République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab, Blida
USDB

Faculté des sciences
Département d'informatique
Master informatique (LMD) ingénierie de logiciel

Réalisation d'un concordancier pour la langue Arabe

Présenté par :

BENAZOUT ABDELHADI

BENHAOUA MANSOUR

promoteur :

MR. NEHAL Djilali

Encadreur :

Dr. MAMMERI Mahmoud Fawzi

Juin 2016

Nous louons allah, notre guide et notre force et la raison de notre existance.

A mes chers parents ma mére et mon père

pour leur patience, leur amour, leur soutien et leurs encouragement.

Un grand merci spécial à mon frère jumeau qui m'a beaucoup encouragé tout au long de mes épreuves

A ma sœur qui m'a encouragé de loin

Un grand merci à mon binôme qui a été toujours optimiste

A tout ce que j'aime

Paroles qui me tiennent à cœur :

إن الحياة علم و عمل******* والعمل فريضة منذ الازل

أتنسى ؟ قول الله أن قال******(إقرأ بسم ربك الذي خلق)

Bénazout Abdelhadi ...

الحمد لله وحده , حمدا يكافئ نعمه يوافي مزيده

في خطواتنا الأخيرة في رحاب الجامعة لابد لنا من وقفة نعود بها الى أعوام

قضيناها مع أساتذتنا الكرام , وقبل أن نمضي أتقدم بوافر الشكر

والعرفان والتقدير الى الذين حملوا أقدس رسالة في الحياة

.. لمن علمني حرفا ..

لجميع الأساتذة الأفاضل

إهداء

أهدي هذا العمل الى من ربتني وأعانتني بالصلوات

والدعوات ,الى والدتي المباركة

الى من عمل بكد في سبيلي و علمنني معنى الكفاح وأوصلني الى ما

أنا عليه أبي الكريم أدامه الله لي وجزاهم الله خير جزاء

والى افراد أسرتي , سندي في الدنيا ولا أحصي لهم فضل

والى القلوب الطاهرة والنوايا الصادقة الى أحبابي وإخواني من

دون إستثناء من عرفت كيف أجدهم و علموني أن لا أ ضيعهم

الى أساتذتي الكرام ورفيقي في العمل وكل رفاق الدراسة

منصور

ملخص

اختراع الحاسوب وتطور برمجياته جاء تزامنا مع التطور الكبير الحاصل في مجال البحث العلمي وتفجر المعلومات وتطور تقنيات الاتصال، اللغة العربية كغيرها من اللغات حظيت باهتمام التربويين والخبراء في سبيل نشرها وتعليمها ودراستها بواسطة الرقمنة .

وتأتي أهمية رقمنة تعليم العربية بوصفها اللغة الأكثر استعمال سواءا كانت كلغة ثانية أو أجنبية وتتجلى أهمية نشر هذه اللغة والاهتمام المتزايد من طرف الراغبين في تعلمها وملاحقة التطور المتعاظم والمتسارع في توظيف الحاسوب في التعليم بصورة عامة وفي تعليم اللغات الثانية والأجنبية بصورة خاصة باستخدام البرمجيات التعليمية.

البرمجيات التعليمية هي برامج لا يتطلب استخدامها خبرة حاسوبية متخصصة، حيث يستخدمها المتعلمون إما لاكتساب مهارة جديدة أو التدرب على مهارة مكتسبة سابقة، ويتم تصميم هذه البرمجيات وفق أنماط متعددة، على أسس تعليمية وتربوية و دراسات رسمية سواءا كانت أبحاث أو دراسات.

إن هدفنا في هذه الأطروحة هو تقديم برنامج المفهرس , الذي يخص المهتمين في تعلم اللغة العربية و الباحثين فيها من الخبراء والعارفين لها لتوسيع معرفتهم لهذه اللغة ودراسة تطورها انطلاقا من« مدونة » للباحثة لطيفة السليطي عن جامعة ليتز .

لا نستطيع إنكار حقيقة أن إنشاء هذا البرنامج ليس بالهين.حيث يكمن المشكل في معالجة اللغة الطبيعية لكون اللغة صعبة في هذا المجال , ولحل هذا المشكل يتطلب دراسة العوائق التي تؤثر على البرمجيات المتواجدة حاليا حيث يكمن الحل في استعمال خصائص اللغة والصرف والنحو الذي يساعد على اقتراح حلول لإنتاج برنامج أفضل.

Résumé

L'invention de l'ordinateur et son développement de logiciels a coïncidé avec le grand développement dans le domaine de la recherche scientifique. Avec l'explosion de l'information, et le développement des technologies de la communication, la langue arabe comme d'autres langues a reçu l'intérêt des éducateurs afin de la diffuser et de l'enseigner par ordinateur.

L'importance d'informatiser l'enseignement de l'arabe en tant que langue secondaire ou étrangère reflète dans l'importance de la diffusion de cette langue, la croissance d'intérêt de vouloir l'apprendre, et poursuivre le développement rapide de plus en plus dans l'utilisation de l'ordinateur généralement dans l'enseignement, et particulièrement dans l'enseignement des langues secondaires et étrangères en utilisant des logiciels éducatifs.

Les logiciels éducatifs sont des logiciels qui ne nécessitent pas une expérience spéciale dans l'informatique pour pouvoir l'utiliser, les apprenants l'utilisent soit pour acquérir une nouvelle compétence, soit pour pratiquer une compétence déjà acquise, ce genre de logiciels sont conçu par multiples styles, Pour des raisons scientifiques, pédagogiques et techniques.

Notre but est de fournir un logiciel soit « concordancier », qui concerne les intéressés d'apprendre la langue arabe, ou ceux qui sont intéressés à élargir leur connaissance de cette langue, tel que les chercheurs ou les experts linguistiques, à partir d'un corpus de la chercheuse Latifa al-solaiti élaboré au sein de l'université Leeds.

Nous ne pouvons pas nier, le fait que la mise en place d'un tel programme n'est pas facile, d'où le problème réside dans le traitement automatique de la langue arabe, car elle est difficile dans ce domaine, pour résoudre ce dernier, cela exige d'étudier les obstacles qui affectent les logiciels existant, la solution se demeure sur l'utilisation de l'analyse morphologique, ce dernier aide a supposé des meilleures résultats qui mène à un meilleur produit.

Mots clés

Langue arabe, corpus, concordancier, analyse morphologique, recherche scientifique, TALN arabe

Abstract

The invention of the computer and its software development coincided with the great development in the field of scientific research.With the explosion of information, and the development of communication technologies, the Arabic language as other languages received the interest of educators to disseminate and teach it by the computer.

The importance of computerize the teaching of Arabic as a second or foreign language reflected in the importance of the dissemination of this language, and the growth of interest who want to learn it, and carry on the quick development of growing in the use of computer in education generally, and particularly in education and secondary languages using educational software.

Educational software are software that does not require special experience in the computer to use it, learners use it to acquire a new skill or to practice a skill already acquired, this kind of software is designed for multiple styles, for scientific, educational and technical

Our goal is to provide software which is a concondencer that concerns who are interested to learn the Arabic language or those who want to spread their knowledge in this language, by using a corpus created by the researcher Latifa al-solaiti from university of leeds.

We can not deny the fact that the establishment of such program is not easy, where the problem lies in Automatic processing of the Arabic language because it is difficult in this domain, to solve this one, the solution lies on the use of morphological analysis, this one help to suppose best results that leads to a better product.

Keywords

Arabic language, corpus, morphological analysis, concondencer, scientific research, arabic NLP

Table des matières

Liste des figures

Liste des tableaux

Introduction général

Introduction général

Avec l'expansion des travaux sur les corpus, et particulièrement la disponibilité de beaucoup de corpus en libre accès, l'utilisation de ces derniers a gagné de l'espace dans l'enseignement des langues. L'importance de l'utilisation des corpus vient du fait que les enseignants utilisent de moins en moins des données langagières créées par des linguistes au profit de données authentiques.

L'objectif d'un corpus sera ainsi de l'utiliser avec l'aide d'une application informatique pour faire l'investigation de la structure de la langue et d'extraire d'autres types d'information pour les objectifs d'enseignement et de recherche. L'une des activités courantes est l'utilisation de la notion de concordance. C'est une technique à travers laquelle les apprenants dans une langue peuvent chercher et classer des données pour obtenir certains types d'information.

La généralisation des logiciels de concordances et l'accès facile à des corpus numériques a permis le développement de la linguistique de corpus au cours des 25 dernières années. Actuellement les méthodes utilisées pour la réalisation des concordances ne donnent pas de résultats satisfaisants sur l'arabe. Ceci est dû au fait que l'arabe est hautement flexionnelle, agglutinante et non vocalisée. Elle contient, de plus, des formes graphiques complexes muettes aux recherches de surfaces. Aujourd'hui le concordancier en tant qu'outil d'exploration en contexte s'avère d'une utilité inéluctable pour le fonctionnement de certaines applications linguistiques :

- « C'est un composant indispensable dans tout système d'enseignement assisté par ordinateur. » (Zaafrani, 2002)
- C'est un composant indispensable pour l'exploitation des corpus.
- « La définition des traits sémantiques associés aux mots nécessite inévitablement une exploration contextuelle des alentours du mot et des usages dans la langue. » (Abbes, 2004 : 23)

Notre travail dans ce projet consiste en la réalisation d'un concordancier. Un concordancier est un outil d'exploration de corpus de textes. Cette application doit fournir à l'utilisateur des fonctions basiques telles que la recherche des occurrences d'un mot, sa catégorie grammaticale, ses concordances, des informations sur l'origine du texte source telles

que le titre et l'auteur. Il permet d'effectuer des recherches sur des mots ou des expressions et d'afficher toutes les occurrences avec un extrait de leurs contextes.

Pour le test et la validation de notre concordancier, nous avons fait appel à un corpus arabe développé à l'université de Leeds par Latifa Al-Solaiti (Al-Sulaiti et al. 2003). Il s'agit d'un corpus annoté en XML et en libre accès. Les textes de ce corpus ont été principalement dérivés à partir de sites Web. Pour ce faire, l'auteure du corpus à procéder à l'identification de plusieurs sites potentiels, ensuite à la récupération manuelle des textes écrits. L'auteure a également inclus certains fichiers correspondants à des séquences orales qu'elle a obtenus à partir de la Radio Qatar. Néanmoins, le nombre de ces fichiers reste faible dans le corpus et ceci est dû au fait que la manipulation et le traitement de ce type de fichier consomment énormément de temps. (AL-Sulaiti et al. 2006)

De plus, les textes qui composent les corpus sont le plus souvent enrichis (ou annotés) avec une variété d'informations. L'annotation d'un corpus peut concerner aussi bien des informations linguistiques qu'extra linguistiques. Le corpus de Latifa Al-Solaiti ne contient aucune information grammaticale sur les mots du corpus telle que la catégorie grammaticale, le genre, le nombre, la personne ou cas grammatical, et n'est donc annoté qu'avec des informations extra linguistiques telles que des informations sur l'auteur du texte, la publication, etc. L'objectif de ce projet est double : principalement, il s'agit de (i) réaliser un concordancier qui fouille dans le corpus de Latifa Al-Solaiti et qui propose les fonctions les plus usuelles, cette tâche elle-même a besoin d'un autre module réclamé par la plupart des applications en TALN qui n'est autre qu'un (ii) analyseur morphologique qui d'une part sera utilisé par le concordancier lui-même lors des recherches des différentes occurrences d'un mot donné et d'autre part par une future annotation grammaticale du corpus.

Ce travail s'intègre dans le cadre d'un projet pour le traitement automatique de la langue arabe qui se résume en la construction d'un corpus arabe annoté. Ce corpus servira dans plusieurs domaines en particulier l'enseignement et l'apprentissage de la langue et la recherche en linguistique. Dans l'enseignement, ce type d'outils permet à l'enseignant de préparer des textes de lecture ou des exemples pour des illustrations ou des exercices d'application. Pour un mot donné, nous pouvons chercher ses différentes cooccurrences avec certains autres mots ou les différents contextes où il peut apparaitre. En analyse linguistique, nous pouvons, par exemple, étudier le comportement du syntagme nominal arabe en analysant les différentes instances du syntagme nominal à travers, non pas d'exemples confectionnés

par un spécialiste d'une manière artificielle, mais plutôt d'exemples de syntagmes nominaux authentiques utilisés dans des situations langagières attestées.

Le mémoire se réparti en trois chapitres. On commence tout d'abord par un état de l'art, on y détaillera tout ce qui concerne la notion de concordancier, les différents outils qui permettent sa construction ainsi que les insuffisances des concordanciers existants. Le deuxième chapitre sera consacré à l'étude de la langue arabe et plus particulièrement la morphologie arabe. Il sera question du mot arabe, sa composition ainsi que l'analyseur morphologique que nous utiliserons dans notre concordancier. Enfin le troisième chapitre sera une démonstration des fonctionnalités de notre programme, accompagné avec des séries de test pour vérifier et valider notre concordancier.

Chapitre 1

Etat de l'art

Chapitre 1 : État de l'art

1. Introduction

L'amélioration de la qualité d'apprentissage des langues semble très important de nos jour, en particulier quand il s'agit d'enseigner une langue étrangère, apprendre une nouvelle langue permet au apprenant d'acquérir tout un nouveau champ de compétence, de décoder une nouvelle phonologie, appréhender des règles de grammaire profondément différents, et enfin rendre visible des concepts parfois cachés dans la langue cible.

« L'enseignement et l'apprentissage d'une langue étrangère visent normalement, avec des variations selon la langue et les buts de l'apprenant, quatre compétences principales : lire, écrire, parler, écouter. Dans le cadre de l'apprentissage des langues assisté par ordinateur, la lecture des textes joue un rôle important pour deux raisons. Du côté pratique les systèmes de traitement automatique de la langue sont actuellement bien équipés dans ce sens (beaucoup plus que pour le traitement de la parole), et du côté pédagogique la lecture permet à l'apprenant de bien comprendre et développer le vocabulaire de la langue cible.

C'est évident pour apprendre une langue reviens sur apprendre ses mots, pour faciliter cette tache la naissance des concordanciers a fait lieu, c'est des logiciels de concordance qui représente un outil de référence très utile aux linguistes, qui permet de faire la recherche dans un corpus d'un mot accompagné de son contexte. Or, L'utilisation des concordances dans l'enseignement ou l'apprentissage de langues étrangères est une pratique commune. Ceci s'est développé dans un domaine de la recherche distinct pour des pédagogues de langue. » (El mezouar et al. 2013)

Le chapitre sera organisé en trois sections. Dans la section 1, nous introduirons la notion de concordance pour avoir une idée de ce qui a causé l'apparition des concordanciers. Dans la section 2, nous aborderons les concordanciers qui sont les applications du TALN[6] (Traitement automatique du langage naturel) par excellence pour la recherche automatique de concordances qui répond aux besoins des linguistes, et effectue des recherches de concordances qui se base essentiellement sur des masses importantes de données, l'objectif de la dernière section est d'introduire la notion de corpus, qui représente un outil d'entré indispensable, avec qui on s'aide a trouvé les résultats souhaité.

6 Traitement automatique du langage naturel

2. Concordance

2.1 Définition

« Historiquement, une concordance est définie comme une liste ordonnée alphabétiquement de tous les mots w d'un texte T. A chaque mot w de T sont rattachés certaines informations à propos de sa position et du contexte dans lequel il apparaît dans T. » (Slaby, 1979), En lexicographie, (Larousse, 1994 : 108) à affirmer que la concordance est un index de mots présentés avec leur contexte. Une fois réalisée, l'indexation des mots d'un texte, d'un auteur, d'une époque fournit des renseignements sur les références des mots et éventuellement sur leur fréquence ; on offre à l'utilisateur la possibilité d'étudier parallèlement les divers emplois du même vocable, pour (Langlois, 1996) l'avènement de l'informatique a influencé sur la forme des concordanciers, il a certifié : « Il n'est plus le livre volumineux, dont la réalisation est très coûteuse en temps et en ressources humaines, il est muté à un logiciel d'extraction de concordance 'au cas par cas, selon les besoins 'de n'importe quel texte écrit. »

2.2 Les concordances dans la littérature

« L'histoire des concordances dans la littérature et l'analyse linguistique a commencé bien avant la naissance de l'informatique. La première concordance était établie au 13e siècle pour la bible. » (Sekhraoui, 1995)

La première concordance de la bible latine date du 16e siècle. Ses bases théoriques ont été mises au point en 1736 par l'écossais Alexander Curden en association avec un licencié du Marischal College (une des deux universités d'Aberdeen). Par la suite Hugo de San Charo a engagé 500 moines pour sa réalisation. (Dundee, 2004) (Tribble et Jones, 1997)

La concordance de la bible a été conçue comme un instrument de travail pour permettre aux exégètes de repérer tout de suite certains thèmes, pour mieux préparer leurs sermons ou leurs commentaires et le cas échéant pour combler des trous de mémoire. (Cameron, 1996)

Les historiens de la langue, les grammairiens, les sociologues… ont vu que les concordances leurs offraient une mine d'informations profitables dans l'étude de tout œuvre, c'est pourquoi ils ont élargie son usage à d'autres textes littéraires d'envergures, comme la concordance de Sheakspeer en 1787et Horas 1916.

Du côté de la langue arabe, l'histoire des concordances a commencé bien plus tard. La pionnière des concordances était celle du Coran, établie par Mohamed Saîd Mustafa en 1811,

publiée à Calcutta sous le titre de «نجوم الفرقان». La seconde concordance a été publiée à Leipzig en 1842 par Gustave Flugel sous le titre de «نجوم الفرقان في أطراف القرآن ».

Ces tentatives ressemblaient plus à un index général qu'à un concordancier. La première véritable concordance du Coran n'est apparue qu'en 1859 sous le titre de « concordance complète du Coran » publiée par Mirza A. Kazem-Bek à St. Petersbourg. Enfin, ces efforts ont été couronnés en 1945 par une nouvelle concordance publiée au Caire par Abdel Baki Mohammed Fouâd intitulé « المعجم المفهرس لألفاظ القرآن الكريم.»(Abbes, 1999)

Dans leurs entreprises de réalisation du concordancier du Coran, les auteurs ont présenté leurs œuvres selon deux modèles. Le premier est inspiré des concordanciers latins de la bible. Il se présente sous la forme d'une liste ordonnée alphabétiquement selon la première lettre de chaque mot. Le second est propre à la langue arabe, il consiste à regrouper tous les mots sous leurs racines et présenter une liste ordonnée alphabétiquement selon la première lettre de chaque racine.

2.3 Les concordances électroniques

L'élaboration de concordances consiste à rechercher dans un texte toutes les occurrences d'un mot ou d'un autre motif linguistique, puis à les présenter, une par ligne, chacune dans son contextes. Les applications relèvent de la lexicographie, de l'apprentissage des langues et de l'exploitation de bases de données littéraires. L'élaboration du dictionnaire Cobuild (Cobuild, 1987), par exemple, a systématiquement fait appel à la recherche d'exemples dans des concordances. (Sinclair, 1991)

2.4 Le programme de concordance

Les fonctionnalités que l'on peut attendre des concordanciers les plus récents sont les suivantes : (Rezeau, 2008 : 3)

- la création de concordances de type KWIC (Key Word In Context)
- le tri en ordre alphabétique à droite ou à gauche du mot recherché

- l'obtention de listes de tous les mots d'un texte, avec tri en ordre alphabétique ou en ordre de fréquence, et
- l'exportation des résultats obtenus pour exploitation ultérieure dans un traitement de texte.

