**Développement d’un moteur de recherche et d’indexation des documents Coraniques**

Assem Chelli

Merouane Dahmani

2009/2010

المدرسة الوطنية العليا للإعلام الآلي

(المعهد الوطني للتكوين في الإعلام الآلي سابقا)

Ecole nationale Supérieure d’informatique

ex. INI (Institut National de formation en Informatique)

**Mémoire de fin d’études**

Pour l’obtention du diplôme d’Ingénieur d’Etat en Informatique

**Option : Systèmes Informatiques**

**Thème**

|  |
| --- |
| **Développement d’un moteur de recherche et d’indexation des documents Coraniques** |

|  |  |
| --- | --- |
| **Réalisé par** | **Encadré par** |
| * Mr CHELLI Assem * Mr DAHMANI Merouane | * Dr BALLA Amar * Mr ZERROUKI Taha |

**Promotion : 2009/201****0**

**Dédicaces**

Louange à Dieu, le seul et unique.

A celle qui m’a porté durant neuf mois,

Qui m'a élevé tout petit, et ‎s'est occupée de moi quand j'ai grandi,  ma chère mère

A mon cher père

Témoignage d’affection et de grande reconnaissance

Que Dieu les garde.

A tous mes frères

Abdellah, Salaheddine et Qalbaddine.

A ma sœur Khadidja.

A tous mes collègues exclus et qui n’ont pas pu terminer leurs études

Lwannas, Omar et les autres dont je n’ai pas cité leurs noms.

A linuxscout, sneetsher, marioBgr et SEband.

A tous mes collègues de l’école et de la cité universitaire.

A tous mes amis et mes proches.

A tous ceux qui m’ont aidé.

A tout musulman.

Avec l’expression de tous mes sentiments de respect,

Je dédie ce modeste travail.

Assem

**Dédicaces**

A mes très chers parents

A mon frère et à mes sœurs

A toute ma famille

A tous mes amis

Et à tous ceux qui m’ont aidé

Je dédie ce modeste travail

Que Dieu les protège tous

Merouane

**Remerciements**

Nous tenons tout d’abord à remercier *DIEU* le tout-puissant qui nous a guidés vers le chemin du savoir et *pour tous ses bienfaits*

Nous adressons nos vifs remerciements à *nos promoteurs* Monsieur *Amar BALLA* et Monsieur *Taha ZERROUKI* de nous avoir proposé ce sujet et de nous avoir encadré tout au long de ce travail.

Nos sincères remerciements s’adressent aux enseignants de l’Ecole nationale Supérieur d’Informatique pour tout le savoir qu’ils ont su nous transmettre durant ces cinq dernières années.

Nous remercions plus particulièrement les membres du jury pour nous avoir honorés en acceptant de juger notre travail.

Nous tenons à remercier également, tous ceux qui ont contribué par leurs conseils ou leurs encouragements à l’aboutissement de ce modeste travail.

Merci à tous et à toutes.

**Résumé :**

Le document coranique est un document de base pour une gamme très riche d'applications autour de Coran, telle que la récitation, la recherche, taffasīr, traduction, etc.

L'existence des différentes applications coraniques exige d'un autre côté le besoin de standardiser le support informatique de Coran et des sciences coraniques. Ces besoins font partie d'un projet global pour la standardisation de support informatique de Coran qui donne naissance à plusieurs sous-projets.

L'indexation et l'interrogation de document coranique font partie de ce projet global.

L'application consiste à implémenter un moteur de recherche pour les documents coraniques, plein d'options de recherche spécifiques pour le texte coranique :

* recherche par mot et par expression régulière ;
* recherche des différentes formes des mots ;
* recherche par racine de mot arabe ;
* recherche par synonyme ;
* recherche par sujet ;
* utilisation des index prédéfinis.

Le travail demandé consiste à concevoir et réaliser un moteur de recherche pour le Coran, qui offre des options de recherche avancées.

Le moteur de recherche doit donner une API (Application Programming Interface) de recherche qui peut être utilisée par les développeurs des applications coraniques.

**Mots-clés :**

*Coran, moteur de recherche, indexation, interrogation.*

**Abstract:**

The Qur’anic document is a document basis for a full range of applications around the Qur’an, such as recitation, research, taffasir, translation, etc.

The existence of different Qur’anic applications requires on the other hand the need to standardize the computer support of the Qur’an and Qur’anic sciences. These needs are part of a comprehensive project for the standardization of computer support giving rise to several sub-projects.

Indexing and querying Qur’anic document are part of this overall project.

The application is to implement a search engine for Qur'anic documents, full of specific search options for Qur'anic text:

* Search by word and regular expression
* Search by different word forms
* Search by Arabic word root
* Search by synonym
* Search by subject
* Search using predefined indexes

The work required is to design and implement a search engine for the Qur'an, which provides advanced search options.

The search engine should provide an API (Application Programming Interface) of search that can be used by developers of Qur'an applications.

Keywords:

[*Quran*](http://code.google.com/hosting/search?q=label:Quran)*,* [*search engine*](http://code.google.com/hosting/search?q=label:searchengine)*,* [*indexing*](http://code.google.com/hosting/search?q=label:indexing)*,* [*querying*](http://code.google.com/hosting/search?q=label:querying)

**ملخّص:**

الوثيقة القرآنية هي وثيقة أساسية لمجموعة زاخرة بتطبيقات حول القرآن الكريم، مثل التلاوة والبحث و التفسير والترجمة وغير ذلك...

وجود تطبيقات قرآنية مختلفة يتطلب من جهة أخرى ضرورة توحيد دعم تكنولوجيا المعلومات للقرآن الكريم والعلوم القرآنية. هذه الاحتياجات هي جزء من مشروع شامل لتوحيد دعم تكنولوجيا المعلومات للقرآن مما يخلق العديد من المشاريع الفرعية. فهرسة و استعلام الوثيقة القرآنية تمثّل جزءا من هذا المشروع الشامل.

التطبيق عبارة عن إنجاز محرك بحث في المستندات القرآنية يوفّر خيارات بحث خاصة بالنص القرآني:

* البحث بالكلمة أو بالعبارة النمطية
* البحث بالأشكال المختلفة للكلمة
* البحث بجذر الكلمة العربية
* البحث بالمترادفات
* البحث في المواضيع
* البحث باستعمال فهارس مُعَدَّة مسبقا

هذا العمل يهدف إلى تصميم وتطوير محرّك بحث للقرآن يوفر خيارات بحث متقدمة و يحتوي على واجهة برمجية للبحث يمكن استعمالها من طرف مطوّري البرامج القرآنية.

**كلمات مفتاحية:**

قرآن، محرّك بحث، فهرسة، استعلام

Table des matières

[Table des matières IX](#_Toc262155033)

[Liste des figures XIII](#_Toc262155034)

[Liste des tableaux XV](#_Toc262155035)

**Introduction Générale 1**

[Chapitre I : Le Coran 5](#_Toc262155036)

[I. Introduction 6](#_Toc262155038)

[II. Définitions 6](#_Toc262155039)

[A. Coran 6](#_Toc262155040)

[B. Document coranique 6](#_Toc262155041)

[III. L'historique du Coran 6](#_Toc262155042)

[A. L’écriture du Coran 6](#_Toc262155043)

[B. Edition du Coran 7](#_Toc262155044)

[C. Informatisation du Coran 7](#_Toc262155045)

[IV. Le Mushaf 8](#_Toc262155046)

[A. Définition 8](#_Toc262155047)

[B. Les informations globales du Mushaf 8](#_Toc262155048)

[C. La structure du mushaf 8](#_Toc262155049)

[D. Les niveaux analytiques de la structure du Coran 10](#_Toc262155055)

[V. Les sciences du Coran 11](#_Toc262155057)

[VI. Les bénéfices de l’informatisation du Coran 15](#_Toc262155058)

[VIII. Les principaux problèmes d’informatisation du Coran 16](#_Toc262155062)

[A. Le script othmani 16](#_Toc262155063)

[B. La vérification et l'authentification du Mushaf 17](#_Toc262155064)

[C. Référence coranique électronique standard 18](#_Toc262155065)

[IX. Conclusion 19](#_Toc262155066)

[Chapitre II :Les moteurs de recherche et le Coran 20](#_Toc262155067)

[I. Introduction 21](#_Toc262155068)

[II. Définition 21](#_Toc262155069)

[III. Les moteurs de recherche 22](#_Toc262155070)

[A. L'exploration (en anglais : crawling) 23](#_Toc262155071)

[B. L'indexation 23](#_Toc262155072)

[C. L’interrogation 23](#_Toc262155073)

[IV. Les fonctionnalités de recherche 24](#_Toc262155074)

[A. La recherche avancée 24](#_Toc262155075)

[1. Recherche par champs (fielded search) 24](#_Toc262155076)

[2. Recherche par une partie de mot 24](#_Toc262155077)

[3. Recherche par les jokers (les caractères spéciaux) 24](#_Toc262155078)

[4. Recherche par les relations logiques 25](#_Toc262155079)

[5. Recherche par une phrase 25](#_Toc262155080)

[6. Recherche par mots proches 25](#_Toc262155081)

[7. Recherche par intervalle 26](#_Toc262155082)

[8. Boosting 26](#_Toc262155083)

[B. Les Fonctionnalités additionnelles 26](#_Toc262155084)

[1. Suggestion de mots-clés alternatifs 26](#_Toc262155085)

[2. Suggestion de mots-clés apparentés 26](#_Toc262155086)

[3. Pagination 27](#_Toc262155087)

[4. Tri (Sorting) 27](#_Toc262155088)

[5. Highlight (sur-lignage) 27](#_Toc262155089)

[C. Les aspects linguistiques Généraux de recherche 27](#_Toc262155090)

[1. La recherche par signification 27](#_Toc262155091)

[2. Recherche par lemme 28](#_Toc262155092)

[3. Recherche par synonymes 28](#_Toc262155093)

[4. Recherche par antonymes 29](#_Toc262155094)

[D. Les aspects linguistiques de recherche pour la langue arabe 29](#_Toc262155095)

[1. La considération de diacritiques الحركات 31](#_Toc262155096)

[2. Les préfixes du mot 32](#_Toc262155097)

[3. Les suffixes du mot 32](#_Toc262155098)

[4. Recherche par la racine arabe 33](#_Toc262155099)

[5. La conjugaison des verbes 33](#_Toc262155100)

[6. La déclinaison des noms 34](#_Toc262155101)

[7. Recherche par les natures des mots arabes 34](#_Toc262155102)

[8. Recherche par les patterns 34](#_Toc262155103)

[9. Respecter les différences de Hamza 35](#_Toc262155104)

[10. Respecter la différence entre Hâ’ et tâ’ marbûtä , entre Yâ’ et Alif maqsûrä 35](#_Toc262155105)

[E. Autres options de recherche dans le Coran 35](#_Toc262155106)

[1. Recherche par le script othmani 35](#_Toc262155107)

[2. Recherche par des index spéciaux 36](#_Toc262155108)

[3. Des options coraniques structurelles 36](#_Toc262155110)

[4. Des options statistiques 36](#_Toc262155111)

[5. Extension de recherche 37](#_Toc262155112)

[V. Un algorithme proposé de la recherche arabe 37](#_Toc262155113)

[VIII. Conclusion 38](#_Toc262155128)

[Chapitre III : Le fonctionnement des moteurs de recherche 39](#_Toc262155129)

[I. Introduction 40](#_Toc262155130)

[II. Définitions 40](#_Toc262155131)

[A. Thésaurus 40](#_Toc262155132)

[B. Mot clé (keyword) 40](#_Toc262155133)

[C. Descripteur (descriptor) 41](#_Toc262155134)

[D. Document 41](#_Toc262155135)

[E. Requête 41](#_Toc262155136)

[F. Relevance 41](#_Toc262155137)

[III. Recherche plein texte 41](#_Toc262155139)

[IV. Exploration 42](#_Toc262155140)

[A. Définition 42](#_Toc262155141)

[B. Les caractéristiques des robots d'exploration 42](#_Toc262155142)

[V. Indexation 43](#_Toc262155146)

[A. Définition 44](#_Toc262155147)

[B. Modes d’indexation 44](#_Toc262155148)

[C. Les index 47](#_Toc262155152)

[1. Différentes formes d’indexes 47](#_Toc262155153)

[D. Mise à jour de l’index 49](#_Toc262155154)

[3. Mise à jour globale 49](#_Toc262155157)

[E. Stockage de l’index 49](#_Toc262155158)

[F. Les étapes d’indexation 49](#_Toc262155159)

[VI. Interrogation 51](#_Toc262155166)

[A. La notion de pertinence 52](#_Toc262155167)

[B. Fonction de similarité 52](#_Toc262155168)

[C. La modélisation de la connaissance 53](#_Toc262155169)

[D. le processus d’interrogation 53](#_Toc262155170)

[VII. Conclusion 54](#_Toc262155171)

[Chapitre IV : Les index de Coran 55](#_Toc262155172)

[I. Introduction 56](#_Toc262155173)

[II. Définitions 56](#_Toc262155174)

[A. Al-mu‛jam 56](#_Toc262155175)

[B. L’index 56](#_Toc262155176)

[III. Historique des index du Coran 56](#_Toc262155177)

[IV. Les index 57](#_Toc262155179)

[A. Par l’unité 58](#_Toc262155180)

[B. Par l’objet 62](#_Toc262155185)

[V. Des projets de construction des index 64](#_Toc262155191)

[A. Midād lbayān 64](#_Toc262155192)

[B. Les Index par M.Taha Zerrouki 65](#_Toc262155196)

[C. The Qur’anic Arabic Corpus 67](#_Toc262155200)

[D. Tanzil Quran Navigator 68](#_Toc262155204)

[VII. Conclusion 70](#_Toc262155208)

[Chapitre V : Conception 71](#_Toc262155209)

[I. Introduction 72](#_Toc262155210)

[II. Le modèle de base 73](#_Toc262155211)

[III. Traitement de texte 76](#_Toc262155212)

[A. Extraction des tokens (Tokenization) 77](#_Toc262155213)

[B. Normalisation 77](#_Toc262155214)

[C. Segmentation (stemming) 78](#_Toc262155215)

[D. Filtrage des stopwords 80](#_Toc262155216)

[IV. Traitement de la requête 81](#_Toc262155217)

[A. Synonymes et antonymes 84](#_Toc262155218)

[E. Dérivations des mots 85](#_Toc262155219)

[F. La considération d'orthographe 86](#_Toc262155220)

[G. La recherche par tuple 86](#_Toc262155221)

[V. Les suggestions 87](#_Toc262155222)

[A. La vérification d’orthographe 87](#_Toc262155223)

[B. Mots-clés relatifs 88](#_Toc262155224)

[VI. Traitement de résultats 88](#_Toc262155225)

[A. Pondération 89](#_Toc262155226)

[B. Tri 89](#_Toc262155227)

[C. Filtrage 89](#_Toc262155228)

[D. Extension 90](#_Toc262155229)

[E. Pagination 90](#_Toc262155230)

[F. Highlight 90](#_Toc262155231)

[VII. L’importation des indexes 91](#_Toc262155232)

[A. Les textes de l’aya 91](#_Toc262155233)

[B. Les informations structurelles 91](#_Toc262155234)

[C. Les informations de la sourate 91](#_Toc262155235)

[D. Les informations statistiques 92](#_Toc262155236)

[E. La classification en sujets 92](#_Toc262155237)

[VIII. L’exploration 92](#_Toc262155238)

[IX. Conclusion 93](#_Toc262155239)

[Chapitre VI : Implementation 94](#_Toc262155240)

[I. Introduction 95](#_Toc262155241)

[II. Python – le langage choisi 95](#_Toc262155242)

[III. Les ressources utilisées 96](#_Toc262155243)

[A. Whoosh : un API d’indexation et de la recherche 96](#_Toc262155244)

[B. Tanzil – Un texte Coranique vérifié 96](#_Toc262155245)

[C. Les Index de M. Zerrouki 97](#_Toc262155246)

[IV. Le modèle de base 98](#_Toc262155247)

[VI. Traitement de texte 98](#_Toc262155249)

[A. Extraction des tokens (Tokenization) 98](#_Toc262155250)

[B. Normalisation 99](#_Toc262155251)

[C. Segmentation (stemming) 99](#_Toc262155252)

[D. Filtrage des stopwords 99](#_Toc262155253)

[VII. Le traitement de la requête 100](#_Toc262155254)

[A. Synonymes et antonymes 100](#_Toc262155255)

[B. Dérivations des mots 101](#_Toc262155256)

[C. La considération d'orthographe 102](#_Toc262155257)

[D. La recherche par tuple 102](#_Toc262155258)

[VIII. Les suggestions 103](#_Toc262155259)

[IX. Le traitement de résultats 104](#_Toc262155260)

[A. Pondération 104](#_Toc262155261)

[B. Tri 104](#_Toc262155262)

[C. Filtrage 104](#_Toc262155263)

[D. Extension 105](#_Toc262155264)

[E. Pagination 106](#_Toc262155265)

[F. Highlight 106](#_Toc262155266)

[X. L’importation des index 107](#_Toc262155267)

[A. Les textes de l’aya 107](#_Toc262155268)

[B. Les informations structurelles 107](#_Toc262155269)

[C. Les informations de la sourate 107](#_Toc262155270)

[D. Les informations statistiques 108](#_Toc262155271)

[E. La classification en sujets 108](#_Toc262155272)

[XII. Les Interfaces 109](#_Toc262155274)

[A. Programming interface (API) 109](#_Toc262155275)

[B. QT Desktop GUI 112](#_Toc262155276)

1. Autres interfaces possibles 116

[XIII. Conclusion 117](#_Toc262155278)

[Conclusion Générale 119](#_Toc262155279)

[Bibliographie 121](#_Toc262155280)

Liste des figures

[Figure 1 : Schéma explicatif de la fragmentation en sourates. 9](#_Toc263084736)

[Figure 2 : Schéma explicatif de la fragmentation en ahzab (أحزاب). 9](#_Toc263084737)

[Figure 3 : Schéma explicatif de la fragmentation en manāzil (منازل) 9](#_Toc263084738)

[Figure 4 : Schéma explicatif de la fragmentation en Ruku'. 10](#_Toc263084739)

[Figure 5 : Les waqfs dans sourate Al-Houmazat en rewayate de Kaloun (برواية قالون عن نافع) 10](#_Toc263084740)

[Figure 6 : Texte coranique écrit en script othmani, selon la rewayat de Kaloun (برواية قالون). 17](#_Toc263084741)

[Figure 7 : Le programme Galoon (برنامج قالون). 17](#_Toc263084742)

[Figure 8 : Le logiciel « mushaf Al-Madina » pour les publications informatiques 18](#_Toc263084743)

[Figure 9 : Le schéma du mushaf électronique. 18](#_Toc263084744)

[Figure 10 : Exemple d’une application du Coran - le tafssīr. 19](#_Toc263084745)

[Figure 11 : Suggestion de mot alternatif pour « algrie » - Google 26](#_Toc263084746)

[Figure 12 : Mots apparentés au mot « coran » - Google 26](#_Toc263084747)

[Figure 13 : Pagination - Google 27](#_Toc263084748)

[Figure 14 : Tri par volume (size) – recherche dans Windows Explorer 27](#_Toc263084749)

[Figure 15 : Highlight (sur-lignage) 27](#_Toc263084750)

[Figure 16 : Exemples des synonymes 29](#_Toc263084751)

[Figure 17 : Exemples des antonymes 29](#_Toc263084752)

[Figure 18 : 3 types de lettre arabes : une graphie, deux graphies, quatre graphies 30](#_Toc263084753)

[Figure 19 : Classification d’affixes en Arabe 30](#_Toc263084754)

[Figure 20 : Le robot d’exploration dans un moteur de recherche web. 42](#_Toc263084755)

[Figure 21 : Processus d’indexation. 44](#_Toc263084756)

[Figure 22 : Le process d’interrogation 54](#_Toc263084757)

[Figure 23 : La classification des index par unité 58](#_Toc263084758)

[Figure 24 : Index prenant le mot comme unité. 59](#_Toc263084759)

[Figure 25 : Index prenant l’aya comme unité. 60](#_Toc263084760)

[Figure 26 : La classification des index par objet 62](#_Toc263084761)

[Figure 27 : Les différentes structures du Saint Coran 63](#_Toc263084762)

[Figure 28 : La hiérarchie des calculs du nombre 64](#_Toc263084763)

[Figure 29 : Le modèle de base 73](#_Toc263084764)

[Figure 30 : L’index de documents et l’index inversé 74](#_Toc263084765)

[Figure 31 : Le comportement de chercheur 75](#_Toc263084766)

[Figure 32 : Le positionnement de la nouvelle comparaison 75](#_Toc263084767)

[Figure 33 : L’emplacement du traitement de texte 76](#_Toc263084768)

[Figure 34 : Traitement de texte 77](#_Toc263084769)

[Figure 35 : Tokenization 77](#_Toc263084770)

[Figure 36 : Normalisation 78](#_Toc263084771)

[Figure 37 : Les niveaux de dérivations 78](#_Toc263084772)

[Figure 38 : La segmentation 80](#_Toc263084773)

[Figure 39 : Segmentation et filtrage des stopwords 81](#_Toc263084774)

[Figure 40 : Le positionnement du traitement de la requête 82](#_Toc263084775)

[Figure 41 : Traitement de la requête 82](#_Toc263084776)

[Figure 42 : Le Parseur 83](#_Toc263084777)

[Figure 43 : Les opérations du parseur: existants et nouveaux 83](#_Toc263084778)

[Figure 44 : Le remplacement par synonymes et antonymes 84](#_Toc263084779)

[Figure 45 : Niveaux de dérivation d’un index 85](#_Toc263084780)

[Figure 46 : La recherche par les dérivés de la racine du mot « فَأَسْقَيْنَاكُمُوه » par la méthode d’index 85](#_Toc263084781)

[Figure 47 : Exemples de considération d'orthographe 86](#_Toc263084782)

[Figure 48 : La recherche par couple (racine, type) 87](#_Toc263084783)

[Figure 49 : Le système de la suggestion 87](#_Toc263084784)

[Figure 50 : Un exemple du n-gram (n=3) sans et avec diacritiques 88](#_Toc263084785)

[Figure 51 : Suggestions du mot « جهنّم » 88](#_Toc263084786)

[Figure 52 : Le positionnement du traitement de résultats 89](#_Toc263084787)

[Figure 53 : Traitement des Résultats 89](#_Toc263084788)

[Figure 54 : La pondération 89](#_Toc263084789)

[Figure 55 : Le tri 89](#_Toc263084790)

[Figure 56 : Le filtrage 89](#_Toc263084791)

[Figure 57 : Extension de la recherche 90](#_Toc263084792)

[Figure 58 : Pagination 90](#_Toc263084793)

[Figure 59 : Highlight 90](#_Toc263084794)

[Figure 60 : Le processus de Highlight 91](#_Toc263084795)

[Figure 61 : Le système d’exploration 93](#_Toc263084796)

[Figure 62 : Transformation d’index de synonymes 101](#_Toc263084797)

[Figure 63 : Exemple de recherche par tuple 103](#_Toc263084798)

[Figure 64 : Highlight avec une balise HTML (mot-clé en jaune) 106](#_Toc263084799)

[Figure 65 : Highlight avec une balise BBcode (mot-clé en rouge) 106](#_Toc263084800)

[Figure 66 : Appercu des informaions de la sourate 108](#_Toc263084801)

[Figure 67 : Aperçu de la classification en sujets 109](#_Toc263084802)

[Figure 68 : Importation et appelle  de l'interface API 110](#_Toc263084803)

[Figure 69 : Exemple de résultats par l'interface API 110](#_Toc263084804)

[Figure 70 : Plus d’informations sur les résultats de l’API 110](#_Toc263084805)

[Figure 71 : Exemple d'utilisation de l'interface API (retrouver les termes utilisés) 110](#_Toc263084806)

[Figure 72 : Exemple d'utilisation de l'interface API (sur-lignage) 111](#_Toc263084807)

[Figure 73 : Exemple d'utilisation de l'interface API (avoir le temps écoulé et le nombre de résultats) 111](#_Toc263084808)

[Figure 74 : Exemple d'utilisation de l'interface API (Trier les résultats) 111](#_Toc263084809)

[Figure 75 : Exemple de suggestion pour une faute d’orthographe 112](#_Toc263084810)

[Figure 76 : Schéma des modules 112](#_Toc263084810)

[Figure 77 : Aperçu de l’interface sous Windows (mots-clés : الحمد لله) 113](#_Toc263084811)

[Figure 78 : Barre de recherche 113](#_Toc263084812)

[Figure 79 : Barre d’information sur les résultats 114](#_Toc263084813)

[Figure 80 : Panneau d'options 114](#_Toc263084813)

[Figure 81 : Panneau de fonctionnalités 114](#_Toc263084813)

[Figure 82 : Exemple d’une réponse du rebot 117](#_Toc263084814)

[Figure 83 : Exemple d’une barre de recherche pour Firefox 117](#_Toc263084815)

Liste des tableaux

[Tableau 1 : Des mots qui se répètent avec le même nombre de fois dans le Coran 16](#_Toc263084900)

[Tableau 2 : Les relations logiques 25](#_Toc263084901)

[Tableau 3 : Le changement de signification en changeant les diacritiques 31](#_Toc263084902)

[Tableau 4 : Le changement de fonction en changeant les diacritiques 32](#_Toc263084903)

[Tableau 5 : Des exemples de préfixes 32](#_Toc263084904)

[Tableau 6 : Des exemples de suffixes 33](#_Toc263084905)

[Tableau 7 : Les différences de graphies de Hamza. 35](#_Toc263084906)

[Tableau 8 : Exemples des différences entre le script standard et le script othmani 36](#_Toc263084907)

[Tableau 9 : Document Index 47](#_Toc263084908)

[Tableau 10 : Forward Index 47](#_Toc263084909)

[Tableau 11 : Inverted Index 48](#_Toc263084910)

[Tableau 12 : N-gram index 48](#_Toc263084911)

[Tableau 13 : Exemple d’un index basé sur les parts du mot 61](#_Toc263084912)

[Tableau 14 : L’index des phrases – surate al-fātiha 62](#_Toc263084913)

[Tableau 15 : Un exemple d’un index sémantique 63](#_Toc263084914)

[Tableau 16 : la structure d’index principal – Midād lbayān 65](#_Toc263084915)

[Tableau 17 : Aperçu sur l’index principal – Midād lbayān 65](#_Toc263084916)

[Tableau 18 : La structure d’index des mots – M.Taha Zerrouk 66](#_Toc263084917)

[Tableau 19 : Aperçu sur l’index des mots – M.Taha Zerrouki 66](#_Toc263084918)

[Tableau 20 : La structure d’index des sujets – M.Taha Zerrouki 66](#_Toc263084919)

[Tableau 21 : Aperçu sur l’index des sujets – M.Taha Zerrouki 66](#_Toc263084920)

[Tableau 22 : La structure d’index des synonymes – M.Taha Zerrouki 66](#_Toc263084921)

[Tableau 23 : Aperçu de l’index des synonymes – M.Taha Zerrouki 67](#_Toc263084922)

[Tableau 24 : La structure d’index de morphologie – Quranic corpus 67](#_Toc263084923)

[Tableau 25 : Exemple d’index de morphologie – Qur’anic corpus 68](#_Toc263084924)

[Tableau 26 : Exemple du Coran modèle simple propre – Tanzil.info 69](#_Toc263084925)

[Tableau 27 : Exemple d’index de surate – Tanzil.info 69](#_Toc263084926)

[Tableau 28 : L’index de sajdah – Tanzil.info 70](#_Toc263084927)

[Tableau 29 : L’index de juz’ – Tanzil.info 70](#_Toc263084928)

[Tableau 30 : L’index de rub’ – Tanzil.info 70](#_Toc263084929)

[Tableau 31: L’index de ruku’ – Tanzil.info 70](#_Toc263084930)

[Tableau 32: L’index de pages - Tanzil.info 70](#_Toc263084931)

[Tableau 33 : L’index de manzil - Tanzil.info 70](#_Toc263084932)

[Tableau 34 : Exemples de la comparaison en considérant les diacritiques 76](#_Toc263084933)

[Tableau 35 : Rôles et syntaxe des opérations existantes de parseur 84](#_Toc263084934)

[Tableau 36 : Rôles et syntaxe des opérations existantes de parseur. 100](#_Toc263084935)

[Tableau 37 : Exemples des fonctionnalités de l’API 112](#_Toc263084936)

[Tableau 38 : Les informations sur les ayas résultant 114](#_Toc263084937)

[Tableau 39 : Politiques d’ajout des mots dans la requête 115](#_Toc263084939)

**Introduction Générale**

**Introduction Générale**

**Contexte et Problématique :**

Le Coran est la parole de Dieu délivrée à toute l'humanité, et le livre sacré qui comporte des connaissances sur tous les aspects de la vie. Depuis sa révélation il y a quatorze siècles, les musulmans continuent à l’apprendre, l’enseigner et réaliser des ouvrages et des applications à son sujet.

Le Coran est également une source importante d’informations vu qu’il traite une diversité de sujets scientifiques, sociaux, historiques, politiques, etc. Cette richesse et la difficulté de l’obtenir ont conduit à l'émergence de plusieurs moteurs de recherche conçus à cet effet, comme Alawfa (الأوفى) et le prospecteur coranique (المنقب القرآني). Bien qu'il y ait beaucoup de moteurs coraniques, nous remarquons plusieurs lacunes dues à un certain nombre de raisons.

Certains de ces défauts sont hérités des moteurs de recherche arabes et leurs particularités par rapport aux moteurs de recherche latins, comme les diacritiques (الحركات) et la ressemblance entre quelques lettres telles que le tâ’ marbûtä (ة) et le Hâ’ (ه).

D’autres défauts sont issus de la sémantique. En fait, les méthodes utilisées par les moteurs de recherche actuels, telles que la recherche par expression régulière ou par racine du mot, ne sont pas suffisantes, car ils négligent des résultats liés au sens des mots de la requête, comme les mots dérivés et les synonymes.

D’autre part, certaines difficultés sont liées au Coran puisqu’il contient beaucoup plus d’informations qu’un texte simple. Ces informations sont relatives à la structure du mushaf (aya, hizb, …) ou à ses propriétés comme les semblants (المتشابهات) et l’abrogeant et l’abrogé (النّاسخ والمنسوخ) ou bien aux autres particularités, comme le scripte othmani (الخط العثماني). Mais, les moteurs de recherche coraniques n’offrent pas actuellement beaucoup d’options spécifiques pour le Coran.

En outre, Le Coran est fortement lié à plusieurs sciences et applications tels que le tafssīr (تفسير), la traduction, la récitation (التّلاوة), etc. Il sera très bénéfique de les assembler, car cela faciliterait l'accès à l'information. Donc, la réalisation d’un nouveau moteur de recherche coranique est une nécessité, mais difficile à accomplir.

Notre travail est une conséquence de l’affluence apparente des moteurs de recherche coraniques, il consiste à les standardiser et offrir aux utilisateurs et aux développeurs un moteur de recherche qui permet d’indexer et interroger les documents coraniques.

Le moteur de recherche proposé est riche en fonctionnalités spécifiques pour le Coran, nous citons à titre d’exemple la recherche par le script othmani, la recherche par les options structurelles du Coran et les options statistiques.

De plus, ce moteur est capable d’importer d’autres index et les inclure avec ses index. Les résultats seront enrichis par des informations additionnelles telles que les traductions, les tafssīrs et les récitations. Il comprend une interface graphique et une interface de programmation API (Application Programming Interface) de recherche qui peut être utilisée par les développeurs des applications coraniques.

**Organisation du mémoire :**

Tout ce qui a été mentionné ci-dessus justifie la répartition de ce mémoire en trois parties. Au début nous décrirons le contexte général dans lequel se situe notre travail, en parlant du Coran, des moteurs de recherche, leur principe de fonctionnement et des index coraniques.

Ensuite, nous passerons en revue notre conception, et enfin et nous présenterons notre application.

Le but de la première partie, **état de l’art**, est d’introduire les concepts de base dans le domaine de Coran et de moteurs de recherche. Elle comporte aussi des études de travaux similaires qui existent déjà. Elle est divisée en quatre chapitres qui sont :

**Chapitre 1 : Le Coran**

Ce chapitre présente une vue globale sur le Coran et ses sciences, il comporte un rappel historique sur l'évolution du Coran, la structure du mushaf, et les principaux problèmes d’informatisation du Coran, notamment le script othmani et l’authentification des textes coraniques.

**Chapitre 2 : Les moteurs de recherche et le Coran**

L’objectif de ce chapitre est de présenter les fonctionnalités des moteurs de recherche en général, et ceux arabes et coraniques en particulier, il a un grand rôle dans notre travail puisque il définit une gamme complète d’options de recherche dans le Coran.

**Chapitre 3 : le fonctionnent de moteurs de recherche**

Pour concevoir un moteur de recherche puissant, il est indispensable de comprendre comment les moteurs de recherche fonctionnent, dans ce chapitre, nous aborderons les différentes parties d’un moteur de recherche, à savoir : l’exploration, l’indexation et l’interrogation. Ainsi que la définition des notions de base dans le domaine des systèmes de recherche d’information.

**Chapitre 4 : Les index du Coran**

Notre moteur de recherche doit permettre d’utiliser les index préétablis, nous avons donc mis ce chapitre pour présenter les index du Coran et expliquer ses différents types. Nous aborderons aussi les différents index informatisés que nous avons collectés, et pour chacun d’eux, nous montrerons un aperçu sur sa structure et une brève évaluation.

Les deux autres parties, **conception** et **implémentation**, décrivent notre contribution à l’issu de ce projet.

**Chapitre 5 : Conception**

Ce chapitre comporte une explication de la démarche à suivre pour la réalisation proposée à l'aide d'un ensemble de schémas, et des détails de chaque composant du système par ordre chronologique de réalisation avec l’argumentation de tout choix effectué.

**Chapitre 6 : Implémentation**

Nous exposons dans ce chapitre les étapes de réalisation de notre moteur de recherche avec des détails techniques comme les outils de développement utilisés, les modules implémentés ainsi que les problèmes que nous avons rencontré.

Enfin, nous terminerons par une conclusion qui récapitule notre travail et les perspectives envisagées.

Ces chapitres sont suivis par quatre annexes que nous avons décidé d’ajouter afin de faciliter la compréhension de ce mémoire.

Chapitre I

**Le Coran**

1. Introduction

Le Coran est le livre sacré des musulmans, il est la base de plusieurs sciences et domaines, comme le tafssīr[[1]](#endnote-1) (التّفسير), la traduction, la récitation[[2]](#endnote-2) (التّلاوة), etc. Les informations contenues dans le texte coraniques sont les sources de ces sciences, d’où la nécessité d’un puissant moteur de recherche qui offre plein de fonctionnalités spécifiques au Coran.

Notre sujet est de réaliser un tel moteur, donc dans ce chapitre, nous allons d’abord préciser la définition du Coran et des documents coraniques, puis nous verrons en bref les principales étapes de l’évolution du Coran à travers le temps et comment il a été écrit, édité et informatisé, ensuite nous allons étudier le mushaf et les sciences du Coran.

Puis, nous verrons les bénéfices et les difficultés de l’informatisation du Coran, notamment, le script othmani, la vérification et l'authentification du mushaf et le besoin d’une référence coranique électronique standard.

1. Définitions
2. Coran

Le mot « Coran » ou « Qur’an », dans la langue arabe, veut dire récitation ou lecture, les savants musulmans le définissent par : les paroles d’Allah y révélées à son Prophète Mohamed s, écrits dans les mushafs et transmis par générations successives[[3]](#endnote-3) (التّواتر).

Le Coran est aussi connu par d’autres noms tels que : Al-Furkān[[4]](#endnote-4), Al-kitāb[[5]](#endnote-5), Al-dhikr[[6]](#endnote-6), Al-wahy[[7]](#endnote-7) et Al-rōuh[[8]](#endnote-8). Il est le livre sacré des musulmans et la première référence de la législation islamique. C’est le miracle divin que Dieu y a doté son prophète s. **[محسين, 73]**

1. Document coranique

C’est un document qui peut correspondre à la totalité du contenu de Coran (mushaf), ou bien à un extrait cohérent (ayas successives) du Coran en langue arabe. Ces extraits peuvent être utilisés comme des citations ou des exemples, et souvent écrits en script arabe standard. **[Hadjhenny, 08]**

1. L'historique du Coran

Dans cette partie, nous allons présenter les principales étapes du développement de mushaf depuis son révélation jusqu'à l’édition et l’informatisation.

1. L’écriture du Coran

Le Coran a été révélé sur le prophète Mohamedsde façon fragmentée pendant 23 ans, Mohamedset beaucoup de ses compagnons[[9]](#endnote-9) (الصّحابة) f l’ont appris par cœur.

Pendant le règne du calife ’Abu-Bakr d, soixante-dix personnes qui préservait le Coran ont été tués dans la bataille d’Ahl Al-Yamāmah[[10]](#endnote-10), ce qui a poussé le calife à appeler Zayd ibn Thābit[[11]](#endnote-11) d et lui confier la compilation du Coran.

A l’ère du calife Othmāne ibn ’Affān[[12]](#endnote-12) d, le Coran était récité de manières[[13]](#endnote-13) (القراءات) différentes, ce qui entrainait des discordes, alors le calife a décidé de faire un mushaf de référence, et faire des copies pour la diffusion et bruler tous les autres mushafs non conformes. **[Zerrouki, 05]** **[محسين, 73]**

1. Edition du Coran

Le Coran a été édité pour la première fois en Europe à Venise en 1573, puis à Hambourg en 1694 et Batavia en 1698. Les informations dans la première édition de Venise n’étaient pas rigoureuses. L’édition de Hambourg en 1694 a été réalisée par l’orientaliste allemand Hinckmani Ebrahami, afin d’étudier l’arabe et l’islam. Mais ces éditions n’ont pas eu d'acceptation parmi les musulmans, à cause des erreurs flagrantes qui violaient les règles du script othmani[[14]](#endnote-14) (الخطّ العثماني).

Actuellement, le mushaf est édité en grand quantité dans le monde arabe et musulman, le nombre des mushafs produits jusqu'à l'an 2000 pour le Complexe du Roi Fahd seulement dépassent les cent cinquante-trois millions d'exemplaires, dont la grande majorité a été distribués gratuitement. Cela montre l’importance exceptionnelle de ce livre pour les musulmans. **[Zerrouki, 05]**  **[آل الشّيخ, 00]**

1. Informatisation du Coran

La langue arabe a été utilisée dans les programmes informatiques depuis les années soixante. Le logiciel ‘Salsabīl’[[15]](#endnote-15) fut le premier à traiter le Coran, et il a été sans diacritiques à cause du bas niveau des logiciels à cette époque. Mais après, les écrans à haut résolution de Mac OS ont permis l'introduction des diacritiques et des polices arabe aux ordinateurs.

Aujourd’hui, le développement des logiciels coraniques a une extrême importance parmi tous les domaines de développement grâce à plusieurs facteurs, notamment :

* Le grand intérêt des musulmans pour le Coran ;
* La pris en charge des systèmes d’exploitation de la langue arabe ;
* L’intégration des symboles du Coran dans le system Unicode ;
* L’apparition des polices de caractères qui permettent d'écrire le mushaf en script othmani.
* Le passage de certains programmeurs au développement en source ouverte (Open Source).
* L’apparition des projets cumulatifs où chacun complète le travail de celui qui le précède.

Nous remarquons que malgré l’existence de nombreux projets dans le domaine du Coran, l’indépendance de ces projets fait perdre beaucoup d'efforts, ce qui appelle à la normalisation dans l’informatisation du Coran, cela exige la coopération des informaticiens avec les spécialistes du Coran et de la langue arabe.

Dernièrement, des efforts sont concentrés pour l’unification du support informatique de Coran, plusieurs projets ont été lancés dans ce sens, principalement, le projet intitulé « une approche ontologique pour le Coran et les sciences coraniques » par Taha ZERROUKI **[Zerrouki, 05]** en utilisant le modèle XML des documents électroniques pour réaliser un modèle de Mushaf électronique standard, notre travail, d'ailleurs, rentre dans le cadre de ce projet global de standardisation. **[خضر, 04]**

1. Le Mushaf
2. Définition

Le mushaf est le livre qui contient la totalité du Coran, le premier mushaf est le mushaf que le calife ’Abu-Bakr[[16]](#endnote-16) d a rassemblé. Les mushafs sont écrits selon le script othmani (الخطّ العثماني) qui est un peu différent du script arabe standard.  
La traduction du Coran ne constitue en aucun cas un mushaf, car le Coran fut révélé par Allah y en arabe dans sa forme comme dans son sens. **[Hadjhenny, 08]**

1. Les informations globales du Mushaf

Chaque mushaf a des caractéristiques faites par les spécialistes en se basant sur des livres de référence **[Zerrouki, 05]** :

* La rewayāte[[17]](#endnote-17);
* L'orthographe[[18]](#endnote-18) (الهجاء) ;
* Le comptage des ayas ;
* Les règles d’écriture[[19]](#endnote-19) (الضّبط) ;
* Fragmentation en ahzāb[[20]](#endnote-20) ;
* L’indication du lieu de révélation (Mecquois[[21]](#endnote-21) مكّي et Médinois[[22]](#endnote-22) مدني);
* L’indication des waqfs[[23]](#endnote-23) (بيان الأوقاف);
* L’indication des sajdahs[[24]](#endnote-24) (بيان السّجدات);
* Le rédacteur du mushaf[[25]](#endnote-25) (خطّاط المصحف);
* Le copieur du mushaf[[26]](#endnote-26) (ناسخ المصحف);
* La commission de révision[[27]](#endnote-27) (لجنة المراجعة) ;
* L’éditeur du mushaf[[28]](#endnote-28) ;
* L’organisme certifiant[[29]](#endnote-29) (هيئة التّرخيص).

1. La structure du mushaf

Le Coran est constitué de 114 sourates, les sourates sont divisées en ayas, en trouve dans le mushaf devant chaque aya son numéro, ce partitionnement et le même depuis le premier mushaf, puisque, comme nous l'avons indiqué précédemment, l’ordre des sourates et des ayas est précisé par le prophète s, cependant il existe d’autres fragmentations du mushaf adoptées par les spécialistes du Coran afin de faciliter l’apprentissage[[30]](#endnote-30) (الحفظ) et la récitation. Les marques de ces fragmentations ne font pas partie du texte coranique, et s’écrivent sur les marges.**[Zerrouki, 05]**

1. Fragmentation en sourates

C’est la fragmentation originale du Coran, chaque sourate comporte un nom, la basmala[[31]](#endnote-31) , – à l’exception de sourate El-Tawba[[32]](#endnote-32) qui ne comporte pas de basmala – et un certain nombre d’ayas séparées par un cercle qui contient le numéro de l’aya dans sa sourate, les sourates sont caractérisées par : un nom, le nombre d’ayas, le type de la révélation (Mecquoiseمكّيّة ou Médinoise مدنيّة), l’ordre dans le mushaf[[33]](#endnote-33) et l’ordre selon la révélation[[34]](#endnote-34).**[Zerrouki, 05]**

Figure 1 : schéma explicatif de la fragmentation en sourates.

1. Fragmentation en ahzab (أحزاب)

Le Coran est fragmenté en 30 parties approximativement égales, chaque partie s’appelle juz’[[35]](#endnote-35) (جزء), cela permet de répartir la lecture du Coran sur un mois, chaque juz’ est devisé de son tour en deux parties, chacun s’appelle hizb[[36]](#endnote-36).

Ainsi, le nisf[[37]](#endnote-37) est un demi de hizb, et le rub’[[38]](#endnote-38) est un demi de nisf, en fin le thumn[[39]](#endnote-39) est un demi de rub’ **[Zerrouki, 05]**

Figure 2 : Schéma explicatif de la fragmentation en ahzab (أحزاب).

1. Fragmentation en manāzil[[40]](#endnote-40) (منازل)

Au Pakistan, le mushaf est divisé en sept parties appelées manāzil qui sont les débuts des sourates suivantes **[Zerrouki, 05]** :

* Sourate Al-fātiha [[41]](#endnote-41);
* Sourate Al-Māida[[42]](#endnote-42);
* Sourate Yūnus[[43]](#endnote-43);
* Sourate Al-Isra[[44]](#endnote-44);
* Sourate Al-Chu’arā’[[45]](#endnote-45);
* Sourate As-Sāffāte[[46]](#endnote-46);
* Sourate Quf[[47]](#endnote-47).

**Figure** 3**: Schéma explicatif de la fragmentation en manāzil (منازل)**

1. Fragmentation en ruku‘s[[48]](#endnote-48) (ركوعات)

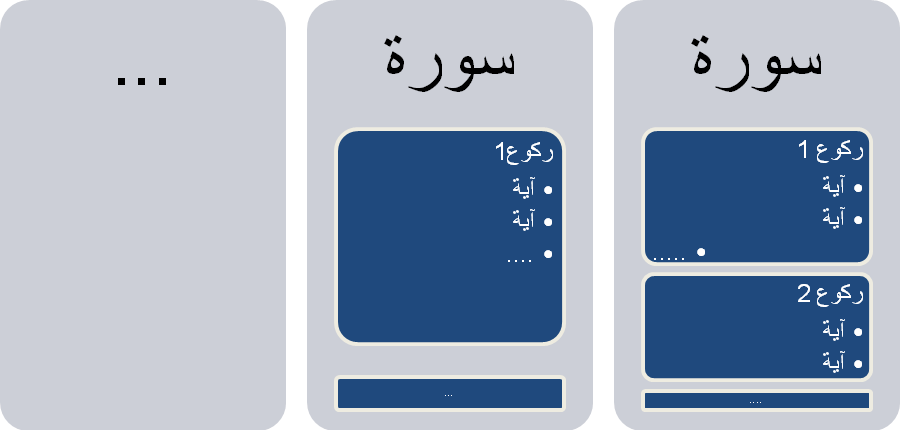
Les sourates sont partitionnées dans les pays de semi continent indien selon le sujet, chaque partie s’appelle ruku’[[49]](#endnote-49) (ركوع), il est indiqué par une Âīn en index **ع** et un autre sur la marge.

