République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb Blida

Faculté des sciences

Département informatique

Mémoire de fin d'étude

Pour l'obtention du diplôme de master en informatique

Spécialité : Traitement automatique de la langue

Thème :

Vers une plateforme d'analyse et traitement de corpus en langue arabe

Mémoire présenté par :

- **Benzineb Asma**
- **Nadour sabrina**

Promotrice : Mme Oukid Lamia Encadreur : Mr Djaidja Yacine

Date de la Soutenance le : **01 Octobre 2019**

2018/2019

Résumé

Les plateformes de traitement statistique des langues permettent d'offrir les fonctionnalités les plus courantes pour un traitement automatique de corpus de données volumineux impraticable à une manipulation manuelle et ouvrant ainsi la voie à l'analyse de millions de textes à la fois. Plusieurs plateformes existent en ligne mais elles restent limitées pour la langue arabe. Notre travail a pour objet de concevoir une plateforme permettant de fournir des traitements sur des corpus en langue arabe. Dans ce mémoire nous proposons une plateforme fournissant plusieurs fonctionnalités : extraction des principales caractéristiques de vocabulaires de corpus décrites à partir du tableau lexical, des index hiérarchiques, des concordances, le calcul des distances entre textes et leur classification, collocation et extraction de groupe de formes …etc., ces outils sont réalisés par le prétraitement classique ou par lexicométrie (une discipline relativement récente). Nous avons finalement développé et mis en œuvre une plateforme réunissant ces différentes fonctionnalités qui concordent aux besoins de l'académie de la langue arabe.

Mots-clés : Langue arabe, plateforme, statistique, concordances , lexicométrie

Abstract

The statistical language processing platforms allow to offer the most common functionalities for an automatic processing of large corpora of data that is impracticable for manual manipulation and thus opening the way to the analysis of millions of texts at a time. Several platforms exist online but they remain limited for the Arabic language. Our work aims to design a platform to provide treatments on corpora in Arabic. In this dissertation we propose a platform providing several functionalities: extraction of the main characteristics of corpus vocabularies described from the lexical table, hierarchical indexes, concordances, the computation of the distances between texts and their classification, collocation and group extraction of forms ... etc., these tools are made by conventional pretreatment or by lexicometry (a relatively recent discipline). We finally developed and implemented a platform bringing together these different functionalities that match the needs of the Arabic language academy.

Keywords: arabic language, platforms, statistical, concordances

ملخص

تسمح منصات معالجة اللغة الإحصائية بتقديم الوظائف الأكثر شيوعًا للمعالجة التلقائية لمجموعة كبيرة من البيانات الغير العملية للمعالجة اليدوية وبالتالي فتح الطريق لتحليل ملايين النصوص في وقت واحد. توجد العديد من المنصات على الإنترنت لكنها تظل محدودة بالنسبة للغة العربية. يهدف عملنا إلى تصميم منصة لتقديم معالجة الية باللغة العربية. في هذه الأطروحة نقترح منصة توفر العديد من الوظائف: استخراج الخصائص الرئيسية لمفردات مجموع النصوص الموصوفة من الجدول المعجمي، فهارس التسلسل الهرمي، التوافق، حساب المسافات بين النصوص وتصنيفها، وترتيبها، واستخراج المجموعة من الأشكال ... وما إلى ذلك، يتم إجراء هذه الأدوات عن طريق المعالجة التقليدية أو عن طريق المعجم (تخصص حديث نسبيا).

أخيرًا، قمنا بتطوير وتنفيذ نظام أساسي يجمع بين هذه الوظائف المختلفة التي تتوافق مع احتياجات أكاديمية اللغة العربية

الكلمات الرئيسية: اللغة العربية، منصة، المعجم اللغوي، التوافقية اللفظية

REMERCIEMENT

Ce mémoire a été réalisé dans le cadre de fin d'études pour l'obtention du diplôme d'ingénieur en informatique. Nous tenons à remercier d'abord dieu de nous avoir donné force et santé pour mener ce travail à terme. Nous tenons à remercier nos parents pour leurs sacrifices, soutien et compréhension durant toutes nos années d'études. Nous tenons à exprimer nos profonds remerciements à **Madame L.oukid** et **monsieur Y. Djaidja** pour leurs conseils précieux, leurs remarques et leurs disponibilités lors de l'élaboration de cette étude.

Nous souhaitons exprimer nos remerciements à tous les membres de jury pour avoir bien voulu accepter de participer à ce jury, pouvant ainsi l'intérêt qu'ils portent à ce travail. Nous exprimons notre gratitude à l'ensemble du corps enseignant, techniques et administratif du département d'informatique à l'université de Blida, pour leur disponibilité et leur gentillesse. Et finalement, mes sincères remerciements à tous ceux qui ont contribué de près ou de loin à la réalisation de ce modeste travail.

DEDICASE

Je tiens à remercier en premier lieu mes parents

A mon symbole de sacrifice, écoles de mon enfance qui était mon ombre durant toutes mes années d'études, toi mon cher père qui a tant investi pour assurer mon Avenir.

À la source d'amour et tendresse à celle qui a tout donné à toi ma chère maman.

A mes frères, A mes sœurs Sihem, Yasmine

A mon petit frère Mouhamed El Amine et à tous les membres de ma famille

Spécialement A celle avec qui j'ai partagé ce travail au cours de cette année, à toi **Sabrina**

A tous mes collègues de deuxième année master. Et à tous mes collègues de l'université.

Benzineb Asmaa

DEDICASE

En signe de respect et de reconnaissance aux personnes les plus chères

À mon cher papa à qui je dois ma réussite, mon bonheur, et tout le respect.

A ma chère et tendre mère à celle qui a tant souffert, sans me faire souffrir.

À mes sœurs Asma, Amina, Amel à qui je souhaite toutes les réussites et le bienêtre, et à mes nièces serine et Sofia que Dieu les protège.

Spécialement à mon binôme **Asma** qui m'a aidé à réaliser ce projet.

À toute ma famille, mes camarades que j'ai connu durant mon cursus universitaire. Ainsi qu'à tous mes amis. Et à tous ceux qui me sont chers.

Nadour sabrina

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

Introduction générale

Contexte et Problématique:

Cette étude est menée au niveau de l'**Académie algérienne de la langue arabe** dans le cadre du **projet de la Dhakhira al-Arabiyya** visant à constituer une **base de données textuelles automatisée** à fin de vulgariser, de façon étendue et profonde le patrimoine arabe ainsi que la production intellectuelle arabe contemporaine et le rendre accessible à tous.

La mise en œuvre d'un projet d'une telle envergure nécessite non seulement la constitution d'un corpus volumineux mais aussi le développement d'outils permettant d'effectuer les traitements et annotations nécessaires en vue d'exploiter les données que peuvent renfermer ces corpus, tout en prenant en considération les particularités de l'analyse linguistique de la langue arabe.

Objectif:

L'objectif de notre travail est triple. Premièrement construire une plateforme composée d'une boîte à outils (toolbox) offrant les services de base pour produire des applications permettant d'analyser statiquement les corpus arabes tels que l'analyse fréquentielle, concordance, collocation, classification… etc. Il s'agit de mettre en place une plateforme logicielle mutualisée par le biais de développement en mode open-source, en s'appuyant sur l'expérience des concepteurs de logiciels de lexicométrie. Deuxièmement, développer deux applications prototypes: une locale, à déployer sur les postes des utilisateurs de Sciences humaines et Sociales (SHS) ou les linguistes : Windows, Linux et Mac OS X 3, et une basée sur le web en mode client/serveur afin de permettre des analyses statistiques sur les corpus arabes. Et enfin, en troisième lieu, permettre des performances suffisantes pour des ressources textuelles complexes et de grande dimension.

Parmi les objectifs les plus importants du projet de la Dhakhira, est de « **servir de source** pour diverses **études linguistiques, sociales, historiques, scientifiques** et autres ». Le projet vise alors, entre autres, à « aider le chercheur dans la recherche des usages, anciens et modernes,

de la langue, à connaître la fréquence des termes et expressions et leur usage sur le plan géographique, ainsi que les transformations sémantiques dans tous leurs contextes ».

De ce point de vue, notre travail traite d'**une thématique d'une importance cruciale** pour le projet, puisque nous envisageons de **fournir une plateforme qui va aider les linguistes à exploiter les corpus de la Dhakhira pour étudier la langue arabe et cela sur les bases de données qualitatives et quantitatives.**

Nous avons réussi à travers cette plateforme à offrir les fonctionnalités les plus courantes pour un traitement automatique de corpus de données volumineux impraticable à une manipulation manuelle et ouvrant ainsi la voie à l'analyse de millions de textes à la fois, ce qui était impossible auparavant. Parmi les fonctionnalités disponibles actuellement dans notre plateforme nous pouvons citer :

- Les outils qui permettent une contextualisation linguistique rapide et organisée des unités (tel que le concordancier), qui permet d'analyser les contextes d'utilisation des unités linguistiques et de définir, sur la base de ces utilisations leur sens. Ce type d'outils est indispensable pour l'élaboration de dictionnaires et dictionnaires historiques.

- Les outils d'analyse fréquentielle qui permettent de déterminer les mots plus utilisés, les mots rares et les mots qui sont sortie de l'usage. Ces outils sont très utiles pour les lexicologues et les concepteurs de manuels scolaires et méthodes d'enseignement des langues.

- Les outils d'analyse de collocations, n-gram / clusters, etc.

Organisation du mémoire :

Ce présent travail sera structuré en cinq chapitres :

Dans le premier chapitre, nous présenterons comme première partie une définition terminologique du traitement automatique des langues et nous discutons le traitement automatique de la langue arabe et les problèmes liés

Dans le deuxième chapitre, nous présentons les différentes approches utilisées pour l'analyse textuelle et nous détaillerons sur la méthode lexicométrie

Dans le troisième chapitre nous avons fait une étude comparative entre les outils TALA d'analyse et traitement statistique existantes

Le quatrième chapitre concerne la conception et l'architecture de notre plateforme

Le dernier chapitre sera la partie de l'implémentation et la réalisation de notre application, où nous présenterons l'environnement de d´développement, la structure de notre application et quelques interfaces de celle-ci.

Enfin, nous clôturons ce mémoire par une conclusion générale et perspective.

.

I. Traitement automatique de la langue arabe

I.1. Introduction

La langue arabe est parlée par près de 377 millions de personnes dans le monde et elle est la langue religieuse pour plus d'un milliard d'individus. Elle appartient au groupe des langues sémitiques parlées depuis la plus haute antiquité. [1]

La langue arabe est devenue aujourd'hui l'une des langues les plus parlées dans le monde. C'est la langue officielle de plus de vingt pays et de plusieurs organismes internationaux, dans l'une des six langues officielles de l'organisation des Nation unies. On distingue l'arabe standard et l'arabe moderne. L'arabe classique est la forme littéraire utilisée partout pour les besoins de l'écriture et de l'imprimerie. C'est aussi la langue de la religion pour les musulmans, quelle que soit par ailleurs leur langue vernaculaire. L'arabe moderne, dérivé de l'arabe classique, est la langue de la presse, des débats politiques, des textes scientifiques et de plus en plus celle des textes littéraires profanes. L'arabe est une langue très riche et différente des langues occidentales. [2]

Ce chapitre est organisé comme suit : nous commençons par la description des caractéristiques et particularités de la langue arabe afin de dégager les problèmes spécifiques de la RI (recherche d'information) liés à la langue arabe.

I.2. Les niveaux du Tal

- **Analyse morphologique :**

La morphologie est la branche de la linguistique qui étudie les types et la forme des mots.

Il existe deux types de la morphologie :

- La morphologie flexionnelle (en interne) :

Est l'ensemble des modifications subies aux mots d'une langue flexionnelle pour dénoter les traits grammaticaux voulus.

Exemple : بنت => بنات

- La morphologie dérivationnelle (en externe) :

La dérivation lexicale est un des procédés de formation des mots, au même titre que le néologisme ou l'emprunt.

Exemple : درس => مدرس

- **Analyse syntaxique :**

La syntaxe est l'étude des contraintes portant sur les successions licites de formes qui doivent être prises en compte lorsque l'on cherche à décrire les séquences constituant des phrases grammaticalement correctes, Les contraintes envisagées sont de nature variée et correspondent à des propriétés sélectionnelles (telles que les règles d'accord en genre, en nombre, en cas, ...) ou positionnelles (telles que celles qui contrôlent les positions relatives des mots dans la phrase,..) [3]

- **Analyse sémantique :**

L'analyse sémantique est une technique proche de l'analyse lexicale, mais elle concerne essentiellement l'étude des sens des mots en se basant sur le sens pas sur l'écriture de celui-ci [4].En particulier, la sémantique possède plusieurs objets d'étude :

- La signification des mots composés.
- Les rapports de sens entre les mots (relations d'homonymie, de synonymie, d'antonymie, de polysémie, d'hyperonymie, d'hyponymie, etc.).
- La distribution des actants au sein d'un énoncé.
- Les conditions de vérité d'un énoncé.
- L'analyse critique du discours.
- La pragmatique, en tant qu'elle est considérée comme une branche de la sémantique.

- **Analyse pragmatique**

L'étude de la signification d'une phrase appartient à la sémantique, tandis que l'étude de la signification dans le contexte appartient à la pragmatique.

En effet, on peut dire que la pragmatique est une branche de la linguistique, un courant dans l'étude du discours, ou plus largement dans le concept de la langue. Elle étudie la contextualisation des phrases et des énoncés ainsi que la manière dont ils réagissent dans les situations langagières. [5]

I.3. Traitement automatique de la langue arabe

I.3.1. Particularités de la langue arabe

- **L'Alphabet arabe :**

L'alphabet arabe s'écrit et se lit de droite à gauche dont l'alphabet est un abjad c'est-à-dire, un alphabet composé uniquement de consonnes (bien qu'il existe des consonnes faibles qui sont utilisées comme des voyelles longues).

L'alphabet arabe comprend les consonnes, voyelles, et plusieurs signes diacritiques. Les lettres sont monocamérales (il n'existe pas de minuscule et de majuscule).La plupart des lettres s'attachent aux lettres qui leur succèdent et aux lettres qui les précèdent ce qui forme l'agglutination sauf (ا , و, ر, ز, د, ذ) avec la particularité que ces dernières s'attachent uniquement aux lettres qui les précèdent .

- **Les consonnes**

L'alphabet de la langue arabe compte 28 consonnes si on ne compte pas le hamza (ء) et 29 si on la considère comme une consonne.

Il y a deux symboles alif, yah , waw (و, ا ,ي) qui sont des semi-consonnes (ils peuvent être considérés comme des consonnes ou des voyelles longues).

Les lettres	Ecriture selon position			Correspondant français	Prononciation
	Initiale	Médiane	Finale		
ا		ئ ,ى ,ؤ ,إ ,أ,ء		A	Alef
ب	بـ	ـبـ	ب، ـب	B	Ba’
ت	تـ	ـتـ	ت, ـت	T	Ta’
ث	ثـ	ـثـ	ث, ـث	Th	Tha’
ج	جـ	ـجـ	ج, ـج	J	Jim
ح	حـ	ـحـ	ح, ـح	H	Hha’
خ	خـ	ـخـ	خ, ـخ	Kh	Kha’
د	د	ـد	د, ـد	D	Dal
ذ	ذ	ـذ	ذ, ـذ	D	Thal
ر	ر	ـر	ر, ـر	R	Ra
ز	ز	ـز	ز, ـز	Z	Zayn
س	سـ	ـسـ	س, ـس	S	Sin
ش	شـ	ـشـ	ش, ـش	Sh	Shin
ص	صـ	ـصـ	ص, ـص	S	Sad
ض	ضـ	ـضـ	ض, ـض	D	Dad
ط	طـ	ـطـ	ط, ـط	T	Tah
ظ	ظـ	ـظـ	ظ, ـظ	Z	Zah
ع	عـ	ـعـ	ع, ـع	‘‘	Ayn
غ	غـ	ـغـ	غ, ـغ	Gh	Ghayn
ف	فـ	ـفـ	ف, ـف	F	Fa
ق	قـ	ـقـ	ق, ـق	Q	Qaf
ك	كـ	ـكـ	ك, ـك	K	Kaf
ل	لـ	ـلـ	ل, ـل	L	Lam
م	مـ	ـمـ	م, ـم	M	Mim
ن	نـ	ـنـ	ن, ـن	N	Nun
ه	هـ	ـهـ	ه, ـه	H	Ha
و	و	ـو	و, ـو	W	Waw
ي	يـ	ـيـ	ي, ـي	Y	Ya

Tableau 1 : *Les lettres de l’Alphabet arabe dans toutes les positions.*

- **Les voyelles :**

Les voyelles ne sont remarquées que dans les textes qui, par leur importance religieuse ou par les difficultés de leur interprétation (poésie, ouvrages techniques), doivent être l'objet d'une attention particulière. [6]

Les voyelles sont nécessaires à la lecture et à la compréhension correcte d'un texte. Elles sont écrites pour lever des ambiguïtés, et elles donnent aussi la fonction grammaticale d'un mot indépendamment de sa position dans la phrase. Autrement dit, les voyelles ont une double fonction : l'une est morphologique ou sémantique et l'autre est syntaxique.

La langue arabe possède deux types de signes de notation des voyelles : les voyelles brèves, qui sont notées au moyen de signes diacritiques secondaires et les voyelles longues. Elles permettent de différencier des mots ayant les mêmes consonnes.

Exemple : نهر - نَهار - نَهَرَ

- Les voyelles brèves :

Les voyelles brèves sont figurées par des symboles appelés signes diacritiques (les harkāt). Ils sont ajoutées au-dessus ou au-dessous des consonnes. Lorsque la consonne n'a aucune voyelle, on marquera une absence de voyelle représentée en arabe par une voyelle muette Sukun.

L'opération qui insère les voyelles brèves par une machine dans un texte est appelée vocalisation automatique ou voyellation automatique. Ces symboles sont transcrits de la manière suivante :

- **Fatha «** ◌َ **»** : elle surmonte la consonne et se prononce comme un «a» français.
- **Damma «** ◌ُ **»** : elle surmonte la consonne et se prononce comme un «ou» français.
- **Kasra «** ◌ِ **»** : elle se note au-dessous de la consonne et se prononce comme un « i » français).
- **Sukun «** ◌ْ **»** : ce signe indique qu'une consonne n'est pas suivie (ou muet) par une voyelle. Il est porté toujours au-dessus de la consonne.

➢ Les voyelles longues :

Les voyelles longues sont des lettres prolongées, elles sont formées par voyelles brèves et une des lettres suivantes (و, ي, ا) comme le montre le Tableau suivant :

voyelles longues	Noms	Transcription	Symboles	Exemple
ـَا	Alif	A	/a:/	une porte : bâb-unبابٌ
ـَو	Waw	U	/ou:/	un hibou : bûm-unبومٌ
ـِي	Ya	I	/i:/	mon père : ‘ab- î أبي

Tableau 2 : *Les voyelles longues* [6]

- **Autres signes diacritiques :**

➢ La gémination «Šadda » :

La chadda est un signe de la gémination il est semblable a une petite forme de «w», peut être placé au-dessus de toutes les consonnes en position non- initiale. La consonne qui la reçoit est analysée en une séquence de deux consonnes identiques géminées, la première sans voyelle avec sukun (ْ), et la deuxième avec une voyelle brève :Fatha , Damma ou Kasra (ـَ,ـُ,ـِ).

Exemple : (درْرَاجة) درّاجة

➢ La nounation « Tanwîn » :

Les trois signes de tanwin : lorsque, les voyelles courtes (la Fatha, la Kasra et la Damma) sont doublées, elles prennent un son nasal, comme si elles étaient suivies de « n » et on les prononce respectivement :

- **an «** ً **»** pour les Fathatan.
- **in «** ٍ **»** pour les Kasratan.
- **un «** ٌ **»** pour les Dammatan.

I.3.2. La grammaire de la langue arabe

L'étude de la grammaire arabe a commencé très tôt au milieu du 11ème siècle de l'hégire et a donné lieu à d'énormes productions, avant de connaître une période de stagnation qui a duré plusieurs siècles. Ces dernières années, elle connaît un regain d'intérêt, entre autres dans le domaine du traitement automatique. [7]

La grammaire se présente traditionnellement comme la science des règles de la langue .Elle comprend deux branches :

- La morphologie alsarf (الصرف)
- La syntaxe alnahw(النحو)

I.3.3. La morphologie de la langue arabe

I.3.3.a. Définition de la morphologie

La morphologie est un domaine de la langue qui permet la description des règles régissant la structure interne des mots (racine d'un mot, affixes, formation du pluriel, etc.)

La morphologie définit le morphème comme le plus petit élément de signification dans un énonce. C'est le morphème qui confère au mot son aspect grammatical (nom, verbe, etc.).

Pour l'arabe, comme nous le voyons dans la définition de la morphologie, le morphème peut être défini de la manière suivante :

≪ Elément de formation, qu'il soit graphique (graphème = lettre ou voyelle courte) et paradigmatique, qui porte une signification. Cette définition s'applique sur la plus petite unité existante, tel que ≪ي≫ qui, infixe dans un nom, exprime l'aspect diminutif. ≫

Chez les grammairiens, la morphologie se divise en :

- **Morphologie dérivationnelle**

La morphologie dérivée ou lexicale est la branche de la morphologie qui s'intéresse à la formation et la construction des mots.

Plus précisément, elle sert à créer des nouveaux mots ; on peut former de mots nouveaux en ajoutant des morphèmes dérivationnels à partir de mots existants, l'unité lexicale ainsi obtenue est un mot dérive, siyag(الصيغ) par exemple «كاتب» ou « كتابة» dérivés du mot de base «كتب»(comme l'analyse des mots français «écrivain» ou «écriture» dérivés du mot de base «écrire»).Les éléments essentiels de la morphologie de la langue arabe sont :

- La racine :

Une partie importante du lexique arabe est structurée autour d'une racine, les racines ou le Jidhr (الجذر) sont à l'origine de la plupart des mots arabes, Elles sont des verbes formés de trois à cinq lettres consonnes le plus souvent, à partir d'une racine trilitère, on forme des mots selon un schème (moule) précis pour désigner formant la base du mot. La racine est l'unité principale pour appliquer la dérivation.

L'application du mécanisme de dérivation sur une racine peut donner naissance à une famille de mots autour d'un même concept sémantique (l'exemple qu'on a cité au-dessus).

- Le schème :

Représente une forme ou modèle général composé de trois consonnes ف [f],ع [‘] et ل [l]et d'autre lettres (affixes, suffixes...), La notion de schème occupe une place importante dans le mécanisme de dérivation de l'arabe par l'utilisation de processus de génération des formes dérivées à partir d'une racine comme on a dit ci-dessus.

L'arabe comprend environ 150 schèmes ou patrons dont certains plus complexes, tel le redoublement d'une consonne ou l'allongement d'une voyelle de la racine, l'adjonction d'un ou de plusieurs éléments ou la combinaison des deux.

Racine	mot dérive	Schème	Prononciation
دَرَسَ	دَارِسْ	فاعل	Daris
	مَدْرَسَةَ	مفعلة	Madrassatoun
	يَدْرُسُ	يفعل	Yadrouso
	دَرْسْ	فعْلْ	Dars
كَتَبَ	كتبتُ	فعلت	Katabto
	كُتُبْ	فُعُلْ	Koutob

	يكتب	يفعل	Yaktoub
	مكتبة	مفعلة	Maktaba
رَسَمَ	رسّام	فعّال	Rassam
	يرسم	يفعل	Yarssoum
	رِسْمٌ	فعْلٌ	Rassm

Tableau 3: *exemple de mécanisme de dérivation*

- **Morphologie flexionnelle**

La morphologie flexionnelle décrit la manière dont les mots varient ou s'infléchissent afin d'exprimer des contrastes ou des catégories grammaticaux, tels que singulier / pluriel ou passé / présent.

Autrement dit, la morphologie flexionnelle ne sert pas à créer de nouveau mots. On ajoute des morphèmes flexionnels pour former les catégories grammaticales dont nous avons parlé ci-dessus. Par exemple « مجتمع» au pluriel « مجتمعات»(en la langue française « société » à la plurielle «société »).

I.3.4. Composition du lexique arabe

L'analyse lexicale consiste à tester l'appartenance de chaque mot du texte au vocabulaire de la langue. L'arabe considère 3 catégories de mots :

- **Le verbe :**

Entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.

La conjugaison des verbes dépend de plusieurs facteurs :

- Le temps (accompli, inaccompli).
- Le nombre du sujet (singulier, duel, pluriel).
- Le genre du sujet (masculin, féminin).
- La personne (première, deuxième et troisième)

- Le mode (actif, passif).

Pour les pronoms personnels, le sujet est inclus dans le verbe conjugué. Il n'est donc pas nécessaire (comme c'est le cas en français) de précéder le verbe conjugué par son pronom. [8]

- **Le nom :**

L'élément désignant un être ou un objet qui exprime un sens indépendant du temps.

La déclinaison des noms se fait selon les règles suivantes :

- Le féminin singulier.
- Le féminin pluriel.

- **Les particules :**

Entités qui servent pour situer les événements et les objets par rapport au temps et l'espace, et autorisent une séquence cohérente du texte. Ces particules incluent les prépositions, adverbes, conjonctions, interjections et les outils d'exceptions, de négation, etc. [9]

Exemple :

- particules qui désignent un temps منذ بعد،قبل (*pendant, avant, après*).
- particules qui désignent un lieu حيث (où).
- particules qui désignent une référence الذين (ceux).

I.3.5. Le modèle du mot graphique en arabe

La plupart des mots arabes (mots maximales) comporte une structure d'objet complexe, ils sont composés par agglutination des morphèmes lexicaux et grammaticaux de base (racine ou mot minimal), ce dernier est délimite par deux séparateurs successifs, portant toutes les marques formelles (éléments flexionnels) qu'il est capable de porter selon sa catégorie grammaticale (nom, verbe, particule...).

Les marques formelles sont des indices d'aspect, de mode, de voix, de personne, de nature, de genre ou de nombre,…utilisés pour la conjugaison du verbe et pour la déclination du nom.