2.5 Concordances KWIC

KWIC est l'acronyme de Key Word In Contexte (mot clef en contexte.), le format le plus commun pour des lignes de concordance. Ses lignes sont représentées sous format d'une liste de toutes les occurrences d'un ou plusieurs mots ou expressions, alignées verticalement en colonne, accompagnées de leur contexte droit et gauche. (Laporte, 2009 : 1)

Dans l'exemple suivant, le mot « الذي » apparaît dans la colonne centrale. À partir d'un corpus de textes, on peut demander la visualisation d'une concordance en particulier (comme « الذي» dans notre exemple), ou bien demander les occurrences ; lorsqu'on les visualise de la manière suivante, on appelle cette visualisation « un mot clef en contexte », ou un « KWIC ». Les contextes gauche et droit correspondent à un découpage linéaire du mot clef en contexte avec ce qui le précède ou le suit pour une ligne donnée de texte.

انتشر تي دمه، تتتي الذي تمؤلم مع مرض السرطان
يصعب على أي مثقث الذي تيث يقاتل الصهيتنية بالمنطق
يبدأ بطبيعة الالتزام تينتتح الذي الحدث تبين الأساس التتري
صدر منذ حتالي ربع الذي "تقد أثر تتابه "الاستشراق
نهش جسد إدتارد سعيد الذي إلى أن سرطان الدم
عذبت قسمات تجهه مطارات الذي إلا ذلث اللاجئ التلسطيني
لم تستطع الغربة محث الذي ترتعدان لأحرث هذا المقدسي
أطلق أتل صرخة عند الذي الآلام عن أقدام الطتل

Figure 1 : *contexte du mot « الذي »*

« La réalisation de la concordance par la méthode de KWIC reste relativement simple, mais le problème de la concordance réside essentiellement dans la méthode de recherche des mots dans le corpus en réponse aux formes demandées par l'utilisateur.» (Garrigues, 1997)

La méthode de recherche d'un mot peut-être (Laporte, 2000) :

i- une simple forme de surface, comme *résolue*, ou une séquence de motif type,
ii- une expression rationnelle sur les lettres, comme réso*, qui reconnaîtra tout mot commençant par « réso », ou une séquence de motifs de ce type, ou
iii- un motif défini par des critères linguistiques : un lemme (<résoudre> pour reconnaître toutes les formes fléchies de ce verbe), et une catégorie grammaticale.

Les motifs de type (i) et (ii) peuvent être confrontés aux occurrences de mots du texte par des fonctions d'appariement des chaînes de symboles. Ils peuvent rendre service, en particulier dans les langues à morphologie pauvre, comme l'anglais. Dans le cas des langues romanes, ils ne permettent pas de simuler les motifs du type (ii) : par exemple, le motif réso* reconnaître aussi bien *résolution* et les formes du verbe *résonner* à celles de *résoudre*. Les concordanciers du commerce se rangent dans ces catégories. Les interrogations du type (i) et (ii) sont possibles à l'aide du KWIC. Le troisième type est plus efficace mais nécessite un étiquetage lexical préalable (Silberztein, 1993)

3. Concordancier

3.1 Introduction

Les concordanciers ou logiciels de concordances ont été considérés parmi les outils les plus simples en ce qui concerne la facilité d'utilisation et les plus puissants en termes de quantité des informations fournies dans le domaine de la linguistique des corpus.

« Ces dix dernières années ont vu l'émergence concomitante de programmes informatiques permettant la recherche rapide des collocations des mots d'une langue d'une part et de corpus de textes sous forme informatisée d'autre part. Le chercheur en linguistique dispose ainsi des outils et des données nécessaires pour mettre en évidence et analyser les propriétés distributionnelles des mots du point de vue de leur combinatoire sémantique, syntaxique et discursive. Quant à l'enseignant de langues sur le terrain, il peut utiliser ces mêmes outils en mettant à la disposition de ses étudiants un type d'apprentissage s'appuyant sur des données authentiques, ce que Tim Johns (université de Birmingham) appelle *Data Driven Learning*. » (Rezeau, 2008 : 2), de nos jour les programmes informatiques facilite clairement l'enseignement, quand il s'agit d'une nouvelle langue le concordancier représente un outil primordial afin de rendre l'effort plus délicat, surtout pour les chercheurs, c'est pour cela « l'utilisation d'outils informatiques tels que les concordanciers, qui permettent d'extraire des collocations de mots et leur contexte, est fréquente dans la communauté des chercheurs en sciences du langage. La taille de corpus actuellement disponibles est en augmentation constate, de même que les schèmes d'annotation, ce qui entraîne la création de nouveaux systèmes pour interroger ces données. » (Barreca et al. ,2004 : 499)

3.2 Synthèse des études précédentes

« Le concordancier a toujours été un outil de grand intérêt pour l'analyse des contextes d'emploi des lexèmes dans des corpus oraux ou écrits. Ses applications théoriques et pratiques sont nombreuses, et intéressent plusieurs disciplines : philologie, littérature (Caballero, 1999) (Magri-mourgues, 2006), syntaxe et sémantique (Gross, 2000), traduction (Jacquet-pfau, 1994) et didactique des langues (Tognini-bonelli, 2001). Pourtant, comme le soulignent (Pincemin et al. 2006), la plupart des concordanciers disponibles n'exploitent pas les informations linguistiques introduites par l'annotation des corpus (p. ex. les catégories morphosyntaxiques, etc.). Les résultats produits par ce genre de concordancier n'aboutissent qu'à de simples collections d'occurrences (p.ex. Lextutor et Lexiqum). » (Barecca et al. p.499), Plusieurs concordanciers existent déjà, dans ce qui suit nous allons décrire les concordanciers les plus connus, tout en sachant qu'il existe deux type : nous nous intéressons a des concordanciers qui prennent en compte le traitement de l'arabe et les concordanciers qui proposent des services en lignes.

3.2.1 Concordancier de l'arabe

Parmi les concordanciers les plus connus qui traitent des corpus arabe :

3.2.1.1 MonoConc

Développé par le linguiste Michael Barlow (Barlow, 1998) pour les chercheurs et les professeurs de langue, MonoConc est un concordancier commercialisé qui a été publié par Athelstan. Son utilisation est relativement simple, il établit des listes de concordances en se basant sur la spécification d'un certain nombre de paramètres de recherche.

MonoConc offre un ensemble limité d'options de recherche (mots simples, lexèmes, ou des séquences de mots), de plus, il ne traite que des textes précédemment chargés. Toutefois, travailler sur des textes du disque local nécessiterait un temps de chargement à chaque fois.

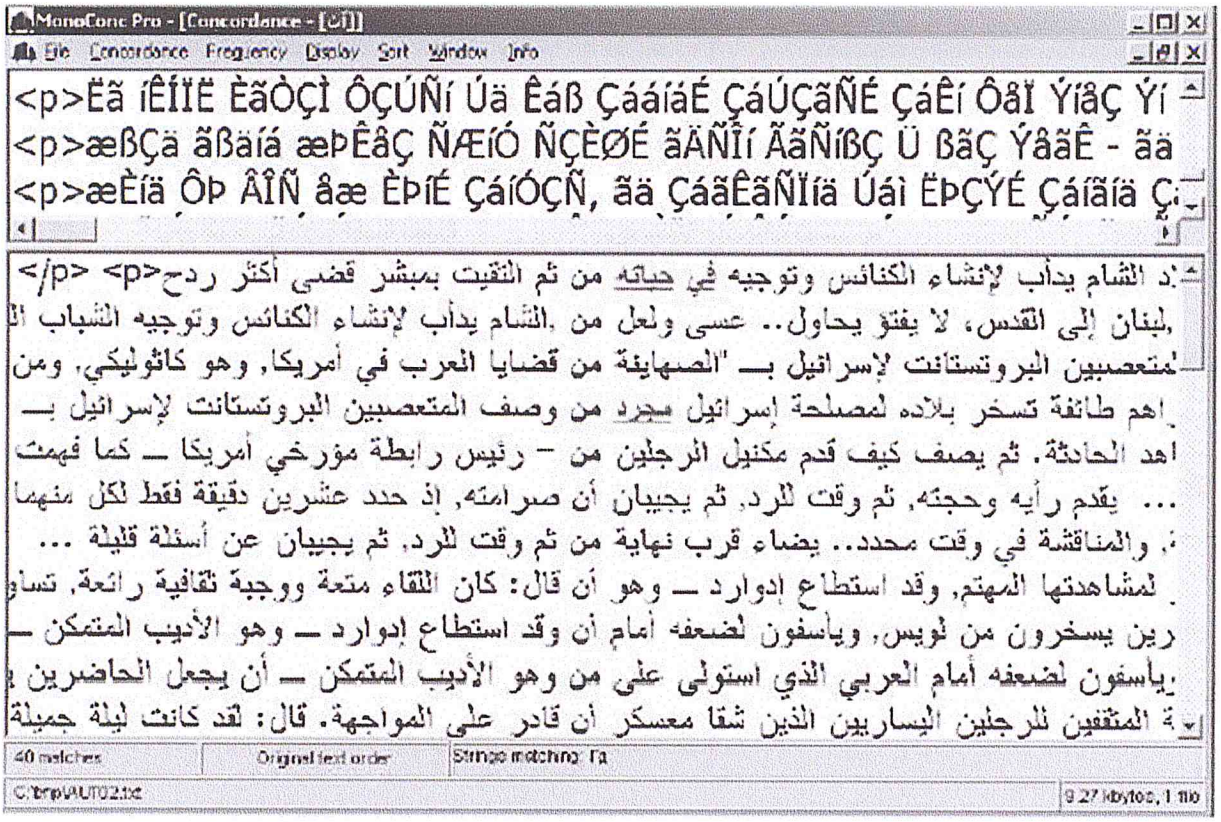

Figure 2 : *Table de concordance de MonoConc*

La figure ci-dessus nous permet de voir clairement les insuffisances de MonoConc qui sont :

- L'apparition des balises lors de l'affichage des contextes
- Inversion de l'ordre d'affichage des contextes droit et gauche autour de la séquence recherchée.
- Affichage erronée des caractères arabes en haut de la fenêtre de concordance.
- La recherche de la particule du subjonctif « أن » a donnée naissance à une reconnaissance de la préposition « من », ce comportement semble être aléatoire car il n'est pas reproductible pour toutes les requêtes.

3.2.1.2 WordSmith

Le programme WordSmith a d'abord été publié en 1996, il a été réalisé à l'université de oxford et est toujours développé par le linguiste Mike Scott (Scott, 2008), il est actuellement à la version 6.0 il inclut trois outils : liste des mots, liste des mots-clés et un concordancier, sa version de démonstration est disponible en ligne[5], affiche correctement les caractères Unicode mais il limite le nombre de résultats lors de l'exécution de requêtes.

[5] La version de démonstration est téléchargeable à l'adresse : http://www.lexically.net/wordsmith

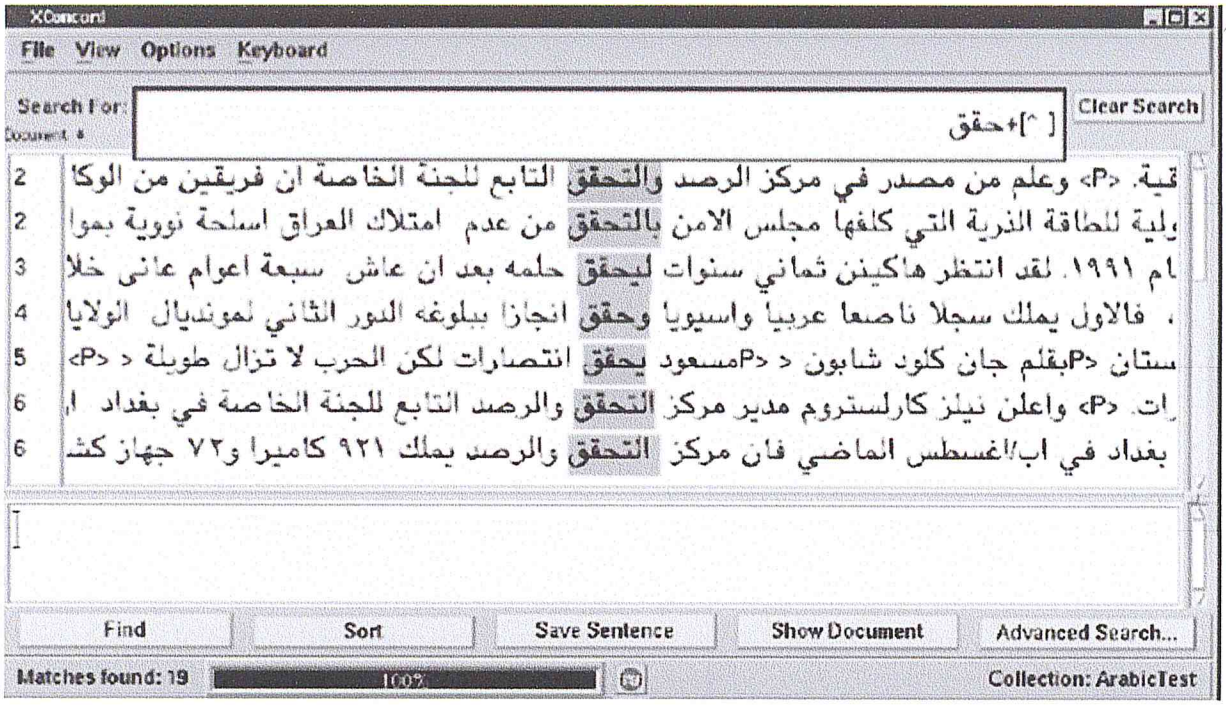

Figure 4 : *Table de concordance de xConcord*

L'exemple proposé sur la figure ci-dessous permet de retrouver tous les mots se terminant par une chaine de caractères, soit « حقق ».

3.2.2 Concordanciers en ligne

Quant aux concordanciers en ligne, nous citons :

3.2.2.1 GlossaNet

GlossaNet est un moteur de recherche spécialisé, pour les linguistes c'est un observateur en même temps, Il vous permet de faire des recherches dans tous les textes publiés sur Internet en forme RSS-feeds : Presse, médias, blogs, forum, etc. Il applique automatiquement les requêtes de l'utilisateur et procède par l'envoi de courriers électroniques contenant les concordances (Fairon, 2001). Il permet aux utilisateurs de rechercher des mots ou des séquences de mot en se basant sur l'utilisation d'un ensemble de ressources linguistiques.

3.2.2.2 WebCorp

WebCorp « the web as corpus » vous permet avoir accès au web comme un corpus, il offre une large gamme d'option de filtrage et de formatage (y compris kwic –keyword in context 'mot-clef en contexte'). Le concordancier repose sur différents moteurs de recherche tels que Google et AltaVista, auxquels il ajoute la possibilité de raffiner les requêtes. Parmi ses limites

le temps de réponse du concordancier car la possibilité de rechercher sur l'ensemble des documents sur le web engendre de sérieux problèmes.

Ces concordanciers décrits, nous a permis de voir quelques différentes méthodes et fonctionnalités existantes. L'intérêt d'une telle démarche étant de pouvoir positionner notre travail par rapport à ceux-ci et de pouvoir profiter des avantages et certainement d'éviter les insuffisances.

3.3 Fonctionnement d'un concordancier

Le concordancier est un moteur qui cherche des mots et des expressions dans un corpus de textes. Lorsqu'on lui soumet une requête, le système fouille dans le corpus et affiche toutes les occurrences trouvées dans des listes alignés verticalement en colonne, accompagnés de leurs contextes droits et gauches.

Tout simplement, « Les concordanciers peuvent être utilisées pour :

- ✓ Effectuer une recherche pour les occurrences d'une chaîne de caractères
- ✓ comparer les divers emplois|sens d'un même terme
- ✓ observer la fréquence des mots
- ✓ déterminer la combinaison des unités lexicals
- ✓ identifier des collocations et des définitions
- ✓ observer des propriétés distributionnelles de certains mots » (Mestivier, 2005)

3.4 Intérêt de concordancier :

D'après nos recherches sur le domaine linguistique, nous avons remarqué que les programmes informatiques tels que les concordanciers, sont nécessaires pour rendre le texte à analyser facile à exploiter, « Des linguistes de plus en plus nombreux utilisent des concordanciers pour extraire de textes préexistants des mots ou expressions en contexte.» (Laporte, 2009 :1)

On peut ajouter que la quantité importante des informations à exploiter, joue un rôle important pour le fonctionnement d'un concordancier en terme de temps de réponse, autrement dit, la taille du corpus est un aspect important en ce qui concerne la représentativité des résultats, « L'utilisation d'un concordancier devient particulièrement intéressante à partir du moment où l'on peut effectuer des recherches en temps réel. Pour arriver à ce genre d'interaction sur un corpus de taille arbitraire, il est nécessaire de faire appel à une structuration particulière du

texte. Quel que soit la nature exacte de celle-ci, elle comprendra forcément une forme d'indexation du texte, qui permettra de localiser toutes les occurrences d'une chaîne de caractères donnée, dans un laps de temps qui soit proportionnel au nombre d'occurrences, plutôt qu'à la taille du corpus. » (Simard et al. 1993 :3)

4. Corpus

4.1 Introduction

Comme nous avons déjà cité, Les concordanciers sont des outils informatiques, qui aident à l'analyse de corpus en offrant des fonctions de recherche avancées, on confirme que le corpus est un élément d'entré pour pouvoir utiliser les concordanciers, c'est une sorte de texte collecté qui répond clairement et rapidement au recherche de l'utilisateur, on ne peut pas aussi négliger la réalité que la richesse du corpus joue un rôle au niveau de résultat de la recherche. Ces dernières années, la construction des corpus a été largement augmenté, « Divers types de corpus ont été développés pour différents objectifs tel que la recherche et l'enseignement (Mcenery et Wilson, 1996). Etant une langue internationale et officielle de plusieurs pays, l'Anglais a reçu la plus grande attention au sein de la communauté de la recherche. Il existe plusieurs types de corpus qui ont été construits, non seulement pour enquêter sur les principales variétés de l'anglais, Britannique (Johannson et al 1986) (Aston et Burnard, 1998) et américaine (Francis et Kucera 1979), mais autres variétés telles que l'Australie (Ahmad et Corbett, 1987)Indien (Shastri, 1988), Camerounais (Tiomajou, 1993), et d'autres. Ces corpus contiennent des échantillons représentatifs de texte anglais ; Beaucoup sont également enrichis à l'aide d'analyse linguistique supplémentaire avec des étiquettes de parties du discours (PoS-tags) sur chaque mot montrant sa catégorie grammaticale ou une fonction dans le contexte (Leech et al. 1983).

Les corpus linguistiques anglais ont été utilisés dans le développement de matériel pédagogique en langue anglaise, ainsi que des systèmes de traitement du langage tels que les modules de reconnaissance de la parole, vérificateur d'orthographe et de grammaire, systèmes de dialogue, etc. (Atwell, 1999)

L'arabe est aussi une langue internationale, rivalisant avec l'anglais dans le nombre de locuteurs de langue maternelle (Graddol, 1997). Toutefois, peu d'attention a été consacrée à l'arabe. Bien qu'il y ait eu un certain effort en Europe et U.S.A, qui a abouti à une production

réussite de certains corpus arabe mais malgré ce dernier les progrès dans ce domaine est encore limitée.