Figure 4 : Schéma explicatif de la fragmentation en Ruku'.

1. Fragmentation en waqfs (أوقاف)

Les marques du waqfs sont utilisés pour savoir quand on doit faire une pause lors de la récitation du Coran, ces marques sont différents afin de distinguer le type du waqf qui peut être : permises, préférables, interdites, etc.

On note que le marquage des waqfs diffère d’une rewayate à une autre, la figure 5 représente une référence des waqfs d’un mushaf en rewayate de Kaloun (برواية قالون). **[Zerrouki, 05]** **[Bourne, 79]**

****

Figure 5 : Les waqfs dans sourate Al-Houmazat[[50]](#endnote-50) en rewayate de Kaloun[[51]](#endnote-51)  
(برواية قالون عن نافع)

1. Les niveaux analytiques de la structure du Coran

La structure du Coran est constituée des niveaux analytiques suivants **[جماعة, --]** :

* la structure principale : la sourate, l’aya, le mot et la lettre ;
* les emplacements spéciaux : le début de la sourate, la fin de la sourate, fin de l’aya, virgule coranique[[52]](#endnote-52) (فاصلة قرآنية), sajdah (سجدة), waqf ;
* la structure du Coran: la page, le thumn, le rub’, le nisf, le hizb, le juz’ ;
* Ecriture du Coran: le script othmani, les al-sawāmit (الصّوامت), les harakāt, Al-hamza, l’écriture des mots, les diacritiques, des signes de distinction des lettres similaires, certains signes phonétiques ;
* structure incorporelle : mot, mot-clé, expression, unité d'objectif ;

La révélation du Coran : la manière de révélation, l'ordre de révélation, le lieu de révélation, le calendrier de révélation, les causes de révélation, le contexte de révélation, etc.

1. Les sciences du Coran

Ce sont les sciences qui prennent le Coran comme un sujet d’étude, pour l’illustrer et découvrir ses secrets, dans certains livres on les trouve sous d’autres noms comme (عِلم التّنزيل) et (عِلم الكتاب), on peut les classer en deux catégories **[الطّيّار, 08]**  :

1. Des sciences provenant originalement du Coran : on ne peut pas les utiliser pour étudier d’autres choses que le Coran comme la connaissance des ayas Mecquoise et Médinoise, les causes de révélation, etc.
2. Des sciences en commun entre le Coran et d'autres sciences : c’est les sciences qui découlent du fait que le Coran peut être vue de deux manières déférentes :
3. Si on le voit comme étant un texte écrit en arabe, les linguistes arabes peuvent ainsi l’étudier en termes de flexion[[53]](#endnote-53) (الإعراب), la conjugaison[[54]](#endnote-54) (الصّرف), la rhétorique[[55]](#endnote-55) (البلاغة), la lexicologie[[56]](#endnote-56) (علم المعاجم), etc.

On note ici que le Coran était le motif principal de l’apparition des sciences de la langue arabe, puisque les arabes n’ont pris soin de leur propre langue qu’après l’arrivée du l’islam. On note également que ce genre de sciences est plus large dans les livres de la langue arabe que dans les livres des sciences du Coran, car les sciences du Coran n’ont pris que dont ils ont besoin.

1. Si on le voit comme étant un texte législatif, alors cette particularité est partagée avec la sunna[[57]](#endnote-57), cela a produit un certain nombre de sciences comme le fiqh (الفقه), l’abrogeant et de l’abrogé (النّاسخ و المنسوخ), le général et le particulier[[58]](#endnote-58) (الخاصّ و العامّ), etc.

Le domaine des sciences coraniques est tellement vaste qu’on ne peut traiter toutes ses branches,  donc nous allons baser sur ce qui est censé nous aider dans le cadre de notre travail :

* **La connaissance des lieux de révélation des ayas :**

Il existe trois définitions pour distinguer ce qui est Mecquois de ce qui est Médinois, mais la définition la plus connue est la suivante :

Le Mecquois est tout ce qui a été révélé avant la migration[[59]](#endnote-59) (الهجرة النّبويّة), même s’il a été révélé à Médine, et le Médinoisest tout ce qui a été révélé après la migration, même s’il a été révélé à Mecque.

Les sourates Mecquoises sont au nombre de 82 sourates, et les Médinoises sont de 20 sourates, tandis que les 12 autres sourates ont soulevés des divergences, les ayas Mecquoises et Médinoises ont des caractéristiques qui les distinguent les unes des autres **[Dr.MAHDI, 09]** :

* **Les caractéristiques du Coran Mecquois :**

Toute sourate contenant :

* Une prosternation de récitation[[60]](#endnote-60) (سجود التّلاوة);
* Le terme non! (kallā كلّا), et il y a 33 occurrences de ce mot dans le Coran ;
* L’expression « ô gens » [يَا أَيُّهَا النّاسُ] ;
* Les contes des prophètes et des communautés passées, sauf dans le cas de sourate Al‑Baqara[[61]](#endnote-61) ;
* Les histoires d’Adam et Satan, sauf, encore une fois, sourate Al-Bakara ;
* Les lettres alphabétiques détachées comme Alif Lām Mīm (أ ل م) et Ḥā Mīm (ح م), sauf sourate Al-Bakara et Āl-‘Imrān[[62]](#endnote-62).

Ces caractéristiques avec leurs exceptions motionnées constituent des signes certains et invariables.

* **Les caractéristiques du Coran Médinois :**

Toute sourate contenant :

* Une obligation ;
* Une peine légale ;
* Ayas parlant des hypocrites ;
* Dialogue avec les juifs et les chrétiens (أهل الكتاب);
* *« ô les fidèles … »* [يَا أَيُّهَا الَّذِينَ آمَنُوا] ;
* **L’importance de savoir le Mecquois et le Médinois :**
* La connaissance de l’abrogeant et de l’abrogé[[63]](#endnote-63) (النّاسخ و المنسوخ).
* La maîtrise de l’historique de la législation islamique et de la finalité juridique en classant les branches avec leurs fondements, en utilisant des principes de réflexion et de psychologie pour prononcer les consultations juridiques de manière fiable. Cela a aidé à l’acceptation de la prédication et les lois islamiques par les peuples nouvellement convertis à l’islam.
* L’explication du Coran et la compréhension de ses enseignements.
* La découverte des divers styles coraniques et son usage dans la prédication islamique.
* La connaissance de la biographie du prophète s à travers les versets coraniques.
* **La connaissance de la conjugaison[[64]](#endnote-64) (التّصريف) :**

Cette science considère la structure des mots, elle est divisée en deux parties :

1. Mettre le mot dans plusieurs modes tel que le morphème[[65]](#endnote-65) (المصدر), nom du fā‛il[[66]](#endnote-66) (اسم الفاعل), nom du mafûl (اسم المفعول), etc.,
2. La modification d’un mot afin d’exprimer un sens particulier comme l’adjonction[[67]](#endnote-67) (الزّيادة), le renversement[[68]](#endnote-68) (القلب), Al-Idghām[[69]](#endnote-69) (الإدغام), etc.

La connaissance de la conjugaison est indispensable pour comprendre le sens de l’aya car parfois, on ne comprend pas les mots ambigus qu’avec la conjugaison, par exemple : [وَأَمَّا الْقَاسِطُونَ فَكَانُوا لِجَهَنَّمَ حَطَباً] et [وَأَقْسِطُوا إِنَّ اللَّهَ يُحِبُّ الْمُقْسِطِينَ], on peut bien voir comment le sens a changé par la conjugaison de l'inégalité à la justice. **[الزّركشي, --]**

* **La connaissance de l’orthographie[[70]](#endnote-70) (علم مرسوم الخطّ) :**

Quand les compagnons ont écrit le Coran à l'époque d’Othmāne d, ils ont divergés dans l’écriture du mot "le cercueil"[[71]](#endnote-71) (التّابوت) dans le mushaf, Zaid d a dit "التّابوه", mais les Koraïchites[[72]](#endnote-72) ont dit "التّابوت". Donc ils ont demandé l’avis d’Othmāne d qui a dit « écrire "التّابوت" car le Coran fut révélé avec le dialecte (la langue) de Koraiche[[73]](#endnote-73) (قريش) ».

En effet, l’écriture coranique est fixée, Abdellāh ibn Droustouīh[[74]](#endnote-74) a dit que dans la langue arabe il y a deux Calligraphies qu’ils ne peuvent pas être jugées : le script du mushaf et le script de découpage de la poésie[[75]](#endnote-75) (تقطيع العروض), donc il n'est pas permis d'écrire le mushaf avec un script autre que le script othmani, car les différences de ce script sont faite intentionnellement, nous donnerons ci-dessous quelques exemples **[الزّركشي, --]**  :

* L’ajout du Alif (أَلِف) : comme le Alif dans [كَأَمْثَالِ الْلُؤْلُؤِا], mais pas dans [كَأَنَّهُمْ لُؤْلُؤٌ] ;
* L’ajout du wāw (واو) : comme [سَأوُرِيكُمْ آيَاتِي] ;
* L’ajout du yā’ (ياء) : tel que [وَالسَّمَاءَ بَنَيْنَاهَا بِأَيْيدٍ] ;
* La suppression de l’Alif : comme dans [بِسْمِ اللَّهِ ], mais pas dans [بِاسْمِ رَبِّكَ] ;
* La suppression du wāw : comme [وَيَمْحُ اللَّهُ الْبَاطِلَ], mais pas dans

[يَمْحُوا اللَّهُ مَا يَشَاءُ];

* La suppression du yā’ : comme [فَلا تَسْأَلْنِ مَا لَيْسَ لَكَ بِهِ عِلْمٌ], mais pas dans

[فَلا تَسْأَلْنِي عَنْ شَيْءٍ حَتَّى أُحْدِثَ لَكَ مِنْهُ ذِكْراً];

* Al-Idghām (الإدغام) : comme [فَإِلَّمْ يَسْتَجِيبُوا لَكُمْ] et [فَإِنْ لَمْ يَسْتَجِيبُوا لَكَ];
* Les lettres porches : tel que le sīn et le sâd dans : [وَزَادَكُمْ فِي الْخَلْقِ بَصطَةً] et [وَزَادَهُ بَسْطَةً فِي الْعِلْمِ وَالْجِسْمِ].
* **L’analyse grammaticale du Coran[[76]](#endnote-76) (إعراب ألفاظ القرآن):**

Le prophète sa dit : « أعربوا القرآن والتمسوا غرائبه » – رواه البيهقي والحاكم عن أبي هريرة – et le mot إعراب veut dire ici la connaissance des sens des mots, et il y a même qui ont analysé grammaticalement tout le Coran mot par mot. **[الزّركشي, --]**

* **La connaissance des causes de révélation :**

L’importance de cette science se manifeste dans les ayas qui concernent des moments spécifiques de la vie du prophète s, par exemple :

[وَلِلَّهِ الْمَشْرِقُ وَالْمَغْرِبُ فَأَيْنَمَا تُوَلُّوا فَثَمَّ وَجْهُ اللَّهِ], si on essaie de comprendre littéralement, on va dire qu’il n’est pas obligatoire de s’orienter vers la Qibla, ce qui est évidement faux. Donc nous ne pourrons saisir le sens de l’aya que si on connait la cause de révélation. **[الزّركشي, --]**

* **La science des** **semblants[[77]](#endnote-77) (علم المتشابه) :**

Les semblants sont les ayas qui ont le même sens mais mentionnées sous des formes différentes et les paroles répétées plusieurs fois, comme dans les exemples suivants  
 **[الزّركشي, --]** :

* l’inversement : comme [وَالنَّصَارَى وَالصَّابِئِينَ] et [وَالصَّابِئِينَ وَالنَّصَارَى] ;
* ressemblance par augmentations ou diminutions : comme [فَأْتُوا بِسُورَةٍ مِّنْ مِثْلِهِ] et [فَأْتُوا بِسُورَةٍ مِثْله] ;
* la définition et l’indéfinition[[78]](#endnote-78) (التّعريف و التّنكير) : comme [وَيَقْتُلُونَ النَّبِيِّينَ بِغَيْرِ الْحَقِّ] et [وَيَقْتُلُونَ النَّبِيِّينَ بِغَيْرِ حَقِّ] ;
* le singulier et le plurielle : comme [لَنْ تَمَسَّنَا النَّارُ إِلاَّ أَيَّاماً مَعْدُودَاتٍ] et [لَنْ تَمَسَّنَا النَّارُ إِلاَّ أَيَّاماً مَعْدُودَةً] ;
* la substitution d’une préposition par une autre : comme [وَمَا أُنْزِلَ إِلَيْنَا] et [وَمَا أُنْزِلَ عَلَيْنَا];
* la substitution d’un mot : comme [مَا أَلْفَيْنَا عَلَيْهِ آبَاءَنَا] et [مَا وَجَدْنَا عَلَيْهِ آبَاءَنَا];
* la répétition littérale : comme [إِنَّ فِي ذَلِكَ لآيَةً] est répétée vingt fois dans le Coran.
* **Les débuts des surates[[79]](#endnote-79) (فواتح السّور) :**

Le Coran contient 114 sourates, chacune commence par l’un des dix types suivants  
 **[الزّركشي, --]**  :

* remerciement du Dieu comme : [الْحَمْدُ لِلَّهِ];
* les lettres de l’alphabet comme : [كهيعص];
* l’appel comme : [يَا أَيُّهَا الَّذِينَ آمَنُوا];
* le serment comme : [وَالنَّازِعَاتِ];
* les phrases narratives comme : [أَتَى أَمْرُ اللَّهِ] ;
* la condition comme : [إِذَا جَاءَ نَصْرُ اللَّهِ];
* l’ordre comme : [اقْرَأْ بِاسْمِ رَبِّكَ];
* l’interrogation comme : [عَمَّ يَتَسَاءَلُون] ;
* la prière (الدّعاء) comme : [وَيْلٌ لِلْمُطَفِّفِينَ];
* la justification utilisée une seule fois dans : [لإِيلافِ قُرَيْشٍ].
* **La connaissance des exemples existants dans le Coran :**

Cette science est importante car il y a plusieurs ayas qui sont éclairées à travers des exemples comme : [كَمَثَلِ الْعَنْكَبُوتِ اتَّخَذَتْ بَيْتاً] et [مَثَلُهُمْ كَمَثَلِ الَّذِي اسْتَوْقَدَ نَاراً], cette science est indiqué dans le Coran dans : [وَتِلْكَ الأَمْثَالُ نَضْرِبُهَا لِلنَّاسِ وَمَا يَعْقِلُهَا إِلاَّ الْعَالِمُونَ]  **[الزّركشي, --]**

* **Le tafssīr du Coran :**

Le tafssīr du Coran est l’explication de ses sens et l’extraction de ses lois[[80]](#endnote-80) (الأحكام), cela nécessite la connaissance des causes de révélation, l’analyse grammaticale du Coran et d’autres sciences coranique. **[Bourne, 79]**

1. Les bénéfices de l’informatisation du Coran

* Rassembler les différentes sciences coraniques : l’amélioration de l’accessibilité et du temps d’accès aux données est un objectif essentiel de l’informatisation en général. En particulier, l’informatisation du Coran facilite énormément l’apprentissage, la récitation, la recherche, le tafssīr, la traduction, et l’enseignement du Coran. Les applications coraniques actuelles nous offrent la possibilité de passer à travers tous ces sciences avec des simples cliques, en économisant de temps et d’efforts que de chercher dans des livres. Nous ne voulant pas dire par la que ces logiciels remplacent les livres et les façons d’apprentissage traditionnels, mais ils les complémentent.
* Étaler les miracles numériques et statistique du Coran : le grand miracle du Coran est son langage éloquent qui a dépassé tous les rhéteurs à l’époque du prophète s, et avec le développement technologique, plusieurs miracles scientifiques ont été découverts, mais il existe aussi un autre type de miracle, c’est les miracles basés sur les statistiques effectués sur le Coran, par exemple :

|  |  |  |  |
| --- | --- | --- | --- |
| Le mot | Le nombre de répétitions | Le mot | Le nombre de répétitions |
| الدّنيا | 115 fois | **الآخرة** | 115 fois |
| الملائكة | 68 fois | **الشّياطين** | 68 fois |
| الصّالحات | 167 fois | **السّيئات** | 167 fois |
| المحبّة | 83 fois | **الطّاعة** | 83 fois |
| الهدى | 79 fois | **الرّحمة** | 79 fois |
| الشّدة | 102 fois | **الصّبر** | 102 fois |
| العقل | 49 fois | **النّور** | 49 fois |
| المصيبة | 75 fois | **الشّكر** | 75 fois |
| الجهر | 16 fois | **العلانية** | 16 fois |

Tableau 1 : Des mots qui se répètent avec le même nombre de fois dans le Coran. [نوفل, 75]

L’informatisation du Coran peut aider à découvrir d’autres miracles statistiques clairs, pourtant, nombreux sont ceux qui doute de l’utilité de ces statistiques, de toute façon un vrai musulman n’est pas besoin de ces statistiques pour croire au Coran, mais c’est juste pour mentionner les merveilles qu’en peut trouver dans le Coran si en l’étudie profondément.

On note ici qu’il existe de nombreuses tentatives modernes pour prouver les miracles numériques dans le Coran qui ont rencontré une forte opposition de la part des savants musulmans, en particulier, l’imposteur Rachād Khalīfa a prétendu être un prophète en se basant sur des observations des phénomènes numériques de ce genre, mais ce qui nous concerne dans cette affaire en tant que développeurs, c’est que nous allons fournir aux bons chercheurs les outils de recherche nécessaires pour leur permettre d'obtenir des résultats certains, et fondées sur des bases scientifiques.**[خضر, --]**

1. Les principaux problèmes d’informatisation du Coran
2. Le script othmani

Quand les compagnons f ont recopié le mushaf à l’époque d’Othmān ibn ‘Affan d, ils ont utilisé le script othmani qui a quelques différences par rapport au script arabe standard utilisé actuellement. Ces différences sont des différences d’orthographe, de waqfs (الأوقاف), indications de récitation comme les diacritiques supplémentaires (al-’išmam[[81]](#endnote-81) الإشمام et al-’imala[[82]](#endnote-82) الإمالة).

Ce script est particularisé pour le mushaf, aucun autre livre ne doit l’utiliser, et le mushaf doit obligatoirement être écrit avec le script othmani, et cela s'applique à ce jour, car si l'on écrit le Coran actuellement avec le script standard, comme certains l’ont fait, nous allons perdre beaucoup d’informations.

Pour remédier à ce problème, certains ont utilisé les images pour afficher les textes coraniques, ce qui rend le traitement de ces textes difficile. La meilleure solution et d’utiliser le codage universel Unicode qui contient 65536 caractères, et support les marques coraniques, et un seul logiciel ou site Internet peut satisfaire simultanément et sans modification les demandes de plusieurs plates-formes, l’Unicode a également alloué l’intervalle de U+0700 à U+08FF pour toute évolution dans les lettres et les symboles arabes.

Pour afficher le texte coranique en Unicode avec le script othmani, il faut utiliser des polices spéciales, par exemple, dans la polices “Msh Quraan” les symboles U+0028 et U+0029 sont utilisées pour encadrer un nombre, et composer ainsi la forme connue pour la numérotation des ayas , ces mêmes codes sont utiliser dans d’autres polices standards pour écrire les parenthèses ().

Donc la compatibilité n’est pas toujours conservée en passant d’une police à une autre, d’où la nécessité d'une police destinée aux textes coraniques comme une solution radicale à ce problème. Dans ce contexte, nous mentionnons le projet de la police TahaArabeyesQr, qui améliore la pris en charge du scripte othmani. La figure 6 représente sourate Al-Nnaçr[[83]](#endnote-83) (سورة النّصر) avec la police TahaArabeyesQr. **[Zerrouki, 05]** **[خضر, 96]**

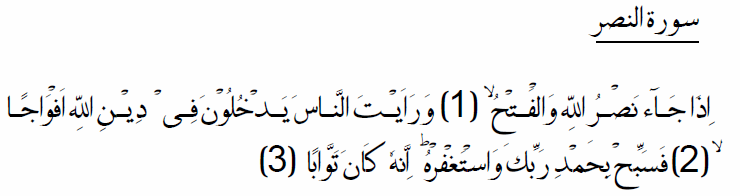
****

Figure 6 : Texte coranique écrit en script othmani, selon la rewayat de Kaloun (برواية قالون). [Zerrouki, 05]

1. La vérification et l'authentification du Mushaf

Les logiciels autour du Coran deviennent de plus en plus nombreux, mais ils ne sont pas tout 100% corrects, nous trouvons parfois des erreurs plus ou moins graves, dans le programme de Galoon[[84]](#endnote-84) à titre d’exemple, nous trouvons l’aya 38 de sourate d’Al-Qalam[[85]](#endnote-85) écrite de cette manière : [إِنَّ لَكُمْ فِيهِ لَمَا يَتَخَيَّرُونَ] (*voir la figure 7*), alors que le correct est de dire : [إِنَّ لَكُمْ فِيهِ لَمَا تَخَيَّرُونَ]. Cette erreur ne se limite pas dans le logiciel seulement, mais elle se propage par le fait de la citation, et si on recherche dans Google sur l’expression "إِنَّ لَكُمْ فِيهِ لَمَا يَتَخَيَّرُونَ", on obtient environ 42 400 résultats, ce qui n’est pas acceptable.

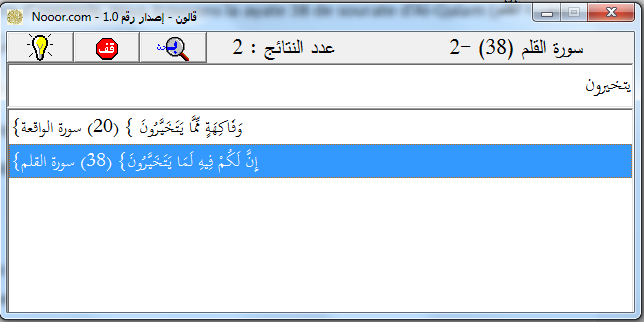


Figure 7 : Le programme Galoon (برنامج قالون). [Galoon, 99]

Mais d’une autre part, il existe des logiciels fiables, tel que le logiciel « mushaf Al-Madina » pour les publications informatiques[[86]](#endnote-86) par le Complexe Roi Fahd pour l'Impression du Saint Coran[[87]](#endnote-87), supervisé par un comité scientifique de haute classe, et il a été révisées plusieurs fois (*voir la figure 8*).  
Ce logiciel est disponible gratuitement à partir de son site web officiel. **[Al-Madina, 00]** **[Galoon, 99]**

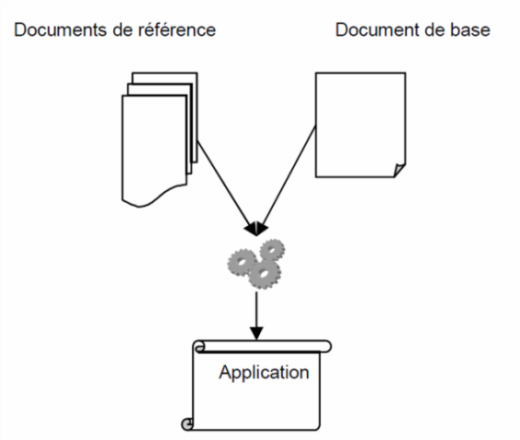
****

Figure 8 : Le logiciel « mushaf Al-Madina » pour les publications informatiques. [Al-Madina, 00]

1. Référence coranique électronique standard

Le mushaf est un document de base pour plusieurs applications autour du Coran, tels que la recherche, le tafssīr, la récitation, etc.

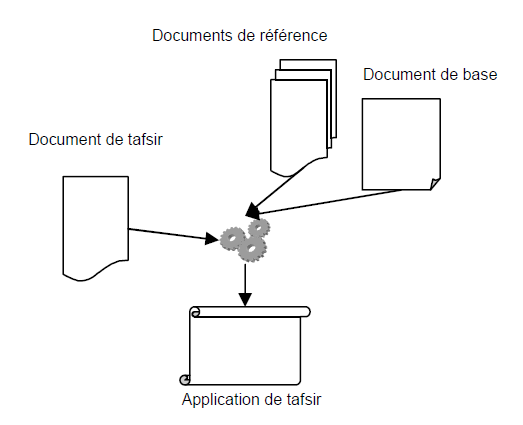
Les concepteurs de ces applications ont besoin d’une référence coranique standard, sinon, chacun serai obligé de réécrire son mushaf, avec tout le gaspillage du temps, ainsi que les risques de faire des erreurs. Le mushaf contient plusieurs types d’information en plus qu’un texte brut, ces informations concernent les ayas, les sourates, le mushaf lui-même, la fragmentation en ahzab ou ruku‘s et les autres caractéristiques du mushaf.

D’autres informations supplémentaires existent sur le mushaf tels que le nombre d’ayas par sourate, le type de révélation et l’ordre de révélation. Ces informations sont décrites séparément du texte coranique.

Pour satisfaire à ces besoins, Mr. Taha Zerrouki a proposé un document électronique de base et un ensemble de documents de référence autour de ce document. Ces documents vont être utilisés pour la conception des applications autour du Coran.

**Figure 9 : Le schéma du mushaf électronique.** **[Zerrouki, 05]**

Après avoir conçu tous les documents de base et de référence, on peut générer des applications sur le Coran. Par exemple, pour générer une application de tafssīr, on a besoin d’un document qui décrit le tafsîr associé à chaque aya. Le résultat est une application qui permet d’afficher le tafssīr correspondant aux ayas *(voir la figure 10)*.



**Figure 10 : Exemple d’une application du Coran - le tafssīr.** **[Zerrouki, 05]**

Ils existent aussi d’autres modalisations du mushaf, comme le projet « Quran » par le groupe de développement Arabeyes.org, il contient le texte du Coran complet sous forme XML. Le texte du Coran est structuré en ayas et sourates, chaque aya est écrite en deux scripts : une fois sans diacritiques (الحركات) pour effectuer la recherche, et une autre fois avec les diacritiques et tout le script othmani, le fichier utilise l’encodage universel des caractères Unicode.

1. Conclusion

Dans ce chapitre, nous avons présenté une vision globale sur le Coran et ses sciences en montrant l'impact positif de l’informatique dans ce domaine, en particulier des technologies comme l’Unicode, les polices de script othmani et l’XML ont contribué énormément dans le développement de plusieurs projets d’informatisation du Coran.

Nous avons constaté qu’Il y a tellement de livres écrits sur le Coran et les sciences coraniques que la plupart du temps des chercheurs se perdent dans la collecte de livres de documentaires et dans la recherche. La solution optimale à ce problème est la collecte et l’indexation de ces écrits puis d’offrir la possibilité d’une recherche bien spécialisée pour ce domaine, et il serait aussi préférable que ce processus soit automatique car cela améliorerait considérablement les conditions de travail pour les chercheurs. Dans le chapitre suivant nous allons étudier les moteurs de recherche pour le Coran.

# Chapitre II

**Les moteurs de recherche et le Coran**

1. Introduction

"*How could the world beat a path to your door when the path was uncharted, uncatalogued, and could be discovered only serendipitously?*" — Paul Gilster, *Digital Literacy*

Le saint Coran[[88]](#endnote-88) est différent de tous les documents que nous connaissons. Avec un trésor important d'informations, nous ne pouvons extraire manuellement qu’une petite partie de ces informations. En raison de cette limitation, le Coran nous oblige à trouver de nouvelles façons d'interagir.   
Par exemple, si je voulais trouver un livre sur l’apprentissage de la langue française, je vais simplement faire une recherche dans Google, sélectionner un livre, le télécharger et c'est tout. Les moteurs de recherche (comme Google) sont utilisés habituellement avec des lettres latines pour des recherches d'informations générales de livres comme titre, éditeur ... ou le texte contenu, l'indexation les rend relativement facile à trouver.

Cependant, dans le saint Coran, ceci n'est pas suffisant, car il y a beaucoup d'informations à trouver, on a besoin de trouver une méthode pour les extraire, mais Il n'y a aucun outil à utiliser sauf quelques lexiques[[89]](#endnote-89) imprimés (sous forme de livres), on doit donc trouver le bon lexique qu’on sera peut être obligé de l'acheter (on peut faire une recherche, mais ça sera juste un coup de chance, de plus, on a besoin de savoir pas où commencer). Parfois, c'est évident : trouvez le nom du lexique, recherche le mot dans l'index, trouver la page, obtenir l'information exacte. Mais que faire dans le cas d’une requête complexe et spéciale sachant que les lexiques existants ne peuvent répondre qu’à des requêtes d’interrogation simples ?

La solution adéquate à ce problème est le moteur de recherche. En fait, c'est l'une des méthodes les plus importantes pour obtenir des informations du Coran. Sachant la quantité d'information qu’un bon moteur de recherche peut gérer, il serait intéressant d’avoir le Coran, les lexiques et beaucoup d’autres choses ... en même temps.

Les moteurs de recherche peuvent fournir des informations beaucoup plus qu'un simple lexique. On trouvera dans ce chapitre toutes les fonctionnalités existantes attendues des moteurs de recherche en général, et celles arabes et coraniques en particulier.

1. Définition

Un moteur de recherche est un [logiciel](http://fr.wikipedia.org/wiki/Logiciel) permettant de retrouver des ressources ([pages web](http://fr.wikipedia.org/wiki/Page_web), [images](http://fr.wikipedia.org/wiki/Image_num%C3%A9rique), [vidéo](http://fr.wikipedia.org/wiki/Vid%C3%A9o), fichiers, informations ... etc.) associées à des mots quelconques. Certains [sites web](http://fr.wikipedia.org/wiki/Site_web) offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors moteur de recherche le site lui-même ([Google](http://fr.wikipedia.org/wiki/Google_Video), Yahoo, Bing… sont des moteurs de recherche).**[Nejjari, 07]**

Un moteur de recherche est aussi un outil de recherche sur le [web](http://fr.wikipedia.org/wiki/World_Wide_Web) constitué de « [robots](http://fr.wikipedia.org/wiki/Robot_d%27indexation) », qui parcourent les [sites](http://fr.wikipedia.org/wiki/Site_web) à intervalles réguliers et de façon automatique (sans intervention humaine, ce qui les distingue des [annuaires](http://fr.wikipedia.org/wiki/Annuaire_web) [URL](http://fr.wikipedia.org/wiki/Adresse_réticulaire)- Uniform Resource Locator-). Ils suivent les liens [hypertextes](http://fr.wikipedia.org/wiki/Hypertexte) (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors [indexée](http://fr.wikipedia.org/wiki/Indexation) dans une [base de données](http://fr.wikipedia.org/wiki/Base_de_données), accessible ensuite par les [internautes](http://fr.wikipedia.org/wiki/Internaute) à partir de [mots-clés](http://fr.wikipedia.org/wiki/Mot_clé). **[Sanan, 08]**

Les moteurs de recherche sont classables en 2 catégories :

* Moteur de recherche à indexation automatique
* Moteur de recherche à inscription

L’indexation automatique part du principe que les pages web contiennent des marques spécifiques suivies du sujet de la page. Altavista, par exemple, un moteur à indexation automatique scanne le web en permanence et repère ces codes puis indexe les pages en questions dans une grosse base de données.

Les moteurs à inscription eux, partent du principe que quelqu’un qui vient de terminer un site web va remplir un formulaire spécifiant l’adresse de son site et le sujet traité puis va envoyer son formulaire au moteur de recherche en question qui, lui, l’indexera. **[menu, 00]**

Les moteurs de recherche ne s'appliquent pas qu’à Internet : certains moteurs sont des [logiciels](http://fr.wikipedia.org/wiki/Logiciel) installés sur un [ordinateur personnel](http://fr.wikipedia.org/wiki/Ordinateur_personnel). Ce sont des moteurs dits desktop qui combinent la recherche parmi les [fichiers](http://fr.wikipedia.org/wiki/Fichier_%28informatique%29) stockés sur le [PC](http://fr.wikipedia.org/wiki/Compatible_PC) et la recherche parmi les sites web — on peut citer par exemple Exalead Desktop, Google Desktop et [Copernic](http://fr.wikipedia.org/wiki/Copernic_%28société%29) Desktop Search…etc.

En décembre 2004, le Dr Tim Berners Lee (l’inventeur du World Wide Web) a parlé d'un nouveau projet : “Web sémantique” qui se base sur l'idée de traitement d’informations disponibles sur le web automatiquement en fonction de leurs significations ; et le motif que 80% du Web contient actuellement des textes destinés à être lus et compris par les humains. Alors que les programmes informatiques, les navigateurs Web et les programmes de recherche sont incapables de comprendre ce contenu, ils sont donc incapables d’accélérer la recherche et le développement web sémantique liée. Dans moins de deux ans de l'article de M. Lee, les Premiers fondations de Web sémantique ont été formées, elles semblaient conduire le monde vers une nouvelle révolution dans l’internet bouleversant une grande partie des constantes actuelles et prévues des changements profonds dans les méthodes des moteurs de recherche. **[أبوالحجاج, 09]**

1. Les moteurs de recherche

De prime abord, rien ne distingue un moteur de recherche “classique”, d’un moteur de recherche “sémantique”. Même interface sobre, avec une zone de texte au centre de la page, dans laquelle l’utilisateur peut entrer sa requête.

En fait, la différence réside dans le mode de recherche. Un moteur “classique”, type Google, fonctionne de la manière suivante : ses robots indexent les “mots” dans les pages qu’ils parcourent. Ces mots sont ensuite rangés dans une gigantesque base de données. Les requêtes que les utilisateurs font sont des recherches dans ce “dictionnaire géant”, un algorithme permettant de classer et présenter les résultats selon un certain ordre de pertinence. **[Mentre, 08]**

Le fonctionnement général d’un moteur de recherche se décompose en trois étapes principales :

1. L'exploration (en anglais : crawling)

Le web est systématiquement exploré par un [robot d'indexation](http://fr.wikipedia.org/wiki/Robot_d%27indexation) suivant [récursivement](http://fr.wikipedia.org/wiki/R%C3%A9cursivit%C3%A9) tous les [hyperliens](http://fr.wikipedia.org/wiki/Hyperlien) qu'il trouve et récupérant les ressources jugées intéressantes. L'exploration est lancée depuis une ressource pivot, comme une page d'[annuaire web](http://fr.wikipedia.org/wiki/Annuaire_web). Un moteur de recherche est d'abord un outil d'[indexation](http://fr.wikipedia.org/wiki/Indexation), c'est-à-dire qu'il dispose d'une technologie de collecte de documents à distance sur les sites web, via un outil que l'on appelle [robot](http://fr.wikipedia.org/wiki/Robot) (encore appelé bot, spider, crawler ou agent). Un robot d'indexation dispose de sa propre signature (comme chaque navigateur web). [Googlebot](http://fr.wikipedia.org/wiki/Googlebot) est le [user agent](http://fr.wikipedia.org/wiki/User_agent) (signature) du crawler de Google La recherche simple est le mode de recherche par défaut, qui interroge l'ensemble des bases revues et jurisprudence du site. **[Sanan, 08]**

1. L'indexation

L'indexation des ressources récupérées consiste à extraire les mots considérés comme significatifs du [corpus à explorer](http://fr.wikipedia.org/wiki/Corpus). Les mots extraits sont enregistrés dans une base de données organisée comme un gigantesque [dictionnaire inverse](http://fr.wikipedia.org/wiki/Dictionnaire_inverse) ou, plus exactement, comme l'[index terminologique](http://fr.wikipedia.org/wiki/Index_terminologique) d'un ouvrage, qui permet de retrouver rapidement dans quel chapitre de l'ouvrage se situe un terme significatif donné. Les termes non significatifs s'appellent des [mots vides](http://fr.wikipedia.org/wiki/Mots_vides). Les termes significatifs sont associés à une valeur de poids. Ce poids correspond à une probabilité d'apparition du mot dans un document. Cette probabilité est indiquée sous la forme d'une "fréquence de terme", ou "inverse de fréquence de terme" et est souvent calculée par la formule [TF-IDF](http://fr.wikipedia.org/wiki/TF-IDF)[[90]](#footnote-1) (term frequency-inverse document frequency). **[Sanan, 08]**

1. L’interrogation

L’interrogation correspond à la partie requêtes du moteur, qui restitue les [résultats](http://fr.wikipedia.org/wiki/Page_de_r%C3%A9sultats_d%27un_moteur_de_recherche). Un [algorithme](http://fr.wikipedia.org/wiki/Algorithmique) est appliqué pour identifier dans le [corpus documentaire](http://fr.wikipedia.org/wiki/Corpus) (en utilisant l'index), les documents qui correspondent le mieux aux mots contenus dans la requête, afin de [présenter les résultats des recherches](http://fr.wikipedia.org/wiki/Page_de_r%C3%A9sultats_d%27un_moteur_de_recherche) par ordre de pertinence supposée. Les algorithmes de recherche font l'objet de très nombreuses investigations scientifiques. Les moteurs de recherche les plus simples se contentent de requêtes [booléennes](http://fr.wikipedia.org/wiki/Bool%C3%A9en) pour comparer les mots d'une requête avec ceux des documents. Mais cette méthode atteint vite ses limites sur des corpus volumineux. Les moteurs plus évoluées utilisent la formule TF-IDF (term frequency-inverse document frequency) pour mettre en perspective le poids des mots dans une requête avec ceux contenus dans les documents. Cette formule est utilisée pour construire des vecteurs de mots, comparés dans un espace vectoriel, par une mesure de Cosinus ([la similarité Cosine](http://fr.wikipedia.org/wiki/Similarit%C3%A9_cosinus)). Pour améliorer encore les performances d'un moteur, il existe de nombreuses techniques, la plus connue étant celle du [pagerank](http://fr.wikipedia.org/wiki/PageRank) de Google qui permet de pondérer une mesure de cosinus en utilisant un indice de notoriété de pages. Les recherches les plus récentes utilisent la méthode dîtes [d'analyse sémantique latente](http://fr.wikipedia.org/wiki/Analyse_s%C3%A9mantique_latente) qui tente d'introduire l'idée de co-occurrentes dans la recherche de résultats (le terme "voiture" est automatiquement associé à ses mots proches tels que "garage" ou un nom de marque dans le critère de recherche).**[Sanan, 08]**

1. Les fonctionnalités de recherche :

La recherche simple est le mode de recherche par défaut, cette recherche est basée sur les mots clés entrés dans l’interrogation. Mais parfois, cette méthode n'est pas suffisante, surtout lorsque la requête devient plus compliquée, par exemple quand on veut chercher les fichiers de type PDF, qui portent le mot mémoire dans leur titre, ce type de recherche doit être effectué en mode avancé.

Les fonctionnalités de recherche en mode avancé plus d’autres d’un aspect linguistiques seront expliqués dans les points suivants :

1. La recherche avancée
2. Recherche par champs (fielded search)

Dans une base de données structurée, les données sont entrées dans des champs ou des zones. Ces champs sont libellés d'après la donnée. Dans une base de données documentaire, il y a 4 champs au minimum : le titre, l'auteur, les sujets et la source. La source peut être divisée en plusieurs champs, comme les lieux d'édition, la maison d'édition et la date. Pour plusieurs moteurs de recherche, le mot champ est remplacé par filtre. Dans une page web, il y a essentiellement deux parties importantes d'où proviennent les champs : les métadonnées et l'adresse URL.**[Duval, 06]**

Pour qu'un champ devient un «champ de recherche», il a besoin d'un identifiant qui est un préfixe. Chaque base de données possède ses propres préfixes de recherche. Plusieurs sites de recherche ont développé ses préfixes de recherche comme Altavista et Google.

La syntaxe dans Google, Yahoo et Ask est : « Préfixe : terme de recherche ».

On peut utiliser cette fonctionnalité pour les index spéciaux comme : « intitle:quran filetype:pdf », cet exemple demande la recherche de mot "quran" situé dans le titre de la page, et le type fichier doit être PDF.

1. Recherche par une partie de mot

Cette fonctionnalité est souvent disponible dans les moteurs de recherche, c’est utile pour rechercher avec un mot en commun dans plusieurs mots qui ont les lettres en plus, par exemple :

* Pour la recherche de "partie", "partition" et "partitionnement ", il suffit d’utiliser le mot "part".

1. Recherche par les jokers (les caractères spéciaux)

L’utilisation des expressions génériques permet la suppression ou l’omission d’une chaîne de caractères dont on veut ignorer la valeur, en utilise ? pour représenter un seul caractère et \* pour représenter n'importe quel nombre de caractères.

Par exemples :

* Pour la recherche du nom propre "Rabah" et "Rabeh", il suffit d’utiliser l’expression "Rab?h".

1. Recherche par les relations logiques

La recherche par les relations logiques fournit plusieurs options pour l'utilisateur comme la recherche de sites qui contient tous les mots sélectionnés (ex : Python et Java) en utilisant la relation « **ET »**, ou des sites qui contiennent au moins l’un des mots en utilisant le lien « **OU »**, ou même les sites qui ne contiennent pas un mot spécifique en utilisant la relation « **SAUF**».  
Il est possible de rechercher avec des relations logiques complexes par l’utilisation des parenthèses « **( )**»

Tel que : (Java OU Python) ET exemples.

|  |  |  |
| --- | --- | --- |
| **ET** permet de sélectionner des éléments communs à plusieurs ensembles ; l'opération réalisée est une **intersection**. | **OU** permet de sélectionner la somme des éléments se trouvant séparément dans plusieurs ensembles ; l'opération réalisée est une **union**. | **SAUF** permet de sélectionner des éléments d'un ou de plusieurs ensembles, en excluant les éléments se trouvant dans un ou d'autres  Ensembles ; l'opération réalisée est une **exclusion**. |
|  |  |  |
| A **ET** B  La partie centrale correspond  Aux éléments communs à A et B | A **OU** B  La partie grisée correspond  À la somme des éléments  Appartenant soit à A soit à B | A **SAUF** B  Seule la partie grisée  Répond à la question |

Tableau 2 : Les relations logiques [phalippon, 2003]

1. Recherche par une phrase

L'utilité de cette fonctionnalité réside dans le maintien de mots voisins pour la recherche des titres des ouvrages, des citations, des proverbes, des noms, etc.

On peut comprendre l'importance des relais dans les mots interrogés à travers l'exemple suivant : « la guerre mondiale II », « Linus Thorvald[[91]](#footnote-2) ».

1. Recherche par mots proches

Cette fonctionnalité permet de rechercher des mots proches dans la même phrase ou le même paragraphe, séparés par quelques mots. Par exemple, on utilise les mots « windows »   et « edition » pour rechercher tous les phrases : "Windows xp home edition", "windows vista home basic edition", "Windows Server 2003 R2 Enterprise x64 Edition", etc.

1. Recherche par intervalle

L’importance de cette fonctionnalité se manifeste dans le cas où le mot que nous recherchons est une valeur numérique telle que les statistiques, les années, les versions, etc. par exemple, si nous voulons rechercher tous les versions d’Ubuntu[[92]](#footnote-3) à l’exception de la dernière version 9, la requête sera : Ubuntu [0 to 8].

1. Boosting

**Boosting query elements**, en français : Stimuler les éléments de la requête. On peut spécifier que certaines parties d'une requête sont plus importantes dans le calcul du score d'un document approprié que d'autres. Par exemple, pour spécifier que le terme « Quranic » est deux fois plus important que d'autres termes, alors que le terme « Engine » l’est beaucoup moins :

Quranic^2.0 Search Engine^0.5

1. Les Fonctionnalités additionnelles
2. Suggestion de mots-clés alternatifs

La possibilité de suggérer des mots pour la recherche lorsque nous écrivons un mot d'interrogation inexistant ou une faute orthographique, par exemple : Le moteur de recherche a besoin d'un moyen pour récupérer "Algérie" plutôt que de "Algrie".



Figure 11 : Suggestion de mot alternatif pour « algrie » - Google

1. Suggestion de mots-clés apparentés

Par l’utilisation des plusieurs stratégies, les moteurs de recherche peuvent suggérer des mots alternatifs.

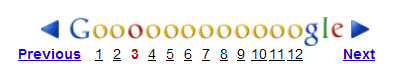
Par exemple : suggérer le mot « Hadith » si l’utilisateur recherche le mot «Coran» car l’apparition du mot « Coran » été, dans la plupart des cas, associé avec le mot « Hadith ».



**Figure 12 : Mots apparentés au mot « coran » - Google**

1. Pagination

C’est la distribution des résultats sur plusieurs pages dans le cas de nombreux résultats et l'affichage complet dépasse l’espace disponible dans une seule page (ex : “l'affichage des résultats 1-20, page 1 de 4”).

  
**Figure 13 : Pagination - Google**

1. Tri (Sorting)

Trier les résultats par ordre d'importance, ou bien par les critères choisis par l'utilisateur.

