République Algérienne Démocratique et Populaire
Ministère De L'enseignement Supérieur Et De La Recherche Scientifique
Université Amar Telidji Laghouat


Faculté des Sciences
Département de Mathématique et Informatique

Mémoire en vue de l'obtention du diplôme de Magister en informatique

Thème :

# Etude statistique des documents Web en langue Arabe

Présenté par :
Abdallah Lakhdari

soutenu publiquement devant le jury composé de :

| | | | | |
|---|---|---|---|---|
| M | M. B. Yagoubi | Président | Université de Laghouat, Algérie | Professeur |
| M | D. Ziadi | Examinateur | Université de Normandie, France | Professeur |
| M | Y. Ouinten | Examinateur | Université de Laghouat, Algérie | M$^{tre}$ de conférences |
| M$^{me}$ | H. Cherroun | Rapporteur | Université de Laghouat, Algérie | M$^{tre}$ de conférences |
| M | A. Nehar | Co-Rapporteur | Université de Djelfa, Algérie | M$^{tre}$ assistant |

**N d'ordre :..../2013 - M / DGI**

## ملخص

تعتبر المعالجة الآلية للغات الطبيعية مجال بحث متعدد التخصصات فهي تشمل الاعلام الآلي واللغويات، كما أن لها تطبيقات عديدة في الحياة اليومية ذات أهمية متزايدة باستمرار.

إن التسلسل الهرمي للمعالجة الآلية للغات الطبيعية يتكون أساسا و على الترتيب من التحليل الصرفي، النحوي ثم الدلالي.

لقد اهتممنا في هذا العمل، بالتحليل الصرفي. لأنه يمثل خطوة أساسية تعنى بدراسة بنية الكلمات في اللغة. من أجل ذلك، قد وظفنا تقنيات تعلم الآلة والاحصائيات لتطوير نموذج إحصائي لبنية الكلمات العربية. تم تدريب هذا النموذج الإحصائي باستخدام الذخيرة اللغوية اوزاك، وهو عبارة عن مجموعة تتألف من الآلاف من الوثائق الالكترونية باللغة العربية على شبكة الإنترنت التي تحوي أزيد من ١٨ مليون كلمة.

لقد قمنا أيضا بتصميم محلل صرفي لاستخراج الجذور الثلاثية يعمل بدون اشراف، و بينت النتائج أن هذا المحلل يحقق دقة تصل الى ٦٥ ٪، والتي تعتبر أفضل من تلك التي حصل عليها المحللون الصرفيون من نفس الفئة. وكانت النتائج حتى منافسة مع تلك المحللات الخاضعة لاشراف و التي تستعمل معارف لغوية مع العلم أن هاته الأخيرة مكلفة لما تحتاج اليه من معارف لغوية.



كلمات البحث :المعالجة الآلية للغات الطبيعية، اللسانيات الحاسوبية، الويب، التحليل الصرفي، العربية، تعلم الآلة، الاحصاءات، بدون إشراف، جذر الكلمة.

**Abstract**

*The Natural Language Processing* -NLP- is a multidisciplinary research area involving computer science and linguistics. Its applications are constantly increasing and their importance is gaining momentum.

The hierarchical processing in the NLP process consists mainly and respectively of morphological, syntactic and semantic analysis.

In this work, we are interested in "*morphological analysis*" as a crucial step which is interested in the study of the structure of words in the language. Indeed, we have instrumented the unsupervised machine learning and statistics to develop a statistical model to capture the regularities of the structure of words in Arabic. The statistical model was trained using the OSAC corpus ; a corpus composed of thousands of Web documents in Arabic containing more than 18 million words.

We have also developed an unsupervised morphological analyzer for extraction of tri-literal root. The results show that our analyzer achieves an accuracy of 65 %, which is better than those obtained by the analyzers of the same class. And the results were even competitive with those supervised analyzers demanding many expensive linguistic knowledge.

**Keywords :** NLP, computational linguistics, Web, morphological analysis, Arabic, machine learning, statistics, unsupervised, word, root.

**Résumé**

*Le Traitement Automatique des Langues Naturelles* -TALN-, est un domaine de recherche multidisciplinaire regroupant l'informatique et la linguistique. Ses applications ne cessent de s'accroitre et leurs importances dans notre vie de tous les jours prennent de l'ampleur.

Le traitement hiérarchique dans le processus TALN se compose principalement d'analyse morphologique, syntaxique et sémantique.

Dans ce travail, nous nous sommes intéressés au niveau "*analyse morphologique*". Vu que c'est une étape cruciale qui s'intéresse à l'étude de la structure des mots de la langue. En effet, nous avons instrumenté les solutions d'apprentissage automatique non supervisé ainsi qu'aux statistiques pour développer un modèle statistique pour capter les régularités de la structure des mots en langue Arabe. Le modèle statistique a été entrainé à l'aide du corpus OSAC ; un corpus composé de milliers de documents Web en langue Arabe comportant plus de 18 millions mots.

Nous avons aussi conçu un analyseur morphologique non supervisé pour l'extraction de la racine tri-littérale. Les résultats obtenus montrent que notre analyseur atteint une précision de 65%. Ils sont meilleurs que ceux obtenus par les analyseurs de même catégorie. Ces résultats sont même compétitive avec ceux des analyseurs supervisés exigeants beaucoup de connaissance linguistique coûteuses.

**Mots clés :** TALN, Linguistique informatique, Web, Analyse morphologique, Langue Arabe, Apprentissage automatique, Statistique, non supervisée, Mot, Racine.

**Dédicaces**

A mes parents,

A tous ceux que j'aime,

A tous ceux qui m'aiment,

A . . . . . ,

je dédie ce modeste travail.

**Remerciemments**

En préambule à ce mémoire, je remercie Dieu tout Puissant de m'avoir permis de mener à terme ce travail.

Je remercie chaleureusement mes parents qui m'ont beaucoup aidé matériellement et moralement, depuis mon enfance jusqu'à ce que je suis arrivé à ce point là.

Je tiens à remercier sincèrement Mme. Cherroun Hadda ; ma directrice de thèse, pour l'encadrement et pour m'avoir encouragé et guidé par son sens d'accueil, et ses multiples conseils, en dépit de ses occupations, pour la disponibilité, la patience, la confiance, les conseils qu'elle m'a prodigué, et pour tout le temps et l'énergie qu'elle a consacré à la réalisation de ce travail.

Je remercie aussi messieurs Ziadi Djelloul et Nehar Attia qui m'ont familiarisé avec ce domaine de recherche passionnant et je leur ai espéré de bonne continuation et plus de conciliation dans leurs carrières de recherche et dans ce domaine en particulier.

Mes remerciements vont en outre à l'adresse des membres du jury qui ont bien voulu nous honorer de leur présence :

- M M. B. YAGOUBI, professeur à l'université de Laghouat,
- M D.Ziadi, professeur à l'université de Normandie, Rouen France,
- M Y.OUINTEN, maître de conférences à l'université de Laghouat.

Merci à tous ceux qui ont participé de près ou de loin à l'aboutissement de ce travail.

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction générale

## Contexte

L'informatique n'a jamais cessé d'envahir les autres disciplines. Elle les séduit par les potentielles de calcul et de traitement dont elle dispose. Comme elle veut en elle-même se rapprocher de l'Homme : le dernier bénéficiaire de ses services. Une preuve de cette raison est la vulgarisation des technologies de l'information et de la communication TIC qui vise toujours la simplification des outils et des moyens d'interaction entre l'Homme et l'information (recherche, traitement, communication etc.). Le support naturel de cette information est la langue. Donc le fait d'amener la machine à la maitriser par un traitement efficace qui permet d'avoir une bonne compréhension et une bonne génération est l'objet d'une filiale de l'informatique, issue de l'intelligence artificielle connues sous le nom « ***Traitement Automatique de la Langue Naturelle -TALN-*** » ou récemment sous un autre nom « ***la linguistique informatique*** ». Ce dernier paradigme indique que la nouvelle discipline est une alliance entre l'informatique et la linguistique, présentant un domaine de recherche assez mature au point qu'il couvre des milliers de travaux, éparpillés en plusieurs sujets selon leurs applications ou les niveaux de traitement linguistique qui sont standardisés et ordonnés par la hiérarchie suivante : morphologie, syntaxe, sémantique et pragmatique.

## Problématique et objectifs

Sans avoir une bonne maîtrise des bas niveaux (morphologie et syntaxe), on n'arrive jamais à une bonne performance de la chaîne de traitement linguistique. C'est la raison qui nous a motivé à faire ce travail, car ces bas niveaux -la morphologie essentiellement- permettent à extraire et reconnaitre les constituants de base d'une langue naturelle (lettre, mot et phrase) par la réponse aux questions suivantes :

- Comment les détecter ?
- Comment elles sont formées ?
- Comment faire pour les analyser (segmenter et étiqueter les entités élémentaires) ?

Deux voies principales ont été suivies ; des analyseurs morphologiques supervisés qui utilisent d'énormes bases de connaissances linguistiques telles que les dictionnaires, les lexiques et les règles grammaticales qui sont généralement construites à la main, cette

tâche coûteuse en terme de temps pour l'inférence et pour les ressources requises. La deuxième voie vise à minimiser l'usage des connaissances linguistiques dont on distingue deux approches : des combinatoires basées sur des ressources assez limitées et des approches non supervisées qui s'appuient seulement sur l'apprentissage automatique. Ces dernières représentent la tendance la plus récente et la plus efficace. Cependant, peu de travaux y sont investis. Plus spécialement pour les langues sémitiques et en l'Arabe en particulier, contexte de notre étude.

Dans ce travail, nous visons à élaborer un modèle statistique pour la morphologie de la langue naturelle, en prenant l'Arabe comme une langage cible [LH13]. C'est une langue qui a une morphologie extrêmement complexe. Le recours à l'apprentissage automatique et aux statistiques est justifié par plusieurs raisons ; profitant le plus possibles des atouts de ces outils qui ont prouvé leurs performances dans plusieurs domaines. Ainsi que nous voulons être indépendants de toutes connaissances linguistiques.

Dans une deuxième contribution, et via cette modélisation statistique qui peut pondérer des analyseurs morphologiques déjà existants dans le cas d'un indéterminisme, nous proposons une méthode de segmentation des mots non supervisée qui peut atteindre la racine tri-littérale d'un mot brut sans aucune utilisation de connaissance linguistique.

## Structure du document

Dans ce document, nous avons voulu suivre une organisation naturelle, dont la structure est récapitulée comme suit :

**Chapitre 1** Contextualisation du problème de l'analyse morphologique du langage naturel.

**Chapitre 2** Nous présentons une description des concepts qui guident le lecteur à comprendre les travaux investis pour résoudre ce problème, on n'a mis l'accent que sur les repères importants dans le développement de TALN. Après cet acheminement, nous avons focalisé notre intérêt sur la langue Arabe. On a commencé par décrire ses spécificités et les challenges qu'elle impose vis-à-vis du problème sujet de l'étude. Ensuite, on a cité quelques travaux traitant de l'analyse morphologique pour la langue Arabe.

**Chapitre 3** Ce chapitre est dédié à notre contribution : la construction du modèle et la méthode de segmentation non supervisée.

**Chapitre 4** Dans le dernier chapitre, nous avons évalué notre conception par une expérimentation dans un benchmark que nous avons conçu et une comparaison avec d'autres algorithmes.

# Chapitre 1

# Généralités

Dans ce chapitre nous allons décrire brièvement le domaine du *TALN* dans le but de répondre aux questions suivantes : Qu'est ce que c'est le *TALN* ? qu'est ce qu'on cherche à toucher ? Commençant par une démonstration de la polyvalence de l'informatique et ses apports pluridisciplinaires afin d'introduire ce nouveau domaine. Eplucher ensuite dans l'histoire pour acquérir une idée de son évolution dans le temps jusqu'à nos jours, où on verra les sujets d'actualité couverts par ce paradigme. Finalement, nous verrons les problèmes encore ouverts pour localiser l'impact de notre travail.

## 1.1 Traitement automatique des langages naturels

Le progrès de l'informatique et sa vulgarisation ont laissé plusieurs disciplines faire appel à des potentiels de traitement de l'information pour approcher leurs dilemmes, ceci d'une part. D'autre part l'informatique quant à elle, doit achever ses propres fins, dont elle ne s'évalue que par la satisfaction du dernier utilisateur qui veut toujours avoir ses résultats de la façon la plus simple possible par des requêtes aussi simplifiées. Un exemple de cette synergie interdisciplinaire est entre la linguistique et l'informatique, dans le but de maitriser la langue naturelle, qu'on a pris comme motivation, dont il n'y a pas de simple à produire et de comprendre par les humains que le langage naturel. Le formalisme, la puissance de calcul et l'interactivité chez les informaticiens et la rigueur des linguistes identifient ensemble un nouvel hybride connu sous le nom de Traitement Automatique de la Langue Naturelle Abr. TALN Anglais : Natural Language Processing Abr NLP.

Evidemment les disciplines impliquées sont l'informatique, par ses avancées surtout dans les domaines de l'intelligence artificielle et l'apprentissage automatique, et bien sur la linguistique, dont chacune d'elles est motivée par les promesses de l'autre.

