .....                                        | 35        |
| III.2. Le langage et l'environnement de programmation .....      | 35        |
| III.2.1. Le langage de programmation (JAVA) .....                | 35        |
| III.2.2 L'environnement de programmation (NetBeans IDE).....     | 35        |
| III.3. Présentation des corpus utilisés .....                    | 35        |
| III.4. La technique utilisée pour la représentation du mot.....  | 36        |
| III.5. Présentation de la nouvelle méthode proposée .....        | 36        |
| III.6. Présentation du processus d'extraction de la racine ..... | 38        |
| IV.7. Présentation de système d'extraction de la racine .....    | 38        |
| IV.7.1. Interface principale .....                               | 38        |
| IV.7.2. Prétraitements sur un texte .....                        | 39        |
| IV.7.3. Traitements sur un texte .....                           | 41        |
| III.8. Évaluation des résultats.....                             | 41        |
| III.8.1. Évaluation de méthode de Yousef et al .....             | 42        |
| III.8.2. Évaluation de nouvelle méthode proposée .....           | 42        |
| III.8.3. Interprétation des résultats obtenus.....               | 42        |
| IV.9.Conclusion.....                                             | 42        |
| <b>CONCLUSION GENERALE.....</b>                                  | <b>43</b> |
| <b>Listes des figures .....</b>                                  | <b>45</b> |
| <b>Listes des tables.....</b>                                    | <b>47</b> |
| <b>Annexe.....</b>                                               | <b>49</b> |
| <b>Bibliographie et webographie.....</b>                         | <b>53</b> |

## INTRODUCTION GENERALE

L'arabe est une langue qui a été particulièrement marquée par l'histoire depuis son apparition au deuxième siècle, elle fait partie de la famille des langues sémitiques. Avec plus de 300 million de locuteurs. L'arabe se positionne à la quatrième place mondiale des langues les plus parlées. Dotée d'une richesse morphologique dû principalement à ses propriétés de dérivation et de flexion forte, tel que à partir d'une racine trilitère (majoritairement) ou quadrilitère on peut créer un nombre important de mot au moyennant des schèmes, comme c'est le cas de la racine "(K+T+B) "كتب à partir de laquelle on peut générer 16 mots représentant 9 catégories grammaticales déférentes.

Il y a beaucoup d'applications informatiques de langue arabe comptent sur l'aide des racines de mots, tels que les systèmes de recherche d'information, classification de textes, résumé de texte, auto-traduction, Data mining, et autres applications. Par exemple La performance de la recherche d'information en langue arabe reste problématique à cause des caractéristiques structurelles de la langue et parce que le problème a fait l'objet de beaucoup moins de travaux que pour l'anglais En effet, les moteurs de recherche les plus utilisés et les plus célèbres, comme Google et Yahoo! n'ont pas les mêmes performances en arabe et dans les langues occidentales. Cette différence est due probablement au fait que le fonctionnement d'un moteur dépend de la nature et de la complexité de la langue traitée. Cette spécificité réside aux niveaux morphologiques et structurels de l'arabe par rapport aux langues occidentales. En effet, l'analyse morphologique d'un mot arabe consiste principalement à déterminer sa racine. Donc, nos études menées sur la méthode d'extraction de la racine d'un mot arabe contribuent à identifier les problèmes et à améliorer les performances de ces moteurs en langue arabe.

Aujourd'hui l'étape d'extraction de la racine est considérée comme un point de passage obligatoire dans le processus de développement d'une application TALN et son importance devient capitale pour les langues dotées d'une richesse morphologique (dû principalement au processus de dérivation et de flexion), comme c'est le cas pour l'arabe.

Si cette richesse morphologique est perçue par les linguistes comme un point fort, cette vision n'est pas partagée par les chercheurs qui travaillent sur l'automatisation de la langue arabe elle est même vue comme le point névralgique de plusieurs problèmes comme l'ambiguïté (que se soient morphologique, syntaxique et même sémantique). Ce qui rend le

processus d'extraction de la racine d'une importance primordiale dans la phase d'analyse (reconnaissance de l'appartenance du mot à la langue).

Notre travail se centralise sur l'extraction de racine du mot sans l'aide de règles morphologiques, mais en utilisant la technique des n- grammes de caractères qui simplifie le processus d'extraction des racines. Et cela afin d'éviter toute complexité causée par la richesse morphologique de l'arabe d'une part, et la diversité de ces règles d'une autre part.