La représentation suivante est une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche [9]

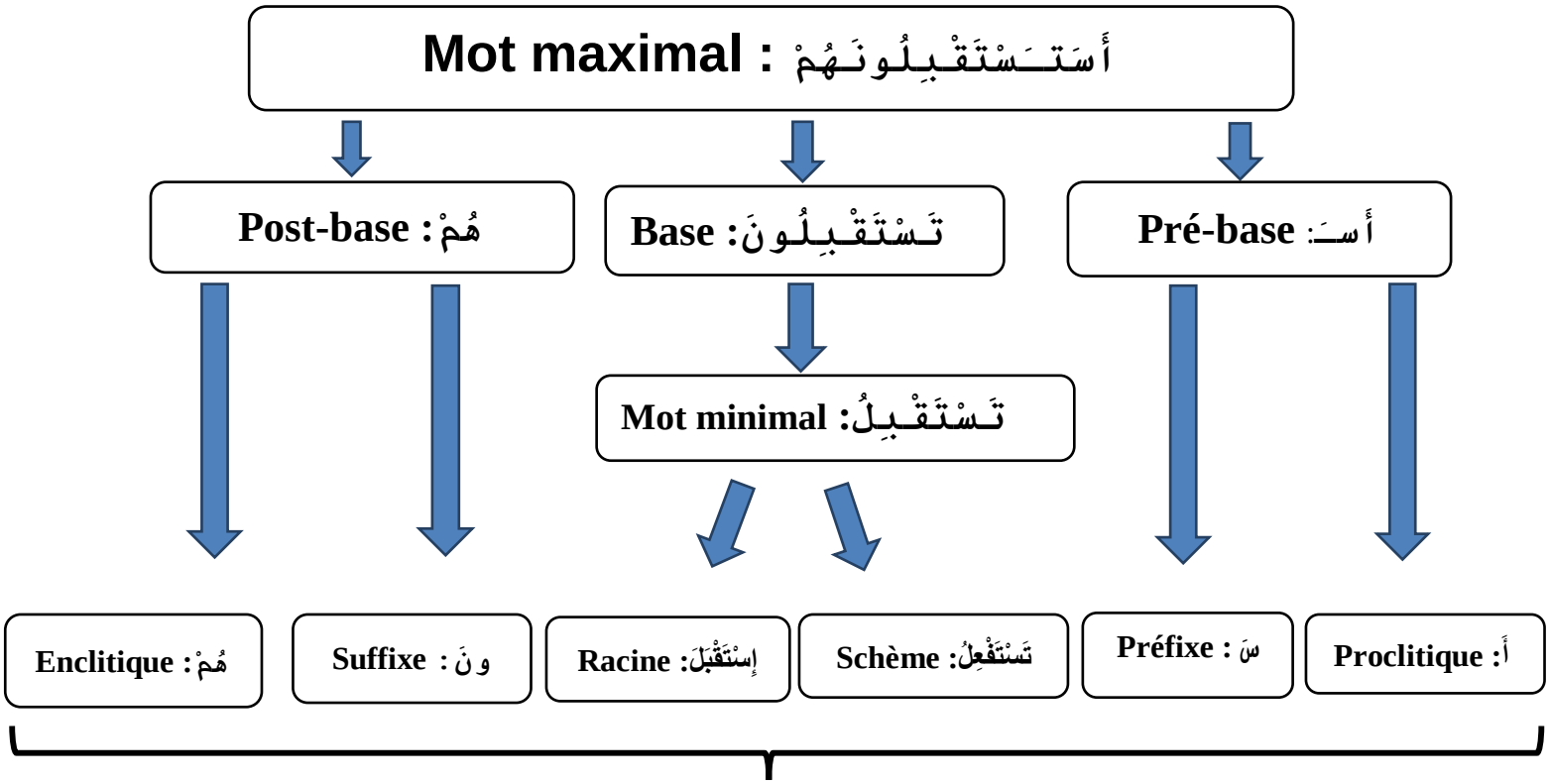


Figure 1 : *Décomposition du mot graphique en arabe أستستقبلونهم asatastakbilūnahu"est ce que vous les accueillerez"*

- **Proclitique :**

Les proclitiques sont des prépositions ou des conjonctions, qu'en combinant, on obtient plus d'informations sur le mot arabe. En effet les proclitiques dépendent exclusivement de l'aspect verbal. Ils prennent donc tous les pronoms et par conséquent ils sont compatibles avec tous les préfixes. [10]

Exemples :

- La coordination par les coordonnants :ف et و
- L'interrogation :أ
- La marque du futur :سـ

- **Enclitiques :**

Les enclitiques sont des pronoms personnels. Ils sont comme les proclitiques, si on combine entre eux, on obtient une post-base composée.

Les enclitiques s'attachent toujours à la fin du mot pour produire des pronoms suffixes qui s'attachent aux verbes, aux noms et aux prépositions.

Mot	Enclitique
أَكَلَهُمْ	هُمُ
كُرَاسِي	ي
لَعِبَّنَ	نَ

Tableau 4 : *exemples d'enclitiques*

- **Préfixes :**

Ils sont liés au mot qui y est attaché, expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,…). En effet, en arabe la majorité des mots commencent par le préfixe (agglutination de l'article) al « الـ », qui est utilisé en tant que terme déclaratif.

Les préfixes sont représentés par un morphème correspondant à une seule lettre en début de mot, qui indique la personne de la conjugaison des verbes au présent.

Préfixe	Signification
أ	Indique la $1^{ère}$ personne du singulier (je).
ن	Indique la $1^{ère}$ personne du pluriel (nous).
ت	Indique la $2^{ème}$ personne au féminin, au masculin et au singulier .
ي	Indique la $3^{ème}$ personne au masculin singulier, pluriel

Tableau 5 : *liste des préfixes arabe*

- **Suffixe :**

Les suffixes en arabe sont utilisés pour des terminaisons des conjugaisons verbales, ainsi que les marques du pluriel et du féminin pour les noms.

Nombre de lemme	L=1	L=2	L=3	L=4	L=5
Les suffixes	ت	تَك	تَها	تَهما	ونهنا
	و	هو	تَهم	وهما	تاهما
	ن	كو	تَهن	نهما	اتهما
	ا	نَه	تَكن	اهما	ينهما
	ي	نَك	تون	يهما	ناهما
	ة	آه	تَنا	ونَهم	تنهما
	ه	آك	وها	ونهن	انهما
	ك	نو	وهَم	ونكن	تموها
		تِي	وهَن	ونكم	تموهم
		تَا	وكن	تاها	تمونا
		آت	وكم	ونَنو	تَماها
		تَه	وون	وننا	تَماهم
					تَماهن

Tableau 6 : *liste des suffixes les plus fréquents en arabe*

- **Les pré-bases :**

Les pré-bases sont le résultat de fusion entre le (s) proclitique (s) et le préfixe. Cette opération se fait d'une façon automatique.

Pré-base	Préfixe	Proclitique
أتَ	تَ	أ
سَتَ	تَ	سَ
أسَتَ	تَ	أفَ
أفَتَ	تَ	أسَ
وَسَتَ	تَ	وَسَ
فسَتَ	تَ	فَسَ
أفسَتَ	تَ	أفَسَ

Tableau 7 : *exemple de groupe de pré-base*

- **Les post-bases :**

Apres la combinaison entre le suffixe et le (s) enclitique (s), on obtient la post-base .le tableau suivant un exemple de groupe de post-base :

Post-base	Suffixe	Enclitique
وك	ك	و
وهُم	هُم	و
ونَنا	نا	ون
ونَني	ني	ون
أنّكُم	كم	أنّ
أنّهُم	هُم	أنّ
تُمُوه	ه	تُمو

Tableau.8 : *exemple de groupe de Post-base*

I.4. Les problèmes de traitement automatique de texte arabe

I.4.1. Le problème de l'agglutination

Dans toute perspective de traitement automatique, le problème est donc de décomposer le mot en ses différentes parties. Cette décomposition nécessite des connaissances de niveau supérieur en cas d'ambiguïtés (si le mot accepte plusieurs segmentations). Une grande partie des mots arabes sont générés en agglutinant des proclitiques et des enclitiques à un radical ce qui cause certains problèmes d'ambigüité spécifiques à la segmentation d'un mot et cela permet d'avoir plusieurs formes (Figure 2).

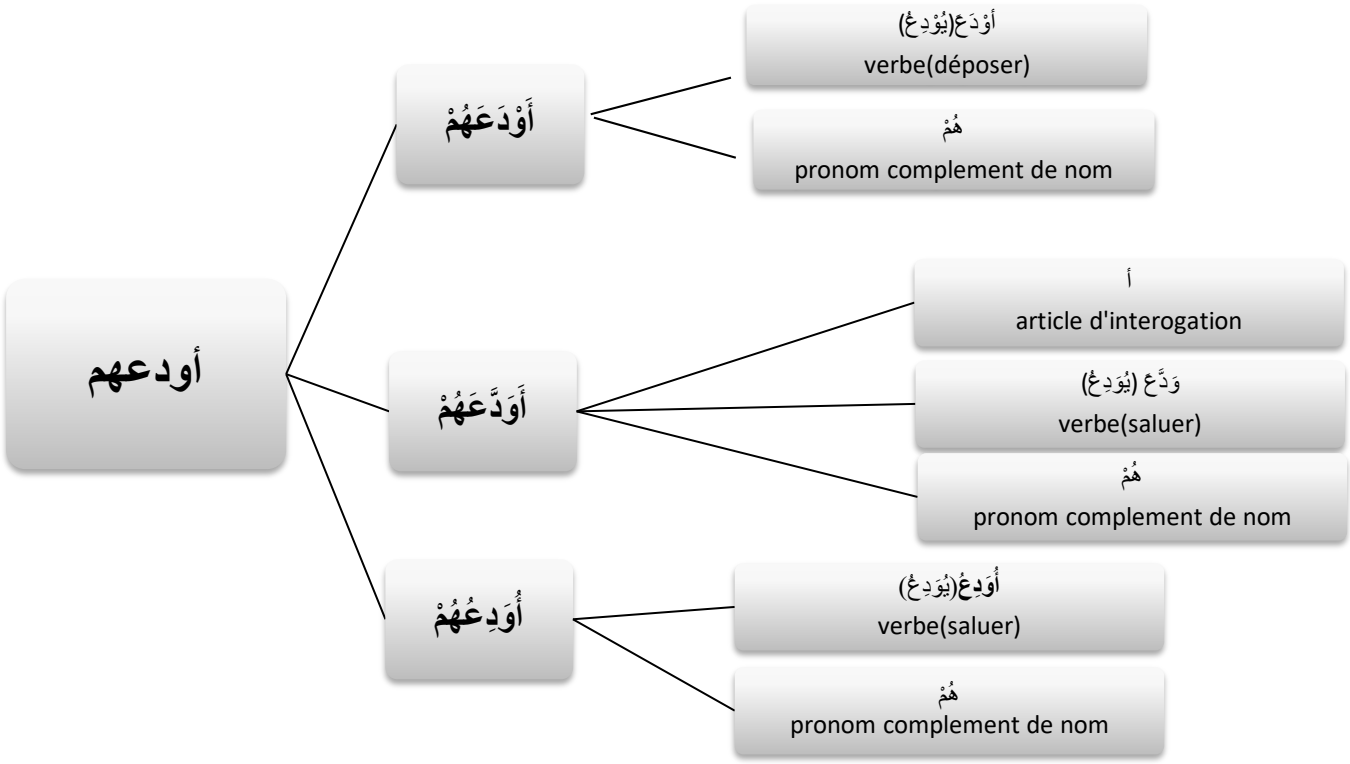


Figure 2 : *Un exemple de plusieurs segmentations d'un mot*

I.4.2. **Le problème de voyellation**

L'écriture arabe courante ne note pas les diacritiques, donc beaucoup de mots en arabe sont homographiques : ils ont la même forme orthographique, bien que la prononciation soit différente et donc certainement la signification soit différente et généralement des catégories grammaticales différentes (Tableau 9). Ce qui cause un taux d'ambiguïté assez élevé donc Pour lire un texte, tout un processus mental est nécessaire il faut identifier le mot comme appartenant au lexique puis lui attribuer ses voyelles dans son contexte, ce qui nécessite la compréhension du texte.

Le mot sans Voyelles	Le mot avec voyelles	Catégorie	Signification
علم	عَلَّمَ	Verbe	a enseigné
	عُلِمَ	Verbe	A été su
	عَلَمٌ	Nom	Drapeau
	عِلْمٌ	Nom	Science
	عُلِّمَ	Verbe	a été enseigné
	عَلِمَ	Verbe	A su

Tableau 9*: exemple de changement de signification selon les voyelles*

I.5. Conclusion

En conclusion, dans ce chapitre, nous avons traité les quatre niveaux de l'analyse Traitement Automatique de Langues Naturelle (morphologique, syntaxique et sémantique) puis nous avons présenté les caractéristiques et les particularités de la langue arabe qui diffère par rapport aux autres langues. Nous avons abordé par la suite, la grammaire de la langue arabe et en final nous avons cité les problèmes du traitement automatique de la langue arabe.

II.Approches et outils statistiques appliqués sur les corpus textuels

II.1. Introduction

Après la phonétique et la phonologie, qui s'occupent des sons, et la morphologie qui s'occupe des unités minimales de forme et de sens, nous arrivons dans la lexicologie, qui s'occupe des masses de mots qui forment le lexique d'une langue, et le stock lexical des individus.

La terminologie de la lexicométrie présente quelques flottements et plusieurs dénominations coexistent : lexicométrie, statistique lexicale, statistique textuelle, approche quantitative des textes, analyse statistique des données textuelles, qui se voient prolongées par la textométrie et la logométrie (prise en compte des textes et des discours et non du seul lexique). Ces méthodes se situent au croisement de plusieurs disciplines : linguistique, statistique et informatique.

II.2. Analyse de données textuelles

L'analyse de données textuelles (ou ADT) est une approche méthodologique des sciences humaines qui envisage les textes comme des données qui peuvent être analysées par un ensemble de manipulations informatiques. Ces analyses, inspirées par la linguistique structurelle et l'analyse de discours, utilisent des approches qualitatives et quantitatives. C'est-à-dire qu'elles cherchent le plus souvent à qualifier les éléments du texte à l'aide de catégories et à les quantifier en analysant la répartition statistique des éléments du texte. [11]

II.3. Définition de corpus

Les corpus sont des collections de données sélectionnées et organisées selon des critères explicites pour servir d'échantillon pour un traitement particulier ou de référence pour fournir une information en profondeur. Généralement, les corpus sont caractérisés par la nature de la langue traitée et par leur contenu. Ainsi, un corpus peut représenter une langue ou plusieurs, comme il peut contenir du texte ou du multimédia [12]

II.4. Méthode d'analyse de données textuelles

- **Méthode quantitatif :**

La méthode quantitative permet de prouver ou démontrer des faits en collectant des données brutes et concrètes. Dans cette forme d'analyse, il est question de calculer la fréquence des unités lexicales (statistiques) qui peuvent être catégorisées. Ces derniers sont utilisés pour créer des graphiques ou des tableaux. [13]

Les études quantitatives permettent à l'utilisateur de tirer des conclusions définitives sur les phénomènes à étudié. Les études quantitatives ont généralement les caractéristiques suivantes :

- Les données sont collectées en utilisant des techniques standardisées.
- On recherche des relations entre les variables en appliquant une analyse statistique sur les données.
- On utilise les données et les analyses pour valider une hypothèse.

- **Méthode qualitative :**

La méthode qualitative est plus descriptive et se concentre sur des interprétations, des expériences et leur signification. Le but n'est pas tant de tester des théories, mais plutôt de mieux comprendre les différentes interprétations de certains événements ou phénomènes. Les résultats d'études qualitatives sont généralement exprimés avec des mots. Les caractéristiques de la recherche qualitative :

- On n'a pas une idée claire des concepts et des résultats qui seront pertinents.
- les résultats des recherches sont plus flexibles avec des études quantitatives.
- La recherche est effectuée dans des environnements «réels».
- La construction de la théorie est plus importante que les tests théoriques.

II.5. L'approche lexicométrique

II.5.1. Définition

On désigne sous le vocable « lexicométrie » la discipline qui prend en charge l'analyse informatisée du discours et du lexique. Cette jeune discipline est appelée également logométrie, statistique lexicale ou encore statistique textuelle.

La lexicométrie est une méthode d'analyse des textes, assistée par ordinateur, qui permet de décrire qualitativement et quantitativement le contenu linguistique d'un corpus. Elle combine entre des outils de recherche ou de compilation documentaire (concordances, recherche de contextes) et des outils statistiques et mathématiques susceptibles de caractériser un texte (dictionnaire de fréquences, classification, calcule de similarité, etc.) [14]

Il s'agit donc d'une technique descriptive. Les différents contextes sémantiques sont des regroupements de fragments de textes nommés Unités de Contexte Elémentaire (dorénavant U.C.E.), correspondent à des phrases dans la mesure où le découpage s'appuie sur la ponctuation forte du corpus en question.

Ils sont nombreux, depuis l'étude des œuvres classiques ou médiévales, l'aide à la constitution de dictionnaires, les études d'occurrences avant ou après lemmatisation, l'analyse des discours politiques ou d'enquêtes contenant des questions ouvertes ou d'enquêtes non directives.

II.5.2. Les étapes d'analyse lexicométrique

L'analyse lexicométrique est de type statistique. Pour pouvoir la mettre en œuvre il est indispensable de postuler, dans un premier temps, l'existence d'unités stables (formes, lemmes ou leurs approximations graphiques) qui garantissent la rapidité et la régularité d'exécution des programmes informatiques d'analyse lexicale

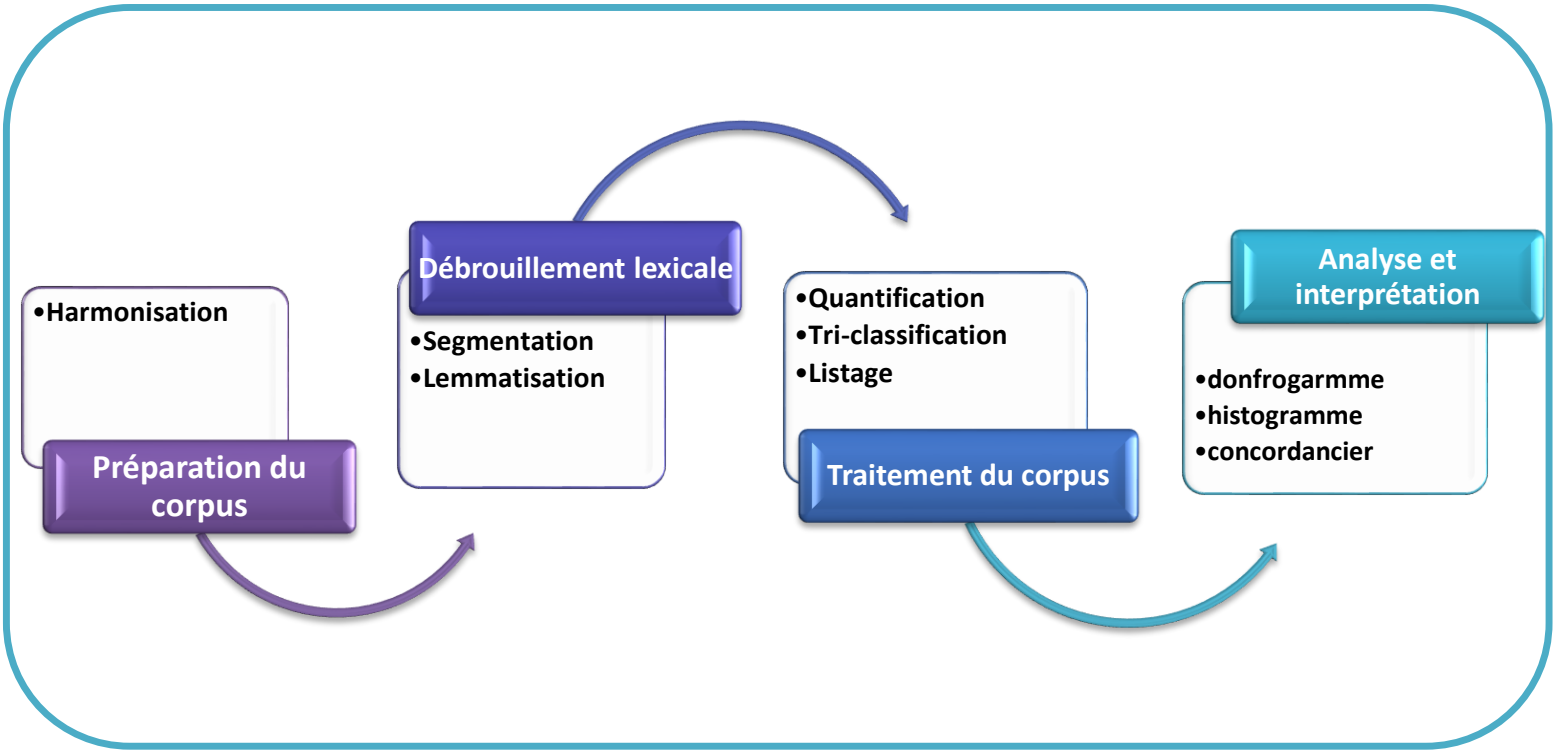


Figure 3 : *Les étapes d'analyse lexicométrie*

Un traitement lexicométrique donnera toujours des résultats, mais pour qu'ils aient un sens il faut passer par les étapes indique dans le cheminement de l'analyste, quatre stations apparaissent incontournables : l'élimination de bruit (harmonisation, suppression tatwil...Etc.) pour la constitution du corpus, la définition des unités linguistiques du texte jugées pertinentes, et le traitement textuel proprement dit : qualitatif ou statistique, enfin l'interprétation de résultats

II.5.3. **Préparation du corpus**

C'est une étape qui consiste à éliminer tout ce qui empêche d'avoir des meilleurs résultats, dans la langue arabe l'étape de l'harmonisation est très importante :

II.5.3.a. **L'Harmonisation :**

L'harmonisation primaire est une opération consistant à homogénéiser toutes les variantes graphiques d'un même mot sous une seule forme graphique d'harmonisation. Et elle est effectuée avant la segmentation, On peut par exemple remplacer systématiquement, lors d'un traitement, toutes les majuscules du texte par un astérisque suivie de la minuscule

correspondante (ex : Analyse devient *analyse).Mais l'harmonisation primaire n'est pas liée aux seules variantes graphiques; il y a un autre phénomène qui est étroitement liées au système d'écriture de l'arabe caractérisé principalement par l'élimination des voyelles.

II.5.4. **Dépouillement lexical**

Après la phase initiale de préparation du corpus (harmonisation) vient le moment critique du dépouillement lexical qui débouchera sur la transformation du texte de départ en une de liste de formes et de lemmes (figure4).

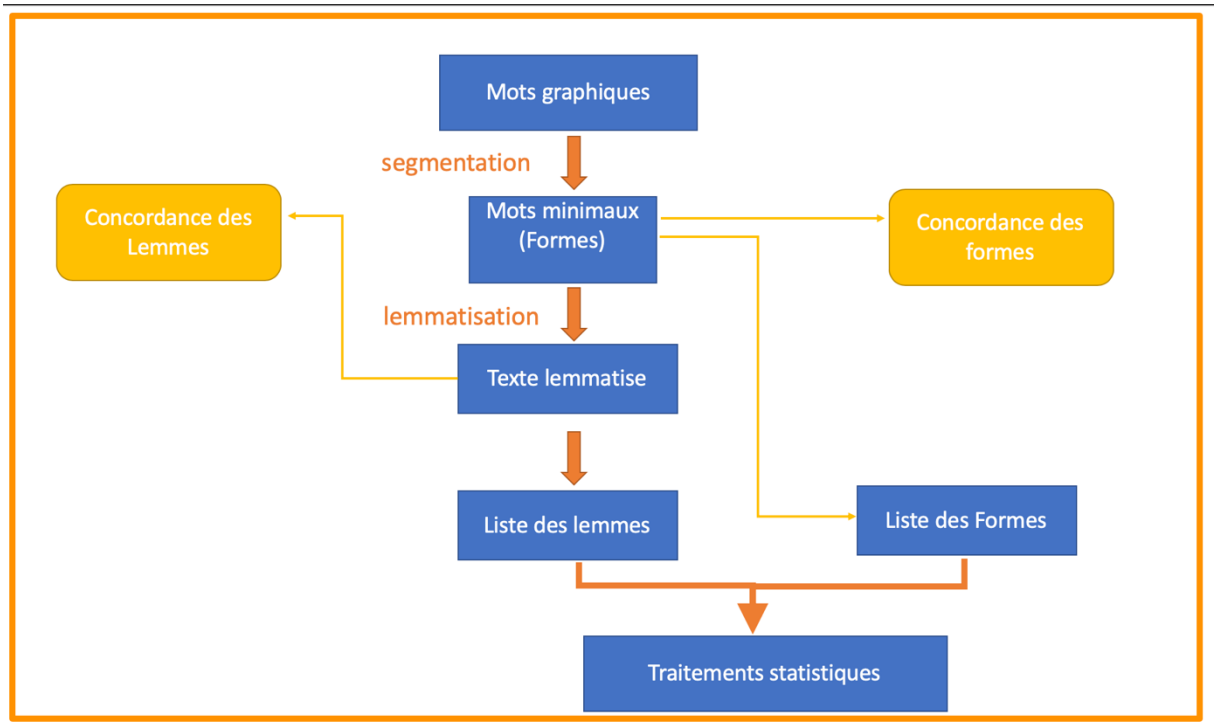


Figure 4 : *Les étapes du dépouillement lexical*

l'analyse lexicale est un processus qui, en analysant le texte, définira les frontières des unités de décompte (segmentation), assigner le Lemme (lemmatisation) à chaque unité, et par la suite faire des concordances des formes ou des symboles des traitements statistiques sur les textes.

II.5.4.a. **La segmentation :**

La particularité de l'approche textométrique est d'aborder le texte au niveau de son support physique en laissant le choix de le traiter mot à mot, phrase par phrase ou encore suivant d'autres niveaux de précision.

Après avoir soigneusement importé et enregistré en ordinateur le corpus global d'analyse, on procède à une segmentation du corpus en formes graphiques ou unités linguistiques .On effectue ensuite une indexation des formes relevées. L'indexation est alphabétique lorsque les occurrences sont classées par ordre alphabétique, ou hiérarchique lorsque nous tenons compte de l'ordre croissant ou décroissant des fréquences des formes répertoriées.