### 1.1.1 La linguistique

La linguistique exploite les méthodes et les procédés offerts par l'informatique avec une éventuelle utilisation des connaissances préétablies afin de découvrir de nouvelles approches sophistiquées pour apprécier leurs phénomènes. On peut citer par exemple :

#### 1.1.1.1 La lexicographie

Tous simplement, elle s'occupe par l'élaboration des dictionnaires, une discipline purement appliquée, en même temps basée sur des fondements théoriques de la lexicologie qui est une science exclusivement théorique par contre la précédente, dont l'aspect d'applicabilité chez la première l'a laissé aller de pair avec l'informatique. Cette dualité est interprétée par un recours des linguistes qui ont défié le challenge disant que *Les très grands dictionnaires tentent de rejoindre le fonctionnement réel de la langue, mais c'est une course dans laquelle le lexicographe est d'avance battu* [Rey77].

#### 1.1.1.2 Les statistiques

Ce n'est pas récent d'avoir utilisé les statistiques lors de l'étude d'une langue naturelle, on trouve une bonne preuve de ça chez les anciens savants Arabes qui réclament des calculs et des recensements dans le but de capturer les caractéristiques de la langue Arabe [AmE69] afin de conserver sa nature et la protéger, surtout après l'émergence des étrangers dans leur nations, et leur influence sur la langue.
Récemment, une tentative pour examiner des constatations sur les propriétés des mots en Arabe qui reviennent à des époques très anciennes comme cité auparavant [SE16] qui n'étaient justifiées que par beaucoup utilisé **كثرة الاستعمال**, mais cette fréquence n'était jamais quantifiée ou normalisée. L'investigation a été faite grâce à l'ordinateur, dont on a pu compter avec une bonne précision les fréquences des occurrences des lettres Arabes pour un dictionnaire référence dans la langue Arabe. Ce travail est documenté dans un bouquin intitulé étude statistique des racines du dictionnaire **معجم الصحاح** en utilisant l'ordinateur [EM67].
Voici quelques exemples de ces constats :

- On ne peut pas trouver les deux lettres ***Djim*** **ج** et ***Qaf*** **ق** en un seul mot d'origine en Arabe, alors le mot ***Menjanik*** **منجنيق** est un terme non-Arabe.

- Même issue pour les deux lettres ***Sad*** **ص** et ***Djim*** **ج** ou ***Taa*** **ط** et ***Djim*** **ج** dont les mots ***Sawlajan*** , ***Djass*** et ***Taadjin*** **صولجان، جص و طاجن** sont étranges.

- Pas de ***Raa*** **ر** après ***Noun*** **ن**, ni de ***Zay*** **ز** après ***dal*** **د** dans un mot en Arabe, tels que ***Nardjes*** et ***Mohandez*** **نرجس و مهندز** l'origine du mot ***Mohandes*** **مهندس** qui veut dire ingénieur.

- On trouve aussi que le ***Zay*** **ز** ou le ***Dhal*** **ذ** ne se rassemblent pas en un seul mot avec ***Sin*** **س**, duquel le mot ***Sadhej*** **ساذج** est arabisé du perse.

- Tout mot dont la racine est à quatre lettres ou cinq contient toujours l'une des lettres suivantes ***Mime*** **م**, ***Lam*** **ل**, ***Raa*** **ر**, ***Noun*** **ن**. Sauf le mot ***Aasdjad*** **عسجد** qui veut dire l'or.

### 1.1.2 L'informatique

Dans plusieurs applications récentes, telles que la traduction automatique et la reconnaissance de la parole etc. le type des données à traiter est la langue naturelle (comprendre, générer ou les deux). Pour cela il faut développer de nouvelles méthodes et algorithmes pour maitriser ce type non structurée. L'analyse linguistique se divise en deux parties dont la première cherche à reconnaitre ce qui est parlé ou écrit et la deuxième cherche à le comprendre. La figure ci-dessous 1.1 montre les deux aspects : la reconnaissance du parlé ou manuscrit et l'autre aspect de la génération de la langue parlée ou écrite.

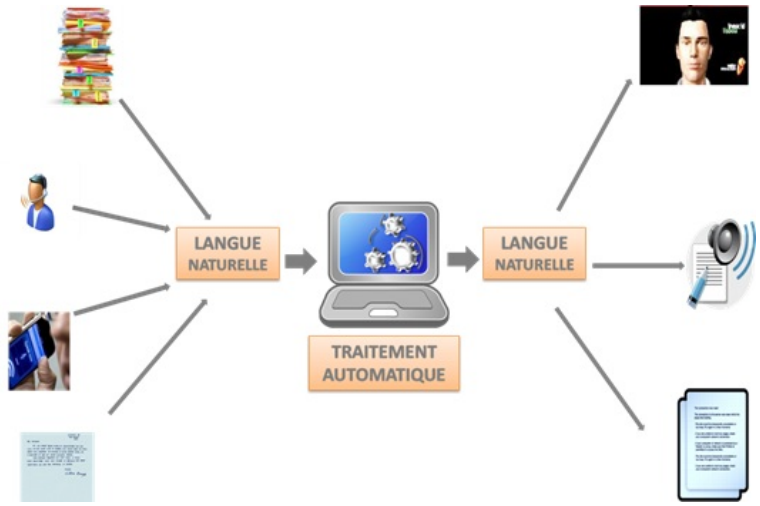


FIGURE 1.1 : Traitement automatique de la langue TALN

#### 1.1.2.1 Reconnaissance

Il se peut qu'elle soit une reconnaissance de la parole (ASR), du manuscrit (OCR) ou n'importe qu'elle forme brute de la langue. Cette opération s'intéresse à la conversion de cette forme brute en une forme lisible par la machine (voir Figure 1.2) pour qu'elle soit traitée et analysée afin de servir à des applications ultérieures à développer prochainement. Le ASR part du signal acoustique qui s'interprète à des phonèmes par un modèle acoustique, une fois on a ces briques de bases, le processus linguistique démarre. Même chose pour les OCR dont on démarre le processus après un traitement d'image dans le but de reconnaitre chaque caractère pour remonter aux mots et de plus en plus au processus.


FIGURE 1.2 : Processus de reconnaissance ASR

#### 1.1.2.2 Génération

C'est la production des textes compréhensibles dans une ou plusieurs langues à partir des données non linguistiques (voir Figure 1.3), en utilisant des connaissances sur la langue cible et le domaine d'application [Mit05]. Ici une définition référence *la génération automatique du langage naturel est le processus de construction d'un texte en langage naturel pour atteindre une communication cohérente* [[McD92], cette anecdote exprime la difficulté du domaine *la compréhension de la langue est comme le comptage de un à l'infini, sa génération est comme le comptage de l'infini à un* [App87].


FIGURE 1.3 : Processus de génération

**Exemple**
***Fonction*** : production de bulletins météorologiques en anglais et en français.
***Entrée*** : Description numérique de données climatiques
***Client*** : Environnent Canada (Canadian Weather Service)
***Statut*** : En production journalière depuis 1992

### 1.1.3 La linguistique informatique

#### 1.1.3.1 Définition

Le TALN a besoin des savoirs-faire suivants : [MT90]

- **La linguistique théorique** fournit des descriptions entièrement explicites organisées dans des théories cohérentes du savoir linguistique, dont les linguistes doivent répondre à ces deux questions suivantes :
  - Que disent les humains ?
  - Que cela demande/dit au monde ?
- **L'informatique théorique** fournit des algorithmes et programmes de traitement efficaces et leurs optimisations.
- **L'étude mathématique** permet d'étudier les propriétés formelles des outils de traitement et des théories linguistiques.
- **Les recherches en sciences cognitives et en intelligence artificielle** sur la représentation des connaissances sont également très importantes, notamment pour

les aspects sémantiques et textuels du TALN.
la figure ci-dessous 1.4 illustre la pluridisciplinarité du domaine TALN.


FIGURE 1.4 : TALN Synergie interdisciplinaire

#### 1.1.3.2 Niveaux de traitement

Par la remontée de plusieurs niveaux (voir Figure 1.5), la linguistique informatique cherche à comprendre l'information ou la connaissance voulu, en imitant le même processus réalisé par le cerveau humain.

1. Niveau morphologique (reconnaître) : Il consiste à décomposer le flux linguistique (parole ou texte) en mots en suite chaque mot est aussi décomposé à son tour en unités élémentaires appelées « Morphème » qui représente la plus petite unité formelle dotée d'une signification ; il est constitué d'un ou de plusieurs phonèmes indécomposables. La tâche n'est pas évidente ; les mots ne sont pas forcément séparés par des espaces.

2. Niveau syntaxique (structurer) : La syntaxe représente la rigueur des contraintes entre les catégories morphosyntaxiques à respecter lors de la description d'une séquences de mots jugée "acceptables" dans la langue. L'analyse syntaxique se fait par la grammaire qui est un ensemble de règles qui traitent la manière dont les mots peuvent se combiner pour former des phrases, ainsi que l'enchainement de ces dernières entre-elles.

3. Niveau sémantique (comprendre) : La sémantique vise à l'étude de sens hors contexte, elle prend comme unité élémentaire la phrase et cherche à mentionner sa signification. Ces phrases se composent d'un certain nombre de mots identifiés par l'analyse

morphologique, et regroupés en structures par l'analyse syntaxique. Ces mots et ces structures constituent autant d'indices pour le calcul du sens : on pourrait dire, que le sens résulte de la double donnée du sens des mots et du sens des relations entre mots [Big06].

4. Niveau pragmatique (contextualiser) : La pragmatique vise à l'étude du sens dans un contexte, car l'aboutissement à la signification des mots dans des phrases isolées en dehors du contexte ne suffit pas pour y avoir la signification complète telle que l'humain l'appréhende lors d'un processus de compréhension. Pour cela l'analyse pragmatique est indispensable, elle consiste à trouver la signification "réelle" des phrases liées aux conditions situationnelles et contextuelles d'utilisation des mots [Del00].

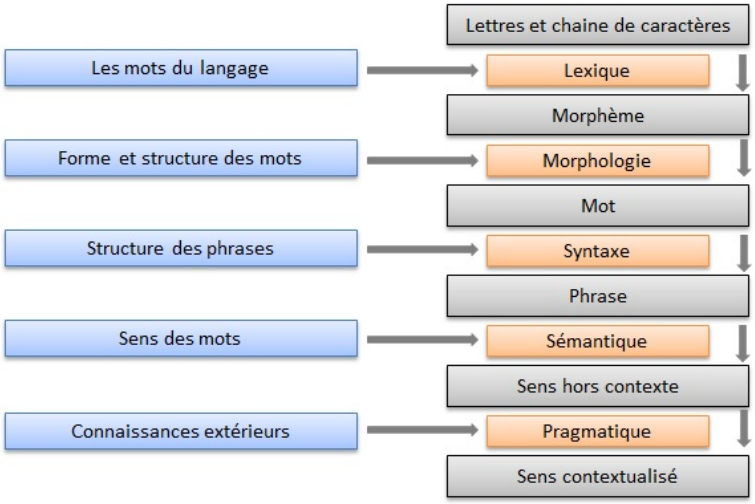


FIGURE 1.5 : Chaine TALN

## 1.2 Histoire des TALN

Afin de tracer la courbe illustrant l'évolution des TALN, on doit d'abord connaitre son repère, Quand est-ce qu'elle a commencé, qu'est-ce qu'elle cherche à couvrir à l'époque et comment elle progresse jusqu'à nos jours. On a préféré d'y venir par plusieurs chemins, en recensant quelques évènements importants pour témoigner du progrès des Travaux en linguistique dans un premier temps, en suite la naissance de l'ordinateur, puis l'apparition du paradigme linguistique informatique. Enfin et finement, on donne un aperçu sur les traitements statistiques du langage naturel.

### 1.2.1 Anciens travaux en linguistique

- 1588-1648 Marin Mersenne aborde l'étude de la phonation, d'un point de vue articulatoire, acoustique et mécanique.
- 1660 Publication de la Grammaire générale et raisonnée (Grammaire de Port-Royal) d'Antoine Arnauld et Claude Lancelot décrivant les règles du langage en termes de principes rationnels universels.
- 1700-1900 Règne de la linguistique comparative.
- 1916 Ferdinand de Saussure publie son Cours de linguistique générale [dSBSR83].
- 1930-1940 Le cercle de Prague prolonge les analyses de Saussure et promeut une linguistique structurale
- 1951-1954 Harris publie ses travaux sur la linguistique distributionnaliste [Har54].
- 1957 L'œuvre de Noam Chomsky marque l'histoire de la syntaxe des 50 dernières années [Cho57].

Par la suite, nous verrons l'évolution du TALN après et tout au long l'émergence de l'informatique, on se focalisant sur l'utilisation des méthodes statistiques et d'apprentissage automatique par la linguistique informatique.

### 1.2.2 Linguistique informatique

- 1949 aux Etats-Unis, Warren Weaver dans Memorandum a parlé de la mécanisation du problème de traduction [Wea55], ce qui lui a fait :

  Aborder les problèmes de l'automatisation du traitement du langage.

  Proposer de résoudre les ambiguïtés syntaxiques et sémantiques en utilisant la redondance du langage écrit dans le cadre de la théorie de l'information (Shannon et Weaver, 1948)
- 1950 Essor fulgurant des recherches en traduction automatique (TA).
- 1954 Première expérience de TA du russe vers l'anglais par IBM, en utilisant l'IBM 701 une machine de traitement électronique des données, c'était le premier ordinateur scientifique commercial de l'entreprise, dont son programme intègre des algorithmes de logique qui ont fait des «décisions» grammaticales et sémantiques imitant le travail d'un personnel bilingue
- 1954 Premiers numéros de la revue « Mechanical Translation »
- 1955 Premier ouvrage sur la TA (Booth et Locke) [LB56].
- 1959 Création de l'Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée (ATALA).
- 1962 Création de l'Association des linguistiques computationelle "Computational Linguistics (ACL)".
- 1965 Première conférence internationale de linguistique informatique biannuelle : Coling
- 1975 Création de la revue Computational Linguistics.