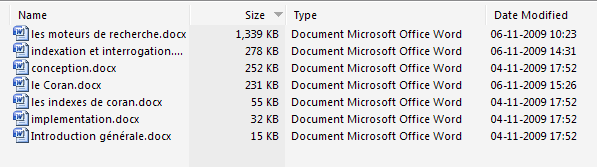


Figure 14 : Tri par volume (size) – recherche dans Windows Explorer

1. Highlight (sur-lignage)

[L](javascript:;)e sur-lignage des mots recherchés dans les résultats pour l’amélioration de la visibilité notamment dans les cas où la recherche n’est pas exacte.



Figure 15 : Highlight (sur-lignage)

1. Les aspects linguistiques Généraux de recherche
2. La recherche par signification

L’origine de ce problème, c'est qu’un mot peut avoir plusieurs sens. Par exemple l’adjectif « light » a plusieurs significations comme : lumineux, léger, faible… Sur cette base, le processus de recherche – s’il ne prend pas en compte le sens du mot – ne fait pas de distinction entre les sens du mot dans :

It was still **light** when we arrived at home.

He was as **light** as a feather to carry.

Leaves were blowing about in the **light** wind.

Le même problème en français, le mot **rayon** a plusieurs significations :

* En [géographie](bword://!!Z8AVT3H7RE,g%C3%A9ographie/), un **rayon** est une subdivision administrative des ex-républiques soviétiques ;
* En [géométrie](bword://!!Z8AVT3H7RE,g%C3%A9om%C3%A9trie/), un [**rayon**](bword://!!Z8AVT3H7RE,rayon%20%28g%C3%A9om%C3%A9trie%29/) est un [segment](bword://!!Z8AVT3H7RE,Segment%20%28math%C3%A9matiques%29/) qui relie le centre d'un [cercle](bword://!!Z8AVT3H7RE,cercle/) ou d'une [sphère](bword://!!Z8AVT3H7RE,sph%C3%A8re/) à un point du pourtour de ce cercle ou de cette sphère. « rayon » désigne également la [longueur](bword://!!Z8AVT3H7RE,longueur/) de ce segment ;
* En [mécanique](bword://!!Z8AVT3H7RE,m%C3%A9canique/), les [**rayon**](bword://!!Z8AVT3H7RE,rayon%20%28technique%29/)**s** d'une [roue](bword://!!Z8AVT3H7RE,roue/) non pleine sont les [tiges](bword://!!Z8AVT3H7RE,tige/) reliant le moyeu à la [jante](bword://!!Z8AVT3H7RE,jante/) ;
* En [optique](bword://!!Z8AVT3H7RE,optique/), un [**rayon**](bword://!!Z8AVT3H7RE,rayon%20lumineux/) est une [courbe](bword://!!Z8AVT3H7RE,courbe/) représentant le trajet de la lumière ;
* Dans les [magasi](bword://!!Z8AVT3H7RE,magasin%20%28commerce%29/)ns, les **rayons** sont les étagères sur lesquelles sont disposés les articles.

La distinction entre ces sens est difficile pour un moteur de recherche. **[al-bawwab, 09]**

1. Recherche par lemme

La lemmatisation est le processus par lequel un mot est réduit à son “lemme” ou à sa forme canonique. Par exemple si on prend le mot « walking » le lemme – ou la forme de base – de ce mot serait « walk ». Ou si o, prend le mot « better » le lemme de ce mot serait « good ». La lemmatisation prend en considération le contexte et la signification du mot lors de la détermination de la forme de base du mot. **[Ferreira da Silva, 07]**

**Remarque :**

**Lemme :**

Le lemme est l’entrée de dictionnaire, il s’agit donc de la forme classique et entièrement vocalisée à laquelle se rattache la forme du texte. Les lemmes composés sont gardés comme composés si le composé lui-même fait l’objet d’une entrée dans les dictionnaires classiques.**[Tuerlinckx, 04]**

**Racine :**

La racine est la suite des consonnes formant le radical du mot ; à chaque racine correspond un champ sémantique. La racine est un élément important dans les langues sémitiques : associée à un schème (pattern), elle forme la base du mot .Les lemmes sont tous rattachés à une racine.**[Tuerlinckx, 04]**

**Lemme-Racine :**

Dans certains cas, sous « racine » se trouve une suite de lettres qui ne constitue pas un radical, mais reprend le lemme dépourvu de ses diacritiques[[93]](#endnote-90) et d’éventuels éléments adventices : ce sont les *lemmes-racines* (noms propres, emprunts, transcriptions, onomatopées, particules, etc.). La case « emprunt » cochée est souvent associée à un lemme-racine.**[Tuerlinckx, 04]**

1. Recherche par synonymes

Les synonymes sont des mots qui portent le même sens, mais ils sont écrits ou proférés de manières différentes. La recherche en utilisant les synonymes aide à trouver des résultats proches du sens qu'on veut **[al-bawwab, 09]** , par exemple, si une recherche du mot « grand » donne des résultats qui contiennent le mot « grand » plus les mots (élevé, haut, éminent, supérieur, éduqué) et d'autres dans les textes indexés. Le tableau suivant montre plus d’exemples :

|  |  |
| --- | --- |
| Mot | Synonymes |
| Véracité | Sincérité, franchise, loyauté, droiture, exactitude … |
| Pardon | Absolution, aumône, grâce, indulgence, miséricorde |
| Chaleur | Ardeur, bouillonnement, chaud, effervescence, élan, entrain, exaltation, excitation, fièvre, flamme, pétulance, trouble… |
| Intelligent | Adroit, astucieux, capable, doué, éveillé, fort, habile, ingénieux, malin… |

Figure 16 : Exemples des synonymes

1. Recherche par antonymes

Les antonymes, sont les mots qui portent des sens opposés, l’utilité de cette fonctionnalité est lorsque, par exemple, on recherche du mot «aller», les résultats doivent contenir les mots (venir, retourner, revenir) et autres mots dans les textes indexés. Le tableau suivant donne plus d’exemples :

|  |  |
| --- | --- |
| Mot | Antonymes |
| Véracité | Fausseté, Hypocrisie, Fourberie, Déloyauté… |
| Pardon | Disgrâce, sévérité, Défaveur, austérité… |
| Chaleur | Fraîcheur, froideur, froidure, tiédeur… |
| Intelligent | Abruti, borné, sot, épais, stupide… |

Figure 17 : Exemples des antonymes

1. Les aspects linguistiques de recherche pour la langue arabe

La langue arabe est l’une des langues sémitiques, et la langue de saint Coran, elle est écrite de droite à gauche. Son alphabet a vingt-neuf[[94]](#footnote-4) lettres consonnes, trois d'entre elles (http://www.isoc.org/inet99/proceedings/posters/085/fig4_1.gif ) sont considérées comme des signes de diacritiques. Optionnellement, l’une des trois diacritiques (علامات الشكل) (http://www.isoc.org/inet99/proceedings/posters/085/fig4_2.gif) peut être placée après certains caractères pour lever l’ambiguïté (dans la prononciation et/ou le sens) quand elle se pose. Dans un texte arabe totalement vocalisé[[95]](#endnote-91), l'absence des diacritiques peut être considérée comme une sokōune[[96]](#endnote-92) (silence) (symbolehttp://www.isoc.org/inet99/proceedings/posters/085/fig4_3.gif). Dans certains cas, une lettre doublée, peut être remplacé par une seule lettre avec le tashdeed[[97]](#endnote-93) (renforcement) (symbolehttp://www.isoc.org/inet99/proceedings/posters/085/fig4_4.gif) placé au-dessus.**[AlKharashi, 99]**

|  |  |
| --- | --- |
| **lettre** | **graphies** |
| hamzä | **ء** |
| *wâw* | **ـو** **و** |
| ‛ayn | **ع** **عـ** **ـعـ** **ـع** |

En outre, il nous importe de signaler que la notion de lettre capitale et lettre minuscule n'existent pas, l'écriture arabe est dite *monocamérale*. En addition, l’arabe est une langue *semi cursive* ; la plupart des lettres s'attachent entre elles, leurs graphies diffèrent selon qu'elles soient précédées et/ou suivies d'autres lettres ou qu'elles soient isolées. Seulement six d’entre-elles ne s'attachent jamais à la lettre suivante : (ا, د, ذ, ر**,** ز,و) et une seule lettre ne s’attache jamais : (ء) **[MESFAR, 08]**

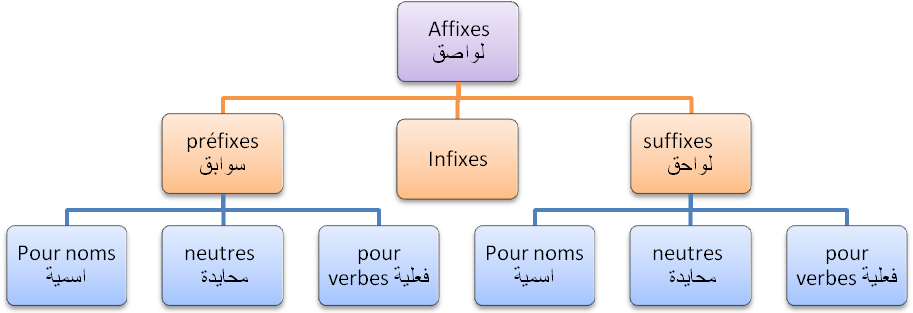
**Figure 18 : 3 types de lettre arabes : une graphie, deux graphies, quatre graphies**

L’arabe est une langue flexionnelle. Elle emploie, pour la conjugaison du verbe et la déclinaison du nom, des indices d’aspect, de mode, de temps, de personne, de genre, de nombre et de cas, qui sont en général des suffixes et préfixes **[Gaudefroy, 75]** **.** Généralement, ces marques flexionnelles permettent de distinguer :

* le mode des verbes : par exemple, pour le verbe " ذهَبَ " (*dahaba* – aller), les formes à l’accompli sont repérables à l’aide de leurs suffixes tel que " ذهَبْتُ " (*dahabatu* – je suis allé) ou de leurs préfixations tel que " أَذهَبُ " (áa*dhabu* – je vais) ;
* la fonction des noms à l’aide des suffixations tels que " رَجُلانِ " (*rağulaáni* – deux hommes au *nominatif*) ou " رَجُلَيْنِ " (*rağulayni* – deux hommes à *l’accusatif* ou *génitif*).**[MESFAR, 08]**

Il y a plusieurs catégories de langages flexionnels, et l’arabes est exactement de la catégorie des langues à "brisure interne" : dans cette catégorie de langues, les consonnes indiquent le sens et les voyelles marquent la flexion du mot, ce système se rencontre surtout dans les langues sémitiques (ex : [arabe](http://fr.wikipedia.org/wiki/Arabe) , [hébreux](http://fr.wikipedia.org/wiki/H%C3%A9breux)) **[Choueiter, 06]**

Morphologiquement, la langue arabe est très riche et fondée sur la structure de patterns[[98]](#endnote-94) et racines[[99]](#endnote-95), la plupart des mots arabes sont générés à partir d’un ensemble fini de racines (environ 7000 racines) transformés en utilisant un ou plusieurs patterns (environ 400-500). Théoriquement, une seule racine arabe peut générer des centaines de mots (substantif, verbes, …). Un mot arabe peut exister sous une centaine de formes dans cent profilés en un texte normal en ajoutant certains suffixes[[100]](#endnote-96) et préfixes[[101]](#endnote-97) (principalement considérées comme des mots vides[[102]](#footnote-5) en langue anglaise). **[AlKharashi, 99]**



**Figure 19 : Classification d’affixes en Arabe**

1. La considération de diacritiques الحركات

La plupart des textes arabes sont écrits sans diacritiques, cela provoque une confusion dans le sens, uniquement le contexte du texte peut lever l’ambigüité.

Par conséquent, les moteurs de recherche basés seulement sur la graphie du mot ne peuvent pas l’enlever.

Si le texte contient le mot (ملك), par exemple. Comment les moteurs de recherche peuvent déterminer le sens de ce mot : Est-il (المَلَك) ou (المُلك) ou (المَلِك) ? **[al-bawwab, 09]**   
Le tableau suivant montre certains mots que ses sens changent en raison du changement de diacritiques :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Le mot sans diacritiques | 1er sens | 2me sens | 3me sens | … |
| القدر | القَدْر | القِدْر | القَدَر | **…** |
| يحرم | يُحَرِّم | يُحْرِم | يَحْرُم | **…** |
| البر | البِرّ | البَرّ | البُرّ | **…** |
| الحجر | الحِجْر | الحَجَر | الحَجْر | **…** |
| العرض | العَرْض | العِرْض | العَرَض | **…** |
| أم | أَمْ | أُمّ | أَمَّ | **…** |
| الملك | المَلِك | المَلَك | المُلْك | **…** |
| من | مِنْ | مَنْ | مَنّ | **…** |

Tableau 3 : Le changement de signification en changeant les diacritiques

En plus de la confusion sur le sens du mot due à l'absence de diacritiques, il y a une plus grande confusion sur la fonction du mot, aussi le contexte du texte seulement peut l’éliminer. Si le mot dans le texte (أخد), par exemple, les moteurs de recherche basée sur la graphie du mot seul ne peuvent pas faire la distinction entre la forme active (أَخَذَ) et la forme passive (أُخِذَ) **[al-bawwab, 09]**

Le tableau suivant montre certains mots qui changent ses fonctions en raison du changement de diacritiques d'alphabets :

|  |  |  |
| --- | --- | --- |
| Le mot sans diacritiques | 1ere fonction | 2eme fonction |
| ملك | (**مَلِكٌ**) اسم | (**مَلَكَ**) فعل... |
| من | (**مَنْ**) اسم موصول | (**مِنْ**) حرف جر... |
| عدوان | (**عُدْوَان**) مفرد | (**عَدُوَّان**) مثنى... |
| فرض | (**فَرَضَ**) فعل ماضٍ | (**فَرْضٌ**) مصدر... |
| عن | (**عَنَّ**) فعل ماضٍ | (**عَنْ**) حرف جر... |
| أدر | (**أَدْرِ**) مضارع مجزوم | (**أَدِرْ**) فعل أمر... |
| أحسن | (**أَحْسَنَ**) فعل ماضٍ | (**أَحْسَنُ**) اسم تفضيل... |
| علم | (**عَلِمَ**) فعل مجرَّد | (**عَلَّمَ**) فعل مزيد... |
| غلبت | (**غَلَبَتْ**) مبني للمعلوم | (**غُلِبَتْ**) مبني للمجهول... |

Tableau 4 : Le changement de fonction en changeant les diacritiques

1. Les préfixes du mot

C’est un autre type de confusion, le mot peut être lu de plusieurs manières. Par exemple : dans le mot (وعد), la lettre wâw [[103]](#endnote-98)(واو) peut appartenir au mot comme : (وَعَدَ : أي الفعل وَعَدَ يَعِدُ), ou bien n'appartient pas au mot comme : (وَعَدَّ = **وَ** + عَدَّ : أي الفعل عَدَّ يَعُدُّ)

Ainsi, les moteurs de recherche basés seulement sur la graphie du mot, ne peuvent pas faire la distinction entre ces deux cas. Ce type de confusion n'existe pas dans la langue anglaise, par exemple :

La traduction de l’expression"والكتاب" en anglais est une phrase de 3 mots libres "and the book".**[al-bawwab, 09]** La table suivante contient d’autres exemples de mots qui peuvent être lus de deux façons :

|  |  |  |
| --- | --- | --- |
| Le mot | 1ère façon | Me façon2 |
| بطاقة | بِطَاقَة (اسم) | **بِطَاقَة** = بـِ + طَاقَة (حرف جر + اسم) |
| وهم | **وَهْمٌ** (مصدر) | **وَهُمْ** = وَ + هُمْ (حرف عطف+ ضمير متصل) |
| فرمت | **فَرَمَتْ** (من فَرَمَ يَفْرِمُ) | **فَرَمَتْ** = فـَ + رَمَتْ (من رَمَى يَرْمِي) |
| لسعت | **لَسَعَتْ** (من لَسَعَ يَلْسَعُ) | **لَسَعَتْ** = لـَ + سَعَتْ (من سَعَى يَسْعَى) |
| أبعث | **أَبْعَثُ** (فعل مضارع) | **أَبَعَثَ** = أَ + بَعَثَ (همزة استفهام + فعل ماضٍ) |
| كرب | **كَرِبَ** (فعل ماضٍ) | **كَرَبّ** = كَ + رَبّ (حرف تشبيه + اسم) |
| الحق | **اِلْحَقْ** (فعل أمر) | **الْحَقّ** = الْ + حَقُّ (ال التعريف + اسم) |

Tableau 5 : Des exemples de préfixes

1. Les suffixes du mot

Les suffixes sont les pronoms relatifs[[104]](#endnote-99) (الضمائر المتصلة [[105]](#footnote-6)) au mot.

La confusion ici ressemble à la confusion des préfixes, cette confusion est que nous pouvons lire le mot de deux façons, selon que les lettres appartiennent au mot ou non.

1. Par exemple, dans le mot (وله), la lettre ha’ (هاء) peut :  
   appartenir au mot comme :( وَلِهَ : الفعل وَلِهَ يَوْلَهُ) ;  
   n’appartient pas au mot comme : (وَلِّهِ = وَلِّ + ـهِ = فعل الأمر من وَلَّى يُوَلِّي).

De même que la confusion précédente, ce type de confusion n'existe pas dans la langue anglaise, par exemple : La traduction de " أَجَبْنَاهُ" en anglais est une phrase de 3 mots séparés "we answered him". **[al-bawwab, 09]**

La table suivante contient d’autres exemples :

|  |  |  |
| --- | --- | --- |
| Mot | 1ère façon | Me façon2 |
| برك | **بَرَكَ** (فعل ماضٍ) | **بَرَّكَ** = بَرَّ + كَ (فعل ماضٍ+ضمير متصل) |
| بكم | **بُكْمٌ** (جمع أبكم) | **بِكُمْ** = بـِ + كُمْ (حرف جر+ضمير متصل) |
| تلفن | **تَلْفَنَ** (فعل رباعي) | **تَلِفْنَ** = تَلِفْ + نَ (فعل ماض+نون النسوة) |
| نبت | **نَبَتَ** (فعل ماضٍ) | **نَبَتْ** = نَبَ + تْ (فعل ماضٍ من نبا ينبو + تاء التأنيث الساكنة) |
| برهن | **بَرْهَنَ** (فعل رباعي) | **بِرِّهِنَّ** = بِرِّ + هِنَّ (اسم+ضمير متصل) |
| لي | **لَيٌّ** (مصدر لَوَى يَلْوِي) | **لِي** = لـِ + ي (حرف جر+ضمير متصل) |
| همهم | **هَمْهَمَ** (فعل رباعي) | **هَمُّهُمْ** = هَمُّ + هُمْ (اسم+ضمير متصل) |

Tableau 6 : Des exemples de suffixes

1. Recherche par la racine arabe

La racine du mot arabe se compose, dans plupart des cas, de trois lettres, parfois de quatre, et plus rarement de cinq lettres.

Le principal avantage obtenu par la recherche au niveau de la racine, réside dans l'obtention de toutes les formes qui peuvent se produire par la dérivation de cette racine. On sait que ces formes multiples du mot remonte à un certain nombre de considérations telles que : les méthodes de la conjugaison, les états flexionnels[[106]](#endnote-100), les relations avec des affixes, etc.

La recherche par la racine offre tous les mots qui reviennent à cette racine : des verbes (muğarrada [[107]](#endnote-101) et mazīda [[108]](#endnote-102); passives[[109]](#endnote-103) et actives[[110]](#endnote-104) ; confirmés[[111]](#endnote-105) et non confirmés ; dans le passé, le présent[[112]](#endnote-106) ou l’impérative ; des morphèmes (originaux[[113]](#endnote-107), mīmiya[[114]](#endnote-108), de forme[[115]](#endnote-109), de fois[[116]](#endnote-110)) ; des noms déclinés (nom du fā‛il[[117]](#endnote-111), nom du maf‛ūl[[118]](#endnote-112), circonstances de lieu et de temps[[119]](#endnote-113), noms d’instrument[[120]](#endnote-114),exagération de nom du fā‛il[[121]](#endnote-115), quasi adjectif[[122]](#endnote-116)) dans le singularité, tathniya[[123]](#endnote-117) (pluralité de deux choses), pluralité ; Raf’, Nasb et ğarr; la définition, l’indéfinition et l’annexion[[124]](#endnote-118), etc.  
Exemple : la recherche par la racine (ح س ب) offre les mots suivants :

حَسَبَ يَحْسُبُ، حَسِبَ يَحْسَبُ، حَسُبَ يَحْسُبُ، حُسِبَ يُحْسَبُ... (أفعال مجردة)

حَاسَبَ يُحَاسِبُ، احْتَسَبَ يَحْتَسِبُ، تَحَسَّبَ يَتَحَسَّبُ، تَحَاسَبَ يَتَحَاسَبُ... (أفعال مزيدة)

حَسْب، حِسَاب، حُسْبَان، حِسْبَة، مُحاسَبَة، احْتِساب، تَحَسُّب، تَحاسُب... (مصادر أصلية)

حَسْبَة، حِسْبَة، مَحْسَب... (مصدر المرة، ومصدر الهيئة، والمصدر الميمي)

حاسِب، مَحْسُوب، حَسَّاب، مُحاسِب، مُحْتَسِب، مُتَحَسِّب، حَسِيب... (أسماء مشتقة)

Chacun de ces mots a des dizaines ou des centaines de formes différentes qui peuvent apparaitre dans les textes.   
Par conséquent, si nous demandons, au moteur de recherche, la recherche par la racine du mot (فحسب), nous allons obtenir des textes qui contiennent tous les mots dérivés de la racine (ح س ب). Il est possible que nous trouvions des mots dérivés que nous ne connaissons pas.

1. La conjugaison des verbes[[125]](#endnote-119)

Par exemple, le verbe قال (dire) dans le passé, apparait dans les textes sous différents formes lorsqu’il est relié avec les pronoms relatifs de Raf’[[126]](#endnote-120) (ضمائر الرفع المتصلة) comme :

(قُلْتُ، قُلْنَا، قُلْتَ، قُلْتِ، قُلْتُمَا، قُلْتُمْ، قالَ، قَالَتْ، قَالا، قَالَتا، قَالُوا، قُلْنَ).

En se basant sur ça, la recherche du verbe (قال) avec tous les pronoms, nécessite la recherche de douze formes que nous avons citées. Bien qu'il soit dans la langue anglaise, cette recherche est commandée par un seul mot (said). Cela est dû à deux raisons : premièrement : il n’y a pas des pronoms relatifs, on dit : (I said, you said, he said, she said, we said, they said).

Deuxièmement : le verbe lui-même ne change pas sa forme en changeant de pronoms, il est toujours (said) avec toutes les pronoms.**[al-bawwab, 09]**

1. La déclinaison des noms[[127]](#endnote-121)

C'est une recherche des mots qui ont le même morphème[[128]](#footnote-7) [[129]](#endnote-122), en tapant directement le morphème ou un nom décliné, et puis retirer sa morphème et l’utiliser pour la recherche.

Par exemple, si on effectue une recherche de mot « (islam) إسلام», le résultat doit contenir les mots أسلم, مسلمون, أسلموا, etc. et pas seulement وإسلام et الإسلام comme dans les moteurs de recherche standards. **[al-bawwab, 09]**

On peut utiliser la recherche par racine pour trouver les noms déclinés, mais cette solution produit deux autres problèmes. Le premier est que les résultats contient les verbes aussi, le deuxième est que par exemple si on veut rechercher les noms déclinés du mot التفريط (perte), on peut trouver les noms déclinés du mot الإفراط (outrance) dans les résultats de recherche par la méthode de racine فرط, mais ce n’est pas correct puisque ces deux mots التفريط et الإفراط sont des antonymes.

1. Recherche par la nature[[130]](#endnote-123) des mots arabes

Les types du mot dans la langue arabes sont : nom, verbe, morphème (مصدر), adjectif, pluriel, Nisba[[131]](#endnote-124)( نسبة),tasġīr[[132]](#endnote-125) ( تصغير) …

Les moteurs de recherche, grâce à la nature des mots, permettent de restreindre et limiter l’intervalle de recherche. Si on limite la recherche aux verbes, par exemple, le moteur de recherche va aussi limiter ses résultats aux verbes seulement, et ignore tous les autres types comme les noms, les morphèmes (المصادر), les adjectifs, les pluriels, etc.**[al-bawwab, 09]**

**Exemple :**

Si nous demandons la recherche des verbes dérivés de la racine (ح س ب), nous obtiendrons les textes contenant les verbes de cette racine : muğarradaet mazīda, passives[[133]](#endnote-126) et actives[[134]](#endnote-127), confirmés et non confirmés, dans le passé, le présent et l’impérative, tels que :

حَسَبَ، حَسَبْتُ، حَسَبْتُمْ، حُسِبَ، حَسِبَتَا، حَسِبْنَ... (أفعال ماضية مجردة)

يَحْسُبُ، تَحْسُبُ، تَحْسُبِينَ، يَحْسُبُونَ، يَحْسُبْنَ... (أفعال مضارِعة مجردة)

حاسَبَ، تَحاسَبَ، احْتَسَبَ، تَحَسَّبَ، احْتَسَبُوا... (أفعال ماضية مزيدة)

يُحاسِبُ، يُحاسِبُون، تَتَحاسَبُون، تَحْتَسِبِينَ، تَتَحَسَّبانِ... (أفعال مضارعة مزيدة)

احْسُبْ، احسُبِي، احسُبُوا، احْتَسِبُوا، حاسِبُوا، تَحَسَبُوا... (أفعال أمر)

حُسِبَ، حُوسِبَ، حُوسِبَتْ، حُوسِبُوا، حُوسِبْنَ... (أفعال ماضية مبنية للمجهول)

1. Recherche par les patterns

Le but de l’utilisation du pattern du mot est de limiter la recherche encore plus. Si la recherche a été limité aux verbes, par exemple, et on limite aussi par un pattern spécifié, le moteur de recherche limite ses résultats à ce pattern, et néglige les résultats d'autres patterns.   
Exemple : Si nous demandons la recherche avec le pattern (**فُعِلَ**) du verbe qui à la racine (ح س ب), nous allons obtenir les textes, dans lequel il y a le verbe (**حُسِب**) et sans les autres verbes.   
Ainsi, nous constatons que l’intervalle de recherche est très large par la recherche par la racine du mot, puis il se rétrécit par la sélection du type, et puis il se rétrécit encore plus par sélection du pattern.**[al-bawwab, 09]**

1. Respecter les différences de Hamza[[135]](#endnote-128)

|  |  |  |
| --- | --- | --- |
| Les formes | Exemples | Les cas |
| **ء** | آلاء | Sur la ligne |
| **أ، إ،آ** | رأيت،إنّ،آمنّا،سبأ | Sur *álif* |
| **ئ** | بئر، عندئذ | Sur nabira[[136]](#endnote-129) |
| **ؤ** | الموؤدة،المؤصدة | Sur wâw |

La Hamza est la lettre la plus souvent utilisé parmi les lettres arabes, mais beaucoup de gens et même les auteurs se trompent dans son écriture **[محبك, 08]** , elle change sa graphie et ses règles d’écriture en fonction de sa position dans la phrase, nous ne détaillerons pas ici les règles d’écriture qui ne font pas partie de notre sujet, mais nous mentionnerons la plupart des graphies existent dans le script standard.

Donc, il est important de prévoir l’option qui néglige les formes de Hamza pour les utilisateurs.

**Tableau 7 : Les différences de graphies de Hamza.**

1. Respecter la différence entre Hâ’[[137]](#endnote-130) et tâ’ marbûtä[[138]](#endnote-131) et entre Yâ’[[139]](#endnote-132) et Alif maqsûrä[[140]](#endnote-133)

De nombreux utilisateurs, ne respectent pas l’écriture correct de la*tâ’ marbûtä* et de Hâ’, cela est due à la similitude de la graphie et aussi à la non-prononciation de Ta’ ficelé dans le dernier mot de la phrase (en raison d’une similitude de prononciation dans ce cas)

La même chose pour *Yâ’* et *Alif maqsûrä,* mais il n’y a aucune similitude de prononciation dans l’arabe standard sauf dans des récitations de warch[[141]](#endnote-134) qui contentent la règle de 'imāala[[142]](#footnote-8) [[143]](#endnote-135) الإمالة))

1. Autres options de recherche dans le Coran
2. Recherche par le script othmani[[144]](#endnote-136)

L’écriture de certains mots est différente dans le script othmani que dans le script standard, par exemple : le mot "نعمة" est écrit dans le script othmani "نعمت"

La table suivant contient d’autres exemples :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Script standard** | **Script othmani** | **Les emplacements** | **Les différences** | |
| **العالمين** | **العلمين** | جميع مواضعها في القرآن | حذف الألف | |
| **الغاوون** | **الغاون** | (الشعراء:94)و موضع آخر | حذف الواو | |
| **النبيين** | **النبين** | جميع مواضعها في القرآن | حذف الياء | |
| **الليل** | **اليل** | جميع مواضعها في القرآن | حذف اللام | |
| **ننجي** | **نجي** | (الأنبياء:88) | حذف النون | |
| **وجيء** | **وجائ** | (الزمر:69) و موضع آخر | زيادة الألف | |
| **سأريكم** | **سأوريكم** | (الأعراف:145) | الزيادة في الواو | |
| **بأيد** | **بأييد** | (الذاريات:47) | الزيادة في الياء | |
| **لتنوء** | **لتنوأ** | )القصص : 76) | الهمزة | |
| **يبدأ** | **يبدؤا** | (يونس:4) | الهمزة | |
| **وإيتاء** | **وإيتائ[[145]](#footnote-9)** | (النحل:90) | الهمزة | |
| **الصلاة،الزكاة** | **الصلوة، الزكوة** | جميع مواضعها في القرآن | مجيء الألف واواً | |
| **يا أسفا** | **يأسفى** | (يوسف:84) | مجيء الألف ياءًا | |
| **عن ما** | **عما** | جميع المواضع | وصل عن- ما | |
| **بئس ما** | **بئسما** | جميع المواضع | وصل بئس- ما | |
| **لكي لا** | **لكيلا** | جميع المواضع | وصل كي-لا | |
| **كلما** | **كل ما** | (النساء:91)و موضعين آخرين | فصل كلما | |
| **أينما** | **أين ما** | (البقرة:148)و 7 مواضع أخرى | فصل أينما | |
| Tableau 8 : Exemples des différences entre le script standard et le script othmani [islamweb, 03] | | | | |

1. Recherche par des index spéciaux

C’est offrir la possibilité de générer différents index à des sujets de Coran, ou en fonction de divers critères tels que le lieu de révélation (Mecquois et Médinois), diverses lectures, des significations, les pluriels …etc.

Nous aborderons les différents index de Coran existantes dans le chapitre IV

1. Des options coraniques structurelles

Comme nous avons vu dans le premier chapitre (**IV.C**) « La structure du mushaf », le Coran se subdivise en Juz’[[146]](#endnote-137), le Juz’ en deux Hezbs, le Hezb[[147]](#endnote-138) en deux Nesfs et le Nesf[[148]](#endnote-139) en deux Rubu’s et le Rubu’[[149]](#endnote-140) en deux Thumuns[[150]](#endnote-141). Nous pouvons aussi le diviser en surates et la sourate[[151]](#endnote-142) aux Ayas. En plus de il est possible de le diviser en Ruku’s[[152]](#endnote-143) et en phrases …  
 La recherche en sélectionnant l’intervalle d'ayas est une option essentielle qui existe dans tous les moteurs de recherche, même si elle est limitée.  
Un exemple d’utilisation de cette option est quand l'utilisateur veut rechercher au début de chaque sourate de Juz’ ‘amma [[153]](#endnote-144)(عمّ) commençant avec un jure, alors on peut choisir le Juz’ et l'aya N°1 et rechercher avec l’expression « وال\* »[[154]](#footnote-10).

1. Des options statistiques

La fourniture des options statistiques peut être avantageuse aux érudits de Shari'a[[155]](#endnote-145) tel que savoir qu’elle est la fréquence d'un mot particulier, d’une lettre particulière, d’une phrase particulière, etc.

Ces options peuvent répondre à quelques types de questions telles que :

* + - Combien de mot « الله » dans surate « المجادلة » ?
    - Combien de mot « بحر » dans le Coran ? Le mot « بحار » sera aussi compté.
    - Quel est le mot le plus fréquent dans le Coran ?
    - Quels sont les mots dont le nombre est supérieur à 1000 fois ?
    - Quels sont les mots dont le nombre est supérieur à 20 fois dans Surate « البقرة » ?
    - Quels sont les mots dont le nombre est inférieure à 10 fois ?
    - Quels sont les mots qui apparaissent 20 fois dans la sourate « الكهف » ?
    - Quelle est la sourate, qui contient le plus grand nombre des noms d’Allah[[156]](#endnote-146) (ألفاظ الجلالة) ?
    - Quel est l'aya la plus longe ?
    - Quelle est la sourate la plus longue ?

1. Extension de recherche

Il est recommandé que le moteur de recherche du Coran puisse s’étendre aux sciences coraniques telles que (tafssīr[[157]](#endnote-147), traduction …) ou même à des sites contenant des ayas coraniques.

1. Un algorithme proposé pour la recherche arabe

Cet algorithme est proposé par **[al-bawwab, 09]** .il est basé sur la fourniture de trois informations morphologiques aux mots de l'index de moteur de recherche qui sont : la racine du mot, son type et son pattern à l’exception de : noms propres, noms d’animaux, de plante …, les mots étrangers arabisés, les particules tels que « إذا », « لو », les pronoms libres et les nombres. Ces mots représentent environ 35% des mots de textes arabes. Les informations morphologiques n’apparaissent pas dans les textes résultants, mais le moteur de recherche les utilise dans ses opérations de recherches.  
Les principaux avantages de cet algorithme sont :

* **Réduire le nombre de résultats** : l’un des avantages de l'algorithme de recherche par (racine, type, pattern) est la réduction de nombre de résultats de la recherche en excluant les résultats non requis. Car lorsqu’on recherche un mot qui peut prendre deux formes où la présence de l’une est beaucoup plus importante que de l'autre, les résultats de la recherche avec la première forme dominent la seconde, et nous serons en face d’un nombre de résultats tellement grand qu’on ne peut pas accéder aux textes qui contiennent la seconde forme. L’exemple suivant illustre cette situation :

Si on cherche le mot « وَفِيّ », les moteurs de recherche basés sur le ciblage des mots clés donnent des millions de résultats qui contient le mot « وَفِيّ » avec ses deux formes (و + في) qui est la forme la plus courante et وَفِيّ (de la fidélité) qui est la moins courante. En revanche, avec la recherche (racine, type, pattern), on obtenir exclusivement les textes qui contiennent le mot « وَفِيّ ».

* **Eviter la divergence des résultats** : la recherche par (racine, type, pattern) restreint les résultats de la recherche à la forme voulue du mot. Vu qu’en recherchant un mot portant plusieurs formes, les résultats de recherche comprendront la totalité de ces formes sans aucun ordre, ou séparation entre eux. Il est impossible d’accéder aux textes qui contiennent le mot avec la forme requise.

Exemple : Si on recherche le mot « فَهْم », les moteurs de recherche basés sur le ciblage des mots clés donnent des millions de textes contenant ce mot avec ses différentes formes tels que « فَهْم », « فُهِّمَ », « فَهَمَّ » … Tandis que la recherche par (racine, type, pattern) recherche avec la racine (ف هـ م), le type (morphème) et le pattern (فَعْل) et on n’obtient en résultat que les textes qui contiennent le mot (فَهْم).

* **Créer de nouveaux types de recherche**: ces types de recherche sont générés en contrôlant l'une des trois composantes : la racine, le type ou le pattern. Voici un exemple :

Si on recherche la racine (ح س ب), et le type (nom d’instrument) sans préciser le pattern, nous allons obtenir les textes contenant les noms d’instrument décliné de cette racine, tels que : « حاسُوب », « مِحْسَب », « حاسِبَة » …

De cette façon, nous n’avons plus besoin de réinterroger autant de fois qu’il ya de patterns. En plus, on peut ne pas connaitre tous les patterns du nom d’instrument.

On note que la recherche (racine, type, pattern) est complémentaire à la recherche basée sur le ciblage des mots clés. Donc, on peut bénéficier des avantages des deux méthodes pour obtenir des résultats efficaces. L'exemple suivant montre comment profiter de ces deux types de recherche : si on veut rechercher le mot « مَلِك » avec toutes ses formes, on demandera la recherche de la racine (م ل ك), le type (nom), et le pattern (verbe) en choisissant ‘pas de correspondance de mot’. On obtiendra les textes qui contiennent le mot (مَلِك), ainsi que les textes qui contiennent les mots (وَمَلِكُهُ), (مَلِكُنا), (للمَلِك) …

Par contre, si nous voulons rechercher le mot (مَلِك) de cette façon seulement, on fait la même chose, mais en choisissant cette fois-ci la correspondance de mot, on obtient seulement les textes qui contiennent le mot (مَلِك).**[al-bawwab, 09]**

1. Conclusion

La comparaison que nous avons faite sur les moteurs de recherche dans le Coran, nous montre la vulnérabilité de ces moteurs et la limitation des méthodes de recherche et des fonctionnalités. Cela peut être dû au fait que la plupart des moteurs de recherche ne sont pas open source ou n'ont pas été créés pour être des moteurs de recherche spécialisés.

L'implémentation d'un moteur de recherche sur le Coran fondée sur des bases modernes, et comprenant toutes les caractéristiques de la recherche qui existent aujourd'hui et la plupart des fonctionnalités supplémentaires et Open Source, unifie les efforts et les rend plus efficaces.

Dans le chapitre suivant, nous expliquerons les étapes de fonctionnement d’un moteur de recherche plus en détails.

# Chapitre III

**Le fonctionnement des moteurs de recherche**

1. Introduction

Notre travail rentre dans le domaine de la recherche d’informations, car il vise à concevoir un moteur de recherche, dans ce chapitre nous allons aborder comment les moteurs de recherche fonctionnent en expliquant ses principaux composants.

L’exploration est la partie qui alimente le moteur de recherche par les documents qu’elle collecte, mais avec le volume d’informations qui deviennt de plus en plus grand, il est nécessaire de développer des méthodes de recherche, seule l’indexation est en mesure d’accélérer la recherche dans des systèmes très volumineux tels que le Web, car elle anticipe la recherche en extrayant les mots clés et en les organisant.

Pour que les résultats de la recherche soient satisfaisants, on doit bien calculer la pertinence des résultats par rapport à la requête, cela est effectué au cours de l’interrogation. L’interrogation doit aussi permettre d’exprimer des questions simples aussi bien que des questions complexes.

La qualité de la recherche est directement liée à la qualité de l’exploration, l’indexation et l’interrogation, ces trois opérations peuvent donc être considérées comme étant le cœur de moteur de recherche, l’objectif de ce chapitre est de définir les principaux concepts de ce domaine, en commençant par définir l’exploration, puis étudier l’indexation, ses méthodes et ses étapes, ensuite nous expliquerons le processus de l’interrogation et la notion de pertinence.

1. Définitions
2. Thésaurus

"C’est **une liste structurée de concepts** destinés à représenter de manière **univoque** le **contenu** de documents et de questions dans un système documentaire déterminé et à assister l’utilisateur dans l'indexation des documents et des questions ; les concepts sont extraits d'une **liste finie**, établie *a priori* ; seuls les termes figurant dans cette liste peuvent être utilisés pour indexer les documents et les questions ; l'assistance à l'utilisateur est apportée par la structure sémantique du thésaurus : relations d'équivalence, de hiérarchie et d'association essentiellement."**[SLYPE, 87]**

"Langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lesquels les notions sont représentées par des *termes* d'une ou plusieurs langues naturelles et les relations et les notions par des signes conventionnels".**[AFNOR, 89]**

1. Mot clé (keyword)

Mot ou groupe de mots choisi en vue de représenter le contenu d’un document, et de le retrouver lors d’une recherche documentaire. Il peut être issu du document (titre, texte, résumé,...) ou d’un vocabulaire contrôlé.**[Hensens, 98]**

1. Descripteur (descriptor)

Mot clé choisi parmi un ensemble de termes équivalents pour représenter sans ambiguïté un concept. Il fait en général partie d’un vocabulaire organisé et hiérarchisé de type "thésaurus".**[Hensens, 98]**

1. Document :

Un document peut être un texte, un morceau de texte, une page web, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d’utilisateur. Pour les documents textuels, il existe beaucoup de formes quant à leur spécification. Un document peut être un texte sans aucune structuration (on l’appelle aussi plein texte) ; il peut aussi être un texte avec une partie structurée (document partiellement structuré ou semi structuré) ou complètement structuré.**[AMROUCHE, 08]**

1. Requête :

Une requête exprime le besoin d’information d’un utilisateur. Divers types de langages d’interrogation ont été proposés pour formuler une requête. Une requête peut être exprimée :

* En langage naturel ou quasi naturel **[Robertson, 97]** **[Salton, 71]** (exemple, "trouver toutes les usines de fabrication de voitures et leurs adresses”)
* Dans un format structuré, appelé aussi interrogation en langage booléen **[Bourne, 79]** (exemple, ”voitures et marque et usines”)
* En langage graphique à partir d’une interface graphique **[Lelu, 92]** .

1. Relevance

*Relevance (Pertinence)* est un mot qui signifie tout simplement de renvoyer l'information considérée comme la plus utile au sommet d'une liste de résultats. Si la définition est simple, obtenir un programme à la pertinence d'ordinateur n'est pas une tâche facile, principalement parce que la notion d'utilité est difficile pour une machine de comprendre.**[Bernard, 08]**

1. Recherche plein texte

La recherche plein texte (‘full-text search’ en anglais) est une technologie basée sur la recherche de documents qui correspondent à un ensemble de mots.

Bien qu'elle semble aussi bête, la recherche plein texte est en fait est plus commune que l’on ne pense, nous l’avons probablement utilisé aujourd'hui. La plupart des moteurs de recherche Internet comme Google, Yahoo et Altavista utilisent des moteurs de recherche plein texte au cœur de leur service. La différence entre chacune d'entre eux est les secrets de recette (et parfois pas si secret), tels que l'algorithme de Google PageRank. PageRank modifie l'importance d’une page web donnée (au niveau des résultats) en fonction du nombre de pages web qui pointent vers elle et à quel point ils sont importants.

Mais attention, ces moteurs de recherche Web sont beaucoup plus qu’un noyau de recherche plein texte : ils ont une interface Web UI, explorent le Web pour trouver des nouvelles pages ou des pages existantes et ainsi de suite. Ils fournissent des spécifications d’entreprises en se basant sur le noyau d'un moteur de recherche plein texte. Étant donné un ensemble de mots (la requête), l'objectif principal de la recherche plein texte est de donner accès à tous les documents correspondants à ces mots. Puisque l'analyse séquentielle de tous les documents pour trouver les mots correspondants est très inefficace, un moteur de recherche plein texte (son noyau) est divisé en deux opérations principales : l'indexation des informations dans un format efficace et la rechercher des informations pertinentes à partir de cet index pré-calculé. De la définition, on peut voir clairement que la notion de mot est au cœur de la recherche plein texte, c’est la pièce atomique d'information que le moteur manipule.**[Bernard, 08]**

1. Exploration
2. Définition

L’exploration (Web crawling en anglais) est le processus par lequel nous recueillons des pages du Web, afin de les indexer et soutenir un moteur de recherche. L'objectif de l'exploration est de recueillir rapidement et efficacement autant de pages web utiles que possible avec la structure des liens qui les interconnecte. L’explorateur est parfois nommé araignée (spider). **[Manning, 09]**

Ce processus se situe dans une phase qui précède l’indexation, *voir la figure 20.*

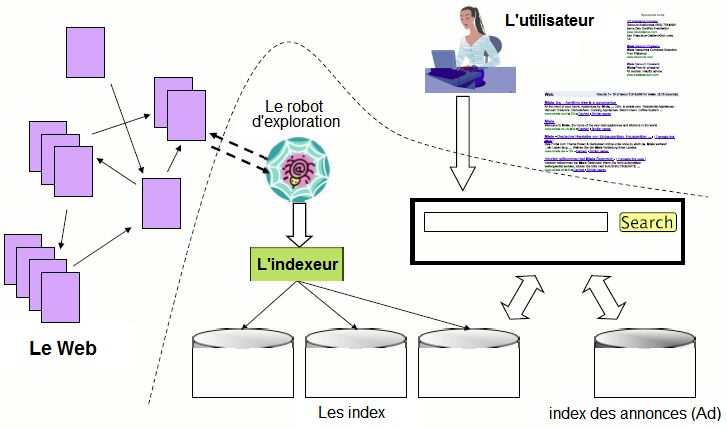


Figure 20 : Le robot d’exploration dans un moteur de recherche web.

1. Les caractéristiques des robots d'exploration

Les caractéristiques des robots d'exploration son devisées en deux catégories : les caractéristiques qu’un robot d'exploration doit fournir (obligatoirement) et les caractéristiques qu'il devrait fournir (moins importantes).**[Manning, 09]**

1. Les caractéristiques obligatoires

* **La robustesse :** Le Web contient les serveurs qui créent des pièges d'araignée, ce sont des générateurs de pages web qui provoquent une confusion dans les robots d'exploration pour qu’ils restent coincés en cherchant un nombre infini de pages dans un domaine particulier. Les robots d'exploration doivent être conçus pour résister à ces pièges. Certains de ces pièges peuvent être dus à la mauvaise conception du cite.
* **La politesse :** Les serveurs web ont des politiques implicites et explicites pour réglementer la vitesse à laquelle un robot d'exploration peut les visiter. Ces politiques de politesse doivent être respectées.