La segmentation est l'opération qui traduit le texte dans un code qu'elle comprend et qu'elle peut ensuite manipuler pour effectuer toutes les sortes de calcul statistique. Cette technique transforme le texte en donnée statistiquement mesurable et pour permettre tout calcul sur les mots, il faut marquer chaque mot d'une étiquette permettant à la fois de l'identifier par rapport aux autres, savoir s'il est unique dans le texte ? Compter combien de fois apparaît-il et de le localiser où le mot apparaît-il dans le texte. [15]

Il faut d'abord que la machine puisse séparer les mots les uns des autres : pour cela, il suffit simplement de ranger les caractères dans deux catégories différentes. Soit un caractère est délimiteur (il sépare les mots les uns des autres), c'est le cas de l'espace et plus généralement des signes de ponctuation, soit il ne l'est pas. En conséquence, une suite de caractères non délimiteurs contenue entre deux caractères délimiteurs est une forme textuelle analysable, mais bien évidemment, il suffit simplement de fournir au logiciel la liste des caractères délimiteurs, beaucoup moins nombreux que les caractères non délimiteurs.

- **Exemple :**

la-langue-arabe-est-la-langue-sémitique-la-plus-parlée-au-monde.

Les caractères délimiteurs (ici en jaune) permettent de découper le texte en " formes graphiques ". Ensuite, la machine va assigner à chaque forme originale un numéro, dans l'ordre d'apparition :

la	-	langue	-	arabe	-	est	-	la	-	langue	-	sémitique	-	la	-	plus	-	parlée	-	au	-	monde.
1		2		3		4		1		2		5		1		6		7		8		9

Les mots “ La ” et “ langue ” figurent deux fois dans le texte, c’est pourquoi le numéro qui leur assigné est répété. C’est ce passage de la forme graphique à la forme numérique qui va permettre d’effectuer des calculs statistiques.

II.5.4.b. **La lemmatisation**

La lemmatisation c’est la fabrication d’une forme réduite du texte, standardisée par des dictionnaires. [16] Elle peut être définie comme étant l’opération qui consiste à rassembler, sous une même forme canonique appelée lemme, toutes les formes fléchies d’un texte ne différant que par des modalités ou flexions grammaticales (conjugaison, déclinaison, genre, nombre, etc.) [17]

II.5.4.c. **Balisage :**

Au cours d’une étude lexicométrique, on cherche à comparer les fréquences des formes dans les différentes parties d'un corpus. Pour rendre possible ces comparaisons, le texte doit comporter des balises indiquant les délimitations logiques du corpus.

L’insertion de clés constitue une phase importante dans la préparation du texte. Les clés introduites permettront ensuite à l’utilisateur d'effectuer des comparaisons à partir des parties du corpus qu'elles découpent.

II.5.5. **Traitement du corpus**

II.5.5.a. **Analyse fréquentielle**

Une analyse fréquentielle montre que certains mots sont très fréquents, tandis que la plupart sont rares, donc elle permet de comparer la fréquence des thèmes. L’hypothèse de cette méthode est : plus haute fréquence de l’idée est élevée, plus cette idée est importante pour le locuteur.

De manière générale, la liste de fréquences de mots (figure 5) est la fonction la plus importante de tout système de traitement de corpus, car elle révèle à la fois les propriétés de la langue et les thèmes principaux du corpus.

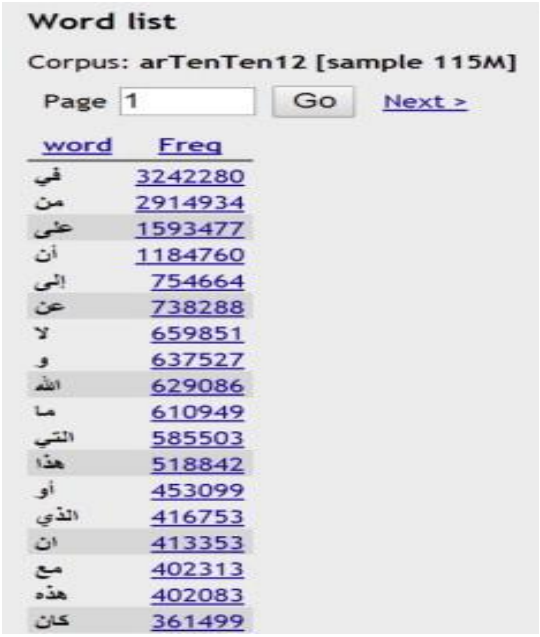


Figure 5: *Liste de fréquence de l'ensemble du corpus arTenTen chargée dans Sketch Engine*

II.5.5.b. **Segments répétés**

Les segments répètes (séquences de mots répétés à l'identique) permet de repérer toutes les occurrences de suite de formes graphiques qui apparaissent plusieurs fois dans un corpus. Autrement ils sont des chaines ou des séries de formes contigües non interrompues par des signes de ponctuation. Ces séries se répètent, semblables, plusieurs fois dans le même texte. [18]

II.5.5.c. **N-Grammes**

Un n-gramme est une séquence de n caractères consécutifs. Pour un document quelconque, l'ensemble des n-grammes (en général n prend les valeurs 2 ou 3) qu'on peut générer est le résultat qu'on obtient en déplaçant une fenêtre de n cases sur le corps de texte. [19]

II.5.5.d. **Concordance**

Plusieurs outils documentaires permettent une contextualisation linguistique rapide et organisée des unités, le plus connu est le concordancier [20]. La concordance est l'une des fonctions les plus importantes pour les linguistes en ce qui concerne les corpus car elle permet d'extraire et de convoquer toutes les phrases du corpus contenant les occurrences d'un mot ou phrase pour en vérifier l'emploi (c'est-à-dire afficher le contexte dans lequel Ces chaînes sont

utilisées dans un ou plusieurs textes électroniques.).Il existe plusieurs types de concordancier comme KWOC (keyword out of context), KWAC (key-word and context), KWUT (Key-word up to text) …etc.

- KWAC (key-word and context) **:**

KWAC affiche les contextes du motif recherché et les aligne de manière à mettre tous les motifs au centre de la ligne. Le contexte à droite et à gauche est paramétrable par l'utilisateur, c'est-à-dire que l'on peut choisir combien de mots afficher avant et après l'occurrence recherchée et trier les résultats par ordre alphabétique par rapport au mot qui suit ou qui précède le motif recherché comme l'illustration ci déçu qui est un extrait de la concordance du mot « طفل » dans un article.

Figure 6 : *image de concordance KAWC de mot « طفل» [Concordance tirée de sketch Engine]*

- KWIC (keyword in contexte) :

KWUT reprend tout le texte d'origine dans lequel se trouve le motif recherché en mettant en évidence la séquence recherchée. Ceci permet d'avoir un contexte plus ample, mais rend parfois la recherche moins immédiate. Voilà pourquoi certains concordanciers combinent les deux techniques en offrant un affichage KWIC qui permet de voir l'occurrence en plein texte si nécessaire.

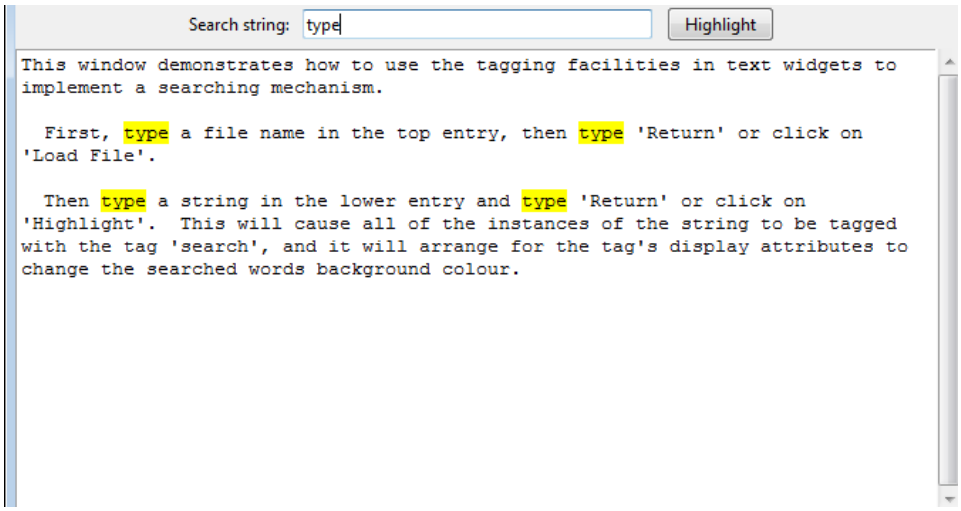


Figure 7*: Image de concordance KWUT de mot « type »*

II.5.5.e. Cooccurrence

La cooccurrence est la coprésence ou présence simultanée de deux unités linguistiques (deux mots par exemple ou deux codes grammaticaux) au sein d'un même contexte linguistique (le paragraphe ou la phrase par exemple). Cette cooccurrence peut être grossièrement constatée, puis vainement exprimée, en fréquence absolue. Elle s'appelle aussi par collocation, corrélation [21]

II.5.5.f. La classification ascendante hiérarchique *(CAH)*

Il existe de nombreuses techniques statistiques visant à partitionner une population en différentes classes ou sous-groupes. La classification ascendante hiérarchique *(*CAH*)* est l'une d'entre elles. On cherche à ce que les individus regroupés au sein d'une même classe soient le plus semblables possibles tandis que les classes soient le plus dissemblables.

- **Principe :**

Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existant entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendrogramme ou arbre de classification (figure 8). La classification est ascendante car elle part des observations individuelles ; elle est hiérarchique car elle produit des classes ou groupes de plus en plus vastes, incluant des sous-groupes en leur sein. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée.

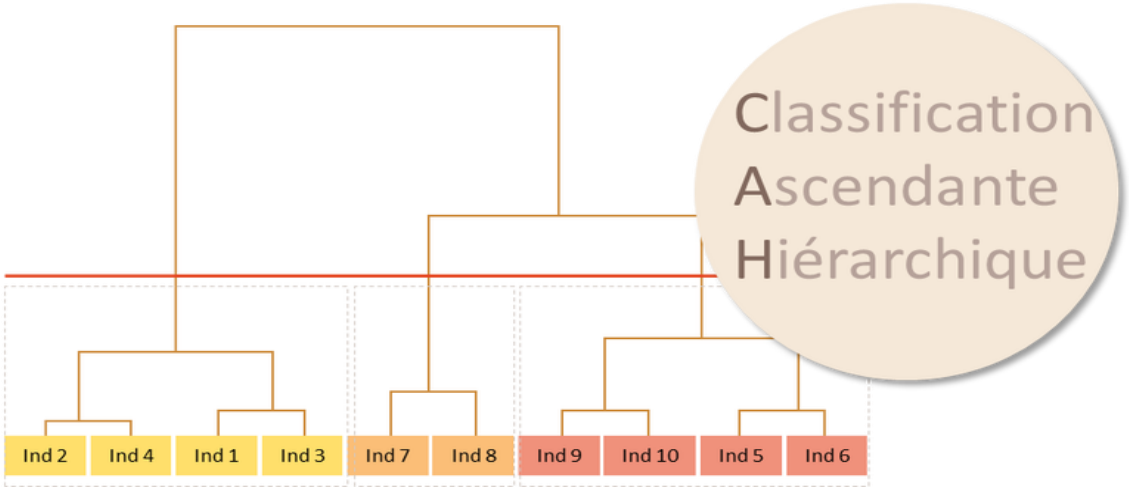


Figure 8: *Regroupement et classification selon la méthode hiérarchique (dendrogramme)*

II.5.5.g. Analyse des spécificités lexicales:

A partir des informations issues des dépouillements quantitatifs en formes graphiques ou sur des formes lemmatisées, ces formes sont souvent plus représentatives d'un corpus que les formes simples .l'analyse de spécificités permet de caractériser les occurrences présentes dans un corpus et de calculer un indice de spécificité pour chaque forme dans chaque segment du texte en distinguant les formes qui sont plus employées [22]. Une fois ces spécificités mises en évidence par la statistique on peut rechercher leurs contextes de réalisation. Ainsi, une forme à spécificité positive est une forme « sur-employée », une forme à spécificité négative est une forme « sous-employée ».

II.5.5.h. Analyse factorielle de correspondance :

AFC est une méthode descriptive d'analyse proposée par Benzecri permettant d'étudier un tableau de contingence conduisant à une représentation graphique où des termes placés en ligne et des documents placées en colonnes, elle considère le tableau de données comme un nuage de points (figure9) [23].Elle est un outil permettant de réduire la dimension des données en conservant le plus d'information possible, autrement la méthode d'AFC s'utilise pour décrire et hiérarchiser les relations statistiques qui peuvent exister entre des individus.

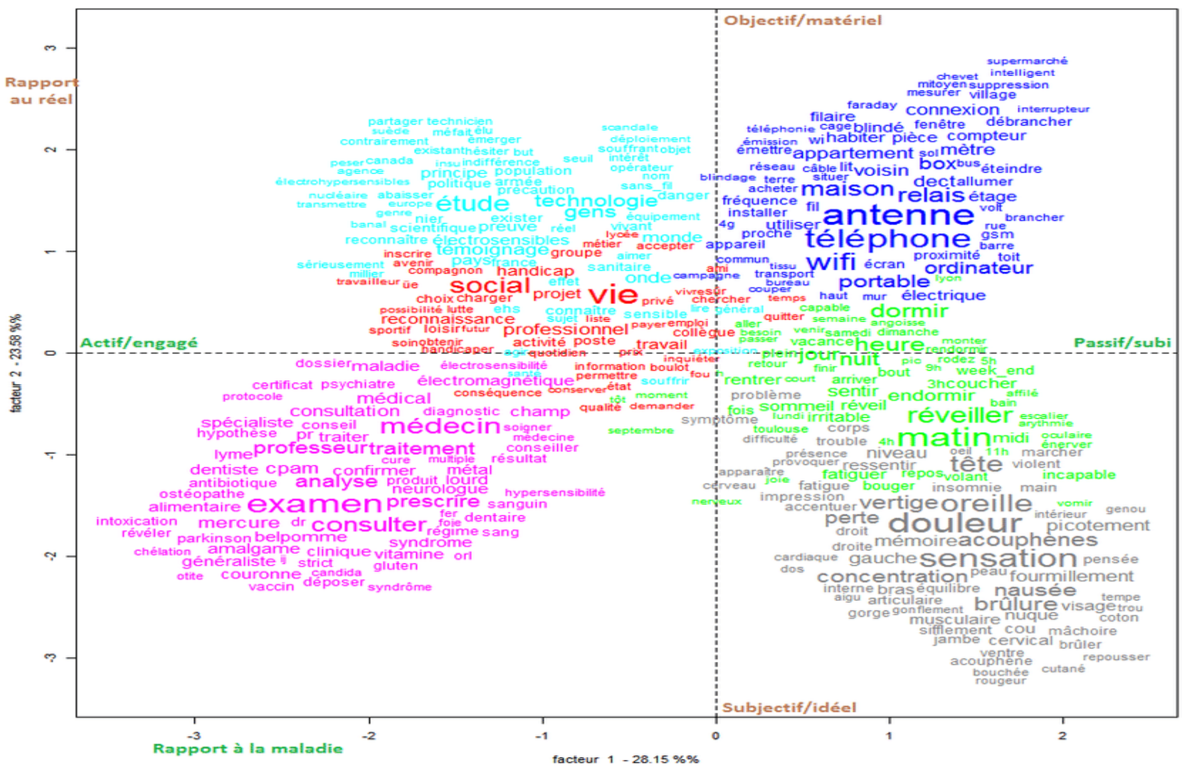


Figure 9 : *Une analyse factorielle des correspondances faite par Iramuteq*

Cette figure présente un nuage de points dans un espace mathématique ayant autant de dimensions qu'il y a de colonnes dans le tableau de données , AFC cherchent à le projeter sur des axes ou des plans (appelés factoriels) de façon que l'on puisse en visualiser et étudier au mieux la forme et donc rechercher globalement des corrélations. La spécificité de l'AFC est qu'elle considère en même temps un nuage de point représentant les lignes (termes) et un autre représentant les colonnes (documents).

II.6. Le rôle et limites de la lexicometrie

Comme nous venons de le voir, les méthodes de statistique multidimensionnelle permettent de traiter et de représenter de grandes masses de données, particulièrement adaptées à l'explosion de données qui caractérise notre époque.

La logométrie a rendu l'analyse de données textuelles accessible via de différents outils et logiciels, cette dernière se démocratise de plus en plus, et un grand nombre d'études s'y intéressent, et cherchent à fournir d'autres applications qui répondent mieux aux exigences multiples, qui se varient en fonction des problématiques à répondre, et la diversité des langues.

Parmi les intérêts les plus évidents de l'utilisation de logiciels d'analyse de données textuelles est précisément de pouvoir traiter des corpus de données volumineux, impraticable à une lecture / annotation manuelle. Le temps de traitement de tels corpus, largement inférieur à celui que réclame une manipulation manuelle, elle permet ainsi d'accéder à des corpus auparavant inaccessibles. Par ailleurs, de nouvelles sources (web) ou un accès diversifié à des sources plus anciennes (numérisation d'archives, politique de libre accès aux données publiques) s'offrent à la recherche, notamment en sciences sociales. Les logiciels d'analyse textuelle constituent dans ce cadre une ressource pour l'exploitation de ces ensembles volumineux de matériaux. [24]

Cette logométrie est également mis une génération de chercheurs intéressés seulement à l'aspect extérieur de la langue, en ignorant que cette technique est un outil descriptif statistique plutôt qu'un outil d'analyse, les résultats obtenus sont parfois classés de manière mixte. Elle s'est restreinte à une prestation statistique d'un contenu, et n'intervient pas pour expliciter ce que qui fait le choix de lexies par le locuteur.

II.7. Conclusion

L'analyse statistique des donnes textuelle a un rôle crucial dans plusieurs recherche et analyse des textes, tel que la détection de la langue par l'analyse fréquentielle, l'extraction des sens des mots à partir de leurs contexte par la concordance et la détection des idées générales des textes… etc. L'amélioration d'un tel système des statistiques influe positivement les résultats rendus par les autres traitements automatiques des textes écrits.

En effet, nous pensons que la combinaison de plusieurs niveaux (par exemple, niveau TAL et niveau statistique) pour les analyses automatique des textes améliorera majoritairement les résultats rendus par les outils statistiques des donnes textuelles. Dans ce qui suit, nous détaillerons l'architecture de notre plateforme pour l'analyse statistique textuelle

III. Étude comparative entre les plateformes TALA d'analyse et traitements statistiques

III.1. Introduction

Pour implémenter un outil approprié pour soutenir l'arabe semble être difficile et nécessite une comparaison entre de multiples outils, leurs potentiels et les fonctions diffèrent en termes de manipulation de l'arabe.

Ce chapitre tente de présenter une évaluation comparative fondamentale de six plateformes de recherche et d'analyse de corpus existants décrits comme supportant plusieurs langues dont l'arabe. Elle est basée sur plusieurs critères qui semblent être les plus essentiels pour la recherche et analyser des corpus arabes.

Le but de cette évaluation est concentré sur les aspects les plus importants être pris en compte dans la réalisation de notre plateforme afin de mieux soutenir le texte arabe.

III.2. Définition d'une plateforme

Une plate-forme est l'ensemble des composants utilisés en commun dans une famille de produits dont les fonctionnalités peuvent être étendues par des tiers [25]

Exemple :

- OS de bureau : Unix, Mac, Windows…
- Appareils mobiles: iPhone, Android…
- Réseaux sociaux: Facebook, MySpace, LinkedIn, Monster, Twitter…
- Voice over Internet Protocol (VOIP) : Skype, Nextiva, Yahoo!...
- Recherche Web: Google, Bing + Yahoo!, Baidu..

III.3. Contexte

De nombreux outils sont utilisés pour rechercher et analyser des corpus. Ils ont généralement pour rôle de fournir quelques fonctions de base (par exemple mots fréquents et concordances), alors que certains de ces outils ont plus de fonctions et statistiques telles que collocations,

n-gram / clusters, mots-clés, etc. Un certain nombre de ces fonctions de recherche et les outils d'analyse sont basés sur le Web, par exemple The Sketch Engine, IntelliText Corpus Queries … etc, afin de les utiliser, les chercheurs doivent rester en ligne de manière permanente. D'autres outils sont basés sur des solutions centraliser, afin qu'ils puissent être téléchargés sur des ordinateurs et utilisés hors ligne, tels que l'outil de traitement des corpus arabes "Khawas" de KACST, aConCorde, Lex&Co, Iramuteq…

En ce qui concerne les corpus arabes, leur nombre augmente constamment, certains de ces corpus arabes sont consultables en ligne et ont leur propres outils d'analyse et autres sont open source et peuvent être téléchargé sur les ordinateurs des utilisateurs.

III.4. Les plateformes étudiées

Nous avons appliqué notre cadre d'évaluation proposé sur six systèmes de traitement de corpus disponibles gratuitement. Les outils sélectionnés ont été conçus pour soutenir L'arabe avec d'autres langues :

III.4.1. AConCorde

AConcorde [26] est un outil de concordance multilingue et gratuit qui a été créé par Andrew Roberts, lorsqu'il était doctorant à l'Université de Leeds, écrit en Java, il est donc fonctionnel sur Java-Runtime-Environnent [27].

Figure 10 : *interface générale de* ***aConCorde***

III.4.2. **LEXICO 5**

LEXICO5 [28] est développé depuis le milieu des années 2010, il représente le premier logiciel abouti de lexicométrie, par André Salem et son équipe parisienne Syled-Cla2t. Il fait suite à Lexico 3 développé à partir de 2003. Lexico 3 était lui-même précédé de Lexico 1 puis Lexico 2 [27] .Il est un excellent logiciel, un pionnier, qui ne tourne que sur les systèmes d'exploitation Windows. On doit à son concepteur dans la diffusion des méthodes de recherche scientifique et de la recherche des segments de marché.

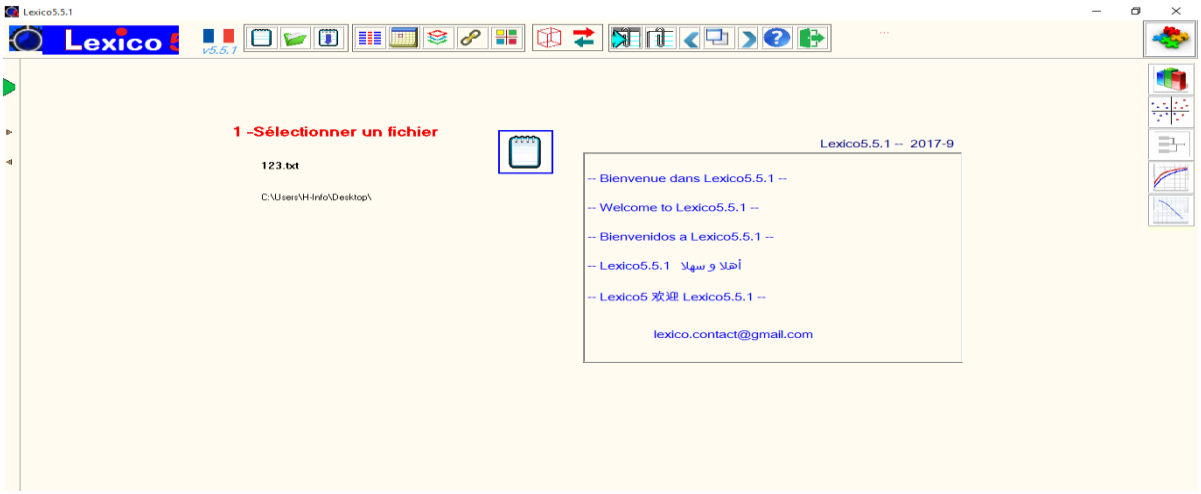


Figure 11: *interface générale de **LEXICO 5***

III.4.3. **Iramuteq**

Iramuteq est un logiciel encore en construction, il a été créé en 2010 en open-source. Il a l'avantage d'être utilisables sur plusieurs systèmes d'exploitation comme Windows, MacOs, Linux, Écrit en python et il utilise la plateforme R. [29]

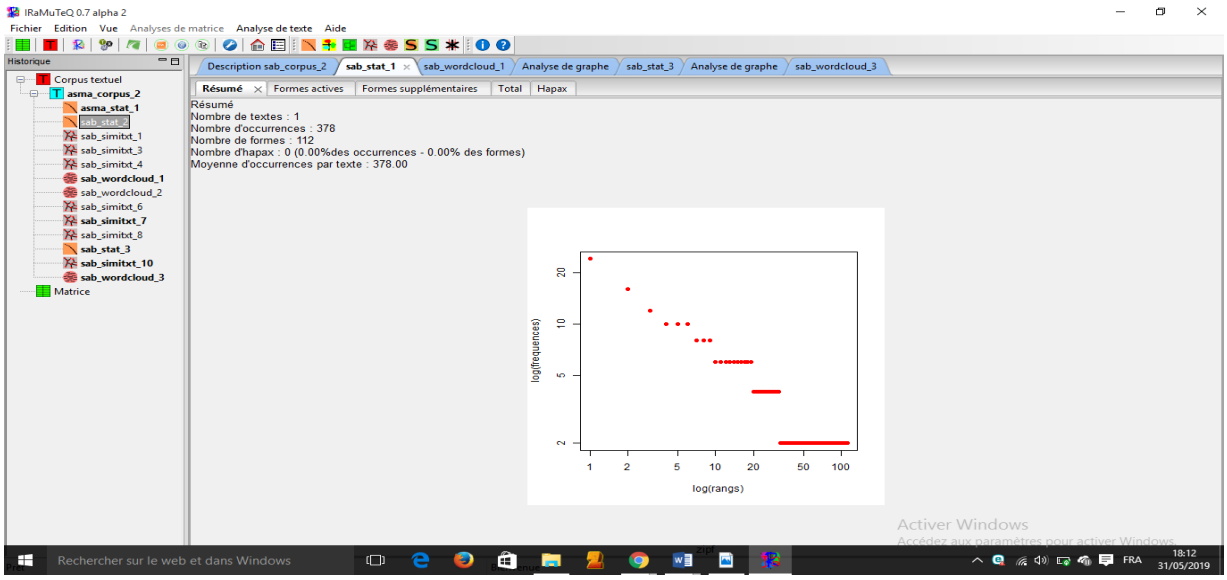


Figure 12: *interface générale de* ***Iramuteq***

III.4.4. **Requêtes de corpus IntelliText**

IntelliText [30] est une application Web basé sur un système développé par le Centre for Translation Studies (CTS) au Université de Leeds dans le but de faciliter et d'améliorer l'enseignement et la recherche dans divers domaines des sciences humaines.