En outre des progrès ont été entravés par le manque d'outils efficaces tels que tagueurs, analyseurs morphologique et lecteurs optiques, qui sont nécessaires pour développer et exploiter un corpus arabe (Atwell et al. 2004). » (Thierry et al. 2010)

Enfin, la nécessité de fondé des corpus est obligatoire pour aider à enseigner la réalité de n'importe quelle langue, La linguistique de corpus apparaît comme un excellent moyen d'atteindre cet objectif.

4.2 Définition d'un corpus

D'après nos recherches sur la définition d'un corpus on a observé que le terme *corpus* y fait l'objet d'un consensus large, et son utilisation devient différent en terme de la quantité d'information dont il contient.

Les linguistes concernés par le TALN définissent le corpus comme une grande collection de documents qui doit servir à mettre au point des projets de traitements linguistiques, ayant des fonctions représentatives précises. (Habert, 2005) penche pour une définition plus restrictive de corpus comme celle emprunté à (Sinclair, 1991) Le corpus est pour ce dernier « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage ». (Muller, 1969) le définit, pour des besoins lexico métriques, comme une collection définie de textes.

Dans le *Dictionnaire de linguistique* de (Mounin, 1995 : 89), le corpus est défini comme "ensemble d'énoncés écrits ou enregistrés dont on se sert pour la description linguistique."

La valeur d'un corpus tient à des critères pour une bonne constitution, ces critères semblent essentiellement envisager le corpus sous l'angle de sa construction :

• Conditions de signifiance : Un corpus est constitué en vue d'une étude déterminée (pertinence), portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue (et non sur plusieurs thèmes ou facettes indépendants, simultanément) (cohérence).

• Conditions d'acceptabilité : Le corpus doit apporter une représentation fidèle (représentativité), sans être parasité par des contraintes externes (régularité). Il doit avoir une

ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse (complétude).

• Conditions d'exploitabilité : Les textes qui forment le corpus doivent être commensurables (homogénéité). Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs.

Ces critères semblent essentiellement envisager le corpus sous l'angle de sa construction, ce qui se justifie pleinement pour pouvoir évaluer sa pertinence en situation, On définit le corpus comme un regroupement de données (textes) correspondant à un choix explicitable et intelligible

« Un corpus peut alors se prêter à plus ou moins d'utilisations intéressantes en fonction de la signification qui lui est attachée : les critères factuels et généraux pourront être préférés aux critères subjectifs, et les sélections systématiques aux recueils irréguliers. Par exemple, un corpus de l'ensemble des articles entrés dans une grande base bibliographique l'année dernière sous la rubrique linguistique peut fournir des résultats d'une portée plus large qu'un corpus des articles cités en référence d'une thèse donnée.

Mais même les corpus les plus subjectifs et les plus irréguliers ont droit à l'existence et peuvent faire l'objet d'études scientifiques, En revanche, un corpus (apparemment) dépourvu de principe de constitution, fut-il volumineux et impeccablement codé, est inutilisable, car il n'y a rien à quoi rapporter les résultats des analyses qui y seraient effectuées. En résumé, le mot corpus est présenté comme étant une collection de textes, qui respect des critères définis, concernant un sujet pour servir à une analyse ou description linguistique. » (Condamines et al. 1999 : 26-36)

4.3 L'intérêt des corpus

Nombreux sont les travaux qui illustrent l'intérêt de s'appuyer sur des corpus, de grands corpus textuels informatisés sont constitués dans certaines langues, notamment européennes, et servent de matière première pour des études diverses, on distingue trois intérêts principales :

4.3.1 Fournir un référentiel

Un corpus est un univers de référence, il est particulièrement bien adapté à la linguistique pour l'étude qualitative et quantitative, il donne aussi la matière concrète pour représenter un axe de lecture.

4.3.2 Permettre l'observation

Le corpus peut être considéré comme un « observatoire » qui permet entre autres d'étudier un sujet désiré au plus près de ses évolutions et variations.

4.3.3 Faciliter la recherche

Un corpus permet l'accès à un regroupement de données, l'informatique permet de structurer ces données, et d'en extraire rapidement les aspects pertinents, un corpus numérique permet de traiter de vastes phénomènes sur le web qui de par leur taille ou leur complexité n'étaient pas autres fois accessibles.

4.4 Les différents types de corpus :

On peut trouver le corpus sous forme de support (papier, électronique, oral, vidéo), version langagière, monolingue, bilingue (comparable ou alignés), multilingue, originaux, traductions, locuteurs natifs ou apprenants de la langue, état de la langue (synchronique ou diachronique), but (corpus de référence ou de spécialité),ouvert ou fermé, et présence d'annotation (textes bruts ou annotés)

5. Conclusion

Les concordanciers sont des logiciels qui utilisent des corpus comme une source d'information en offrant des fonctionnalités qui aident les utilisateurs à mieux apprendre une langue.

Beaucoup de chercheurs ou experts en linguistiques ou même des apprenants utilisent systématiquement des concordances pour recueillir des exemples. Cette pratique tend à rendre plus rigoureuse la collecte des exemples, et à orienter l'étude vers les formes réellement en usage. Elle facilite aussi la connaissance de la grammaire des mots, ainsi exploiter les informations d'une manière libre, afin d'explorer l'utilisation d'un mot, d'une expression ou d'un terme technique dans un type de texte donné ou dans un domaine technique, à l'aide d'un corpus de textes et d'un concordancier.

Chapitre 2

Analyse morphologique de l'arabe

Chapitre 2 : Analyse morphologique de l'arabe

1. Introduction

Il est difficile d'aborder l'étude d'une langue sans référer à l'histoire qu'elle véhicule, « l'origine de la langue arabe remonte au II[e] siècle, dans la péninsule Arabique, dans une forme assez proche de l'arabe standard moderne actuel. Le radical *ʿarab*, désigne le désert et c'est un mot araméen « arâbâh » et peut également dériver de la racine sémitique Abhar « se déplacer » »

« utilisé comme étant une langue officiel, l'arabe est l'une des principales langues parlées au monde avec plus de 300 millions de personnes dans divers pays arabes. »(Abu-absi, 1986) Par ses propriétés morphologiques et syntaxiques, la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique de la langue.

2. La richesse de la langue arabe

La langue arabe est un monde à part entière, Si on procède par le nombre de mots, c'est la langue qui occupe la première place et de loin, sa richesse créa un vocabulaire fort étendu et une rare souplesse de formes, comme étant une « langue savante, compliquée et très riche, elle s'écrit au moyen de 28 lettres (voir ***figure 5***). Cet alphabet est un *abjad* : terme décrivant un système d'écriture ne notant que les consonnes de la langue. » Ces consonnes peuvent devenir également des longues voyelles comme 'ا', 'و' et 'ي'

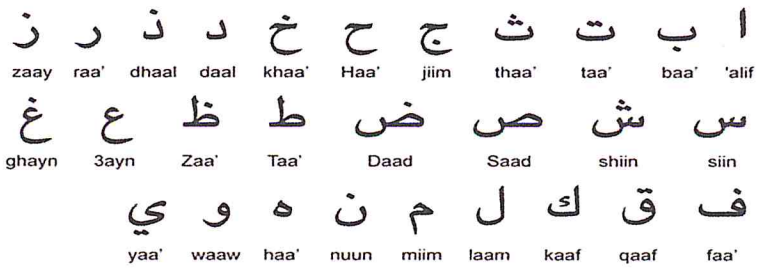

Figure 5 : *L'alphabet arabe*

L'écriture arabe comporte également des voyelles qui ne sont pas essentielles à l'écriture, il existe de plus une série d'autres diacritiques de syllabation dont les plus courants sont l'indication de l'absence de voyelle « ° » (السكون) et la gémination des consonnes « ّ » (الشدة).

« Les grammairiens arabes prétendent que toutes les racines de leur langue, au nombre de 6 000, ont été primitivement des verbes. Les mots composés de ces racines se complètent, soit

au moyen de lettres, soit par le redoublement des radicales, ou encore par le changement des voyelles. C'est ainsi qu'une même racine peut donner des verbes, des substantifs, des adjectifs, des adverbes, enfin des dérivés de toute sorte.

L'arabe dispose également d'une richesse de vocabulaire et de figures rhétoriques inimaginables. Il y aurait ainsi 80 termes différents pour représenter le miel, 200 pour le serpent, 500 pour le lion, 1000 pour le chameau, autant pour l'épée, et jusqu'à 4000 pour rendre l'idée de malheur. De la même manière. Cette richesse sémantique témoigne du raffinement de la civilisation et de la littérature arabo-islamique. Il existe une multitude de nuances d'idées, chacune traduite par un terme spécifique. » (Chouikha, 2010)

Malheureusement, Malgré la richesse de la langue arabe, mais elle n'arrive pas à suivre l'évolution et le développement de certaines langues comme le français ou l'anglais, ce qui pousse les arabophone à reprendre des mots à ces langues étrangères pour désigner certaines choses au lieu de créer des mots propres à la langue arabe.

Dans ce qui suit, nous décrivons les caractéristiques essentielles de la langue arabe, qui sont en relation avec notre travail.

3. L'arabe standard modern

Quand on se réfère à la langue arabe, on ne peut certainement pas parler d'unicité, c'est en réalité une entité complexe. Certes, l'arabe appartient à la famille des langues sémitiques, mais cette affirmation pose d'emblée un certain nombre de problèmes, dans notre cas on fait référence à l'arabe standard moderne, qui est la variété écrite de la langue, plus ou moins commune à tous les arabophones du monde. (Baccouche, 2009 : 19) affirme que « ce qu'on appelle aujourd'hui l'arabe moderne est le fruit d'une évolution qui a duré plus d'un millénaire, avec une interaction entre l'arabe littéral et ses divers dialectes ». Cependant, il n'est la langue maternelle de personne puisqu'il est appris à l'école à partir de l'âge de 6 ans, certains chercheurs affirment (Holes, 1995) qu'on ne peut pas généraliser ou dire que l'école constitue la première exposition ou le premier contact de l'enfant avec l'arabe standard cela dépend de l'environnement social et culturel dans lequel évolue l'enfant.

4. Les formes de base d'un mot arabe

« Les caractéristiques d'un mot arabe sont nombreuses. On y trouve, le type (nom, verbe, particule, ….), le nombre (singulier, duel, pluriel), le genre (masculin, féminin, neutre), l'état (défini, indéfini), le cas (nominatif, accusatif, génitif, …), le cas grammatical etc. » (Marsi et al., 2005), (Khoja et al., 2001) (Al Ameed, et al., 2008) Cette classification montre ses limites, quand il s'agit d'un traitement informatisé de la langue. En effet, « la classe des particules a été étendue pour inclure des morphèmes grammaticaux qui, en réalité appartiennent à d'autres classes telles que les pronoms démonstratifs ou relatifs qui constituaient des entrés nominales particulières, cette extension a abouti à une réorganisation en trois sous-ensembles du lexique arabe : Les verbes, les noms, les particules.»(Kouloughli, 1991) (khoja et al. 2001)

4.1 Le verbe

Un verbe est une entité exprimant un sens dépendant du temps, la majorité des verbes arabes sont formés sur des radicaux de 3 consonnes et éventuellement 4 consonnes, ces deux derniers peuvent donner naissance à des schèmes à la suite d'une ou plusieurs transformations morphologiques telles que le redoublement d'une consonne, (El-Dahdeh, 1999) affirme que « Le verbe de la langue arabe est classé selon deux critères principaux : le schème et la racine .»

4.2 Le nom

Le nom est un élément désignant un être ou un objet qui exprime un sens indépendamment du temps, Il peut être propre, commun ou dérivé d'un verbe. Il s'exprime au singulier, au duel ou au pluriel, au féminin ou au masculin. Il peut être agent, objet, instrument ou lieu.

4.3 La particule

La particule est une entité qui sert à situer les événements par rapport au temps et par rapport à l'espace. Elles peuvent être des conjonctions de coordination « و,أو »ou de subordination « إذا » Les particules sont généralement des mots outils, bien que jouant un rôle important dans la cohésion d'une phrase, sont souvent associées à des mots vides qui ne véhiculent pas un sens spécifique à un domaine donné.

5. Le mot arabe et ses compositions

Sur le plan linguistique, cette question a fait couler beaucoup d'encre et elle a donné lieu à plus d'un exposé. Plusieurs livres ont été consacrés à cette question, sans chercher à être exhaustive nous citons (Gaudin et Guespin, 2000 : 355) et (Mitterand, 2000 : 127) qui

présentent un bon récapitulatif des grandes théories sur le sujet et offrent une bonne entrée en matière pour les lecteurs désireux d'approfondir ce point.

Le mot graphique[1] en arabe peut provenir d'une structuration assez complexe, auquel cas il est désigné de « mot maximal ». Cette appellation a été attribuée par (Cohen, 1970) à un mot graphique décomposable en : proclitique(s), forme fléchie, enclitique(s). La forme fléchie désignée de mot minimal, est le noyau lexical du mot graphique, les autres constituants étant des extensions.

« La définition du mot du point de vue du traitement automatique (Pearson, 1998) se heurte à des considérations syntaxiques, sémantiques voir pragmatiques. La nature agglutinante de la langue : l'ensemble des lettres collées les unes aux autres et constituant un mot graphique véhiculent plusieurs informations morphosyntaxiques. Sur le plan pratique le mot graphique en arabe est porteur de beaucoup plus d'informations que les mots latins. Certains mots se traduisent par l'équivalent d'une phrase en français par exemple le mot « أستتذكرونهم » est traduit par « est-ce que vous allez vous souvenir d'eux » » (Abbes, 2004 : 39), En effet la séquence de lettres entre deux espaces est formée de proclitique + préfixe + base + suffixe + enclitique (Cohen, 1970) (Descles et al. 1983) (Dichy, 1990).

Le mot graphique arabe est considéré comme une structure d'objet complexe contenant une suite de morphèmes[2]. Chaque mot graphique peut se décomposer en une suite ordonnée de : proclitique(s), préfixe, base, suffixe(s), enclitique(s) (Zaafrani, 2002)

5.1 Proclitiques et enclitiques

Il y a des particules qui s'additionnent au début ou à la fin d'un mot pour en changer le sens ou pour avoir un effet sur la rection du mot.

Elles s'appellent les enclises (لواصق). Celles qui viennent au début de mot sont les proclitiques (لواصق قبلية) comme : « ل » (لام التوكيد), (لام الأمر) comme dans les entités « لكريم », « ليضرب »et aussi le signe de la détermination «ال».

---

[1] Nous désignons par mot graphique toute séquence de caractères arabes délimités par deux séparateurs (blanc ou autre)

[2] Les morphèmes constitutifs de l'unité-mot sont appelés des formants de mot, c'est-à-dire, des signes linguistiques minimaux dont les relations de contextualisation sont limitées aux autres morphèmes inclus dans l'unité composée que constitue le mot dans sa manifestation graphique (Dichy, 1987).

Alors que celles qui viennent à la fin sont appelées enclitiques (لواصق بعدية), par exemple les pronoms affixes compléments comme (ها, هم) dans : (مجلسها,ضربهم)

Remarque :

Les proclitiques et les enclitiques ne sont pas libres d'apparaître au hasard, mais il existe une certaine compatibilité entre eux c'est-à-dire le découpage du mot en " proclitiques + base1+ enclitiques" ne se limite pas à la recherche d'un proclitique (respectivement un enclitique) parmi la liste au début (respectivement la fin) du mot, mais ainsi à une certaine compatibilité entre les proclitiques et les enclitiques repérés dans le mot à décomposer. Le tableau (voir **Table 1)** représente la liste des proclitiques et enclitiques, avantageusement elle est limitée, elle a été proposé par (Darwish, 2003), le tableau indique aussi les cas d'incompatibilité (*), c'est-à-dire proclitique et enclitique compatibles donne une décomposition correcte.

Proc/enc	"	ه	ي	ك	هم	هن	هما	ها	كم	كن	كما	ني	نا
"													
ب												*	
ك												*	
ل													
ف													
س			*										
أ			*										
ال		*	*	*	*	*	*	*	*	*	*	*	*
بال		*	*	*	*	*	*	*	*	*	*	*	*
كال		*	*	*	*	*	*	*	*	*	*	*	*
لل		*	*	*	*	*	*	*	*	*	*	*	*
فب												*	
فس			*										
فال		*	*	*	*	*	*	*	*	*	*	*	*
فك												*	
فل												*	
فلل		*	*	*	*	*	*	*	*	*	*	*	*
أف													
أس			*										
فبال		*	*	*	*	*	*	*	*	*	*	*	*
فكال		*	*	*	*	*	*	*	*	*	*	*	*

***Table 1** : table de compatibilité entre proclitiques/ enclitiques*

5.2 Préfixes et suffixes

Comme les enclises, les affixes sont concaténées au début (préfixes) ou à la fin (suffixes) du mot.

Les grammairiens arabes catégorisent les préfixes par les lettres qui sont ajoutées aux verbes pour exprimer l'inaccompli c'est-à-dire les lettres assemblées dans le mnémonique « أنيت » (حروف المضارعة).

Les suffixes sont les lettres qui donnent une information sur le genre comme le «تاء » dans «الكريمة» (la généreuse) et le nombre «ان» ,«ات» dans « المجلسان »: (مثنى ***المجلس***) et « الكريمات » : (جمع مؤنث ***الكريمة***) et les pronoms affixes sujets comme le «تم» dans «ضربتم». *(Sadik et al. 2007 : 43-46)*

Remarque

Pour les préfixes et les suffixes, il existe aussi une certaine compatibilité. Le tableau (voir **Table 2**) représente les préfixes et les suffixes existant ainsi les cas d'incompatibilité (*), préfixe et suffixe compatibles donne une décomposition correcte.

Préfixe/ Suffixe	‘‘	ات	ية	ة	يات	نا	ت	تما	تم	تن	ن	ين	ان	ون	وا	ا	ي
‘‘																	
ا		*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
ت		*	*	*	*		*	*	*	*							
ي		*	*	*	*		*	*	*	*		*					*
ن		*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
إ		*	*	*	*		*	*	*	*		*	*	*			

Table 2 : *table de comptabilité entre préfixes / suffixes*

5.3 Base

La base du mot est, construite selon un procédé appelé *dérivation interne* (Zaafrani, 2002) La dérivation consiste à créer une forme à partir d'une autre forme même si les deux se rejoignent dans le sens, dans la matière et dans la façon de se construire pour exprimer à travers la deuxième [forme] le sens d'origine.

La dérivation interne est constituée selon des schèmes[3], Quel que soit le mot il est donc issu

[3]La grammaire arabe fait traditionnellement usage، pour cela، d'une convention qui consiste à faire appel à une racine tri-consonantique théorique (ف،ع،ل) = («faire»). On utilise également de nos jours d'autres représentations symboliques comme $R_1R_2R_3$ (où « R » signifie consonne radicale et le chiffre en indice indique la position dans la racine).

d'une racine et inséré dans un schème. En fait le schème est une sorte de moule.