1. Les caractéristiques moins importantes

* **Distribué :** Le robot d'exploration devrait avoir la capacité d'exécuter de manière ‘distribuée’ sur plusieurs machines.

* **Evolutif** : L'architecture du robot d'exploration devrait permettre l'élargissement de la vitesse d'exploration en ajoutant des machines supplémentaires et de bande passante.

* **Performance et efficacité**: Le système d’exploration devrait faire un usage efficace des différentes ressources du système, tels que le processeur, le stockage et la bande passante du réseau.

* **La qualité**: Etant donné qu'une grande partie des pages web sont d'une utilité médiocre pour servir les besoins des requêtes des utilisateurs, le robot d'exploration devrait être biaisé à la récupération des pages "utiles" tout d'abord.

* **La nouveauté**: Dans de nombreuses applications, le robot devrait fonctionner en mode continu: il devrait obtenir de nouveaux exemplaires des pages déjà téléchargées. Un explorateur de moteur de recherche, par exemple, peut ainsi faire en sorte que l'index du moteur de recherche contient une représentation assez récente de chaque page web indexée. Pour une telle exploration continue, un robot d'exploration doit être en mesure d'explorer une page avec une fréquence qui se rapproche du taux de changement de cette page.

* **L’extensibilité**: Les robots d'exploration devraient être conçu pour pouvoir évoluer à plusieurs égards : faire face aux nouveaux formats de données, de nouveaux protocoles de chercher, et ainsi de suite. Cela exige que l'architecture du robot d'exploration soit modulaire.

1. Indexation

Pour que le coût de la recherche soit acceptable, il convient d’effectuer une étape primordiale sur la base documentaire. Cette étape consiste à analyser chaque document de la collection afin de créer un ensemble de mots-clés : on parle de l’étape d’indexation. Ces mots-clés seront plus facilement exploitables par le système lors du processus ultérieur de recherche. L’indexation permet ainsi de créer une représentation des documents dans le système. Son objectif est de trouver les concepts les plus importants du document (ou de la requête), qui formeront le descripteur du document.**[Sauvagnat, 05]**

1. Définition

L’indexation n’a pas une seule définition, la norme AFNOR NF Z 47-102 1996 définit l’indexation comme *«*l’opération qui consiste à décrire et à caractériser un document à l’aide de représentations des concepts contenus dans ce document, c’est-à‑dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse ». **[ABAR, 09]**

Pomart et alii **[Pomart, 97]** définissent l’indexation comme étant une opération ayant « pour but de faciliter l’accès au contenu de documents ou d’un ensemble de documents à partir d’un sujet ou d’une combinaison de sujets (ou toutes autres entrées utiles à la recherche). Cela s’applique aussi bien à l’élaboration d’index situés généralement en fin d’ouvrage qu’à l’usage des langages documentaires pour analyser le contenu d’une collection de documents et permettre par la suite, grâce aux fichiers où à la banque de données ainsi alimentée, la recherche des informations répondant à une préoccupation particulière. Ainsi la notion d’indexation se trouve‑t‑elle au cœur de la pratique documentaire : elle constitue l’opération centrale du traitement des documents au sein de la chaîne documentaire classique (repérage, et acquisition des documents ; traitement et exploitation ; recherche et diffusion) ».**[Mustafa, --]**

L’indexation est le plus souvent utilisée pour effectuer des recherches d’informations. Mais elle peut servir également dans d’autres domaines tels que la classification automatique des documents, proposer des mots-clés, faire une synthèse automatique des documents, calculer les co-occurrentes des termes, générer des résumés automatiquement, etc. (*voir la figure 21*) **[ABAR, 09]**

**Base de données**

**Index**

Recherche d’information

Synthèse automatiques

Extraction de mots clés

Classification de documents

Documents

**Figure 21 : Processus d’indexation.**

1. Modes d’indexation
2. L’indexation manuelle

L’indexation manuelle est réalisée par l'expert humain (documentaliste ou spécialiste du domaine) qui analyse le contenu du texte pour identifier les termes représentatifs du document.

L’indexation manuelle garantit une meilleure pertinence dans les réponses, car elle permet de repérer d’une façon plus précise les mots clés décrivant un document.

Elle présente toutefois plusieurs inconvénients, il y a le problème du vocabulaire utilisé et la dépendance par rapport à l’état de connaissance de l’indexeur sur le sujet c’est-à-dire le même document peut être indexé de plusieurs manières (selon la vision de la personne qui fait l’indexation), et un indexeur à deux moments différents peut présenter deux termes distincts pour représenter le même concept.

L’inconvénient majeur de cette méthode reste le coût en matière du temps, cette méthode n’est pas donc appropriée quand le nombre de documents à indexer est important. **[Sauvagnat, 05]** **[ABAR, 09]** **[AMROUCHE, 08]**   
L’indexation manuelle se base sur quatre points essentiels **[CHARTRON, 89]** :

* la lecture du document dans son intégralité pour l'élaboration de l'indexation ;
* la prise en compte de choix des descripteurs, des objectifs du centre de documentation (applications) et des besoins des utilisateurs ;
* la complémentarité permanente entre les termes de l'indexation manuelle et le résumé;
* en l'absence de descripteur approprié, et lorsque l'émergence d'un nouveau concept n'est pas suffisamment explicite pour proposer un candidat descripteur, la possibilité d'utiliser un descripteur voisin ou générique

On a donc assez vite pensé à avoir recours à l’ordinateur.**[Mustafa, --]**

1. L’indexation automatique

L’indexation automatique, que nous décrivons en détail dans ce qui suit, regroupe un ensemble de traitements automatisés sur un document. On distingue : l’extraction automatique des mots des documents, l’élimination des mots vides, la lemmatisation (radicalisation ou normalisation), le repérage de groupes de mots, la pondération des mots et enfin la création de l’index.**[Sauvagnat, 05]**

La première approche d’indexation automatique KWIC (Key-Word In Context) a été introduite par Luhn (1957). Il a été question ensuite de pondérer les index. Aux débuts de la recherche d’information, les méthodes statistiques ont été fondées sur la fréquence de termes dans le document. Plus tard, cette mesure a été étendue pour tenir compte de la spécificité d’un terme pour le document. À cette fin, d’autres méthodes ont été exploitées, telles que 2-Poisson (Nie, 2003).**[Mustafa, --]**

Les systèmes d’indexation automatique utilisent plusieurs méthodes d’analyse :

* **Analyse linguistique :**

Technologie issue du « text mining », cette dernière consiste à mettre en œuvre un modèle simplifié des théories linguistiques dans des systèmes informatiques d’apprentissage. Cela fait partie du domaine de l’intelligence artificielle.**[ALLAB, 08]**

La méthode linguistique est composé de plusieurs modules d’analyses linguistiques : morphologique, lexical, syntaxique et pragmatique.

Le fait que certains systèmes d’indexation utilisent les techniques du traitement automatique des langues, démontre la pertinence d’une approche linguistique.**[El-Hachani, 97]**

* **Analyse statistique :**

L’initiateur des méthodes d’indexation automatique est H.P. Luhn avec son célèbre article “*The automatic creation of literature abstracts”* paru en 1958 dans le “Journal of Research and Development” d’IBM. Il déclare : « (...) *au lieu de tirer l’information au hasard comme le fait normalement le lecteur, la nouvelle méthode automatique choisit les phrases d’un article qui représentent le mieux l’information pertinente »*. **[CHAUMIER, 90]** , H. P. Luhn ouvrit la voie aux travaux sur l’indexation automatique par voisinage appelée aussi méthode statistique.

L’indexation automatique comporte les étapes suivantes **[ALLAB, 08]** :

**Extraction des mots simples**

1. Extraction des termes (tokenization) : les règles d’extraction dépendent de la langue.
2. Suppression des mots vides (stop words) : ce sont des mots trop fréquents mais inutiles. Exemple : le, un, de, or, on, il…

**Normalisation**

1. Lemmatisation (radicalisation) : par exemple inter + classe + ment, ici le radical est classe.
2. Règles de transformation : suppression des terminaisons des pluriels.
3. Troncature : choisir une valeur optimale de troncature des mots. Il vaut mieux tronquer les suffixes. Il n’y a pas de règle absolue pour cela.

L’indexation automatique permet **[Hadj henni, 08]** :

* de limiter les choix parfois subjectifs de l’indexeur.
* d’alléger le travail requis par une indexation manuelle.
* d’éviter les incohérences résultant des interprétations différentes entre plusieurs indexeurs.
* de réaliser une recherche exhaustive des sujets traités dans l’information analysée.

1. L’indexation semi-automatique

Les deux techniques précédentes peuvent être combinées, un premier processus automatique permet d'extraire les termes du document. Cependant le choix final reste au spécialiste du domaine ou au documentaliste pour établir les relations sémantiques entre les mots clés et choisir les termes significatifs en utilisant un thésaurus ou une base terminologique, qui est une liste organisée de descripteurs (mots clés) obéissant à des règles terminologiques propres.

On peut mentionner comme un exemple de cette technique les deux logiciels : SINTEX et ALEXDOC qui utilisent l’indexation assistée par ordinateur.**[ABAR, 09]** **[Sauvagnat, 05]** **[Hadj henni, 08]**

1. Les index

L’index est la sortie du processus d’indexation, il existe plusieurs types d’index selon la technique utilisé et la fonction souhaitée :

1. Différentes formes d’indexes

**Document Index**

L'index des documents conserve des informations sur chaque document. Il est un index ISAM (Index sequential access mode) d'une largeur fixe, ordonné par l'ID du document. Les informations stockées dans chaque entrée comprennent des données, une somme de contrôle de documents et diverses statistiques. Si le document a été exploré, il contient aussi un pointeur vers un fichier d'une largeur variable appelé les informations du document qui contient l'URL et le titre. Cette décision de conception a été guidée par le désir d'avoir une structure de données relativement compact, et la possibilité de chercher un enregistrement dans une seule parcoure du disque, lors d'une interrogation **[Brin, 98]**

Le tableau suivant est une illustration simplifiée d'un index des documents :

|  |  |  |
| --- | --- | --- |
| Tableau 9 : Document Index | | |
| Document ID | **texte** | **lien** |
| **Document 1** | The cow says moo | /ex/doc1.txt |
| **Document 2** | The cat and the hat | /ex/doc2.txt |
| **Document 3** | The dish ran away with the spoon | /ex/doc3.txt |

**Forward Index**

Forward index (ou index transmis) stocke une liste de mots pour chaque document. Ce qui suit est une forme simplifiée de forward index :

|  |  |
| --- | --- |
| Tableau 10 : Forward Index | |
| **Document** | **Mots** |
| **Document 1** | the, cow, says, moo |
| **Document 2** | the, cat, and, the, hat |
| **Document 3** | the, dish, ran, away, with, the, spoon |

La raison derrière le développement d'un indice avancé est, pendant que les documents sont parcourus, il est préférable de stocker immédiatement les mots par document. La délinéation active un système de traitement asynchrone, ce qui contourne partiellement l’index inversé en goulot d'étranglement[[158]](#footnote-11). L’index transmis est sauvegardé pour le transformer en un index inversé. L'indice à terme est essentiellement une liste de couples constitués d'un document et un mot, collectés par le document. La conversion de l’index transmis en un index inversé n'est qu'une question de trier les couples par les mots. A cet égard, l'index inversé est un mot-triés de l’index transmis.**[Brin, 98]**

**Inverted index**

Plusieurs moteurs de recherche intègrent un index inversé lors de l'évaluation d'une requête de recherche pour localiser rapidement des documents qui comportent les mots dans la requête et puis les classer par pertinence. Puisque l’index inversé sauvegarde la liste des documents contenant chaque mot, le moteur de recherche peut utiliser un accès direct pour trouver les documents associés à chaque mot dans une requête afin de récupérer les documents qui répondent rapidement. Le tableau suivant est une illustration simplifiée d'un index inversé :

|  |  |
| --- | --- |
| Tableau 11 : Inverted Index | |
| Mot | **Documents** |
| **the** | Document 1, Document 3, Document 4, Document 5 |
| **cow** | Document 2, Document 3, Document 4 |
| **says** | Document 5 |
| **moo** | Document 7 |

Cet index ne peut déterminer que si un mot existe dans un document particulier, car il ne stocke aucune information concernant la fréquence ou la position du mot, il est donc considéré comme un index de booléen. Un tel index décide quels sont les documents qui correspondent à une requête, mais ne les classe pas. Dans certains modèles, l'index comprend des informations supplémentaires telles que la fréquence de chaque mot dans chaque document ou les positions d'un mot dans chaque document. Les renseignements de position permettent à l'algorithme de recherche d’identifier le voisinage du mot pour supporter la recherche de phrases. La fréquence peut être utilisée pour aider dans le classement de la pertinence des documents par rapport à la requête. **[Grossman, 02]** **[Tang, 04 ]**

**N-gram index**

Un n-gramme est une séquence de n caractères consécutifs. Pour un document quelconque, l’ensemble des n-grammes (en général n prend les valeurs 2 ou 3) qu’on peut générer est le résultat qu’on obtient en déplaçant une fenêtre de n cases sur le corps de texte. Ce déplacement se fait par étapes, une étape correspond à un caractère. Ensuite on compte les fréquences des n-grammes trouvés. Par exemple la phrase "La nourrice nourrit le nourrisson" se représente par :

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n-grammes | la\_ | a\_n | \_no | nou | our | urr | rri | ric | ice | \_ce | e\_n | rit | it\_ | t\_l | … |
| Fréquences | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | … |

Tableau 12 : N-gram index

Nous avons utilisé le caractère “\_” à la place des blancs, pour faciliter la lecture.**[Jalam, 02]** L’un des avantages des n-grammes est la capture automatique des racines les plus fréquentes **[Grefenstette, 95]** : dans l’exemple précédent, grâce aux techniques basées sur les n-grammes nous trouvons la racine commune de : Nourrir, nourri, nourrit, nourrissez, nourriture, etc.

La tolérance aux fautes d’orthographes et aux déformations est également une propriété importante. Enfin, ces techniques n’ont pas besoin d’éliminer les mots-outils (Stop Words) ni de procéder à la lemmatisation, ni au stemming. **[Sanan, 08]**

1. Mise à jour de l’index

Par mise à jour d’index on entend le comportement de l’index face aux changements que subit la base de données. Des changements qui peuvent être des insertions, des modifications ou des suppressions. Et ceci peut concerner les documents comme les nœuds des documents. Un index peut être plus ou moins capable de s'adapter à ces changements. Cette adaptation peut se présenter sous trois formes :

1. Mise à jour incrémentale

Dans le cas d’une mise à jour incrémentale (également appelée par augmentation) la structure de l'index est mise à jour en lui rajoutant les index des nouvelles données (documents ou nœuds à indexer) sans modifier ceux qui existent déjà. Le nombre de changements supportés est cependant, souvent limité, dans ce cas. **[Dahak, 06]**

1. Mise à jour progressive

Dans le premier cas la structure de l'index existante peut être modifiée et les modifications sont limitées uniquement à une petite partie de l'index. Cet index supporte donc des mises à jour progressives. **[Dahak, 06]**

1. Mise à jour globale

Le troisième cas, et le plus défavorable, c’est quand la structure de l'index entière doit être reconstruite à partir de zéro.**[Dahak, 06]**

1. Stockage de l’index

Le stockage des structures d’index est caractérisé essentiellement par la taille de l’index et l’organisation des éléments de celui-ci. Les structures d'index varient largement dans leur consommation d'espace qui est étroitement liée à l’organisation des données dans l’index.

Cette organisation a un impact significatif sur la latencede recherche. Plus les éléments liés sont près les uns des autres dans l’espace de stockage moins est la latence de recherche, c’est ce qu’on appelle le concept de localité. Il est également très important que l’index puisse tenir dans la mémoire centrale, cela évite les accès disque au système et réduit considérablement la latence de recherche.

L’index idéal est donc celui qui occupe moins d’espace et qui permet de réduire au maximum la latence de recherche.**[Dahak, 06]**

1. Les étapes d’indexation

Le processus d’indexation se compose des étapes suivantes :

1. Extraction des mots-clés

Appelée *tokenization* en anglais, l’extraction des mots-clés est une étape qui peut sembler triviale au premier abord, et qui pourtant constituera la base de tout le reste du processus d’indexation. Il faut donc que cette phase soit d’une qualité maximale **[Meylan, 01]** . Certains systèmes de recherche d’information utilisent une liste de mots-clés prédéfinie. Cette liste est conçue manuellement et, dans la plus part des cas, construite par rapport à un thème spécifique. Cette méthode permet entre autre de contrôler la taille des index. L’utilisation d’une extraction automatique des mots-clés ou l’utilisation d’une liste de mots-clés prédéfinie, détermine le type d’indexation. Orientée document dans le premier cas et orientée requête dans le second (aller à**[Berrut, 97]** pour plus de détails sur ces deux concepts).**[Dahak, 06]**

1. Normalisation des mots-clés

Ce traitement consiste à retrouver pour un mot sa forme normalisée (généralement le masculin pour les noms, l’infinitif pour les verbes, le masculin singulier pour les adjectifs, etc.). Ainsi, dans l’index ne sont conservées que les formes normalisées, ce qui offre un gain de place appréciable, mais surtout, si le même traitement est effectué sur la requête, cela permet d’être beaucoup plus souple et rapide dans la recherche: par exemple, si l’utilisateur effectue une recherche avec un verbe, les documents comportant ce verbe dans toutes ses formes conjuguées seront pris en compte, et pas seulement les documents contenant le verbe dans la forme fournie par l’utilisateur. Cette étape est également appelée « traitement morphologique des mots-clés » **[Denoyer, 04]** .

Cette phase peut également être enrichie avec un traitement syntaxique et sémantique des mots-clés. Le premier consiste à identifier et regrouper un ensemble de mots dont la signification dépend de leur union. Par exemple, les mots « maison blanche » ne signifient habituellement pas qu’on a affaire à une maison qui est blanche, mais plutôt au siège de la présidence des États-Unis. Elle consiste aussi à éliminer des ambiguïtés comme par exemple les problèmes d’homographie.

Le traitement sémantique a pour but de faire des distinctions entres les différents sens possibles d’un même mot (polysémie). Par exemple, cette phase permet de différencier le mot « pièce » qui peut correspondre à une pièce de monnaie par exemple, ou à une pièce dans une maison. C’est une tâche ardue qui n’est pas aujourd’hui bien maîtrisée et dont l’intérêt en termes de hausse des performances des systèmes n’est pas toujours démontré.**[Dahak, 06]**

1. Élimination des mots vides

Cette étape revêt une importance certaine dans la mesure où elle constitue un facteur d’une grande influence dans la précision de la recherche. Le fait de ne pas éliminer les mots vides provoque inévitablement du bruit. L’élimination des mots vides qui sont des mots du langage courant et qui ne contiennent pas beaucoup d'information sémantique doit se faire aussi bien à l’indexation qu’à l’interrogation (élimination des mots vides de la requête).**[Dahak, 06]**

1. La pondération

Cette étape est entièrement dépendante du modèle de recherche d’information utilisé. Elle permet de définir l’importance qu’a un terme dans un document donné, elle est également utilisée pour filtrer l’index résultant du processus d’indexation (c’est-à-dire : éliminer les index dont le poids est inférieur à un certain seuil).**[Dahak, 06]**

De manière générale, la majorité des formules de pondération des termes est construite par combinaison de deux facteurs. Un facteur de pondération locale quantifiant la représentativité locale d’un terme dans le document, et un second facteur de pondération globale mesurant la représentativité globale du terme vis-à-vis de la collection des documents.**[AMROUCHE, 08]**

**Types de pondération :**

**Pondération locale**

La pondération locale prend en considération les informations locales du terme qui ne dépendent que du document. Elle correspond généralement à une fonction de la fréquence d'apparition du terme dans le document, notée ***tf*** (*term frequency*). Un terme qui apparaît fréquemment dans un document sera considéré comme pertinent pour décrire son contenu. **[Dahak, 06]**

**Pondération globale**

La pondération globale mesure l'importance d'un terme au sein de l'ensemble des documents de la collection. Elle vise à représenter son caractère discriminant ou en d’autres termes sa capacité à différencier les documents de la collection. En effet, un terme apparaissant dans peu de documents est considéré comme plus discriminant et doit être privilégié par rapport à un terme présent dans de nombreux documents. Le calcul de la pondération globale s'appuie sur le nombre de documents de la collection dans lesquels un terme apparaît. L'une des mesures les plus utilisées est l’***idf*** (*inverse document frequency*), représentée par la formule suivante:

Où *i n* est le nombre de documents contenant le terme *i t i* et *N* le nombre total de documents dans la collection.

La mesure ***tf*** \****idf*** donne une bonne approximation de l’importance d’un terme dans le document, particulièrement dans les corpus de documents de taille homogène. **[Dahak, 06]**

1. Interrogation

L'interrogation est la phase d'interaction entre le système et l'utilisateur. Ce dernier exprime son besoin d'information via un langage de requête que le système va se charger de traduire. Cette traduction se fait selon le modèle de requête et a pour but de comprendre les besoins de l'utilisateur et de les exprimer dans un formalisme similaire à celui mis en œuvre lors de l'indexation des documents. Ce processus fournit une requête interne. Suite à cette phase de compréhension de la requête, un modèle de correspondance calcule la correspondance entre la requête interne et chaque index des documents. Ce calcul, établi par la fonction de correspondance, a classiquement pour résultat une liste ordonnée des documents de la base. Il faut, à ce niveau, établir une comparaison sémantique (et non une égalité) entre les concepts figurants dans un document et ceux figurants dans la requête. La comparaison entre requête et document aboutit rarement à des équivalences strictes, mais plutôt à des équivalences partielles : le document correspond à une partie seulement de la requête. Le premier document de la liste renvoyée par le système est celui qui est considéré par le système comme le plus pertinent, c’est-à-dire celui qui répond le mieux à la requête, toujours d'après le système. Le dernier document est celui qui est considéré par le système comme le moins pertinent. Cette notion de pertinence repose sur la proximité entre les besoins exprimés par l'utilisateur et les résultats fournis par le système. On différencie la pertinence utilisateur de la pertinence système. **[Dahak, 06]**

1. La notion de pertinence

Pertinence est une notion centrale de l’interrogation car toutes les évaluations s'articulent autour de cette notion. Mais c'est aussi la notion la plus mal connue, malgré de nombreuses études portant sur cette notion tel que celle figurant dans **[Denos, 97]** . Voyons quelques définitions de la pertinence pour avoir une idée de la divergence.

La pertinence est:

* La correspondance entre un document et une requête, une mesure d'informativité du document à la requête;
* Un degré de relation (chevauchement, relativité, …) entre le document et la requête;
* Un degré de la surprise qu'apporte un document, qui a un rapport avec le besoin de l'utilisateur;
* Une mesure d'utilité du document pour l'utilisateur; …etc.

Même dans ces définitions, les notions utilisées (informativité, relativité, surprise, …) restent très vagues et ceci, parce que les utilisateurs ont des besoins très variés. Ils ont aussi des critères très différents pour juger si un document est pertinent. Donc, la notion de pertinence est utilisée pour recouvrir un très vaste éventail de critères et de relations.

Dans la suite, nous appelons pertinence système l'ensemble des principes qui sous-tendent la fonction de correspondance (donc le document est jugé pertinent par le système) **[Denos, 97]** , par opposition à la pertinence utilisateur, qui correspond à l'ensemble des jugements de pertinence que produit l’utilisateur qui utilise le système (le document est jugé pertinent par l’utilisateur). **[Dahak, 06]**

1. Fonction de similarité

La comparaison entre le document et la requête revient à calculer un score, supposé représenter la pertinence du document vis-à-vis de la requête. Cette valeur est calculée à partir d’une fonction ou d’une probabilité de similarité notée **rsv**(q,d) *(retrieval status value),* où ***q*** est une requête et ***d*** un document et dont la formule de calcul dépend entièrement du modèle de recherche d’information utilisé. Cette mesure tient compte du poids des termes dans les documents, déterminé en fonction d’analyses statistiques et probabilistes. La fonction d’appariement est très étroitement liée aux opérations d’indexation et de pondération des termes de la requête et des documents du corpus. D’une façon générale, l’appariement document - requête et le modèle d’indexation permet de caractériser et d’identifier un modèle de recherche d’information. La fonction de similarité permet ensuite d’ordonner les documents renvoyés à l’utilisateur. La qualité de cet ordonnancement est primordiale. En effet, l’utilisateur se contente généralement d’examiner les premiers documents renvoyés (les 10 ou 20 premiers). Si les documents recherchés ne sont pas présents dans cette tranche, l’utilisateur considérera le tri comme mauvais vis-à-vis de sa requête.

Le but de tout tri est donc évidemment de rapprocher la pertinence système de la pertinence utilisateur (qui comme nous l’avons vu précédemment, est fortement subjective)**[Sauvagnat, 05]** **[Dahak, 06]** .

1. La modélisation de la connaissance

Selon **[Berrut, 97]** et **[Mathias, 02]** la modélisation de la connaissance apparaît sous la forme d’un thésaurus, et sert de référence aussi bien au processus d’interrogation qu’au processus d’indexation. Le processus d’interrogation l’utilise afin d’augmenter ou de restreindre les requêtes des utilisateurs, agissant ainsi directement sur le nombre et la qualité des réponses données par le système. Le processus d’indexation l’utilise essentiellement comme norme de définition des termes d’indexation **[Dahak, 06]** .

1. le processus d’interrogation

La recherche tient une requête de l’utilisateur et retourne efficacement la liste des résultats correspondants classés par pertinence. Comme l'indexation, la recherche est un processus à plusieurs étapes, comme le montre la figure 22. La première opération consiste à l’établissement de la requête. Le moyen d'exprimer la requête est :

1. Basé sur les chaines de caractères : langage d’interrogation basé sur le texte. En fonction de l’objectif, un tel langage peut être aussi simple que la manipulation des mots ou aussi complexe que d'avoir des opérateurs booléens, des opérateurs de rapprochement, restriction de champ, et bien plus encore!
2. Basé sur une API de programmation : pour une requête avancées et étroitement contrôlée, une API de programmation est très approprié. elle donne au développeur un moyen souple pour exprimer des requêtes complexes et décider comment exposer la flexibilité de la requête aux utilisateurs (ça peut être une fonction exposée avec une interface de la REST[[159]](#footnote-12) Representational State Transfer).

Certains outils vont concentrer sur la requête basée sur les chaines de caractères, d'autres sur l'API de programmation, et d'autres sur les deux.  
La deuxième opération, appelons-la *l’analyse*, est chargé de prendre des phrases ou des listes de mots et les applique l'opération similaire à celle du moment d'indexation (classer en mots, radicaux, ou description phonétique). Ceci est crucial car le résultat de cette opération est le langage en commun que indexation et la recherche utilisent pour parler entre eux, et se trouve d’être celui stocké dans l'index. Si la même série d'opérations n’est pas appliquée, la recherche ne trouvera pas les mots indexés, ce qui n’est pas si utile!

Basé sur le langage commun entre l'indexation et la recherche, la troisième opération (recherche de documents) lit les index et recherche l’information d'index associées à chaque mot correspondant. N'oublions pas que pour chaque mot, l'index peut stocker la liste des documents correspondants, la fréquence, les positions du mot dans un document, et ainsi de suite. L'accord implicite ici est que le document en lui-même n'est pas chargé, et c'est une des raisons pour laquelle la recherche à texte intégral (full-text search) est efficace : le document n'a pas à être chargé pour savoir s'il correspond.

L'opération suivante (filtrer et trier) traitera les informations récupérées à partir de l'index et construira la liste des documents (ou plus précisément, les gestionnaires ‘handlers’ de documents). A partir des informations disponibles (les documents correspondants par mot, la fréquence du mot, et la position du mot), le moteur de recherche est en mesure d'exclure les documents de la liste. Plus important encore, il est capable de calculer le score de chaque document. Plus son score est élevé, plus un document sera dans la liste des résultats. Jetons un regard sur certains facteurs qui influent sur sa valeur :

* Dans une requête qui comporte plusieurs mots, plus ils sont proche dans un document, plus le rang (rank) est supérieur.
* Dans une requête comportant plusieurs mots, plus on les trouve dans un document unique, plus le rang est supérieur.
* Plus la fréquence d'un mot dans un document est supérieure, plus le rang est supérieur.

Requête de texte

Requête programmatique

Traiter les mots

Trouver IDs des documents correspondant à chaque mot

Filtre et ordre IDs des documents par pertinence

Exposer les résultats à l’utilisateur

Récupérer les documents (de l'extérieur d'index)

Liste ordonnée du résultat

**Index**

* Le une approximation moindre d'un mot donne un rang supérieure.

Selon la façon dont la requête est exprimée et comment le produit calcule le score, ces règles peuvent s'appliquer ou non. Cette liste est là pour donner une idée de ce qui peut affecter le score, donc la pertinence d'un document. Une fois la liste ordonnée des documents est prête, le moteur de recherche à plein texte expose les résultats à l'utilisateur à travers une API de programmation ou dans une page web.**[Bernard, 08]**

Figure 22 : Le processus d’interrogation

1. Conclusion

Dans ce chapitre, l’étude a porté sur le mécanisme de travail des moteurs de recherche web et des systèmes de recherche d’information, en se basant sur l’indexation en raison de son importance. En effet, elle constitue l’étape la plus importante dans le processus de recherche car elle permet l’extraction et le traitement des mots clés.

L’interrogation ne constitue pas seulement la phase d’interaction entre les utilisateurs est le système, mais calcule aussi la correspondance entre la requête et les documents pour fournir les résultats les plus pertinents.

Dans le chapitre suivant, nous allons voire les index et les ma‛ājim du Coran.

# Chapitre IV

**Les index du Coran**

1. Introduction

Au fil des siècles, les musulmans étaient très intéressés au Coran, par le tafssīr, l'étude, la préservation, la récitation, et la calligraphie. Le domaine des sciences coraniques été l’objet d’une grande attention. Donc, des index ont été crées pour le vocabulaire du Coran, la recherche des mots similaires et les différences entre eux, trouver leurs racines et leurs dissimilitudes.

Dans l'ère moderne, « le mu‛jam indexé des mots du Saint Coran » par Mohammed Fouad Abd El‑bāki avait comblé une lacune importante dans la vitesse d’accès aux ayas du Coran à partir de la racine du mot ou de rechercher n'importe quel mot à travers la racine, le mu‛jam indexe les noms et les verbes dans le Coran selon la racine.

Aujourd’hui, il y a eu de nombreuse voix appelant à l’informatisation de tels index pour les utiliser dans les vastes domaines de l’informatiques, y compris la recherche, ainsi plusieurs projets ont été lancés à cet effet d’ou le sujet de ce chapitre.

1. Définitions
2. Al-mu‛jam

Le mu‛jam ou « lexique » est une référence qui contient des informations (sujets ou mots) triées dans un certain ordre ainsi que leurs explications ou traductions, le mu‛jam peut être général (d’une langue par exemple) ou spécialisé. Dans notre étude nous nous intéresserons aux ma‛ājim spécialisés pour le Coran.

1. L’index

L’index est comme le mu‛jam car il contient des mots triés, mais la différance est que l’index n’explique pas les mots, mais il indique la position de toutes ses occurrences pour qu’on puisse y accéder facilement. Les index du Coran donnent la position par la sourate et l’aya contentant un terme spécifique. **[الخميسي, 04]**

1. Historique des index du Coran

Beaucoup de savants musulmans sont intéressé par l’indexation du Coran en développant les méthodes et en renouvelant la manière et le but. Ils ont mis des index de termes coraniques et leurs positions dans le Coran, puis ils ont construit des lexiques pour expliquer les significations de ces mots après avoir rappelé les ayas dans lesquelles ils figurent. Certains savants ont construit des lexiques spécialisés, chacun contient le vocabulaire d’un seul sujet.

L’évolution des index du Coran s’est faite en quatre étapes principales **[الخميسي, 04]**  :

1. **Indexation des mots du Coran**

Le développement des index des mots du Coran est une nécessité, car ils permettent aux chercheurs d'accéder facilement aux mots du Coran. Les musulmans et les orientalistes se ont axé leurs recherches dans le premier tiers du XXe siècle, au développement des index pour les mots du Coran et les parties des ayas.

Parmi les principaux ouvrages dans ce domaine nous citons :

* « nujoum al-furkān fī atrāf al-coran[[160]](#endnote-148) » par l’orientaliste allemand Leberecht Flugel ;
* « Le mu‛jam des ayas du Coran[[161]](#endnote-149) » par Houssine Nassār[[162]](#endnote-150) ;
* « le mu‛jam indexé des mots du Saint Coran [[163]](#endnote-151) » par Mohammed Fouad Abd El-bāki[[164]](#endnote-152).

L’index de Mohammed Fouad est le plus important, il a bénéficié de l'expérience de Flugel, a traduit son livre nujoum al-furkān fī atrāf al-coran et l’a perfectionné en éliminant les affixes des mots et en les arrangeant par ordre alphabétique. Son mu‛jam « Le mu‛jam des ayas du Coran » a été publié en 1945 en Egypte, et les numéros des ayas mentionnées dans ce livre correspondaient bien au mushaf officiel. Puis il a été édité dans plusieurs pays arabes et musulmans y compris Istanbul en 1982.

Ce livre a connu beaucoup de succès, néanmoins il ne contient pas les pronoms[[165]](#endnote-153) et les particules[[166]](#endnote-154). **[الخميسي, 04]**

1. **Les ma**‛**ājim des mots du Coran**

Les index du Coran offrent la possibilité de trouver la position des mots, puis le chercheur doit rechercher le tafssīr dans les livres appropriés par lui-même, d’où le besoin de produire des ma‛ājim qui donne les mots avec l’explication afin de rendre la recherche encore plus souple.

De nombreux travaux collectifs ont été lancés dans ce sens. En 1941 Dr. Mohamed Hussein Heikal,[[167]](#endnote-155) membre de l'Académie de la Langue Arabe au Caire[[168]](#endnote-156) a proposé de mettre en œuvre un mu‛jam des termes du Coran. En 1949, un comité a commencé à élaborer ce mu‛jam en s’inspirant de celui de Mohammed Fouad Abd El-bāki. **[الخميسي, 04]**

1. **Les ma‛ājim des mots du Coran spécialisés**

Les ma‛ājim cités précédemment sont généraux, et il sera difficile de chercher les ayas contenant des mots qui appartiennent à un domaine spécifique. Cela a conduit à la création des ma‛ājim spécialisés tels que « le mu‛jam de vocabulaire des plantes dans le Coran » par le leader dans ce domaine Mokhtar Fawzi Al-na’āl[[169]](#endnote-157), qui a achevé ensuite d’autres ma‛ājim spécialisés pour l’homme, les animaux, les éthiques, le commerce, etc. **[الخميسي, 04]**

1. **Les index du Coran et l’ordinateur**

Quand l'ordinateur est entré dans la vie quotidienne des gens, il a fait une révolution dans le monde de l’information. Les savant musulman ont profité de ces services, et ils ont fait de lui un excellant index pour le Coran. En fait la vitesse et l’exactitude de l’ordinateur lui permettent de donner rapidement l’emplacement et le nombre d’occurrences de n’importe quel mot dans le Coran, il suffit seulement qu’on l’alimente avec les index et les ma’ājim.

1. Les index

Les Index existants diffèrent considérablement les uns des autres, nous pouvons les classer en utilisant certain critères, y compris : l'objet ou le but de chaque index et l'unité sur laquelle l'index se base ... etc.   
classées selon les critères cités ci-dessus, nous expliquerons les index, existants ou possibles, les plus importants:

1. Par l’unité:

Les index du Coran se basent sur plusieurs unités, *voir la figure 23* :

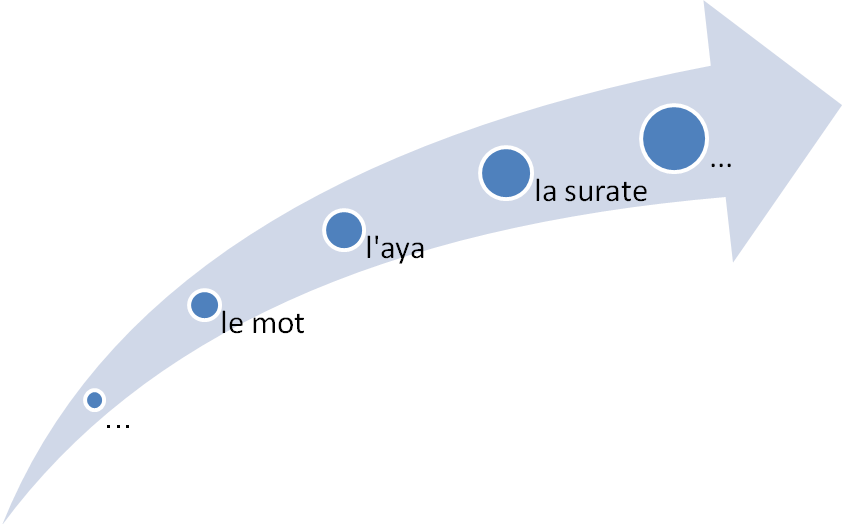


Figure 23 : La classification des index par unité

1. Le mot

Ces index peuvent prendre deux aspects, le premier est grammatical, il consiste à découper les mots selon leur structure fondamentale, en indiquant les affixes et ces origines tels que la racine et le morphème. Le deuxième aspect a pour but de faciliter la lecture du mushaf par un index des mots dont l’écriture diffère entre le script othmani et le script standard et un index des mots qui s’écrivent de la même manière mais se prononcent différemment.

De même, pour faciliter la lecture des livres des sciences du Coran, un mu‛jam électronique des termes des sciences coraniques a été établi, et un autre pour les sciences du Coran dans la sunna, parmi ces index il existe :

1. Des index de tous les mots et en les reliant avec toutes leurs parties : préfixes + noyau + affixes, ainsi qu’avec la racine, le morphème, le verbe à l’infinitif ou la forme singulière dans le cas des mots pluriels, plus les différentes informations flexionnelles[[170]](#endnote-158), comme le type du mot est sa fonction flexionnelle. Ces index ont une utilité dans le traitement automatique du langage naturel.
2. Un index des mots dont l’écriture diffère entre le script othmani et le script standard, de 6419 mots, en reliant chaque entrée de l’index avec les ayas et les mots qu’on veut atteindre dans les pages du mushaf, en vue de faciliter la lecture dans le Coran pour les débutants et ceux qui ne connaissent pas les règles d’écriture des mots en script othmani.
3. Un index des mots dont l’écriture diffère entre le script othmani et le script standard (ceux qui nous sont parvenus des savants), qui sont de 1318 mots, pour faciliter l'apprentissage, en les reliant avec les pages du mushaf et la récitation correcte d'un célèbre récitant. Cet index – comme son prédécesseur – vise les débutants dans la préservation du Coran et ceux qui veulent le mémoriser.
4. Un Index des mots dont les structures (les lettres) se ressemblent mais la prononciation varie en fonction de la forme convenue, avec distinction entre la prononciation de ces mots et en les reliant avec les pages du mushaf, plus la possibilité d'entendre selon le choix de l'utilisateur.
5. Des ma‛ājim électroniques de termes de sciences coraniques, qui peuvent être affichés par ordre alphabétique, ou classés par les sciences liés au Coran, comme la récitation, le tafssīr, l’abrogation[[171]](#endnote-159), les causes de révélation, et tout ce qui concerne la langue du Coran et son éloquence. La préparation d’un tel mu‛jam exhaustif nécessite l'utilisation des références patrimoines, telles que: « la métrisation en sciences du Coran[[172]](#endnote-160) »par Al‑sayouti, « la preuve en science du Coran[[173]](#endnote-161) » par Al-zarkachi et « les sources de gratitude[[174]](#endnote-162) » par Al-zarkāni, ainsi que des références modernes tels que : « les termes de sciences du Coran[[175]](#endnote-163) » par Anwar Al-baaz,  « le mu‛jam de sciences du Coran[[176]](#endnote-164) » par Ibrāhīm Mohamed Al-djarmī,  le mu‛jam des termes de la science des récitations et connexes, le mu‛jam  des termes des sciences de tajuīd et de récitation et ainsi de suite. **[الميمان, --]**

Exemple d’index qui prend le mot comme unité, cet index va être détaillé par la suite.

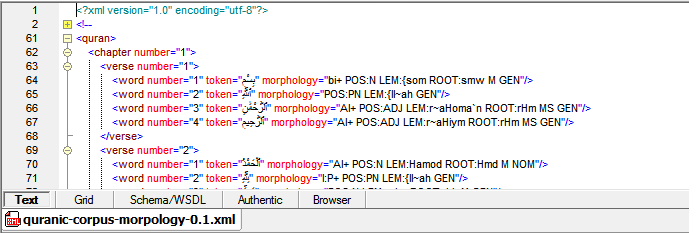


Figure 24 : Index prenant le mot comme unité.

1. L’aya

Ces index prend les ayas comme une unité d'indexation dans le but de les classer par sujets et d’examiner la similarité et la coïncidence et élaborer un mu‛jam basé sur un analyseur morphologique des racines du Coran, ainsi que de faire une indexation des débuts de mots des ayas, leurs fins[[177]](#endnote-165), etc. Voici quelques index comme exemple :

1. Un arbre de thèmes complet et global constitué de quatre niveaux contenants les sujets abordés par le Coran (plus de 1100 sujets) ainsi on obtient un index informatisé des sujets du Coran qui relie chaque sujet principal ou secondaire avec les ayas qui peuvent être placés sous ce sujet. Cet index est considéré comme une excellente aide pour les prêcheurs, les chercheurs et les étudiants en sciences islamiques pour accéder à toutes les ayas d'un sujet ou sous-sujet lié au texte coranique. A noter que l’aya peut être liée à plusieurs sous-sujets, le premier niveau de cet index contient les sujets suivants : la foi[[178]](#endnote-166), le jour de la résurrection, le paradis et l'enfer, les cultes, les interactions entre les gens, l’intimidation et l’encouragement, les morales et les éthiques, les nations, les tribus et les peuples, les anges, les gins et les humains, le Coran, la biographie du prophète s, les histoires des prophètes, des vertueux et des anciens, la science, l'univers et des créatures.
2. Un index des ayas identiques et similaires dans le Coran pour servir les préservateurs du Coran et les chercheurs, où les ayas identiques et similaires sont identifiées et liées en groupes textuels. Ces ayas et passages sont accessibles à partir d’une liste des ayas identiques et similaires pour chaque sourate (s’ils existent), et en choisissant une aya (ou un passage) en peut accéder à la liste des ayas (ou passages) identiques ou similaires dans tout le Coran.
3. Un index des ayas coraniques disposées selon les lettres des premiers mots, les fins des ayas et les deux derniers mots, pour faciliter à la fois l'accès à n’importe quelle aya, et le processus de préservation et mémorisation du Coran, tout en assurant la validité de cet index en le révisant avec les ouvrages correspondants, tels que: « le répertoire indexé des ayas  du Coran[[179]](#endnote-167) » et « l’index coranique[[180]](#endnote-168) ».
4. Un mu‛jam indexé de mots du Coran, plus précis, plus complet et plus facile à utiliser que le mu‛jam en papier préparé par le Pr. Muhammad Fuad Abd al-Baqi un pionnier louable à ses efforts malgré l’absence des nouvelles technologies dont nous profitons actuellement. Dans ce mu‛jam on peut rechercher par racine de mots contenus dans le Coran, qui s'élevaient à plus de 1700 racines, en arrivant à tous les mot associés à une racine, on peut aussi accéder à un groupe de mots à une seule racine avec n'importe quel mot d’entre eux en utilisant l'analyseur morphologique, dans les deux cas on lie le mot avec son emplacement dans le mushaf.
5. Un index de l’abrogeant et l’abrogé parmi les ayas du Coran : comprend le listage des ayas abrogeant ou abrogés dans les sourates du Coran, en reliant ces ayas avec leurs emplacements dans les pages du Coran. **[الميمان, --]**

Exemple d’index qui prend l’aya comme unité.