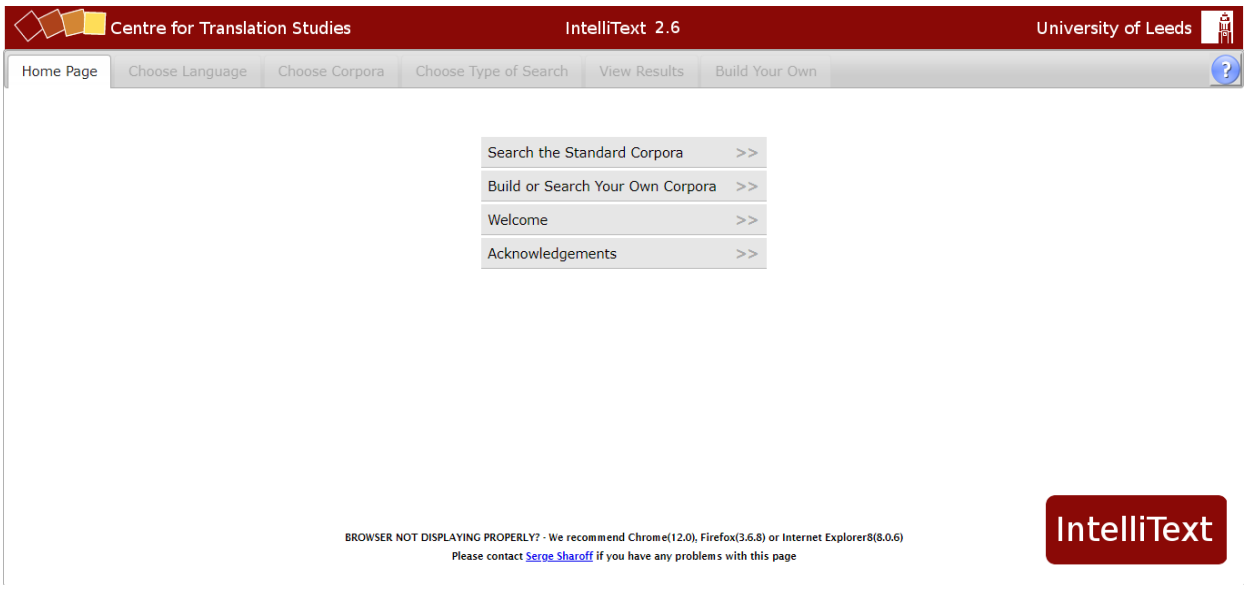


Figure 13 : *interface générale de* ***IIntellitext***

III.4.5. **Le Sketch Engine**

Sketch Engine [31] est un outil de corpus de premier plan. Il a été largement utilisé dans lexicographie. Dix ans se sont écoulés depuis son lancement [32]. Sketch Engine est un logiciel basé sur le Web qui permet d'analyser le corpus.

Figure 14 : *interface générale de* **Sketch Engine**

III.4.1. **ACPTs (version 3.0)**

ACPTs (Arabic Corpus Processing Tools) est un outil autonome de traitement des corpus Al-thubaity, il a été conçu par une équipe de recherche chez KACST (Cité du roi Abdulaziz pour la science et la technologie) en février 2013 dirigée par AlThubaity et al. ACPTs a été développé en Java, ce qui signifie qu'il peut être exécuté sur de nombreux systèmes d'exploitation. [33]

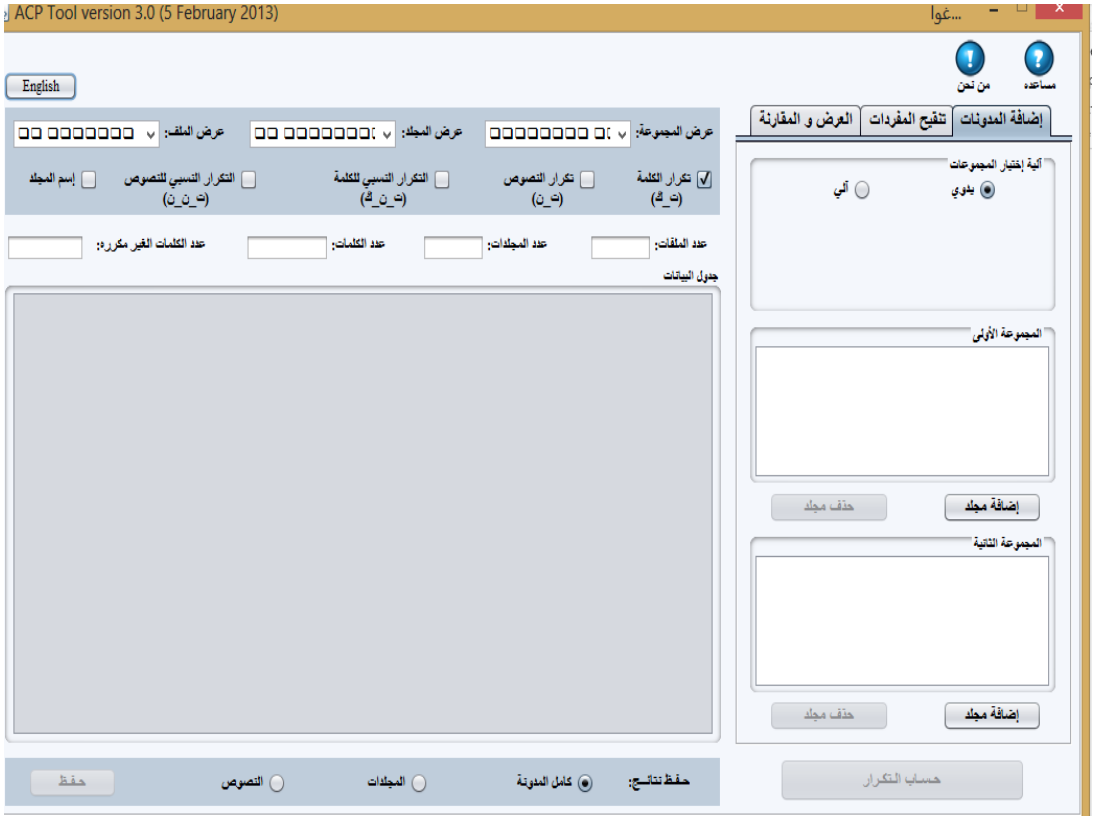


Figure 15: *interface générale de* **KACST**

III.5. Critères d'évaluation

Étant donné que les fonctions des outils examinés ici diffèrent d'un outil à l'autre, de manière générale, l'utilisabilité et les fonctionnalités sont les facteurs les plus importants pour la crédibilité des logiciels et basés sur des caractéristiques linguistiques de la langue arabe. Plusieurs points ont été sélectionnés semblant être le critère le plus essentiel pour la recherche et l'analyse de l'arabe corpus.

Les sous-sections suivantes décrivent les critères d'évaluation proposés pour les systèmes de traitement de corpus arabe.

III.5.1. **Utilisabilité**

Pour vérifier l'utilisabilité, on va tester la facilité avec laquelle un utilisateur peut apprendre à utiliser, préparer des entrées et interpréter les sorties d'un système ou d'un composant. On a fait une liste de contrôle comprenant quatre dimensions comme suit :

- **Ecran principale :**

Cette dimension concerne l'écran principal de l'interface du système. Par exemple, s'il est convivial et facile à comprendre et rend toutes les options disponibles facilement accessibles et il consiste à savoir s'il a été conçu de manière professionnelle.

- **Orientation des tâches :**

Cette dimension couvre la manière dont les tâches sont effectuées à l'aide de l'outil. Par exemple, il examine à quel point les tâches sont simples, si les résultats sont clairs et si un débutant peut les utiliser sans assistance.

- **Formulaires et saisie de données :**

Cette dimension concerne la conception de formulaire. Les éléments permettent notamment de déterminer si les champs du formulaire sont clairement expliqués, si les étiquettes et les en-têtes sont clairs et si les formulaires sont validés avant d'être soumis.

- **Recherche :**

Cette dimension concerne la fonction de recherche et les résultats de la recherche.

III.5.2. **Les fonctionnalités**

Les utilisateurs recherchent des systèmes fournissant des profils de fréquence ou la concordance de mots spécifiques, d'autres utilisateurs recherchent la distribution des mots ou des *N*-grams dans les textes de corpus, ou ils peuvent rechercher les mots les plus Co-localisés. En utilisant la même technique d'évaluation qu'on a faite avec le critère précédent (liste de contrôle) comprend six dimensions comme suit :

- **Fournir une interface arabe :**

Cette dimension a permis de déterminer si l'interface utilisateur peut être utilisée dans une seule langue ou si elle peut être commutée dans d'autres langues (anglais, arabe ou autres langues).Les systèmes comportant des interfaces multilingues attireront évidemment d'avantage d'utilisateurs qui parlent et travaillent dans ces langues, ce qui facilitera leur utilisation.

Cela permet de déterminer si ces outils fournissent une interface utilisateur arabe. Pour les utilisateurs arabes, étant donné que certains chercheurs ne peuvent pas utiliser un outil si l'interface est dans une langue différente de leur langue maternelle, donc ils ne peuvent pas bénéficier de ses fonctionnalités.

- **Lecture de fichiers de texte arabe au format UTF-8 :**

Des textes de corpus peuvent être créés en utilisant diverses méthodes d'encodage. Ce point examine si les outils testés sont capables de lire l'arabe fichier texte au format UTF-8 et affichent les caractères correctement, la norme Unicode comporte trois UTF : UTF-16, UTF-8 et UTF-32 (dans l'ordre chronologique), l'UTF-16 est appelé "Unicode", et UTF-8 est supérieur aux deux autres.

- **Lecture de fichiers de texte arabe au format Unicode :**

Il s'agit d'examiner si les outils sont capables de lire des fichiers de texte arabe format Unicode et afficher les caractères correctement. Malgré le fait qu'UTF-8 est recommandé pour la construction de corpus, Les applications Microsoft conseillent à l'utilisateur d'utiliser UTF-16. Notepad est une application en particulier sur laquelle beaucoup de gens comptent pour créer et enregistrer leurs fichiers de corpus. Notepad conseillé à l'utilisateur d'utiliser "Unicode" (qui fait référence à UTF-16), en ignorant UTF-8, qui est également disponible parmi les autres formats de codage. Ainsi, les corpus les outils peuvent ou non être capables de gérer le format de codage Unicode outre le format UTF-8 qui est le plus largement utilisé dans la construction de corpus. Pour cette raison, la capacité de lire des caractères arabes en Unicode était incluse dans cette évaluation.

- **Format de texte de saisie :**

Les fichiers de corpus sont généralement au format texte brut (extension .txt) en tant que format standard pouvant être traités facilement. Cependant, les matériaux de corpus originaux peuvent exister dans une variété d'autres formats, tels que des pages Web ou des documents Microsoft Word.

La possibilité d'inclure différents autres formats de texte que le format .txt standard dans un corpus permettra aux utilisateurs de gagner beaucoup de temps et d'effort.

- **Permettre aux utilisateurs de télécharger ou d'ouvrir leurs corpus personnels en arabe :**

Les chercheurs peuvent souhaiter utiliser certains corpus arabes, voire même construire leur propres corpus à partir de zéro et utiliser des outils pour rechercher et analyser ces Ressources. Par conséquent, les outils ici sont examinés pour voir s'ils acceptent des fichiers de données externes.

- **Prétraitement du texte**

Cette dimension représente une fonctionnalité facultative, mais elle est également importante car elle permet aux utilisateurs d'omettre certains mots ou caractères des résultats du traitement du corpus. L'accessibilité de telles fonctionnalités peut aider les utilisateurs à interpréter facilement les données et à se concentrer d'avantage sur la cible de leur enquête. Notre évaluation inclut la possibilité de supprimer certaines listes de mots (mots vides) et de supprimer les chiffres, la ponctuation et d'autres symboles spéciaux. De plus, nous considérons les caractéristiques de la langue arabe pouvant nécessiter une prise en compte lors de la phase de prétraitement, telles que les signes diacritiques et la normalisation de Hamza

- **Normaliser les signes de diacritiques** :

C'est pour vérifier si l'outil est capable de normaliser les signes diacritiques, de sorte que l'utilisateur a la possibilité de rechercher des textes arabes comprenant des signes diacritiques à l'aide d'une forme de mot unique dans la requête. Par exemple, si un texte contient le mot"كَتَّبَ" (avec diacritiques) et le mot "" (sans diacritiques), l'utilisateur est-t-il capable de rechercher à la fois en utilisant le formulaire unique « كتب » ? Ceci est important dans

recherche dans les corpus arabes, une forme pouvant avoir plusieurs sous-formes avec les signes diacritiques.

- **Normaliser Hamza "ء" :**

Ceci est similaire à la référence précédente. Ici, nous vérifions si l'outil a la capacité de normaliser les mots qui ont Hamza, de sorte que l'utilisateur ait une option pour rechercher des textes en arabe, qui incluent Hamza en utilisant un seul mot forme dans la requête. Par exemple, si un texte contient le mot "إستمرار" (avec Hamza) et le mot "استمرار" (sans Hamza), est-ce que l'utilisateur est capable de rechercher les deux en utilisant le formulaire unique "استمرار" ?

- **Listes de fréquence de mots simples :**

Dans notre évaluation, nous avons considéré l'importante fonction de la fréquence relative ; les informations de fréquence relative sont réduites pour les comparaisons de profils de fréquence de mots entre différents corpus.

- **Listes de fréquence de *N*-grams :**

N-grams désigne tout N mots consécutifs. La fréquence des *N*-grams montre comment ces mots sont liés en général et comment ils s'unissent pour donner un sens. La fréquence des graphes N est également utile dans la modélisation du langage dans les applications statistiques de traitement automatique du langage naturel (TALN).

- **Afficher le texte arabe dans le bon sens (de droite à gauche) :**

Comme l'arabe est écrit de droite à gauche, les outils ont été examinés pour déterminer s'ils peuvent afficher le texte arabe dans la bonne direction, en particulier dans la concordance, où les contextes doivent également être ordonnés correctement.

- **Affichage correct des signes diacritiques :**

La possibilité de montrer des diacritiques en arabe s'il y'en a, par exemple "هِمَةٍ". Afficher des signes diacritiques peut être essentiel dans certains cas, en particulier avec des formes similaires qui ne peuvent pas être distingués s'ils n'ont pas les signes diacritiques.

Exemples :

- ذَهَبَ(le passé du verbe «Aller») et ذَهَبْ (Nom: Or)

- **Concordance :**

Nous avons examiné la disponibilité des problèmes de fonctionnalité liés à la concordance de mots simples et de *N*- grammes, y compris la possibilité de trier les résultats de concordance...

- **Collocation :**

En général, Différentes mesures statistiques peuvent être utilisées pour valider la force des collocations de mots. Ici, nous prenons en compte la capacité à produire des collocations de mots simples et de N- grammes.

- **Sauvegarde de la sortie :**

La capacité de préserver les résultats du traitement de corpus peut bien sur faire gagner du temps aux utilisateurs en ne les contraignant pas à retraiter le corpus s'ils veulent examiner encore les résultats. Dans notre estimation, nous avons vérifié la disponibilité de six formats principaux d'archivage des données, à savoir .txt, .svc, .htm, .html, .rtf et .xml, et s'il existait un autre format

III.6. Tableau comparatif des modes de fonctionnement des plateformes étudiés

En conclusion, nous présentons ci-dessous un tableau synthétique, résumant les méthodes sur lesquelles se fondent les plateformes étudiés, ainsi que leurs fonctionnalités :

Les critères \ Plateformes		ACPTs	AConCorde	Sketch Engine	Intellitext	Iramuteq	Lexi&Co
Interface arabe		×	×	**Non**	**Non**	**Non**	×
Lire fichier UTF-8 en arabe		×	×	×	×	×	×
Lire fichier Unicode en arabe		×	×	×	×	×	**Non**
Télécharger et ouvrir un corpus personnel		×	×	×	×	×	×
Format du texte de saisie		**.txt /.doc/.docx/.html**	**.txt**	**Tous les formats**	**.txt /.doc**	**.txt**	**.txt**
Normalisation	Diacritique	×	×	×	×	**Non**	**Non**
	Hamza "ء"	×	×	×	×	**Non**	**Non**
Liste de fréquence	Mots simple	×	×	×	×	×	×
liste des segments répétés		**Non**		×	**Non**	**Non**	×
Groupe de formes		**Non**	**Non**	**Non**	**Non**	Non	×
Représentation graphiques des résultats	Dendrogramme	Non	Non	Non	Non	×	Non
	Histogramme de fréquence	Non	Non	Non	Non	**Non**	Non

		Nuage des mots	Non	Non	Non	Non	×	Non
	Concordance		×	×	×	×	×	×
	Colocation		×	×	×	×	**Non**	×
	Classification		**Non**	**Non**	**Non**	**Non**	×	×
	Analyse des Spécificités		**Non**	**Non**	**Non**	**Non**	**Non**	×
	Analyse Factorielle des Correspondances (AFC)		**Non**	**Non**	**Non**	**Non**	×	×
	Affichage correcte de texte	De droite à gauche	×	×	×	×	×	×
		Signes diacritiques	×	×	×	×	×	×
	Sauvegarde de la sortie		×	×	**Non**	**Non**	×	×

Tableau 10 : *Tableau comparatif entre les plateformes*

III.7. Résultats et discussion

Sur la base des critères d'évaluation d'utilisabilité et de fonctionnalité susmentionnés. Chaque outil sera exploré en détail avec ses résultats de référence, qui seront suivies d'une brève comparaison globale fournie à la fin de cette section.

- **AConCorde(version 0.8)** :

AConCorde est relativement basique par rapport aux autres inclus dans ce chapitre, car il ne fournit aux utilisateurs que des concordances et une liste de fréquence de mots. Développé à l'origine pour la concordance en arabe natif, il possède une fonctionnalité de concordance de base, ainsi que des interfaces en anglais et en arabe.

- **LEXICO5 (version 5.5.1)** :

La première qualité de LEXICO5 est sa rapidité d'exécution puisqu'il ne faut pas plus de quelques secondes pour traiter de très gros corpus. Son deuxième avantage est son ergonomie moderne. Avec sa méthodologie les unités (un mot du texte par exemple) sont tirés vers les fonctions pour faire apparaître un graphique, une concordance, une AFC. La seule faiblesse de LEXICO 5 est son incapacité actuelle à traiter les sorties lemmatiseurs. On le favorisera donc seulement pour un traitement des formes graphiques et des segments répétés.

- **Iramuteq (version 0.7):**

Iramuteq il fournit également un éventail de statistiques descriptives sur l'ensemble du corpus : le nombre et la fréquence des mots, le nombre d'U.C.E (unité contextuelle élémentaire) classées, les formes lexicales, et beaucoup plus. Le point fort de Iramuteq c'est la fonctionnalité d'analyse centrale est l'analyse thématique de corpus, par la classification descendante de segments de textes D'une façon générale il est particulièrement travaillé la qualité, la richesse informationnelle et la diversité des visualisations, avec un usage caractéristique de nombreuses couleurs, associées aux différentes classes thématiques.

- **IntelliText (version 2.6):**

IntelliText fournit un nombre de corpus, y compris l'arabe, ainsi que de nombreux outils tels que concordances, collocations, affixes, comparer fréquences, mots clés et phrases…,mais concernant la performance ,il est trop lent surtout qu'on utilise des corpus volumineux

- **Sketch Engine :**

Sketch Engine à un large éventail de fonctions dont les principales comme Concordance, Word Sketch (cela nous donne un résumé des types de mots et des modèles de mots trouvés avec le mot de recherche.), liste de mots et un concordance de lemme et basique comme il donne aussi des listes de collocations des mot précis .

- **ACPTS (version 3.0):**

ACPTs est un outil spécifiquement pour le traitement de la langue arabe avec Interface arabe / anglais. Il est gratuit à télécharger. Il est sans doute devenu l'outil le plus fiable pour le traitement de textes arabes en raison de sa capacité à lire les caractères arabes. L'interface des ACPT est fournie avec des options en anglais et en arabe. Il donne la fréquence et la fréquence relative des types de forme et des documents (s'ils sont sélectionnés au cours du processus) dans les dossiers sélectionnés. De plus, il supporte les formats TXT. DOC. DOCX et HTML et codage ANSI ou UTF-8.

III.8. Conclusion

Six outils de recherche et d'analyse de corpus arabes ont été abordés et évalués selon plusieurs critères dans ce chapitre. Les résultats ont montré que trois de ces outils satisfaisant à la plupart des critères d'évaluations et atteignent des scores élevés, tandis que les autres valent des scores minimum.

Ce chapitre nous a permis de souligner les fonctionnalités les plus importante afin de nous inspirer dans la conception de notre plateforme , et ainsi la rendre plus appropriées à l'usage avec les corpus arabes, offrant d'avantage de fonctions compatibles avec les caractéristiques de la langue arabe, telles que les signes diacritiques et le hamza. Il a révélé aussi que bien que les outils informatiques aient des scores plus élevés que ceux basés sur lexico5, Sketch Engine était un puissant concurrent des outils informatiques. Cela peut indiquer qu'en principe il n'existe pas de Différences techniques importantes entre les outils basés sur PC et les outils basés sur le Web en termes de manipulation de la langue arabe. Ce qui est donc nécessaire, c'est que Les développeurs de concordance accordent assez d'attention aux caractéristiques de la langue arabe

IV. Vers une plateforme d'outils statistiques linguistiques appliqués sur des corpus arabe

IV.1. Introduction

En exploitant les différentes plateformes existantes de traitement statistique de la langue arabe cité dans le chapitre précédent, ces plateformes sont, plus ou moins, matures pour les langues latines. Néanmoins, il reste beaucoup de travail élaboré pour le cas de la langue Arabe. D'où la nécessité de converger vers une plateforme de développement dédiée au traitement statistique des corpus arabes standardisant l'ensemble des traitements et garantissant une meilleure flexibilité .notre objective est d'implémenter une plateforme web qui regroupe la plupart des outils statistiques.

Dans le présent chapitre, nous mettons le point sur la conception et l'architecture de notre plateforme en spécifiant ses principales caractéristiques .Nous discuterons ensuite les étapes essentielles pour développer notre outils .

IV.2. Architecture de plateforme développée

IV.2.1. Architecture globale

Pour désigne la structure générale inhérente à notre plateforme, l'organisation des différents éléments du système et des relations entre les éléments. La figure suivante (figure16) représente l'architecture globale de la plate-form

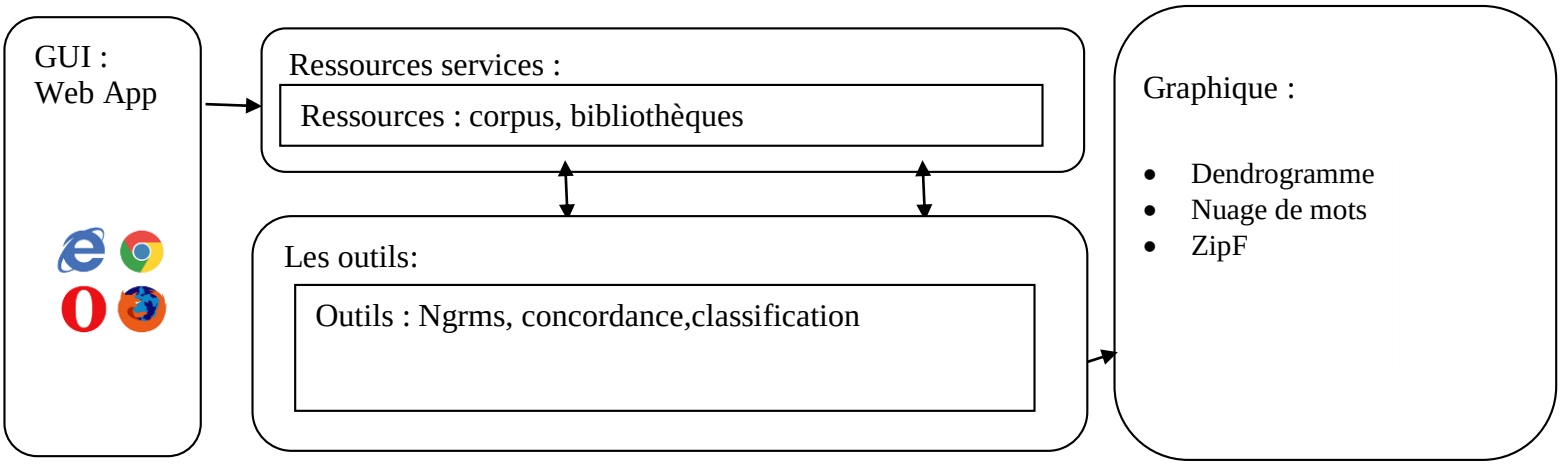


Figure 16: *architecture globale de plateforme*

Chaque module définit par :

- GUI : interface graphique intégré dans la plateforme pour utiliser directement les services
- Ressources services : Fournit des services de consultation des ressources linguistiques telles que le dictionnaire d'équivalence et les corpus
- Les outils : Regroupe une série des services techniques comme concordance, collocation, analyse fréquentielle, classification ….
- Graphique : permet de représenter les résultats des outils graphiquement en utilisant des diagrammes tels que dendrogramme, histogramme, nuage du mot …..

IV.2.2. **Architecture détaillé**

La figure 17 suivante représente l'architecture détaillée de notre système.