Pour l'organisation de notre mémoire, nous proposons le suivant :

- Introduction générale
- Chapitre I : Les caractéristiques de la langue arabe
- Chapitre II : Méthodes d'extraction de la racine en langue arabe
- Chapitre III : Réalisation et expérimentation
- Conclusion générale

## Conclusion générale

Ce mémoire porte sur l'extraction de racine dans le cas de la langue arabe, Cette langue est riche et sensible au contexte mais qui présente des nombreux défis pour des domaines divers tels que la recherche d'informations et le traitement automatique du langage naturel.

Nous avons utilisé dans ce mémoire un nouveau algorithme purement statistique (différente d'une méthode morphologique) qui permet de chercher les racines des mots arabes en basant sur la technique de n-grammes de caractères sans besoin d'utiliser des règles morphologiques de la langue, et cela afin d'éviter toute complexité causée par la richesse morphologique de l'arabe d'une part, et la diversité de ces règles d'une autre part.

Notre travail rencontre plusieurs difficultés lors des différentes étapes de constructions, nous avons souligné les difficultés suivant : le problème de la reconnaissance des formes fléchies puisque la langue arabe étant fortement flexionnelle et aussi le problème d'absence de travaux réalisés sur l'extraction de racine issues de textes arabes en utilisant des méthodes statistiques, le manque d'ouvrages spécialisées dans ce domaine, s'ajoute à tout cela, la non suffisance du temps attribué à ce projet de recherche, tous ces problèmes et autres influence le processus d'extraction et dégrade la qualité des résultats obtenus.

Malgré tous ces difficultés, Ce travail a été très bénéfique à plus d'un point et Il a permis d'apprendre beaucoup de choses qui sont nouvelles pour nous telles que : les techniques et les méthodes d'extraction de la racine, et les outils de programmation en langage Java NetBeans. Mais puisque rien n'est parfait, notre travail pourra subir des améliorations dans des projets futurs, notamment : l'utilisation des règles morphologiques et les techniques d'intelligence artificielle, pour améliorer les résultats préliminaires qui ont émergé après l'extraction de la valeur de similitude.

## BIBLIOGRAPHIE

- [1] Soraya Zaidi–Ayad , Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran), mémoire de doctorat d'informatique, Univ. Badji Mokhtar Annaba, 2013.
- [2] Fouad Soufiane Douzidia, Résumé automatique de texte arabe, Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique, Septembre, 2004.
- [3] ABDELOUAHAB Soumia , Processus de classification supervisée de textes arabes par la méthode KPPV Application aux articles de presse, mémoire de Master En Informatique, Université de M'sila, 2012.
- [4] Aïda KHEMAKHEM, ArabicLDB : une base lexicale normalisée pour la langue arabe, mémoire de master en Systèmes d'Information et Nouvelles Technologies, Université de Sfax Faculté des Sciences Economique et de Gestion ,2006.
- [5] الشيخ مصطفى الغلاييني, جامع الدروس العربية، صيدا بيروت ، ب:8355 ، 1994/1414
- [6] Abd El Salam AL HAJJAR, extraction et gestion de l'information a partir des documents arabes, mémoire de doctorat d'informatique, Université paris VIII – saint denis, Décembre 2010.
- [7] Tahar DILEKH, Implémentation d'un outil d'indexation et de recherche des textes en arabe, mémoire de Magister en système d'information et de connaissance (SIC), Université Hadj Lakhdar – Batna, 2011.
- [8] Mesfar Slim. Analyse morpho-syntaxique automatique et reconnaissance des entites nommees en arabe standard, mémoire de doctorat d'informatique, université de Franche-Comte, 2008.
- [9] Sofiane Baloul, développement d'un système automatique de synthèse de la parole à partir de texte arabe standard voyelle, mémoire de doctorat d'informatique, Université du Maine-France, 2003.
- [10] Lazhar FAREK, Identification d'opinions dans les journaux arabes, mémoire de Magister d'informatique, Badji Mokhtar Annaba, 2009.
- [11] Hajjar, M., Al Hajjar, A., Zreik, K. and Gallinari, P., An Improved Structured and Progressive Electronic Dictionary for the Arabic Language: iSPEDAL. The Fifth International Conference on Internet and Web Applications and Services, ICIW, 2010.