Les racines et les schèmes constituent deux grands systèmes croisés, enveloppant dans leur réseau toute la masse du vocabulaire sémitique, (Cohen, 1970 : 48) précise : « la plus grande partie, et de loin, du vocabulaire se définit en effet par le croisement d'une racine et d'un schème ». Tout mot est analysé selon ces deux systèmes et appartient à chacun d'eux. Il s'agirait là d'une caractéristique commune et profonde des langues sémitiques

Exemples de schème (l'étoile représente les 3 lettres de la racine) :

م***ة ou م*** ➡ (selon le mot est féminin ou masculin il aura un ة مربوطة)

Ce schème est celui des noms de lieu اِسْمُ الْمَكَانِ, c'est à dire que tout mot qui sera inséré dans ce schème représentera un lieu qui a un lien avec la racine dont est issu le mot.

En associant la racine 'درس' au schème de nom de lieu م***ة nous obtenons le lieu où l'on étudie en d'autres terme l'école ➡ مدرسة = درس+ م***ة

Il existe évidemment beaucoup d'autre schème, alors par commodité les grammairiens arabes ont remplacé les trais qui correspondent, dans les schèmes aux trois consonnes de la racine, par une racine type فعل : où ف représente la 1ère consonne de la racine,ع la 2ème consonne, etل la 3ème consonne. Ainsi le schème des noms de lieu م***ة \م*** sera désormais identifié comme le schème 'مفعلة' et 'مفعل'

Ce procédé s'applique pour la totalité des verbes, des dérivés nominaux immédiats (nom verbal, participe actif, etc.) et pour une partie importante des noms. Toutefois, un sous-ensemble important des noms, c'est le cas des mots empruntés aux autres langues et des noms propres, ne sont pas construits selon ce procédé. Ces noms correspondent à des pro-bases (Dichy, 1997).

6 L'arabe utilisé dans le corpus de Latifa Al-solaiti

Le corpus construit par Latifa Al-solaiti, a été généralement tiré à partir des sites Web, elle a obtenus ses textes écrites après avoir identifié plusieurs sites utiles, le corpus contient aussi, des textes extrait à partir des fichiers parlés obtenus de la radio. Elle a aussi enrichie son corpus avec des différentes magazines, et journaux en ligne, cela a influencé sur la qualité de corpus, on ne peut pas négliger que le corpus contient plusieurs erreurs graphiques, dû au différentes sources, qui n'ont pas été parfaitement qualifier pour une bonne qualité de la

langue arabe, on peut rencontrer des mots qui n'existe pas dans la langue, même des mots qui ne fais pas partie de l'arabe classique, c'est-à-dire des mot utilisé dans le dialecte «el 3amiyah».En effet le corpus a été construit avec des textes non voyellès, (pas tout à fait) car parfois, on peut rencontrer des mots voyellès avec la chadda « الشدة » etc.

Certainement on ne peut pas échappée à la réalité que La non-vocalisation due à une absence des voyelles brèves dans les textes qui forme le corpus entraîne un haut degré d'ambiguïté.

Par exemple le mot 'كتب ' a deux sens « il a écrit » et « des livres »

7 Faut-il identifier le mot arabe ?

Nous avons montré l'importance de la reconnaissance des compositions du mot pour son identification. Mais en raisons de l'écriture non-vocalisé de l'arabe une même forme graphique peut avoir plusieurs analyses morphologique d'où l'importance de l'identification grammatical des mots arabes.

Avec l'ambiguïté née dans la langue arabe, l'analyse morphologique est indispensable pour réaliser une concordance automatique, qui doit répondre à plusieurs heuristiques pour réduire la multiplicité des solutions, et doit rester interactive et assister l'inévitable intervention experte.

Dans la section suivante, nous allons définir l'analyseur morphologique et son principe, nous présentons aussi par la suite les éléments essentiels pour la construction de l'analyseur morphologique arabe.

8 L'analyse morphologique de l'arabe

8.1 Introduction

Aujourd'hui, avec le développement qu'a subit l'informatique que ce soit en terme de vitesse de traitement ou de support de stockage, le traitement automatique du langage naturel est devenu un domaine à la fois technologique due à l'émergence d'un nombre important d'applications, tels que : les traducteurs automatiques, générateurs automatiques de résumé, correcteurs orthographiques d'erreurs, …etc. Mais aussi un domaine scientifique traitant des problématiques de plus en plus complexes comme celle de l'ambiguïté.

Dans la littérature l'ambiguïté est comparée à un état de confusion, cet embrouillement se manifeste sous différentes formes et selon les différents niveaux de traitements que ce soit

lexical, morphologique, syntaxique et même sémantique. L'une des formes d'ambiguïté la plus persistant en traitement automatique de la langue arabe est l'ambiguïté morphologique

Comme nous avons constaté que l'analyse morphologique n'est pas non seulement indispensable pour réaliser la concordance automatique mais elle est aussi importante pour avoir le bon résultat au niveau grammatical, on peut à tout moment se demander comment peut-on appliquer la morphologie ? Cela provoque plusieurs heuristiques qui mènent à la résolution de cette question, en effet pour répondre à ce dernier il est opportun d'étudier dans un premier temps le principe de l'analyseur morphologique.

8.2 L'analyse morphologique

« La morphologie est un domaine de la langue qui permet la description des règles régissant la structure interne des mots (unités lexicales), chez les grammairiens la morphologie est l'étude des formes des mots (flexion et dérivation), en d'autres termes, la morphologie est l'étude des mots considéré isolément (hors contexte) sous le double aspect de la nature et des variations qu'ils peuvent subir .» (Hoceini, 2002) « En langue arabe, l'analyse morphologique est d'autant plus importante que les mots sont fortement agglutinés[4], c'est-à-dire qu'ils sont formés dans leur majorité par assemblage d'unités lexicales et grammaticales élémentaires.» (Balou, 2003) Ainsi Le traitement morphologique est considéré comme une introduction principale à la compréhension globale d'une langue naturelle ; il joue un rôle très important aussi bien du côté linguistique que du côté technique.

8.3 Description de la méthode

Ce travail s'inscrit dans le cadre de l'analyse morphologique de la langue arabe, nous nous intéressons au traitement d'un texte non voyellès et à son apport pour l'analyse morphologique, En effet la plupart des analyseurs morphologiques existants provoque un phénomène d'ambiguïté morphologique dans le traitement du texte arabe, parmi eux on a flashé sur un analyseur morphologique assez compréhensible utilisé dans le mémoire de magister de (Sadik et al. 2007)

Nous commençons par présenter cet analyseur, en nous focalisant sur les erreurs amenées lors de l'analyse, Ensuit nous décrivons en détail la méthode avec laquelle nous visons à améliorer les performances de l'analyseur morphologique proposé dans (Sadik et al. 2007).

---

[4] Processus d'ajout d'affixes à un mot qui exprime ses différentes relations grammaticales

8.4 L'analyseur morphologique proposé dans (Sadik et al. 2007)

Il procède en premier lieu par une segmentation qui découpe le mot pour que chacune des parties obtenues soit une entité lexicale. Cette segmentation isolera les préfixes et les suffixes du mot. La partie restante correspondra à la racine dans le cas où la segmentation est poussée jusqu'à la fin (par l'utilisation de la notion du schème).

Dans le cas contraire la partie restante est appelée une base. Cette étape est une tâche délicate du fait que l'arabe est une langue flexionnelle et fortement dérivable.

L'analyseur morphologique ne peut fonctionner sans l'aide d'un dictionnaire contenant les unités lexicales. Cette étape est l'analyse lexicale qui permet de vérifier si l'unité lexicale appartient bien à la langue, mais qui doit aussi vérifier la compatibilité entre les différents constituants du mot. Une troisième étape interprète la base obtenue par la segmentation en se basant toujours sur la notion de schème, et retourne comme résultat la valeur morphologique en se basant sur différentes dictionnaires pour vérifier chaque résultat obtenus.

Malgré l'utilisation des différents dictionnaires mais le problème se pose lors de différentes élimination, la procédure se base sur l'effacement des enclises (respectivement les affixes) pour obtenir la base voulu, on a rapidement remarqué que ça arrive dans la plupart des temps, que les racines a trois lettres (racine trilitère) perd l'information lors de ce découpage

Exemple

La décomposition du mot 'فسمعهم' donne :

Proclitique	فس
Enclitique	هم
'base'	مع

Figure 6 : *Décomposition erronée du mot « فسمعهم »*

C'est une décomposition fausse qui a clairement crée un cas d'ambiguïté, malgré le mot 'مع' existe dans la langue arabe, Mais la racine de 'سمعهم' est 'سمع ' et (non pas ' مع '). Le problème qui s'est posé dans ce cas est dû au fait que 'س' est un radical de 'سمع' et un proclitique en même temps.

Remarque

Dans ce cas, il n'existe pas les infixes (l'exemple évoqué ci-dessus.)

La décomposition du mot 'فسأحسنه' donne :

Proclitique	فس
Enclitique	ه
'base'	أحسن
Préfixe	أ
Suffixe	ن
Base	حس

Figure 7 : *Décomposition erronée du mot «* فسأحسنه *»*

Comme c'est montré la base 'حس' n'est pas une base valide, l'élimination du suffixe 'ن' qui fais partie du radical de la base correct 'حسن' a donné une décomposition incorrecte.

Dans ce qui suit nous allons présenter d'une manière plus compréhensible tout la procédure qu'on va utiliser, en évitant les erreurs produit par la méthode précédente.

8.5 Méthode proposée

Pour remédier à ce problème, la méthode que nous proposons consiste à vérifier le nombre de lettres qu'on va obtenir dans la base,

On doit veiller à obtenir une base qui sera supérieur ou égale à trois lettre, car comme nous avons déjà mentionner que nous intéressons qu'aux racines trilitères, si la base est égal à trois(racine), elle est probablement un verbe de trois lettre, cela seras vérifier dans le dictionnaire des verbes, sinon elle sera prise comme étant un schème qui se base sur une procédure spécifique pour extraire la racine.

Pour ce faire, nous avons créé un certains dictionnaires nécessaires (échantillon) dont chacun a une structure spécifique (si obligatoire), nous avons aussi utilisé les différentes liste qui contient les enclises et les affixes, ainsi un mode de représentation obligatoire des enclises et des affixes, ce mode de représentation va permettre de réaliser les tests de compatibilité nécessaire qui veille à l'obtention d'une base correcte ;

Dans un premier temps, nous représentons dans un tableau (voir **Table 3**) les fichiers utilisés pour la bonne conduite de notre analyseur morphologique, Ensuite on va appliquer notre méthode proposé en montrant les détails nécessaire pour une meilleure compréhension.

Fichiers	Type	Description
Schèmes	**dictionnaire**	une codification des schèmes pour faciliter la recherche du schème et de la racine de l'entité
Racines	**dictionnaire**	Dans ce dictionnaire sont stockées les racines représentatives de la langue, dans notre système on a utilisé un échantillon des racines de trois lettres, ces racines sont les plus utilisées dans la langue arabe. Ceci nous permettra de vérifier la compatibilité entre la racine et le schème, évitant ainsi toute décomposition erronée
Mots outils	**dictionnaire**	tout mot qui reste invariant quel que soit son contexte (à l'exception des noms propres et mots communs) tel que les pronoms, particules,…etc. Nous ne mettons dans ce dictionnaire que les mots-outils isolés
Mots spécifiques	**dictionnaire**	formée de mots qui n'ont pas une origine arabe comme les noms propres et communs ; On a proposé un échantillon de quelque catégorie tel que pays, les capitaux, et nom propres.
Mots inconnu	**dictionnaire**	On le considère comme étant un dictionnaire imaginaire, il se caractérise par des mots non porteurs de sens, on pose les décompositions mises en échec comme étant mot inconnu.
Proclitique	**Liste**	liste fini des proclitiques
Enclitique	**Liste**	liste fini des enclitiques
Préfixes	**Liste**	liste fini des préfixes
Suffixes	**Liste**	liste fini des suffixes
Compatibilité des enclises	**Liste**	Liste fini contient des chaines fusionnées à partir de deux sous chaines (proclitique&&enclitique) qui présente les cas d'incompatibilité
Compatibilité des affixes	**Liste**	Liste fini contient des chaines fusionnées à partir de deux sous chaines (préfixe&&suffixe) qui présente les cas d'incompatibilité

***Table 3 :** fichiers utilisé pour l'analyseur morphologique*

Remarque

Rappelons que chaque mot du lexique arabe est associé un schème qui est le même mot sauf les lettres de sa racine qui sont remplacées par les lettres de la racine 'فعل'.

Le dictionnaire des schèmes qu'on a utilisé a une structure spécifique :

Exemple

Wazn i	افتعل
Liste-infixe i	13
Catégorie	verbe

Table 4 : *Structure de recherche de schème*

Le champ Wazn i contient la chaîne consonantique du schème. Le champ liste-infixe i contient la position des lettres autres que les lettres de la racine dans le schème et le champ catégorie donne la catégorie grammaticale (nom, verbe, ...). Cette structure est adaptée pour faciliter la recherche des schèmes qu'on va présenter dans la suite

Ainsi le dictionnaire des mots utiles se décompose à plusieurs types qui sont :

- Adet-chart 'أداة شرط' comme : [لو,لولا].
- Dharf-makan ' ظرف مكان' comme : [تحت, أمام,وسط].
- Dharf-zaman 'ظرف زمان' comme : [قبل, عندما,يوم].
- Fi'l madhi-na'ess 'فعل ماض ناقص' comme : [يكون, كان,كنت].
- Harf-nidaa 'حرف نداء' comme : [أيها, يأيها,يا].
- Adet-istifham 'أداة استفهام' comme : [لماذا,أين,هل].
- Adet-Tawkid 'أداة توكيد' comme : [إن, لقد].
- Adet-nafy 'أداة نفي' comme : [لم,كلا].
- Adet-el 'Ad 'أداة العد' comme : [كل,بعض,جل].
- Adet-el-rabt 'أداة الربط' comme : [مع,و,ثم].
- Adet-jarr ' أداة جر' comme : [من,في,به].
- Dhamir-el-moutakalim 'ضمير المتكلم' comme : [نحن , انا].
- Dhamir-el-moukhatib 'ضمير المخاطب' comme : [أنتم,أنتما,أنتي].
- Dhamir-el-ghaib ' ضمير الغائب' comme : [هو,هي,هن].
- Issm mawssoul 'إسم موصول' comme : [الذي,التي].
- Issm ichara 'إسم إشارة' comme : [ذلك,هذه,هذا].

8.6 Application de la méthode proposée

Pour une meilleur compréhension on a préféré d'expliquer chaque décomposition du mot appart, le mot passe pas trois principales types de découpage pour arriver au résultat souhaité.

8.6.1 Eliminer les enclises

Premièrement on va présenter les listes des proclitiques et enclitiques ainsi la liste de leur incompatibilité qu'on a utilisé, ces listes sont extraites de la **Table 1**(table de compatibilité entre proclitiques/enclitiques) :

Proclitique = [‘’, فكال,فبال,أس,أف,فلل,فك,فال,فس,بال,فب,لل,كال,أ,س,ف,ل,ك,ب]

Enclitique = [’’,ه, ي, ك, هم, هن, هما, ها, كم, كن, كما, ني, نا]

Ensuite la liste d'incompatibilité des enclises est comme suit :

[بني,كني,سي,أي,اله,الي,الك,الهم,الهن,الهما,الها,الكم,الكن,الكما,النا,الني,الن,باله,بالي,بالك,بالهم,بالهن,بالهما,بالها,بالكم,بالكن,بـ
الكما,بالني,بالنا,كاله,كالي,كالك,كالهم,كالهن,كالهما,كالها,كالكم,كالكن,كالكما,كالني,كالنا,لله,للي,للك,للهم,للهن,للهما,للها,للكم,لل
كن,للكما,للني,للنا,فبني,فسي,فاله,فالي,فالك,فالهم,فالهن,فالهما,فالها,فالكم,فالكن,فالكما,فالني,فالنا,فكني,فلني,فلله,فللي,فللك,فللله
م,فللهن,فللهما,فللها,فللكم,فللكن,فللكما,فللني,فللنا,أسي,فباله,فبالي,فبالك,فبالهم,فبالهن,فبالهما,فبالها,فبالكم,فبالكن,فبالكما,فبالني,
فبالنا,فكاله,فكالي,فكالك,فكالهم,فكالهن,فكالهما,فكالها,فكالكم,فكالكن,فكالكما,فكالني,فكالنا]

Le mot passe par une phase qui va éliminer les proclitiques (respectivement enclitiques) qui existe pour obtenir la base.

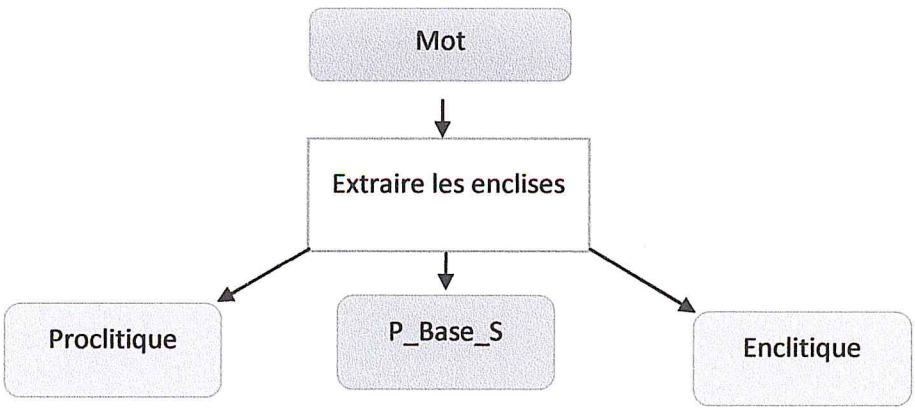

Figure 8 : *Extraire les enclises*

Le processus identifie le plus long proclitique (respectivement enclitique) à éliminer, ce dernier s'effectue en respectant la règle de compatibilité, pour vérifier la compatibilité on fusionne le proclitique identifié avec l'enclitique identifié dans une chaine ensuite vérifier

l'existence de cette dernière en accédant à la liste de compatibilité des enclises. Si la chaine existe on ne peut pas effectuer l'élimination sinon l'effacement de proclitique identifié (respectivement enclitique) sera effectuer, les deux cas vont donner une décomposition correcte.

Exemple

Mot	كمساجدهم
Proclitique	ك
Enclitique	هم
Fusion (enc+proc)	كهم
Compatibilité	Non identifié
P_Base_S	مساجد

Figure 9 : *Décomposition correcte du mot «* كمساجدهم »

La P_Base_S obtenue 'مساجد' est tout à fait correcte, le processus a éliminé les enclises car la fusion de ses derniers qui est 'كهم' n'existe pas dans la liste de compatibilité des enclises.

8.6.2 Eliminer les affixes

Les listes suivantes sont extraites à partir de la **Table 2** (table de compatibilité entre préfixes suffixes) :

Préfixe = ['',ت , ن , ي , إ , ا]

Suffixe = ['', ات , ة , ية , يات , نا , ت , تما , تم , تن , ن , ين , ان , وا , ا, ي, ون]

La liste de compatibilité des affixes est comme suit :

[
ات,تات,يات,نات,إات,اية,تية,يية,نية,إية,اة,تة,ية,نة,إة,ايات,تيات,ييات,نيات,إيات,ات,تت,يت,نت,إت,اتما,تتما,يتما,نتما,إتما,
اتم,تتم,يتم,نتم,إتم,اتن,تتن,يتن,نتن,إتن,اان,نن,اين,ييين,نين,إين,ان,نان,إان,اون,نون,إون,او,نوا,اا,نا,اي,يي,ني]

La P_Base_S obtenus de la phase précédente qui se charge par l'élimination des enclises passe par la deuxième segmentation qui va éliminer les affixes qui existe, c'est-à-dire identifier le préfixe (respectivement suffixe) à supprimer.