### 1.2.3 TALN statistique

- 1940s, 1950s : Langage est un processus séquentiel, modèle de Markov
- 1950s, 1960s : Chomsky ; les méthodes statistiques ne sont pas adéquates pour la langue naturelle.
- 1970s, 1980s : L'émergence du TALN statistique par IBM Watson group qui ont instillé ses grains.
- 1990s : Le paradigme IBM Watson est adopté par l'association de linguistique informatique.
- 2000s : Trois communautés sont apparues dans le TALN statistique.
  Linguistique informatique traditionnelle.
  Un tas important de chercheurs utilisent méthodes statistique simple.
  Un petit groupe de chercheurs qui font des recherches actives sur les méthodes d'apprentissage automatique.

## 1.3 Ressources linguistiques

Pour une langue donnée, un système TALN peut avoir besoin des connaissances linguistiques, soit bien structurées connues sous le noms de lexique, soit un jeu important de documents connu sous le nom de corpus.

### 1.3.1 Les lexiques

Un lexique est constitué d'une liste d'entrées lexicales auxquelles peuvent être associées des informations linguistiques(règles ou entrée des dictionnaires). Dans la littérature il y a deux types de lexique monolingue et multilingue, chacun et ses utilités. Les informations de chaque entrée relèvent ses propriétés dans la morphologie, la syntaxe, ou la sémantique ainsi que sa fréquence d'usage, des exemples d'emploi, etc. Elles peuvent être aussi l'interprétation de ce mot aux langues dans le cas des lexiques multilingues les unités sont regroupées selon deux liens :

1. les liens morphologiques permettent de lier l'unité lexicale à sa forme de base. Ils regroupent les informations flexionnelles et dérivationnelles (lien entre une forme fléchie et son lemme).
2. les liens sémantiques lient l'entrée lexicale avec ses informations sémantiques.

### 1.3.2 Les corpus

C'est l'anneau crucial dans la chaine de traitement automatique statistique de la langue, auquel s'exerce toute sorte d'expérimentation pour les approches empiriques dont

elles sont par définition basées sur des corpus, il est défini par Laporte : « l'ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique » [Lap00]. Pour la création d'un corpus on passe par plusieurs étapes : collecte, étiquetage et purification, ces étapes différent selon les sources des documents collectés (archives, base de données ou du WEB), la figure suivante 1.6 résume le processus de création d'un corpus à partir du WEB. Le premier corpus organisé par l'ordinateur était "Brown University Standard Corpus "pour l'anglais américain contemporain (réputé par the Brown Corpus), compilé en 1960s par les linguistes Henry Kuera et W. Nelson Francis. Selon Pincemin le corpus doit vérifier trois types de conditions : des conditions de signifiance, des conditions d'acceptabilité, et des conditions d'exploitabilité [Pin99].

- Conditions de signifiance : un corpus est constitué en vue d'une étude déterminée, portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue. Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse.
- Conditions d'acceptabilité : le corpus doit apporter une représentation fidèle, sans être parasité par des contraintes externes. Il doit avoir une ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse.
- Conditions d'exploitabilité : les textes qui forment le corpus doivent être Semblable. Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme).

On distingue deux catégories de corpus : spécialisé dans un domaine particulier (corpus techniques, médicaux) ou généraliste rassemblant des textes plus diversifiés, dans le but d'avoir une bonne représentation de la langue.

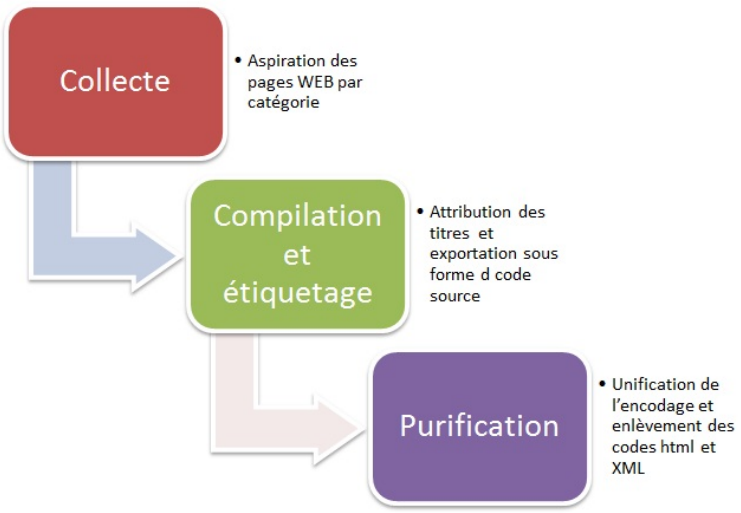


FIGURE 1.6 : Processus de creation d'un corpus

## 1.4 Enjeux des TALN

Dans cette section on argumente la difficulté du domaine, en mentionnant d'où elles sont engendrées, puis on expose les deux voies adoptées pour la maitriser

### 1.4.1 Pourquoi le TALN est difficile ?

Une réponse directe et succincte à la question "pourquoi le TALN est difficile ? " est de dire : reproduire tous le système de cognition des humains, dès la première perception (lire ou entendre ...), puis l'analyse implicite de la structure lexicale et syntaxique jusqu'à la compréhension à partir du sens et du contexte. Et réciproquement vers l'expression des idées en langage naturel [MS99].
En effet, c'était simple à dire mais en réalité on est amené à reconfirmer que l'aspect multidisciplinaire est à l'origine de cette difficulté. Dans un laboratoire de recherche de TALN on trouve que chaque étape présente un bain de recherche pour un tas de spécialistes qui maitrisent les moyens à utiliser dans ce champ ou tentent de les maitriser. Dont on trouve les psychologues et les cogniticiens lors de l'étude de l'interaction homme-machine (IHM), les électroniciens pour le traitement du signal brut acoustique ou optique ..., et les informaticiens qui interviennent avec les linguistes pour l'analyse et la compréhension.

Ici on s'intéresse à l'étape qui nous concerne en tant que des informaticiens, dont on cherche à structurer les données en entrée, pour qu'elles soient manipulables par la machine, autrement dit le déterminisme (la désambiguïsation), et on cherche aussi à représenter toute la connaissance amenée par l'entrée ainsi que toutes les connaissances requises pour l'analyse de cette entrée que ce soit explicite ou implicite.

#### 1.4.1.1 La désambiguïsation

L'ambiguïté réside dans tous les niveaux de traitement cités auparavant, son enlèvement est sollicité par plusieurs travaux du TALN, qui visent à formaliser l'entrée linguistique à travers une bonne représentation pour maitriser ensuite l'ambigüité et la déterminiser par des approches régulière, ce qui est étonnant de l'achever, ou empirique ce qui reflète la réalité de ce stade.

#### 1.4.1.2 L'implicite

L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, sensément dotés d'une connaissance du monde et de son fonctionnement telle que l'immense majorité des éléments de contexte nécessaires à la désambiguïsation mais aussi à la compréhension d'un énoncé naturel peuvent rester implicites. La situation change du tout au tout dès qu'une machine tente de s'insérer dans un processus de communication naturel avec un humain : la machine ne dispose pas de cette connaissance d'arrière-plan, ce

qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible. [Yvo10]

Donc la majorité des applications du TALN sont dédiées à un domaine bien spécifié tels que textes juridiques, textes scientifiques, serveur d'information spécialisé dans les informations sportives

### 1.4.2 Le régulier et l'empirique pour maitriser la langue

Après avoir vu les problèmes associés au TALN, voyons maintenant comment les résoudre. Par intuition on pense à découvrir et utiliser les règles linguistiques qu'on doit respecter lors de la construction d'un mot (règles lexico morphologiques), ou pour la génération d'une phrase (la syntaxe et la sémantique). En effet cette intuition a tracé un chemin pour les linguistes dans un premier temps, dont leur but était –est toujours- la capture des régularité d'une telle langue, puis les informaticiens qui vise à induire les règles grammaticales ou les expressions régulières ou algébriques depuis le langage naturel par analogie aux langages évolué même si la complexité du naturel est apparente. En fait les résultats obtenus n'étaient plus encourageants, c'est la raison qui a réorienté les chercheurs vers l'utilisation des heuristiques dans leurs propositions d'où l'émergence de l'empirique, dont la concurrence est jugée par les résultats atteints, en évaluant la proximité de la réalité, le deuxième critère qui est aussi indispensable est la bonne couverture ou la bonne représentation de la langue traitée.

### 1.4.3 Fondements théoriques et mathématiques

Nous exposons ici les préliminaires de l'apprentissage automatique Machine learning lors du traitement automatique de la langue indépendamment de toute ressource de connaissances linguistique (les approches data-driven), et on se restreint aux méthodes utilisées lors de l'analyse morphologique qui présente le but de notre travail, ces méthodes sont considérées comme des modèles mathématiques, le plus haut niveau d'abstraction qu'on puisse atteindre pour représenter la réalité, plus précisément, les modèles dont on est en train de formuler sont des modèles statistiques, où on essaye de représenter le modelé que par des distributions de probabilités dans le but de prédiction de future observations ou du cluster dont il appartient le modèle, etc.

#### 1.4.3.1 La modélisation statistique

Dans la modélisation statistique, on identifie une formule de distribution de probabilité par des paramètres,[Can11] dont l'estimation de ces derniers est la pierre angulaire des statistiques qu'on fasse lors de l'apprentissage. Voilà trois méthodes d'estimation reconnues dans le TALN.

– **Estimation de la probabilité maximale (MLE)** émet l'hypothèse qui a la plus grande probabilité dans un jeu de données. Et on appelle La probabilité d'avoir ces données sous cette hypothèse par la vraisemblance. Dans un jeu de données D, la vraisemblance V d'une hypothese $\theta$ est égal à la probabilité d'avoir ce jeu sachant qu'on a l'hypothese $\theta$.

$$L(\theta/D) = p(D|\theta)$$

la vraisemblance maximale $\hat{\theta}$ est donnée par :

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} P(D|\theta)\\ \hat{\theta} &= \arg\max_{\theta} P(x_1, x_2, x_3, \ldots x_k|\theta)\\ \hat{\theta} &= \arg\max_{\theta} \prod_{i=1}^{k} p(x_i|\theta)\end{aligned}$$

– **Description de la longueur minimale** en Anglais Minimum Description Length (MDL). Cette technique est basée sur la théorie de l'information. Elle était proposée par Rissanen [Ris78]. MDL cherche à découvrir l'hypothèse qui conduit à la meilleure compression des données qui se fait par la détection des régularités dans les données [Gru04]. Ce dernier a dit, plus qu'il y ait de compression plus qu'on apprenne des données.
**Exemple** appliquons le MDL pour la segmentation morphologiques, prenant l'hypothèse conduisant au corpus ayant l'espace minimale comme sortie du processus de compression sachant que les régularités sont déterminées (morphèmes), voici un ensemble D = walked ; talked ;walking ; talking. Si les données sont compressées comme ci D = walk ; talk ; ed ; ing, dont on a aucune perte d'information, donc l'analyse morphologique de ce corpus sera achevée.
– **les n-grammes** est une sous-séquence de n éléments construite à partir d'une séquence donnée. L'idée semble provenir des travaux de Claude Shannon en théorie de l'information. Son idée était que, à partir d'une séquence de lettres donnée (par exemple "par exemple") il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante. À partir d'un corpus d'apprentissage, il est facile de construire une distribution de probabilité pour la prochaine lettre avec un historique de taille $h = 1$. Cette modélisation correspond en fait à un modèle de Markov d'ordre un où seules les dernières observations sont utilisées pour la prédiction de la lettre suivante. Ainsi un bigramme est un modèle de Markov d'ordre 2. **Exemple** À partir du petit corpus "par exemple", nous obtenons :
  – Pas d'historique (unigramme) :
    p : 2 occurrences sur 10 lettres = 1/5 ;
    e : 3 occurrences sur 10 lettres = 3/10 ;
    x : 1 occurrence sur 10 lettres = 1/10 ;
    ...

La somme des probabilités étant nécessairement égale à 1.

– Historique de taille 1 (on considère la lettre et un successeur) :
  p-a : 1 occurrence sur 9 couples = 1/9 ;
  p-l : 1 occurrence sur 9 couples = 1/9 ;
  p-e : 0 occurrence sur 9 couples = 0 ;
  ...
  La somme des probabilités étant toujours nécessairement égale à 1.

Nous obtenons des probabilités conditionnelles nous permettant de connaître, à partir d'une sous-séquence, la probabilité de la sous-séquence suivante. Dans notre exemple $p(a|p) = 1/9$ est la probabilité d'apparition de l'élément a sachant que l'élément p est apparu. À titre d'exemple, le bi-gramme le plus fréquent de la langue française est de, comme dans l'article de, mais aussi comme dans les mots demain, monde ou moderne.