**Figure 25 : Index prenant l’aya comme unité.**

1. La surate

L’objet de base de ces index est les sourates du Coran de sorte qu’ils donnent à l'étudiant une idée générale sur les sourates, avec des informations suffisantes pour acquérir ce qui doit être appris : des données de base des sourates et leurs vertus (فضائلها), les plus importantes bases et les sujets de ces index sont :

1. Un index (liste) de sourates du Coran et leurs emplacements dans le mushaf.
2. Un index de données de base des sourates, qui comprend la sourate, ses noms, la raison d’attribution du nom, son type (mecquoise ou médinoise), son ordre dans le mushaf, son ordre selon la révélation, sa précédente et sa suivante (selon l’ordre de mushaf et de révélation).
3. Un index des vertus des sourates coraniques dans la sunna du Prophète s, avec la narration des hadiths dans le but de définir la sourate et ses vertus.
4. Un index des clés[[181]](#endnote-169) des sourates du Coran, dont l'objectif est de définir la sourate en termes de noms, la raison du nommage, la cause de révélation (si elle existe), etc. avec une présentation générale de la sourate, et des principales bases et sujets qu’elle contient.
5. Un index de sujets pour chaque sourate du Coran afin de diviser la sourate en sujets, chaque sujet comprend plusieurs ayas. **[الميمان, --]**
6. Autre

En plus des index mentionnés précédemment basés sur le mot, l’aya et la sourate, il existe des index basés sur d'autres unités, y compris :

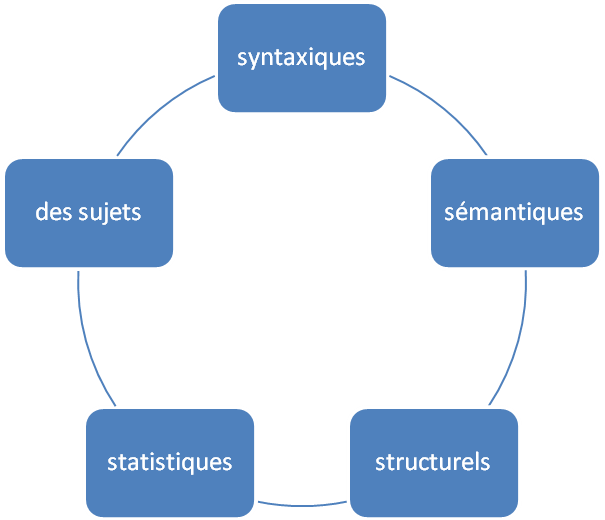
* 1. L’index grammatical basé sur les parties du mot, il fournit des informations sur le type de chaque partie et ses caractéristiques grammaticales : si elle est invariable[[182]](#endnote-170) ou fléchi[[183]](#endnote-171), sa flexion[[184]](#endnote-172), sa marque flexionnelle si elle est fléchi et ainsi de suite.
  2. L’index de syllabes ou unités phonétiques, il fournit des informations sur la prononciation qui sont très utiles dans la science de récitation. Aucun index de ce genre n’existe encore.
  3. L’index de phrases coraniques (*voir le tableau 13*), qui montre la manière de division en phrases, car il est bien connu que l’aya du Coran ne représente pas nécessairement une phrase.

|  |  |  |  |
| --- | --- | --- | --- |
| اللواصق | الحركة | رقم الكلمة | تسلسل جزء الكلمة |
| فـ | فتحة | 2597 | 1 |
| ـسـ | فتحة | 2597 | 2 |
| يكفي | سكون | 2597 | 3 |
| ك | فتحة | 2597 | 4 |
| هم | سكون أو ضم عند التقاء الساكنين | 2597 | 5 |

Tableau 13: Exemple d’un index basé sur les parts du mot

|  |  |
| --- | --- |
| رقم الجملة | الجملة |
| 1 | بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ@ |
| 2 | الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ@الرَّحْمَنِ الرَّحِيمِ@مَالِكِ يَوْمِ الدِّينِ@ |
| 3 | إِيَّاكَ نَعْبُدُ |
| 4 | وَإِيَّاكَ نَسْتَعِينُ@ |
| 5 | اهْدِنَا الصِّرَاطَ الْمُسْتَقِيمَ@صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ الْمَغْضُوبِ عَلَيْهِمْ وَلاَ الضَّالِّينَ@ |

Tableau 14 : L’index des phrases – surate al-fātiha

1. Par l’objet

La science d'indexation a abordé le Coran à plusieurs égards. De l’indexation de mots pour qu’un musulman puisse connaître l’aya une fois qu’il connaît l’un de ses mots, à l’indexation des sujets qui prend les ayas séparément, à d'autres index que nous citons comme suit :

Figure 26 : La classification des index par objet

1. Les index syntaxiques (des mots)

Ce sont les index qui traitent les mots en termes de fonctionnalités de la langue arabe pour montrer l’affixe du mot sa racine, son verbe s’il est dérivé d’un verbe, sa forme singulière s’il est pluriel, etc. Le but est de rechercher par mots et remplacer les analyseurs morphologiques en effectuant la plupart de leurs fonctions. Ces index sont construits avec ou sans la répétition de mots. Parmi les meilleurs livres qui ont indexé le Coran le mu‛jam de Pr. Muhammad Fuad Abd El-baqi ainsi que le livre « nujoum al-furkān fī atrāf al-coran » et le livre « Le guide du confus aux ayas du Coran[[185]](#endnote-173) ». L’index de mots par Taha Zerrouki, l’index principal du projet ‘Midād lbayān’ et le corpus coranique peuvent être considérés de cette catégorie, nous allons les expliquer par la suite dans ce chapitre.

1. Les index sémantiques

Ce sont les index qui suit une approche sémantique, afin de relier des mots du Coran avec leurs sens, la première base de données des significations peut être considérée comme celles qui est fondée sur les racines, puisque en termes de sens, les racines sont étroitement liés entre eux, par exemple, les racines : « جيء », «  أتى», « قدم », «  حضر», « وصل », «  ولي» et « دبر » ainsi que : « دخل », « ولج », «  قبل», « ورد » et «  وصل» . Sur cette base, les verbes se réunissent en groupes qu’on peut décrire comme ‘rassembleurs généraux’, ensuite les diviser en plus petits groupes, qui se caractérisent par un rassembleur plus particulier que le premier et ainsi de suite. Les groupes finaux n’ont pas toujours le même sens. Il y a beaucoup d’ouvrages dans le patrimoine arabe, qui mentionnent des mots de ce genre **[الجياني, 91]** . Il y a également des dictionnaires modernes qui donnent de tels groupes et les mots étrangers correspondants. **[ناصر, 89]**

Les bases de données sémantiques visent à suivre les relations sémantiques préalablement établies entre les notions coraniques, voici un modèle illustratif :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| م | الحقل الدلالي الأول | الحقل الدلالي الثاني | اشتمالية نوعية | نوعية  اشتمالية | اشتمالية عضوية | عضوية  اشتمالية | كلية  جزئية | جزئية  كلية | ترادف | تضاد |
| 1 | الإنسان | الرَّجُل | + |  |  |  |  |  |  |  |
| 2 | الإنسان | الرِّجْل |  |  |  |  | + |  |  |  |
| 3 | الإنسان | السَّاق |  |  |  |  | + |  |  |  |
| 4 | الرَّجُل | الإنسان |  | + |  |  |  |  |  |  |
| 5 | الرَّجُل | الرِّجْل |  |  |  |  | + |  |  |  |

Tableau 15 : Un exemple d’un index sémantique

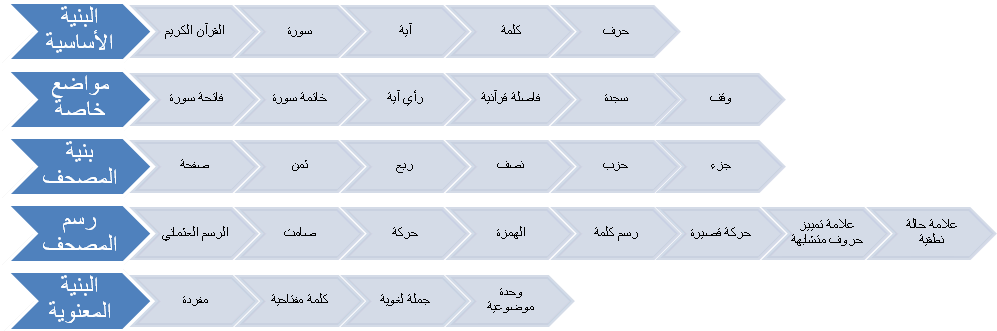
1. Les index des sujets

Ce sont les index qui divisent le Coran en unités traitant le même sujet, la division prend une forme de pyramide de sujets. Parmi les livres qui ont indexé les sujets coraniques le livre « zad al-moâllifīn min kitābi rabbi al-‘ālamin » par Abd Allah Mohammed Al-daruīche, et parmi les index de sujets qui ont été informatisés, l’index des sujets par Taha Zerrouki que nous allons aborder dans ce chapitre.

1. Les index structurels

Ce sont les index qui mettent l'accent sur les relations entre les différentes divisions du mushaf comme la division en ayas, sourates …, ou en thumns, rub’ …, ou encore les divisions spéciales comme en ruku’, phrases … et ainsi de suite, *voir la*  *figure 27*.

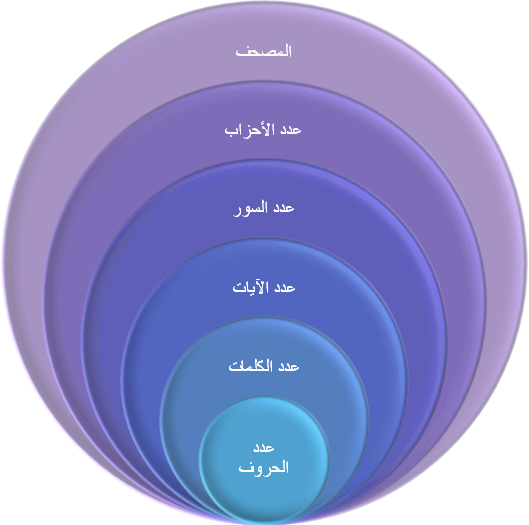
Ces index sont fondés sur l’une des unités de la division.



**Figure 27 : Les différentes structures du Saint Coran**

1. Les index statistiques

Ce sont ceux qui visent principalement à recueillir des statistiques sur les différentes unités du Coran à partir des lettres, cyclables, … jusqu’au Coran comme un tout. Parmi les statistiques possibles : le comptage du nombre, de la moyenne et de la composante dominante.

Ces index sont fondés sur l'élément qui représente la base de la statistique. Un exemple de statistiques est le nombre de caractères dans les mots, dans ces sourates, dans les ayats …

**Figure 28 : La hiérarchie des calculs du nombre**

1. Des projets de construction des index
2. Midād lbayān
3. Définition

C’est un grand projet initié par le Dr. Mohamed Zaki Khadher qui vise à construire une base de données coranique dans toute la mesure du possible et décrit tout le détail du Coran à partir des plus petits termes qui sont les syllabes du mot puis le mot, les expressions, les phrases, les ayas et les sourates. Cette base contient l’écriture, la prononciation, la morphologie, la grammaire, et la signification **[خضر, --]** .

Nous allons aborder l'une des versions précédentes de la base de données, en raison du manque d'accès à la version actuelle.

1. Les index et leurs structures

**L’index principal**

**Structure :**

|  |  |
| --- | --- |
| Champ | Commentaire |
| Seq | Code séquentiel qui sert d’un index pour les mots |
| Stem | Le noyau[[186]](#endnote-174) du mot |
| Word | Le mot en script standard (vocalisé[[187]](#endnote-175)) |
| Othmani | Le mot en script othmani |
| Root | La racine du mot |
| Prefix1 | Premier préfixe du mot (s’il existe) |
| Prefix2 | Deuxième préfixe du mot (s’il existe) |
| Prefix3 | Troisième préfixe du mot (s’il existe) |
| Prefix4 | Quatrième préfixe du mot (s’il existe) |
| Postfix1 | Premier suffixe du mot (s’il existe) |
| Postfix2 | Deuxième suffixe du mot (s’il existe) |
| Postfix3 | Troisième suffixe du mot (s’il existe) |
| Sura | Le numéro de la sourate |
| Aya | Le numéro de l’aya |

Tableau 16 : La structure d’index principal – Midād lbayān

**Aperçu :**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Seq | Word | Stem | Root | Othmani | Prefix1 | Prefix2 | Prefix3 | Prefix4 | Postfix1 | Postfix2 | Postfix3 | Sura | Aya |
| 1 | # | # | # | # |  |  |  |  |  |  |  | 1 | 0 |
| 2 | @ | @ | @ | @ |  |  |  |  |  |  |  | 1 | 1 |
| 3 | بِسْمِ | اسْم | سمى | بِسْمِ | بِ |  |  |  | ِ |  |  | 1 | 1 |
| 4 | اللَّهِ | اللَّه | ءله | ا?للَّهِ |  |  |  |  | ِ |  |  | 1 | 1 |
| 5 | الرَّحْمَنِ | رَحْمَن | رحم | ا?لرَّحْمَ!نِ | ال |  |  |  | ِ |  |  | 1 | 1 |
| 6 | الرَّحِيمِ | رَحِيم | رحم | ا?لرَّحِيمِ | ال |  |  |  | ِ |  |  | 1 | 1 |
| Tableau 17 : Aperçu sur l’index principal – Midād lbayān | | | | | | | | | | | | | |

1. Évaluation

Midād lbayān est considéré comme un grand accomplissement, il contient dans ses index beaucoup de renseignements sur les mots coraniques et ses caractéristiques et a également des perspectives d’ajout d'autres index pour étudier les phrases du Coran, l'indexation des sujets, etc.

Les auteurs de projet Midād lbayān confirment que la base est destinée pour l’utilisation gratuite, mais ils ne donnent pas plus de détails sur la licence utilisée. Ils ne fournissent qu’un programme d’exposition manuelle de la base, sans références, en plus ils ne répondent pas aux messages, bien qu’ils disent qu'ils les attendent. **[Midād lbayān, 10]**

1. Les Index par M.Taha Zerrouki
2. Définition

C’est un effort individuel de Mr. Taha Zerrouki dans le cadre du projet de Coran et ses sciences, que notre travail en fait partie, il a construit trois index qui sont l'index de mots, l'index de synonymes, et l’index de sujets, nous parlerons de chacun d'eux séparément par la suite :

1. Les index et sa structure

**L’index du mot:**

**Structure :**

|  |  |
| --- | --- |
| Champ | Commentaire |
| الكلمة الأصلية | Le mot original |
| الجذر | La racine |
| المصدر | Le morphème |
| النوع | Le type |
| الجمع | Le pluriel |
| المؤنث | Le féminin |
| المثنى | Le pluriel du deux personnes |
| العلم | Le nom propre |
| الكتابة العثمانية | Le mot en script othmani |

Tableau 18 : La structure d’index des mots – M.Taha Zerrouk

**Aperçu :**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| الكلمة الأصلية | الجذر | المصدر | النوع |  | الجمع | المؤنث | المثنى |  | العلم | كتابة عثمانية |
| البسط | بسط | بسط | اسم |  |  |  |  |  |  | البسط |
| مبسوطتان | بسط | مبسوطة | اسم | مثنى |  |  | مبسوطتان | مبسوط |  | مبسوطتان |
| فيبسطه | بسط | بسط | فعل |  |  |  |  |  |  | فيبسطه |
| يبسطوا | بسط | بسط | فعل |  |  |  |  |  |  | يبسطوا |
| باسقات | بسق | باسقة | اسم | جمع | باسقات | باسق |  |  |  | باسقت |

Tableau 19 : Aperçu sur l’index des mots – M.Taha Zerrouki

**L’index des sujets :**

**Structure :**

|  |  |
| --- | --- |
| Champ | Commentaire |
| الفصل | le thème |
| الموضوع | le sous-thème |
| الباب | la rubrique |
| رقم السورة | Le numéro de la sourate |
| آية البداية | Le numéro de la première aya |
| آية النهاية | Le numéro de la dernière sourate |

Tableau 20 : La structure d’index des sujets – M.Taha Zerrouki

**Aperçu :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| الفصل | الموضوع | الباب | رقم السورة | آية البداية | آية النهاية |
| أسماء الله تعالى وصفاته | صفات الله تعالى | مشيئته تعالى | 13 | 31 |  |
| أسماء الله تعالى وصفاته | صفات الله تعالى | مشيئته تعالى | 16 | 93 |  |
| أسماء الله تعالى وصفاته | صفات الله تعالى | مشيئته تعالى | 17 | 54 |  |
| أسماء الله تعالى وصفاته | صفات الله تعالى | عدله تعالى | 2 | 272 |  |
| أسماء الله تعالى وصفاته | صفات الله تعالى | عدله تعالى | 2 | 281 |  |
| Tableau 21 : Aperçu sur l’index des sujets – M.Taha Zerrouki | | | | | |

**L’index des synonymes**

**Structure :**

|  |  |
| --- | --- |
| Champ | Commentaire |
| ملاحظة | Remarque |
| النوع | Le type des mots du synonyme |
| المفردات | Les synonymes |

Tableau 22 : La structure d’index des synonymes – M.Taha Zerrouki

**Aperçu :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ملاحظة | النوع | المفردات |  |  |  |
| ألف | فعل | أَبِقَ | فَرَّ | نَاصَ | هَرَبَ |
| جمع | اسم | إِبِلٌ | عِيرٌ |  |  |
|  | اسم | أَبْ | وَاِلْد |  |  |
|  | فعل | أَتَى | جَاءَ | حَضَرَ | أَقْبَلَ |
|  | فعل | آثَرَ | اخْتَارَ | اصْطَفًى | فَضَّلَ |

Tableau 23 : Aperçu de l’index des synonymes – M.Taha Zerrouki

1. Évaluation

Ces index sont offerts à nous, sans rentrer dans les détails de licences et de leurs problèmes. Ce qui les distingue est leur diversité : d’un index pour les racines à un l'index des sujets et un autre de synonymes. Mais ils ne sont pas basés sur une unité claire (contrairement à Midād lbayān et au corpus coranique qui sont basés sur les mots du Coran en prenant tous ses occurrences) cela augmente la complexité et la probabilité d’erreur dans le processus de jointure avec les autres bases de données.

1. The Qur’anic Arabic Corpus
2. Définition

Qur’anic Arabic Corpus (le Corpus Coranique Arabe) est une ressource linguistique annotée qui montre la grammaire arabe, la syntaxe et la morphologie de chaque mot dans le Coran, faite par Kais Dukes en 2009 à l’école d’informatique, Université de Leeds. Nous allons aborder la version disponible en téléchargement. **[Corpu, 10]**

1. Les index et sa structure

**Index de la morphologie :**

**Structure :**

|  |  |
| --- | --- |
| Champ | Commentaire |
| chapter | Le numéro de la sourate dans le Coran |
| verse | Le numéro de l’aya dans la sourate |
| word | Le numéro du mot dans l’aya |
| token | Le mot en script othmani (Unicode) |
| morphology | Une expression codée qui décrit le mot morphologiquement et grammaticalement |

Tableau 24 : La structure d’index de morphologie – Corpus coranique

**Aperçu :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| chapter | verse | word | token | morphology |
| 1 | 1 | 1 | بِسْمِ | bi+ POS:N LEM:{som ROOT:smw M GEN |
| 1 | 1 | 2 | ٱللَّهِ | POS:PN LEM:{ll~ah GEN |
| 1 | 1 | 3 | ٱلرَّحْمَٰنِ | Al+ POS:ADJ LEM:r~aHoma`n ROOT:rHm MS GEN |
| 1 | 1 | 4 | ٱلرَّحِيمِ | Al+ POS:ADJ LEM:r~aHiym ROOT:rHm MS GEN |
| 1 | 2 | 1 | ٱلْحَمْدُ | Al+ POS:N LEM:Hamod ROOT:Hmd M NOM |
| 1 | 2 | 2 | لِلَّهِ | l:P+ POS:PN LEM:{ll~ah GEN |
| 1 | 2 | 3 | رَبِّ | POS:N LEM:rab~ ROOT:rbb M GEN |
| 1 | 2 | 4 | ٱلْعَٰلَمِين | Al+ POS:N LEM:Ea`lamiyn ROOT:Elm MP GEN |

Tableau 25 : Exemple d’index de morphologie – Qur’anic corpus

1. Évaluation

Cet Index est disponible sur Internet avec d'autres index simplifiés sous la licence GPL (GNU Public License), mais il impose des restrictions supplémentaires lors du téléchargement y compris l’interdiction de modification. Le corpus coranique fournit des informations grammaticales précieuses sur les mots du Coran sur lesquels l’index est basé, mais il les formule d'une façon spéciale et complexe qui rend difficile leur extraction.

1. Tanzil Quran Navigator
2. Définition

Tanzil est un projet coranique qui vise à fournir un texte précis du Coran et hautement vérifié. Il a été développé par Hamid Zarrabi-Zadeh et révisé par une équipe de vérification du texte.

Tanzil fournit plusieurs textes du Coran en deux grandes catégories décrites ci-dessous :

* **Script othmani**
  + **Othmani** : texte du Coran en script othmani similaire au « Mushaf Al-Madina ».
  + **Othmani minimal**: texte en script othmani avec le minimum de signes de diacritiques et de symboles.
* **Script simple**
  + **Simple** : le texte du Coran en script standard plein texte, recommandé pour l'affichage et la recherche.
  + **Simple amélioré** : texte simple avec une meilleure démonstration d’ikhfā’ et d’idghām[[188]](#footnote-13), simplifié pour la lecture.
  + **Simple minimal** : texte simple avec un nombre minimum de signes de diacritiques et de symboles. Convenable pour l’intégration dans d'autres textes.
  + **Simple propre** : texte simple sans signes de diacritiques ou de symboles. Convenable pour la recherche facile. **[Zarrabi-Zadeh, 09]**

Tanzil fournit également certains index comme :

* index de la surate ;
* index du sajdah ;
* index structurel en :
* Juz’ ;
* Rub’ ;
* Manzil ;
* Ruku’ ;
* Page.

1. Les index

**Un modèle du Coran comme un index d’ayas :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sura index | sura name | aya index | aya texte | bismillah |
| 1 | الفاتحة | 1 | بسم الله الرحمن الرحيم |  |
|  |  | 2 | الحمد لله رب العالمين |  |
|  |  | 3 | الرحمن الرحيم |  |
|  |  | 4 | مالك يوم الدين |  |
|  |  | 5 | إياك نعبد وإياك نستعين |  |
|  |  | 6 | اهدنا الصراط المستقيم |  |
|  |  | 7 | صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين" |  |
| 2 | البقرة | 1 | الم | بسم الله الرحمن الرحيم |
| … | … | … | … | … |

Tableau 26 : Exemple du Coran modèle [simple propre](http://tanzil.info/twiki/index.php?title=Simple_Clean&action=edit) – Tanzil.info

**L’index de surate :**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| index | ayas | start | name | tname | ename | type | order | rukus |
| 1 | 7 | 0 | الفاتحة | Al-Faatiha | The Opening | Meccan | 5 | 1 |
| 2 | 286 | 7 | البقرة | Al-Baqara | The Cow | Medinan | 87 | 40 |
| 3 | 200 | 293 | آل عمران | Aal-i-Imraan | The Family of Imraan | Medinan | 89 | 20 |
| 4 | 176 | 493 | النساء | An-Nisaa | The Women | Medinan | 92 | 24 |
| … | … | … | … | … | … | … | … | … |

Tableau 27 : Exemple d’index de surate – Tanzil.info

**L’index de sajdah :**

|  |  |  |  |
| --- | --- | --- | --- |
| index | sura | aya | type |
| 1 | 7 | 206 | recommended |
| 2 | 13 | 15 | recommended |
| 3 | 16 | 50 | recommended |
| 4 | 17 | 109 | recommended |
| 5 | 19 | 58 | recommended |
| 6 | 22 | 18 | recommended |
| 7 | 22 | 77 | recommended |
| 8 | 25 | 60 | recommended |
| 9 | 27 | 26 | recommended |
| 10 | 32 | 15 | obligatory |
| 11 | 38 | 24 | recommended |
| 12 | 41 | 38 | obligatory |
| 13 | 53 | 62 | obligatory |
| 14 | 84 | 21 | recommended |
| **Tableau 28 : L’index de sajdah – Tanzil.info** | | | |

**Les index structurels :**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | index | sura | aya | | 1 | 1 | 1 | | 2 | 2 | 142 | | 3 | 2 | 253 | | 4 | 3 | 93 | | 5 | 4 | 24 | | … | … | … |   **Tableau 29 : L’index de juz’ – Tanzil.info** | |  |  |  | | --- | --- | --- | | index | sura | aya | | 1 | 1 | 1 | | 2 | 2 | 26 | | 3 | 2 | 44 | | 4 | 2 | 60 | | 5 | 2 | 75 | | … | … | … |   **Tableau 30 : L’index de rub’ – Tanzil.info** | |  |  |  | | --- | --- | --- | | index | sura | aya | | 1 | 1 | 1 | | 2 | 2 | 1 | | 3 | 2 | 8 | | 4 | 2 | 21 | | 5 | 2 | 30 | | … | … | … |   **Tableau 31: L’index de ruku’ – Tanzil.info** |
| |  |  |  | | --- | --- | --- | | index | sura | aya | | 1 | 1 | 1 | | 2 | 2 | 1 | | 3 | 2 | 6 | | 4 | 2 | 17 | | 5 | 2 | 25 | | … | … | … |   **Tableau 32: L’index de pages - Tanzil.info** |  | |  |  |  | | --- | --- | --- | | index | sura | aya | | 1 | 1 | 1 | | 2 | 5 | 1 | | 3 | 10 | 1 | | 4 | 17 | 1 | | 5 | 26 | 1 | | 6 | 37 | 1 | | 7 | 50 | 1 |   **Tableau 33 : L’index de manzil - Tanzil.info** |

1. Évaluation

Tanzil est un projet libre à utiliser et téléchargeable sur son site public. Il bénéficie d'une grande crédibilité vu qu’un groupe de chercheurs veille à le vérifier.

1. Conclusion

Il existe beaucoup d’index (ou ma‛ājim) de Coran qui varient d'index pour les mots à des index pour les sujets et d'autres spécialisés … Mais leur plus grand problème est qu’ils ne sont pas entrés dans le domaine de l’informatique, puisque la plupart des ma‛ājim sont imprimés, ou destinés à la lecture ; seule une petite partie a été informatisée et publiée sur des fichiers de stockage d'informations tels que XML, EXCEL ou sous forme de bases de données. Et parmi les travaux les plus importants dans l'informatisation des index (dont nous avons parlé dans ce chapitre) le travail de Taha Zerrouki et le projet de ‘Midād lbayān’. Cependant, ces projets sont encore en cours de développement pour vérifier ce qui a été et achevé et compléter ce qui manque ...

# Chapitre V

**Conception**

1. Introduction

Nous présentons dans cette partie, la conception de notre moteur de recherche.

Notre travail sera basé dans un premier temps sur la forme générale des moteurs de recherche ordinaires, afin de bénéficier des fonctionnalités qu’ils fournies, ensuite nous allons le personnaliser pour ajouter les fonctionnalités de recherche dans le Coran ainsi que les fonctionnalités linguistiques mentionnées dans le deuxième chapitre.

Ce chapitre utilise des schémas illustratifs pour décrire le rôle et la situation de chaque composant du système, mais sans détailler comment le réaliser et les outils utilisés, ces détails seront abordé au chapitre VI.

Dans cette partie :

* Nous commencerons par le modèle de recherche basique, puis nous ferons des ajouts et des modifications de façon progressive selon les fonctionnalités souhaités.
* Nous ne détaillerons pas trop les parties qui s’appuient sur la méthode des moteurs de recherche ordinaires, mais ils seront plutôt présentés comme des boites noires.
* Certains éléments nécessitent une étude spéciale et approfondie et d'autres nécessitent des projets parallèles, nous allons nous référer à tout cela par la suite.

Les fonctionnalités de recherche que nous essayons de réalisé se résume dans :

1. Les options de la recherche avancée : elles incluent la recherche par champs, partie du mot, jokers, relations logiques, mots proches, intervalle, et le boosting (privilégier des termes sur d’autres).
2. Les fonctionnalités additionnelles : leur but est des faciliter la tache de recherche pour l’utilisateur, se sont : la suggestion de mots-clés alternatifs, la suggestion de mots-clés apparentés (tel que la correction la considération d'orthographe), la pagination, le tri (Sorting), et surligner les mots recherchés (Highlight).
3. Les aspects linguistiques : ils comportent tous options liés à la langue de manière générale comme : la recherche par signification, par lemme, par synonymes et par antonymes.
4. Les aspects linguistiques de recherche pour la langue arabe : ce sont les options spécifiques pour la langue arabe, tels que : La considération des signes diacritiques الحركات, les préfixes et suffixes du mot, la recherche par racine, la conjugaison des verbes et déclinaison des noms, recherche par la natures des mots, recherche par les patterns, respecter les différences de Hamza, respecter la différence Hâ’ et tâ’ marbûtä et entre Yâ’ et Alif maqsûrä.
5. Les options spécifiques du Coran : ils incluent la recherche par le script othmani, par des index spéciaux, les options coraniques structurelles, les options statistiques et l’extension de la recherche.
6. Le modèle de base :

La recherche simple est le mode de recherche par défaut ou le mode de base.  
Pour atteindre cet objectif, nous appuyions sur la méthode générale des Systèmes de Recherche d’Information (*voir la figure 29*).



**Figure 29 : Le modèle de base**

Chaque moteur de recherche récupère les données à partir d'une source particulière, notamment les pages web et des fichiers existants dans les ordinateurs. Par conséquent, ils ont besoin essentiellement du processus de l'exploration. Par contre, dans le cas de notre moteur de recherche, l’unique source de données est le Coran, et donc pas besoin de ce processus. Au moins dans la recherche simple.

Le texte coranique est divisé principalement en ayas et en surates. Nous considérons l’aya comme l’unité de recherche, vu que le nombre de sourate est limité et certains ont une grande taille. L’aya est définie par son numéro et le nom de la sourate à laquelle elle appartient, ces deux informations devraient donc être incluses dans les index.

Le texte du Coran s’écrit par nature en script othmani, il est riche en symboles de diacritiques et de récitation, et nécessite un traitement particulier que nous aborderons ultérieurement. Nous allons adopter, pour fournir une recherche simple, un texte prétraité et sans aucun signes de diacritiques.

Le tableau ci-dessous montre le format par défaut d'index de documents et comment il est converti en index inversé :



Figure 30 : L’index de documents et l’index inversé

Comme son nom indique, le chercheur est l’élément chargé de la recherche. Il reçoit les requêtes, qui sont dans ce cas des mots du Coran, puis recherche dans l'index inversé pour les mots de l’aya, afin d’obtenir les **Doc-IDs,** et les utiliser pour trouver les informations complètes de l'index des documents, qui sont le texte et le numéro de l’aya et le nom de sourate auquel elle appartient, ensuite envoie les résultats vers les interfaces.

La figure suivante montre le comportement de chercheur :



Figure 31 : Le comportement de chercheur



Le texte coranique est complètement vocalisé[[189]](#endnote-176) (مشكول), cependant les requêtes des utilisateurs ne seront pas nécessairement vocalisées de la même manière que le texte du Coran. Donc, nous avons mis la considération et la négligence des signes de diacritiques comme un de nos objectifs.

A cette fin, nous avons proposé la modification du processus de comparaison entre les mots-clés indexés et celles recherchés, qui est d’origine une opération de comparaison littérale.  
Le positionnement de la nouvelle comparaison est décrit dans le schéma ci-dessous :

Figure 32 : Le positionnement de la nouvelle comparaison

Notre modification vise à ajouter l’option de considérer/négliger les signes de diacritiques d'une manière partielle et intelligente comme suite :

* Les diacritiques sont séparées des lettres, en gardant pour chaque diacritique l’index de la lettre qui la porte.
* La comparaison entre les deux mots s’effectue après avoir enlevé les signes diacritiques.
* Si le résultat de la comparaison est négatif, l'algorithme s’arrête ici avec un résultat négatif.
* Si le résultat de la comparaison est positif, les diacritiques sont comparés un par un selon les règles suivantes :
  + dans le cas où au moins l'un des mots n’est pas vocalisé le résultat est positif ;
  + dans le cas d'un caractère vocalisé différemment dans les deux mots, le résultat est négatif ;
  + la chedda[[190]](#endnote-177) (الشّدّة) est traité particulièrement. Elle ne peut être comparée qu'avec elle-même, et est considérée obligatoire si elle est trouvée dans l'un des mots.
* Si la comparaison précédente réussit pour touts les caractères, alors le résultat est positif, sinon le résultat est négatif.

|  |  |  |
| --- | --- | --- |
| Le résultat | Le deuxième mot | Le premier mot |
| Vrai | المٌلكٌ | المُلك |
| Faux | المٌلكٌ | المَلك |
| Faux | الخَنَّاس | الخناس |
| Vrai | الخَنَّاس | الخنّاس |

Tableau 34 : Exemples de la comparaison en considérant les diacritiques

**Remarque** : l’indexation est refaite avec un texte coranique vocalisé.

1. Traitement de texte

Nous avons motionné qu’on ait utilisé un texte coranique prétraité dans le modèle de base, nous passons maintenons au traitement textuel au niveau des processus de l’indexation et la recherche. Ce traitement ce fait globalement selon les propriétés de texte arabe, et selon les propriétés de texte coranique en particulier. Il nous apporte l’option de rechercher sans suffixes.

Le schéma suivant montre l’emplacement du traitement de texte.



Figure 33 : L’emplacement du traitement de texte

Le traitement de texte est généralement constituer de deux phases principales :  
La première est la phase d’extraction des tokens (tokenization) où les phrases complexes sont devisées en tokens. La seconde est l'étape le traitement des tokens, elle consiste à la normalisation de caractères et à ignorer les stopwords. Ce qui reste passe par le processus de la segmentation.

Nous nous appuierons sur ces mêmes étapes dans le traitement du texte coranique, mais en les personnalisant.

****

Figure 34 : Traitement de texte

1. Extraction des tokens (Tokenization)

La tokenization est l’opération de fractionner le texte en unités indépendantes appelées tokens. Dans les textes littéraires, souvent les tokens représentent les mots.

Pour le texte coranique, le mot avec tous ses affixes littérales et nominaux constitue l’unité qui peut être indépendante car elle n’est pas liée avec son voisinage en matière de grammaire.

Il est possible de séparer les affixes nominaux comme les pronoms, mais cela nécessitent une étude plus large des différentes situations où le pronom peut être séparé et les procédures qui suivent cette séparation, spécialement qu’il peut figurer en tant que noyau d’un mot et pas affixe comme dans l'exemple : « لهنّ »

Figure 35 : Tokenization

1. Normalisation

Comme nous avons indiqué précédemment dans le chapitrer 2 (IV.D) « Les aspects linguistiques de recherche pour la langue arabe » à propos des caractéristiques de la langue arabe, les lettres changent leur forme en fonction de leur emplacement (ce qui est appelée façonnage). Par conséquent, le code de la lettre peut être différent dans certains programmes qui enregistrent le texte avec le code des lettres à l’état attaché et pas à l’état général. Le problème majeur du façonnage est le symbole "لا" qui doit être décomposé en deux symboles "ل" + "ا", idem pour «لأ», Il y a aussi le mot «الله» qui peut être représenté comme un seul symbole en Unicode.

Donc parmi les fonctions du normalisateur est de retourner les lettres à leur état générale et séparer les symboles qui représentent plus d'une lettre.

Certain moteurs de recherche arabes suppriment les diacritiques[[191]](#endnote-178) à ce stade, mais pour nous ce n'est pas utile car cela néglige totalement les signes de diacritiques, alors que nous voulons avoir l’option d’ignorance/prise en compte des diacritiques.

La même chose pour la forme du hamzä[[192]](#endnote-179) et les fautes d’écriture (tâ’ marbûta[[193]](#endnote-180) – hâ’[[194]](#endnote-181), alif maqsûra[[195]](#endnote-182) – yâ’[[196]](#endnote-183)), l’unification de leurs code à ce stade désactive l'option de recherche avec le symbole correspondant, pour cela nous reportons la réalisation de ces option à la phase de recherche, et exactement, au moment du traitement de la requête.

Figure  : Normalisation

1. Segmentation (stemming)

Le processus du segmentation diffère en langue arabe qu'en anglais en termes de degré de complexité, car le Segmentation, en anglais, supprime uniquement les affixes pour obtenir le stem qui est la plus petite partie qui regroupe les mots appartenant à la même famille, mais dans la langue arabe, la suppression des affixes ne donne pas à l’origine due à la nature des langues flexionnelles à brisure interne, mais retourne le mot dérivé. On obtient ce dernier avec une opération de dérivation ou une série d’opérations de dérivation. A partir de notre observation de la langue arabe et les différents index du mot, nous avons vu de la diviser en 3 niveaux de traitement :

1. enlever les affixes littérales[[197]](#endnote-184) et nominaux[[198]](#endnote-185) ;
2. retourner les noms à ses morphèmes (مصدر) et les verbes au temps passé avec la troisième personne du singulier (il) ;
3. retourner les mots à leurs racines, la racine étant la plus petite unité qui rassemble les familles du mot dans la langue arabe.

Figure 37 : Les niveaux de dérivations

Il existe une incertitude dans la définition de la segmentation. Il y a ceux **[Duwairi, 08]** qui le considèrent comme l’élimination des affixes, et d’autres **[Alchalabi, 05]** le considèrent comme la remonte du mot à son origine. Mais nous avons tendance à la première définition c.-à-d. enlever les affixes du mot arabe sans le retourner à sa racine ou se qu’on appelle « segmentation légère » pour plusieurs raisons qui sont **[Thabet, 04]** :

* retourner le mot au niveau de la racine élargie considérablement la recherche, parce que la racine a de nombreux mots dérivés. Ils peuvent être limités en nombre dans le Coran, mais leur nombre reste important spécialement que certains dérivés sont différents dans le sens et parfois antonymes ;
* retourner le mot au niveau de la racine désactive d’autres options linguistiques comme la recherche par mot identique, par les déclinaisons[[199]](#endnote-186) des noms et par les conjugaisons des verbes ;
* le calcul des options statistiques que nous voulons offrir dépend des mots indexés, donc retourner les mots de leurs racines fait perdre ces options, car les statistiques ne peuvent être construites que sur des mots identiques.

L’élimination des affixes seulement ne manque pas l'occasion de la recherche par déclinaison, conjugaison et racines car elle garde l’information dont nous avons besoin pour ce faire, nous l’expliquerons par la suite dans le traitement de la requête.

Les statistiques vont être améliorées car ils s’appuieront sur des mots à unité abstraite[[200]](#endnote-187) libre de touts les affixes qui ajoutent des informations au mot, mais sans changer son sens. Par exemple, nous ne somme pas intéressés par les statistiques du mot «فقال», mais plutôt du mot « قال ».

Alors la segmentation va supprimer les affixes seulement, mais les affixes sont de toutes sortes, y compris le nom et la particule[[201]](#endnote-188), lesquels est-ce que nous supprimerons et comment le faire ?

Spécialement dans des cas particuliers tels que ceux-ci :

* seuls les particules (les prépositions[[202]](#endnote-189) et les conjonctions de coordination[[203]](#endnote-190)) et les pronoms relatifs[[204]](#endnote-191) sont supprimés, les terminaisons des verbes conjugués ne sont pas considérées comme des affixes, exemple : « فعل », « فعلن » et « فَعَلَا » ;
* dans le cas où l’affixe modifie la forme du mot, telle que la suppression du Lâm[[205]](#endnote-192) dans « لله » d'origine « الله », ainsi que l’ajout du chedda dans « الشّمس » d'origine « شمس » ;
* dans le cas où le mot est constitué de particules et pronoms qui sont touts considérés comme des affixes, telles que « لهنّ ».

Le traitement de ces cas dépend de l’appréciation des chercheurs qui travail sur ce domaine, car ils connaissent mieux les sens du Coran. Ce qui importe est que le mot résultant représente le meilleur cas de généralisation.

Pour le traitement des diacritiques lors de la segmentation, elles sont recopiées telles qu’elles sont, sauf dans les cas suivants :

* les diacritiques sur les affixes son éliminées avec eux ;
* les diacritiques résultant de l’ajout des affixes comme la chedda sur les lettres chamsiyya[[206]](#endnote-193) sont aussi éliminées ;
* la marque de la flexion sur la dernière lettre exprime l’état flexionnelle[[207]](#endnote-194) et le mot ne se change pas (sémantiquement) quand elle se change, donc il vaut mieux la négligées.

Figure 38 : La segmentation

1. Filtrage des stopwords

Les stopwords sont des mots ignorés lors du processus de recherche, ils varient en fonction de la langue et le domaine. Cette phase consiste à détecter ces mots et les ignorer.

Nous n’avons pas utilisé la traduction du mot stopwords en français qui est « mots vides » car chaque lettre et mot du Coran portent un sens et une valeur et ne sont pas vide.  
**[خضر, 08]**

Mais pour le Coran, y-t-il vraiment des stopwords dans le Coran ? Si oui, alors comment choisir les mots à exclure ?

Pour répondre à la possibilité d’existence des stopwords, nous devons rappeler les objectifs du processus de leur filtrage. Le but n’est pas de les négliger pour que l'utilisateur ne soit pas capable de rechercher en utilisant ces mots, mais plutôt d'éviter que ces mots submergent les résultats de la recherche, sans les autres mots qui peuvent être ce que l'utilisateur cherche. En arabe, il y a une ambigüité entre quelques particules[[208]](#endnote-195) fréquentes d’une part, et les noms et les verbes rares d'une autre part, dû à l'absence de la vocalisation, tels que : « مِن » et « مَنّ ».

C'est la principale raison de l'utilisation du filtrage des stopwords, elle est facultatives et sans supprimer totalement les mots, mais seulement les nommer pour un traitement spécial dans le calcul du score et dans la recherche.

D'abord il faut considérer les points suivants :

* l’utilisation des stopwords non vocalisés (مشكولة), conduit à ignorer un grand nombre de mots qui ne sont pas des stopwords, mais qui ont les mêmes lettres d’un stopword ;
* le Coran a un nombre limité de mots, et donc certains stopwords en arabe ne seront pas des stopwords dans le Coran où leur présence est rare ;
* la plupart des conservateurs du Coran le mémorise tel qu'il est et non pas avec le sens comme dans le cas des textes arabes, et donc un mauvais choix de mots exclus peut produire une confusion à l'utilisateur, lorsqu’il ne trouve pas l’aya qu’il cherche parmi les premières pages.

Un exemple de ça, est la recherche de la phrase : « لو كان البحر ».

Dans la langue arabe on peut considérer « لو » et « كان » comme des mots exclus, et il ne reste donc que le mot « البحر » pour la recherche, en conséquence l’aya ci-dessous sera perdu dans les autres ayas qui contiennent le mot « البحر » seulement sans « لو كان » :

**[قُل لَّوْ كَانَ الْبَحْرُ مِدَاداً لِّكَلِمَاتِ رَبِّي لَنَفِدَ الْبَحْرُ قَبْلَ أَن تَنفَدَ كَلِمَاتُ رَبِّي وَلَوْ جِئْنَا بِمِثْلِهِ مَدَداً ]**

La conclusion est que nous devons choisir des stopwords de la liste des mots les plus fréquents dans le Coran, où ils peuvent être des particules (حرف) ou des pronoms[[209]](#endnote-196) (ضمير) et ils devaient être pris avec les diacritiques.  
L’option de désactivation de cette phase est fournie dans le cas d’un processus de recherche à des fins statistiques.

**Figure 39 : Segmentation et filtrage des stopwords**

1. Traitement de la requête

La recherche avancée est constituées d’un ensemble d’options de recherche. Chez plusieurs moteurs de recherche, la recherche avancée permet d’exprimer des requêtes assai complexes comme le choix de la langue ou le type des résultats. Par exemple, le moteur de recherche Google comporte une page dédie à cet effet, de plus on peut indiquer ces options de cherche directement à travers une syntaxe bien déterminée.

Les options de recherche qui existe actuellement dans les moteurs de recherche sont :

* recherche par champs (fielded search) ;
* recherche par une partie de mot ;
* recherche par les jokers (les caractères spéciaux) ;
* recherche par les relations logiques ;
* recherche par une phrase ;
* recherche par mots proches ;
* recherche par intervalle ;
* boosting.

Nous profitons de la recherche avancée dans la mise en œuvre des fonctionnalités linguistiques de recherche qui sont considérés parmi nos objectifs et qui sont :

* recherche par synonymes et antonymes ;
* recherche par lemme ;
* recherche par racine ;
* conjugaison des verbes et déclinaison des noms ;
* recherche par la nature des mots arabes ;
* recherche par les patterns ;
* respecter les différentes forme de la Hamza ;
* respecter la différence entre Hâ’ et tâ’ marbûtä et entre Yâ’ et Alif maqsûrä.

****

Figure 40 : Le positionnement du traitement de la requête

En bref, nous allons expliquer les étapes les plus importantes de la requête à partir de la saisie par l'utilisateur jusqu'à ce qu’elle atteigne le chercheur. La première opération est le parseur qui explique la requête et fait sortir les termes et les classe selon les champs de recherche.

Les tokens sont traités de la même manière que les mots des ayas qui ont été traités lors de l'indexation (voir traitement de texte), puis ils sont envoyés au chercheur.



Figure 41 : Traitement de la requête

Le parseur est la partie principale ici, il reçoit la requête complexe, son rôle est d’interpréter la requête et la retourner sous la forme d'un arbre de termes, puis attribuer une opération à chaque terme, on aura ainsi un arbre d’opérations, son exécution récurrente donne un ensemble de couples constitués de champs et de mots-clés utilisés dans la recherche et le highlight ainsi que les suggestions …



**Figure 42 : Le parseur**

Nous allons expliquer seulement les nouveaux processus que nous avons ajoutés, et mentionner juste le rôle des processus qui existent déjà.



Figure 43 : Les opérations du parseur: existantes et nouvelles

Les opérations existantes déjà dans les moteurs de recherche sont :

|  |  |
| --- | --- |
| Opération | rôle |
| terme simple | Son but est d’identifier le terme de base qui n'est pas soumis à des ajustements dans le parseur. C’est le résultat final de toutes les opérations en son sein. |
| relations logiques | Définir les opérations logiques entre les termes qui sont : ET, OU, et NON. |
| Champs | Spécifier la recherche pour l’un des termes à un champ donné. |
| Groupage | Grouper plusieurs termes dans le cas des opérations logiques. |
| Intervalle | Recherche par intervalles |
| Jokers | Trouver les mots qui correspondent à l’expression régulière. |
| Boosting | Contrôler l'importance de l’un des termes par rapport aux autres. |

Tableau 35 : Rôles et syntaxe des opérations existantes de parseur

Pour réaliser les options linguistiques dont nous avons parlé précédemment, nous ajoutons quatre opérations, à savoir :

* **Synonymes et antonymes**
* **Dérivations des mots :** Pour réaliser la recherche par racine, morphème, conjugaison des verbes er déclinaison des noms.
* **La considération d'orthographe :** pour respecter/ignorerles différentes forme de la Hamza, la différence entre Hâ’ et tâ’ marbûtä et entre Yâ’ et Alif maqsûrä.
* **La recherche par tuple:** pour larecherche par la nature des mots arabes, la recherche par les patterns et la recherche par racine.