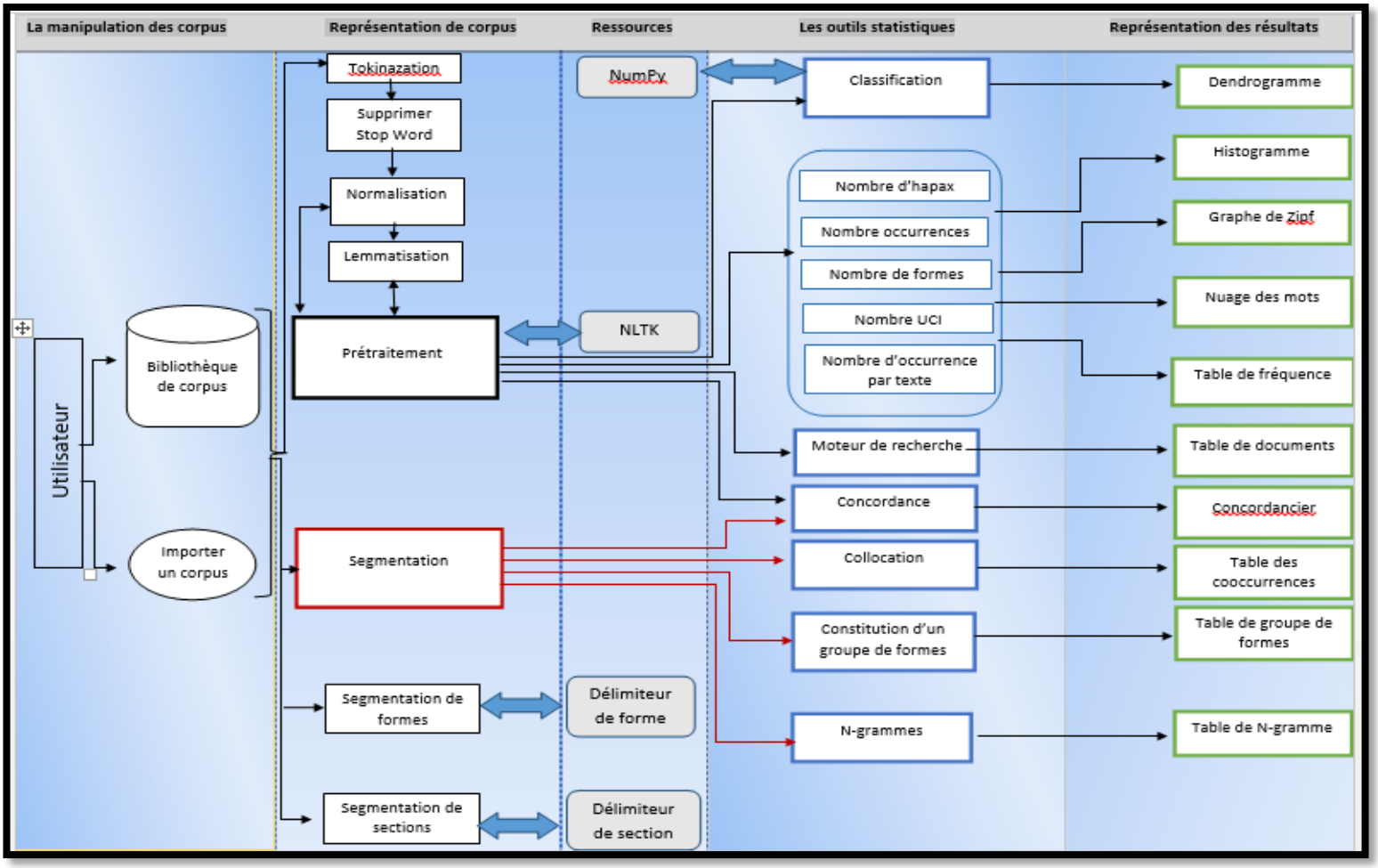


Figure 17 :*L'architecture détaillée de notre système*

Dans l'architecture détaillée, il y'a quatre principaux modules :

- Manipulation de corpus
- Représentation de corpus
- Les outils statistiques
- La représentation graphique

IV.2.2.a. **Manipulation de corpus**

Dans la conception de ce composant, nous avons défini un seule périmètre principal de qualité de représentation des données qui est TXT (texte brut), l'utilisateur au choix d'importer son propre corpus ou de sélectionner un des corpus existant dans notre bibliothèque qui en contient plusieurs.

IV.2.2.b. **Représentation de corpus**

Afin de bien manipuler les textes il est nécessaire d'utiliser une technique de représentation efficace. Dans notre projet l'utilisateur a le choix de traiter son corpus avec le prétraitement classique ou avec l'approche de la lexicométrie.

- **Prétraitement classique**

Cette étape consiste généralement en la représentation de chaque document par un vecteur, dont les composantes sont par exemple les mots contenus dans le texte, afin de le rendre exploitable par les algorithmes d'apprentissage. Une collection de textes peut être ainsi représentée par une matrice dont les lignes sont les termes qui apparaissent au moins une fois et les colonnes sont les documents de cette collection.

Un grand nombre de chercheurs dans le domaine ont choisi d'utiliser une représentation vectorielle dans laquelle chaque texte est représenté par un vecteur de *n* termes pondérés. A la base, les *n* termes sont tout simplement les *n* différents mots apparaissant dans les textes de l'ensemble d'entraînement, avant de faire ça il faut d'abord appliquer un prétraitement qui consiste sur quelques étapes comme montre la figure ci-dessus

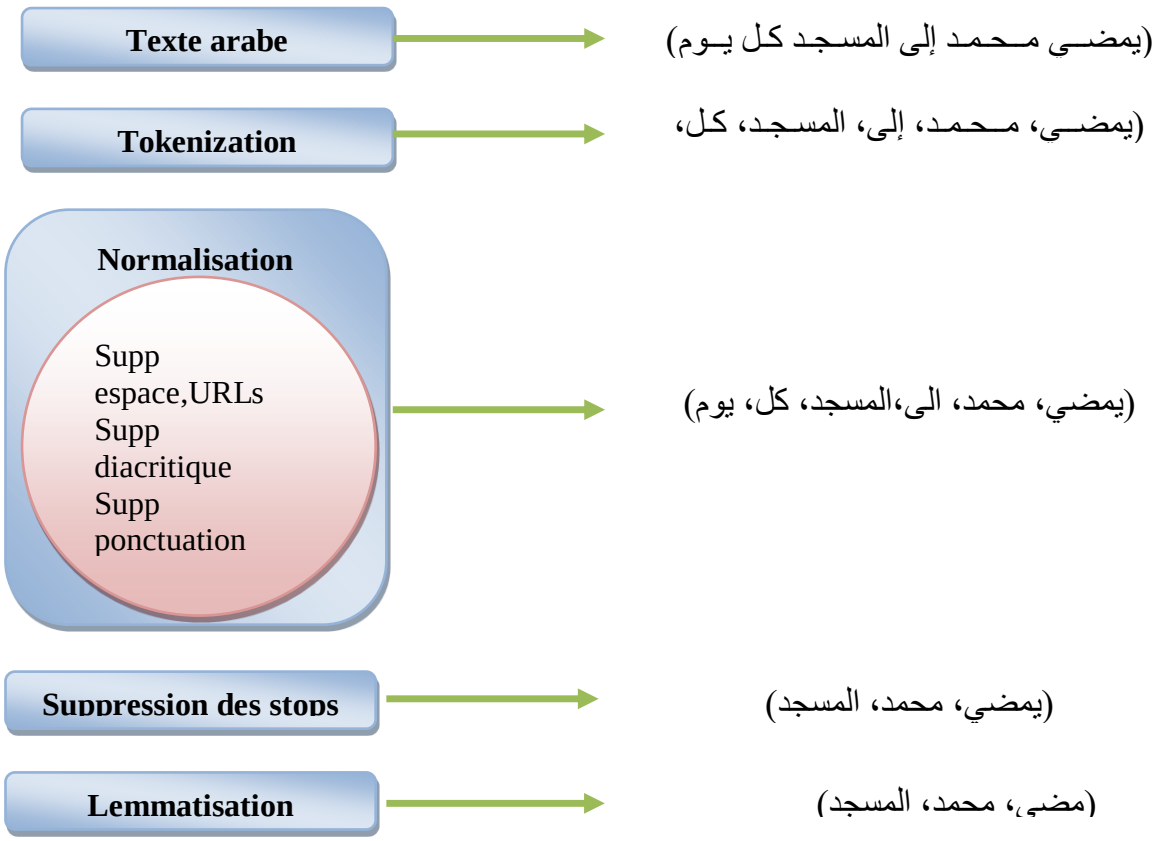


Figure 18 : *Un exemple d'étapes de prétraitement d'un texte arabe*

- **Tokenization :**

Le processus de décomposition d'un paragraphe de texte en jeton tels que des mots ou des phrases s'appelle Tokenization, il est la première étape de l'analyse de texte. Le jeton est une entité unique constituant des blocs de construction pour une phrase ou un paragraphe.

- **Supprimer les stops_words :**

Les stops_words sont les mots vides, ils étaient utilisés dans n'importe quel texte pour cela ils sont considérés comme sans importance différences entre les documents et comme du bruit dans le texte .

Le texte peut contenir des mots tels que من ،الى،و،بعد ...etc. Ils sont enlevés en ordre à obtenir plus significatif résultats en réduisant le nombre de faux positifs

- **Normalisation :**

La normalisation du texte est une technique importante pour l'analyse statistique. Il comprend de nombreuses étapes de prétraitement, notamment suppression des espaces, suppression tatwil, suppression de diacritique…etc (figure 19).Dans cette partie, la normalisation du texte joue un rôle important dans de nombreuses procédures telles que la segmentation du texte, l'extraction de caractéristiques et la reconnaissance des caractères.

Figure 19: *Exemple de normalisation de texte arabe*

- **Lemmatisation :**

La lemmatisation réduit les mots à leur mot de base, ce qui correspond à des lemmas corrects sur le plan linguistique (figure20). Il transforme le mot racine en utilisant du vocabulaire et une analyse morphologique. La lemmatisation est généralement plus sophistiquée que la radicalisation. Stemmer travaille sur un mot individuel sans connaître le contexte. Par exemple, le mot "mieux" a "bon" comme lemme. Cette chose manquera par la racine car elle nécessite une recherche dans le dictionnaire.

Figure 20: *Exemple de lemmatisation des mots arabe*

IV.2.2.c. La segmentation

La technique de la lexicométrie déjà défini dans le chapitre 2 mais avec quelques changements La lecture informatique du corpus doit permettre d'identifier d'une part les unités lexicales et d'autre part les unités contextuelles. Pour rendre possible cette double lecture du texte il est nécessaire de transformer ce dernier par l'introduction de deux types de marqueurs : les signes délimiteurs et les clefs de balisage.

- **Segmentation en unités contextuelles (sections) :**

La textométrie propose un deuxième niveau de segmentation avec le découpage contextuel en phrases et en paragraphes. On pourra retenir pour le marquage méta-textuel des paragraphes le signe '§' et pour celui des phrases un signe tel que '$' pour éviter toute ambiguïté avec les points d'abréviation particulièrement fréquents dans le texte.

- **Segmentation en unités lexicales (formes) :**

La technique de base de la lexicométrie consistait en l'expression d'une forme par une position. Rappelons qu'une forme textuelle analysable est une suite de caractères non délimiteurs contenue entre deux caractères délimiteurs dont ni le graphisme exact, ni la signification n'ont d'importance pour ce qui suit.

La détection automatique des formes graphiques s'effectue à partir de la division fondamentale des caractères composant le texte en deux catégories : les caractères délimiteurs et les caractères non-délimiteurs. Outre l'espace, considéré comme caractère séparateur par défaut, la liste de délimiteurs en vigueur dans le cadre de l'analyse du

corpus comprend les marques habituelles de ponctuation forte et faible, telles que [. ! ? , ; : …]. [34]

Pour une simple illustration de notre segmenteur, nous avons travaillées sur un "texte" appelé "corpus T" en utilisant:

- Au lieu de déterminer la forme par une seule position, ça sera par plusieurs positions (figure 22).
- Les dites positions seront maintenues chacune comme références pour la même forme (figure 21).

Figure 21 : La segmentation du "corpus T"

Ce "corpus" compte 19 occurrences, numérotées de 1 à 19 (position) .Parmi ces occurrences, il en est qui sont des occurrences de la même forme graphique. Ainsi, les occurrences qui portent les numéros : 5, 11 et 16 sont toutes des occurrences de la forme " **اللغات** " et on va les garder comme des références.

Au total on dénombre dans ce corpus 15 formes différentes qui constituent le vocabulaire du corpus P.

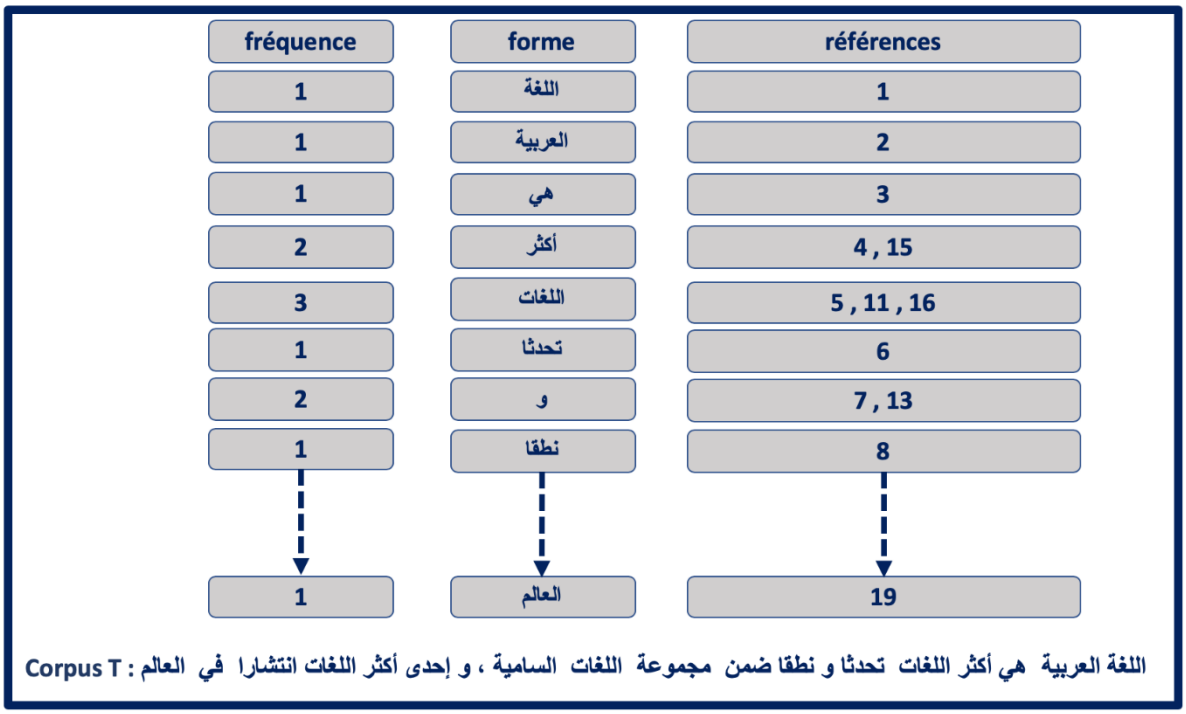


Figure 22 : *L'index hiérarchique du corpus T*

IV.2.2.d. **Les outils statistiques**

- **Analyse fréquentielle :**

L'analyse fréquentielle répond à une première question évidente : une unité de lexique est-elle présente dans le corpus analysé et, dans l'affirmative, avec quelle densité ?

L'information de base sur un texte ou une collection de textes est la liste de toutes les unités du lexique ou vocables (entrées du dictionnaire et catégories grammaticales correspondantes) avec leurs effectifs (c'est-à-dire le nombre de leurs occurrences), présentons d'abord les limites quantitatives globales du corpus à travers le tableau des dimensions lexicométriques ,nous avons met trois tableaux déférents de formes actives, formes supplémentaires et un tableau de toutes les formes de corpus (voir tableau11). Pour avoir ce dernier il faut d'abord que l'utilisateur choisit la méthode de représentation du corpus (segmentation ou prétraitement classique)

Rang	Mot	Occurrence
0	و	4
1	حرب	3
2	ايران	2
3	في	2
4	في	1
5	كبيرة	1
6	لقد	1

Rang	Mot active	Occurrence
0	حرب	3
1	ايران	2
2	في	1
3	كبيرة	1

Rang	Mot sup	Occurrence
0	و	4
1	في	1
2	لقد	1

Tableau 11 : *Table de fréquence d'un corpus arabe*

Le prétraitement classique donne comme résultat un sac de mots comme on a déjà annoncé, elle constitue une première étape nécessaire qui permet d'obtenir une première estimation des principales caractéristiques du corpus qui sont déjà définit dans l'Annexe :

- nombre d'occurrences
- nombre de formes
- nombre d'hapax
- nombre fréquence maximale …

Pour faire l'analyse fréquentielle mais sans devoir parcourir le corpus à chaque fois on a utilisé la méthode de la segmentation

- nombre d'occurrences : C'est la taille de corpus donc c'est la somme de toutes les fréquences.

  Exemple : dans la figure 21 nombre d'occurrences =19

- nombre de formes : La longueur de notre dictionnaire est le nombre de formes de corpus

  Exemple: dans la figure 21nombre de formes =15

- nombre des hapax : Les formes (اللغة, العربية, هي ….. etc.) n'apparaissent qu'une seule fois dans le corpus T donc c'est des hapax et pour les obtenir il suffit de chercher les formes de fréquence 1 dans la figure 21

Exemple: dans la figure 21 nombre des hapax =12

- nombre fréquence maximale : C'est la forme qui a le plus d'occurrence donc c'est le max des fréquences

Exemple: dans la figure 22 nombre fréquence maximale =3

La figure ci-dessous montre les principales caractéristiques quantitatives d'un corpus arabe exprimées en nombre de formes et en nombre d'occurrences, ainsi que le rappel des caractères délimiteurs choisis lors de la segmentation

```
Lexico3.1 PC DUCH
nbetiq=0
196125 196125 11023 142185 10859 6130 4953 5000000 14 8 143 0 0
*** Résultat de la segmentation du fichier: DUCH.TXT ***
Délimiteurs #— :;/\\.,?¿!¡*$\'"+=(){}[]§
nombre des occurrences : 142185
nombre des formes : 10859
frequence maximale : 6130
nombre des hapax : 4953
nombre des clés(type) : 8
nombre des clés(ctnu) : 143
*** Fin de la segmentation du fichier: DUCH.TXT ***
```

Figure 23 :*Exemple de fichier caractéristique fait par le logiciel Lex&Co5*

- **Les segments répétés**

Les formes graphiques ne sont évidemment pas des unités de sens. En effet, on observe les récurrences d'unités comme : sécurité sociale, niveau de vie, etc. dotées en général d'un sens qui leur est propre et que l'on ne peut déduire à partir du sens des formes qui entrent dans leur composition.

Dans les études textuelles, il sera utile de compléter les résultats obtenus à partir des décomptes de formes graphiques par des comptages portant sur des unités plus larges, composées de plusieurs formes .Il est alors utile de soumettre ces unités aux mêmes traitements statistiques que les formes graphiques.

Figure 24: *image de segment répète [tirée de lexico 3]*

- **Collocation :**

Précisons d'abord que les collocations ne sont pas à confondre avec la notion de segments répétés souvent utilisé dans les statistiques lexicales. La raison en est qu'un segment répété peut être une collocation et que toute collocation n'est pas pour autant un segment répété.

Les recherches sur les segments répèté se sont développées à partir des difficultés rencontrées dans le domaine de la recherche des cooccurrences (attirances particulières entre couples de formes au sein d'une unité de contexte donnée). les méthodes utilisées dans ce but varient fortement en fonction des domaines d'application, on mentionnera très brièvement, dans les paragraphes qui suivent, quelques- unes des méthodes élaborées pour répondre à cette préoccupation.

La méthode des collocation affecte un indice de probabilité au nombre des rencontres, chaque fois à l'intérieur d'une phrase (séquence de formes entre deux ponctuations fortes), de chaque couple et de chaque paire (couple non-orienté) de formes du texte. Cette méthode permet de sélectionner, à l'aide d'un seuil en probabilité des ensembles de couples et de paires de formes présentant des affinités dans le corpus de textes que l'on étudie.

- ***La* classification ascendante hiérarchique *(*CAH*)***

Dans notre plateforme nous détaillerons une méthode du type « clustering» qui est une des techniques statistiques largement utilisées dans la Fouille de Données et parmi ces techniques : la classification ascendante hiérarchique (CAH)

La classification ascendante hiérarchique ascendante est l'idée de créer à chaque étape une partition obtenue en agrégeant 2 à 2 (figure25) les éléments représenté sous forme des vecteurs les plus proches en sachant que un Elément est un individus ou groupe d'individus donc par principe chaque point ou individus ou cluster est progressivement « absorbé » par le cluster le plus proche. [35]

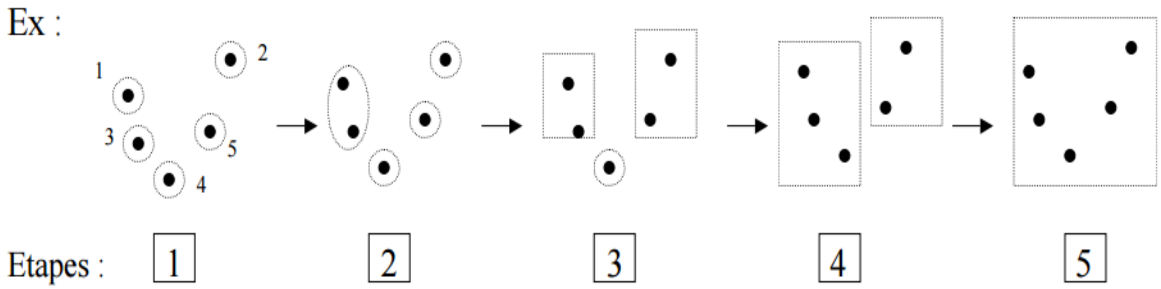


Figure 25: *Les étapes d'une classification hiérarchique non supervisée*

- **Algorithme de CHA :**

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative, son principe est de rassembler des individus selon un critère de ressemblance défini au préalable comme on a déjà expliqué (voir chapitre 2) dont l'algorithme est simple :

1. Initialisation :

Chaque individu est placé dans son propre cluster, Calcul de la matrice de ressemblance avec les mesures de similarités M entre chaque couple de clusters (ici les points).

2. Répéter :

   - Sélection dans M des deux clusters les plus proches Ci et Cj.

- Fusion de Ci et Cj par un cluster CG plus général.
- Mise à jour de M en calculant la ressemblance entre CG et les clusters existants Jusqu'à fusionner les 2 derniers clusters.

Ces regroupements successifs produisent un arbre binaire de classification (dendrogramme), dont la racine correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions.

- **TF_ IDF:**

La première étape pour appliquer l'algorithme de CAH il existe une méthode typique qui consiste à calculer le poids des termes des deux documents en utilisant TF-IDF, organisés dans les matrices documents-termes. Chaque document sera donc représenté par un sac de mots (bag of Words) ou similaire (racines, lemmes, termes composé).

Le modèle de sac de mots est le moyen le plus simple d'extraire des caractéristiques du texte. Nous pouvons créer une matrice de document et de mots en comptant l'occurrence de mots dans le document donné. Donc il convertit le texte en matrice d'occurrence de mots dans un document (tableau12). Ce modèle concerne la question de savoir si des mots donnés apparaissent ou non dans le document. Comme toujours, c'est plus simple avec un exemple, Prenons 2 documents où chaque phrase est un document séparé:

Doc1 : ذهب الكاتب إلى المؤتمر

Doc2 : ذهب القراء إلى عين المكان حيث المؤتمر

terme / doc	ذهب	الكاتب	إلى	المؤتمر	القراء	عين	المكان	حيث
Doc1	1	1	1	1	0	0	0	0
Doc2	1	0	1	1	1	1	1	1

Tableau 12 : *matrice d'occurrence*

- Fréquence de terme (TF) :

Dans TF on ne compte que le nombre de mots contenus dans chaque document. Le principal problème de cette fréquence de terme est qu'elle donnera plus de poids aux documents plus longs, donc TF est le nombre d'occurrences (en utilisant la matrice d'occurrence) du terme au sein du document [36]

$$\boldsymbol{tf_{i,j}} = \frac{\boldsymbol{n_{i,j}}}{\sum_k \boldsymbol{n_{i,j}}} \text{ [37]}$$

- Fréquence de document inverse (IDF) :

L'IDF mesure la quantité d'informations qu'un mot donné fournit sur le document. IDF est le rapport inverse, mis à l'échelle logarithmique, du nombre de documents contenant le mot et du nombre total de documents.

$$\boldsymbol{idf(w)} = \boldsymbol{log}\left(\frac{\mathrm{N}}{\mathrm{df_t}}\right) \text{ [37]}$$

- Terme Fréquence-Fréquence Document Inverse (TF-IDF) :

TF-IDF normalise la matrice de termes du document. C'est le produit de la TF et de l'IDF. Mot avec tf-idf élevé dans un document, il s'agit le plus souvent de documents donnés et doit être absent des autres documents. Donc, les mots doivent être un mot de signature.

$$\boldsymbol{W_{i,j}} = \boldsymbol{tf_{i,j}} \times \boldsymbol{log}\left(\frac{\mathrm{N}}{\mathrm{df_i}}\right) \text{ [37]}$$

Où :

$\boldsymbol{tf_{i,j}}$: nombre d'occurrence de i dans j

$\boldsymbol{tf_i}$: Nombre de document qui comtien i

N : nombre totale de document

Nous allons maintenant calculer le TF-IDF pour les deux documents ci-dessus, qui représentent notre corpus :

Terme	TF		IDF	TF*IDF	
	Doc1	Doc2		Doc1	Doc2
ذهب	1/4	1/7	Log(2/2)=0	0	0
الكاتب	1/4	0	Log(2/1)=0.33	0.082	0
إلى	1/4	1/7	Log(2/2)=0	0	0
المؤتمر	1/4	1/7	Log(2/2)=0	0	0
القراء	0	1/7	Log(2/1)=0.33	0	0.047
عين	0	1/7	Log(2/1)=0.33	0	0.047
المكان	0	1/7	Log(2/1)=0.33	0	0.047
حيث	0	1/7	Log(2/1)=0.33	0	0.047

Tableau 13 : *Exemple de calcule de TF-IDF*

A la fin de cette étape (calcule de TF-IDF) on obtient une matrice terme-document (DTM), qui contient pour chaque terme sa valeur TF-IDF dans chaque document (tableau 14)

terme / doc	ذهب	الكاتب	إلى	المؤتمر	القراء	عين	المكان	حيث
Doc1	0	0.082	0	0	0	0	0	0
Doc2	0	0	0	0	0.047	0.047	0.047	0.047

Tableau 14: *Matrice document-terme (DTM)*

- **Fonctions de similarité :**

Pour comparer l'homogénéité, la ressemblance ou la similarité entre deux objets (points, images, classes, phonème..), il faut pouvoir mesurer la similarité (ou la dissimilarité) entre eux.