#### 1.4.3.2 Représentation des connaissances linguistiques

Le TALN fait appel, d'une façon ou d'une autre, à des objets formels qui représentent la connaissance linguistique soit une partie du sens des textes, soit une partie du savoir humain, soit une partie du raisonnement humain. Ces objets sont des modèles très simplifiés de la réalité. Au suivant quelque techniques de modélisation et de formalisation :

1. la logique du premier ordre ou les prédicats et les arguments avec notamment les connecteurs logiques, les quantificateurs, la déduction automatique et l'inférence automatique,
2. les réseaux sémantiques, dans lesquels les nœuds représentent les mots et les transitions des relations entre eux,
3. la résolution des références, c'est-à-dire établir le lien entre les expressions dans un texte et les objets formels correspondants dans les modèles de représentation des connaissances, la modélisation du dialogue.

Ceci était pour le haut niveau concerné par la compréhension sémantique et la contextualisation, lorsqu'on parle de niveau bas, on est sensé alors par la représentation de la forme des données linguistiques, mots dans le palier lexical, et morphologique(voir Figure 1.8), et phrase dans l'analyse syntaxique (voir Figure 1.7). La forme naturelle pour représenter un mot ou une phrase (elle peut être vue comme un mot dont ses lettres sont les mots de cette phrase) est les chaines et les sous chaines, comme on trouve aussi les machines à état finis qui offrent aussi un format naïf de représentation et simple à manipuler tels que les automates, les transducteurs et les treillis.

## 1.5 Applications du TALN

Plusieurs applications sont couvertes par le domaine TALN. Elles sont abordées depuis quelques décennies, quelques années ou juste dernièrement. Jurafsky et Manning ont

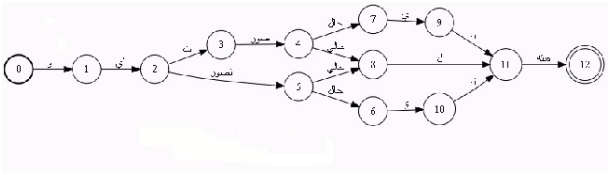

FIGURE 1.7 : La phrase **واي تصور حالي لهته** segmentée par un treillis

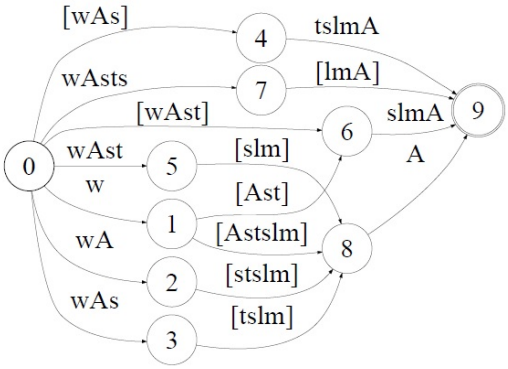


FIGURE 1.8 : Le mot **واستسلما** [wastslma] représenté par MEF

proposé une classification selon le degré du progrès (voir Figure 1.9) : "bon avancement ou presque résolus", "en bon essor" ou "encore difficile". Autrement dit dans ce qui suit, nous allons dévoiler quelques domaines d'application de cette discipline.

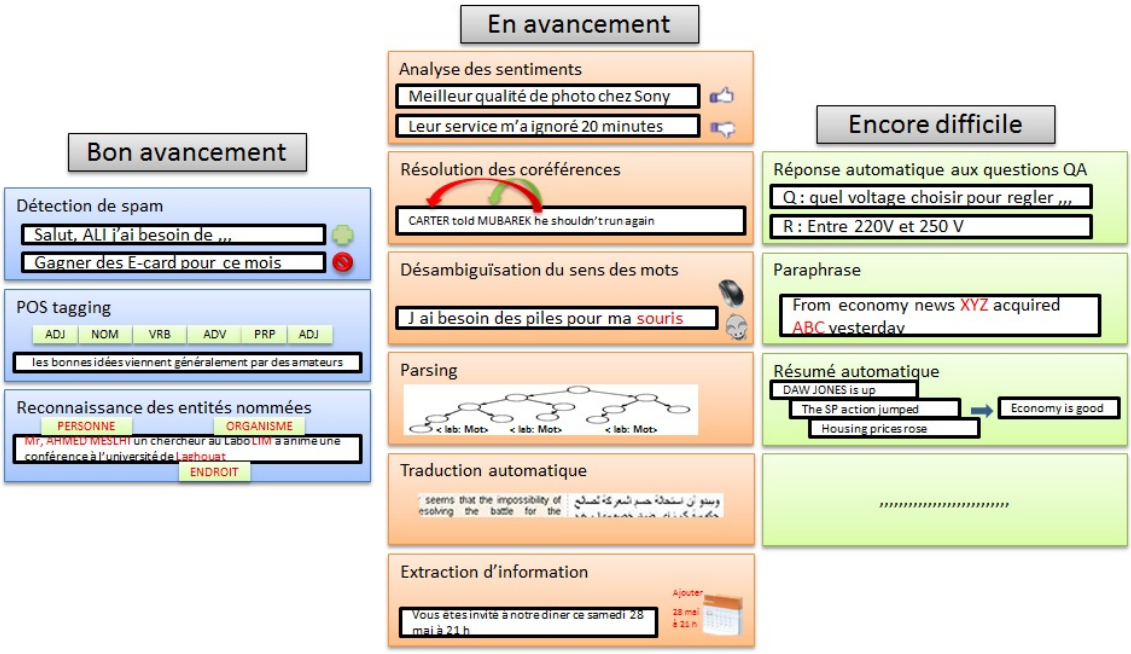


FIGURE 1.9 : Applications TALN

### 1.5.1 Bon avancement

#### Détection de spam

Elle vise à détecter les courriers indésirables d'une façon automatique et les répertorier en une classe particulière, cela se fait par des heuristiques visant à analyser la forme du message (les IDS. les systèmes de détection d'intrusion) ou le contenu dont le TALN s'intéresse.

#### Reconnaissance des entités nommées

C'est l'identification et la catégorisation des noms propres, en se basant sur des règles linguistiques qui utilisent des informations morphosyntaxiques, ou des méthodes statistiques par apprentissage sur des corpus étiquetés manuellement.

### 1.5.2 En avancement

#### Traitement de la parole

C'est l'étude des signaux de parole et leur traitement. Ces signaux sont généralement traitées sous une forme numérique, la filière est considérée comme un cas particulier de traitement numérique du signal, elle comprend les aspects de l'acquisition, la manipulation, le stockage, le transfert et la production des signaux vocaux numériques. Elle est aussi étroitement lié au traitement du langage naturel (NLP), comme son entrée peut provenir / sortir des applications du TALN. Par exemple la synthèse (text-to-speech) peut utiliser un analyseur syntaxique de texte en entrée comme on peut appliquer les techniques d'extraction d'information sur la sortie d'un système de reconnaissance de la parole.

#### Analyse des sentiments

C'est une filière de la linguistique informatique qui s'intéresse par l'extraction des émotions du texte depuis n'importe quelle source de langage naturel. On trouve plusieurs applications de cette analyse comme : donner des avis sur des produits -Product reviews-, le suivi des marques -brand tracking-, les informations financière et les blogs politiques ... . La branche est assez récente, elle est apparu la première fois comme une extension de l'extraction des connaissances -knowledge discovery-, pour cela elle est connus aussi sous le nom de Opinion mining. Le domaine n'était qu'un nouveau-né incubé par les laboratoires de recherche jusqu'à l'apparition du fameux article " twitter mood predicts the stock market " [BMZ11].qui a bouleverser le cas lorsqu'il a mentionner l'impact de ces analyses

### La décomposition analytique

En Anglais "*Parsing*" , c'est l'analyse formelle par ordinateur d'une phrase ou autre chaîne de mots en ses constituants afin de construire un arbre montrant les relations entre ses composants, qui peut également contenir des informations sémantiques ou autres.

### L'étiquetage des parties de discours

En anglais "*Part of speech tagging -P O S tagging-*", une opération simple sur les mots, elle consiste à attribuer à chaque mot des différentes caractéristiques morphologiques syntaxiques ou sémantiques elle est utilisée par la majorité des technique du TALN.

### Extraction d'information

Lorsque les techniques de compréhension des textes dans le TALN sont encore en progrès, une alternative qui minimise ses exigences et rétrécit ses résultats, ne cherche qu'extraire des éléments pertinents d'un texte donnée qui doivent répondre aux besoins définis auparavant par le modèle. Une définition succincte de cette filière est donnée par T. Poibeau comme suit " L'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle ".

### Recherche d'information

L'intitulé semble générique, il ne donne pas un aperçu sur les tâches et les traitements à accomplir, mais il parvient à leur objectif, qui est la collecte après avoir retrouvé dans un corpus tout genre d'informations conforme à une réponse pertinente d'une telle requête qui est en langue naturelle ou juste des mots clés. Le fouillis qu'on l'épluche est composé de documents qui sont décrits par une forme et un contenu. La forme est exprimée par leurs appartenances (bases de données, documents WEB : XML, Html ..., reliés ou non par des liens hypertextes ... etc.), et le contenu qui peut être du texte, des sons, ses images ou des données.

1. (Salton, 1968) [Sal68] Recherche d'information c'est le domaine qui étudie la structure, l'analyse, l'organisation, le stockage, la recherche et la récupération d'informations.
2. (Lancaster, 1968) [Lan68] Un système de recherche d'information n'informe pas (c'est à dire changer la connaissance de) l'utilisateur sur l'objet de son enquête. Il ne fait que l'informer seulement sur son existence ou pas, et lui restituer les documents relatifs à sa demande.
3. (Needham, 1977) [Jon97] : La complexité découle de l'impossibilité de décrire le contenu d'un document, ou en induire l'intention de la requête précisément et sans ambiguïté.

#### Classification automatique des textes

TC acronyme de l'anglicisme text classification, également connu comme la catégorisation de texte. C'est la tâche de trier automatiquement un ensemble de documents en catégories, classes ou des sujets à partir d'un ensemble prédéfini. Cette tâche, qui relève au carrefour de la recherche d'information (RI) et l'apprentissage automatique (ML), a connu un intérêt en plein essor par et les chercheurs et les développeurs.[Seb99]

#### Traduction automatique

Elle était l'une des motivations en informatique dans son premier ère, apparue la première fois par IBM en 1954, la tâche est simple à dire et simple à exprimer "'J'ai un texte en face de moi qui est écrit en russe, mais je vais faire semblant que c'est vraiment écrit en anglais et qu'il a été codé dans certains symboles étranges. Tout ce que j'ai à faire est de dépouiller le code afin de récupérer les informations contenues dans le texte "' [Wea55].

#### Résolution des coréférences

C'est la détection automatique des noms, phrases et pronoms qui réfèrent à la même entité (Personne, objet, idée, endroit ... etc.)[NK08]. Elle est utilisée par plusieurs applications du TALN tels que l'analyse des sentiments, la traduction automatique et l'extraction d'information ... .Le processus se fait par plusieurs approches heuristiques, dont on trouve des techniques basée sur des règles et des connaissances linguistiques et d'autres méthodes basées sur l'apprentissage automatiques (data-driven).

#### Désambiguïsation de sens des mots

Comme son nom l'indique, le but de cette procédure est d'avoir un déterminisme lors de l'extraction du sens d'un matériau linguistique, ce problème était considéré comme un AI- complète [BH60], dont sa résolution revient à résoudre des problèmes liés à l'intelligence artificielle, comme la représentation des connaissances. En se focalisant sur ce dernier point pour comprendre le coup de pousse important qu'a eu la désambiguïsation grâce à l'émergence des réseaux sémantiques et leurs évolutions.[IV98]

### 1.5.3 Encore difficile

#### Dialogue homme machine

C'est une perspective pour le TALN, simple à exprimer, mais très difficile à y parvenir, elle présente un domaine de recherche pluridisciplinaire qui fait appel à la philosophie, sciences cognitives et sociales, informatique, télécommunication, ... [Ngu05] pour produire un système capable de cognition au même titre qu'un homme, en utilisant la parole

comme un support de communication, dont son traitement revient au TALN dans des stages postérieurs. Lors de la compréhension et antérieurs lors de la génération.