L’explication de ces opérations est dans les points suivants :

1. Synonymeset antonymes

C’est un processus simple du remplacement du mot par l’ensemble des ses synonymes et antonymes.

**Figure 44 : Le remplacement par synonymes et antonymes**

1. Dérivations des mots

C’est une option linguistique, son objectif est d'extraire les dérivés d’un mot donné selon un niveau de dérivation déterminé, elle se passe en 3 étapes :

* Déterminer le niveau de la dérivation du mot, est-ce que c’est un :
  + mot complet ;
  + mot dépourvu d'affixes ;
  + morphème ou verbe ;
  + racine.
* Puis, récupérer l'origine du mot du niveau requis.
* Après la récupération de l'origine du mot, on détermine tous ses dérivés.

La méthode précédente peut être implémentée en utilisant des index spécifiques qui contiennent les différents niveaux de dérivation de tous les mots coraniques, on peut observer un exemple d’index dans la figure suivante :

Figure 45 : Niveaux de dérivation d’un index

Une autre méthode, par les analyseurs morphologiques s’ils disposent d’un taux de précision élevé et atteindre l'objectif requit, notamment dans la première étape, où la méthode de l’index reste insuffisante car elle ne peut pat extraire les origines des mots qui ne figurent pas dans le Coran.

Cependant, la méthode d’index est la meilleure dans la deuxième étape, car les analyseurs morphologiques génèrent énormément de résultats qui ne sont pas inclus dans le Coran, et causent une dégradation de la performance.

**Figure 46 : La recherche par les dérivés de la racine du mot « فَأَسْقَيْنَاكُمُوه » par la méthode d’index**

1. La considération d'orthographe

Avec cette option, le moteur de recherche peut ignorer toutes les fautes d'orthographe spécifiques de la langue arabe[[210]](#footnote-14) :

* les fautes dans la forme de la hamzä ;
* l’ambiguïté entre tâ’ marbûta (ة) et hâ’ (ه) et entre alif maqsûra (ى) et yâ’ (ي).

Les mots saisis sont traités comme suit :

* + la tâ’ marbûta se transforme en hâ’ ;
  + l’alif maqsûra se transforme en yâ’ ;
  + toutes les hamzäs se transforment en hamzä individuelle (ء).

Ensuite, le moteur de recherche examine tous les mots indexés et les traite de la même manière pour les comparer avec les mots saisis. Tous les mots qui donnent des résultats de comparaison positifs sont ajoutés à la liste des mots sortants.

Le traitement temporaire : il est uniquement pour la comparaison, les mots sortants sont les mots corrects et non pas ceux qui ont été traités.

Figure 47 : Exemples de considération d'orthographe

1. La recherche par tuple:

Nous avons parlé dans le deuxième chapitre (V) « Un algorithme proposé de la recherche arabe » de la recherche avec le triple (racine, type, pattern), nous appliquerons la même idée ici et nous la développerons pour inclure d'autres options.

Le comportement général de cette propriété est que l'utilisateur entre le triple contenant les caractéristiques linguistiques d'un mot, puis le moteur de recherche cherche dans l'index pour extraire tous les mots correspondants à ces caractéristiques.

Ces derniers sont les mots-clés externes.

Nous pouvons ici utiliser toutes les caractéristiques linguistiques du mot, tels que :

La racine, le type, le pattern, la signification, synonyme, antonymes

Nous pouvons rendre n’importe quelle option facultative, par exemple :

**Figure 48 : La recherche par couple (racine, type)**

1. Les suggestions

La méthode de fonctionnement des modules de suggestion est généralement : pendre les mots-clés résultant du parseur et non pas directement de la requête brute entrée par l'utilisateur car certains mots-clés sont traités et leur forme change et aussi, de nouveaux mots sont ajoutés, comme dans le cas des synonymes et antonymes. Le module, grâce à une certaine politique, trouve les mots qui ont une relation particulière avec les mots-clés.



Figure  : Le système de la suggestion

Nous avons motionné dans le chapitre 2 (IV.B.1 « Suggestion de mots-clés alternatifs » et IV.B.1 « Suggestion de mots-clés apparentés ») deux options de suggestion qui sont : la suggestion avec considération d'orthographe et la suggestion des mots qui sont en relation avec les mots-clés, nous expliquerons leur Implémentation par la suite :

1. La vérification d’orthographe

Un vérificateur d’orthographe permet de proposer une liste de mots similaires au mot mal orthographié. Nous choisirons l’implémentation basée sur la méthode n-gramme et la distance de Levenshtein. Nous allons expliquer l'incapacité du n-gram avec la langue arabe et comment la résoudre.

Le vérificateur d’orthographe construit un dictionnaire basé sur tous les mots indexés, il prend chaque mot et génère son propre n-gramme, mais sans tenir compte des diacritiques en raison des problèmes que nous montrons dans l'exemple suivant : c'est le mot « اَلْجِنَّةِ » en gram3 :

Figure 50 : Un exemple du n-gram (n=3) sans et avec diacritiques

La première chose que nous remarquons dans l'exemple précédent est la différence significative en nombre de grams entre les deux cas où il vaut 9 lors de la considération des signes diacritiques, pour un mot de petite taille, quel sera le résultat pour un mot de la taille de « **فَأَسْقَيْنَاكُمُوهُ**» ?  
La seconde observation est que plusieurs grams ne portent aucune valeur significative, ce sont celles qui précédent les signes diacritiques des lettres telles que « ـّـَة ».

Le recours à la négligence des diacritiques est la seule solution qui existe avant de trouver une nouvelle formule au n-gramme qui convient la langue arabe et distingue entre les lettres et les signes de diacritiques.

1. Mots-clés relatifs

Pour trouver des mots relatifs, nous devons nous appuyer sur l’une des ontologies du Coran, qui contient tous termes du Coran et les relations entre eux. La suggestion sera selon le dégrée de la relation.

Figure 51 : Suggestions du mot « جهنّم »

1. Traitement de résultats

Pour réaliser les fonctionnalités additionnelles qui incluent : la pagination, le tri, et le sur-lignage des mots recherchés (Highlight), les propriétés de filtrage et d’extension de la recherche, nous avons modifié notre modèle de sorte que les résultats passent par des opérations de traitement qui améliorent leur qualité et les prépare pour la transmission vers les interfaces :



Figure  : Le positionnement du traitement de résultats

Ces opérations sont les suivantes :



Figure 53 : Traitement des résultats

1. Pondération

Un score est attribué à chaque document selon une méthode de pondération, puisque ce moteur de recherche coranique n'est pas différent dans la façon d'évaluation des autres moteurs de recherches, nous allons utiliser TF\*IDF[[211]](#footnote-15) ou l'un de ses algorithmes développés.



Figure 54 : La pondération

1. Tri

La liste des documents résultants est ordonnée d'abord par score, puis par un critère déterminé par l'utilisateur. Parmi les critères recommandés il ya :

* + **l'ordre de mushaf** ;
  + **l’ordre de révélation**;
  + **l'ordre alphabétique** des champs textuels, et **l'ordre numérique** pour les champs avec des valeurs numériques.



Figure 55 : Le tri

1. Filtrage

Le filtrage se fait selon un critère donné, seuls les résultats correspondants sont conservés. Par exemple, par la sourate.



**Figure 56 : Le filtrage**

1. Extension

Ici, des informations supplémentaires sont liées avec chaque aya de résultats, notamment :

* le tafssīr de l’aya;
* la traduction de l’aya dans n'importe quelle langue ;
* Les abrogeants et les abrogés de l’aya ;
* les données audio de récitation de l’aya par réciteur ;
* … etc.



Figure 57 : Extension de la recherche

1. Pagination

A cette phase, on répartie les résultats à un flux de pages pour être envoyée à l'interface page par page, car il n'est pas nécessaire d'envoyer tout si l'utilisateur ne veut que les premiers résultats.



Figure 58 : Pagination

1. Highlight

Pour le highlight (le sur-lignage), nous n’allons pas nous appuyer sur la méthode simple basée sur la coloration des mots-clés qui coïncident exactement avec les textes des résultats uniquement, car nous utilisons les options de recherche comme la diacritique, les synonymes, les antonymes, les tuples et les dérivations qui se basent dans la recherche sur d’autres mots que ceux introduits.



Figure 59 : Highlight

La méthode que nous avons adoptée (existe dans des moteurs de recherche tels que Google) est basée sur la répétition d’un processus qui ressemble à celui de la recherche tout en conservant des indicateurs des positions (*voir la figure 60*) :

Figure 60 : Le processus de Highlight

1. L’importation des index

Pour accomplir les options spécifiques du Coran, à savoir :

* La recherche par le script othmani ;
* Les options coraniques structurelles ;
* Les options statistiques ;
* La recherche par des index spéciaux ;
* La recherche par sujet.

On doit invoque d’autres informations que le texte du Coran, le nom de la sourate et le numéro de l’aya que nous avons utilisées dans le modèle de base. Ces informations sont :

1. Les textes de l’aya :
   * Comprennent le texte de l’aya sous différentes formes tels que le script standard ou othmani. L’objectif est de rechercher avec différents scripts.
2. Les informations structurelles :
   * Comprennent les diverses fragmentations contenues dans le mushaf comme la fragmentation en ahzab et en ruku’. L’objectif est de faire une recherche en utilisant les options structurelles.
3. Les informations de la sourate :
   * Inclue tout ce qui est attribué à la sourate comme le nom, le lieu et la date de révélation et ainsi de suite. C’est une partie de la recherche qui utilise des index spéciaux.
4. Les informations statistiques :
   * Comprennent toutes les statistiques des sourates et ayas tels que le nombre de mots. L’objectif est de faire une recherche par l’utilisation des options statistiques.
5. La classification en sujets :
   * Contient la catégorisation des ayas en sujets d’une manière ramifiée ou arborescente. L’objectif est de faire une recherche par sujet.
6. L’exploration

L'exploration est considérée comme la partie la plus importante dans les moteurs de recherche web, mais cela est différent pour notre moteur du recherche, parce que la recherche va être essentiellement dans le Coran. Néanmoins, l’option de rechercher dans les documents qui contiennent du Coran reste requise.

Pour réaliser le Système d'exploration, nous avons proposé ce modèle de base qui nécessite une future étude approfondie.

Une personne établie la liste de sites à explorer. Des robots explorent cette liste à la recherche des pages contenants des ayas, puis les stockent après qu’ils les compressent dans des barils.

L’explorateur appuie sur un fichier de configuration qui contient la politique de son fonctionnement, comme les caractéristiques des pages ciblées, les options d’exploration, le maximum de ressources système, etc.

L’indexeur analyse syntaxiquement (parse) les pages à la recherche des ayas qui se distinguent par une balise html qui indique ainsi, et il existe deux moyens pour le faire :

1. La balise contient le texte de l’aya et son numéro : les inconvénients de cette approche réside dans la difficulté de savoir si le verset est distordue[[212]](#endnote-197) (محرّفة) ou non, c’est pour ça qu’on ajoute un module pour authentifier les ayas correctes ;
2. La balise contient le numéro de l’aya et le texte d'une façon dynamique en interrogeant l'index du moteur de recherche.

Quand on extrait le numéro d’aya de la page, l’URL de la page est ajoutée à la table urllists dans la ligne qui corresponde au numéro de l’aya.

Le Ranker a comme rôle de classer les pages récupérées par l’explorateur selon leur importance.

Le ranker attribue un rank à chaque page en fonction de certains paramètres, tels que le nombre de pages pointantes vers la page en question. Nous parlons ici de ranker indépendant de la requête.  
L’optimisateur fait l’opération de fusion sur les index inversés résultants de l'urllists au besoin ou à des intervalles spécifiques.



**Figure 61 : Le système d’exploration**

1. Conclusion

Dans ce chapitre, Nous avons décrit les étapes nécessaires pour développer le système qui sert à rechercher dans le Coran avec de nombreuses fonctionnalités.

Pour réaliser ces options, nous avons profité énormément des méthodes qui existent au niveau des moteurs de recherche standards. Nous avons vu, cependant, que ces méthodes ne peuvent pas être appliquées directement sur le Coran, ce qui nous a obligé d’apporter plusieurs modifications à certains d'entre eux, et d'ajouter d’autres.

Pour un bon fonctionnellement de chaque option de recherche, il est crucial de déterminer l'emplacement approprié du système où ont doit implémenter le traitement correspondant.

Dans le chapitre suivant nous allons expliquer comment réaliser les différents modules cités dans ce chapitre avec toutes les détailles techniques de programmations

# Chapitre VI

**Implémentation**

1. Introduction

Après avoir faire la conception de notre système, nous allons passer à la phase de réalisation. Le chapitre présent documente pour les détails techniques de programmation.

Dans le but de concevoir une interface de programmation API qui offre les fonctions de recherche et d’indexation des documents coraniques nous avons besoin de plusieurs outils. On a vu dans les chapitres précédents certains de ces outils qu’on peut exploiter, notamment, les index de mots, de synonymes et de sujets de M. Taha Zerrouki, les index du Corpus coranique, les index de Midād lbayān, et les versions XML du texte coranique et ses propriétés réalisées par le projet Tanzil.

Nous nous appuierons principalement sur l’index de mots pour réaliser les fonctions de segmentation (stemming), tandis que les autres index permettraient d'enrichir le moteur avec plus d’options. Pour le texte coranique, nous avons utilisé celui du projet Tanzil car c’est le plus authentique, les propriétés du Coran nous servent à réaliser les options structurelles.

Nous allons aborder dans ce chapitre les outils utilisés, le langage de programmation adopté, les problèmes rencontrés et les arguments de tout choix effectué. Ensuite, nous présenterons des captures d’écrans des interfaces d’utilisateur pour bien expliquer la manière de manipulation du système.

Enfin, nous terminerons par un certain nombre d’essais pour tester la fiabilité et la performance du système, et éliminer d’éventuels bugs.

1. Python – le langage choisi

Nous avons opté pour le langage de programmation Python en raison de ces caractéristiques qui sont commodes pour nous, Python est **[Python, 10]** :

* sous licence open source, ce qui le rend gratuit et librement utilisable et distribuable ;
* indépendant de toute plate forme, il fonctionne sur tous les principaux systèmes d'exploitation, notamment : Windows, Linux / Unix, OS / 2, Mac, Amiga. Il existe même des versions qui s'exécutent sur. NET, la machine virtuelle Java, et les téléphones cellulaires Nokia Séries 60. En faite, le même code source s’exécutera à travers les plateformes sans avoir besoin de le changer ;
* un langage qui support l’Unicode naturellement, en conséquence il n’y a pas besoin d'outils supplémentaires pour pouvoir manipuler la langue arabe ;
* il s’intègre facilement avec les serveurs web tels que ‘Apache’ et ‘IIS’ grâce à sa bibliothèque standard CGI (*Common Gateway Interface*).

Python a été conçu pour être un langage lisible. Il utilise des mots-clés anglais fréquents alors que d’autres langages utilisent la ponctuation, et possède également moins de constructions syntaxiques[[213]](#footnote-16) que de nombreux langages structurés tels que C, Perl ou Pascal.

Cette simplicité du langage algorithmique permet de passer facilement à d’autres langages de programmation.

1. Les ressources utilisées
2. Whoosh : un API d’indexation et de la recherche

Whoosh a été créé par Matt Chaput. Il a commencé comme un serveur rapide de recherche pour la documentation en ligne du progiciel « Houdini 3D animation ». ‘Side Effects Software’ a aimablement autorisé à Matt qu'il ouvre le code source dans le cas où il pourrait être utile à quelqu'un qui a besoin d'un moteur de recherche très flexible ou purement Python (ou les deux!).

La puissance principale de la conception de Whoosh est le fait qu'il est pur Python. Il devrait être possible d'utiliser Whoosh partout où on peut utiliser Python, pas de compilateur ou Java requis.

Comme l'un de ses ancêtres, Lucene[[214]](#footnote-17), Whoosh n'est pas vraiment un moteur de recherche, c'est une bibliothèque de programmation pour créer un moteur de recherche.

Pratiquement aucun comportement de Whoosh n’est drument codé. L’indexation du texte, le niveau des informations stockées pour chaque terme dans chaque domaine, l'analyse des requêtes de recherche, les types de requêtes autorisés, les algorithmes d’attribution du score, etc. sont tous personnalisables, remplaçable et extensible.

Il serait sûrement possible de construire un moteur de recherche prêt à l'emploi en se basant sur Whoosh, comme Nutch et Solr utilise Lucene.**[Whoosh, 10]**

* Whoosh est rapide, et n’utilise que de pur Python, il s’exécutera partout où Python s’exécute, sans avoir besoin d'un compilateur.
* Par défaut, Whoosh utilise la fonction de classement « Okapi BM25F », mais comme la plupart des choses, la fonction de classement peut être facilement personnalisée.
* Whoosh crée des index assez fiables par rapport à beaucoup d'autres bibliothèques de recherche.
* Touts les textes indexés dans Whoosh doivent être en Unicode.
* Whoosh permet de stocker arbitrairement des objets Python avec les documents indexés.

1. Tanzil – Un texte Coranique vérifié

On a choisis d’utiliser le texte coranique réalisé par le projet Tanzil car c’est le texte informatisé du Coran le plus authentique, il a passé par les phases de vérification suivantes :

* **Extraction automatique du texte :** à cette étape, une collection de textes du Coran provenant de diverses ressources authentiques est recueillie, puis un programme converties chaque texte dans un format canonique pour l’analyser et extraire un texte coranique de base.
* **Vérification à base de règles :** dans cette étape, un programme est développé afin de vérifier le texte coranique de base suivant un ensemble de règles de grammaire et de récitation. Le programme a été en mesure de produire des textes coraniques sous la forme imlāi et othmani.
* **Vérification** **manuel**: dans cette étape, les textes du Coran produits sont examinés manuellement avec le mushaf de Médine.

Le projet a utilisé l'aide d'un groupe d'experts comprenant des spécialistes du Coran, hafizes (حفظة) et savants (علماء), afin d'assurer de l'exactitude et de la précision des textes du Coran obtenus. Depuis la sortie de la première version du texte du Coran Tanzil, aucune faute de frappe n’a été trouvé, par des millions d'utilisateurs qui l’ont utilisé. La précision du texte a été également confirmée par plusieurs projets et applications[[215]](#footnote-18) qui utilisent le texte du Coran ‘Tanzil’.

De plus, il comprend une gamme riche de types de textes qui sont **[Tanzil, 10]**  :

* othmani ;
* othmani minimal ;
* simple amélioré ;
* simple minimal ;
* simple propre.

Le choix entre ces types se fait en fonction du but : par exemple, le texte en script othmani est adéquat pour afficher les pages du mushaf, par contre le simple propre est convenable pour la recherche facile. (Voir chapitre 4 : VI.D « Tanzil Quran Navigator »)

Ces textes sont tous encodés en format UTF-8 ce qui est convenable pour nous, car la bibliothèque Whoosh l’utilise aussi.

1. Les Index de M. Zerrouki

M. Zerrouki nous a fournit ces index dans un projet global autour de sciences coraniques. Ils sont originalement sous format de tables Excel pour faciliter leur lecture. Le chapitre 4 (VI.B) « Les Index par M. Taha Zerrouki » contient la structure de ces index et un bref aperçu.

Ils incluent 3 formes d’index, à savoir :

* **Un index des mots** : il contient pour chaque mot du Coran, son morphème, sa racine, son type (verbe, nom, …) et le mot en script othmani, ainsi que sa forme plurielle et féminine ou le nom propre s’il y a lieu.
* **Un index de sujets** : il est organisé hiérarchiquement en trois niveaux qui sont le thème, le sous-thème et la rubrique. L’emplacement dans le Coran est indiqué avec le numéro de la sourate et le numéro de l’aya de début et de fin.
* **Un index de des synonymes** : il contient pour chaque mot un ensemble de taille variable de synonymes (jusqu'à treize), plus le type du mot.

1. Le modèle de base

La première étape que nous prenons est de construire une fondation de base du moteur de recherche qui fournit la recherche simple. Pour ce faire, nous avons utilisé la bibliothèque Whoosh, qui est une API (interface de programmation) pour la recherche/indexation écrite en langage Python. Elle fournit un environnement de base pour le développement des moteurs de recherche.

Pour le texte du Coran, nous avons opté pour la version téléchargeable du projet Tanzil, parce que c'est le texte coranique informatisé le plus fiable, nous avons choisi la version en script standard (imlāi[[216]](#endnote-198)) qui ne contiennent ni de signes diacritiques ni de symboles supplémentaires.

Ce modèle de base doit être enrichi pour prendre en charge la vocalisation partielle qui consiste à mettre certains signes de diacritiques sans autres par l'utilisateur. La comparaison se fait en considérant seulement les diacritiques existantes, et pas celles négligées.

La méthode décrite dans la conception a été mise en œuvre sans problèmes, mais le problème se pose lorsqu'elle est reliée au moteur de recherche. Puisque en s’appuyant sur la bibliothèque Whoosh, il n’est pas possible de remplacer la comparaison standard des chaines de caractères sans changer dans le code de la bibliothèque, ce qui annule son indépendance du moteur et nécessite ainsi la répétition du processus de remplacement avec chaque version réalisée par Whoosh, spécialement qu’il est encore dans les versions expérimentales.

1. Traitement de texte

A la phase du traitement de texte, le texte coranique importé est analysé en vue de construire les index inversés. Ce traitement se compose des opérations suivantes :

1. Extraction des tokens (Tokenization)

La tokenization est l’opération de fractionnement du texte en unités indépendantes appelées tokens. Dans les textes littéraires, les tokens souvent représentent les mots.

Puisque le texte coranique que nous utilisons est libre de symboles spéciaux et chiffres, il nous suffit un extracteur de tokens (tokenizer) simple qui divise les termes selon les espaces contenue dans la phrase.

1. Normalisation

Il n'existe pas une définition unique pour la normalisation de texte en arabe, mais, elle consiste généralement à réduire toutes les variantes de caractères tels que les formes de hamzä, alif maqsûra et yâ’ sous une seule forme. **[Benajiba, 06]**

Nous avons ajouté une autre fonction au normalisateur qui supprime la lettre de tatwīl[[217]](#footnote-19) (ـ), et en cas du texte coranique en script othmani, on supprime aussi les lettres qu'elles ne peuvent pas être écrites avec le clavier comme l’álif khinjarïa (ــٰـ).

1. Segmentation (stemming)

La segmentation a pour but soit de regrouper les mots qui sont des variantes d'une même racine (allomorphes) **[AMROUCHE, 08]** ou bien d’éliminer les affixes du mot et ne garder que le stem, cette dernière s’appelle la segmentation légère (light stemming).

Dans la partie conception, IV.C Segmentation (stemming), nous avons proposé de se limiter à la segmentation légère, il existe deux méthodes pour atteindre cet objectif :

* Utiliser un analyseur morphologique qui enlève les affixes en s’appuyant sur des algorithmes particuliers, parmi les analyseurs morphologiques en code source libre, il y a Tashaphyne[[218]](#footnote-20) qui effectue des opérations linguistiques simples sur le mot arabe telles que la segmentation légère, mais son taux d’erreur actuel nous empêche de l’appliquer sur le texte coranique.
* Trouver des index de mots du Coran sans affixes établis manuellement par les chercheurs dans ce domaine. Il existe l’index coranique « Corpus » qui énumère les affixes selon ses caractéristiques et ses fonctions. Mais la complexité de cet index rend difficile le processus d’extraction des mots.

1. Filtrage des stopwords

Les stopwords sont les mots que l’on ignore lors de la recherche. Nous avons soulevé le problème de choix des stopwords concernant le Coran, et comment que la sensibilité du choix impose qu’il soit précis.

Nous avons examiné le projet « Arabic Stop words »**[arabicstopwords, 10]** , qui est un projet gratuit et en code source libre pour les stopwords dans la langue arabe.

La liste de mots dans le projet ci-dessus contient les mots qui peuvent être fréquentes dans la langue arabe contemporaine, mais rares dans le Coran, comme le mot « ثمانين ».

Pour résoudre ce problème, nous avons comptabilisé les mots-clés et considérer seulement les particules ensuite les trier en ordre croissant. On filtre les mots les plus fréquents et qui existent au niveau du projet « Arabic Stop words », parmi les mots que nous avons sélectionné nous citons : « مِن », « فِي » et « مَا ».

1. Le traitement de la requête

On utilisant cette bibliothèque, nous avons fourni le processus d'indexation et de recherche en plus des fonctionnalités de la recherche avancée :

|  |  |  |
| --- | --- | --- |
| opération | rôle | Syntaxe |
| terme simple | Son but est d’identifier le terme de base qui n'est pas soumis à des ajustements dans le parseur. C’est le résultat final de toutes les opérations en son sein. | Mot sans caractères spéciaux |
| relations logiques | Définir les opérations logiques entre les termes qui sont : ET, OU, et NON. | Terme opération Terme |
| champs | Spécifier la recherche pour l’un des termes à un champ donné. | Field : terme |
| Groupage | Grouper plusieurs termes dans le cas des opérations logiques. | (terme opération terme) opération terme |
| intervalle | Recherche par intervalles | [début TO fin] |
| jokers | Trouver les mots qui correspondent à l’expression régulière. | Mots qui contient \* ou ? |
| boosting | Contrôler l'importance de l’un des termes par rapport aux autres. | Terme ^ nombre |

Tableau 36 : Rôles et syntaxe des opérations existantes du parseur.

1. Synonymes et antonymes

La recherche par synonymes et antonymes consiste à trouver des mots synonymes ou antonymes du mot entré par l'utilisateur, puis faire la recherche avec les mots résultants en plus de ceux introduits.

Pour les synonymes, nous avons adopté un index par M. Zerrouki, nous avons modifié celui-ci pour qu’il devient sous la forme d'un dictionnaire où chaque mot correspond à ses synonymes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| المترادفات | | | | |
| السعير | النار | الجحيم | جهنم | سقر |

|  |  |
| --- | --- |
| الكلمة | مترادفاتها |
| السعير | ]السعير, النار, الجحيم, جهنم, سقر[ |
| النار | ]السعير, النار, الجحيم, جهنم, سقر[ |
| الجحيم | ]السعير, النار, الجحيم, جهنم, سقر[ |
| جهنم | ]السعير, النار, الجحيم, جهنم, سقر[ |
| سقر | ]السعير, النار, الجحيم, جهنم, سقر[ |

Figure 62 : Transformation d’index de synonymes

En ce qui concerne les antonymes, aucun index informatisé d’antonymes dans le Coran n’est disponible pour le moment.

1. Dérivations des mots

Dans la recherche par dérivation, l'utilisateur saisie l'origine du mot ou le mot lui-même en indiquant le niveau de dérivation, puis le moteur de recherche trouve les mots qui appartiennent à la même famille et au niveau de dérivation requis et effectue la recherche à l’aide de ces mots.

Les niveaux de dérivation que nous avons adoptés sont :

* le mot sans affixes ;
* le morphème (مصدر) ou les verbes à l’infinitif ;
* la racine.

Il existe deux approches de réalisation :

* La première est d’utiliser les analyseurs morphologiques[[219]](#endnote-199) qui s'appuient sur des algorithmes spéciaux pour séparer les affixes et extraire les origines du mot. Cependant, les analyseurs existants actuellement pour la langue arabe sont faibles dans l'extraction de tous les niveaux de la dérivation, surtout que nous traitons un texte sacré.
* La deuxième est d’utiliser les index coraniques des mots préparées manuellement par les chercheurs. L’avenage des index résident dans la possibilité de les contrôler et leurs résultats stables. Leurs inconvénients résident dans l'incapacité de trouver les dérivés de la famille si le mot n'est pas cité de la même forme dans le Coran. Par exemple :   
  Si on cherche le mot « وواعدناكم » au niveau de la racine, on va trouver tous les ayas qui contiennent ce mot, plus les mots de la même famille, comme « وعدناهم », « وواعدنا », etc. Mais si on cherche le mot « وعدهما » on va rien trouver car ce mots n’existe pas dans le Coran et donc dans les index.

La recherche par dérivation se déroule en deux phases : premièrement, déterminer le niveau de dérivation du mot, ensuite extraire les mots dérivés de la même famille que le mot introduit selon le niveau requis.

Nous avons adopté la méthode des index dans les deux phases en raison de l'inefficacité des analyseurs morphologiques, mais la première étape se basera à l'avenir sur les analyseurs dans le cas où le taux de réussite sera presque complet. Nous avons utilisé un index de mots du Coran de M. Zerrouki qui contient les racines de mots et leurs morphèmes, il fournie ainsi deux niveaux de dérivation.

1. La considération d'orthographe

Dans la langue arabe, il y a des lettres dont la forme se ressemble, mais certains utilisateurs ne font pas de distinction lorsqu’ils saisissent les requêtes, et les moteurs de recherche doivent en tenir compte. L’option de considération d’orthographe peut régler ce problème.

A chaque opération de recherche par négligence d’orthographe, la hamzä est convertie en une hamzä indépendante, tâ’ marbûta (ة) en hâ’ (ه) et alif maqsûra (ى) en yâ’ (ي).  
Ensuite, nous prenons tous les mots indexés et nous les traitons de la même façon et on compare.

La complexité de cet algorithme pour la recherche simple est de O(*n*) où *n* est le nombre de mots indexés, ce qui augmente significativement la lenteur du processus de recherche, nous devons donc abandonner cette option pour gagner de la performance et de la vitesse, surtout que son importance est limités en présence du système de suggestions.

1. La recherche par tuple :

La recherche par tuple s’appuie sur la présence des informations sur les propriétés du mot dont les valeurs divisent les mots du Coran en un nombre considérable de groupes, le moteur de recherche utilise l'intersection de ces groupes au lieu des mots introduits par l'utilisateur.

Parmi les caractéristiques que nous avons pu trouver ses index, la racine et le type. Nous les avons importés de l’index des mots de M. Zerrouki.

Avec la disponibilité d’une propriété la cherche s’effectue avec un couple de type :

{racine: ملك, type: إسم)}.  
Pour le pattern[[220]](#endnote-200) (الوزن), il n'y a pas d’index, que ce soit avec ou sans vocalisation.

**Figure 63 : Exemple de recherche par tuple**

1. Les suggestions

Les moteurs de recherche suggèrent des fois des mots plus fréquents et proches de ceux introduits dans la requête. Il existe deux types de suggestions qui sont :

1. La vérification d’orthographe

La vérification d’orthographe est invoquée lorsque la recherche avec la requête originale n’aboutie à aucun résultat. Dans ce cas, le système propose à l’utilisateur les mots les plus similaires qui sont correctement écrits.

Nous avons utilisé Whoosh pour construire deux dictionnaires n-gramme, le premier pour les mots coraniques en se basant sur les mots issus du processus de normalisation et de tokenization du texte coranique. Le second est pour les sujets en se basant sur tous les mots inclus dans les champs de la subdivision thématique.

La vocalisation n'est pas pris en compte car elle alourdie les index en vain. Pour plus de détails, voir le chapitre conception V.A.

1. **Mots-clés relatives**

Contrairement à la vérification d’orthographe, les mots-clés relatives sont proposés même si la requête aboutie à des résultats, quand il y a des mots souvent associés ou sémantiquement proches des mots le la requête.

Afin de réaliser cette fonctionnalité, nous utilisons l’ontologie du Coran, et parmi les ontologies que nous avons trouvées il y a :

* L’ontologie de Hadj henni : une ontologie du domaine des documents coraniques généré à partir d’un schéma XML.**[Hadj henni, 08]**   
  Elle n'est pas utilisable ici, car elle se soucie surtout des informations concernant le mushaf.
* L’ontologie du corpus coranique : utilise la représentation des connaissances (contenus dans les sources de hadith et de tafsīr) pour définir les concepts clés du Coran et les relations eux en utilisant la logique des prédicats.**[Corpu, 10]**   
  Elle n'aide pas beaucoup parce qu’elle est basée sur les termes anglais.

1. Le traitement de résultats
2. Pondération

La pondération des termes est une mesure statistique qui peut être modifiée par différents critères (mots clés, domaines d’information, ...). En fait, le principe de la pondération est simple, il est basé sur une hypothèse selon laquelle « lorsqu’un auteur écrit un texte, il répète certains termes pour développer un aspect du sujet » **[Luhn, 58]** .

La méthode de pondération dans le texte coranique ne diffère pas de celles de tout autre texte, donc nous allons choisir l’un des algorithmes de pointe, qui est : BM25F.

BM25F est l’un des modèles les plus connus en recherche d’information, et un des plus anciens également, a été le système, ou plutôt les systèmes Okapi. Le schéma de pondération BM25 (BM pour « Best Match ») a été développé par Robertson et al. **[Robertson, 95]**

Ce choix est opté pour les raisons suivantes :

* la complexité apparente de la formule (par rapport au modèle vectoriel) et sa grande efficacité (ce modèle est l’un des plus performants actuellement) montre bien les avantages de la modélisation probabiliste ; **[Karbasi, 07]**
* ce modèle est destiné à la recherche multi-champs ;
* ce modèle est activé par défaut sur la bibliothèque Whoosh.

1. Tri

Le tri fonctionne selon des critères spécifiés par l'utilisateur, nous avons proposé plusieurs critères dans la conception qui sont les suivantes :

* + **l'ordre de mushaf** :

Pour réaliser ce tri, nous ajoutons un champ de numéro de la sourate à l’index de document, il sera ainsi selon le couple (numéro de la sourate, numéro de l’aya).

* + **l’ordre de révélation** :

Pour réaliser ce tri, nous ajoutons un champ de l’ordre de révélation pour chaque sourate, il sera ainsi selon le couple (numéro\_de\_la\_sourate, numéro\_de\_l’aya).

* + **l'ordre alphabétique** des champs textuels, et **l'ordre numérique** pour les champs avec des valeurs numériques.

1. Filtrage

Le filtrage se passe selon un critère donné, seuls les résultats correspondants sont conservés.

Parmi ces critères :

* **Le filtrage selon le sujet :**

Pour se faire, nous nous appuyons sur les index de sujets du Coran. Voir importation des index – la classification en sujets.

* **Le filtrage par options structurelles :**

Pour se faire, nous nous appuyons sur les index structurels du Coran. Voir importation des index – les informations structurelles.

1. Extension

L’extension la recherche est le fait de relier des informations supplémentaires avec chaque aya de résultats. Nous avons proposé certaines informations supplémentaires dans la conception, à savoir :

* **Les tafassīr et les traductions :**Pour les traductions et les tafassīr, il y a deux modèles qui sont disponibles sur l'Internet :
  1. Le modèle d’Arabeyes : basé sur le langage XML par l’équipe Arabeyes[[221]](#footnote-21). Plusieurs traductions sont disponibles en utilisant ce modèle, mais manquent de fiabilité et ne sont pas unifiées par un site officiel pour les réviser au moins superficiellement.
  2. Le modèle de Zekr : Ce modèle est structuré en deux fichiers, l’un contient les traductions (ou les tafassīr) aya par aya et l’autre fichier pour définir la traduction et son auteur. Ce modèle a été établi par les développeurs du programme « Zekr pour le saint Coran » et a adopté un grand nombre de traducteurs de différentes langues du monde.  
     Ces traductions sont rassemblées et publiées sur le site officiel de Zekr[[222]](#footnote-22), mais restent sans révision et la responsabilité est assumée par sur celui qui les télécharge.
* **Les récitations de l’aya :**Il existe sur le net beaucoup de récitations par différents réciteurs, qui sont disponibles dans des logiciels ou libre sous forme d’un fichier audio. Ils sont souvent une récitation d’une sura entière ou d’une partie, et non pas de chaque aya séparément.  
  Le projet VerseByVerse[[223]](#footnote-23) est un projet qui est conçu pour donner un accès facile et gratuit aux récitations du Coran en format MP3 séparés en ayas individuelles, il est utilisable par les développeurs de sites web et de logiciels et fournit des modèles pour les différents recteurs et en haute qualité et peuvent être utilisés et analysées sans téléchargement « à la demande » et donc nous débarrassons de la grande taille qui peuvent occuper et qui peut aller jusqu'à plusieurs gigaoctet.

Nous avons gardé la phase d’extension de la recherche comme une perspective pour le future proche afin de garder notre concentration sur la recherche dans le Coran.

1. Pagination

Malgré que la pagination est un processus simple, elle donne un grand avantage, surtout dans le Web, où la performance du serveur n’est pas consommée pour télécharger tous les résultats, même ceux que l'utilisateur ne vas pas les consulter.  
Pour réaliser cet objectif, nous avons adopté les fonctions de type ‘générateur’ fourni par le langage Python.

Un générateur est tout simplement, une fonction qui peut arrêter tous ce qu'elle est entrain de faire en un point arbitraire de son corps, et retourner une valeur à l'appelant, et reprendre, plus tard, du point où elle été ‘gelé’ de procéder comme si rien ne s'était passé. **[PramodeC.E, 04]**

Le nombre de résultats par page est déterminé par l’utilisateur.

1. Highlight

Le highlight (sur-lignage) consiste à distinguer des mots-clés dans le texte, la méthode que nous avons adoptée est présentée dans la conception (VII.F Highlight).  
Les langages que nous choisissons pour implémenter le sur-lignage sont :

1. HTML :

Le langage HTML est le plus employé pour réaliser le sur-lignage, il est utilisable sur le Web et les documents Office. Nous avons développé le sur-lignage par HTML pour corriger le problème de la dépendance du texte surlignée avec le fichier CSS (feuilles de style en cascade).

Figure : Highlight avec une balise HTML (mot-clé en jaune)

1. BBcode :

Le BBcode est utilisé par la plupart des forums notamment vBulletin[[224]](#footnote-24), sachant que les forums sont considérés comme les sites les plus actifs dans le monde arabe.

**Figure 65 : Hhighlight avec des balises BBcode (mot-clé en rouge)**

1. L’importation des index

Les index coraniques constituent un élément important de notre moteur de recherche et la principale source de données qui permet de réaliser un grand nombre d’avantages.  
Parmi les index que nous avons importés :

1. Les textes de l’aya :

Constitués de plusieurs textes du Coran sous différentes formats, nous les importons du projet Tanzīl et nous avons choisis :

* simple propre : texte coranique non vocalisé ;
* simple vocalisé : inclus les signes de diacritiques, mais pas le script othmani ni les marques du waqf ;
* othmani : le texte coranique en scripte othmani (tel qu’il est imprimé sur les mushafs) ;

1. Les informations structurelles :

Ce sont les informations relatives aux différentes fragmentations du mushaf comme la division en juz’, rub’ …, ou les devisions spéciales comme en ruku’, manāzil, etc.

Nous les avons importés à partir du projet Tanzīl qui inclut les fragmentations suivantes :

* **rub’ (ربع) :** c’est l’un quart du hizb, le Coran contient 60 hizbs, soit 240 rub’ de tailles approximativement égales ;
* **page :** la fragmentation des mushafs en pages est unique dans la plupart des mushafs d’aujourd’hui ;
* **ruqu’ (ركوع) :** c’est une réparation des sourates selon les sujets abordés, elle est utilisée en particulier chez les pays de semi continent indien ;
* **manzil (منزل) :** le mushaf est divisé en sept parties appelées manāzil, cette fragmentation est utilisée au Pakistan.

1. Les informations de la sourate :

Ce sont les informations relatives à chaque sourate du mushaf telles que le nom et l’ordre de révélation.

Nous les importons du projet Tanzīl avec l'arabisation des informations qui y sont contenues, elles incluent :

* le nom de la sourate ;
* le lieu de révélation (Mecquoiseمكّيّة ou Médinoise مدنيّة) ;
* l’ordre de révélation.

|  |  |  |
| --- | --- | --- |
| **nom de la sourate** | **lieu de révélation** | **l’ordre de révélation** |
| الفاتحة | Meccan | 5 |
| البقرة | Medinan | 87 |
| آل عمران | Medinan | 89 |
| النساء | Medinan | 92 |
| المائدة | Medinan | 112 |
| الأنعام | Meccan | 55 |
| الأعراف | Meccan | 39 |
| الأنفال | Medinan | 88 |
| التوبة | Medinan | 113 |

Figure 66 : Appercu des informaions de la sourate

1. Les informations statistiques :

Ces informations inclues les calcules statistiques effectués sur le Coran ou une partie du Coran, comme le nombre de lettres ou de mots dans une sourate donnée. Ils ont une grande importance pour vérifier les miracles numériques du Coran, elles peuvent aussi jouer un rôle dans l’authentification des textes coraniques **[Zerrouki, 05]** .

Il n'y a actuellement aucun index statistiques informatisé fiables et la plupart d'entre eux manquent de la source et n’expliquent pas les détails des fondations des calculs.

Par conséquent, nous comptons sur la génération automatique des statistiques simples qui sont :

* le nombre de caractères dans la sourate et l’aya : le calcul s’appuis sur le texte othmani après la suppression des symboles supplémentaire et des diacritiques ;
* le nombre de mots dans la sourate et l’aya : calculé par la tokenization basée sur les espaces ;
* le nombre de noms d’Allah [[225]](#endnote-201) (ألفاظ الجلالة) dans la sourate et l’aya : calculé par l’expression « الله » ;
* le nombre de ruku’ dans la sourate : pris du projet Tanzīl ;
* le nombre de versets dans la sourate : pris du projet Tanzīl.

Le processus d'extraction des statistiques doit être soutenu par des index manuels vérifiés soigneusement.

1. La classification en sujets :

Elle consiste à classer les ayas du Coran en sujets, pour chaque aya on fait correspondre un ou plusieurs sujet, cette classification est définie d’une maniéré arborescente, toutefois une aya peut appartenir directement à un sujet du premier niveau.

Nous les avons importés à partir des index de sujets de Zerrouki qui contient la répartition arborescente sur trois niveaux :

* **le thème :** il divise le Coran en 22 thèmes principales, comme la religion, les droits de l’homme et les relations sociales ;
* **le sous-thème :** chaque thème est divisé en sous-thèmes, par exemple le thème des relations sociales est divisé en sous-thèmes comme l’industrie et l’héritage ;
* **la rubrique**: chaque sous-thème est divisé en rubriques qui sont plus précis que le sous-thème, par exemple le sous-thème de l’industrie est devisé en rubriques comme la construction navale et l'industrie des barrages.

Figure 67 : Aperçu de la classification en sujets

1. Les Interfaces
2. Interface de programmation (API)

L’interface de programmation constitue notre objectif principal, elle est organisée selon le concept de classes. Les développeurs des applications coraniques peuvent l’utiliser pour la recherche plutôt que de redévelopper de nouveau, ici le rôle de Python se met en évidence comme un langage supporté par la plupart des langages de la programmation.

L’API est emballée en un fichier EGG de python qui est un format de fichier unique pour la distribution des projets liés à Python.

« Les EGGs sont à Pythons comme les JARs sont à Java ... », les EGG sont, en fait, plus riches que les JARs, ils détiennent des métadonnées intéressantes telles que les détails de licence et les dépendances.

Les principaux avantages de Python EGG sont **[PEAK, 10]** :

* les fichiers EGG sont de format à "installation zéro" c.à.d. qu’aucun d’étape d’installation n’est nécessaire, il suffit seulement de les mettre dans PYTHONPATH ou sys.path ;
* ils peuvent inclure des métadonnées, tels que les autres EGG desquels ils dépendent ;
* ils permettent aux "packages namespace" (les paquets qui ne contiennent que d'autres paquets) d’être divisés en distributions distinctes, cela permet aux packages énormes d’être distribués en tant qu'éléments distincts ;
* ils permettent aux applications de spécifier la version requise d'une bibliothèque, de sorte qu’on peut par exemple requière ("Twisted-Internet> = 2.0") avant de faire une importation de twisted.internet ;
* c’est un format génial pour distribuer des extensions ou des plugins pour les applications extensibles.