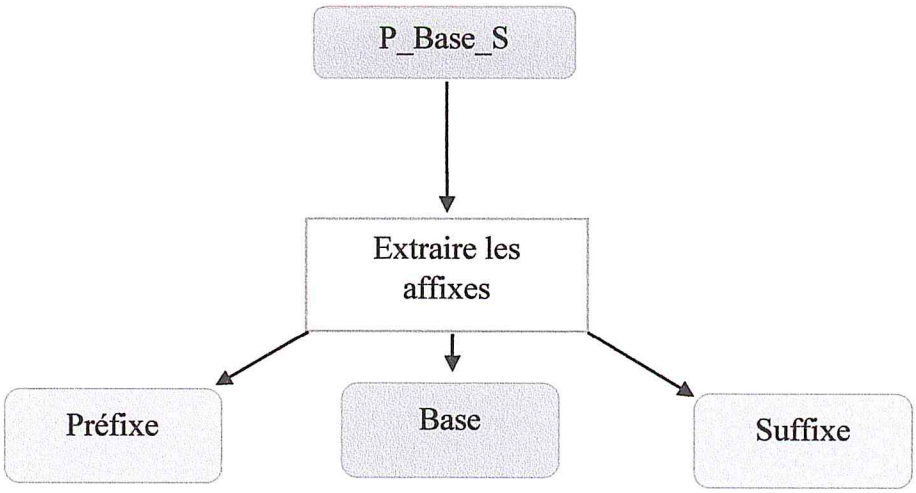

Figure 10 : *Extraire les affixes*

L'extraction des affixes se fait de la même manière que les enclises, la ***figure 11*** représente un exemple qui passe par les deux phases de segmentation précédentes dont le mot qu'on va représenter contient des enclises et des affixes.

Exemple :

La décomposition du mot 'فسأسمعه' d'après le processus de décomposition des enclises (en proclitique+ P_Base_S +enclitique) et celui des affixes (préfixe+ base +suffixe) donne :

Proclitique	فس
Enclitique	ه
Fusion enclises	فسه
Compatibilité enclises	Non identifié
P_Base_S	أسمع
Préfixe	أ
Suffixe	‘ ‘
Fusion affixes	أ
Compatibilité affixe	Non identifier
Base	سمع

Figure 11 : *Décomposition correcte du mot «* فسأسمعه *»*

Remarque :

Rappelons que la décomposition aura lieu pour les enclises (respectivement pour les affixes) si seulement si la compatibilité est identifiée.

Revenant à notre exemple de la ***figure 6*** (La décomposition du mot 'فسمعهم') par l'analyseur morphologique proposé dans (Sadik et al. 2007) de TALN, rappelons que cette méthode a donné la racine incorrecte 'مع', en appliquant notre méthode proposée, le résultat que va donner notre analyseur est correcte, comme c'est montré dans la ***figure 12.***

Comme c'est déjà mentionner le processus identifie le plus long proclitique (respectivement enclitique) du mot, si le résultat n'est pas bon, on propose une solution :

On doit vérifier si la base obtenu est inférieur ou égale à 3 lettres, dans ce cas on procède à une nouvelle autre identification du proclitique mois long par rapport au première (respectivement enclitique) évidement ce dernier peut se dérouler en cas d'existence d'un proclitique (respectivement enclitique) moins long.

Proclitique	ف
Enclitique	هم
Fusion enclises	فهم
Compatibilité enclises	Non identifié
Base	سمع

Figure 12 : *Décomposition correcte du mot « فسمعهم »*

C'est une décomposition correct, on peut voir que la base obtenu 'سمع' est le résultat souhaiter, cela été possible après l'élimination du proclitique 'ف' qui est moins long que le proclitique'فس' qui a été éliminer dans la première décomposition.

Un deuxième exemple est inévitable, *la* ***figure 13*** représente aussi la solution correcte de l'exemple précèdent (***figure 7***) :

Proclitique	فس
Enclitique	ه
Fusion enclises	فسه
Compatibilité enclises	Non identifié
P_Base_S	أحسن
Préfixe	أ
Suffixe	‘’
Fusion affixes	أ
Compatibilité affixes	Non identifié
Base	حسن

***Figure 13** : Décomposition correcte du mot «فسأحسنه »*

La racine 'حسن' est correcte, lors de l'élimination des affixes le processus n'élimine le suffixe 'ن' car la taille de mot ne peut être inférieur à 3 pour donner le bon résultat, on peut dire que c'est la même réflexion qui a été déjà cité pour les enclises.

8.6.3 Traiter la base

La base obtenue ne peut pas donner toujours une taille de 3 lettres (racine), par exemple le mot 'كمساجدهم' donnera une base de 4 lettres qui est 'مساجد' (***figure 9***), dans ce cas, la base sera prise comme étant un schème qui se base sur une procédure pour extraire la racine. Il est claire que la base s'analyse en racine et schème Le principe suivant va décrire la tâche.

La base sera prise en charge pour essayer de trouver le schème (ou les schèmes) ayant la même longueur que celle-ci, une fois trouver, l'analyseur vérifie si toutes les lettres correspondantes aux positions dans le champ liste-infixe se trouvent dans le mot "M" aux mêmes positions révélées par ce champ (liste-infixe). Cette procédure est déjà mentionnée dans la **Table 4** (Structure de recherche de schème).

Voici un exemple qui nous aide à comprendre le processus :

Mot = 'ساجد' le processus de recherche de schème parcourt tous les enregistrements qui ont la même taille avec le mot jusqu'à la rencontre du schème'فاعل' . Le champ liste-infixe correspondant est '2 ' la lettre 'ا' trouve à la position 2 du mot 'ساجد' donc c'est probablement le bon schème, la figure ci-dessous décrit l'identification du schème :

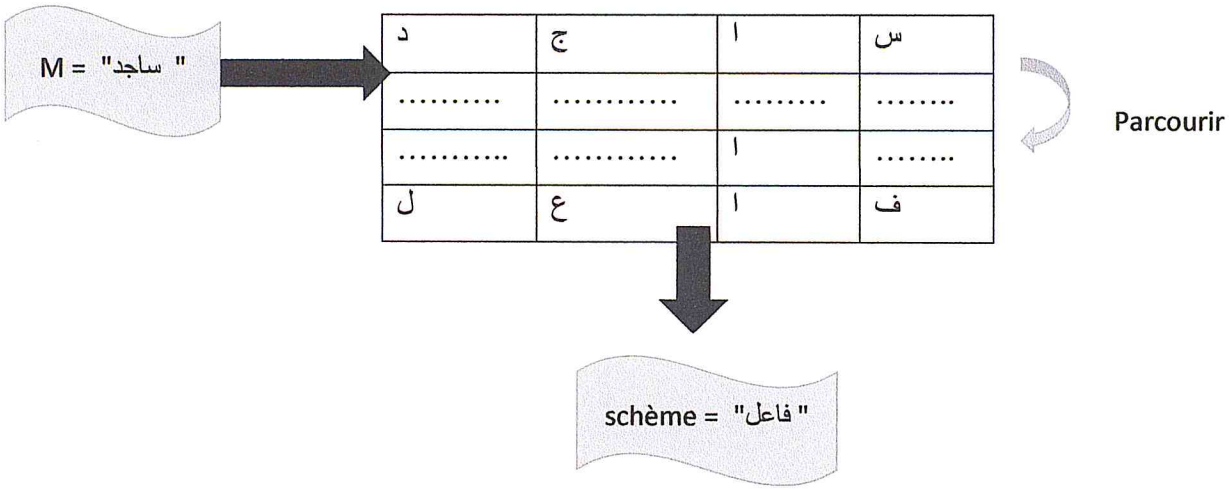

***Figure 14 :** identifier le schème*

Après la détermination du schème, l'extraction de la racine se limite à la suppression de toutes les lettres correspondantes aux positions de champs liste-infixe dans le mot à décomposer

Exemple

Le mot ' ساجد' a pour schème 'فاعل', le champ liste-infixe est '2'. L'élimination de la lettre 'ا' de la position 2 du mot 'ساجد' qui est la même du champ liste-infixe donne 'سجد', de cette façon on a retrouvé la racine correcte du mot 'ساجد' qui est la racine 'سجد'. la figure ci-dessous décrit l'identification de la racine.

Remarque

Après extraction de la racine, le champ catégorie représente la catégorie grammatical du mot ce dernier est générer en fonction de chaque schème. (La structure de dictionnaire des schèmes approprie chaque schème à une catégorie)

La racine de trois lettres non reconnu sera considéré comme un mot vide, sinon si l'expert linguistique vois que l'ajout de cette dernière est nécessaire, car pour lui elle est correcte, on lui donne la main d'ajouter la racine au dictionnaire, la figure qui suit éclaircie la procédure.

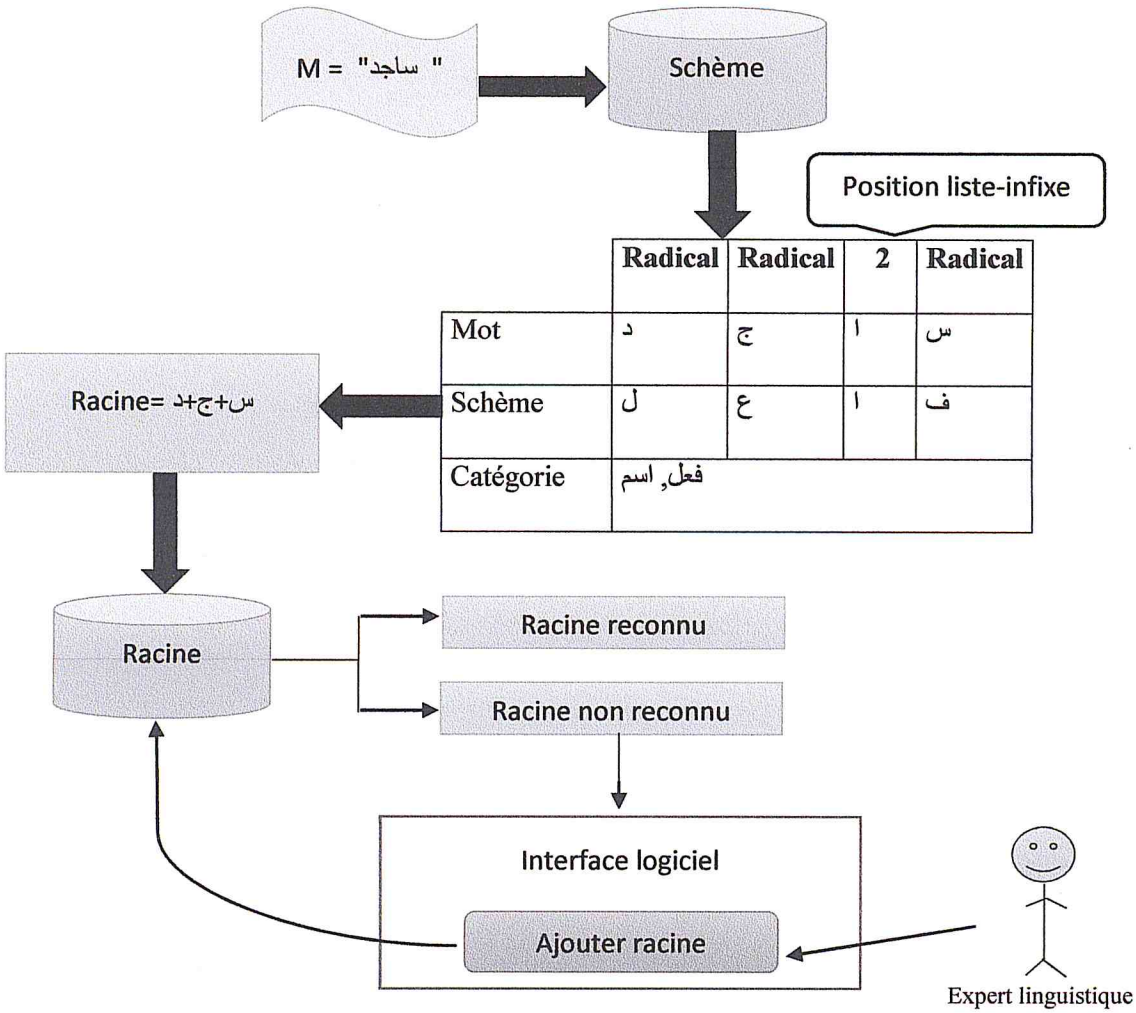

Figure 15 : *Extraire la racine*

Le schéma qui suit récapitule les étapes de l'analyse. Il représente le passage de mot dans chaque station, qui conduit soit à son identification avec succès, ou soit à un échec de reconnaissance de ce mot. Le mode échec à la fin conclu que le mot sera considéré comme inconnu.

Remarque

Le dictionnaire des mots vide se déroule en fonction des autres dictionnaires, plus ils sont riche plus on aura moins de résultat des mots vides.

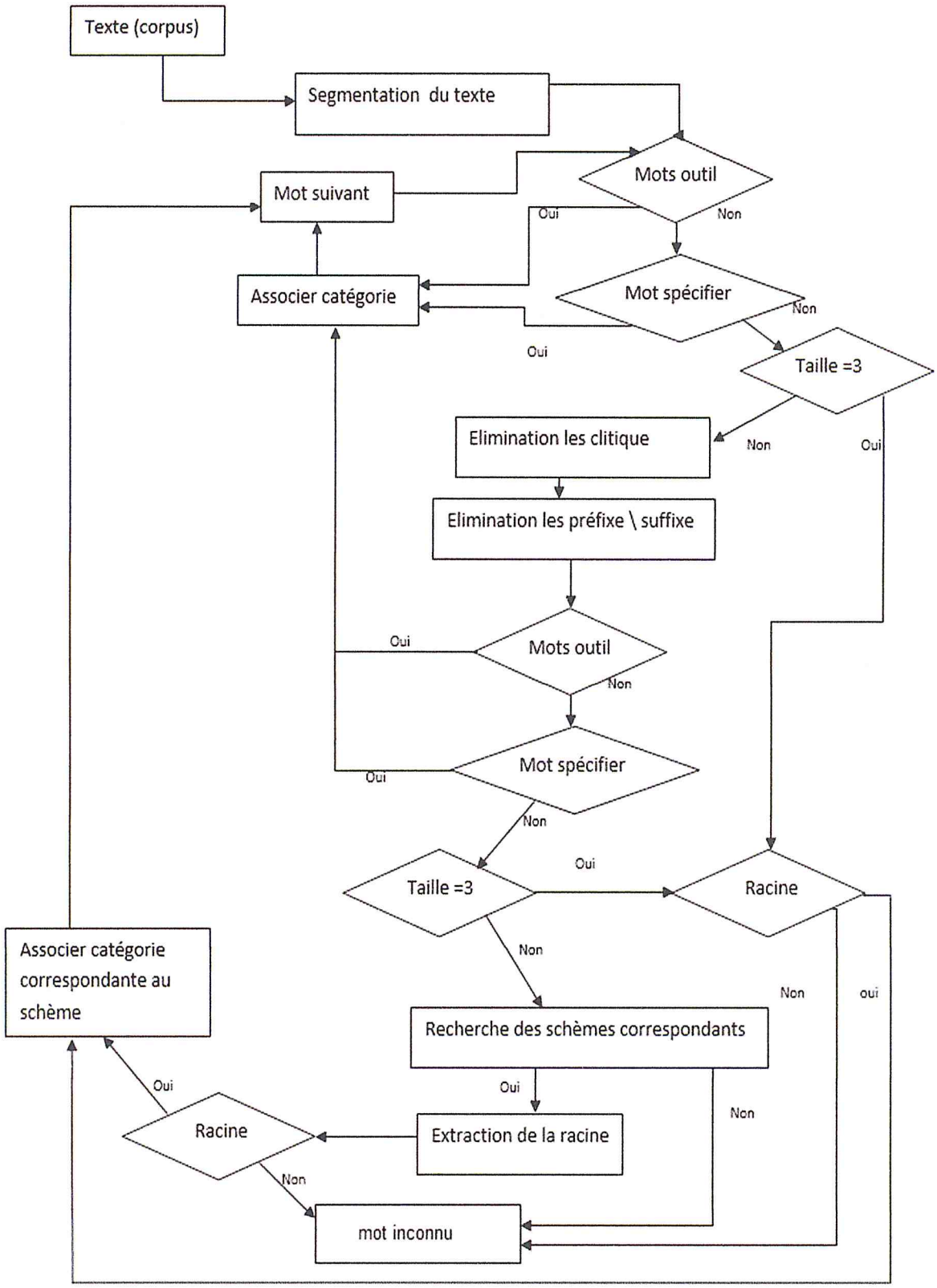

Figure 16 : *Schéma général de l'analyse morphologique*

9 Conclusion

La recherche d'information et la fouille de textes en langue arabe constitue un défi majeur. La langue arabe inclut des mots composés, c'est-à-dire accompagnés d'une structuration assez complexe, ce qui engendre des cas d'ambigüité.

La grammaire arabe est une branche de la linguistique contemporaine qui nécessite l'étude de la morphologie ainsi que la composition des mots en phrase, l'analyse morphologique s'est imposé comme une solution afin d'éviter les différentes cas d'ambiguïté, l'analyse suit un principe de différentes segmentation qui veille à identifier le mot composé

Nous avons cherché, dans ce travail, à présenter les outils qui veillent à contribuer très fortement à l'analyse morphologique en se basant sur un analyseur morphologique existant tout en profitant de ses avantages et évitant ses insuffisances.

Enfin on conclut qu'aujourd'hui l'analyseur morphologique est considéré comme un point de passage obligatoire dans le processus de développement d'une application TALN et son importance devient capitale pour les langues dotées d'une richesse morphologique, comme c'est le cas pour l'arabe.

Chapitre 3

Implémentation du concordancier

Chapitre 3 : Implémentation du concordancier

1. Introduction

Des linguistes de plus en plus nombreux utilisent des concordanciers pour extraire de textes préexistants des mots ou expressions en contexte. L'objectif de ce chapitre est de présenter notre concordancier afin de tester plus particulièrement ses fonctionnalités, Nous expliquons en quoi consiste chaque opération et ce qu'elle implique par l'utilisation de l'analyse morphologique.

La première méthode à prendre en considération pour produire un logiciel tel que le concordancier est bien qu'il soit souple, c'est au produit qui doit s'adapter à l'utilisateur et non à l'utilisateur à s'adapter au logiciel.

2. La notion XML

Comme notre travail consiste a utilisé un corpus format XML, décrire la notion xml est inévitable.

Le XML ou *Extensible Markup Language* est un langage informatique de balisage générique.

« Un langage de balisage est un langage qui s'écrit grâce à des balises. Ces balises permettent de structurer de manière hiérarchisée et organisée les données d'un document, il a été créé pour faciliter les échanges de données entre les machines et les logiciels. » (Ludovic, 2008)

3. Encodage UTF-8 utilisé

Tout document contenant du texte, et c'est le cas également pour les fichiers en .xml, sont enregistrés avec un jeu de caractères précis. Ce jeu de caractères, utilisé pour créer ou enregistrer le document, correspond à l'encodage réel du document, parmi eux l'encodage utf-8 qui permet théoriquement d'encoder toutes les langues, dont l'arabe fait partie.