### Résumé automatique

Ce domaine aide à contribuer à une meilleure compréhension de la façon dont les gens produisent et comprennent la langue, car il peut résoudre les besoins croissants d'information de synthèse dans notre société. La tâche de résumé semble être intrinsèquement interprétée dans le sens où différentes personnes produisent généralement des résumés très différents pour un texte donné. Ainsi, la qualité des résumés peut être jugée très différemment [DCFVM+07] En matière de résumé automatique, on peut distinguer trois principales approches à savoir, l'approche par compréhension appelée l'approche symbolique, l'approche par extraction appelée l'approche numérique et l'approche qui combine les deux approches précédentes appelée l'approche hybride. [IK12]

### Réponse automatique aux questions

(Anglicisme : Question answering) une filière enclenchée par la recherche d'information (RI) où on vise à structurer une bonne requête interprétant la question posée en langage naturel, puis on cherche à synthétiser une réponse agrégée à partir du résultat obtenu par un processus de RI plus fin. Répondre à une question en langage naturel revient à traiter et analyser cette question, et de créer une certaine représentation de l'information demandée. Cette représentation nécessite un traitement afin de déterminer :

- Le type de question, généralement basé sur une taxonomie des questions possibles déjà codées dans le système ;
- Le type de réponse attendue, grâce à une certaine transformation profonde sémantique de la question, et
- La question principale, qui représente l'information nécessaire

Ces étapes servent à passer un ensemble de termes présentant la requête à chercher dans les textes indexés. [Lam04]

### Paraphrase

Elle cherche à résoudre le problème de décèlement de l'équivalence de sens entre segments textuels qui est au cœur des besoins du Traitement Automatique des Langues. La capacité à déterminer si deux mots, ou deux groupes de mots, ont la même signification dans leur contexte respectif permet de résoudre, au moins localement, les difficultés posées par la variation en langue. L'acquisition de groupes d'équivalence permet de constituer des ressources pouvant être utiles. [Bou12]

## 1.6 Problèmes ouverts

Même si tout cet essor dans les applications des TALN mais la recherche reste toujours encore ouverte, on n'est pas encore arrivé à un modèle générique pour le TALN. Mais toutes les applications sont plus aux moins circonscrites. Donc la structure naturelle elle-même d'un processus de TALN nous parle des domaines des recherches à fouiner :

- Recherches fondamentales en matiere de :
  - Compréhension de texte
  - Génération de texte
- TAL porte sur la forme et le contenu
  - Modèles d'interprétation
  - Acquisition des corpus
- Connaissances linguistiques complètes
  - Morphologie
  - Syntaxe
  - Sémantique
  - Pragmatique

## 1.7 Conclusion

Dans ce chapitre, on a présenté une nouvelle discipline, qui s'intéresse au traitement automatique du langage naturel. Elle est issue de l'intelligence artificielle, dont elle rassemble de l'apprentissage automatique, et la représentation des connaissances, où on a profité le créneau pour donner quelques concepts préliminaires avec des exemples d'applications dans le TALN, avant ça nous avons défini le domaine comme une synergie entre la linguistique et l'informatique, en donnant ses ancêtres depuis les deux filières, ensuite on a donné une classification des objectifs visés par ce domaine, ainsi que ses niveaux de traitement, nous avons exposé son histoire depuis les premiers travaux apparus qui s'intéressaient par la langue naturelle jusqu'à l'arrivée à nos jours où on a clôturé le chapitre par les sujets d'actualité de la linguistique informatique, comme on a démontré les problèmes encore ouverts, qui font de la recherche fondamentale de cette discipline afin de repérer notre intérêt sur l'analyse morphologique, dont on va s'attarder sur l'exploration des concepts dans le prochain chapitre.

# Chapitre 2

# L'analyse morphologique

Dans le chapitre précédent, nous avons donné un aperçu général sur le TALN, depuis ses fondements jusqu'à ses applications, nous avons mis aussi les points sur ses problèmes et ses défis rencontrés tout au long de l'évolution de cette discipline.

Nous enchainons dans ce chapitre, par les détails au niveau morphologique, l'une des issues encore ouvertes, qui présente à la fois une étape primordiale dans plusieurs applications de TALN d'un coté, et un domaine de recherche fondamentale pour la linguistique informatique de l'autre coté, dont il n'a jamais cessé d'améliorer ses performances par l'accumulation des résultats de plusieurs travaux qui reviennent aux moins à cinq ou six décennies de nos jours [Oet60].

L'organisation du chapitre suit un plan incrémental, on commence par la clarification des concepts et des notions de base, définitions, terminologie et la situation de la morphologie et son interaction dans la hiérarchie d'un processus de traitement automatique de la langue. Par la suite, on rappelle les applications de l'analyse à ce bas niveau. Puis, on énoncera les approches adoptées pour faire cette analyse, en s'étalant sur les méthodes non supervisées, comme un point d'intérêt et on termine par l'exposition de l'état actuel de l'analyse morphologique dans la langue Arabe.

## 2.1 Concepts et terminologie

Nous commençons par l'explication de l'analyse morphologique (AL) en donnant quelques définitions bien répandus dans la littérature du TALN, puis on présente la terminologie standard de ce domaine. Par la suite on discute l'opération de dérivation et on verra ses interactions avec les autres niveaux linguistiques.

### 2.1.1 L'analyse morphologique

La morphologie est l'étude, l'identification, l'analyse et la description des plus petites unités portant un sens qui constitue un mot, ces unités s'appellent **"morphèmes"** [JM08].

L'analyse morphologique est le processus de catégorisation et la construction d'une structure représentative des morphèmes composants d'un mot, la prise en compte des règles orthographiques et morphologiques est indispensable ; par exemple le pluriel du

mot anglais « *party* » est « *parties* », dont la règle d'orthographe nous force à changer le *–y* à un *–i* et ajouter un *–es*. Les règles morphologiques nous informe que le mot "*fish*" n'a aucun pluriel [JM08].

L'analyse morphologique automatique est apparue aux années cinquantes du siècle précèdent 1950s, pour offrir un support à la traduction automatique. Le stemmer [1] de Porter [Por80] est un ancien exemple d'un analyseur morphologique qui est utilisé beaucoup dans la recherche d'information. Plusieurs applications de TALN développées anciennement ont sollicité l'automatisation de l'analyse morphologique tels que les correcteurs d'orthographes "*spell cheker*", les systèmes de reconnaissance vocale ou optique, les systèmes de génération de texte. Cependant à l'époque il y avait peu d'intérêt quant à l'évaluation de l'exactitude des résultats obtenus par les analyseurs morphologiques dans les premières applications. La préoccupation était sur la justesse des résultats plutôt que sur la vérification et l'analyse des méthodes en elles même [RS07].

Les approches basées sur des machines à états finis MEF pour l'analyse morphologique automatique ont été dominantes depuis les années 1980, à l'origine cette méthodologie a été étudié chez Xerox. En effet, la première application était due à Koskenniemi [Kos83], dans le but était de développer des analyseurs morphologiques ayant une couverture étendue sur plusieurs langues.

### 2.1.2 Les constituants d'un mot

En général, pour n'importe quelle langue donnée, un mot est composé d'un ensemble de morphèmes, dans ce qui suit, on abordera en plus détail la notion de morphème.

#### Morphème

La plus petite unité porteuse d'un sens dans un mot est appelé un morphème. On trouve un ou plusieurs morphèmes dans un seul mot tandis qu'il est souvent dépourvu d'autonomie linguistique, ce petit élément significatif peut être capturé par la segmentation d'un mot qui est l'opération présentant l'objet crucial de l'analyse morphologique.
***Remarque*** : on peut utiliser aussi le terme morphe en alternance avec morphème en se référant à la forme physique d'un morphème comme un ensemble de phonèmes où chacun présent un son distinct [Kat06].
***Exemple*** : le mot «*atomisation*» peut être segmenté comme suit *"atom+is+ation"*.

On distingue deux types de morphèmes : libres ou liés. Ceux qui peuvent apparaitre librement, sans s'associer avec d'autres morphèmes sont appelés morphèmes libres. Morphèmes liés ne peuvent être qu'une partie attachée d'un mot, ou coexister en combinaison avec d'autres morphèmes.
***Exemple*** : en anglais "*Pen, effet, sleep*" sont des morphèmes libres qui peuvent se produire sans s'associer avec d'autres morphèmes, tandis que les *"in-, de-,-isme"* sont des morphèmes liés qui ne peuvent être qu'une partie d'un mot comme *"inaptitude, désactiver*

1. Dans ce qui suit du mémoire, on utilisera le terme *stemmer* en alternance avec le terme *analyseur morphologique*

*et le déterminisme"*. Dans le mot "heures", il y a deux morphèmes, dont l'un est un morphème lié *-s* et l'autre est un morphème libre *heure*.

### Allomorphe

c'est une variante d'un morphème qui se diffère de ce dernier juste dans la forme, on veut dire par la forme la représentation phonétique qui interprète la prononciation même si l'écriture est la même [HS02].
***Exemple*** : en anglais le pluriel peut être prononcé par des manières différentes dans les mots suivants : *cats [s], in dogs [z], et faces [az]*, et le morphème est le même *"S"*.

### Racine

"*root* en anglais" c'est l'élément de base, irréductible commun à tous les mots de même origine. En Arabe Une racine se compose généralement de trois lettres portant l'aspect du contenu sémantique. Bauer donne une autre définition basée sur le concept morphème, dit que la racine est tout morphème libre qui ne peut plus être analysé encore, et pour lequel les morphèmes liés sont attachés [Bau03].

### Radical

c'est l'une des formes prises par la racine dans les réalisations diverses, ex. racine verbale : ven (venir), deux radicaux : ven/vien (venons, vient). Mais en Arabe un radical c'est l'une des lettres constituant la racine ex.ب \ ت \ ك كتب .

### Lemme

c'est la forme canonique d'un mot, c'est une réduction amoindrie, elle ne cherche pas à remonter jusqu'à la racine mais elle veut juste le format standard de ce mot à cette position, comme les verbes sont réduits à l'infinitif et les noms au masculin singulier ..., la lemmatisation est utilisée avec le POS tagging lors de l'analyse syntaxique.

### Stem

La différence entre un stem et une racine, c'est que les racines ne peuvent plus être divisées encore et que le stem peut être divisé encore à des morphèmes même s'il est produit à partir d'une segmentation. Khorsi définit le stem comme étant le plus petit segment du mot contenant toutes les lettres de la racine [Kho12].
***Exemple*** : on segmente le mot مكتبات à un stem مكتب et un morphème ات , le stem peut être réduit encore à une racine كتب .

### Affixes

les morphèmes liés qui n'apparaissent qu'attachés à des mots, et ils portent un sens abstrait distinct sont appelés affixes [HS02], dont ceux qui viennent au début du mot sont

des préfixes, ceux à la fin présentent des suffixes, il y a aussi des affixes qui s'insèrent entre les lettres d'un stem qui s'appellent des infixes comme-ci le cas dans la langue Arabe ا dans le stem كتاب dans le mot الكتاب.
***Exemple*** : Dans le mot anglais *"unintentionally"*, *un-* est un prefixe, *-ion , -al and -ly* sont des suffixes.

**Stop words**

sont des mots ayant une sémantique amoindrie. Leurs existence est importante dans la structure des phrases. Ils définissent les relations grammaticales des autres mots dans la phrase. Elles indiquent également les relations structurelles des mots les uns aux autres. Ces petit mots sont les pronoms, les prépositions, déterminants, conjonctions, auxiliaires et les verbes modaux [BHM06]. Dans certaines langues, certains mots ne sont pas de fonction autonome, mais clitiques attachés et concaténés à des mots, comme dans la langue Arabe par exemple " الضمائر المتصلة ".

**Motif**

c'est le modèle à suivre lors de la formation d'un mot, pour maitriser la combinaison des consonnes et des voyelles dans un mot. Les consonnes représentent slots pour les radicaux de la racine (lettres de la racine) à insérer, les voyelles représentent le vocalisme, comme ils présentent les infixes eux-mêmes dans la langue Arabe, Le motif n'est qu'une séquence Consonne-Voyelle(CV). L'approche CV pour les motifs est largement utilisé dans toutes les langues [MP90] [Hab10]. La représentation originale des modèles a été proposée par des chercheurs de grammaire Arabes comme الميزان الصرفي «l'échelle morphologique» qui utilise le verbe passé "fait" pour représenter les radicaux de la racine "فعل" [Ali87].

#### 2.1.2.1 Formation des mots

Pour la plupart des langues naturelles, il existe trois procedes pour former des mots.

- **L'affixation** L'une des façons les plus productives pour former des nouveaux mots est l'affixation qui forme les nouveaux mots par la combinaison des affixes et des morphèmes libres. Il existe trois types d'opération d'affixation :
  préfixation : l'affixe est ajouté au mot en avant.
  suffixation : l'affixe est ajouté à la fin du mot.
  infixation : où un affixe est placé à l'intérieur du stem.
  Comme l'anglais et les langues latines utilisent principalement la préfixation ou la suffixation, de nombreuses autres langues utilisent l'infixation comme les langues semetiques(Arabe, Hebreu, ...) et le tagalog par exemple, une langue des Philippines, l'infixe ***"um"*** est utilisé dans la forme infinitive des verbes voir le tableau ci-dessous 2.1.
  **Exemple**
- **La composition** "*Ang. compounding*", c'est la formation de nouveaux mots non pas à partir affixes, mais à partir de deux mots ou plus indépendants, les mots

TABLE 2.1 : L'infixation en Tagalog

| | | | |
|---|---|---|---|
| sulat | 'write' | sumulat | 'to write' |
| bili | 'buy' | bumili | 'to buy' |
| kuha | 'take' | kumuha | 'to take' |

peuvent être des morphèmes libres, des mots engendrés par affixation, ou des mots composés eux-mêmes (voir tableau 2.2).

**Exemple**

TABLE 2.2 : La composition en Anglais

| | | |
|---|---|---|
| air-conditioner | air | conditioner |
| blackbird | black | bird |
| looking-glass | looking | glass |
| textbook | text | book |
| watchmaker | watch | maker |

– **La réduplication** c'est la formation de nouveaux mots, soit en doublant un morphème libre tout entier (réduplication totale) ou une partie d'un morphème (réduplication partielle). Beaucoup de langues utilisent le redoublement. En Indonésie, par exemple, la réduplication totale est utilisée pour former les pluriels (voir tableau 2.3).