Les figures ci-dessous sont des exemples simples d’utilisation de l’API :

* Importation et appelle :

>>> from qse.main import QuranicSearchEngine

>>> QSE = QuranicSearchEngine()

<\_\_main\_\_.QuranicSearchEngine instance at 0xb77d168c>

>>> QSE.OK

True

Figure : Importation et appelle  de l'interface API

* Trouver les résultats d’une requête :

>>> results,\_ =QSE.search\_all(u"الحمد")

>>> for r in results:

>>> print r["aya"]

الحمد لله رب العالمين

Figure : Exemple de résultats par l'interface API

* Voir plus d’informations sur les résultats : le nom de la sourate, le numéro de l’aya et le texte vocalisé de l’aya :

>>> for r in results:

>>> print r["sura\_name"],":",r["aya\_id"],":",r["aya\_"]

الفاتحة:2:الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ

**Figure 70 : Plus d’informations sur les résultats de l’API**

* Trouver les termes utilisés pour la recherche à partir de la requête :

>>> \_,terms =QSE.search\_all(u"الحمد")

>>> print terms

[("aya",u"الحمد")]

>>> print ",".join([term[1] for term in list(terms)])

الحمد

Figure : Exemple d'utilisation de l'interface API (retrouver les termes utilisés)

* Surligner des termes dans le texte de l’aya :

>>> Qhighlight(u"الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ", [u"الحمد", u"لله"])

<span style="color:red."><b>الْحَمْدُ</b></span>

<span style="color:green;"><b>لِلَّـهِ</b></span> رَبِّ الْعَالَمِينَ

Figure : Exemple d'utilisation de l'interface API (sur-lignage)

* Avoir le temps écoulé pour effectuer une recherche ainsi que le nombre de résultats :

>>> print "time=%f,number=%d" % (results.runtime,len(results))

time=0.100318,number=160

Figure : Exemple d'utilisation de l'interface API (avoir le temps écoulé et le nombre de résultats)

* Trier les résultats par un ordre donné : score, mushaf, tanzil, ou n’importe quel autre champ :

>>> results,\_ =QSE.search\_all(u"الله",sortby="mushaf")

Figure : Exemple d'utilisation de l'interface API (Trier les résultats)

* Suggérer les mots mal orthographiés dans la requête :

>>> for key, value in QSE.suggest\_all(u"ابراهام ").items():

>>> print key, ":", ",".join(value)

ابراهام : إبراهيم

Figure 75 : Exemple de suggestion pour une faute d’orthographe

|  |  |
| --- | --- |
| Recherche simple | سماكم |
| Dérivation niveau 1 (stem) | <سماكم |
| Dérivation niveau 2 (racine) | <<سماكم |
| Joker ? | ال؟لك |
| Joker \* | مالك\* |
| Phrase | "رب العالمين" |
| Relation logique ET | الصلاة و الزكاة |
| Relation logique OU | الصلاة أو الزكاة |
| Relation logique ET NON | الصلاة وليس الزكاة |
| Synonymes | ~السعير |
| Antonymes | #السعير |
| Par champ | سورة:الفاتحة |
| Par intervalle | رقم\_السورة:[112 إلى 114[ |
| propretés du mot | }ملك,فعل{ |
| Tableau 37 : Exemples des fonctionnalités de l’API | |

L’utilisateur peut utiliser des requêtes avancées pour trouver exactement le résultat qu’il veut, voir le tableau ci-dessus.

**Figure 76 : Schéma des modules**

L’API comporte d’autres extensions facultatives qui ne sont pas terminées, mais sont en cours de réalisation, parmi lesquelles :

* système d'importation ;
* système d’exploration ;
* une API pour utiliser le projet Tanzil ;
* une API pour utiliser le projet AyaByAya ;
* une API pour lire les modèles de traduction du projet Zekr

1. QT Desktop GUI

C’est une interface destinée à l’utilisateur final pour qu’il l’utilise sur son bureau sans avoir besoin de se connecter à l'Internet, elle est faite de QT afin d’être multiplateforme supportée par la plupart des systèmes d’exploitation comme Windows, Linux et Mac, en plus des appareilles mobiles de type Nokia.

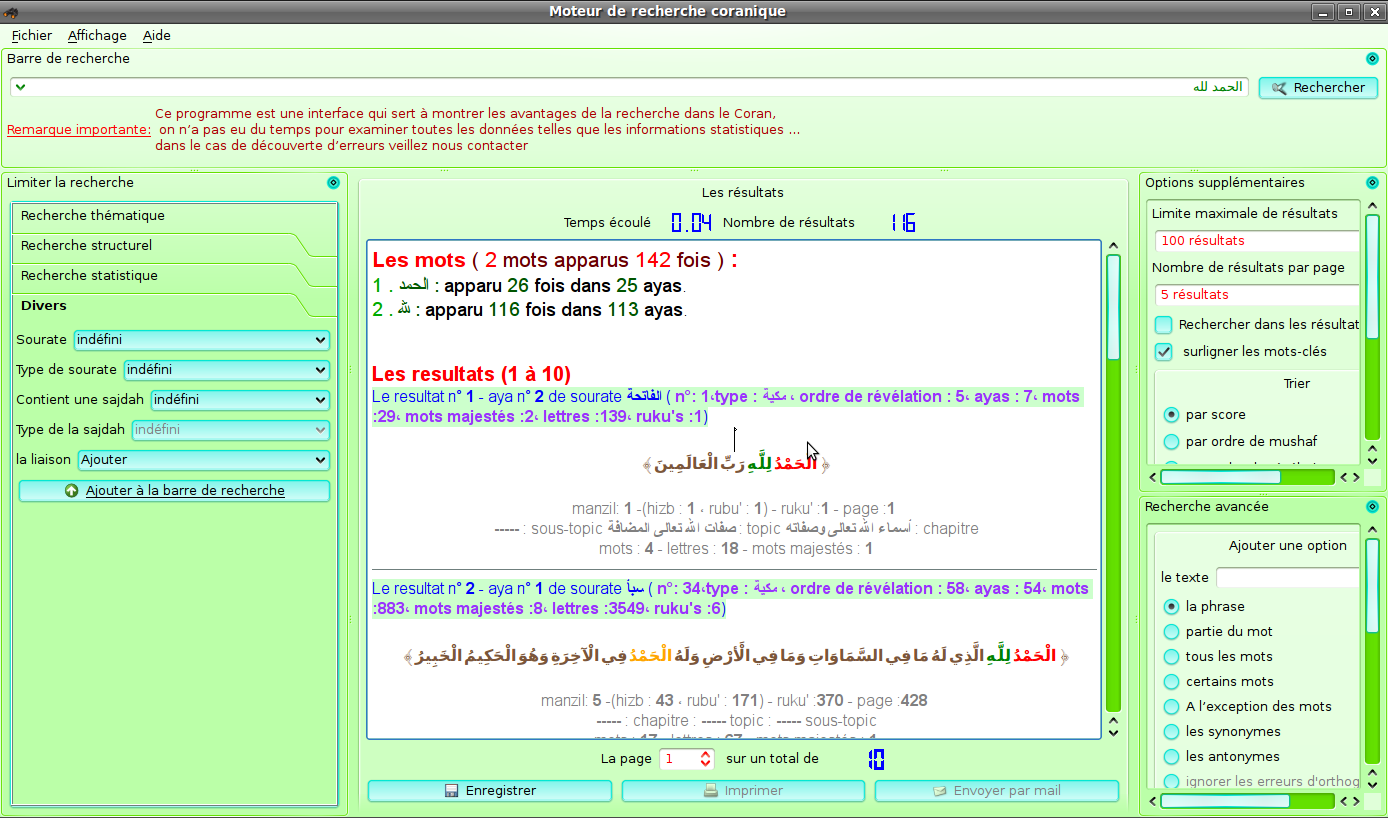


Figure : Aperçu de l’interface sous Windows (mots-clés : الحمد لله)

L’interface se compose principalement de quatre parties, deux principales et deux secondaires :

1. Barre de recherche

Constitue l’emplacement où on entre la requête et l’unité d'entrée est principale, elle accepte les requêtes simples et complexes.



**Figure 78 : Barre de recherche**

1. Panneau des résultats

Où on affiche les résultats et diverses informations sur le processus de la recherche en cours, c’est l’unité de sortie principale, les informations affichées sont constitués de :

Divisé en différentes informations présentées à:

* **Les informations sur l’opération de la recherche :** Inclure le temps écoulé, le nombre de résultats, la page en cours et le nombre total des pages



**Figure 79 : Barre d’information sur les résultats**

* **Suggestions :** Suggestions sont pour les mots qui n’ont pas été trouvés, par exemple :

**مءصدة :** مؤصدة،مقتصدة.

**ابراهام :** إبراهيم،وإبراهيم.

* **Statistiques des mots :**  
   Inclure le nombre d’occurrences d’un mot-clé et le nombre d’ayas où ils sont cités. Le « mot-clé » ne réfère pas au mot introduit par l'utilisateur, mais aux mots qui sont réellement utilisés lors de la recherche, par exemple, si l'utilisateur entre ‘les dérivations du mot قال », les statistiques incluront tous les dérivés du mot « قال » et pas le mot « قال » seulement, exemple :

**الكلمات** ( 2 كلمة وردت 142 مرة )**:**

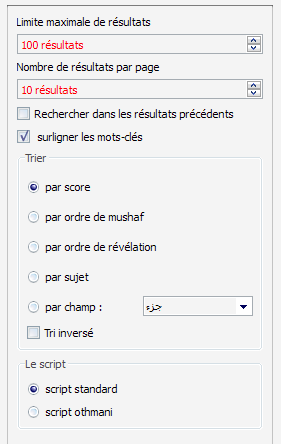
1 . الحمد : وردت 26 مرة في 25 آية.

2 . لله : وردت 116 مرة في 113 آية.

* **Les ayas résultant :**  
   Comprend une liste d’ayas qui correspondent à la requête. Chaque résultat comporte les informations suivantes :

Tableau 38 : Les informations sur les ayas résultant

|  |  |
| --- | --- |
| **Information** | **Exemple** |
| Identificateur de résultat | النتيجة رقم  **1** - الآية رقم **2** من سورة **الفاتحة** |
| Traduction utilisé | **الترجمة (Mohammed Marmaduke William Pickthall-English):**  **Praise be to Allah, Lord of the Worlds,** |
| Informations sur la sourate | ( الرقم: 1،مكان النزول : مكية ، ترتيب النزول : 5، الآيات : 7، الكلمات :29، ألفاظ الجلالة :2، الأحرف :139، الركوعات : 1 ) |
| Le texte de l’aya | [**الْحَمْدُ لِلَّـهِ رَبِّ الْعَالَمِينَ**] |
| Informations structurelles | المنزل: 1 -(الحزب : 1 ، الربع : 1) - الركوع :1 - الصفحة : 1 |
| Subdivision thématique | الفصل : **أسماء الله تعالى وصفاته** الفرع : **صفات الله تعالى المضافة** الباب : **---** |
| Informations statistiques | عدد الكلمات : 4 - عدد الأحرف : 18 - عدد ألفاظ الجلالة : 1 |

1. Panneau d'options

Contient les différentes options de la recherche et non les fonctionnalités, y compris :

1. Le nombre maximal de résultats et le nombre de résultats par page ;
2. filtrage des résultats précédents ;
3. l’ordre des résultats (par score, selon le mushaf, selon la révélation, selon un champ textuel ou numérique) ;
4. script utilisé: othmani ou standard ;
5. traduction associées.

Figure 80 : Panneau d'options

1. Panneau de fonctionnalités

Les fonctionnalités sont entées grâce à l'adoption de règles dans la barre de recherche (*voir le tableau ci-dessous*), mais pour faciliter la tâche et le processus d'apprentissage, nous avons créé ce panneau. L’idée est basée sur la composition de la requête visuellement ensuite l’ajouter à la barre de recherche selon l'une des politiques suivantes en considérant A la requête dans la barre de recherche et B la composante :

|  |  |
| --- | --- |
| **résultat** | **Politique d’ajout** |
| A B | Concaténation direct |
| B | Remplacement |
| A et B | Conjonction ET |
| A ou B | Disjonction OU |
| A et non B | Exclusion ANDNOT |

Tableau 39 : Politiques d’ajout des mots dans la requête

Le panneau contient plusieurs onglets, chacune comporte un ensemble de fonctionnalités :

|  |  |  |
| --- | --- | --- |
| ***Onglet de la recherche avancée*** | ***Onglet de la recherche avec les propriétés de mot*** | ***Onglet de la recherche structurelle*** |
| ***Onglet de la recherche statistique*** | ***Onglet des divertissements*** | ***Onglet de recherche thématique*** |

Figure  : Panneau de fonctionnalités

1. Autres interfaces possibles

L’API permet de créer des interfaces multiples dans différents types d'ordinateurs et d’environnements. Les points suivants sont des exemples des différentes interfaces possibles

1. Interface web

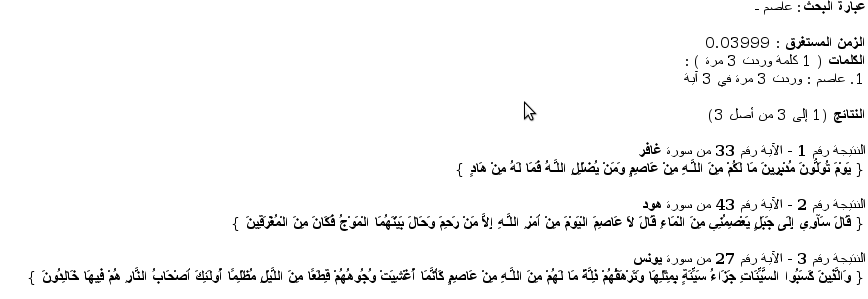
L’interface web est la plus populaire entre les interfaces des moteurs de recherche.  
Elle est semblable à celle de Google comme une page principale simple et une page pour la recherche avancée. Toutes les fonctionnalités disponibles dans l'interface de bureau sont aussi disponibles dans l'interface web, en plus de la récitation qui se base sur le Flash Player et le projet VerseByVerse. L’interface web fournit les résultats sous plusieurs formats possibles, notamment l’html pour la présentation, le BBcode pour les utilisateurs des forums et Wiki pour les éditeurs des encyclopédies libres. La traductibilité permet de personnaliser l’interface selon le pays et la langue de l’utilisateur.

1. Agent conversationnel (Chat bot)

L’agent conversationnel est un programme intelligent qui peut répondre aux questions des utilisateurs 24/24 ces repenses sont le plus souvent simples et précises. Python nous offre un ensemble complet d’APIs des programmes de chat les plus connus, comme Gtalk, Yahoo et Jabber ainsi que les réseaux sociaux tels que Facebook et Twitter.

Dans le cas de notre projet, l’agent conversationnel peut constituer un outil de présentation et fournit des opérations de recherche simples avec la possibilité de développement en cas de succès, car c’est bien connu que les réseaux sociaux et réseaux de messagerie instantanée (chat) sont actuellement le moyen le plus puissant pour diffuser des nouvelles et des publicités.

L'exemple suivant est la réponse de l’agent à une interrogation du mot-clé **« عاصم » :**

****

**Figure 82 : Exemple d’une réponse du rebot**

1. Barre de recherche pour les navigateurs

Avec le développement des navigateurs web à code source libre, notamment Firefox, la tache de création des extensions est devenue facile et organisée. Une barre de recherche peut être faite pour être un raccourci vers l'interface web, et un outil de présentation et de publicité pour le projet.

**J:\releases-alfanous\screenshots\toolbar.png**

**Figure 83 : Exemple d’une barre de recherche pour Firefox**

1. Interface pour les appareils mobiles

Après l’entrée du Python sur les appareils mobiles et exactement le système Symbian (sur lequel les appareils Nokia sont basés), les appareils mobiles ont devenus un terrain fertile pour adopter notre projet. C’est un peut similaire à l'interface de l’agent conversationnel mais avec plus d'options et de souplesse grâce une interface visuelle simple.

Nous n'avons pas pu faire une interface expérimentale pour des raisons de contraintes de temps et la difficulté de préparer un émulateur sous Linux, qui est la plateforme de développement de notre projet.

1. Conclusion

Dans ce chapitre, nous avons présenté les détails de l’implémentation du moteur de recherche et les outils utilisés et les problèmes rencontrés au cours de la réalisation. Nous avons également vu comment manipuler le système à travers ses deux interfaces : API pour les développeurs et QT pour les utilisateurs finaux.

Grace à la grande flexibilité des outils choisis, le système est capable de fonctionner sous divers plateformes et en multi-langage. En outre, le fait qu’il est en code source libre lui permet d’évoluer et d’acquérir de nouvelles options de recherche et de nouvelles formes d’interfaces

**Conclusion Générale**

Conclusion Générale

Le problème que nous avons traité consiste à développer un moteur de recherche coranique standard. Ce système doit être réutilisable par d’autres applications.

De nombreuses applications coraniques existent déjà, mais leur indépendance nécessite que chaque application reconstruise ces services depuis le début. Cela conduit à une perte considérable en termes du temps et d'efforts, en plus du risque accru de commettre des erreurs. Alors que ce domaine, en particulier, exige de l’exactitude.

Dans le cadre de notre projet, nous avons proposé un nouveau moteur de recherche qui s’appuie sur les index coraniques que nous avons collecté. Le moteur de recherche conçu doit couvrir les besoins de la recherche dans le Coran. En plus de la recherche de base il offre plusieurs fonctionnalités, notamment :

* la recherche avancée par :
  + champs,
  + partie du mot,
  + jokers (expressions réguliers),
  + opérations logiques,
  + phrase,
  + mots proches,
  + intervalles,
  + boosting ;
* les fonctionnalités additionnelles :
  + suggestion de mots-clés alternatifs et apparentés comme la correction automatique des erreurs d’orthographe,
  + la pagination des résultats sur plusieurs pages,
  + le tri des résultats selon le choix de l’utilisateur,
  + le sur-lignage des mots recherchés dans les résultats ;
* les options linguistiques telles que :
  + la recherche par synonymes et antonymes,
  + la considération de diacritiques,
  + la recherche par la racine du mot arabe,
  + la conjugaison des verbes,
  + la déclinaison des noms,
  + la recherche par les natures des mots arabes,
  + la possibilité de respecter les signes de diacritiques (الحركات),
  + la possibilité de respecter les formes de la hamza,
  + la possibilité de respecter la différence entre Hâ’ et tâ’ marbûtä, entre Yâ’ et Alif maqsûrä ;
* plus d’autres options de recherche dans le Coran :
  + la recherche par scripte othmani,
  + la possibilité d’exploiter les divers index spéciaux pour ajouter de nouvelles options, telles que la recherche par sujet,
  + la recherche par les propriétés structurelles du Coran vu qu’il est fragmenter en hizb, juz’, ruku’, etc.,
  + fournir les options statistiques du coran comme le calcule d nombre de mots et de lettres ;
  + extension de recherche pour inclure également d’autres ressources comme les récitations et les traductions.

De manière globale, notre système peut être considéré comme une solution de la problématique initiale, car il aboutit aux objectifs fixés. Néanmoins, plusieurs améliorations peuvent être établies pour le complémenter d’avantage. Parmi ces perspectives, nous citons :

* l’amélioration du traitement linguistique de texte
  + la segmentation légère qui conserve mieux le sens des mots ;
  + la détection automatique du niveau de dérivation d’un mot arabe ;
* l’exploration des sites web coraniques
  + l’algorithme d’authentification d’ayas
  + la politique d’exploration et de correspondance
* amélioration du system de suggestion
  + personnaliser la distance de Levenshtein pour s'adapter avec la longue arabe;
  + l’utilisation d’une ontologie de mots du Coran pour la suggestion des mots relatifs ;
* amélioration des calculs statistiques
  + fournir les nombres de la présence des racines des origines du mot arabe ;
  + fournir des statistiques selon les propriétés du mot comme le type et le pattern (الوزن) ;
  + renforcer le système par des index manuels révisés ;
* l’extension de la recherche
  + localiser des ressources utilisables de tafssīr ;
  + l’index d’antonymes ;
  + l’index de patterns.
* Localisation de touts les index et traductions vers la langue anglaise.

Finalement, nous espérons que notre système réponde aux besoins des utilisateurs et fonctionne sur tous les interfaces, API pour les programmeurs et Web et QT pour les utilisateurs finaux

Bibliographie :

|  |  |
| --- | --- |
| **[ABAR, 09]** | ABAR Ali*, La capitalisation du profil des experts du CENEAP sur la base de la classification des études*: memoire d’ingénieur ESI, 2009. |
| **[AFNOR, 89]** | AFNOR *, AFNOR : Vocabulaire de la documentation 2ème édition NF Z 47-100*, Paris, 1989. |
| **[al-bawwab, 09]** | al-bawwab merouane*, محركات البحث في النصوص العربية وصفحات الإنترنت.*, damas: مجمع اللغة العربية, 2009. |
| **[Alchalabi, 05]** | Alchalabi Riyad*, Pattern-based Stemmer for Finding Arabic Roots*, Irbid, Jordan: Yarmouk University, 2005. |
| **[AlKharashi, 99]** | AlKharashi Ibrahim*, A Web Search Engine for Indexing, Searching and Publishing Arabic Bibliographic Databases*: King Abdulaziz City for Science and Technology, 1999. |
| **[ALLAB, 08]** | ALLAB KAMEL*, Classification Automatique de Documents XML*, 2008. |
| **[Al-Madina, 00]** | Al-Madina *, Le logiciel « mushaf Al-Madina » pour les publications informatiques (مصحف المدينة النّبويّة للنّشر الحاسوبي) v. 1.0*: Site web officiel : http://www.qurancomplex.org, 2000. |
| **[AMROUCHE, 08]** | AMROUCHE Karima*, Passage à l’échelle en Recherche d’Information : Méthode d’élagage pour la réduction de l’espace de recherche*, 2008. |
| **[arabicstopwords, 10]** | *Arabic Stop words*, URL: http://sourceforge.net/projects/arabicstopwords, visité en 2010. |
| **[Araby, 09]** | *Araby*, visité en 2009. |
| **[Benajiba, 06]** | Benajiba Yassine Manuel G´omez*, Adapting the JIRS Passage Retrieval System to the Arabic Language*, Spain: Universidad Polit´ecnica de Valencia, 2006. |
| **[Bernard, 08]** | Bernard Emmanuel*, Hibernate Search in Action*, Greenwich : MANNING, 2008. |
| **[Berners-Lee, 01]** | Berners-Lee T.*, The Semantic Web*: Scientific American, 2001. |
| **[Berrut, 97]** | Berrut C*, Indexation des données multimédia, utilisation dans le cadre d'un système de recherche d'informations*: Université Joseph Fourier, 1997. |
| **[Bourne, 79]** | Bourne C*, In Lockheed Information Systems*, Californie, 1979. |
| **[Brin, 98]** | Brin Sergey*, The Anatomy of a Large-Scale Hypertextual Web Search Engine*: Stanford University, 1998. |
| **[Chang, 07]** | Chang B. S.*, Using Ontologies to Search*, Berlin : Computational Science and Its Applications – ICCSA 2007, 2007. |
| **[CHARTRON, 89]** | CHARTRON Ghislaine*, Indexation manuelle et indexation automatique. Dépasser les oppositions*: Documentaliste, vol. 26, n° 4-5, juillet-octobre 1989, p.181-187, 1989. |
| **[CHAUMIER, 90]** | CHAUMIER J.*, L’indexation documentaire, de l’analyse conceptuelle humaine à l’analyse morphosyntaxique*, 1990. |
| **[Choueiter, 06]** | Choueiter Ghinwa F.*, MORPHEME-BASED LANGUAGE MODELING FOR ARABIC LVCSR*: MIT CS and AI Laboratory 32 Vassar St., Cambridge, MA 02139 et IBM T. J. Watson Research Center Yorktown Heights, NY 10598, 2006. |
| **[Corpu, 10]** | *Quranic Arabic Corpus*, URL: http://corpus.quran.com, visité en 2010. |
| **[Dahak, 06]** | Dahak Fouad*, Indexation des documents semi-structurés : Proposition d’une approche basée sur le fichier inversé et le”Trie“*, Alger : Institut National de formation en Informatique (I.N.I), 2006. |
| **[Denos, 97]** | Denos N*, Modélisation de la pertinence en recherche d'information : modèle conceptuel, formalisation et application*: l'Université Joseph Fourier, 1997. |
| **[Denoyer, 04]** | Denoyer L*, Apprentissage et inférence statistique dans les bases de documents structurés : Application aux corpus de documents textuels*: l’université de PARIS 6, 2004. |
| **[Dr.MAHDI, 09]** | *Dr. MAHDI*, URL: http://taharmahdi.free.fr/, visité en 2009. |
| **[Duval, 06]** | Duval Marc*, Les langages des automates de recherche:Les champs de recherche*: dsi-info.ca, 2006. |
| **[Duwairi, 08]** | Duwairi Rehab*, Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization*, QATAR : QATAR UNIVERSITY, 2008. |
| **[El-Hachani, 97]** | El-Hachani Mabrouka*, L'indexation automatique*: DEA Ecole Nationale Supérieure des Sciences de l’Information et des Bibliothèques (ENSSIB), 1997. |
| **[Ferreira da Silva, 07]** | Ferreira da Silva João Ricardo Martins*, SHALLOW PROCESSING OF PORTUGUESE: FROM SENTENCE CHUNKING FROM SENTENCE CHUNKING*, Portugal: Faculdade de Ciências da Universidade de Lisboa, 2007. |
| **[Galoon, 99]** | Galoon *, Le programme Galoon (برنامج قالون) v.1.0*: par le group Sémaphore, 1999. |
| **[Gaudefroy, 75]** | Gaudefroy R. Demombynes*, Grammaire de l'arabe classique (Morphologie et syntaxe)*, Paris: éd. G.P.Maisonneuve et Larose, 1975. |
| **[Grefenstette, 95]** | Grefenstette G.*, Comparing Two Language Identification Schemes.*, Rome, Italy: Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT’95), 1995. |
| **[Grossman, 02]** | Grossman Frieder, Goharian*, IR Basics of Inverted Index*, 2002. |
| **[Gwen, 09]** | Gwen Harris*, Search Syntax*, 2009. |
| **[Hadj henni, 08]** | Hadj henni M’hamed*, Approche ontologique pour la modélisation sémantique, l’indexation et l’interrogation des documents Coraniques*: Memoire de magister ESI, 2008. |
| **[Hensens, 98]** | Hensens Hanka*, Traitement des fonds de Laboratoires : Indexation Principes*, 1998. |
| **[Hernandez, 07]** | Hernandez N.*, Modeling context through domain ontologies*: Journal of Information Retrieval, Springer, Numéro spécial Contextual Information Retrieval Systems, 2007. |
| **[Hernandez, 08]** | Hernandez N.*, A Model to Represent the Facets of Learning Objects*,  Santa Rosa - USA: Interdisciplinary Journal of E-Learning and Learning Objects, Informing Science Institute, 2008. |
| **[Hubert, 09]** | Hubert Gilles*, Modèle d’indexation dynamique à base d’ontologies*, 2009. |
| **[islamweb, 03]** | islamweb *, الخط القرآني والخط الإملائي*, 2003. |
| **[Jalam, 02]** | Jalam Radwan*, Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l’aide des n-grammes caractéristiques*, France: JADT 2002 : 6es Journées internationales d’Analyse statistique des Données Textuelles, 2002. |
| **[Karbasi, 07]** | Karbasi Soheila*, Pondération des termes en recherche d’Information*: Université Paul Sabatier - Toulouse III, 2007. |
| **[Lelu, 92]** | Lelu A*, Information retrieval based on neural unsupervised extraction of thematic fuzzy clusters*, 1992. |
| **[Luhn, 58]** | Luhn H.*, The automatic creation of literature abstracts*: IBM Journal of Research and Development, 1958. |
| **[Manning, 09]** | Manning Christopher D.*, An Introduction to Information Retrieval*, England: Cambridge University Press, 2009. |
| **[Mathias, 02]** | Mathias G*, Indexation et interrogation de chemins de lecture en contexte pour la Recherche d’Information Structurée sur le Web*: l’université JOSEPH FOURIER, 2002. |
| **[Meek, 08]** | Meek Colin*, Web 3.0: what it means for journalists (part 1)*: Online Journalism News, 2008. |
| **[Mentre, 08]** | Mentre Marc*, Les moteurs de recherche sémantique : un pas dans le web 3.0*: Media Trend, 2008. |
| **[menu, 00]** | menu nicolas*, INFORMATIQUE (THEORIE)*, 2000. |
| **[MESFAR, 08]** | MESFAR Slim*, ANALYSE MORPHO-SYNTAXIQUE AUTOMATIQUE ET RECONNAISSANCE DES ENTITES NOMMEES EN ARABE STANDARD*: UNIVERSITE DE FRANCHE-COMTE, 2008. |
| **[Meylan, 01]** | Meylan Ed*, Introduction théorique à la gestion de données textuelles*: Haute Ecole Spécialisée de Suisse Occidentale, 2001. |
| **[Midād lbayān, 10]** | *مداد البيان مشروع قاعدة البيانات القرآنيّة*, URL: http://mbayan.net/, visité en 2010. |
| **[Mustafa, --]** | Mustafa El Hadi Widad*, INDEXATION HUMAINE ET INDEXATION AUTOMATISÉE : LA PLACE DU TERME ET DE SON ENVIRONNEMENT*: UFR IDIST/CERSATES, UMR 8529. |
| **[Nejjari, 07]** | Nejjari Youssef*, Recherche documentaire sur le web*, 2007. |
| **[PEAK, 10]** | *PEAK - The Python Enterprise Application Kit*, URL : http://peak.telecommunity.com/DevCenter/PythonEggs, visité en 2010. |
| **[Pomart, 97]** | Pomart Paul-Dominique*, » Indexation «, article du Dictionnaire encyclopédique de l’information et de la documentation*, Paris: Nathan, 1997. |
| **[PramodeC.E, 04]** | *Python Generator Tricks*, URL : http://linuxgazette.net/100/pramode.html, visité en 2004 Mars. |
| **[Psyché, 04]** | Psyché V.*, Apport de l'ingénierie ontologique aux envirennements de formation à distance*, 2004. |
| **[Python, 10]** | *Le site web officiel du Python*, URL : http://python.org, visité en 2010. |
| **[Robertson, 95]** | Robertson S. E.*, Okapi at trec-3*: In The Third Text REtrieval Conference (TREC-3) NIST, 1995. |
| **[Robertson, 97]** | Robertson S*, Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track*, Maryland, 1997. |
| **[Salton, 71]** | Salton G*, The SMART Retrieval System - Experiment in Automatic Document Processing*, 1971. |
| **[Sanan, 08]** | Sanan Majed*, Etude des methodes de la recherche d’information et de l’indexation sur les documents electroniques : cas de la langue arabe*, Paris: UNIVERSITE PARIS VIII – SAINT DENIS, 2008. |
| **[Sauvagnat, 05]** | Sauvagnat Karen*, Modéle flexible pour la Recherche d’Information dans des corpus de documents semi-structurés*, Toulouse: Doctorat de l’Université Paul Sabatier, 2005. |
| **[SLYPE, 87]** | SLYPE G. VAN*, Les langages d’indexation : conception, construction et utilisation dans les systèmes documentaires*, Paris: Les Editions d’organisation, 1987. |
| **[Song, 05]** | Song J. F.*, Ontology-based Information Retrieval Model for the Semantic Web*: International Conference on e-Technology, e-Commerce and e-Service, 2005. |
| **[Tang, 04 ]** | Tang Hunqiang*, Hybrid Global Local Indexing for Efficient Peer to Peer Information Retrieval*, San Francisco, California : University of Rochester, 2004 . |
| **[Tanzil, 10]** | *Le site web du projt Tanzil*, URL : http://tanzil.info, visité en 2010. |
| **[Thabet, 04]** | Thabet Naglaa*, Stemming the Qur’an*, Newcastle: School of English Literature, Language and Linguistics - University of Newcastle, 2004. |
| **[Tuerlinckx, 04]** | Tuerlinckx Laurence*, La lemmatisation de l’arabe non classique*, Belgique: Centre d’Études sur Grégoire de Nazianze, 2004. |
| **[Whoosh, 10]** | *Le site web de la bibliothèque Whoosh*, URL: http://whoosh.ca, visité en 2010. |
| **[Zarrabi-Zadeh, 09]** | Zarrabi-Zadeh Hamid*, tanzil.info - Quran Text Types*, 2009. |
| **[Zerrouki, 05]** | Zerrouki Taha*, Un modèle de mushaf électronique standard*: mémoire du magister ESI, 2005. |
| **[أبوالحجاج, 09]** | أبوالحجاج محمد بشير*, الويب الدلالية - ثورة الانترنيت المقبلة*, 2009. |
| **[آل الشّيخ, 00]** | آل الشّيخ معالي الشّيخ صالح بن عبد العزيز بن محمّد*, كتابة المصحف الشّريف وطباعته, مجمّع الملك فهد لطباعة المصحف الشّريف*, 2000. |
| **[الجياني, 91]** | الجياني جمال الدين*, الألفاظ المختلفة في المعاني المؤتلفة*, 1991. |
| **[الخميسي, 04]** | الخميسي أحمد حسن*, حركة التأليف المعجمي في مفردات القرآن*, دمشق: مجلة التراث العربي -مجلة فصلية تصدر عن اتحاد الكتاب العرب- , 2004. |
| **[الزّركشي, --]** | الزّركشي بدر الدّين محمّد بن عبد الله بن بهادر*, البُرْهانُ في عُلوم القرْآن*. |
| **[الطّيّار, 08]** | الطّيّار د. محمّد بن سليمان بن ناصر*, المحرّر في علوم القرآن, مركز الدّراسات و العلوم القرآنيّة* : بمعهد الإمام الشّاطبي, 2008. |
| **[الميمان, --]** | الميمان سليمان بن عبد الله*, توظيف التقنيات الحاسوبية لإعداد فهارس هامة و مبتكرة لخدمة القرآن الكريم و علومه*: مجمّع الملك فهد لطباعة المصحف الشريف. |
| **[جماعة, --]** | جماعة محمّد بن*, منهجيّة تعاونيّة لإنجاز موسوعة إلكترونيّة شاملة للقرآن الكريم و علومه*, المدينة الموّرة: مجمّع الملك فهد لطباعة المصحف الشّريف, ندوة القرآن الكريم و التّقنيّات المعاصرة. |
| **[خضر, --]** | خضر محمد زكي محمد*, عرض أولي لمشروع مداد البيان في خدمة القرآن الكريم*. |
| **[خضر, 96]** | خضر د. محمّد زكي*,  المعلوماتيّة في خدمة القرآن الكريم*: الجامعة العالميّة الإسلاميّة – ماليزيا – , 25/06/2004, 1996. |
| **[خضر, 04]** | خضر د. محمّد زكي*, التّعامل مع القرآن الكريم في عصر المعلوماتيّة*: مؤتمر كليّة الآداب الخامس – جامعة الزّرقاء الأهليّة – حضارة الأمّة وتحدّي المعلوماتيّة, 2004. |
| **[خضر, 08]** | خضر الدكتور محمد زكي*, ضوابط الإعجاز العددي في القرآن الكريم*, 2008. |
| **[محبك, 08]** | محبك أحمد زياد*, كتابة الهمزة في اللغة العربية*, 2008. |
| **[محسين, 73]** | محسين الدّكتور محمّد سالم*, تاريخ القرآن الكريم*: دار الأصفهانيّ للطّباعة بجدّة, 1973. |
| **[ناصر, 89]** | ناصر كاظم عادل*, قاموس المفردات المتضادة – انكليزي – عربي*, 1989 . |
| **[نوفل, 75]** | نوفل عبد الرزّاق*, الإعجاز العددي للقرآن الكريم*, 1975 . |

Annexe A

**Étude des moteurs de recherche existants**

**Étude des moteurs de recherche existants**

L’internet contient de nombreux moteurs de recherche : générales ou spécialisées, qui sont diffèrent entre eux en matière de vitesse, le nombre de résultats, ses langues, ses domaines et ses méthodes ; mais la plupart d’entre eux ne prennent pas une partie importante du marché de l'Internet, contrairement aux moteurs de recherche puissants comme : Google et Yahoo. Selon ‘Market Share by Net Applications’, le meilleur moteur de recherche en août 2008 a été Google avec une part incroyable de 79,03%. Yahoo est arrivé en deuxième position en remportant 11,58% et MSN s'est classée le troisième au niveau mondial avec 3,46% du part de marché.  
Le graphique ci-dessous représente les taux de marché des six principaux moteurs de recherche rapporté par ‘Market Share’ Août 2008.

**Figure 84 : Les meilleures moteurs de recherche (août 2008) – source : MARKET SHARE**

**Autres :** représente le pourcentage combiné des moteurs de recherche AltaVista, Excite, tout le Web et Lycos.

Comme nous avons vu dans la figure ci-dessus, Google a la plus grande part. Google est Fondé en 1998 par deux étudiants, Larry Page et Sergey Brin, de Stanford. Il est le plus rapide parmi les moteurs de recherche web, en plus, l'avantage réside dans sa précision, son exhaustivité et sa multiplicité des langues de recherche (il peut rechercher dans 66 langues y compris l'arabe). Il peut traiter plus de 120 millions de requêtes par jour (Media Metrics). La figure suivante montre l'interface de recherche simple de Google :



**Figure 85 : L'interface de recherche simple sur Google**

Dans notre étude, nous nous intéresserons aux moteurs de recherche arabe et en particulier les moteurs de recherche spécifique au Coran.

1. les moteurs de recherche de la langue arabe

Après le grand développement des moteurs de recherche, les caractéristiques pour la langue arabe sont restées non-pertinentes, ce qui a conduit à l’apparition des moteurs de recherche arabes spécialisés tels que « Ayna[[226]](#footnote-25) » et « Araby ».

Nous présenterons quelques moteurs de recherche arabes, en expliquant les caractéristiques les plus importantes de chacun d’eux :

1. Araby (Araby.com)

Araby est un projet développé par le groupe Maktoob (récemment acheté par Yahoo). Il fournit des services de recherche spécialisés en arabe et en tenant compte des règles grammaticales afin d'atteindre tous les contenus en langue arabe sur Internet. Il reconnaisse toutes les structures des phrases arabes, corrige les erreurs grammaticales et suggère aux utilisateurs d'autres termes connexes dans la signification ou la structure, ce qui aide à obtenir des résultats précis.

Araby indexe les sites web en langue arabe (Il utilise un certain nombre de serveurs) sur Internet contrairement à d'autres qui indexent une liste très limitée de sites. Il affiche les résultats les plus proches.

Araby comprend des sections de recherche spécialisés qui comprennent : les nouvelles, l'islam, les forums, les blogs et les images. La section d’islam inclut de nombreux thèmes comme le Coran, les forums islamiques et les études islamiques, ce qui offre aux utilisateurs l’opportunité de profiter de tous les contenus des sites islamiques sur Internet. **[Araby, 09]**



**Figure 86 : Le moteur de recherche Araby**

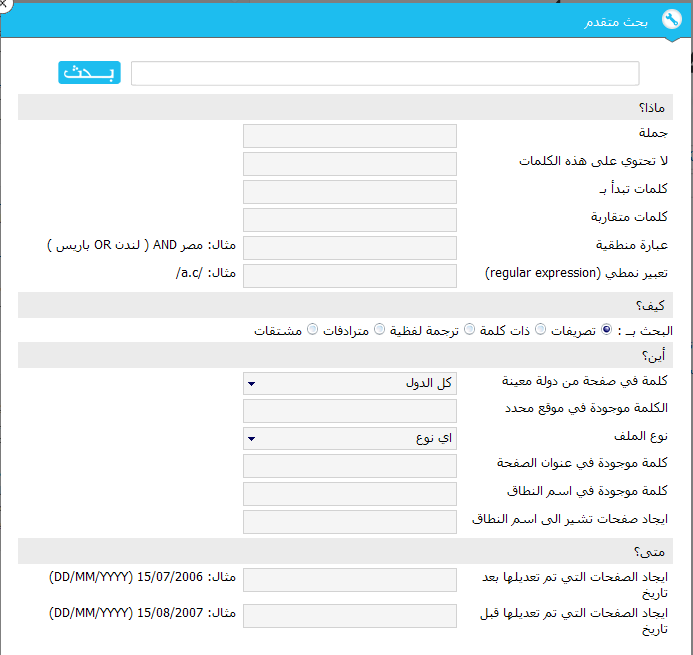
1. **Taya-it  & Tama (tayait.com)**

C’est un produit d’une société américaine appelée « Taya enterprise » spécialisé en linguistique et en traitement du langage naturel. Il comporte essentiellement un analyseur morphologique appelé “Tama” qui fonctionne comme un plug-in aux moteurs de recherche pour améliorer les résultats arabes. Mais à la fin, la société a décidé d’entrer dans le marché arabe comme un moteur de recherche sous le nom de « Taya-it » et pas comme un service supplémentaire.

Taya-it utilise les technologies de recherche d’Exalead (un moteur de recherche français). Il a un volume d’index plus petit que Araby, et sa vitesse de recherche est relativement lente.   
Taya-it fournit les fonctionnalités de recherche suivantes :

* + - * recherche morphologique ;
      * recherche par synonymes ;
      * recherche en anglais dans le contenu arabe ;
* recherche avancée.



**Figure 87 : Aperçu de Tayait.com**

**Figure 88 : Le panneau de recherche avancée – tayait.com**

1. **Ksearch & KIndexer (**[**AlKhawarizmy.com**](http://www.AlKhawarizmy.com)**)**

Ksearch a une technologie avancée d'indexation avec une gestion complète d’index. Sa vitesse d'indexation est de 20.000 mots par seconde. La taille de l'index n’est pas limitée en raison de l'utilisation de la technologie de 64 bits. Ce dernier donne la recherche une grande vitesse.  
Les caractéristiques les plus importantes de Ksearch sont :

* recherche par les différentes formes de conjugaison ;
* distinction entre les différentes significations des mots, ex : le mot « هاتف », a deux sens (téléphoner) ou (clamer) ;
* recherche par des expressions régulières.



**Figure 89: Aperçu de Ksearch.**

1. **IDRISI (Sakhr Software)**

Idrisi (Intelligent Document Retrieval & Information Search) est un moteur de recherche bilingue arabe-anglais pour les données électroniques, documents, bases de données et sites web. Lancé par le propriétaire arabe Sakhr de traitement du langage naturel (Natural Language Processing NLP), la technologie IDRISI prévoit une capacité exceptionnelle de recherche en langue arabe.

Le moteur de recherche IDRISI permet aux organisations d'établir une vue exacte, complète, sûre et intelligente de toutes leurs informations dans des systèmes de fichiers, sites Web, bases de données ou bases de connaissances.

Le moteur de recherche utilise IDRISI NLP pour la correction d’orthographe, l'analyse morphologique et l'homonymie, ce qui permet aux utilisateurs d'effectuer des recherches en langue arabe, basé sur le morphème, les racines et les mots déclinées sur des gros volumes de données pour obtenir des résultats de recherche précis.

IDRISI supporte la recherche par translittération c.-à-d. la recherche en caractères arabes correspondante au terme de recherche anglais, et vice versa. Par exemple, lorsqu’on effectue une recherche pour « Sherihan », les résultats de recherche incluent les mots « Sherihan » et "شريهان".

1. ***Les moteurs de recherche du Coran***

Le besoin de moteurs de recherche supportant bien l'arabe, ou bien des moteurs spécialisés en arabe, s'étend à des besoins arabes spécifiques, comme le besoin d'un moteur de recherche plus intelligents dans les textes islamiques surtout pour le Coran.  
Donc, la diffusion de versions électroniques du Coran, des programmes qui prennent soin de lui et la nécessité de la recherche, a conduit de l'émergence de quelques programmes de potentiel limité vers l'émergence des sites de recherche compétentes comme : *holyquran.net (المنقب القرآني) et alawfa.com (الأوفى)*

Contrairement aux moteurs de recherche arabes, les moteurs de recherche coraniques n’ont pas bénéficiés des caractéristiques modernes ou d’options élaborées de la langue arabe. Cela est peut être dû au fait que la plupart des projets d’arabe sont à but lucratif et non pas open source. Et par conséquent, n'importe quel programmeur qui veut participer dans le développement, doit répéter le même travail par lui-même et nous allons résoudre ce problème par la publication de notre programme avec une licence open source pour l'utilisation d'organismes non-lucratifs.

Nous allons montrer brièvement les caractéristiques de certains de ces moteurs, et nous nous limiterons à ceux qui ont leurs propres fonctionnalités :

1. Holyquran.net

C’est un site web du Coran qui contient le service de «prospecteur coranique » (المنقب القرآني) pour la recherche dans le Coran.  
Il est caractérisé par une interface de recherche avancée avec beaucoup d’options, comme l’ignorance de la forme des lettres qui porte la hamza, la recherche avec la transcription, la recherche dans les sourates, etc. Il ne fournit aucune analyse morphologique, mais se restreint simplement à la recherche à partir de la racine et donne à l’utilisateur le choix entre la recherche par mot entier ou en tant que partie d'un mot.

L’affichage des résultats dans le prospecteur est assez bon, un bouton apparaît pour le tafssīr avec chaque aya, plus le sur-lignage des mots recherchés, et la distribution des résultats sur des pages. Le prospecteur n’est pas basé sur la réaction automatique et ne fait aucune suggestion, l'utilisateur reste comme l’aveugle s’il tape le mot d’une façon autre que laquelle le mot a été sauvegardé (par exemple : écriture en script othmani ou des erreurs d'écriture)   
Ce qui peut être également pris en défaut sur le site est le recours à des références chiites pour le tafssīr.



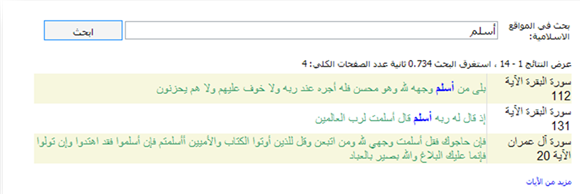
Figure  : Le panneau de recherche avancée - holyquran.net

1. **alawfa.com (الأوفى)**

C’est un moteur de recherche spécialisé en Coran, hadiths et les sites islamiques, il se considère comme étant le premier dans ce domaine. Ce qui le distingue du reste des autres sites qui fournissent la recherche dans le Coran est le fait qu’il a été spécifiquement conçu à cet effet et a été construit sur les bases des moteurs de recherche.