4. Fichier utilisé

4.1 corpus XML

Forte heureusement que XML s'appuie sur Unicode, le codage universel de plus de 5000 caractères dans toutes les langues, parmi eux la langue arabe.

Comme nous avons déjà cité, Le fichier que nous avons utilisé est le corpus de Latifa el-solaiti, il se diffuse sur plusieurs catégories (11 catégories), dont chaque catégorie contient

plusieurs documents XML, on a déjà mentionné dans le chapitre 2(la section 6) que le corpus est loin d'être parfait, dû à sa construction qui été manuellement rédigé. Notre programme consiste à utiliser :

-un texte non voyellès,

-travailler avec des documents xml,

Mais un problème est déjà née sur ces deux derniers, on peut rencontrer des mots voyellès dans le texte, ainsi la structure des différentes documents xml n'est pas normalisé, c'est à dire qui ne suit pas la même structuration, on sait que le document xml se caractérise par des balises, qui vont certainement influencer sur le résultat de notre travail. Ces majeurs problèmes, nous a conduit à effectuer un prétraitement sur chaque document chargé à notre concordancier. Le prétraitement consiste à :

Lors du chargement de document xml a utilisé, enlever tout voyelles rencontré ainsi que les balises, pour la structure des documents xml on doit utiliser certains documents normalisé, pour un future besoin, afin d'éviter ce genre de problème, on affirme que ces insuffisances nécessite tout une étude qui consiste à utiliser des robot pour construire automatiquement des corpus qui respect une certains normes, notre collègue Mr Mahmoudi Lakhdar est entrain d'établir cet étude.

4.2 Dictionnaires

Appart le corpus chargé, le concordancier a besoin de différentes ressources pour pouvoir répondre au requête de l'utilisateur, ces ressources semble indispensable. C'est une sorte de différents dictionnaires, dont l'utilité de chaque dictionnaire revient à effectuer une reconnaissance dynamique et intelligente des différents mots de la langue, la qualité du dictionnaire en question joue également. Sa couverture lexicale et l'exactitude des informations qu'il renferme ont des conséquences directes sur la fiabilité des concordanciers.

Malheureusement, la non disponibilité des dictionnaires riches arabe utilisé dans ce domaine, nous cause des insuffisances au niveau des résultats, car aujourd'hui ce genre de dictionnaires ne sont plus gratuits, ni à la disposition pour le but de recherche, mais cela nous a pas empêché de tenter et produire notre concordancier, on a utilisé quelques dictionnaires qu'on voyer les plus nécessaire par rapport au corpus(voir '**Table 3**', chapitre 2), dont chaque dictionnaire est juste un échantillon de chaque catégorie, On a créé cinq sous format xml qui sont :

(1) Dictionnaire des pays
(2) Dictionnaire des capitales
(3) Dictionnaire des noms propre
(4) Dictionnaire des racines
(5) Dictionnaire des mots outils

Ces dictionnaires sont utiles en termes de vérification, avant d'utilisé le concordancier on doit les charger, afin qu'ils permettent de voir si le mot appartient à la langue ou pas, c'est évident que la richesse de ces dictionnaires joue un rôle important, plus ils sont riche plus on obtient le résultat souhaité.

5. Environnement de travail

Le choix du langage de programmation représente une étape très importante dans la réalisation de n'importe quel logiciel. C'est dans cette étape qu'on fait la correspondance entre les solutions que le langage nous offre et les résultats souhaités.

Pour implémenter notre logiciel ConcArabe, nous avons utilisé :

- **Python (Version 3.4.4) :** Python est un langage de programmation interprété, c'est-à-dire que les instructions que vous lui envoyez sont « transcrites » en langage machine au fur et à mesure de leur lecture (à ne pas confondre avec un langage compilé). Il permet de créer toutes sortes de programmes, comme des jeux, des logiciels, des progiciels, etc. On peut aussi lui associer des bibliothèques afin d'étendre ses possibilités, parmi ses avantages la simplicité, (on ne passe pas par une étape de compilation avant d'exécuter son programme) et la portabilité, c'est à dire qu'il peut fonctionner sous différents systèmes d'exploitation (Windows, Linux, Mac OS X,…) cette version est disponible depuis 16 mars 2014, elle peut être téléchargée gratuitement à partir du lien suivant :

  https://www.python.org/download/releases/3.4.4/

- **lxml** : Est la bibliothèque la plus riche en caractéristique (fonction), elle permet de faciliter le traitement des fichiers XML et le HTML dans la langue de programmation *python*, La dernière sortie marche avec toutes les versions *python* de 2.6 à 3.5.

- **Qt creator** : ressemble à un framework lorsqu'on l'utilise pour concevoir des interfaces graphiques ou que l'on conçoit l'architecture de son application, Qt permet la portabilité des applications qui n'utilisent que ses composants par simple recompilation du code source. Les environnements supportés sont les Unix (dont GNU/Linux) qui utilisent le système graphique X Window System ou Wayland, Windows, Mac OS X et également Tizen. Le fait d'être une bibliothèque logicielle multiplateforme attire un grand nombre de personnes qui ont donc l'occasion de diffuser leurs programmes sur les principaux OS existants. Qt supporte des bindings avec plus d'une dizaine de langages autres que le C++, comme Java, Python, Ruby, Ada, C#, Pascal, Perl, Common Lisp, etc.

6. Technique de parse

XML permet de définir la structure du document uniquement, ce qui permet d'une part de pouvoir définir séparément la présentation de ce document, d'autre part d'être capable de récupérer les données présentes dans le document pour les utiliser.
Toutefois la récupération des données encapsulées dans le document nécessite un outil appelé *analyseur syntaxique* (en anglais *parser*), permettant de parcourir le document et d'en extraire les informations qu'il contient.

Dans ce qui suit nous allons décrire les fonctionnalités de notre concordancier qui est construit principalement de cinq taches fondamentales en respectant ce qui nous a été demandé.

On va montrer aussi les méthodes utilisé pour le bon déroulement des taches ainsi l'ajout de quelques options qu'on voyer nécessaire.

7. Concordancier

Notre concordancier ConcArabe est utilisé dans l'analyse des corpus en linguistique, Il prend en charge la langue arabe qui est gérée par le codage UTF8, il peut analyser des fichiers simples ou des corpus de plusieurs fichiers. Le logiciel prend en charge le format .xml, donc il peut travailler avec les textes balisés, sa tache principal consiste à faire la recherche dans un corpus d'un mot accompagné de son contexte, que ce soit pour attester son usage ou l'étudier, pour les linguistiques les concordanciers les sert à analyser des discours, l'utilisé dans l'enseignement/apprentissage de langues, ainsi qu'une recherche général. Dans ce qui suit nous allons présenter les fonctionnalités adoptées sur l'ConcArabe.

Remarque

-*Afficher fichier* est un onglet de visualisation de fichier chargé, il permet juste l'affichage du contenu. Ainsi *Fichiers corpus* qui est en bas de la fenêtre affiche le nom du corpus chargé.

7.1 Concordance

7.1.1 Avec l'option *recherche par mot*

Cet outil permet de faire la recherche par mot à qui vous intéressé sur votre corpus, le résultat obtenus seras une suite de lignes, dont chaque ligne représente l'affichage des contextes droit et gauche de la séquence recherchée, en se basant sur la spécification des nombre de mot a affiché sur chaque côté avec l'option « *taille de recherche* ».ce dernier nous permet d'avoir un peu plus de contexte. Dans cet onglet vous pouvez trier les résultats par « kwic », c'est-à-dire, identifier plusieurs mots à gauche et à droite des résultats, la figure suivante montre le résultat de recherche du mot « الخلايا » en précisant le nombre de mot a affiché dans le contexte pour chaque côté qui est égal à 7.

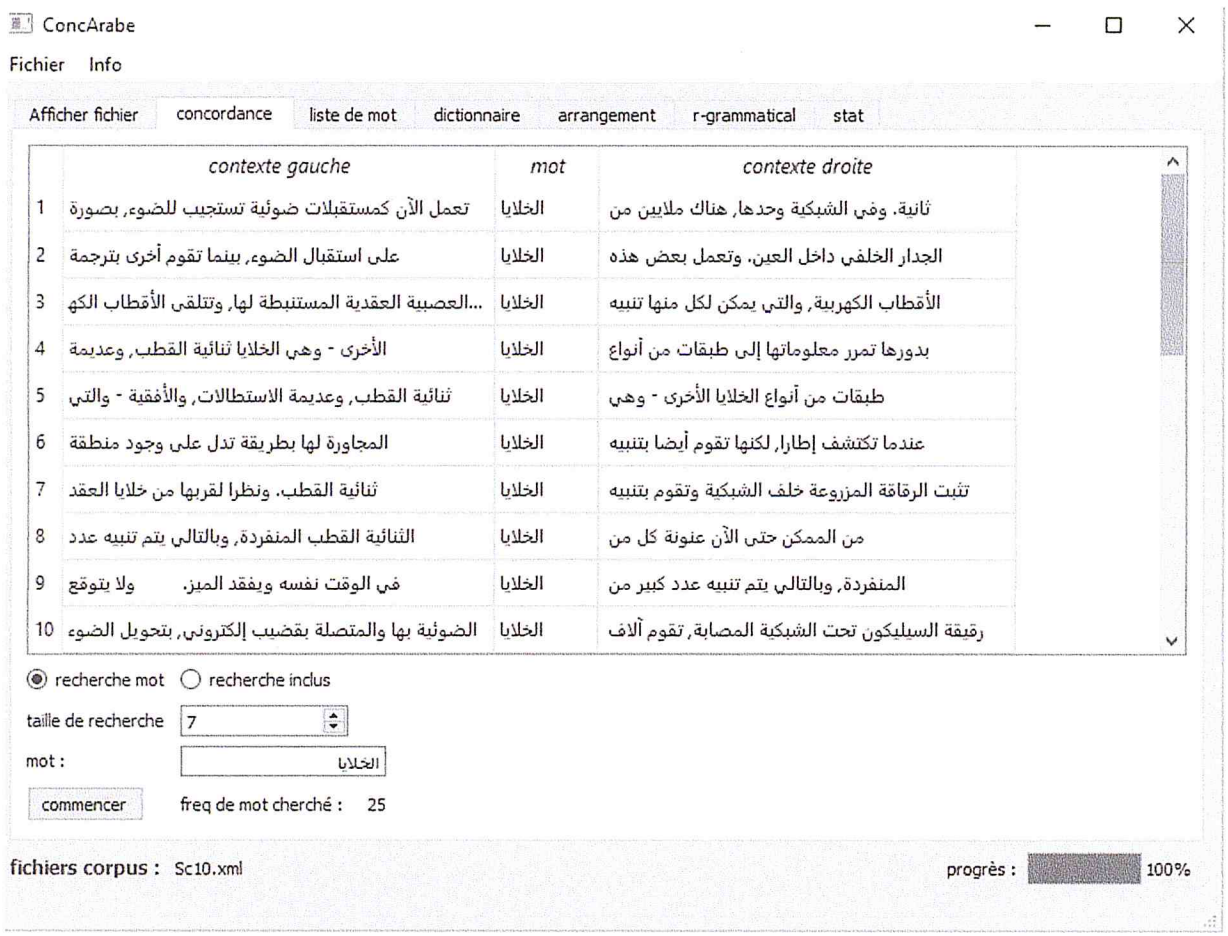

***Figure 17** : Table de concordance de ConcArabe*

7.1.2 Avec L'option *recherche inclus*

Cette option fournit la même tache que précédente, sauf avec une expression rationnelle sur les lettres, comme le mot « قال », qui reconnaîtra tout mot commençant par « قال ».

Remarque

L'option *freq de mot* affiche le nombre d'occurrence de la séquence recherchée.

7.2 Liste des mots

L'outil permet de créer un index des mots contenus dans le corpus et de les compter. Il affichera par la suite la liste de tous les mots avec leur information sur la fréquence d'apparition ainsi que leur pourcentage. Pour effectuer cette opération il suffit juste de cliquer sur « *commencer liste* » sans écrire aucun mot, la figure suivante montre le résultat du parcours de corpus.

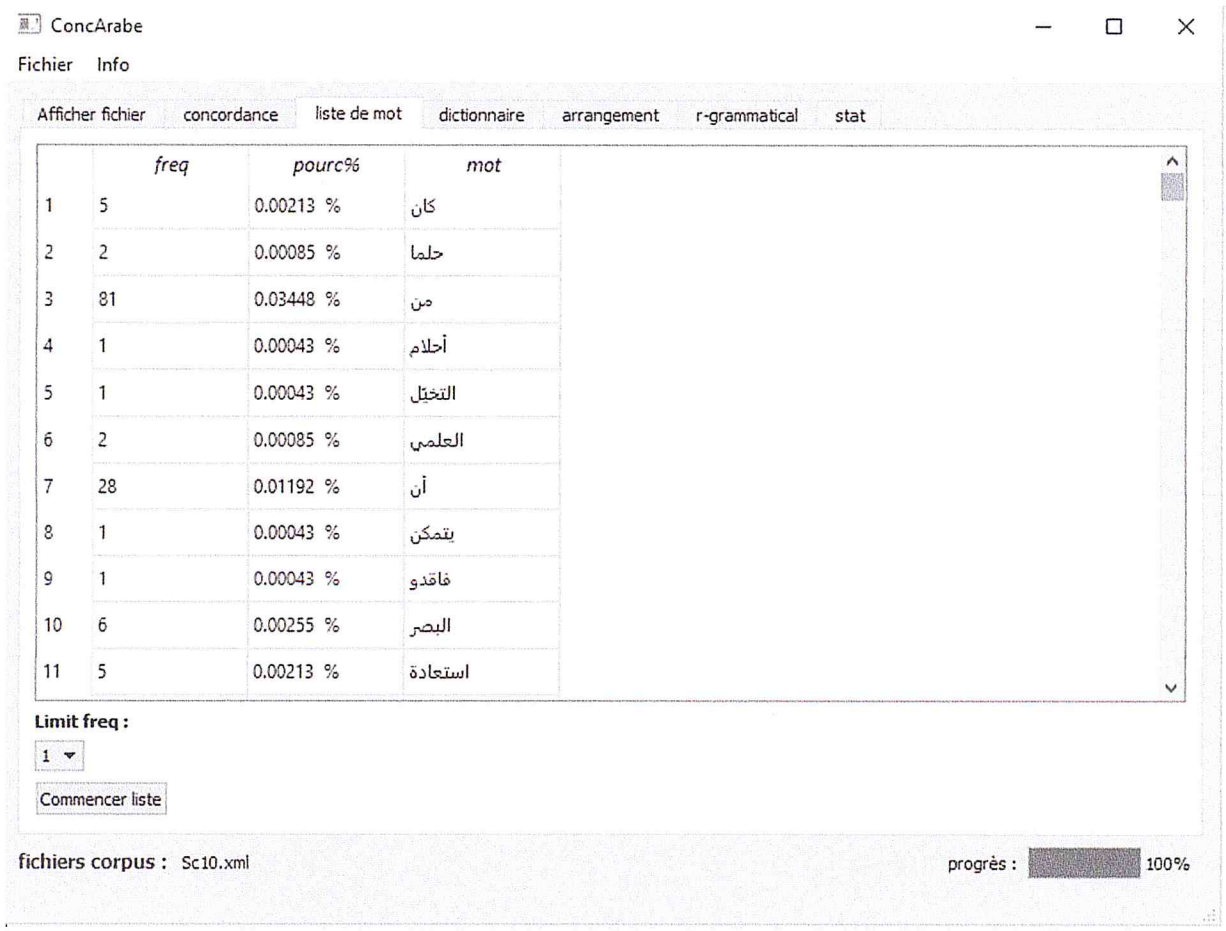

Figure 18 : *Liste de mot de ConcArabe*

L'option *limit freq* nous permet de manipuler les résultats au niveau des fréquences d'apparition de chaque mot, c'est-à-dire trier les résultats par rapport a une certaines fréquence, la figure suivante montre l'utilisation de cet option, dans ce cas on a limité la fréquence a 4, c'est-à-dire la table affiche les mots qui ont des occurrences supérieur ou égal à 4.

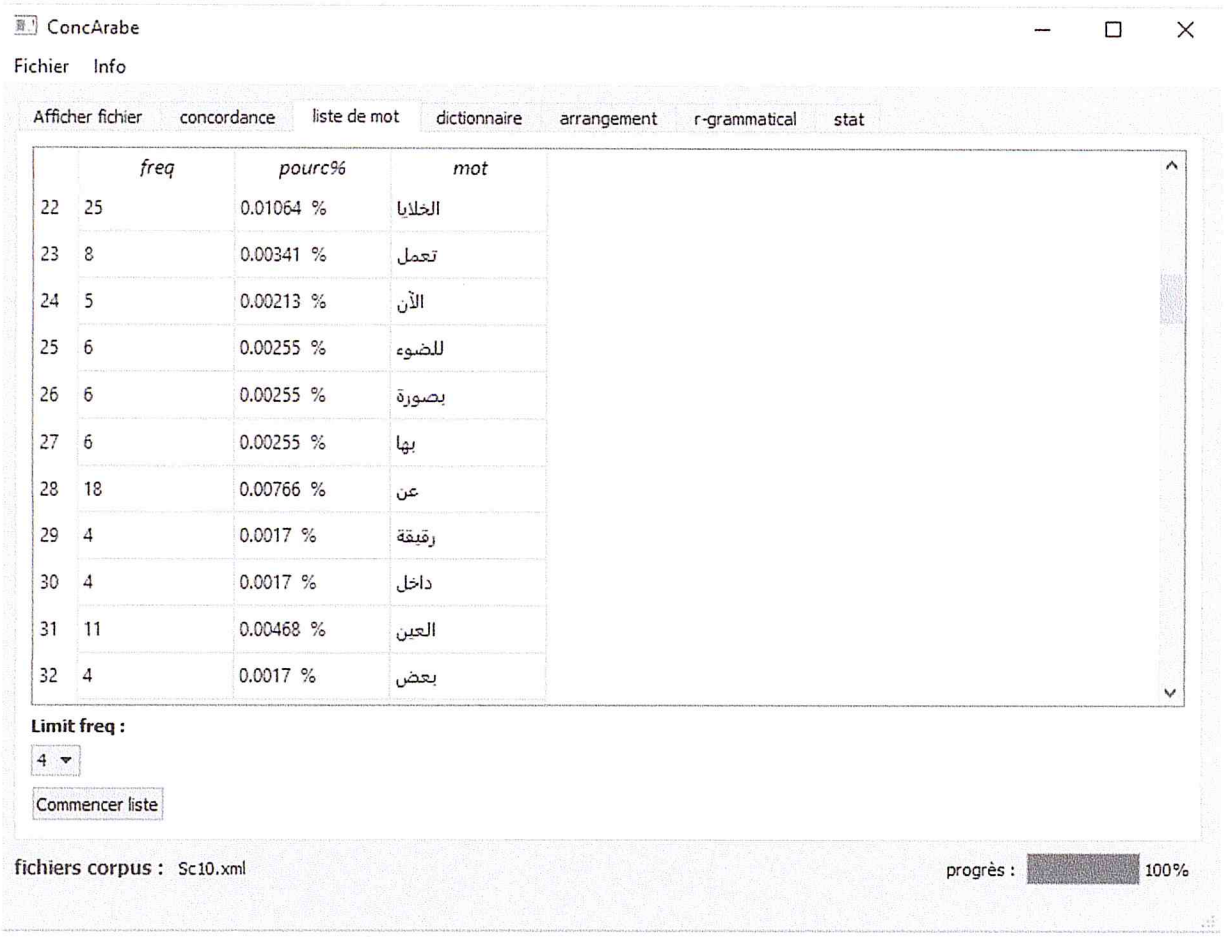

***Figure 19** : listes de mot limité de ConArab*

7.3 Le dictionnaire

7.3.1 Avec L'option *Premier dernier*

Le dictionnaire a pour but de chercher tous les mots qui existent dans le corpus, cette opération est effectué par certains critères, l'utilisateur doit déterminer sur le champs de saisie *indice de recherche :* la première et la dernière lettre des mots a cherché sur le corpus, tout en

précisant (si voulu) la longueur souhaité de ces mots, par l'option *long,* qui est limité à 12 lettres maximum, le résultat obtenus sera une liste des termes qui respectent les critères de l'utilisateur avec leur fréquences d'apparition dans le corpus. La figure suivante montre la liste de tous les mots qui existent dans le corpus dont chaque mot commence par le préfixe « ا » et se termine par le suffixe « ل » avec leur fréquence d'apparition.