**Exemple**

TABLE 2.3 : La réduplication en Indonésie

| | |
|---|---|
| rumah | 'house' |
| rumahrumah | 'houses' |
| ibu | 'mother' |
| ibuibu | 'mothers' |
| lalat | 'fly' |
| lalatlalat | 'flies' |

#### 2.1.2.2 Opérations morphologiques

Plusieurs niveaux peuvent être considérés lors de l'analyse morphologique ( voir Figure 2.1), moyennant plusieurs opérations telles que :

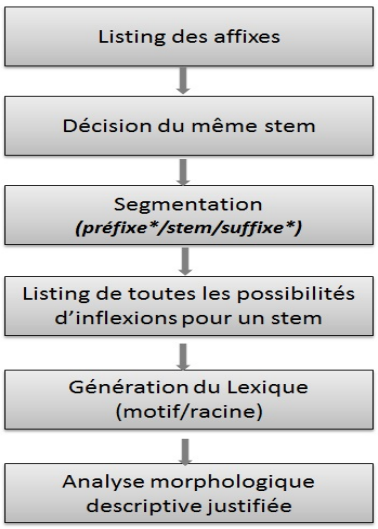


FIGURE 2.1 : Niveaux de l'analyse morphologique

– **Tokenisation ou segmentation** : c'est le processus de définition ou génération des morphèmes d'un mot. Ces morphèmes peuvent être classés en cinq catégories : *proclitiques, préfixes, suffixes, stem et enclitiques.*
Un mot doit avoir au moins un morphème stem, et des combinaisons de clitiques et affixes attachés. Un analyseur morphologique est chargé de définir les variations possibles de la segmentation d'un mot en morphèmes.

– **Lemmatisation** : est le processus de regroupement d'un ensemble de mots à leur forme canonique, celle du dictionnaire qui est appelée le lemme. Par exemple, en anglais, "*course, courses*" sont les formes du même lexème "*course*". Par exemple les mots *l'écriture, écrite, écrivain et écrivez* sont réduits à le lemme "*écrire*".

– **Stemming** : est le processus d'affectation des variantes morphologiques des mots en classes d'équivalence, de sorte que chaque classe correspond à un seul stem. Il est également défini comme la réduction des mots fléchis à leur stem, base (lemme) ou la racine. Plusieurs auteurs confondent les définitions ***lemme*** = ***stem*** .

– **Reconnaissance des motifs** : *ou Ang. Pattern matching* c'est le processus de rapprochement des mots avec leurs motifs possibles, par l'extraction de leurs modèles morphologiques "**الأوزان**". L'algorithme de recherche des motifs fait face à trois types de changements : intégration des infixes, leur substitution et la suppression des voyelles.

- **Diacritisation ou vocalisation** : c'est le processus d'ajout correct des voyelles courtes et des signes diacritiques aux mots. La voyellation est une caractéristique importante de la langue Arabe. Elle permet de déterminer certaines caractéristiques morphologiques de mot. La présence de la voyelle courte sur la dernière lettre permet de déterminer la situation du mot, et la présence d'une voyelle sur la première lettre détermine si le verbe est actif ou passif. La présence de signes diacritiques tels que *Shaddah "الشدة" et El-mad "المد"* (extension) aide à résoudre certaines ambiguïtés dans le sens des mots.

- **L'étiquetage des parties du discours** : *Ang. Part Of Speech abr. POS tagging*, c'est le processus d'attribution des étiquettes de catégories grammaticales(nom, verbe, préposition ...) aux mots d'un corpus. Le marquage est effectué automatiquement en utilisant des programmes *POS tagger*, dont leurs élaboration est un domaine de recherche dans le TALN.

- **L'analyse syntaxique** : ou *Ang. Parsing*, c'est le processus d'analyse de la structure grammaticale d'une séquence de mots. Cette analyse est automatiquement réalisée en utilisant des Parser, ils sont des programmes analyseur syntaxique qui génèrent en sortie les arbres syntaxiques du texte analysé.

### 2.1.3 Dérivation

Les affixes sont en deux catégories : flexionnels ou dérivationnels. Affixes flexionnels construisent des nouvelles formes du mot dérivé, tandis que les affixes dérivationnels créent des nouveaux mots. Ici se pose un autre terme pour faire la distinction. Chaque nouvelle forme résultante d'une inflexion appartient au même lexème, tandis que chaque mot qui est créé par une dérivation est un mot distinct. Donc :

- La morphologie dérivationnelle étudie la construction des unités lexicales et leur transformation selon le sens voulu. Ainsi, la dérivation morphologique est décrite sur une base morphosémantique : d'une même racine, se dérivent différentes unités lexicales.
- La morphologie flexionnelle concerne le marquage pour le nom et l'adjectif ou la conjugaison du verbe, elle ne crée que des formes variées du même mot.

La figure ci-dessous 2.2 montre comment un mot est généré par inflexion et dérivation est grammaticalement correcte. Nouveaux lexèmes sont d'abord construits via une dérivation, puis l'affixation flexionnelle suivra la dérivation.

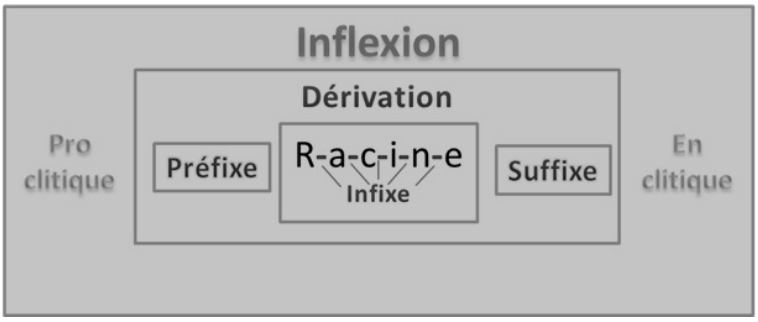


FIGURE 2.2 : Dérivations d'un mot

***Exemple*** : les formes : *buying, buys, bought* appartiennent au même lexème *buy* et les mots en Arabe **يقولون،يقولان** sont derivés du stem **قول** par l'ajout des affixes **ي،ون،ان**. Autrement dit ces différentes formes sont obtenues par l'ajout des affixes d'inflexions

## 2.1.4 Interaction de la morphologie avec les autres niveaux linguistiques

### 2.1.4.1 Interaction morphologie-phonologie

La morphologie est influencée par la phonologie, qui détermine l'utilisation des sons dans les morphèmes. Les allomorphes et l'harmonisation de voyelle sont deux exemples de cette interaction entre la morphologie et la phonologie.
***Exemple*** : l'article indéfini *a ou an* en anglais. Quand les mots commencent par une consonne, "*a*" est utilisé, alors que l'article "*an*" s'utilise dans les mots qui commencent par une voyelle [Kat06].

### 2.1.4.2 Interaction morphologie-syntaxe

L'interaction entre la morphologie et la syntaxe est déterminée par l'inflexion et la dérivation des mots. L'inflexion d'un mot est déterminée par les règles syntaxiques dans une phrase. En d'autres termes, la forme de mot approprié est choisie en fonction de la structure syntaxique de la phrase.
***Exemple*** le mot "*faire*" est exposé à l'affixation pour former une phrase grammaticalement correcte, il peut être sous la forme du passé "*faisait*", participe présent "*faisant*", présents "*fait*" etc, Il est à noter que toutes ces formes appartiennent à la même catégorie syntaxique, qui est "verbe". Contrairement à l'inflexion, la dérivation peut changer la catégorie syntaxique d'un mot, comme les mots *personal, personally, personalize* appartiennent à des catégories syntaxiques différentes (adjective, adverbe, et verbe respectivement) qui sont générés de la racine *person*.

#### 2.1.4.3 Interaction morphologie-sémantique

Certainement, chaque forme de mot a un sens distinct selon les morphèmes lié, mais souvent la classe d'équivalence construite par le stemming porte une sémantique plus ou moins rapprochée, par conséquent, ce qui donne une interaction étroite entre la morphologie et la sémantique.

## 2.2 Domaines d'application

L'analyse morphologique est appliquée partout dans les systèmes TALN, les techniques d'analyse morphologiques forment une étape indispensable servant comme une base de la plupart des systèmes de traitement du langage naturel, (Kiraz (2001) [Kir01], Al-Sughaiyer and Al-Kharashi (2004) [ASAK04], Jurafsky and Martin (2008) [JM08], Pauw and Schryver (2008)) [DPDS08] dans des applications tels que :

### Recherche dans le WEB

Pour les langues ayant une morphologie riche et complexes, l'analyse morphologique permet la recherche de la forme fléchie d'un mot même si la requête ne contient que sa forme de base.

### l'étiqueteur de parties du discours

L'analyse morphologique aide les *Part-of-speech taggers* à déceler le marquage le plus approprié dans un contexte donné.

### Correcteurs orthographiques

Le rôle des *spell-cheker* est la vérification automatique de l'orthographe qui est une application nécessite une étape d'analyse morphologique.

### Reconnaissance de la parole

Elle sollicite fortement la segmentation morphologique. Les systèmes de reconnaissance de la parole sont souvent basés sur un dictionnaire de mots avec un modèle de langue qui fait explorer les caractéristiques de la langue en étudiant les séquences de ses constituants ( morphèmes, mots, ...). Former un dictionnaire de mots est une tâche laborieuse et leur performance est insuffisante, surtout Turc, Arabe, etc). A ce moment-là, la morphologie est le premier niveau qu'on utilise pour faire face au problème du vocabulaire non déterminée par le rétrécissement du nombre infini de formes de mots dans la langue. Puis les modèles de langue considèrent les séquences de morphèmes plutôt que des séquences de mots, dans un souhait de capturer tous les mots enclenchés du même morphème.

- Creutz et al. (2007) [CHK+07] propose la modélisation par morphème pour la reconnaissance de la parole pour le finnois, l'estonien, le turc et un dialecte Arabe parlé en Egypte, qui sont tous considérés comme des langues morphologiquement riches.
- Kirchhoff et al. (2006) [KVB+06] présente plusieurs approches où les morphèmes sont modélisés, afin de contrecarrer la pauvreté des données pour l'Arabe.
- Berton et al. (1996) [BFRB96]et Roeland & Ordelman Jong (2003) [OvHdJ03] exploitent la segmentation des mots composés dans leurs sous-mots pour les langues agglutinantes, par exemple la famille germanique des langues allemand, suédois, néerlandais, etc.
- (Heintz 2010) [Hei10] donne un bon état de l'art des travaux réalisés dans la segmentation des mots pour la reconnaissance de la parole.

## La Traduction automatique

*Ang. Machine Translation* elle utilise la segmentation morphologique, soit à l'étape de prétraitement, de post-traitement, ou par l'intégration de l'information morphologique lors du processus de traduction. Il y a des systèmes qui utilisent la segmentation morphologique pour intégrer des connaissances supplémentaires sur les mots pour créer des modèles de traduction (Yang & Kirchoff 2005) [KY05], (Koehn & Hoang 2007) [KH07], (Avramidis & Koehn 2008) [AK08], dans leurs modèles, différents types de connaissances tels que le lemme et la categorie de partie du discours sont utilisés, ainsi que des informations morphologiques.

Contrairement aux approches mentionnées auparavant, d'autre systèmes de traduction n'utilisent la segmentation qu'au post-traitement, une fois la traduction terminée (Minkov et al 2007 ; Kristina Toutanova 2008) [51]

## La recherche d'information RI

les chercheurs dans la RI exploitent également la segmentation morphologique en raison de l'ambiguïté des mots. La troncature simple des affixes et le stemming qui sont des approches simples, peuvent être adoptées en recherche d'information pour faire correspondre les mots de la requête avec les mots du document (Harman, 1991) [Har91], (Krovetz 1993) [Kro93], (Järvelin & Pirkola 2005) [JP05] Cependant, ces approches sont trop simples à manipuler mais elles ne sont pas suffisantes pour les langues riches morphologiquement, et par conséquent, elles ne peuvent pas gérer l'ambiguïté et les mots hors-vocabulaire.

La génération de stems est également adoptée dans la recherche d'information en engendrant des différentes formes de mots, avant le matching des stems avec les mots du document (Kettunen et al. 2005) [KKJ05]. Pourtant, la production de stems ne donne pas des résultats fiables non plus, surtout dans les langues riches morphologiquement. La lemmatisation est une approche importante qui résout le problème d'ambiguïté en extrayant les formes de base des mots, en tenant compte du contexte.

### Réponse automatique aux Questions

*Ang.Question Answering QA*, un autre domaine qui utilise beaucoup la segmentation morphologique, pour l'extraction des questions, ainsi que pour récupérer les réponses. Les approches utilisées dans la *QA* sont similaires à celles utilisées dans la recherche d'information, le stemming (Bilotti et al. 2004) [BKL04], la lemmatisation (Aunimo et al. 2003) [AHK+03]. L'extension des requêtes est également adoptée au réponse aux questions, où toutes les formes de mots sont indexées, puis les mots sont développés avec leurs variantes morphologiques lors de la récupération des réponses (Bilotti et al. 2004) [BKL04].

## 2.3 Approches existantes

Vu l'importance de l'analyse morphologique dans les systèmes de TALN, plusieurs travaux ont vu le jour. Nous pouvons les classer selon plusieurs critères ; les connaissances linguistiques et le savoir faire informatique qu'ils utilisent.

### 2.3.1 Utilisation des connaissances linguistiques

Le formalisme à deux niveaux est la méthode théorique la plus utilisée dans l'analyse morphologique. Il est basé sur la construction d'un ensemble de transducteurs à états finis dont chacun d'eux encode une règle morphologique particulière, afin de tracer la surface (le motif) d'un morphème donné.

Les deux méthodes d'analyse morphologique suivantes sont basées sur ce formalisme :

- **Morphologie Basée sur les Syllabes** : *Ang.Syllabe-Based Morphology SBM*, elle s'appuie sur les syllabes pour former les motifs des mots qui seront utilisés plutard par le formalisme *transducteur* décrit ci-dessus.
- **Modélisation des racines** : dans cette méthode la racine du mot est extraite en faisant correspondre le mot avec des listes des stems et des affixes.