Alawfa porte quelques bonnes caractéristiques, telles que l’exploration des sites et la suggestion de mots, le sur-lignage, etc. Mais elles restent très modestes par rapport à ce que les moteurs de recherche étrangers ont. L’Awfa ne prévoit pas un panneau de recherche avancé, mais base sur la spontanéité, ce qui est bien, mais n'aide pas les chercheurs qui préfèrent utiliser des options spécifiques sans d'autres. L'analyse morphologique des mots entrés est mauvaise ou presque inexistante, en plus du mauvais traitement avec les signes de diacritiques.

L’Awfa s’intéresse à l’affichage des résultats de recherche sur les sites musulmans et néglige le traitement des résultats de recherche dans le Coran comme si ils étaient des résultats secondaires. Il fonctionne aujourd'hui avec une version de test, et en attendant la version complète, on espère qu’il remédiera aux faiblesses de la version expérimentale.



**Figure 91 : La recherche du mot « أسلم»- alawfa.com**

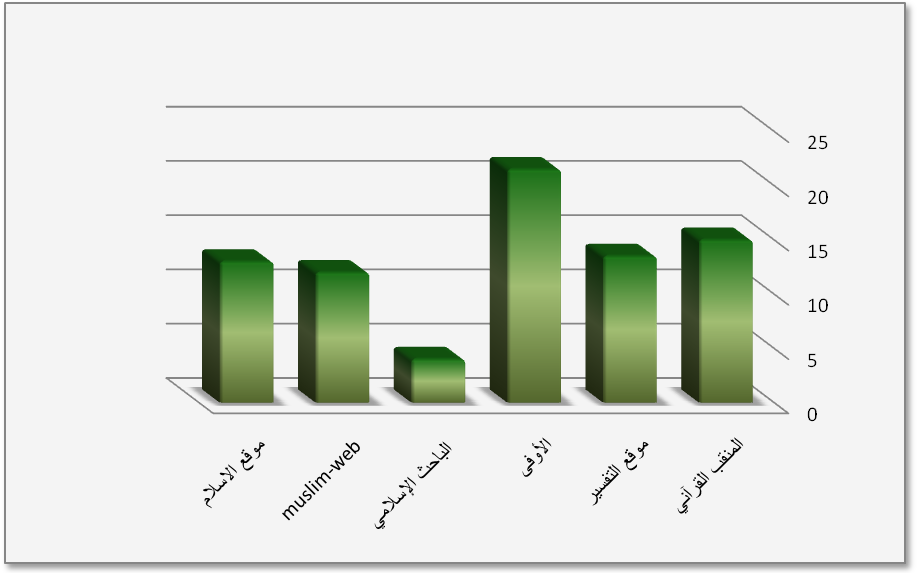
1. **altaffsir.com**

Il n’est pas très différent de l’exploiteur coranique (المنقب القرآني), sauf qu’il est caractérisé par la recherche au niveau de la racine et la recherche dans les tafssīrs. Mais malheureusement, il ne fournit pas l’option d’ignorer la hamza (الهمزة) ni le sur-lignage (highlight).



**Figure 92 : Le panneau de recherche – altaffsir.com**

L’histogramme suivant exprime l’évaluation approximative des moteurs de recherche coraniques (pour plus de détails voir le tableau de comparaison à la fin de cette annexe) :



**Figure 93 : Comparaison entre les sites de recherche du Coran**

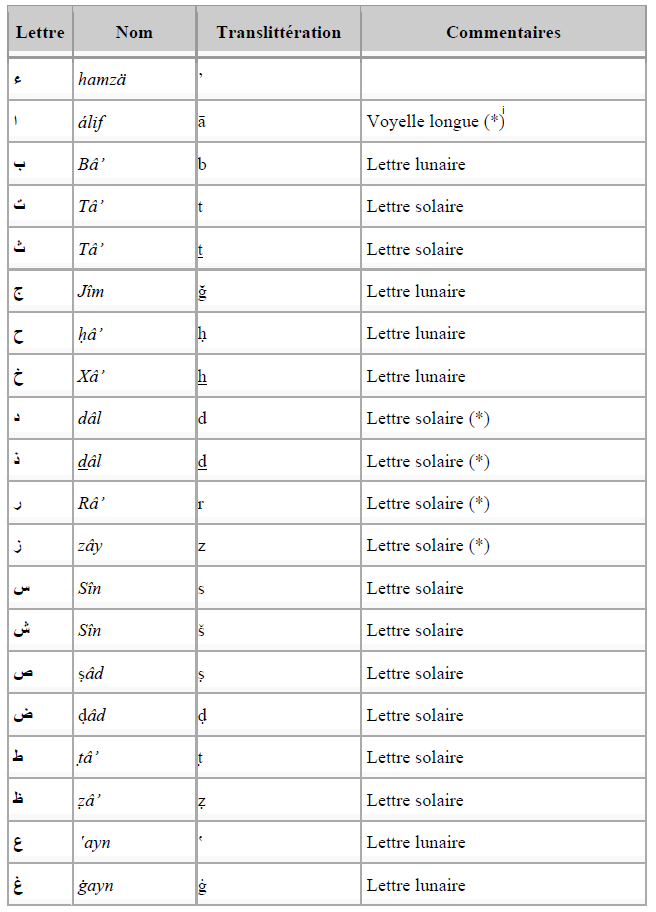




Annexe B

**Tableau de translittération de l’alphabet arabe**

**Tableau de translittération de l’alphabet arabe [MESFAR, 08]**



1. Le symbole (\*) désigne les six lettres de l’alphabet qui ne s'attachent jamais à la lettre suivante.



Annexe C

**L’approche sémantique**

**L’approche sémantique**

1. **Les moteurs de recherche sémantiques**

Les moteurs de recherche sémantiques reposent sur une nouvelle approche, liée à ce qu’on appelle le Web 3.0, ou Web sémantique, comme l’a baptisé Tim Berners-Lee, l’un des créateurs de la world wide web (les fameux “www” que l’on trouve devant la plupart des adresses de sites web).

Le web sémantique vise à lier (*link up*) différents types d’information, horaires d’avion, prévisions météo, bookmarks de réseaux sociaux, informations, qui sont toutes publiées dans différents formats et accessibles par des applications différentes.**[Meek, 08]**

Il s’agit de faire en sorte que l’on puisse consulter son relevé de banque sur son calendrier, par exemple, alors que ces données dépendent de deux applications distinctes. Atteindre cet objectif, implique une formalisation, qui passe par des “outils sémantiques” et des langages [définis par le W3C](http://www.w3.org/2001/sw/) (*World Wide Web Consortium*).**[Mentre, 08]**

1. Évolution vers le web sémantique : le web 3.0

De plus en plus de producteurs de [contenu](http://fr.wikipedia.org/wiki/Contenu), à la suite des recommandations du [W3C](http://fr.wikipedia.org/wiki/W3C) sur le [Web sémantique](http://fr.wikipedia.org/wiki/Web_s%C3%A9mantique), indexent leurs bases avec des [métadonnées](http://fr.wikipedia.org/wiki/M%C3%A9tadonn%C3%A9e) ou des [ontologies](http://fr.wikipedia.org/wiki/Ontologie_%28informatique%29), en vue de permettre aux moteurs de recherche de s'adapter aux [analyses sémantiques](http://fr.wikipedia.org/wiki/Analyses_s%C3%A9mantiques).

Il convient de préciser que ces formes de recherches et d'analyses de corpus d'informations par voie informatique ne sont encore que des potentialités. De nombreux algorithmes ou méthodes de travail devront être mis au point avant de pouvoir créer un système de recherche d'information véritablement sémantique.

En effet, par comparaison avec des [recherches plein texte](http://fr.wikipedia.org/wiki/Recherches_plein_texte), de véritables recherches réalisées sur le [web sémantique](http://fr.wikipedia.org/wiki/Web_s%C3%A9mantique) devraient être beaucoup plus conviviales pour l'utilisateur: contrairement à un moteur interrogé en mode similarité (requête par mot clé appelant la fourniture de documents pertinents) un système sémantique n'impose pas à l'utilisateur de fournir les éléments de la réponses sous forme de mots clés.

* L'utilisateur d'un système sémantique doit pouvoir directement poser sa question en langue naturelle.
* Un véritable moteur de recherche sémantique ne fournira pas de liste de pages répondant à une question mais la réponse précise.

Il n'existe pas encore à proprement parler de moteur de recherche [sémantique](http://fr.wikipedia.org/wiki/S%C3%A9mantique). Mettre au point un véritable moteur de recherche capable de comprendre et de fournir du sens et capable de comprendre une question en langue naturelle ou d'adapter une réponse en fonction d'un ensemble d'information) est encore difficile. Quelques tentatives existent néanmoins pour chercher à répondre par des formes intermédiaires à cette problématique du sens dans la recherche d’information :

* [GREPER](http://fr.wikipedia.org/w/index.php?title=GREPER&action=edit&redlink=1), qui répond aux questions en langue naturelle
* [Powerset](http://fr.wikipedia.org/w/index.php?title=Powerset&action=edit&redlink=1) de Microsoft, qui répond aux questions en langue naturelle
* [Nlgbase](http://fr.wikipedia.org/w/index.php?title=NLGbAse&action=edit&redlink=1), qui permet d'interroger une ontologie extraite depuis Wikipedia
* Le projet de recherche Edelweiss, de l'inria qui développe des outils exploitant les triplets [RDF](http://fr.wikipedia.org/wiki/RDF).
* [Kartoo](http://fr.wikipedia.org/wiki/Kartoo) de la société Kartoo qui affiche des graphes sémantiques dans le cadre de ses **cartes de recherche**
* Sinequa CS de [Sinequa](http://fr.wikipedia.org/wiki/Sinequa) qui fût l'un des premiers moteurs de réponses implémenté en situation réelle, sur le site institutionnel de [Gaz de France](http://fr.wikipedia.org/wiki/Gaz_de_France).

1. ***La recherche par la langue naturelle***

Une recherche en langage naturel est une recherche en utilisant la langue parlée ordinaire, comme l'anglais. En utilisant de ce type de recherche nous pouvons poser une question à la base de données, ou taper une phrase qui décrit l'information que nous recherchons. La base de données utilise ensuite une logique programmée pour déterminer les mots-clés par leur position dans la phrase.

Un moteur de recherche en langage naturel pourra en théorie retrouver des réponses appropriées aux questions des utilisateurs (par opposition à la recherche par mots-clés). Par exemple, lorsqu'ils sont confrontés à une question de la forme « Quel est l’état américain qui a le plus d’impôt sur les revenus ? ». Les moteurs de recherche classiques ignorent la question, et au lieu de répondre, ils effectuent une recherche sur les mots-clés : ‘état’, ‘revenu’ et ‘impôt’. La recherche en langage naturel, d'autre part, tentent d'utiliser le traitement du langage naturel pour comprendre la nature de la question, ensuite pour rechercher et retrouver un sous-ensemble du web qui contient la réponse à la question. Si cela fonctionne, les résultats auraient une pertinence beaucoup plus élevé que les résultats d'un moteur de recherche par mots-clés.

**Examples**:

* **Google** : Utilise des phrases naturelles, il maintient les termes dans leur ordre naturel. Puis il ajoute des mots du contexte pour distinguer entre les sens, tels que « Apple » l'ordinateur et « Apple » qui veut dire pomme en anglais.

1. « First woman to climb Mount Everest » - Résultats : L'histoire du mont Everest, la page montre la première femme (Junko Tabei du Japon en 1975) et la première femme américaine (Stacy Allison en 1987). Sharon Wood, un Canadien, prétend également avoir été la première femme américaine, sa montée a été en 1986.
2. « Budget accommodation in Paris France » - Résultats : Google trouve des centaines de possibilités. Les liens commerciaux pourraient être utiles aussi. Cette requête fonctionne même si on a fait des fautes d’orthographe. Cela fonctionne avec n'importe quelle destination.
3. Convertir les gallons en mètres cubes - Résultats : Il ya plusieurs résultats sur la première page qui peuvent être utilisé pour convertir les volumes. Google donne également une réponse instantanée.

* **Autres moteurs de recherche :** Tous les moteurs de recherche accepteront les phrases du langage naturel. Beaucoup vont aussi faire des calculs et des conversions : Ask, Yahoo, Live (mais pas AOL).
* **Hakia** se nourrit de mots. Il répond bien sur les questions de fait. Souvent, la réponse est dans le texte en surbrillance. On remarque la diversité des documents indexés par Hakia. Les recherches floues montrent comment Hakia cherche le sens plutôt que les mots exacts, par exemple :
  1. « using technology to work together »  – Résultats : Les résultats web correspondent bien à la phrase. Il existe également plusieurs résultats à partir de sites crédibles (les sources identifiées par les bibliothécaires). Il s'agit d'une recherche universelle - il ya des résultats à partir d'images et d’actualités.
  2. « Diagnosing concussion » – Résultats : Hakia nous montre sa galerie de sujets avec les informations qui ont été assemblées pour former un paquet : information de base, les symptômes et les diagnostics, etc.
* **Ask** cherche des réponses. Il n'hésite pas à répondre à chacune des questions ci‑dessous :
  1. Première femme à escalader le mont Everest – Résultats : le premier résultat est la bonne réponse.
  2. Who was tecumseh ? - Résultats : Ask a une réponse intelligente.
  3. What are the qualities of an online course ? – Résultats : Ask renvoie des résultats très pertinents. **[Gwen, 09]**

1. ***Le futur des moteurs de recherche sémantiques***

Les moteurs de recherche actuels acceptent les mots-clés comme des entrées et affichent la liste des liens vers les documents et les pages qui contiennent ces mots, la qualité des résultats est toujours déterminée par les niveaux de classification de la liste. Par contre les moteurs de recherche sémantiques seront complètement différents, ils contiendront deux avantages majeurs par rapport aux moteurs de recherche traditionnels, la première est qu'ils accepteront les requêtes de l'utilisateur qui sont élaborés en langage naturel, deuxièmement, le résultat donné sera celui recherché par l’utilisateur, et non une liste d’ informations hétérogènes.

Ainsi, dans le web sémantique, les moteurs de recherche auront une fonction différente. Les capacités seront différentes de ce qui existe aujourd’hui. La fonction de moteur de recherche sémantique se concentrera dans le traitement des informations très volumineuses. Contrairement à ce que font les autres moteurs de recherche à savoir, l’ouverture de chaque document à partir d’une liste, puis la lecture rapide du contenu et puis décider si le contenu est pertinent pour l’interrogation de la recherche ou non. Alors que les moteurs de recherche sémantiques vont faire une révolution dans la façon de recherche de l’information numérique, car ils passeront d'une période de recherche et de récupération des documents vers une période de réponse aux questions.

L’un des défis qui se posent aux moteurs de recherche sémantiques c’est de se limiter aux résultats contenant la bonne réponse, mais ils souffrent de l’incapacité de parcourir ou lire une énorme quantité de données pour extraire la bonne réponse. Imaginer les énormes possibilités et applications, si cela s'est produit. Si un moteur de recherche est en mesure de combiner un moteur sémantique et un moteur de recherche classique, alors il sera en mesure de trouver des résultats limités qui ont une relation directe avec l’interrogation, et le moteur pourra accéder à des index (ou des listes) contenant la liste complète de toutes les emplacements où un terme particulier appartient. Puis il pourra utiliser la logique pour exclure tout ce il n’a pas une relation avec une interrogation spécifique.

Il est possible à l’avenir que les moteurs de recherche sémantiques deviennent comme des clients ou des agents de la recherche d’information. Ils pourront résoudre le problème de l’énorme quantité d’informations à travers la réalisation de différents niveaux de comportement intelligent, en commençant par une simple réponse jusqu’au problème d'adaptation et d'apprentissage c’est le désire de l'utilisateur qui n’existe pas encore. Ça permettra d’éviter le temps perdu dans la lecture des informations non requises et d’éviter également de faire des efforts sauf dans des informations qui ont une valeur réelle pour le but demandé.

Les moteurs de recherche sémantiques devraient être en mesure d'échanger et de traiter les données même si chacun d'eux s’est développé séparément. Ce qui crée un espace pour que les navigateurs sémantiques naviguent d'une page à une autre en se basant sur le contenu de ces pages et effectuent des tâches complexes au profit des utilisateurs. Ces programmes sauront que un site apparient à un dispensaire de physiothérapie, par exemple, et que le médecin (son nom) travaille à la clinique tous les jours de lundi, mercredi et vendredi, et qu'il doit entrer les dates dans le site sous la forme (jour - mois - année) pour afficher l’emploi de réservation. Donc, le web sémantique n'est pas une section distincte, mais une extension des pages actuelles, où des significations particulières d'informations ont été utilisées, permettant aux ordinateurs et aux humains de travailler ensemble avec plus de bénéfices.**[أبوالحجاج, 09]**

1. **Indexation sémantique**

Suite au développement du web sémantique et de ses technologies **[Berners-Lee, 01]** , différents travaux s’intéressent à son application en RI. L’utilisation d’une ontologie lors de la phase d’indexation permettrait de lever les ambiguïtés des sens des termes utilisés et de mieux représenter les connaissances inhérentes dans les documents. En termes d’indexation sémantique, des concepts de l’ontologie sont associés à chaque document selon les sémantiques qui y sont véhiculées.

Différents travaux ont montré l’intérêt d’utiliser une indexation sémantique à base d’ontologie. Dans le domaine de l’apprentissage en ligne, **[Chang, 07]** propose une indexation basée à la fois sur une ontologie du domaine de l’apprentissage et sur une ontologie dérivée de LOM (Learning Object Metadata), qui représente les métadonnées décrivant les ressources pédagogiques. Dans le cadre des recherches d’objets pédagogiques relatifs aux mathématiques en secondaire, les résultats montrent une meilleure efficacité en termes de rappel et de précision par rapport aux mêmes recherches basées sur mots-clés. De même,**[Hernandez, 08]** utilise les termes d’une ontologie de domaine, associée à une ontologie de tâche et de scenario d’apprentissage comme valeurs des métadonnées de LOM. Afin d’accéder aux instances d’ontologie d’une part et aux index associés aux documents d’autre part, **[Hernandez, 07]** propose de les stocker dans une base de données relationnelles.

Par ailleurs, **[Song, 05]** propose un modèle de RI basé sur des ontologies de domaine, définies avec OWL lite. Les différentes ontologies de domaines sont intégrées pour former une ontologie unique. Les termes définis dans l’ontologie sont alors utilisés d’une part comme métadonnée pour annoter les contenus du web et d’autre part comme termes d’indexation de la collection.  
Dans ces différents travaux, les structures de données utilisées permettent d’associer des concepts issus d’une ontologie aux documents de la collection.

Cependant, le suivi de la dynamique (ajout, suppression, modification) des documents de la collection ainsi que l’impact de ces évolutions au niveau de l’index sémantique n’a pas été traité. Dans la section suivante, nous présentons le modèle de données que nous avons défini et qui permet une indexation sémantique dynamique.**[Hubert, 09]**

1. Définition de l’ontologie

La définition qui revient le plus souvent dans les écrits est celle de Gruber qui définit les ontologies comme étant la spécification explicite d’une conceptualisation d’un domaine de connaissance :

« An ontology is an explicit specification of a conceptualization » **[Psyché, 04]** .Cette définition a été précisée par Borst en 1997 pour devenir: « la spécification formelle et explicite d’une conceptualisation partagée ».

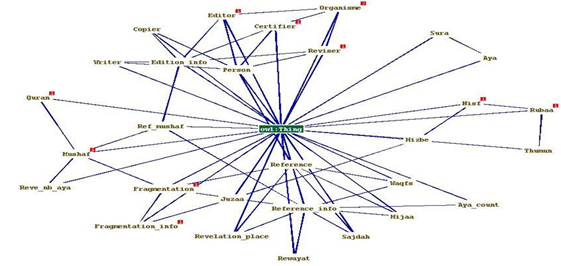
Le terme « spécification explicite «  employé dans ces définitions signifie qu’une ontologie est un ensemble de concepts, de propriétés , d’axiomes ,de fonctions et de contraintes explicitement définis. **[Psyché, 04]** **[Hadj henni, 08]**

1. ***L’ontologie du Coran faite par M. Hadj henni :***

Il existe plusieurs possibilités d’ontologie pour le contenu du Coran. M. Hadj henni a procédé à la construction d’une ontologie de domaine des documents coraniques par la méthode de mapping du schéma XML du modèle standard du Mushaf électronique vers le langage OWL. Ce choix est justifié par les avantages très intéressants de ce langage.

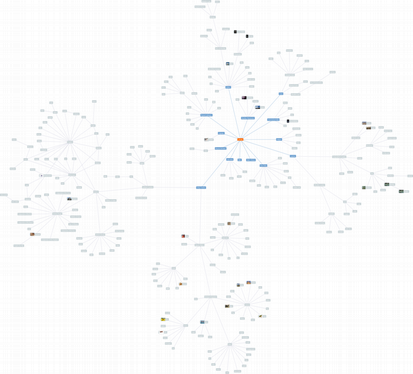
Dans un premier temps, il a réalisé ce mapping automatiquement, ensuite manuellement. Il a constaté que ce dernier s’avère satisfaisant et plus performant que le mapping automatique. L’ontologie obtenue par le mapping manuel est enrichie pour constituer la première ontologie de domaine des documents coraniques.

Cette ontologie répond aux besoins de modèle standard sémantique, réutilisable et modulaire ; elle contiendra les concepts du domaine des documents coraniques et les contraintes sémantiques et servira de base à la construction d’autres ontologies en relation avec le contenu du Coran.**[Hadj henni, 08]**



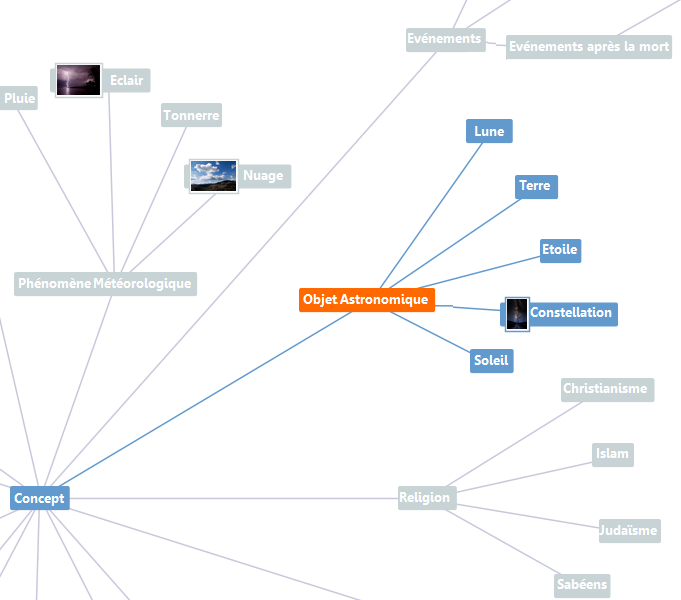
**Figure 94 : Schéma de l’ontologie de domaine des documents Coraniques**

1. ***L’ontologie du Corpus coranique***

Il existe aussi une autre ontologie coranique qui est celle du Corpus coranique. Les concepts fondamentaux dans cette ontologie sont basés sur les informations contenues dans les sources traditionnelles telles que le hadīth du prophète Mohamed set le tafssīr d’Ibn Kathīr (ابن كثير). Les entités citées dans les ayas, comme les noms des personnes et lieux historiques mentionnés dans le Coran, sont liées à des concepts dans l'ontologie dans le cadre de l'étiquetage des entités citées. Le diagramme ci-dessous montre une représentation visuelle de l'ontologie. Le schéma est une partie d’un réseau de 300 concepts liés avec 350 relations. **[Corpu, 10]**

**Figure 95: L’ontologie coranique. Chaque nœud du graphe représente un concept unique dans le Coran.**

Aussi bien que lister des principaux concepts dans le Coran, l'ontologie définit également un ensemble de relations sémantiques entre ces concepts. La relation la plus importante est la relation d‘appartenance « *membre* » dans laquelle un concept est défini comme étant une instance ou d'un membre individuel d'un autre groupe. Par exemple la relation «*Satan* est un *djinn* » dans l'ontologie représente l’information contenue dans le Coran que l'individu connu sous le nom de *Satan* appartient à l'ensemble de créatures vivantes nommé les *djinns*. D'autres concepts dans l'ontologie sont regroupés en catégories, en fonction des propriétés qu'ils partagent. Par exemple, le Soleil, la Terre et la Lune sont classées sous « objets astronomiques » :



**Figure 96: une vue plus proche, montrant les relations sémantiques entre les concepts.**

**Annexe D**

Lexique des termes arabes

On liste ici les noms propres et les termes d’origine arabe utilisés selon l’ordre de leur première occurrence, ainsi que leur écriture en lettres arabes, afin de faciliter la révision des traductions choisies :

1. Le tafsīr : التّفسير [↑](#endnote-ref-1)
2. La récitation : التّلاوة [↑](#endnote-ref-2)
3. Par génération successive : التّواتر [↑](#endnote-ref-3)
4. Al-Furkān : الفرقان [↑](#endnote-ref-4)
5. Al-kitāb الكتاب [↑](#endnote-ref-5)
6. Al-dhikr : الذكر [↑](#endnote-ref-6)
7. Al-wahy : الوحي [↑](#endnote-ref-7)
8. Al-rōuh : الروح [↑](#endnote-ref-8)
9. Les compagnons : الصّحابة f [↑](#endnote-ref-9)
10. La bataille d’Ahl Al-Yamāmah : أهل اليمامة [↑](#endnote-ref-10)
11. Zayd ibn Thābit : زيد بن ثابت d [↑](#endnote-ref-11)
12. Le calife Othmāne ibn ’Affān d : d الخليفة عثمان بن عفّان [↑](#endnote-ref-12)
13. Les façons de lecture : القراءات  [↑](#endnote-ref-13)
14. Le script othmani : الخطّ العثماني [↑](#endnote-ref-14)
15. Le logiciel ‘Salsabīl’ : سلسبيل [↑](#endnote-ref-15)
16. Le calife ’Abu-Bakr d: d أبو بكر الخليفة [↑](#endnote-ref-16)
17. La rewayāte : الرّواية [↑](#endnote-ref-17)
18. L'orthographe : الهجاء [↑](#endnote-ref-18)
19. Les règles d’écriture : الضّبط [↑](#endnote-ref-19)
20. Fragmentation en ahzāb : أحزاب [↑](#endnote-ref-20)
21. Mecquois : مكّي [↑](#endnote-ref-21)
22. Médinois : مدني [↑](#endnote-ref-22)
23. L’indication des waqfs : بيان الأوقاف [↑](#endnote-ref-23)
24. L’indication des sajdahs : بيان السّجدات [↑](#endnote-ref-24)
25. Le rédacteur du mushaf : خطّاط المصحف [↑](#endnote-ref-25)
26. Le copieur du mushaf : ناسخ المصحف [↑](#endnote-ref-26)
27. La commission de révision : لجنة المراجعة [↑](#endnote-ref-27)
28. L’éditeur du mushaf : ناشر المصحف [↑](#endnote-ref-28)
29. L’organisme certifiant : هيئة التّرخيص [↑](#endnote-ref-29)
30. L’apprentissage : الحفظ [↑](#endnote-ref-30)
31. La basmala : البسملة [↑](#endnote-ref-31)
32. Sourate El-Tawba : سورة التّوبة [↑](#endnote-ref-32)
33. L’ordre dans le mushaf : حسب ترتيب المصحف [↑](#endnote-ref-33)
34. L’ordre selon la révélation : التّرتيب حسب النّزول [↑](#endnote-ref-34)
35. Le juz’ : جزء [↑](#endnote-ref-35)
36. Le Hizb : حزب [↑](#endnote-ref-36)
37. Le nisf : نصف [↑](#endnote-ref-37)
38. Le rub’ : ربع [↑](#endnote-ref-38)
39. Le thumn : ثمن [↑](#endnote-ref-39)
40. Les Manāzil : منازل [↑](#endnote-ref-40)
41. Sourate al-fātiha: سورة الفاتحة [↑](#endnote-ref-41)
42. Sourate Al-Māida : المائدة [↑](#endnote-ref-42)
43. Sourate Yūnus : سورة يونس [↑](#endnote-ref-43)
44. Sourate Al-Isra : سورة الإسراء [↑](#endnote-ref-44)
45. Sourate Al-Chu’arā’ : سورة الشّعراء [↑](#endnote-ref-45)
46. Sourate As-Sāffāte : سورة الصّافّات [↑](#endnote-ref-46)
47. Sourate Quf : سورة ق [↑](#endnote-ref-47)
48. Ruku‘s : ركوعات [↑](#endnote-ref-48)
49. Ruku ’ : ركوع [↑](#endnote-ref-49)
50. Al-Houmazat : الهُمزة [↑](#endnote-ref-50)
51. En rewayate de Kaloun : برواية قالون [↑](#endnote-ref-51)
52. Une virgule coranique : فاصلة قرآنية [↑](#endnote-ref-52)
53. Flexion : الإعراب [↑](#endnote-ref-53)
54. La conjugaison : الصّرف [↑](#endnote-ref-54)
55. La rhétorique : البلاغة [↑](#endnote-ref-55)
56. La lexicologie : علم المعاجم [↑](#endnote-ref-56)
57. La sunna : السّنّة [↑](#endnote-ref-57)
58. Le général et le particulier : الخاصّ و العامّ [↑](#endnote-ref-58)
59. La migration : الهجرة النّبويّة [↑](#endnote-ref-59)
60. Une prosternation de récitation : سجود التّلاوة [↑](#endnote-ref-60)
61. Sourate Al‑Baqara : سورة البقرة [↑](#endnote-ref-61)
62. Sourate Āl-‘Imrān : سورة آل عمران [↑](#endnote-ref-62)
63. L’abrogeant et de l’abrogé : النّاسخ و المنسوخ [↑](#endnote-ref-63)
64. La conjugaison : التّصريف [↑](#endnote-ref-64)
65. Le morphème : المصدر [↑](#endnote-ref-65)
66. Nom du fā‛il: اسم الفاعل [↑](#endnote-ref-66)
67. L’adjonction : الزّيادة [↑](#endnote-ref-67)
68. Le renversement : القلب [↑](#endnote-ref-68)
69. Al-Idghām : الإدغام [↑](#endnote-ref-69)
70. La connaissance de l’orthographie : علم مرسوم الخط [↑](#endnote-ref-70)
71. Le cercueil : التّابوت [↑](#endnote-ref-71)
72. Les Koraïchites : القرشيّون [↑](#endnote-ref-72)
73. Koraiche : قريش [↑](#endnote-ref-73)
74. Abdellāh ibn Droustouīh : عبد الله ابن درستويه [↑](#endnote-ref-74)
75. Le découpage de la poésie : تقطيع العروض [↑](#endnote-ref-75)
76. L’analyse grammaticale du Coran : إعراب ألفاظ القرآن [↑](#endnote-ref-76)
77. La science des semblants : علم المتشابه [↑](#endnote-ref-77)
78. La définition et l’indéfinition : التّعريف و التّنكير [↑](#endnote-ref-78)
79. Les débuts des surates : فواتح السّور [↑](#endnote-ref-79)
80. Les lois : الأحكام [↑](#endnote-ref-80)
81. Al-’išmam : الإشمام [↑](#endnote-ref-81)
82. Al-’imala : الإمالة [↑](#endnote-ref-82)
83. Sourate Al-Nnaçr : سورة النّصر [↑](#endnote-ref-83)
84. Le programme de : Galoon برنامج قالون [↑](#endnote-ref-84)
85. Sourate d’Al-Qalam : سورة القلم [↑](#endnote-ref-85)
86. Le logiciel ‘mushaf Al-Madina’ pour les publications informatiques : مصحف المدينة النّبويّة للنّشر الحاسوبي [↑](#endnote-ref-86)
87. Le complexe Roi Fahd pour l'impression du Saint Coran : مجمّع الملك فهد لطباعة المصحف الشّريف [↑](#endnote-ref-87)
88. Le Saint Coran : القرآن الكريم [↑](#endnote-ref-88)
89. Lexiques : معاجم [↑](#endnote-ref-89)
90. Une méthode de pondération [↑](#footnote-ref-1)
91. Linus Thorvald est un [informaticien](http://fr.wikipedia.org/wiki/Informaticien) [finlandais](http://fr.wikipedia.org/wiki/Finlande). Il est connu pour avoir créé en [1991](http://fr.wikipedia.org/wiki/1991) le [noyau Linux](http://fr.wikipedia.org/wiki/Noyau_Linux) dont il continue à diriger le [développement](http://fr.wikipedia.org/wiki/D%C3%A9veloppement_de_logiciel). [↑](#footnote-ref-2)
92. Ubuntu est une distribution Linux basée sur Debian et destinée à proposer un système convivial et ergonomique [↑](#footnote-ref-3)
93. Diacritiques : علامات الشكل [↑](#endnote-ref-90)
94. Les scientistes de la langue arabe considèrent le hamza comme une lettre [↑](#footnote-ref-4)
95. Vocalisé : مشكول [↑](#endnote-ref-91)
96. Sokone : سكون [↑](#endnote-ref-92)
97. Tashdeed : التشديد [↑](#endnote-ref-93)
98. Pattern : وزن الكلمة العربية [↑](#endnote-ref-94)
99. Racine : جذر الكلمة العربية [↑](#endnote-ref-95)
100. Suffixes : لواحق [↑](#endnote-ref-96)
101. Préfixes : سوابق [↑](#endnote-ref-97)
102. Les mots vides (ou stop words, en anglais) sont des [mots](bword://!!Z8AVT3H7RE,mot/) qui sont tellement communs qu'il est inutile de les indexer ou de les utiliser dans une recherche. En français, des mots vides évidents pourraient être « [le](bword://!!Z8AVT3H7RE,Article%20d%C3%A9fini/) », « [la](bword://!!Z8AVT3H7RE,Article%20d%C3%A9fini/) », « [de](bword://!!Z8AVT3H7RE,D%C3%A9terminant%20%28grammaire%29/) », « [du](bword://!!Z8AVT3H7RE,D%C3%A9terminant%20%28grammaire%29/) », « [ce](bword://!!Z8AVT3H7RE,D%C3%A9terminant%20%28grammaire%29/) » … [↑](#footnote-ref-5)
103. Wâw : حرف الواو [↑](#endnote-ref-98)
104. Les pronoms relatifs : الضمائر المتصلة [↑](#endnote-ref-99)
105. الضمائر المتصلة ما تلحق الاسم أَو الفعل أَو الحرف فتكون مع ما تتصل به كالكلمة الواحدة، وذلك مثل التاء والكاف والهاء في قولنا: (حضرتُخطابـك الموجه إليـه ( [↑](#footnote-ref-6)
106. L’état flexionnel : الحالة الإعرابيّة [↑](#endnote-ref-100)
107. Muğarrada : مجردة [↑](#endnote-ref-101)
108. Mazīda : المزيدة [↑](#endnote-ref-102)
109. Passives : مبنية للمجهول [↑](#endnote-ref-103)
110. Actives : مبنية للمعلوم [↑](#endnote-ref-104)
111. Confirmés : المؤكَّدة [↑](#endnote-ref-105)
112. Présent : المضارع [↑](#endnote-ref-106)
113. Originaux : أصلية [↑](#endnote-ref-107)
114. Mīmiya : ميمية [↑](#endnote-ref-108)
115. Forme : الهيئة [↑](#endnote-ref-109)
116. Fois : المرة [↑](#endnote-ref-110)
117. Nom du fā‛il : اسم الفاعل [↑](#endnote-ref-111)
118. Nom du maf‛ūl : اسم المفعول [↑](#endnote-ref-112)
119. Circonstances de lieu et de temps : أسماء المكان و الزمان [↑](#endnote-ref-113)
120. Noms d’instrument : أسماء الآلة [↑](#endnote-ref-114)
121. Exagération de nom du fā‛il : مبالغة اسم الفاعل [↑](#endnote-ref-115)
122. Quasi adjectives : الصفات المشبهة [↑](#endnote-ref-116)
123. Tathniya : التثنية [↑](#endnote-ref-117)
124. L’annexion : الإضافة [↑](#endnote-ref-118)
125. La conjugaison des verbes : تصريف الأفعال [↑](#endnote-ref-119)
126. Raf' : رفع [↑](#endnote-ref-120)
127. La déclinaison des noms : اشتقاق الأسماء [↑](#endnote-ref-121)
128. Le morphème est la plus petite unité linguistique qui effectue une interprétation sémantique [Porter, 2001] [↑](#footnote-ref-7)
129. Morphème : مصدر [↑](#endnote-ref-122)
130. Nature : نوع [↑](#endnote-ref-123)
131. Nisba : نسبة [↑](#endnote-ref-124)
132. tasġīr: تصغير [↑](#endnote-ref-125)
133. Passive : مبني للمجهول [↑](#endnote-ref-126)
134. Active : مبني للمعلوم [↑](#endnote-ref-127)
135. Hamza : همزة [↑](#endnote-ref-128)
136. Nabira : نبرة [↑](#endnote-ref-129)
137. Hâ’ : هاء [↑](#endnote-ref-130)
138. tâ’ marbûtä : تاء مربوطة [↑](#endnote-ref-131)
139. Yâ’ : ياء  [↑](#endnote-ref-132)
140. Aleph maqsura ألف مقصورة [↑](#endnote-ref-133)
141. Warch : ورش مقصورة اءt rche ternes [↑](#endnote-ref-134)
142. **الإمالة:** هي أن تنحوا بالفتحة نحو الكسرة، وبالألف نحو الياء من غير قلب خالص ولا إشباع مبالغ فيه [↑](#footnote-ref-8)
143. 'imāala : الإمالة [↑](#endnote-ref-135)
144. Script othmani الخط العثماني [↑](#endnote-ref-136)
145. Le hamza est au dessous de nabira [↑](#footnote-ref-9)
146. Juz’ : جزء [↑](#endnote-ref-137)
147. Hezb : حزب [↑](#endnote-ref-138)
148. Nesf : نصف [↑](#endnote-ref-139)
149. Rubu’ : ربع [↑](#endnote-ref-140)
150. Thumun : ثمن [↑](#endnote-ref-141)
151. Sourate : سورة [↑](#endnote-ref-142)
152. Ruku’ : ركع [↑](#endnote-ref-143)
153. Juz’ ‘amma : جزء عمّ [↑](#endnote-ref-144)
154. Recherche par expression régulière [↑](#footnote-ref-10)
155. Shari'a : شريعة  [↑](#endnote-ref-145)
156. noms d’Allah : ألفاظ الجلالة   [↑](#endnote-ref-146)
157. tafssīr: تفسير [↑](#endnote-ref-147)
158. Un point d'un système limitant les performances globales. [↑](#footnote-ref-11)
159. Une manière de construire une application pour les systèmes distribués. [↑](#footnote-ref-12)
160. Nujoum al-furkān fī atrāf al-coran : نجوم الفرقان في أطراف القرآن [↑](#endnote-ref-148)
161. Mu3jam des ayas du Coran : معجم آيات القرآن  [↑](#endnote-ref-149)
162. Houssine nassar : حسين نصار [↑](#endnote-ref-150)
163. Le mu3jam indexé des mots du Saint Coran : المعجم المفهرس لألفاظ القرآن الكريم [↑](#endnote-ref-151)
164. Mohammed Fouad Abd El-bāki : محمد فؤاد عبد الباقي [↑](#endnote-ref-152)
165. Les pronoms : الضّمائر [↑](#endnote-ref-153)
166. Les particules : الأدوات (في اللّغة) [↑](#endnote-ref-154)
167. Mohamed Hussein Heikal : محمد حسين هيكل [↑](#endnote-ref-155)
168. L’Académie de la Langue Arabe au Caire : مجمع اللغة العربية في القاهرة [↑](#endnote-ref-156)
169. Mokhtar Fawzi Al-na’āl: مختار فوزي النعال [↑](#endnote-ref-157)
170. Flexionnelles : إعرابيّة [↑](#endnote-ref-158)
171. L’abrogation : النّسخ [↑](#endnote-ref-159)
172. La métrisation en sciences du Coran : الإتقان في علوم القرآن للسيوطي [↑](#endnote-ref-160)
173. La preuve en science du Coran : البرهان في علوم القرآن للزركشي [↑](#endnote-ref-161)
174. Les sources de gratitude : مناهل العرفان للزرقاني [↑](#endnote-ref-162)
175. Les termes de sciences du Coran : مصطلحات علوم القرآن لأنور الباز [↑](#endnote-ref-163)
176. Le mu‛jam de sciences du Coran : معجم علوم القرآن لإبراهيم محمد الجرمي [↑](#endnote-ref-164)
177. Fins des ayas : رؤوس الآيات [↑](#endnote-ref-165)
178. La foi : العقيدة [↑](#endnote-ref-166)
179. Le répertoire indexé des ayas  du Coran : الدليل المفهرس لآيات القرآن الكريم [↑](#endnote-ref-167)
180. L’index coranique : الفهرس القرآني [↑](#endnote-ref-168)
181. Les clés des sourates : مفاتيح السّور [↑](#endnote-ref-169)
182. Invariable : مبني [↑](#endnote-ref-170)
183. Fléchi: مُعرَب [↑](#endnote-ref-171)
184. Flexion : إعراب [↑](#endnote-ref-172)
185. Le guide du confus aux ayas du Coran : مرشد الحيران إلى آيات القرآن [↑](#endnote-ref-173)
186. Noyau : جذع الكلمة [↑](#endnote-ref-174)
187. Vocalisé : مشكّل [↑](#endnote-ref-175)
188. ikhfā’ (الإخفاء) et idghām (الإدغام) : des règles de tajwīd appliquées quand le noun as-sākina ou le tanwīn est suivi par certaines lettres. [↑](#footnote-ref-13)
189. Vocalisé : مشكول [↑](#endnote-ref-176)
190. Chedda : الشّدّة [↑](#endnote-ref-177)
191. Les diacritiques : الحركات [↑](#endnote-ref-178)
192. Hamzä : همزة [↑](#endnote-ref-179)
193. Tâ’ marbûta : تاء مربوطة [↑](#endnote-ref-180)
194. Hâ’ : هاء [↑](#endnote-ref-181)
195. Alif maqsûra : ألف مقصورة [↑](#endnote-ref-182)
196. Yâ’ : ياء [↑](#endnote-ref-183)
197. Littérales : حرفيّة [↑](#endnote-ref-184)
198. Nominaux : إسميّة [↑](#endnote-ref-185)
199. Déclinaison : اشتقاق (الأسماء) [↑](#endnote-ref-186)
200. Unité abstraite : وحدة معنويّة [↑](#endnote-ref-187)
201. Particule : حرف (أداة في اللّغة) [↑](#endnote-ref-188)
202. Prépositions : حروف الجرّ [↑](#endnote-ref-189)
203. Conjonctions de coordination : حروف العطف [↑](#endnote-ref-190)
204. Les pronoms relatifs: الضمائر المتّصلة [↑](#endnote-ref-191)
205. Lâm : لام [↑](#endnote-ref-192)
206. Chamsiyya : الحروف الشّمسيّة [↑](#endnote-ref-193)
207. Flexionnelle : إعرابيّة [↑](#endnote-ref-194)
208. Particules : (في اللّغة) حروف [↑](#endnote-ref-195)
209. Pronom : ضمير [↑](#endnote-ref-196)
210. Voir le chapitre 2 (IV.D.9 Respecter les différences de Hamza et IV.D.10 Respecter la différence entre Hâ’ et tâ’ marbûtä, entre Yâ’ et Alif maqsûrä) [↑](#footnote-ref-14)
211. TF\*IDF (term frequency-inverse document frequency) : une méthode de pondération qui permet d'évaluer l'importance d'un mot par rapport à un document extrait d'une collection. [↑](#footnote-ref-15)
212. Distordue : محرّفة [↑](#endnote-ref-197)
213. Constructions syntaxiques (Boilerplate code) : section de code qui doit être inclue dans de nombreux endroits avec peu ou pas d'altération, le plus souvent utilisé dans les langues ‘verbeux’, c.à.d. le programmeur doit écrire beaucoup de code pour faire des travaux minimes. [↑](#footnote-ref-16)
214. Bibliothèque de moteur de recherche libre écrit en Java qui permet d'indexer et de rechercher du texte. [↑](#footnote-ref-17)
215. Par exemple : Muslim Web, Quranic Corpus, Guided Ways, Quran.com, Quran Reader, Zekr, Noor, alQuran, Quran Verses … [↑](#footnote-ref-18)
216. Le script standard (imlāi) : الخط الإملائي [↑](#endnote-ref-198)
217. Tatwīlتطويل ou Kashida : un glyphe utilisé pour mettre un allongement dans le mot. [↑](#footnote-ref-19)
218. API arabe de segmentation et segmentation légère pour python, <http://pypi.python.org/pypi/Tashaphyne> [↑](#footnote-ref-20)
219. Les analyseurs morphologiques : المحللات الصرفية [↑](#endnote-ref-199)
220. Pattern : وزن الكلمة العربية [↑](#endnote-ref-200)
221. Arabeyes.org : un projet qui vise à supporter pleinement la langue arabe dans l'environnement Unix / Linux [↑](#footnote-ref-21)
222. Zekr.org: logiciel multiplateforme et open source pour explorer et rechercher dans le Coran [↑](#footnote-ref-22)
223. URL : http://versebyversequran.com [↑](#footnote-ref-23)
224. vBulletin (ou vB) : un logiciel de forum de discussions commercial développé par Jelsoft Entreprises Ltd. Écrit en PHP et utilisant la base de données MySQL. [↑](#footnote-ref-24)
225. Les noms d’Allah  : ألفاظ الجلالة [↑](#endnote-ref-201)
226. Ayna, qui signifie «où» en arabe, est le premier moteur de recherche arabe apparu. [↑](#footnote-ref-25)