Remarque

Dans ce cas on n'a pas précisé la longueur des mots.

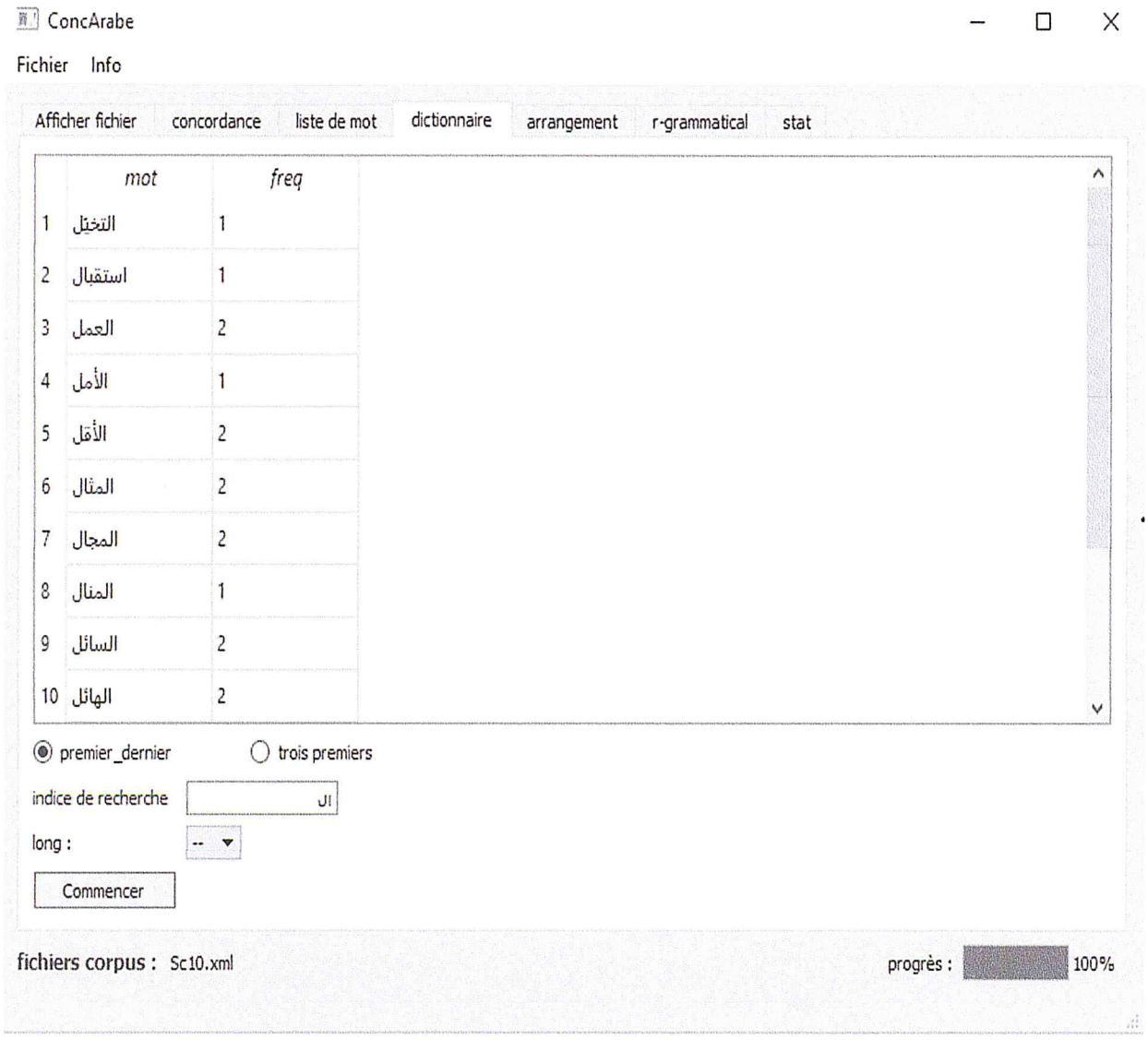

Figure 20 : *Table de dictionnaire par la méthode premier-dernier*

7.3.2 Avec l'option *trois premiers*

Cette option fournit la même tache que précédente, sauf la différence se situe dans l'*indice de recherche*, l'utilisateur cette fois doit déterminer les trois premiers lettres des mots a cherché sur le corpus, La figure suivante montre la liste de tous les mots qui existent dans le corpus avec leur fréquences d'apparition, dont chaque mot commence par le préfixe constituer de trois lettres « الأ » tout en précisant la longueur des mots qui est égal à 5.

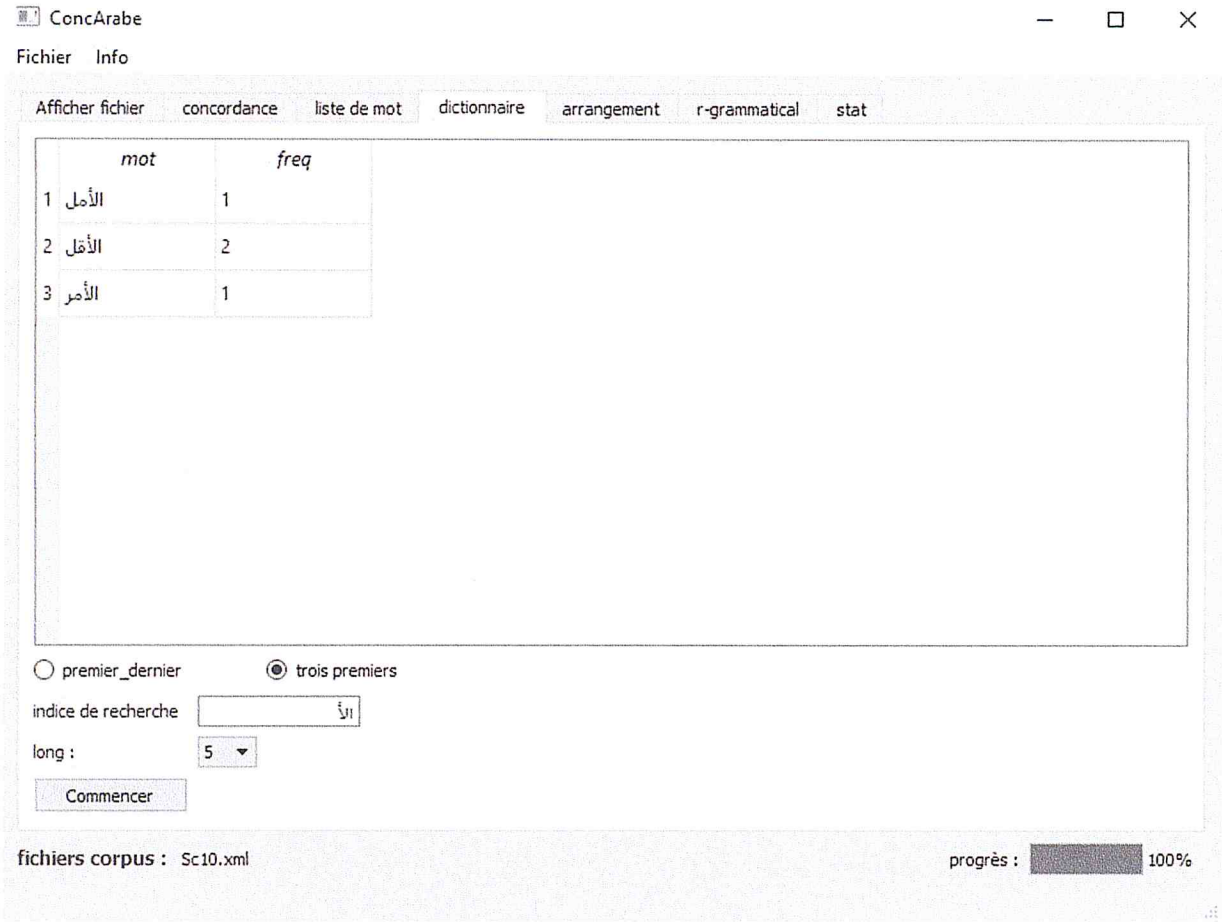

Figure 21 : *Table de dictionnaire par la méthode trois-premiers*

7.4 Arrangement

Cet outil permet de chercher des mots qui ont des connexions ou des associations plus ou moins directes dans un texte de la séquence saisie. L'utilisateur doit déterminer le nombre de mot à gauche (respectivement à droite) à apparaitre, cet option permet la propagation des résultats, par conséquence l'outil génère une liste des mots voisins avec leur fréquences d'apparition de chaque côté. La figure suivante montre le résultat de l'arrangement du mot « الأمل».

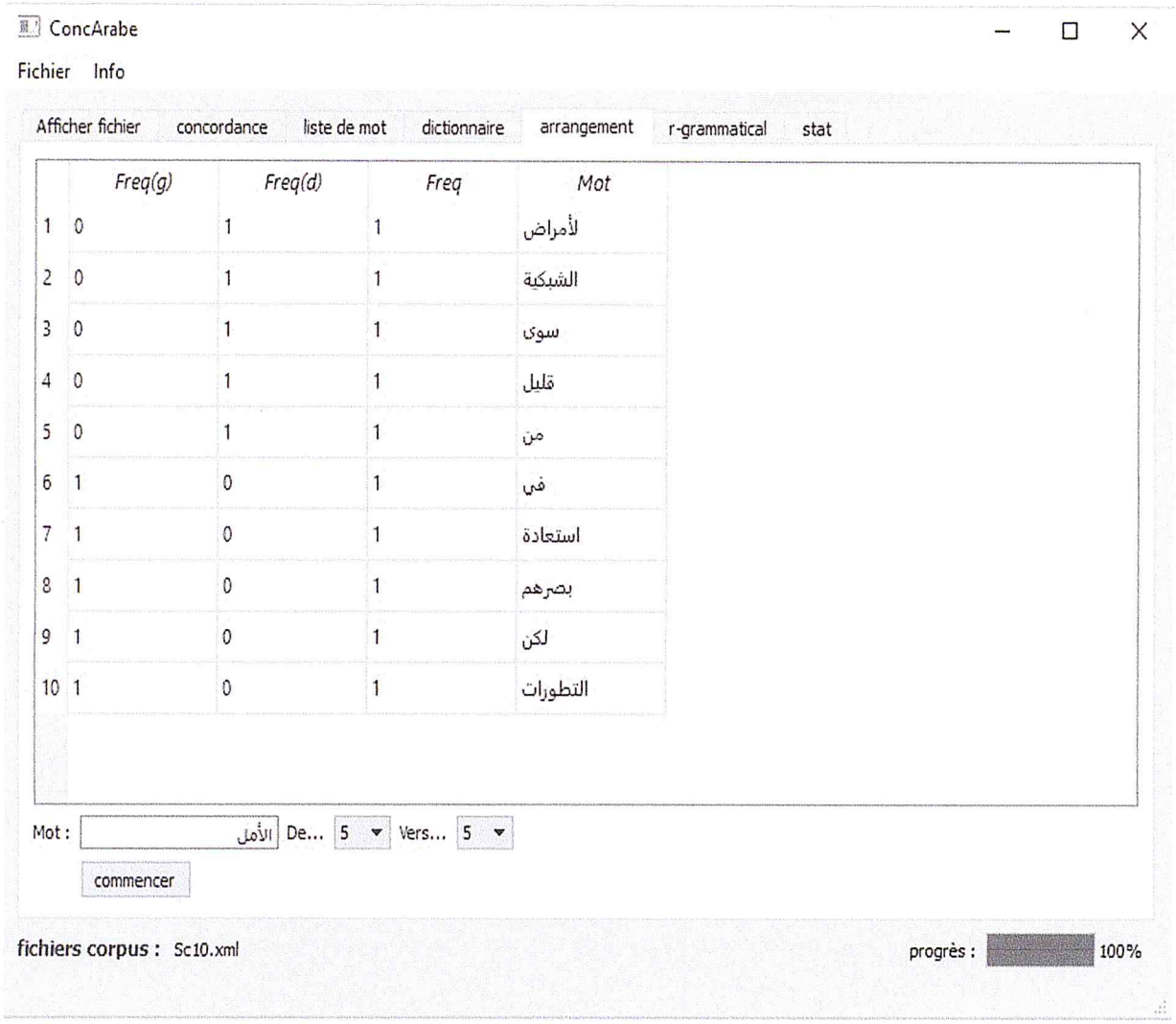

Figure 22 : *Table de l'arrangement du mot « الأمل »*

7.5 Recherche grammatical

Parmi nos réflexions sur ce projet se réside dans la fusion du concordancier avec l'analyseur morphologique.

Notre outil permet plusieurs fonctionnalités qui veillent à non seulement enrichir le dictionnaire des racines en plus avoir « مشتقات الكلمة».

Le résultat dans cet outil va prendre le même principe que celle de la concordance, la différence se situe sur la séquence cherché elle sera cette fois-ci les mots obtenus dans « مشتقات الكلمة» du mot saisie.

Le résultat obtenu sera une suite de lignes, dont chaque ligne représente l'affichage des contextes droit et gauche des séquences trouvés.

L'option *morphologie de mot* veille à afficher des informations supplémentaires du mot saisie tel que le schème, la catégorie grammaticale, catégorie du mot etc.

L'option *charger dictionnaire* permet d'extraire les dictionnaires nécessaires pour la recherche grammaticale.

L'option *ajouter au dictionnaire* permet de donner la main à l'expert linguistique qui va ajouter des mots sur le dictionnaire des racines si nécessaire. Les deux figures suivantes montrent la recherche grammaticale des mots « مجد » et « حاضر».