Nous induisons deux inconvénients de cette approche :

- Le formalisme dépend fortement de la langue.
- La construction manuelle des transducteurs pour chaque règle.

Ce qui laisse le développement d'un tel analyseur morphologique très coûteux [DPDS08].

### 2.3.2 Techniques basées sur des corpus

Elles utilisent des corpus annotés morphologiquement pour construire une base de données au lieu de dépendre de connaissances linguistiques préétablies. Les techniques basées sur des corpus peuvent être utilisés pour fournir une base de données morphologiques qui est utilisée dans le traitement statistique ou dans les techniques d'apprentissage automatique dans l'analyse morphologique.

L'avantage de ces dernières méthodes est qu'elles sont indépendantes des connaissances linguistiques, donc, en principe, elles peuvent être appliquées aux différentes langues.

En outre, les approches *data-driven* dans l'analyse morphologique surpassent celles axées sur des règles construites à la main [DPDS08].

### 2.3.3 Les techniques non supervisées

Récemment, des approches non supervisées ont été explorées dans l'analyse morphologique, basée sur l'utilisation des méthodes reconnues dans l'apprentissage automatique telle que la métrique de la distance d'édition minimale et la recherche de motifs afin de deviner automatiquement les propriétés morphologiques d'une langue, en se basant juste sur des textes brutes non annoté [DPDS08].

Cependant, Harald [HB11] affirme qu'il est impossible de développer un système d'analyse morphologique ne dépendant nullement de connaissance linguistique.

Dans le but de donner aux lecteurs une idée sur la difficulté de l'analyse morphologique non supervisée, nous présentons les résultats donnés par la compétition ***Morphochallenge*** ; ce concours fait appel à compétition de tout algorithme d'apprentissage non supervisé qui peut renvoyer les morphèmes de chaque mot d'une liste donnée contenant des mots de plusieures langues. Par exemple dans Morphochallenge 2009, les langues étaient l'Arabe, l'Anglais, le Finnois, l'Allemand et le Turc. Le meilleur algorithme doit être aussi indépendant que possible de la langue. Tous les mots dans le corpus d'apprentissage sont donnés dans des phrases, pour que l'algorithme puisse utiliser des informations sur le contexte.

- Le corpus d'apprentissage était de 3 millions de phrases pour l'Anglais, le Finnois et l'Allemand,
- et 1 million de phrases pour le Turc dans des fichiers texte sans annotation.
- Le corpus d'apprentissage de l'Arabe était le Coran, ce qui est un petit corpus composé de seulement 78K mots. Le texte du corpus Coran est disponible aux formats voyellé et non voyellé.

Harald [HB11] a donné une définition succincte de l'analyse morphologique non supervisée ; elle n'exige en entrée que des données texte brutes (non annotées, non sélective) en langue naturelle, elle fournit en sortie une description de la structure morphologique de la langue du texte d'entrée, moyennant le très peu de supervision possible. La restriction annoncée ne limite que l'utilisation des connaissance linguistiques, mais elle offre une liberté dans le choix des méthodes à utiliser. De ce fait, on remarque clairement la diversité de techniques proposées dans l'état de l'art de cette approche (voir Figure 2.3).

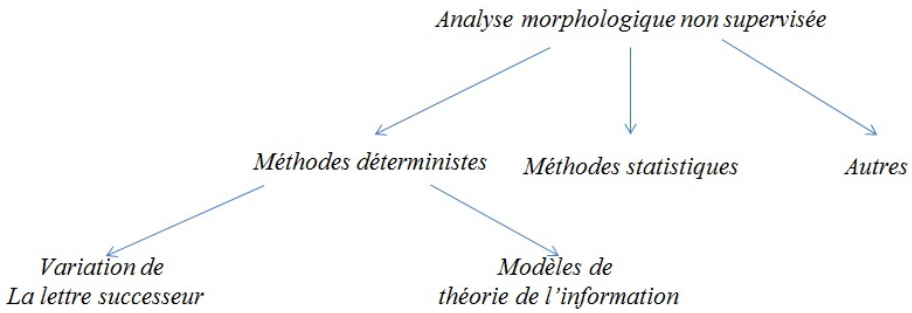


FIGURE 2.3 : Différentes méthodes non supervisées pour l'AM

#### 2.3.3.1 Les méthodes déterministes

Les modèles déterministes définissent les variables d'une manière déterministe, où aucun caractère aléatoire n'est engagé. Les modèles déterministes utilisés pour la segmentation morphologique dans la littérature, sont classés en deux grandes approches : les modèles de variétés de la lettre successeur et des modèles inspirés de la théorie de l'information.

**Les modèles de variation de la lettre successeur** En 1955, Harris [Har55] était le pionnier qui a entrepris cette idée. Par exploiter la distribution des lettres dans un mot. Il identifie les limites des morphèmes en fonction de la variation du caractère successeur après chaque lettre (voir Figure 2.4). Si le nombre de types de successeurs augmenté de manière significative à une position au sein d'un énoncé, alors une nouvelle frontière de morphème est définie.

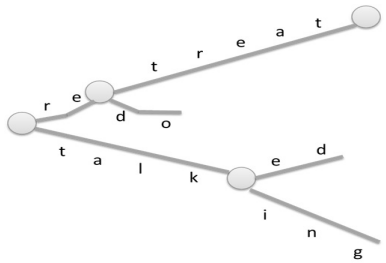


FIGURE 2.4 : Variation de la lettre successeur

- Hafer et Weiss (1974) [HW74] suivent la même idée en utilisant les propriétés statistiques des variation du caractère successeur et prédécesseur pour définir un point de segmentation dans un mot. Ils utilisaient une valeur seuil dans le choix d'un point de segmentation.
- Déjean (1998) [Déj98] utilise un dictionnaire de morphèmes qui se compose des morphèmes les plus fréquents dans un corpus et le processus de segmentation est effectué par ce dictionnaire.

– Bordag (2005) [Bor05] utilise la même approche en intégrant des informations de contexte aux variation du caractère successeur pour éliminer le bruit. Bordag emploie la similitude du contexte pour analyser les mots similaires ensemble. Cela permet d'éliminer une quantité importante de bruit.

**Modèles inspirés de la théorie d'information** Le principe de la longueur de description minimale est largement appliqué dans l'apprentissage automatique de la morphologie. Bien que le principe puisse également être défini dans un cadre probabiliste, il sera présenté comme un modèle de théorie de l'information, en raison de son inspiration de laquelle. Le modèle est longuement évoqué dans les modèles statistiques, en essayant de bénéficier de simple principe disant que « plus les données sont réduites plus on apprend d'elles ».

**Exemple** Le jeu de mots en Arabe suivant (القول، الكتب، يقولون،يفعلون) sera réduit comme suit (و ن،ي،ال،كتب،قول،فعل), dont on peut en déduire que (و ن،ي،ال) sont des affixes.

– Linguistica (Goldsmith 2001, 2006) [Gol06] est un système d'apprentissage non supervisé pour la morphologie. Goldsmith introduit les signatures morphologiques des structures pour coder les données. Une signature représente la structure interne d'une liste de mots qui ont la même morphologie inflectionelle.
– Argamon et al. (2004) [AAAK04] introduit un nouvel algorithme récursif utilisant le principe MDL pour segmenter les mots de façon récursive en suggérant de multiples points de division en un mot.

#### 2.3.3.2 Les méthodes statistiques

La modélisation statistique traite les données comme un phénomène aléatoire lors de leur entrée et même lors de la production des résultats. Contrairement à la modélisation déterministe, une attention particulière est accordée à la modélisation statistique, surtout dans les langues latines, pour lesquelles on trouve un nombre important de travaux basés sur cette méthode.

– Goldwater et al. (2006) [GGJ06] introduisent un modèle à deux étapes dans lequel les mots sont initialement générés par un générateur de composant, puis les fréquences des mots sont estimées par un adaptateur.
– Snyder et Barzilay (2008) [SB08] ont élaboré un autre modèle bayésien non paramétrique qui utilise des corpus parallèles bilingues pour induire des morphèmes fréquentes à l'intérieur de courtes phrases parallèles, au lieu d'induire des morphèmes dans chaque langue individuellement.

#### 2.3.3.3 Autres approches

Il existe des travaux dans cette approche, qui utilisent d'autres techniques qui n'appartiennent pas aux catégories décrites précédemment.

- Neuvel et Fulop (2002) [NF02] proposent un algorithme basé sur les mots. Dans cette approche, les relations morphologiques entre les mots sont définies au lieu d'inciter les morphèmes, afin d'apprendre des formes de nouveaux mots.
- Keshava et Pitler (2006) [KP06] conçoivent un nouvel algorithme appelé *RePortS* qui construit des arbres par les lettres du lexique. Un arbre vers l'avant est utilisé pour les suffixes, tandis que l'arbre vers l'arrière est utilisé pour les préfixes. En utilisant des arbres, ils définissent certains critères basés sur les probabilités conditionnelles, afin d'identifier les suffixes et préfixes en leur affectant des scores. Enfin, les mots sont segmentés en utilisant ces scores.
- Demberg (2007) [Dem07] améliore l'algorithme original des *RePortS* pour les morphologies complexes, en ajoutant une étape supplémentaire à l'algorithme, pour trouver une liste des stems candidats.
- Lavallée et Langlais (2009) [LL09] utilisent des analogies formelles pour trouver la relation entre les différentes formes de mots, Toutefois, en raison de la complexité de recherche des analogies, l'algorithme n'est pas pratique, lorsque les lexiques sont grands.
- Lignos (2009) [LCMY09] améliore une version précédente de leur algorithme en introduisant modèle d'inférence de base qui apprend les formes de base lorsque ces formes n'existent pas dans le corpus

## 2.4 L'analyse morphologique pour la langue Arabe

Dans la première partie de ce chapitre ; nous exposons le niveau morphologique d'une façon générale ; indépendante de la langue, dont on l'a mentionné comme une étape fondamentale dans une chaine de traitement automatique d'un langage naturel ; nous avons présenté des travaux en morphologie computationnelle ; terminologie et opérations fondamentales telles que la dérivation et la troncature des mots en morphèmes, par la suite on a mentionné ses domaines d'application et on a terminé par un recensement et une classifications des méthodes adoptées pour l'analyse à ce niveau linguistique.

### 2.4.1 La langue Arabe

La langue Arabe est la cinquième langue parlée au monde, en nombre de ses locuteurs qui font plus de 422 millions en tant qu'une langue native, et environ 250 millions comme une seconde langue [75] ; Comme son nom l'indique, la langue Arabe est la langue parlée à l'origine par le peuple Arabe, c'est une langue sémitique (comme l'hébreu, l'araméen et le syriaque). Aujourd'hui, Elle est langue officielle d'au moins 22 pays, C'est aussi la langue de référence pour plus d'un milliard de musulmans. L'alphabet Arabe se compose des 28 lettres suivantes : ا , ب , ت , ث , ج , ح , خ , د , ذ , ر , ز , س , ش , ص , ض , ط , ظ , ع , غ , ف , ق , ك , ل , م , ن , و , ه , ى. Contrairement à la langue latine, l'orientation de l'écriture en Arabe est de droite à gauche. Les mots Arabes sont en deux genres, masculin et féminin, et en trois nombres, singulier, duel et pluriel, et trois cas grammaticaux : nominatif, accusatif, génitif.

- nominatif quand il est un sujet,
- accusatif quand il est l'objet d'un verbe,
- et génitif quand il est l'objet d'une préposition.

Les mots sont classés en trois grandes parties du discours, les noms (en incluant les adjectifs et les adverbes), verbes et particules. Le développement de la langue Arabe a été associé à la naissance et la diffusion de l'Islam. L'Arabe s'est imposée, depuis l'époque Arabo-musulmane, comme langue religieuse. De nos jours, aux pays parlant par laquelle c'est la langue formelle de l'administration, du journalisme, de la culture et de la pensée, des dictionnaires, des traités des sciences et des techniques. Ce développement s'est accompagné d'une rapide et profonde évolution (en particulier dans la syntaxe et l'enrichissement lexical). L'Arabe peut être considérée comme un terme générique rassemblant plusieurs variétés :

- **L'Arabe classique** : la langue du Coran, parlée au VIIe siècle.
- **L'Arabe standard moderne (l'ASM)** : une forme un peu différenciée de l'Arabe classique, et qui constitue la langue écrite de tous les pays arabophones. L'ASM reste le langage de la presse, de la littérature et de la correspondance formelle, alors que l'Arabe classique appartient au domaine religieux.
- **Les dialectes Arabes** : malgré l'existence d'une langue commune, chaque pays a développé son propre dialecte. Issus de l'Arabe classique, leurs systèmes grammaticaux respectifs affichent de nettes divergences avec celui de l'ASM. On peut regrouper ces dialectes en quatre grands groupes :

  les dialectes Arabes, parlés dans la Péninsule Arabique : dialectes du Golfe, dialecte du NAJD, Yéménite,

  les dialectes maghrébins : algérien, marocain, tunisien, hassaniya de Mauritanie,

  les dialectes proche-orientaux : égyptien, soudanais, syro-libano-palestinien, irakien (nord et sud),

  la langue maltaise est également considérée comme un dialecte Arabe.