Nous allons décrire maintenant des mesures de similarité pour prouver la similarité entre les objets. Donc la similarité est une partie importante de la définition d'une méthode de clustering, elle consiste en effet à définir et formaliser une mesure de similarité adaptée aux caractéristiques des données

- **La distance euclidienne** :

Aussi appelée la distance à vol d'oiseau la mesure de la distance la plus courante dans les études publiées dans ce domaine de recherche est la distance euclidienne ou la distance au carré euclidienne.

$$D\ 2\ (x1, x2) = \sum i\ (x1i - x2i)\ 2 = (x1 - x2)\ (x1 - x2)\ [38]$$

- **La distance de Manhattan :** (appelée aussi taxi-distance)

$$D\ 2\ (x1, x2) = \sum i\ |\ x1i - x2i\ |\ [38]$$

- **La distance cosinus :**

La distance ou la similarité entre le Di et Dj deux documents est calculé de cette manière :

$$\cos(Di, Dj) = \frac{|Di \cap Dj|}{\sqrt{|Di||Dj|}}\ [39]$$

Ou :

- Di et Dj sont des descriptions de deux documents distincts.
- Deux documents sont similaires s'ils contiennent des termes similaires
- Deux termes sont similaires s'ils apparaissent dans des documents similaires

- **Méthodes d'agrégations:**

  - **saut minimum :**

La méthode du saut minimum correspond au choix du "meilleur des cas". Dans cette méthode, on choisit comme représentant du groupe le document qui est le plus similaire avec les documents de l'autre groupe. Ainsi la similarité entre les clusters est le maximum de la

similarité entre les documents des clusters. Cette méthode (connue sous le nom de « single linkage » en anglais ») consiste à écrire que :

$$\Delta(Doc1, Doc2) = min_{i \in Doc1\ , j \in Doc2}\ d(i,j)$$ [39]

- **saut maximum :**

Il s'appelle aussi diamètre ou complete linkage .La distance entre deux documents est la plus grande distance entre éléments des deux documents

$$\Delta(Doc1, Doc2) = max_{i \in Doc1\ , j \in Doc2}\ d(i,j)$$ [39]

- **Concordance :**

Comme on déjà dit, la concordance est un outil de contextualisation qui joue un roule très important dans la statistique textuelle, Il existe quatre types d'affichage de concordance KWAC, KWUT, KWOC... (Voir chapitre 2) .Dans notre application nous avons conçu deux concrodancier dont le type est le plus demandé par les linguistes (KWAC, KWUT) avec les deux méthodes présentées dans ce chapitre.

Nous avons conçu deux concordanciers de type KWAC l'un en utilisant la segmentation lexicometrique et l'autre en utilisant prétraitement classique que nous allons expliquer ci-dessous :

- **Un concordancier suivant le prétraitement classique :**

Nous avons utilisé la méthode basique de représentation de données (prétraitements classique), mais sans éliminer les mots vides ou séparer les mots (tokinisation) .pour créer ce concordancier de KWAC nous avons parcouru le corpus de la première lettre jusqu'au mot recherché en gardant toujours l'index du dernier caractère du mot recherche, en stockant au fur et à mesure séparément les contextes d'avant et les contextes d'après.

Le contexte avant est la suite des caractères qui commence par la lettre de l'index de début (dans l'exemple l'index est 0) jusqu'à la dernière lettre (index dernière lettre =index de dernière lettre de mot recherché –taille du mot).

Le contexte après est la suite de caractères qui commence par l'index de dernière lettre de mot recherché jusqu'à la fin. (Figure 26)

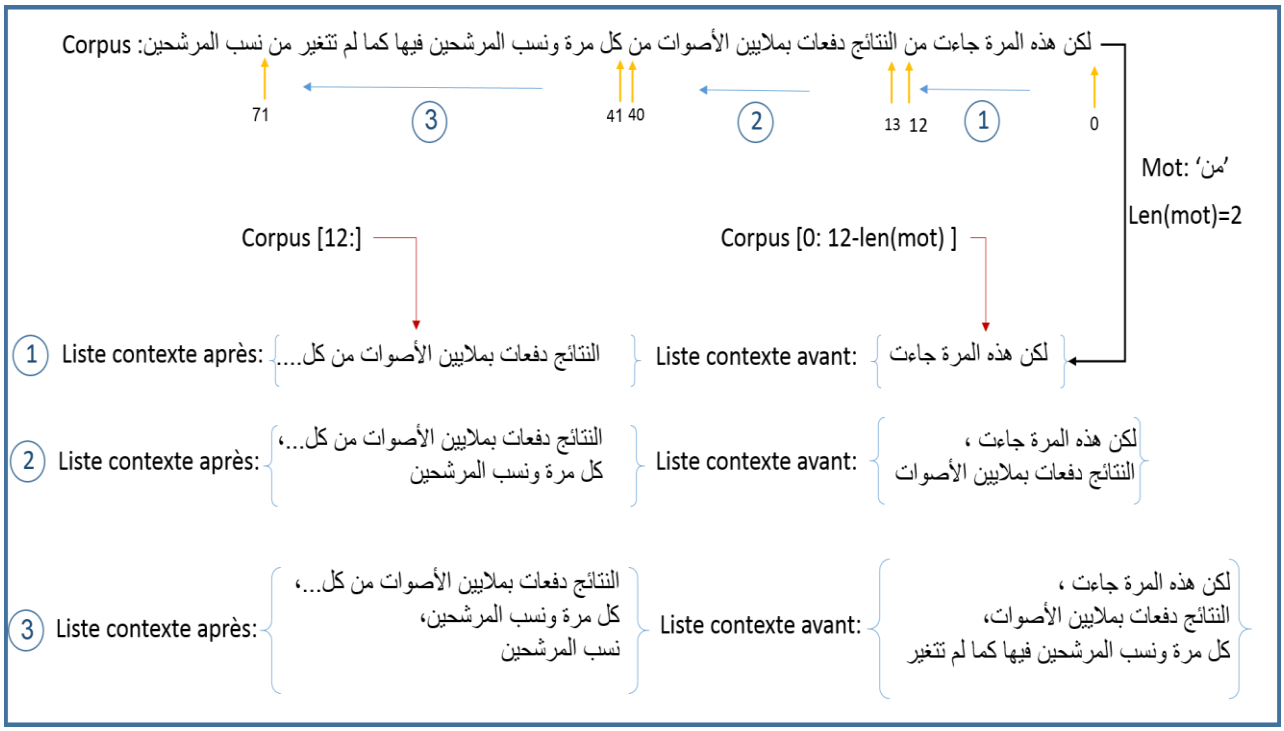


Figure 26: Exemple illustre les principales étapes de création de concordance

Nous avons répété ces étapes en boucle jusqu'à la fin de corpus.

après on peut limiter la taille de contexte avant et après selon le choix d'utilisateur et on crée une liste finale (figure27) qui contienne toutes les informations essentielle pour le concordancier sou forme :

[Contexte après, mot, contexte avant]

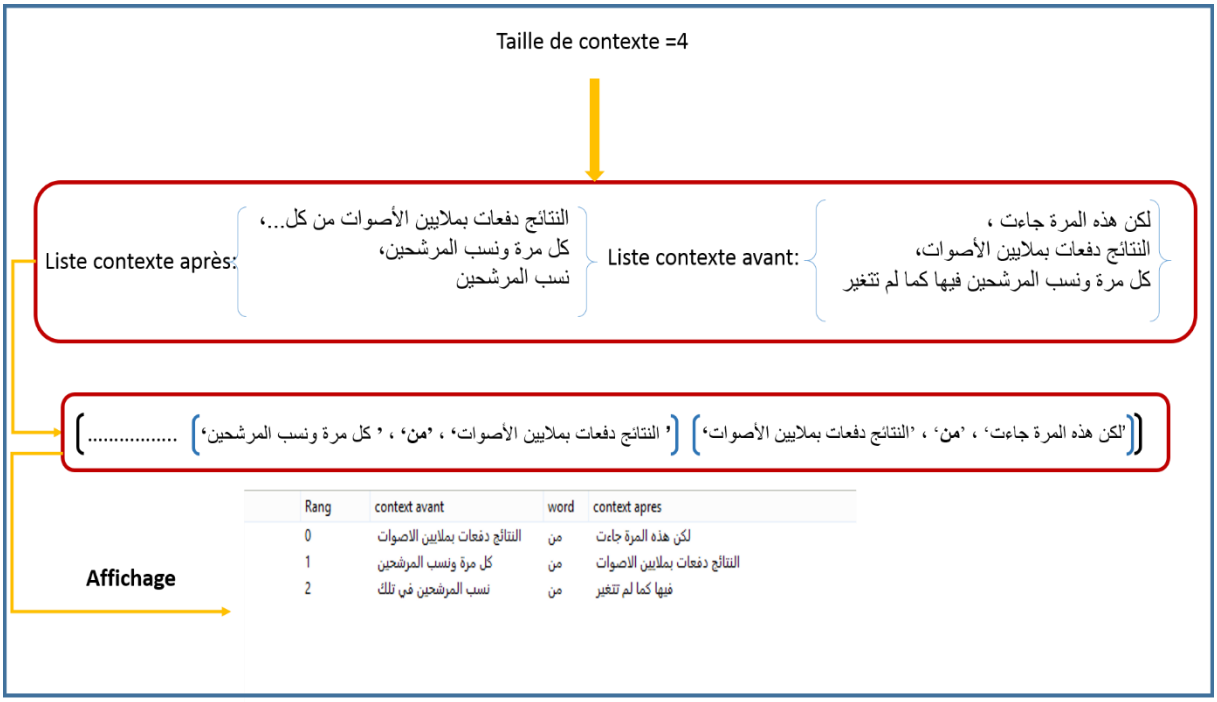


Figure 27 : *Dernière étape de création d'un concordance de 'من'*

- **Un concordancier suivant la segmentation lexicometrique :**

Les index (positions) qui constituent par rapport au corpus de départ une réorganisation des formes et des occurrences, permettent de repérer immédiatement, pour chacune des formes, tous les endroits du corpus où sont situées ses occurrences. Pour retrouver ces occurrences dans le texte de départ, on utilise un système de coordonnées numériques liées à la position de l'occurrence dans la ligne. Ces renseignements qui permettent de retourner plus facilement au document d'origine sont les références associées à chacune des occurrences (figure 28). [34]

La figure 27 contient l'index hiérarchique réalisé à partir du corpus T. Dans cet index, c'est le numéro d'ordre (position) dans le corpus qui sert à référencier chacune des occurrences.

Il peut être intéressant d'étudier systématiquement les contextes immédiats dont les occurrences d'une forme ont été extraites et pour localiser l'ensemble de ces derniers dans le texte d'origine (corpus T), On se sert des références constitué par rapport au corpus de départ.

On appelle forme-pôle la forme dont on regroupera les contextes. On trouve ci-dessous l'exemple d'une concordance qui regroupe toutes les occurrences relatives à une même forme-pôle, la forme ''**اللغات**'' du corpus T. [34]

Dans cet exemple, les lignes de la concordance sont triées d'après l'ordre hiérarchique de la forme qui suit la forme-pôle. La référence située à droite est la position d'occurrence dans le corpus T.

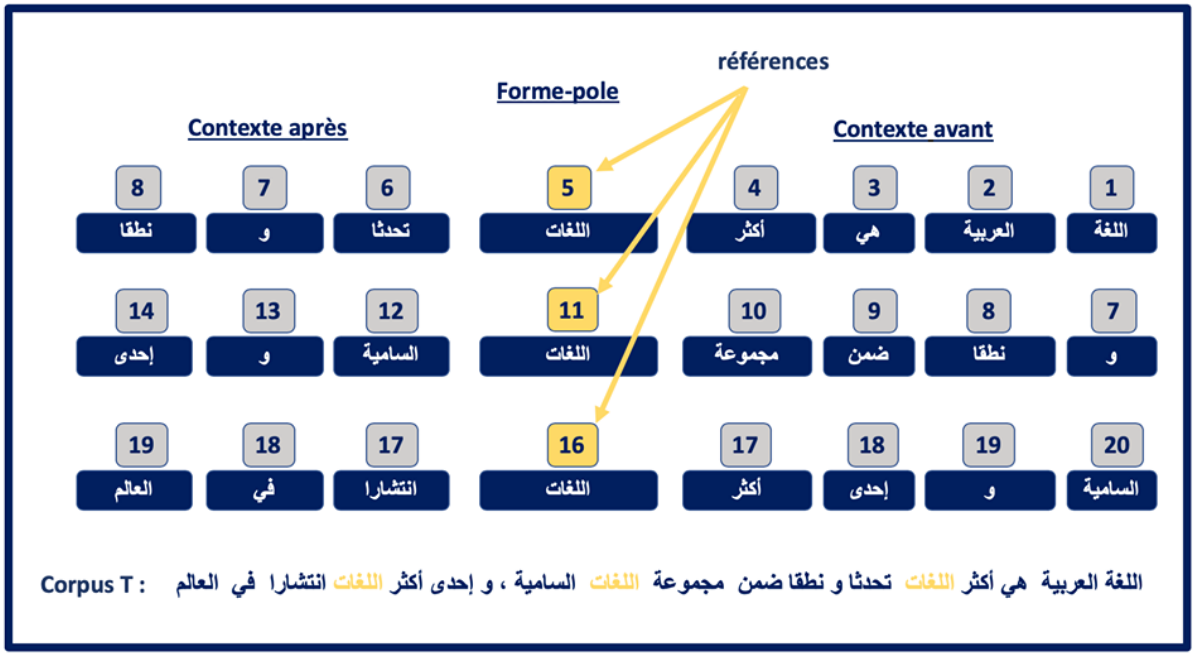


Figure 28: *Concordance de la forme اللغات dans le corpus T*

La Figure 28 contient les lignes d'une concordance réalisée pour les contextes de la forme **اللغات** dans le corpus T. La forme enfants compte 3 occurrences. Les lignes de contexte sont ici triées par ordre hiérarchique en fonction de la forme qui suit la forme-pôle.

IV.2.2.e. **Représentation des résultats :**

Dans un premier temps, nous avons effectuées une analyse lexicale afin de déterminer les éléments principaux de notre corpus .Nous obtenons les entités les plus fréquemment utilisées dans la presse. Il ne s'agit donc pas ici simplement d'une liste de mots, mais de l'évaluation de la fréquence d'apparition de différents objets.

Nous traduisons cette liste par plusieurs graphes qui facilite la lecture et l'interprétation :

- Dendrogramme
- Histogramme
- Nuage de mots
- Graphe de zipf

IV.3. Conclusion :

Dans ce chapitre, nous avons présenté les méthodes utilisée pour concevoir et développer notre plateforme d'analyse statistique des données textuelle, en décrivant l'architecture du notre système avec une description des deux différents modules utiliser pour la représentation des corpus, nous avons par la suite expliqué les différentes étapes de développement des outils statistique de notre plateforme. En finale nous avons défini les différentes méthodes de représentation des résultats.

V. Implémentation et test

V.1. Introduction

Nous arrivons dans ce chapitre à la description de l'aspect pratique de notre travail. Dans la description de notre plateforme qui suivra, nous mettrons l'accent sur le côté visuel (les interfaces) afin montrer sa facilité d'utilisation qui nous a été un objectif principal. En effet, Nous avons essayé de concevoir une interface intuitive et pratique. Nous décrivons aussi dans ce chapitre l'ensemble des moyens technologiques utilisés dans le développement de notre plateforme.

V.2. Environnement de développement

Chaque plateforme nécessite, lors de sa création et de son exécution, un appui matériel et un autre logiciel afin de réaliser les fonctionnalités pour lesquelles il a été conçu.

V.2.1. Environnement Matériel

Pour développer l'application, nous avons utilisé comme environnement matériel deux ordinateurs :

- **PC HP Pavilion g6 Notebook PC:**
  - Processeur : Intel(R) Pentium(R) CPU B960 @ 2.20GHz
  - Mémoire vive : 4Go.
  - Disque dur : 698 Go
  - Taille d'écran :15,6 pouces.
  - Type de système : Windows 8 avec système d'exploitation 64 bits
- **MacBook Pro (Retina, Mid2012):**
  - Processeur : Intel Core i7 2.30GHz
  - Mémoire vive : 8Go.

- Disque dur : 512 Go
- Taille d'écran : 15,4 pouces.
- Type de système : Mac OS X 10.7 Lion

V.2.2. **Environnement langages de programmation utilisés**

V.2.2.a. **Python**

Python est un langage et une plateforme de développement logiciel complète et généraliste, très facile d'accès et capable de se spécialiser de manière très pointue dans la majorité des domaines informatiques. Python est utilisé par un public très large : des développeurs web professionnels, des chercheurs en intelligence artificielle ou en bio-informatique, des administrateurs systèmes…

Figure 29: *Python logo*

V.2.3. **Flask**

Flask est un micro-framework web python qui permet de réaliser des sites web dynamiques. Il est utilisé dans Analytex, car il permet de faire le lien entre le code python utilisé au sein du projet Analytex et l'interface.

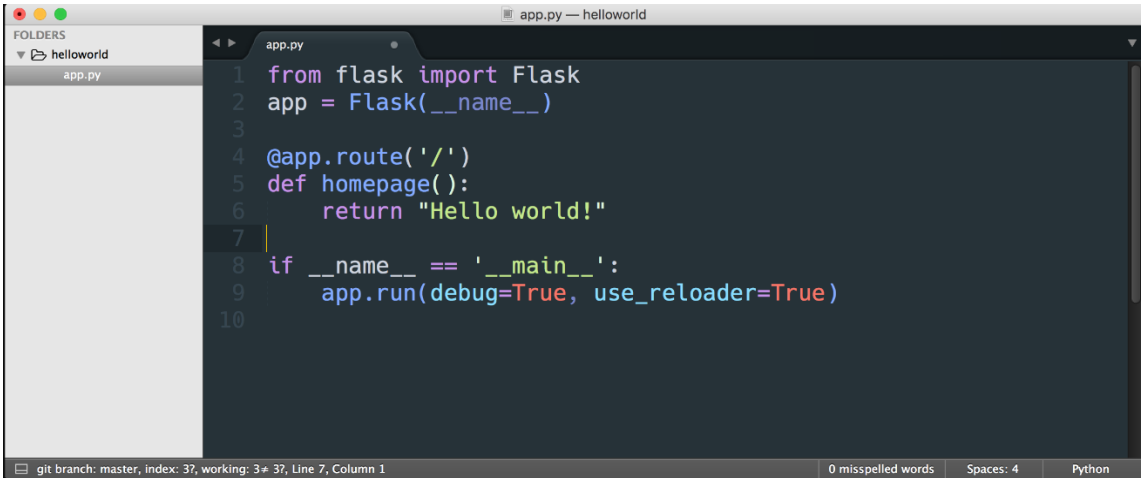


Figure 30 : *Hello world avec Flask*

V.3. Format des données en entrée

Les fichiers d'entrée doivent être au format texte brut (.txt), et de préférence contenir les caractères de ponctuation. Dans ce formatage, l'unité de base est appelée « Texte ». Un texte peut représenter un entretien, un article, un livre ou tout autre type de documents. Un corpus peut contenir un ou plusieurs textes (mais au minimum un).

V.4. Jeux de données

La base qu'on a adopté pour nos tests à fin d'établir nos outils c'est un base de documents à base de fichiers textes (.txt) de taille 4 755 fichiers textes (7 dossiers), des fichiers qui sont déjà classifier en sous classes (figure31), afin de vérifier notre classification au résultat.

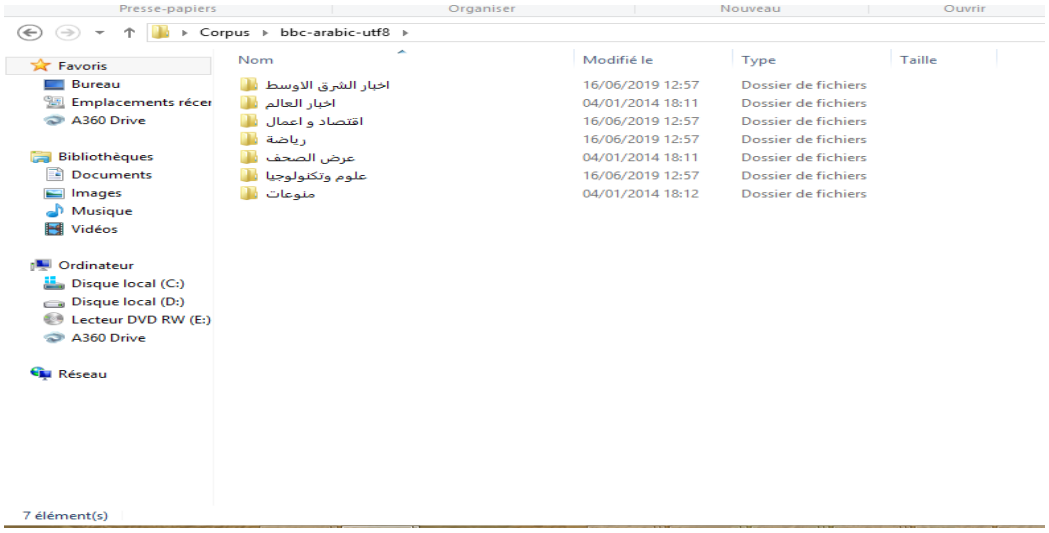


Figure 31 : *base de document*

V.5. Base de données

On a créé une base de données pour pouvoir garder l'historique de chaque utilisateur et enrichir notre bibliothèque de corpus, elle nous garantit aussi la gestion d'utilisateur.

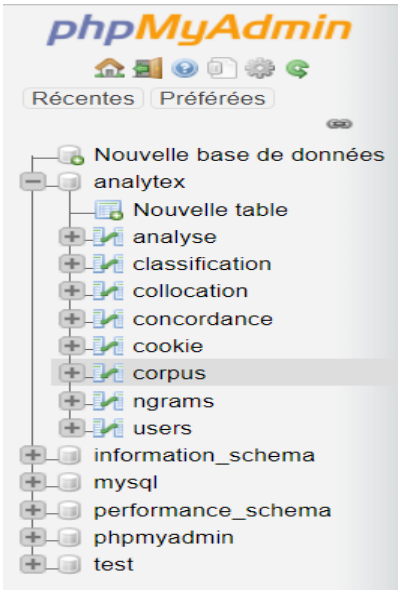


Figure 32 : *les tables de la base de donnée de Analytex*

- **Exemple de Table d'historique d'analyse fréquentielle :**

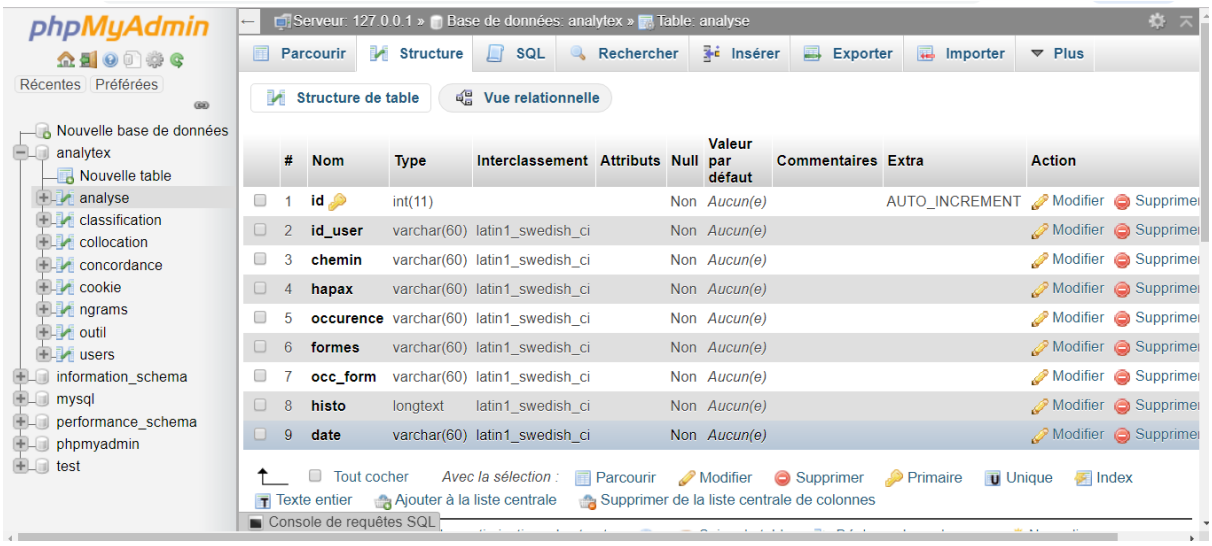


Figure 33 : *table de fréquence (historique)*

V.6. L'architecture de l'implémentation :

Notre projet est composé de 2 répertoires, un seul fichier Backend relié à une base de données, et un serveur Windows.

- Le 1er répertoire, nommé **Static**, contient les fichiers suivants : « CSS, JS, et un fichier image qui contient toutes les images statique de notre projet (icones, ..).
- 2ième répertoire, nommé **Templates**, contient tous les fichiers HTML.
- Le fichier Bachend, nommé **All** qui dans le dossier **ANALYSE STATISTIQUE,** avec l'extension .py (c'est un fichier de langage Python), nous avons utilisé un système modulaire qui englobe tous nos outils (un élément peut être remplacé ou modifié sans devoir changer toute l'architecture)

- La base de données est une BD MySQl, contient 8 tables.