- [12] Kadri, Y. and Nie, J., Effective Stemming for Arabic Information Retrieval. Proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.
- [13] Semmar, N., Elkateb-Gara, F. and Fluhr, C. Using a Stemmer in a Natural Language Processing system to treat Arabic for Cross-language Information Retrieval, Fifth Conference On Language Engineering, Cairo, Egypt, 2005.
- [14] Aljlal, M. and Frieder, O., On arabic search: Improving the retrieval effectiveness via a light stemming approach. In Proceedings of ACM Eleventh Conference on Information and Knowledge Management, CIKM, Mclean, 2002.
- [15] Al Hajjar, A., Hajjar, M., AND Zreik, K. 2009. Classification of Arabic Information Extraction methods, 2nd International Conference on Arabic Language Resources and Tools Cairo (Egypt), p. 71 – 74, 2009.
- [16] Thabet, N., Stemming the Qur'an, WORKSHOP ON Computational Approaches to Arabic Script-based Languages, University of Newcastle, 2004.
- [17] Goweder, A., Poesio, M., De Roeck, A. and Reynolds, J., Identifying Broken Plurals in Unvowelised Arabic Text, Proceedings of EMNLP, pp. 246-253, 2003..
- [18] Beesley, K. R., Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001". In The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France, 2001.
- [19] Yaghi, J. and Yagi, S. Systematic Verb Stem Generation For Arabic. Workshop On Computational Approaches To Arabic Script-Based Languages. 2004.
- [20] Yousef, N. Abu-Errub, A. Odeh, A and Khafajeh, H, An improved arabic word's roots extraction method using n-gram technique, Journal of Computer Science 10 (4): 716-719, 2014.
- [21] Al-Nashashibi, Y. Neagu, A. Yaghi, A, An improved root extraction technique for Arabic words, International Conference on Computer Technology and Development, ICCTD, 2010.
- [22] Khreisat, L., Arabic Text Classification Using N-gram Frequency Statistics a Comparative Study. *Conference on Data Mining | DMN'06* |, pp. 78-82, 2006.
- [23] Hammo, B., Abu-Salem, H., Lytinen, S., AND Evens, M., A Question Answering System to Support the Arabic Language. Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages Philadelphia, Pennsylvania, pp. 1 – 11, 2002.
- [24] Larkey, L., Ballesteros, L., and Connel, M. E., Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis, SIGIR'02 Proceeding of

the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275 - 282, 2002

## WEBOGRAPHIE

[S1] [https://www.java.com/fr/download/faq/whatis\\_java.xml](https://www.java.com/fr/download/faq/whatis_java.xml) Consultée le 21/05/2015.

[S2] <http://www.techno-science.net/?onglet=glossaire&definition=5346> Consultée le 21/05/2015.

### **Abstract :**

The purpose of this Master project is to propose an efficient algorithm which permit to find the roots from Arabic words based on n-grams technique of characters without using morphological rules of Arabic language which are many, variables, and very difficult to understand. And for the implementation step, we opted for the language JAVA NetBeans.

**Keywords:** Extraction from the Arabic root, Treatment of Arabic language, N-grams.

### **Résumé :**

Le but de ce projet de Master est de proposer un algorithme efficace qui permet de trouver les racines des mots arabes en basant sur la technique des n-grammes de caractères sans besoin d'utiliser les règles morphologiques de la langue arabe qui sont généralement nombreuses, variables, et très difficile à maîtriser. Concernant la phase d'implémentation, nous avons opté pour le langage JAVA NetBeans.

**Mots clés :** L'extraction de la racine arabe, traitement de la langue arabe, n-grammes.

### **ملخص :**

الهدف من هذا المشروع الذي يندرج في إطار التحضير للحصول على شهادة الماستر في التخصص هو اقتراح خوارزمية بسيطة وفعالة تمكّن من العثور على جذور الكلمات العربية باستعمال تقنية (N-grams) دون الحاجة إلى استخدام القواعد الصرفية للغة العربية التي عادة ما تكون كثيرة العدد، متغيرة، و صعبة الفهم. أما بالنسبة للغة البرمجة المستعملة فقد وقع اختيارنا على لغة JAVA NetBeans.

**الكلمات المفتاحية :** استخراج الجذور العربية، معالجة اللغة العربية، تقنية N-grams