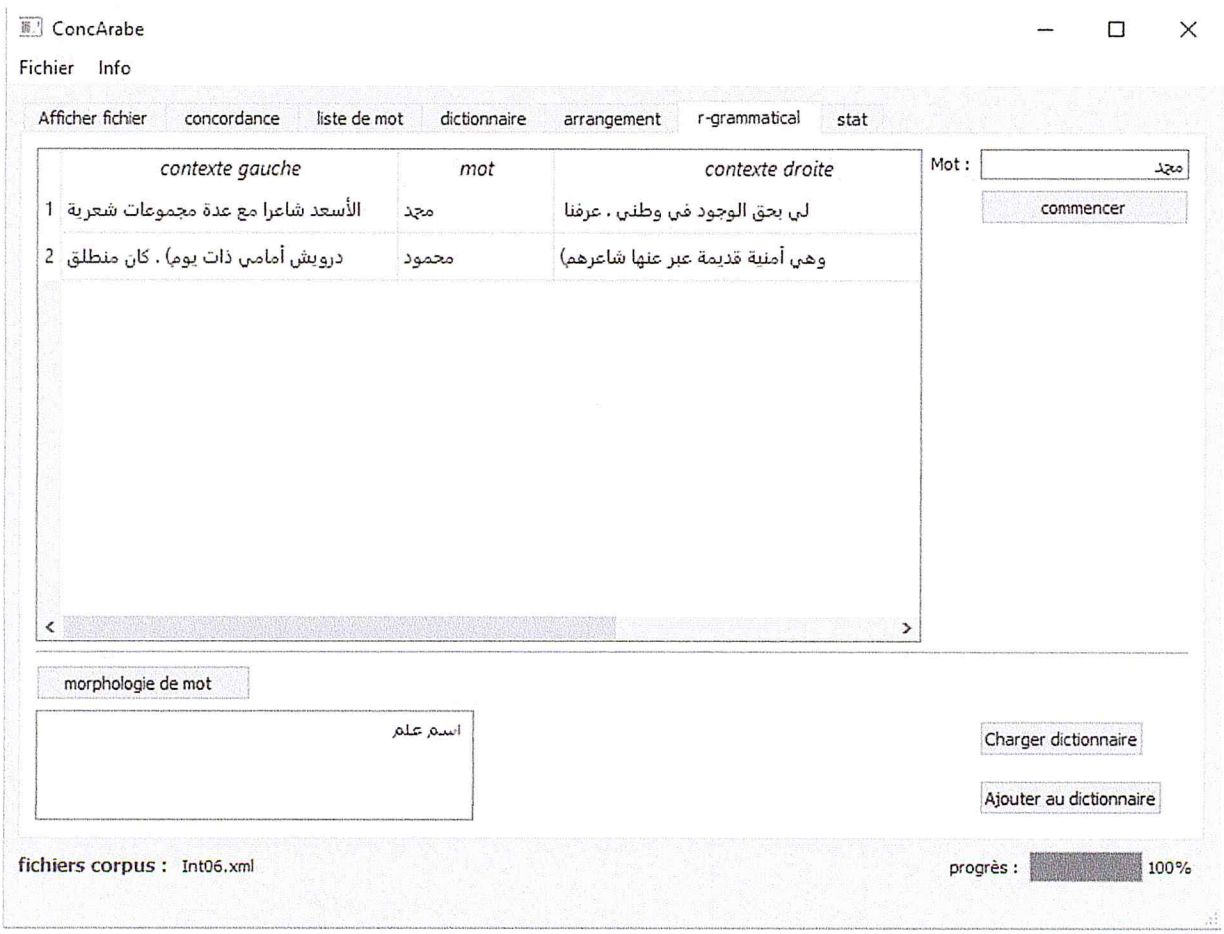

Figure 23 : *la recherche grammaticale du mot « مجد»*

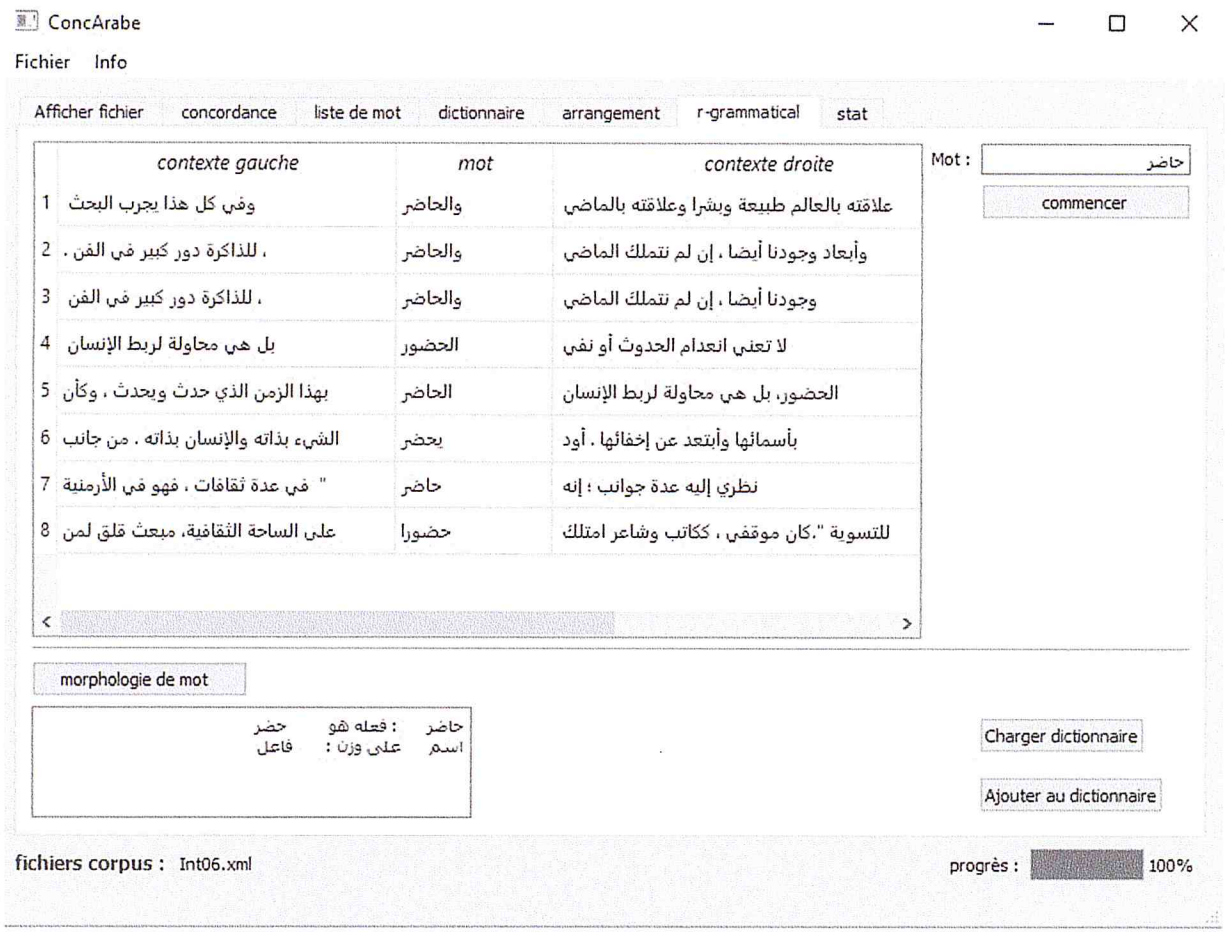

Figure 24 : *la recherche grammaticale du mot «حاضر»*

7.6 Statistiques et information

Les informations sur l'origine du texte source telle que le titre et l'auteur, etc. semble nécessaire pour n'importe quel chercheur sur la langue, car on sait très bien que la fiabilité des informations joue un rôle important dans la valorisation des informations obtenus.

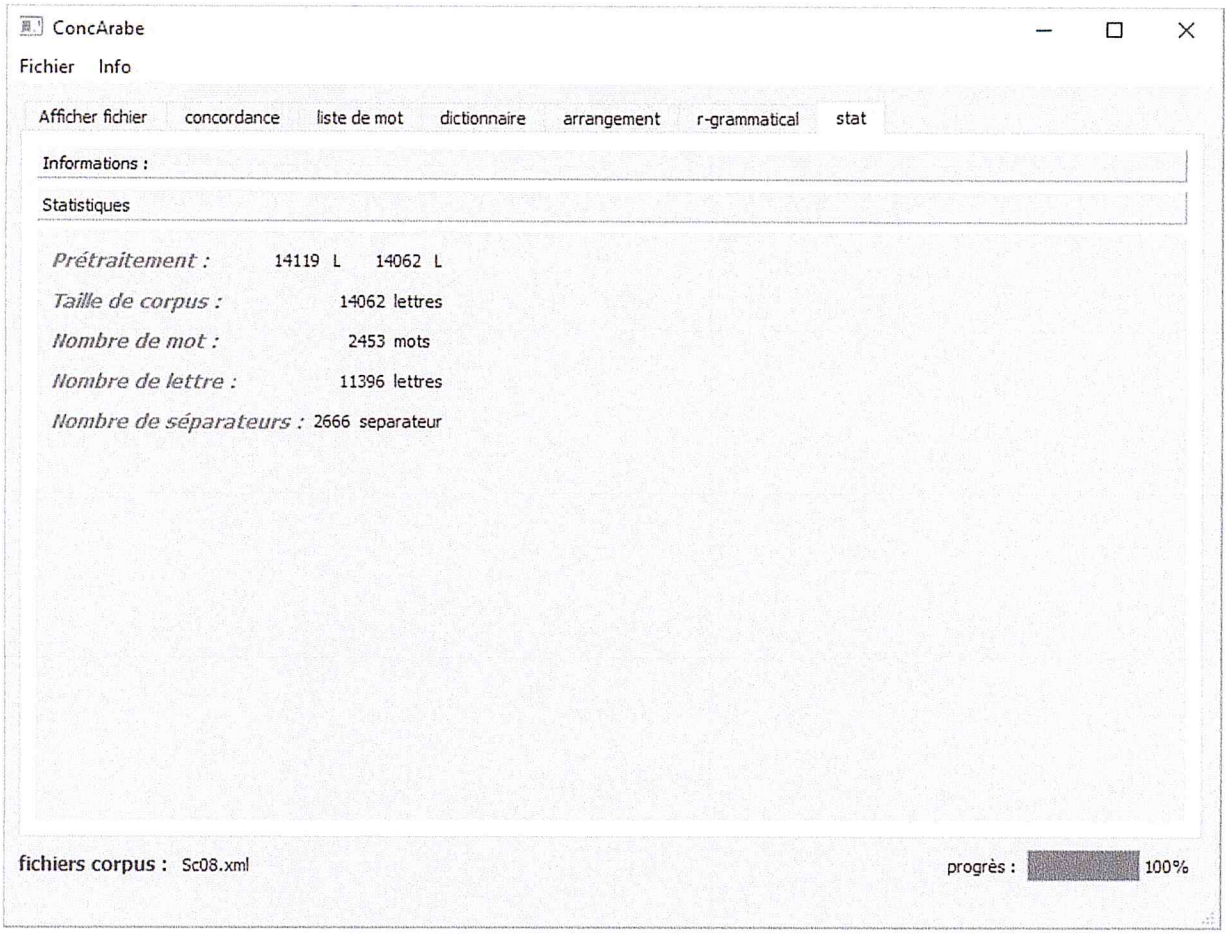

Figure 25 : *les statistiques de corpus chargé*

Remarque

Le ***prétraitement*** montre la différence entre le nombre de lettres lors du chargement de corpus, et le nombre de lettres de corpus chargé après avoir effectué le prétraitement.

Figure 26 : *les informations concernant le corpus chargé*

8. Conclusion

Dans ce chapitre nous avons réalisé une série des fonctions pour analyser et exploiter le corpus de Latifa el-solaiti, nous avons effectué les fonctions de base demandé d'un concordancier en ajoutant une amélioration qu'on voyait nécessaire pour les linguistes, c'est-à-dire des fonctionnalités supplémentaires telles que le dictionnaire et l'arrangement d'un mot. Cela nous a permis d'approfondir nos connaissances sur les fonctions attendues par les linguistes, ainsi les réaliser nous a permis d'avoir des compétences supplémentaires sur la programmation d'un tel logiciel qui traite des texte en arabe.

Conclusion général

Conclusion général

La réalisation des concordanciers était jusqu'à ces dernières années un travail de grande envergure envisageable uniquement pour les œuvres pérennes. Le traitement automatique a facilité la tâche et a étendu leurs champs d'application à de nombreuses disciplines scientifiques. Dans le cas de la langue arabe, l'aboutissement d'un concordancier électronique nécessite un travail préalable faisant appel à des ressources lexicales et des outils pour reconnaitre la morphologie de la langue.

Au cours de ce projet, nous avons essayé d'apporter une amélioration pour les concordanciers qui traitent des corpus arabe, cela en se basant sur l'analyse morphologique qui est considéré parmi les outils les plus efficaces pour développer et exploiter un corpus. Nous avons implémenté la méthode de découpage de mot du modèle proposé par (Sadik et al. 2007) ; en évitant ses insuffisances, nous avons proposé une méthode qui mène à des cas d'ambigüité plus réduites.

A l'issue de l'étape de découpage, nous avons utilisé un échantillon de différents dictionnaires pour l'attestation de l'appartenance de chaque composant du mot à la langue arabe et pour l'extraction des traits morphologique correspondants, tout en connaissant que l'existence des mots voyellès sur le corpus non-voyellès nous a obligé à traiter ce dernier.

Par conséquent notre concordancier tient a non seulement connaitre la grammaire d'un mot mais aussi de chercher dans la concordance tout séquence dérivé d'un mot saisie, en donnant le contexte de chaque mot résultant.

On espère par la suite que notre méthode sera appliquée sur les concordanciers qui traitent les corpus arabe dans le futur.

L'apport que nous avons essayé d'apporter au domaine de la réalisation des concordanciers pour la langue arabe est petit, ce domaine est très vaste es nécessite la collaboration de plusieurs travaux pour proposer des solutions aux différentes problèmes existant, pour rendre le texte à analyser plus facile à exploiter.

Bibliographie

-(AL-Sulaiti et al. 2006)- Latifa AL-Sulaiti, Eric Atwell. : . Designing and Developing a Corpus of Contemporary Arabic. England.university of Leeds School of computing, 2006.

-(Abbes, 2004 : 23) - Ramzi Abbes. La conception et la réalisation d'un concordancier électronique pour l'arabe. Lyon France. Université de lumière. L'institut national des - sciences appliquées de Lyon.2004.p.23

-(EL Mezouar et al. 2013) - Vers un modèle pour le recueil d'un corpus d'apprentissage d'une langue étrangère peu dotée de ressources, Ecole de Science et Génie Université Al Akhawayn Ifrane, Maroc, 2013, p.23

-(Slaby, 1979) - Slaby wolfgang. a. Concordances to the Greek New Testament and to the Bad Quartos to the Works of Shakespeare : two Strategies for an Automatic Selection of Context. In : R.J.Dilligan E.G.Bedford. Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research, Birmingham. p 117-127.

-(Larousse, 1994 p. 108) - Dictionnaire de linguistique et des sciences du langage (1994), Paris : Larousse, p. 108.

-(LANGLOIS, 1996) - Langlois lucie. Bi-texte, Bi-concordance et collocation [en ligne]. Thèse en Traduction. Ottawa : université d'Ottawa, 1996, [En Ligne] Disponible sur http://www.dico.uottawa.ca/theses/langlois/ll-debut.htm

-(Zaafrani, 2002) – Riadh zaafrani. Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère PhD. Dissertation, ENSSIB/université Lyon2.2002

-(Abbes, 2004) - Ramzi Abbes. La conception et la réalisation d'un concordancier électronique pour l'arabe. Lyon France. Université de lumière. L'institut national des sciences appliquées de Lyon.2004.p.23

-(Sekhraoui, 1995) - Sekhraoui majid. Concordance: historique, méthode et pratique. Paris:Paris 3, 1995, 204 p.

-(Dundee, 2004) - Dundee. What is concordance? En ligne ; Disponible sur: <http://www.dundee.ac.uk/english/wics/cncintro.htm>.

-(Tribble et Jones, 1997) - Tribble chris et Jones glyn. Concordances In The Classroom : a resource book for teachers. 2e édition. Houston: Athelstan, 1997, p.114

-(Cameron, 1996) - Cameron keith. De la concordance au dictionnaire. Dictionnaires électroniques du français des XVIème et XVIIème siècles, 14-15 juin 1996, Université Blaise

Pascal, Clermont-Ferrand, . [En Ligne] Disponible sur <http://www.chass.utoronto.ca/~wulfric/siehlda/clermont/index.html>

-(Abbes, 1999) - Abbes ramzi. Conception et réalisation d'un prototype de concordancier électronique de la langue arabe. DEA Science de l'Information de la Communication. Lyon:ENSSIB, 1999, p.100.

-(Cobuild, 1987) - Cobuild, Dictionary of the english language. first édition.: Collin's, 1987

-(Sinclair, 1991) - Sinclair john. Corpus, Concordance and Collocation. Oxford: Oxford University, 1991, p.180.

*-(Rezeau, 2008 : 3) - **Joseph** Rézeau. Applications des concordanciers à l'enseignement de la grammaire anglaise en DEUG. 15-18 | 1997.p. 3*

-(Garrigues, 1997) - Garrigues mylene. Lemmatized concordances of complex utterances: application to language learning. In: A.K. Korsvold and B. Rüschoff. New technologies in language learning and teaching. Strasbourg:Council of Europe Publishing, 1997, p.87-98.

-(Laporte, 2000) - Laporte eric. Mot et niveau lexical. In: Jean-marie pierrel. Ingénierie des langues. Paris:Hermes, 2000, p. 25-46.

-(Silberztein, 1993) - SILBERZTEIN MAX. Dictionnaires électroniques et analyse automatique de textes. le système INTEX,. Paris; Milan;Barcelone: Masson, 1993, p.233

*-(Rezeau, 2008 : 2) - **Joseph** Rézeau. Applications des concordanciers à l'enseignement de la grammaire anglaise en DEUG. 15-18 | 1997.p. 2*

-(Barreca et al. ,2004. P.499) - Giulia Barreca, George Christodoulides. un concordancier multi-niveaux et multimédia pour des corpus oraux.p.499 -500.

-(Barlow, 1998) – M.Barlow, MonoConc, Houston TX :Athelstan, 1998.

-(Scott, 2008) M.Scott, Introduction to wordsmith tools : onlibe manuel, 2008.

-(Boualem et al. 1999) – M.Boualem, M.Leisher et B. Ogden, Concordancer for Arabic, Actes de la conférence internationale ATLAS'99, 1999.

-(Fairon, 2001) – C.Fairon, GlossaNet en ligne : service de concordances automatique, CENTAL, Université Catholique de Louvain, 2001.

-(Mestivier, 2005) - Alexandra Mestivier (Volanschi), Qu'est-ce qu'un concordancier et à quoi il sert ?, cours, 2005-2006.

-(Simard et al. 1993. P.3) - Michel Simard, George F. Foster, François Perrault, TransSearch: un concordancier bilingue, october1993, p.3)

-(Thierry et al. 2010) – thierry chanier, maud ciekkanski, Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage, 2010.

-(Ferreira, 2002) – En ligne ; Disponible sur : http://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2002.ferreiraqueiros_rm&part=57607

-(Condamines et al. 1999. p.26-36) - Condamines, Marie-Paule PERY-WOODLEY et Cécile FABRE, Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative, 1999. p.26-36

-(Habert, 1997) - Benoît Habert, Adeline Nazarenko, André Salem, Les linguistiques de corpus, Paris, 1997, p.3

-(Abu-absi, 1986) - Samir Abu-Absi, La modernisation de la langue arabe : problèmes et perspectives, la linguistique anthropologique, automne 1986.

-(Chouikha, 2010) – chouikha, la langue arabe, son histoire, son originalité, et son influence.juin2010

-(Baccouche, 2009 : 19) – Taïeb baccouche « Dynamique de la langue Arabe », Synergies Tunisie, n°1, 2009, p.19.

-(Holes, 1995) - HOLES, C., Modern Arabic: Structures, Functions and Varieties, London/New-York, Editions Longman, 1995.

-(Marsi et al., 2005) – Marsi, E., Bosch, A.v.d. et Soudi, A., Memory-based morphological analysis generation and part-of-speech tagging of Arabic". In Computational Approaches to Semitic Languages Workshop Proceedings, University of Michigan Ann Arbor, Michigan, USA, 2005.

-(Al Ameed, et al., 2008) - Al Ameed, H., Al Ketbi, S., Al Kaabi, A., Al Shebli, K., Al Shamsi, N., AL Nuaimi, N. and Al Muhairi, S., UNITED ARAB EMIRATES UNIVERSITY - Faculty of Information Technology, Arabic Search Engines Improvement: A New Approach using Search Key Expansion Derived from Arabic Synonyms Structure. 2008.

-(Kouloughli, 1991) – kouloughli, lexique fondamental de l'arabe standard moderne, paris : le Harmattan, 1991.p.287.

-(khoja et al. 2001) - S.Khoja, R.Garside, G.Knowles, A tagset for the morph-syntactic tagging of arabic, actes de la conférence internationale corpus linguistics 2001, lancaster(2001).

-(El-Dahdeh, 1999) – Antoine El-Dahdeh, a dictionary of universal arabic grammar, 1999

-(Cohen, 1970) – D.Cohen, Essai d'une analyse automatique de l'arabe. Dans : David Cohen. Etudes de linguistique sémitique et arabe. Paris : Mouton, p.49-78, 1970.

-(Gaudin et Guespin, 2000) Gaudin Francois et Guespin Louis. Initiation à la lexicologie française : De la néologie aux dictionnaires. Bruxelles : Edition duculot, 2000.p. 355.

-(Mitterand, 2000) Mitterand Henri. Les mots français. 10e édition. Paris : PUF, 2000.p.127

-(Abbes, 2004 : 39) - Ramzi Abbes. La conception et la réalisation d'un concordancier électronique pour l'arabe. Lyon France. Université de lumière. L'institut national des - sciences appliquées de Lyon.2004.p.39

- (Descles et al. 1983) Descles Jean-pierre, Adaab h, Dichy joseph, Kouloughli djamel eddine and Ziadah M.S. conception d'un synthètiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement Assisté par Ordinateur. Paris : Rapport rédigé à la demande du Ministère des Affaires étrangères. 1983.

-(Dichy, 1990) Dichy joseph. L'Ecriture dans la représentation de la langue : la lettre et le mot en arabe. Thèse pour le doctorat d'état (ès Lettres). Lyon : Université Lumière-Lyon 2, 1990.

-(Zaafrani, 2002) – Riadh zaafrani. Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère PhD. Dissertation, ENSSIB/université Lyon2.2002

-(Sadik et al. 2007 : 43-46) Bessou Sadik, Louail Mohamed, Refoufi Allaoua Kadem Zehour& Touahria Mohamed. Un système de lemmatisation pour les applications de TALN, 2007.Lp.43-46.

- (Cohen, 1970) Cohen david. Essai d'une analyse automatique de l'arabe. In: David Cohen. Etudes de linguistique sémitique et arabe. Paris : Mouton, 1970.

-(Dichy J, 1997) Dichy joseph, Pour une lexicomatique de l'arabe : l'unité lexicale simple de l'inventaier du mot. META - journal de traduction, 1997.

- (Hoceini, 2002) : Un système d'analyse morphologique de la langue arabe, mémoire magister, école nationale supérieure d'informatique, (2002).

- (Balou, 2003) S. Baloul : Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé, thèse de doctorat, université du Mans, 2003.

-(Darwish, 2003) K.Darwish, Probabilistic Methods for Searching OCR-Degraded Arabic Text, Doctoral dissertation, University of Maryland, 2003.

-(Ludovic, 2008) –Roland Ludovic, Structurez vos données avec XML, 2008.