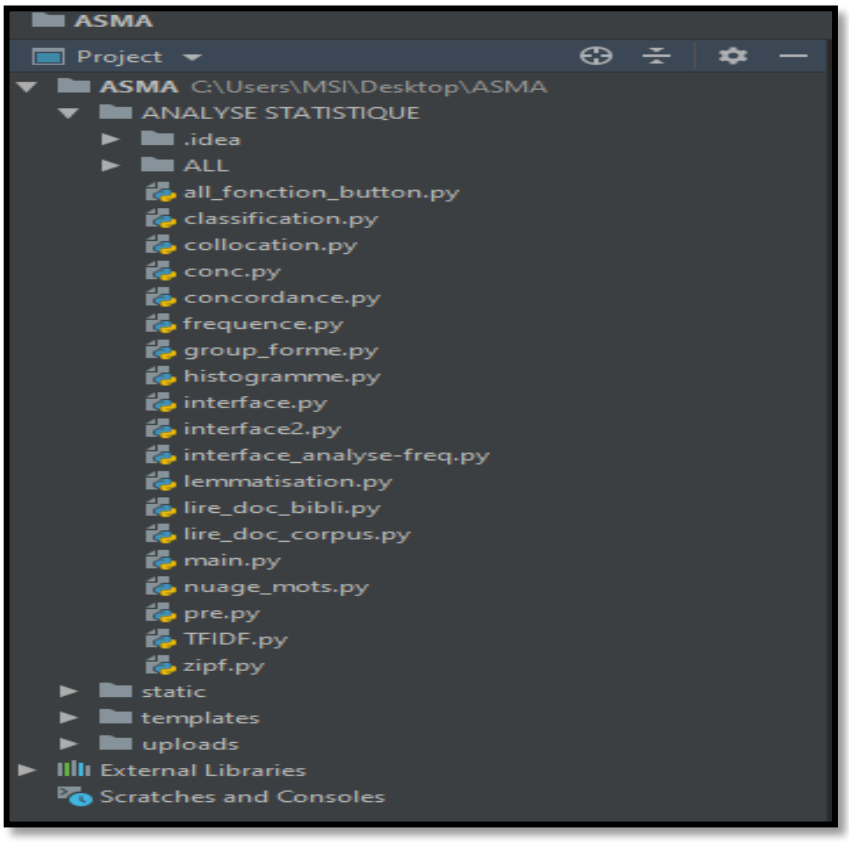


Figure 34 : *base de document*

V.7. Les Besoins non fonctionnels :

Pour donner plus de performance à notre site, on a décidé de traiter les besoins non fonctionnels suivants :

V.7.1. Gestion de cookies :

Dans notre plateforme on a géré les cookies, ce dernier est un fichier qui est déposé par le navigateur sur l'ordinateur lorsqu'on surfe sur internet. Donc à chaque fois qu'un utilisateur effectue un Login le serveur de site consulté lui affecte un cookie, ce cookie sera envoyé au navigateur internet. Ce cookie est un moyen pour suivre la connexion de la session de l'utilisateur. Chaque cookie a un temps, à chaque fois que l'utilisateur visite une page, et en cours de la visite le site va vérifier si le temps courent est supérieur au temps de cookie ou non,

si oui l'utilisateur sera rediriger vers la page Login, sinon ce temps s'incrémente avec 1000 secondes.

V.7.2. **Gestion d'image :**

La plateforme contient un nombre important d'images, il est nécessaire de les enregistrer dans la Base de données afin de garder l'historique de chaque utilisateur. Il existe l'ancienne méthode d'enregistrement qui consiste de stocké le chemin de l'image dans la BD. Mais dans notre projet on a préféré d'utiliser la nouvelle méthode : On récupère l'image à partir de l'input de type file, cette image sera codé avec l'encodage base 64 et stocké dans la BD.

V.7.3. **Gestion de sécurité :**

Quand vient le temps de protéger des informations sensibles stockées dans une base de données, comme les mots de passe, on utilisera le cryptage. Dans notre plateforme nous avons utilisé SHA-1 (Secure Hash Algorithm) comme algorithme de hachage, après que l'utilisateur taper le mot de passe, nous lui appliquons ce dernier, nous obtenons une chaine de caractère hexadécimale de 40 caractères, et c'est ce qui est placé dans la base de données au lieu de mot de passe saisie par l'utilisateur.

V.7.4. **Bibliothèque de corpus :**

Dans notre plateforme nous avons enregistré les corpus dans un dossier qui s'appelle **'UPLOAD'** à chaque fois que l'utilisateur importe un corpus afin de créer une bibliothèque de corpus

V.8. **Interfaces de l'application et test**

Dans ce qui suit, nous allons présenter quelques interfaces de l'application " Analytext ".

V.8.1. **Page d'Accueil**

Cette page est consacrée à la présentation de notre projet (statistique textuelle)

Figure 35 : *Page d'accueil partie 1*

Cette figure informer nos utilisateurs de notre objectives et les avantage de notre plateforme

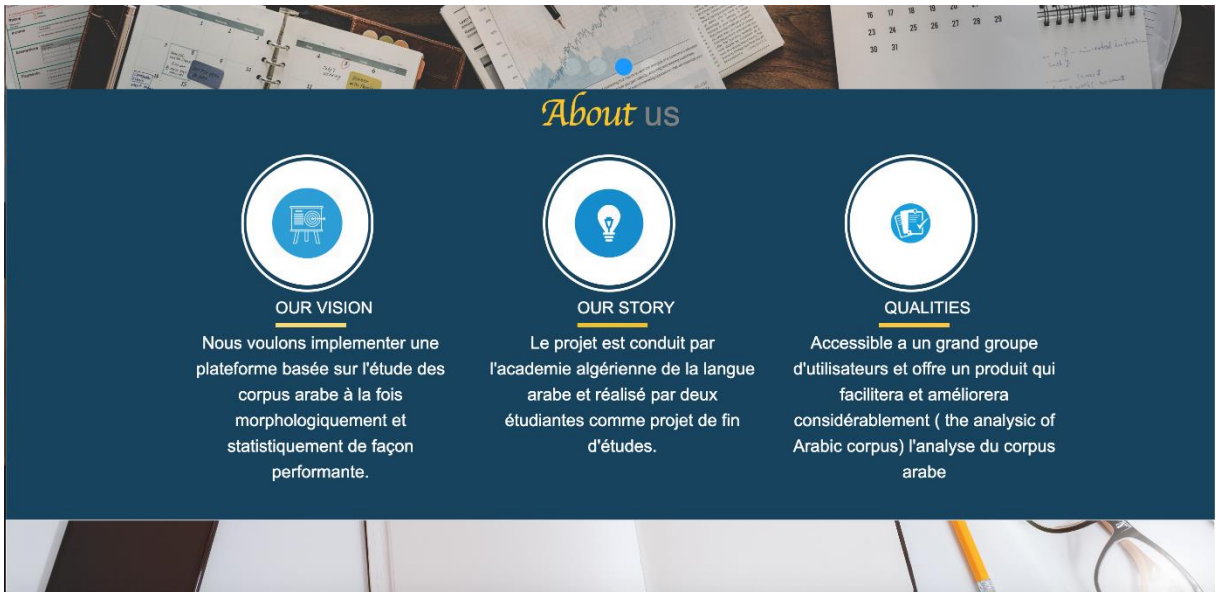


Figure 36: *Page d'accueil partie 2*

Dans cette partie l'utilisateur peut contact nous.

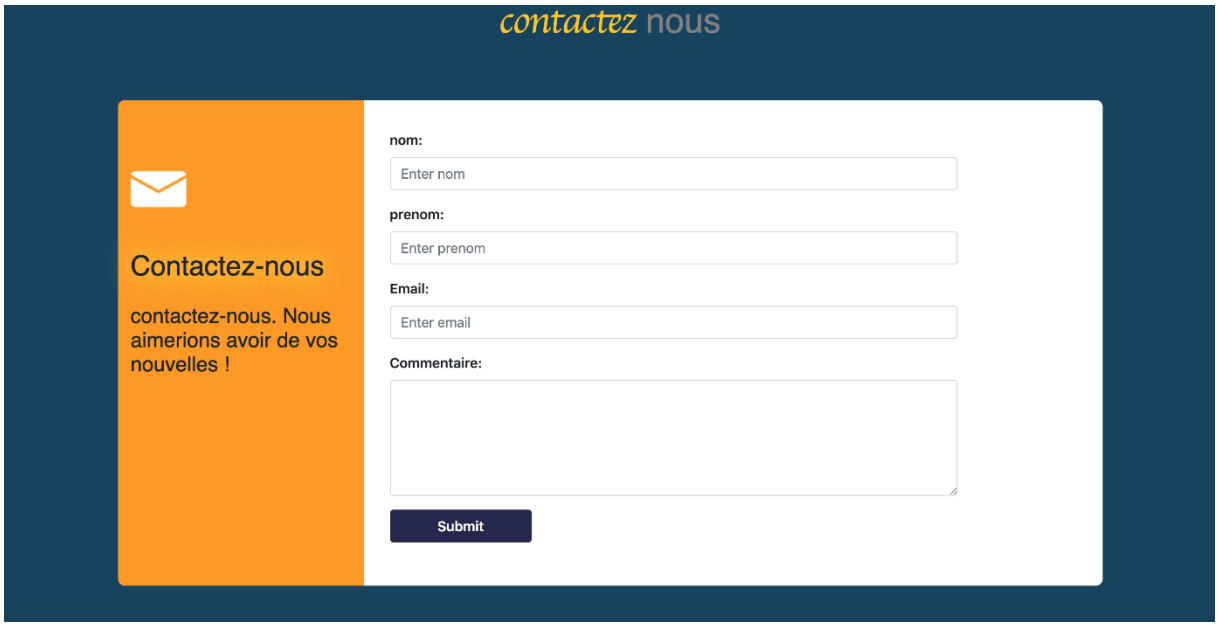


Figure 37 : *Page d'accueil partie 3*

V.8.2. **Fenêtre d'Authentification**

Apres que l'utilisateur appuiera sur le bouton **'outil'** dans la page home, une page authentification sera afficher .Cette page permet à l'utilisateur de s'authentifier avec un login et mot de passe pour se connecter au serveur de la base de données.

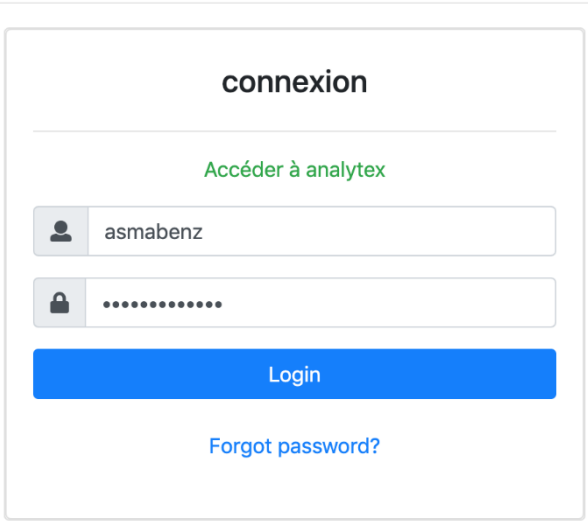


Figure 38: ***page authentification***

En cas d'erreur un message d'alerte s'affiche, elle renvoie une fenêtre avec l'un des messages suivant :

- Authentification obligatoire : si on n'a rien saisi.
- Vérifiez votre login et/ou mot de passe : si on a saisi un login et/ou mot de passe incorrect.
- Vérifiez votre mot de passe : si on a saisi un mot de passe incorrect.

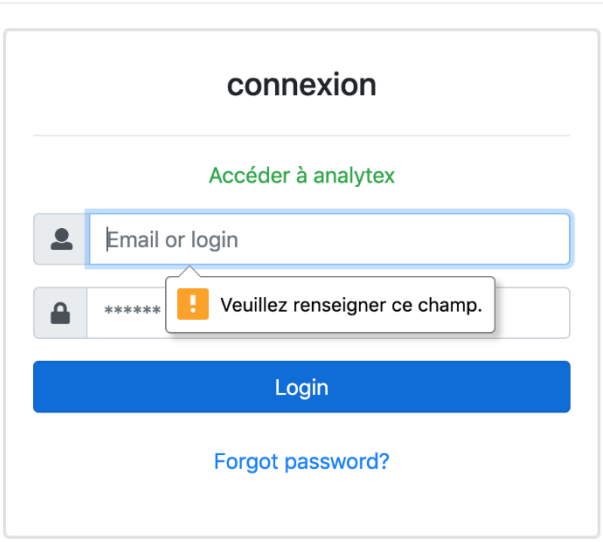


Figure 39: *message d'erreur en cas de champ vide*

V.8.3. **Page inscrire**

La figure ci-dessus présente page d'inscription, l'utilisateur peut s'inscrire en cliquant sur bouton **'inscrire'**, un formulaire comportant plusieurs champs s'affiche pour l'entrée des données. L'application gère le contrôle de saisie ainsi que la sauvegarde dans les bases des données.

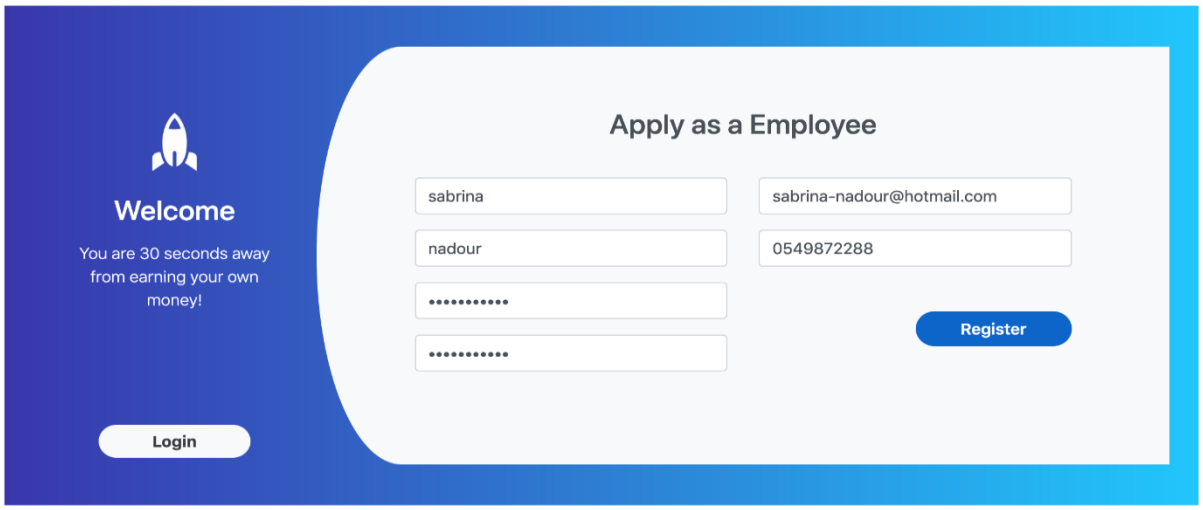


Figure 40 : *Page d'inscription*

V.8.4. **Page outils**

Cette page permet à l'administrateur d'accéder aux différentes fonctionnalités de l'application " Analytex " qui sont principalement tous les outils annoncés dans le chapitre quatre et un espace qui affiche tout l'historique de l'utilisateur trie par date et par type d'analyse

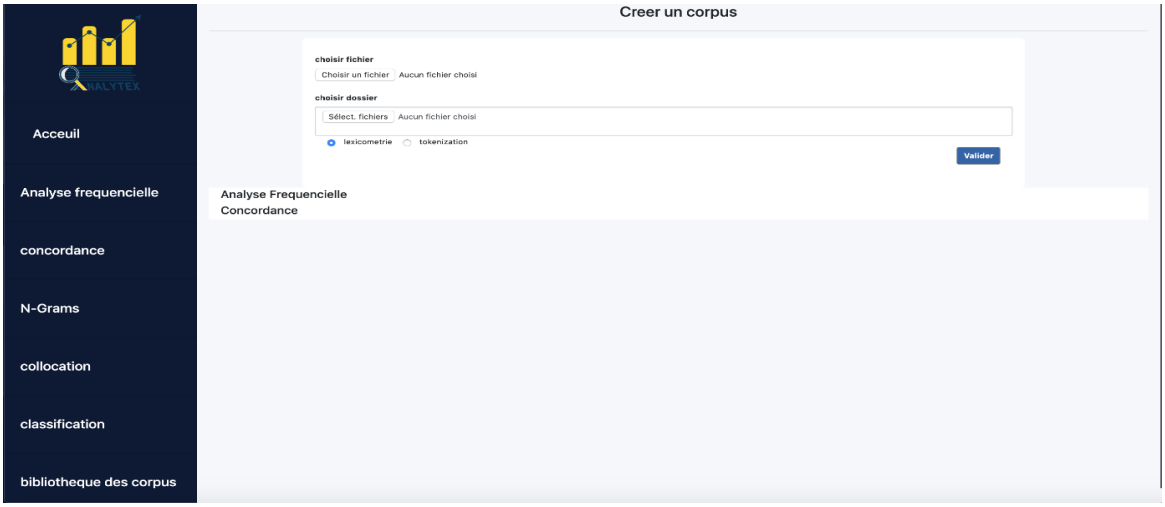


Figure 41 : *Page d'outils statistiques*

V.8.5. **Page de concordance**

Cette page fournit deux affichages de concordance qui permet de visualiser toutes les occurrences d'une forme ou d'un type généralisé en contexte, avec un paramétrage de nombre de contexte avant et après. Nous avons implémenté trois types de concordanciers (voir chapitre4)

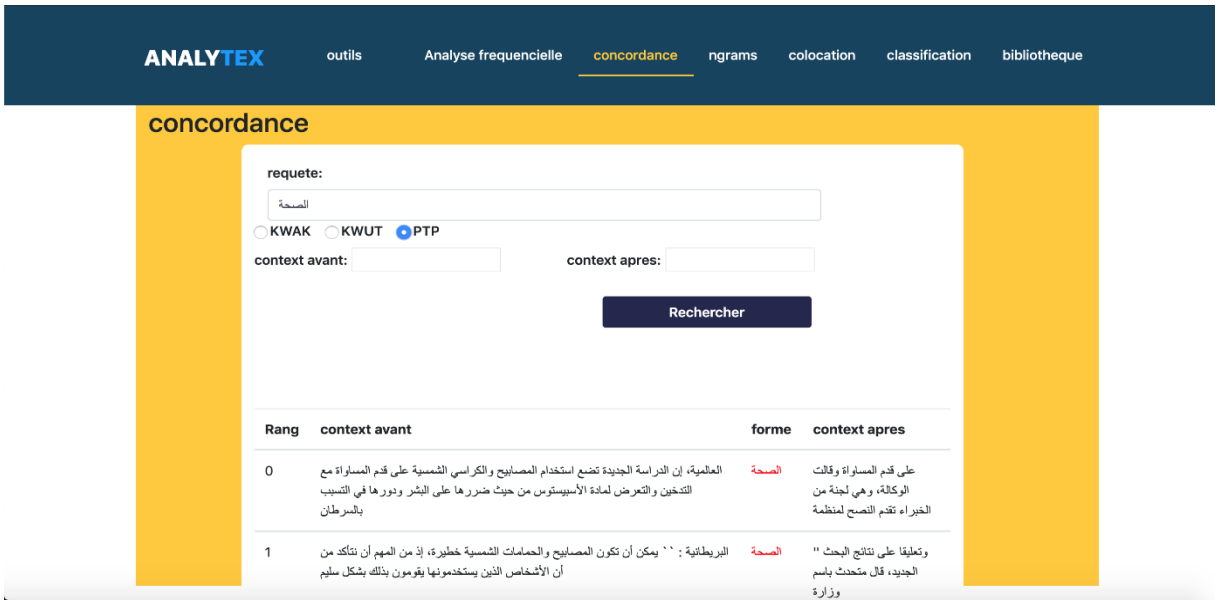


Figure 42 : *Page de concordance*

V.8.6. Page d'analyse fréquentielle

Cette page contienne Cinq ongle qui affichent tous les résultats d'analyse statistique sur le corpus téléchargé :

- **Onglet Global**

Cette page propose des statistiques sur les corpus textuels : effectif de toutes les formes, effectif des formes actives et supplémentaires, liste des hapax…etc. elle offre aussi une recherche de groupe de forme qui est gérée par plusieurs paramètres (commence par, se termine par ….)

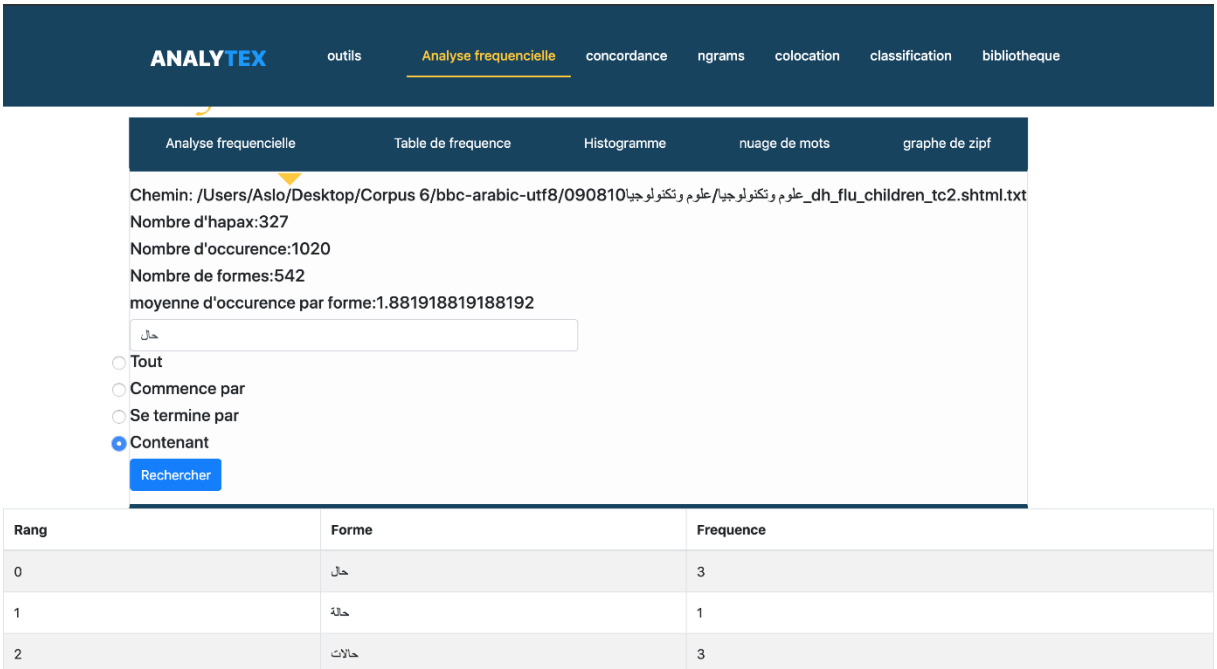


Figure 43 : *Page d'analyse fréquentielle*

- **Onglet tables de fréquences :**

Représente trois tableaux qui Listent les formes du corpus (colonne forme) avec leurs effectifs (colonne nb) et leurs rangs (colonne rang) :

- Tableau de forme active
- Tableau de forme supplémentaire

- Tableau total (Liste de toutes les formes actives et supplémentaires)

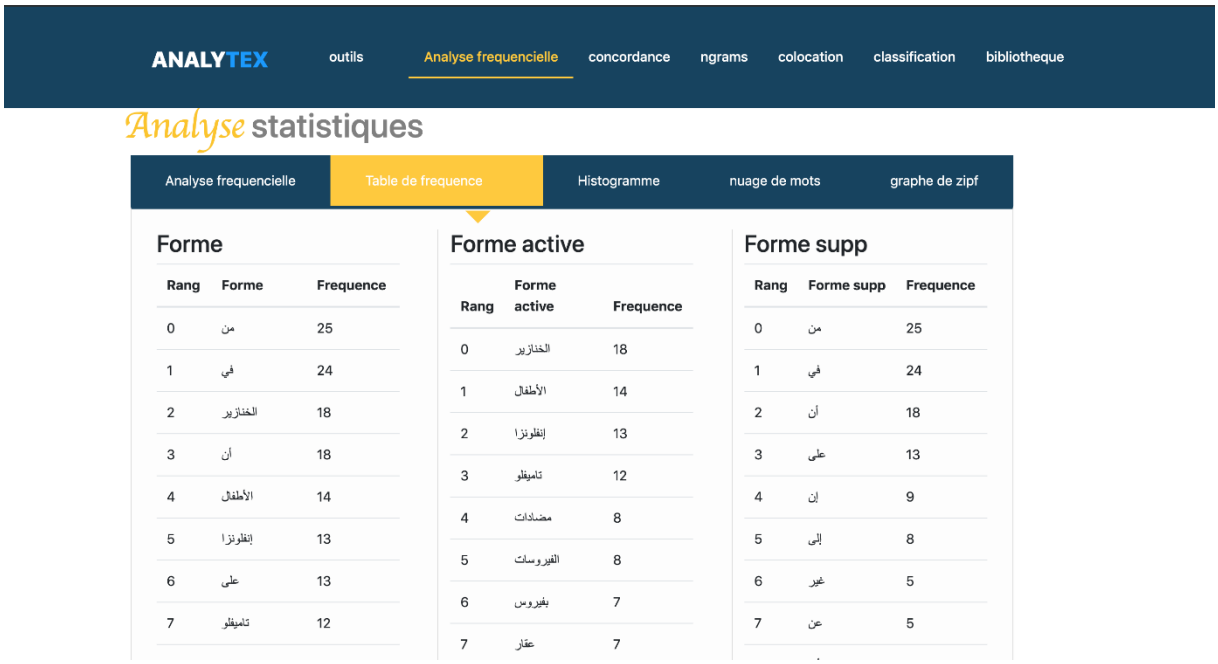


Figure 44 : *ongle tables de fréquences*

- **Onglet nuage de mot**

Le nuage de mots suivant (Word Cloud) est un outil de lexicométrie qui ne représente que la structure significative du corpus. Toutes les occurrences de « hapax » ont été éliminées de manière à ne retenir que les mots fréquentes. Plus la fréquence du mot dans le corpus est élevée, plus la taille du mot est importante, après un test sur notre jeux de donnée avec la plateforme 'Analytex' le mot 'الخنازر' a la plus haute fréquence (voir figure 44).

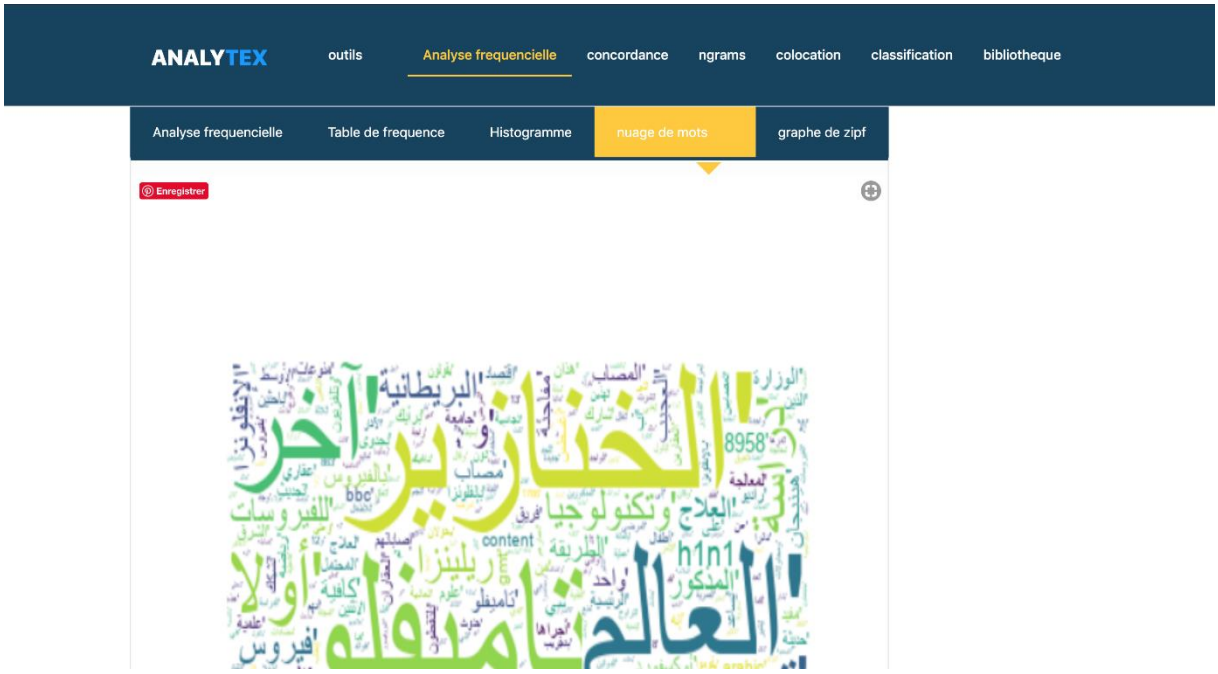


Figure 45: *Nuage de mots réalisé par Analytex*

Ce graphique permet donc de percevoir directement la diversité du contenu du corpus et les mots les plus significatifs (permet au préalable d'observer les formes lexicales pleines les plus utilisées du corpus).

- **Onglet histogramme**

Comme on a utilisé plusieurs graphes pour chaque outils, on a choisi l'histogramme aussi pour représenter le tableau de fréquence (le résultat du l'analyse fréquentielle).

Cet histogramme s'interprète de la façon suivante: La hauteur des bâtons est déterminée par la valeur de la fréquence relative affecté à chaque terme, après l'analyse fréquentielle. Plus les bâtons sont hauts, plus les fréquences sont élevés, plus l'importance du mot est forte.

- **Onglet graphe de zipf**

La loi de Zipf décrit la relation entre la distribution de fréquence d'un mot et son rang dans un corpus qui représente une langue. Il est une loi empirique basée sur une observation qui déclare que la distribution de fréquence de tout mot dans un corpus est inversement proportionnelle à son rang

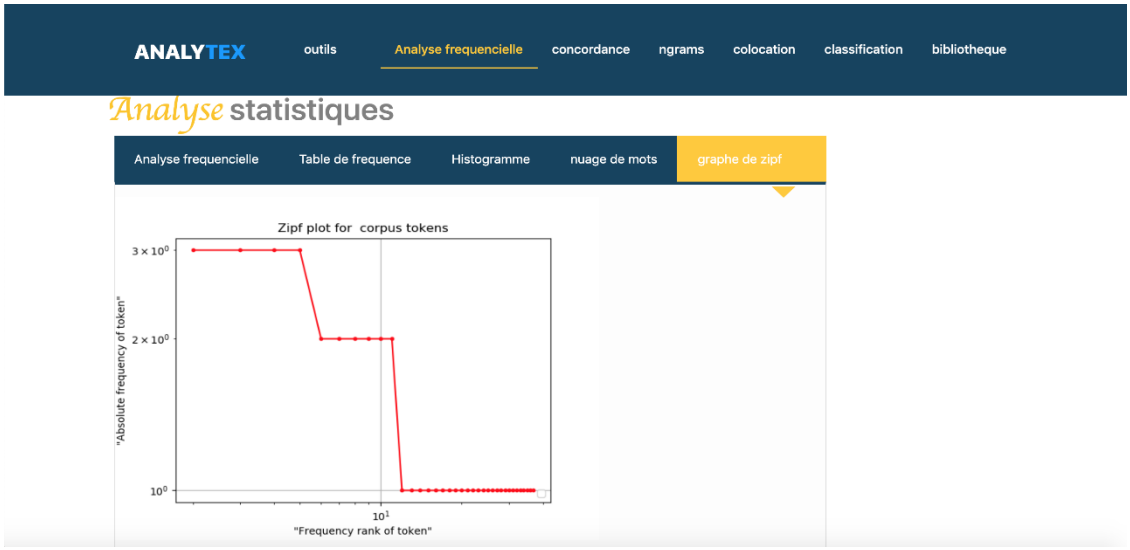


Figure 46 : *Graphe de zipf réalisé par Analytex*

Le graphique présente en abscisse les logarithmes des rangs et en ordonnées les logarithmes des fréquences des formes .Les valeurs indiquées sont par contre exprimés dans leur unité de départ. Les mots les plus fréquents sont ceux qui ont le rang le plus élevé. Par exemple, le mot le plus fréquent d'un corpus aura le rang 1.L'échelle logarithmique permet de réduire l'échelle pour les fréquences et rangs élevés.

V.8.7. **Page de N-Grams**

Cette page permet de listes de fréquences N-Grammes(les expressions multi-mots). L'utilisateur dispose d'un choix d'options de filtrage, pour spécifier en détail les n-grammes dont la fréquence doit être générée.

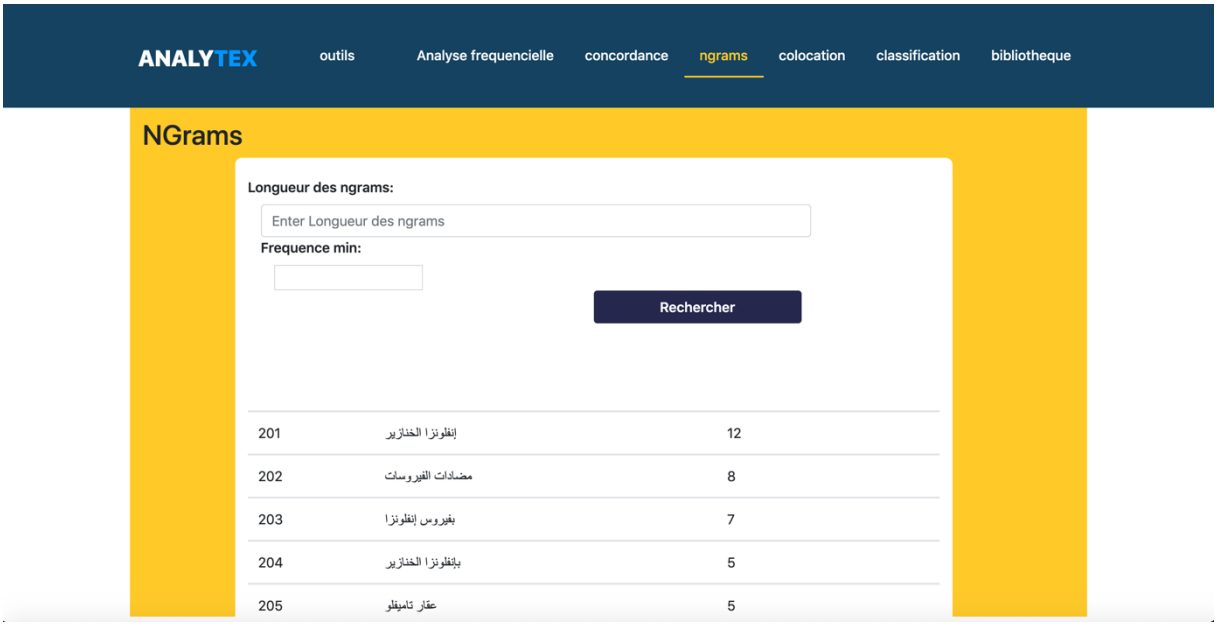


Figure 47 : *Page d'extraction des N_grams*

V.8.8. **Page de classification**

Cette page offre à l'analyste une classification automatique ascendante de documents de corpus, elle donne comme résultat un dendrogramme, Il se présente souvent comme un arbre binaire dont les feuilles sont des textes alignés sur l'axe des abscisses. Lorsque deux classes ou deux textes se rejoignent avec l'indice d'agrégation T, des traits verticaux sont dessinés de l'abscisse des deux classes jusqu'à l'ordonnée T, puis ils sont reliés par un segment horizontal.

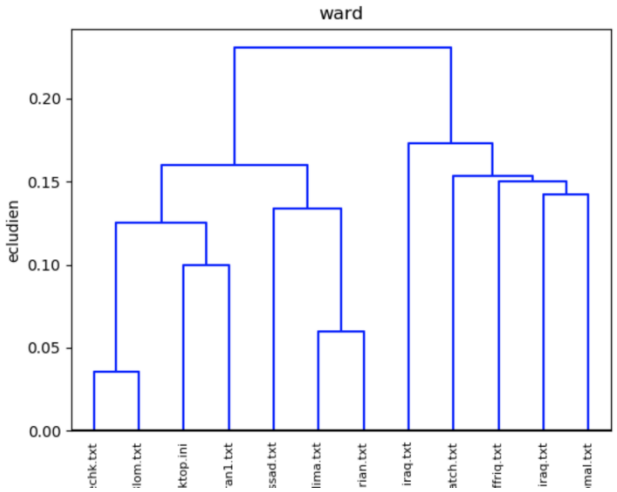
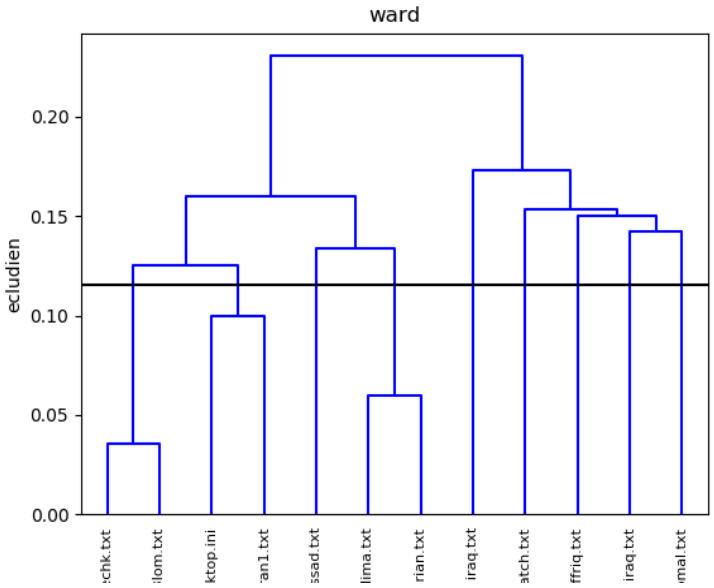


Figure 48: *Le dendrogramme due d'une classification hiérarchique ascendante réaliser par Analytex*

Lorsque l'utilisateur s'appuie sur le bouton **'couper le dendrogramme'**, il doit choisi un paramètre, ce dernier est un pourcentage pour couper le dendrogramme a fin d'afficher le dendrogramme couper et les classes obtenues avec des informations de chaque classe (mots clés, nom des documents, nom de classe)



num classe	mot cle	doc de classe
1	الى القانون مجلس	['sabrian.txt']
2	الصومال ان الحكومة	['ee.txt']
3	العالم كاس جنوب	['sportsouthaffriq.txt']
4	ان الخنازير الى	['teckno3lom.txt']
5	الى ان القانون	['3oloum clima.txt', 'coatch.txt']
6	ان الإيرانية الى	['iran1.txt', 'techk.txt']
7	المنة عقود بنسبة	['i9tissad huil de iraq.txt', 'japon i9tissad.txt']
8	الصومال ان الحكومة	['somal.txt']

Figure 49: *le dendrogramme et son résultat après le couper*

V.8.9. **Page bibliothèque de corpus**

Cette page fournit à l'utilisateur un moteur de recherche, qui permet de rechercher des corpus ou des requêtes dans la base de documents de notre plateforme et la possibilité de l'analyser et consulter le corpus choisi. L'utilisateur doit choisi un des modes de filtrage de documents

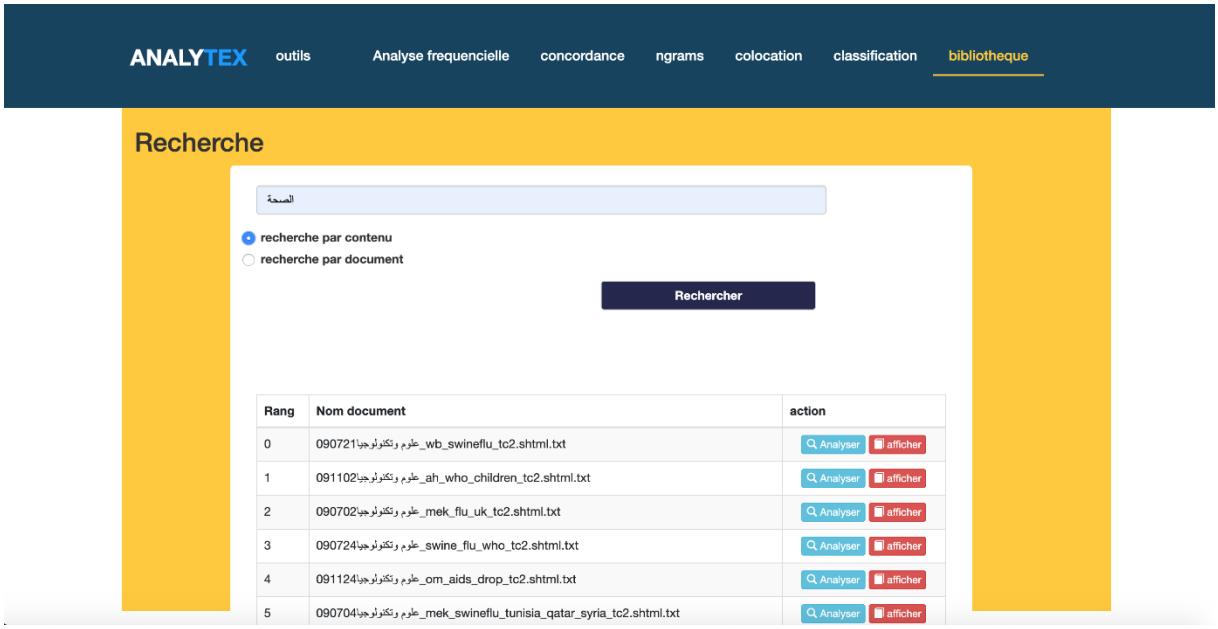


Figure 50: *la page bibliothèque de corpus*

V.9. Conclusion :

Ce dernier chapitre de notre projet décrit la phase de réalisation. Nous y avons précisé en particulier les outils de développement utilisés pour l'application "Analytext". Puis nous avons présenté les interfaces les plus importantes dans notre application

Dans ce chapitre, nous avons décrit l'aspect pratique de notre projet. Tout d'abord, nous avons listé l'ensemble des moyens technologiques utilisés (matériels et logiciels). Puis, nous avons présenté les différentes interfaces de notre application ainsi que leurs comportements. A travers cette réalisation, nous avons pu atteindre les objectifs fixés lors de la phase d'analyse des besoins nous avons pu découvrir plusieurs outils informatiques. Lors de développement, nous avons essayé de fournir un ensemble d'interfaces intuitives et simples à utiliser.

Conclusion général

Notre projet consiste à concevoir et mettre en œuvre une plateforme de traitement statistique et automatique de corpus arabe. Dans ce travail, nous avons présenté la plupart des outils d'analyse statistique des données textuelle et l'approche et la méthode suivie pour leur réalisation et conception. Comme nous venons de le voir, les outils statistiques sont fondamentales pour la recherche dans plusieurs domaine de recherche, tel que la détection de la langue par l'analyse fréquentielle, l'extraction des sens des mots à partir de leurs contexte par la concordance et la détection des idées générales des textes,…

Un recensement des plateformes TALA et d'analyse statistique nous as permis de faire une étude évaluative et comparative sur la base d'un nombre de critères (fonctionnalités offertes, volume de données traitées, possibilité d'intégrer de nouveaux corpus, types de corpus traités, utilisation en ligne ou hors ligne,…).

Nous avons aussi ressenti l'importance d'une plateforme de développement dédiée au traitement statistique des corpus arabes standardisant l'ensemble des traitements et garantissant une meilleure flexibilité. De notre point de vue, les outils TAL et les outils statistiques fournissent des résultats pertinents, les premiers en termes de qualité et les deuxièmes en termes de volume et du temps du traitement, donc la combinaison des deux peut améliorer les résultats fournis pour l'analyse automatique des textes comparé au cas où seuls les outils statistiques sont utilisés.

Au final étant donné que nul ne peut se prétendre aborder un domaine dans son ensemble nous envisageons de:

- Améliorer notre plateforme en ajoutant des outils qui facilitent d'avantage l'analyse des corpus textuels, en ajoutant l'analyse de correspondance, la classification des mots et l'analyse factorielle, l'analyse des spécificités
- Héberger l'application sur un serveur afin d'élargir le corpus de **Dhakhira al-Arabiyya**
- Intégrer d'autres format d'importation des donnes textuelles (CSV ,HTML , .DOCX , .DOC).

- Poursuivre le développement d'outils de collocation en enregistrant un score pour d'améliorer nos résultats

Annexe :

- **Hapax :** fait référence à l'apparition unique d'une occurrence d'un terme à l'intérieur d'un corpus.
- **Une occurrence** : est suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs.
- **La forme graphique :** c'est l'unité minimale de travail et de comptage dans le corpus est la plus brute et la plus simple possible
- **U.C.I (Unité de Contexte Initial) :** correspond à l'ensemble textuel que l'on décide de repérer à l'intérieur d'un corpus à l'aide d'indicateurs étoilés. L'U.C.I constitue la plus grande unité statistique. Elle peut donc être un segment textuel du corpus ou même représenter le corpus dans sa totalité.
- **Caractère :** signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.
- **Le lexique** : est formé de toutes les occurrences possibles de toutes les formes articulées que constituent l'ensemble des occurrences (ou formes lexicales) potentielles ou réalisées d'une langue.
- **Fréquence (d'une unité textuelle) :** le nombre de ses occurrences dans le corpus.
- **Segment :** toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur de séquence
- **Caractères délimiteurs / non-délimiteurs :** distinction opérée sur l'ensemble des caractères qui entrent dans la composition du texte, permettant aux procédures informatisées de segmenter le texte en occurrences
- **Les formes supplémentaires :** présentent des mots de liaison (الى, و, في ...etc.) ou des déterminants (ال...etc.)

REFERENCES BIBLIOGRAPHIQUE

[1] A. Djebbar, «les cahiers de l'Islam,» 30 septembre 2012. [En ligne]. Available: https://www.lescahiersdelislam.fr/Histoire-et-evolution-de-la-langue-arabe_a137.html. [Accès le avril 2019].

[2] F. DARDOUR, *LANGUE ENSEIGNEE ET DIALECTE ARABE: QUELLE METHODOLOGIE ET QUELLE FORMATION POUR L'ACQUISITION DE LA COMPETENCE COMMUNICATIVE EN ARABE STANDARD ?,* UNIVERSITE NANCY 2, 15 décembre 2008.

[3] D. Bouneffouf, *Etat de l'Art sur la Reconnaissance de l'Inférence Textuelle,* Télécom SudParis: Department of Computer Science, 14 Janvier 2015.

[4] I.Tellier, Introduction au TALN et `a l'ingenierie linguistique, Univer-sité de Lille, 2010.

[5] Z.Danijel, *UNE APPROCHE PRAGMATIQUE DE L'ANALYSE DU DISCOURS ET SON APPLICATION À LA DIDACTIQUE DU FRANÇAIS SUR OBJECTIF SPECIFIQUE* , Belgrade, Serbia: Union University, Faculty of Law and Business study, 2017.

[6] A.ROMAN, Grammaire de l'arabe, Paris: Presse universitaires de France, 1990, p. 10.

[7] S.Baloul, *thèse de doctorat en informatique : Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé,* Université du Maine, France, soutenue le 27 mai 2003.

[8] M. Maraoui, M.Zrigui, G.Antoniadis, *" Un système de génération automatique de dictionnaires étiquetés de l'arabe",* Rabat, Maroc., 2007.

[9] Douzidia, Fouad Soufiane, *Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique:Résumé automatique de texte arabe,* Université de Montréal ,Canada, Septembre 2004.

[10] R. Abbes, *La conception et la réalisation d'un concordancier électronique pour l'arabe,* l'institut national des sciences appliquées de Lyon, Lyon., 2004.

[11] R.Semin, *au Museum national d'histoire naturelle,* SeminR : Journée R, 2013 .

[12] I.Miftah,N. Ataa Allah et F.Taghbalout , «Corpus multilingue pour les langues un peu moins Rabat, Maroc, LRIT, Faculté des sciences, Université Mohammed V,» 2016. [En ligne]. Available: https://www.researchgate.net/publication/311377456. [Accès le 15 mai 2019].

[13] J. Debret, 5 mars 2018. [En ligne]. Available: https://www.scribbr.fr/memoire/recherche-qualitative-ou-quantitative-quelles-differences/. [Accès le juin 2019].

[14] D. Mayaffre, *L'analyse du discours assistée par ordinateur,* 2009.

[15] P. Catherine, *GUIDE D'UTILISATION DE LEXICO3 Application à la langue arabe,* 2010.

[16] B. FALLERY, *La lemmatisation c'est la fabrication d'une forme réduite du texte, standardisée par des dictionnaires.,* Montréal, Juin 2007.

[17] Z. MOUELHI, *Essai de lexicométrie d'une œuvre arabe,* Lyon , 22 novembre 2008.

[18] C. SOUTI, *La lexicométrie ou l'analyse du discours assistée par ordinateur : ce que l'informatique et les mathématiques peuvent apporter à la littérature et la linguistique,* Laboratoire Slaad Université Constantine 1, Juin 2015.

[19] J.Radwan,C. Jean-Hugues , *Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques,* Laboratoire ERIC – 5, Pierre Mendès-France, 2002.

[20] B.Pincemin, Concordanciers : thème et variations, Besançon : Presses universitaires de Franche-Comté: Viprey (éd.), 8èmes JADT 06, 2006.

[21] D. Mayaffre, *L'entrelacement lexical des textes. Cooccurrences et lexicométrie,* France: Journées de Linguistique de Corpus Lorient, 2007.

[22] W.Martinez, M.Zimina., *Première approche textométrique de l'analyse contrastive du corpus trilingue anglais-français-allemand GUIDES,* 2007.

[23] D. Salles, analyse factorielle des correspondances simples sosiologie, 2009, p. 129 à 239.

[24] D.DEMAZIERE , C. BROSSAUD, P. TRABAL et K.M. METER, *Analyses textuelles en sociologie Logiciels, méthodes, usages,* Presses Universitaires de Rennes, 2006.

[25] K. Boudreau, *"Opening the Platform vs. Opening the ComplementaryGood? The Effect on Product Innovation in HandheldComputing.",* 2008.

[26] [En ligne]. Available: https://www.andy-roberts.net/coding/aconcorde . [Accès le 15 mai 2019].

[27] R.Andrw, L. Al-Sulaiti et E. Atwell, *aConCorde: Towards an open-source, extendable concordancer for Arabic Corpora,* 2006.

[28] [En ligne]. Available: http://www.lexi-co.com/. [Accès le avril 2019].

[29] [En ligne]. Available: http://www.iramuteq.org. [Accès le avril 2019].

[30] [En ligne]. Available:] http://corpus.leeds.ac.uk/itweb/htdocs/Query.html. [Accès le mai 2019].

[31] [En ligne]. Available: https://auth.sketchengine.eu/#login?next=https%3A%2F%2Fapp.sketchengine.eu%2F%23dashboard. [Accès le 15 mai 2019].

[32] A.Kilgarriff,M. Jakubíček,J. Pomikálek,T. Berber Sardinha, et P.Whitelock, *The Sketch Engine: Ten Years On,* 2014.

[33] [En ligne]. Available: https://sourceforge.net/projects/kacst-acptool/. [Accès le 22 mai 2019].

[34] L .LEBART, A. SALEM, « Statistique textuelle »,Préface de Christian Baudelot, Paris, 1994.

[35] Dagenais, V. Ridde et Christian, Approches et pratiques en évaluation de programme, Montreal, 2013.

[36] L.Gulen, *A study of neural networks for natural language processing,* Genève, 2018.

[37] C. Maklin, «TF IDF | Exemple Python TFIDF,» 5 mai 2019. [En ligne]. Available: https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76. [Accès le juillet 2019].

[38] Mourad, Raphaël, C. Sinoquet, and P.Leray., *"Apprentissage de réseaux bayésiens hiérarchiques latents pour les études d'association pangénomiques,* Nantes: Proc. JFRB 2010, 5th French-speaking meeting on Bayesian networks, 2010.

[39] E. Negre, *Comparaison de textes: quelques approches...,* paris, 2013.

[40] E.Selab, and A. Guessoum, *Building TALAA, a Free General and Categorized Arabic Corpus,* alger: Université des Sciences et de la Technologie Houari Boumediene (USTHB), 2016.

[41] T.Lecroq et S. Faro, «Recherche de motif,» 12 Décembre 2010. [En ligne]. Available: https://www.irif.fr/~carton/Enseignement/Algorithmique/Programmation/Pattern/MorrisPratt/.