### 2.4.2 L'Arabe standard moderne -ASM-

La langue Arabe est un ensemble complexe dans lequel s'étendent des variétés écrites et orales répondant à un spectre très varié d'usages sociaux. Mais au-delà de cette variété, les sociétés Arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène, d'où l'importance de l'ASM qui forme un terrain commun pour cette large population. L'ASM est la langue des médias officiels, de la communication écrite et de tout type de communication non spontanée. Elle se distingue des dialectes Arabes par son système grammatical partagé avec l'Arabe classique. L'ASM, quoique qu'elle soit considérée comme le symbole le plus puissant de l'unité Arabe, possède des variations régionales.

### 2.4.3 Les challenges de la morphologie en langue Arabe

L'Arabe est une langue morphologiquement complexe et hautement inflexionnelle. Son modèle de dérivation (racine-motif) non concaténative rend les tâches de traitement

au niveau morphologique des textes en Arabe extrêmement difficile que ce soit du côté théorique ou informatique.

Sawalha a recensé quelques caractéristiques de cette langue qui présentent des raisons ou des sources de cette complexité dans son analyse [Saw11] :

#### 2.4.3.1 L'orthographe

Lorsqu'on parle de l'orthographe de la langue Arabe, on cible le script de l'Arabe standard. L'alphabet Arabe se compose de 25 consonnes, 6 voyelles divisées en trois voyelles longues (ا و ي) (a, w, y) et trois voyelles courtes écrites en signes diacritiques (a,u, i), et un Hamzah ء qui a un coup de glotte. En outre, le script Arabe contient d'autres formes de lettres tels que «alif maqsourah ى. Les lettres Arabes changent leurs formes en fonction de leur position dans le mot. Une autre question orthographiques en langue Arabe qui l'utilisation de signes diacritiques au-dessus ou au-dessous des lettres. Ces signes diacritiques sont sukun pour marquer des lettres muettes (c.à.d. absence de voyelle courte) et la gémination, ou l'incorporation Shaddah الشدة pour indiquer une lettre doublée, et tanwin qui est une marque syntaxique en cas d'un nom indéterminée. Hamza ء, taa marbuah ة qui partage les propriétés phonétiques de deux consonnes taa ت et haa ه, elle est aussi utilisé pour marquer des noms singuliers féminins. La maddah المد ou de l'extension d'une lettre composée de Hamza ء et alif ا.

#### 2.4.3.2 La nature Non concaténative

Le modèle de dérivation "racine-et-motif" est non concaténative (ou non linéaire) ses mots résultants sont souvent infixés selon le motif utilisé pour la dérivation et non pas un ajout juste à les deux côtés ; le cas des langues latines. Dans un processus de formation des mots très complexe de racines et de motifs. Des centaines de mots peuvent être dérivé d'une seule racine, en suivant certains modèles. Ces modèles sont des modèles abstraits où les radicaux seront substitués. Un morphème dans la terminologie occidentale est une unité lexicale indivisible "atomique" et le stem est le morphème de base d'un mot. En Arabe, le stem combine la racine et le modèle.

#### 2.4.3.3 Les Clitiques dans l'Arabe

Les clitiques sont des conjonctions, prépositions, particules, qui sont attachés aux débuts et à la fin des mots. On distingue les clitiques de les affixes car on considère les clitiques ne s'ajoutent que par la dérivation inflexionelle pour désigner le genre ou le nombre ou le marquage que le mot doit respecter dans sa position dans la phrase. Mais les affixes (préfixe ou suffixes et même infixe) s'insèrent dans le mot lors de sa formation depuis sa racine.

#### 2.4.3.4 L'ambiguïté

C'est le critère perçu clairement lors du traitement de la langue Arabe, d'où on juge la complexité de cette langue. Il est du à plusieurs facteurs tels que :

- L'assimilation ou l'élision des voyelles : si la racine présente de voyelles longues parmi ses radicaux, souvent elle va être changée ou totalement éliminée lors d'une dérivation, par exemple :**قال** → **قل** \ **يقول**.
- Interaction entre les affixes, les clitiques et les radicaux de la racine.
- L'écriture des textes sans diacritiques.
- L'interaction inter niveaux linguistique
- ...

#### 2.4.3.5 La Ponctuation

La ponctuation a été introduite récemment dans le système d'écriture Arabe. Les textes MSA est caractérisé par l'incohérence et de l'irrégularité dans l'utilisation des signes de ponctuation. En plus de l'introduction tardive de la ponctuation au texte MSA, l'absence d'une prise en charge globale de la ponctuation dans les grammaires Arabes augmente le problème de l'incohérence dans l'utilisation de la ponctuation dans le texte MSA.

## 2.5 Analyseurs morphologiques pour l'Arabe

Dans la littérature, l'analyse morphologique pour la langue Arabe n'a vu le jour que dans les quinze dernières années, peu de travaux ont été réalisés, en comparaison avec les autres langues. Pour analyser les mots Arabes, certains chercheurs ont suggéré à atteindre leurs racines tandis que d'autres suggère de les analyser juste à leurs stems. L'analyse des mots à leurs racines est la plus utile mais elle est la plus difficile, d'autres applications peuvent bénéficier juste des stems telles que des systèmes de RI [ASAK04].

Dans cette section nous donnons une classification des analyseurs morphologiques pour l'Arabe (voir Figure 2.5), ainsi qu'on expose quelques travaux par ordre chronologique. A la fin, on récapitule par la discussion d'une table comparative de tous les travaux recensés.

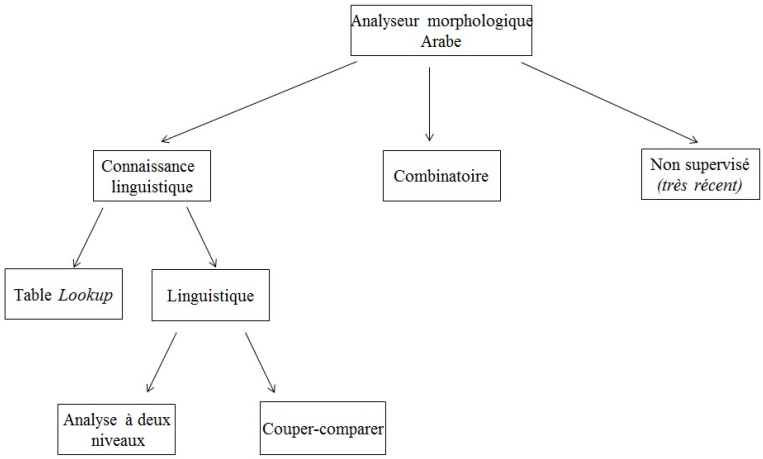


FIGURE 2.5 : Classification des *stemmers* pour l'Arabe

### 2.5.1 Approches basées sur des connaissances linguistiques

Les approches de cette catégorie simulent le comportement d'un linguiste en considérant le système morphologique Arabe par une analyse approfondie des mots en entrée selon leurs composants morphologiques. Dans ces approches, le préfixe et le suffixe d'un mot donné sont éliminées par comparaison des caractères de début et de fin de mot avec une liste d'affixes préétablie. La partie restante est soit acceptée comme le stem requis ou modifiée à l'aide délétion, addition ou substitution de lettres internes pour générer le stem valide, le stem résultant est transformé en une racine simplement par le processus de filtrage en utilisant une liste de motifs valables ou des machines à état finis implémentant ces motifs. La plupart des travaux publiés repose principalement dans cette catégorie.

#### 2.5.1.1 Générateur à états finis pour la morphologie Arabe

Ce système de Xerox [Bee01] traite les textes en Arabe standard moderne . Il accepte comme entrée des textes entièrement voyellés, partiellement voyellés ou non voyellés pour générer en sortie la racine, le modèle, et les affixes du mot analysé avec d'autres attributs tels que : la partie du discours (personne, nombre, la situation, la voix etc). Le système Xerox vise à résoudre trois challenges imposés par la langue Arabe : l'inflexion, les voyelles courtes et les choix dans un lexique Arabe. Le système Xerox est basé sur un lexique de représentation de root-motif, il contient plus de 5000 racines et 400 motifs phonologiquement distincts. Il est basé sur deux analyseurs morphologiques ALPNET et Xerox calculus à états finis qui a été utilisée pour insérer les radicaux d'une racine dans des motifs dans un processus de génération, qui a pu effectivement générer 85.000 stems valides. Le lexique des transducteurs contient également les préfixes et les suffixes appropriés qui sont ajoutés à des stems par concaténation.

**Avantages**

– Une bonne couverture du vocabulaire Arabe,

– la reconstruction des voyelles courtes,
– et le petit glossaire en Anglais fourni pour chaque mot.

**Inconvénients**
– Sur génération en dérivation des mots en raison de la répartition exhaustive des motifs pour les racines
– et le haut niveau d'ambiguïté où il produit de nombreuses analyses pour la plupart des mots.

#### 2.5.1.2 le Stemmer de Shereen Khoja

Le Stemmer de Khoja [Kho01] supprime le suffixe et le préfixe le plus long, Il compare ensuite le mot restant avec des motifs des verbes et des noms, pour en extraire la racine. Le stemmer fait appel à plusieurs fichiers de données linguistiques ; une liste de tous les caractères diacritiques, des caractères de ponctuation, des articles définis, et 168 mots vides (Larkey et Connell, 2001) [LBC02]. En résumé, le stemmer de Khoja performe trois procédures :

1. L'enlevement des affixes les plus longs,
2. La recherche du motif correspondant au reste du mot pour en extraire la racine,
3. La verification de la racine obtenue via une liste des racines valides.

#### 2.5.1.3 L'analyseur morphologique de Tim Buckwalter

Tim Buckwalter [Buc02] a développé un analyseur morphologique pour l'Arabe. Il a compilé trois lexiques simples ; préfixes, suffixes et stems, en prenant compte des voyelles courtes et des signes diacritiques dans ces lexiques. L'analyseur (i) divise le mot en trois parties, en prenant toutes les possibilités(le préfixe, le stem, le suffixe), ensuite, (ii) il vérifie la validité de chaque segmentation en utilisant les trois dictionnaires cités auparavant (préfixe, stem et suffixe), ainsi que trois autres tableaux contenant des paires des catégories morphologiques afin de (iii) vérifier la compatibilité.

#### 2.5.1.4 Morphologie fonctionnelle pour la langue Arabe

ElixirFM [Smr07] est une implémentation d'un nouveau modèle computationnelle des processus morphologiques en Arabe moderne. Il est lié à "*Prague Arabic Dependency Treebank (PADT)*" (Hajic et al 2004) [HSZ$^+$04], (Smrz et al 2008) [SBK$^+$08]. ElixirFM fournit à l'utilisateur quatre fonctionnalités différents :
– **Résoudre** fournit la segmentation et l'analyse morphologique du texte inséré,
– **Inflect** transforme les mots dans des formes requises par le contexte.
– **Dériver** convertit les mots en leurs homologues ayant un sens similaire, et une différente catégorie grammaticale
– **Lookup** cherche des entrées lexicales par la forme de citation et la racine.

**Avantages** l'utilisation de motifs morpho-phonologique"الجذور مع الأوزان" qui permettent d'éviter la conception des règles spéciales pour éviter les problèmes d'assimilation (suppression de quelques caractères du mot) et l'énumération des formes de r chaque item lexical.

**Inconvénient** La taille du lexique des motifs morpho-phonologique qui est de 4290 entrées, ce qui affaiblit la couverture du vocabulaire.

#### 2.5.1.5 Analyseur morphologique pour l'Arabe *MIDAD*

MIDAD [SAM09] utilise des connaissances linguistiques de la morphologie Arabe pour développer des algorithmes informatiques qui simulent des méthodes et des règles humaines pour générer et analyser les mots.

L'analyseur utilise une base de données des racines Arabes et des mots irréguliers qui ont besoin d'un traitement spécial. Cette base de données peut être utilisé pour générer une autre base de données plus grande qui comprend plus de vocabulaire Arabe par l'utilisation des racines et des mots irréguliers déjà requis ce qui rend le programme petit, rapide et robuste.

#### 2.5.1.6 Analyseur morphologique pour l'Arabe *MORPH2*

MORPH2 [KBH10] est un analyseur morphologique pour la langue Arabe, il est une extension de MORPH (Hadrich et CHAABEN 2006) [BC06]. Cette extension avait pour but l'ajout d'une étape de la vocalisation. MORPH2 utilise un modèle standard de la morphologie Arabe ; le modèle interprète toutes les règles possibles qui régissent la dérivation d'un mot de son morphème (root), MORPH2 prend en compte les questions orthographiques des mots Arabes tels que l'incorporation, la substitution, la vocalisation et l'omission. Les entrées sont des mots soit entièrement voyellés, partiellement voyellés ou non voyellés.

Les sorties sont stockées dans un fichier XML ou. Feuille de style XSL dans un format structuré. MORPH2 dépend d'une liste préétablie de motifs et de modèles générés pour traiter les cas de substitution et de vocalisation. L'analyse de mots est effectuée en cinq étapes :

1. **Tokenisation** : elle est basée sur l'exploration contextuelle de ponctuation qui divise le texte en phrases, puis la détection de mots dans les phrases.

2. **Prétraitement morphologique** : elle extraits les clitiques des mots analysés. Ensuite, un processus de filtrage classifie les stems des mots analysé, les particules, le numéro, la date ou nom propre.

3. **Analyse des affixes** : elle décèle les éléments fondamentaux de la parole, à savoir : racines et affixes. Ce processus est accompli à son tour en cinq étapes : (i) la suppression des éventuels préfixes et suffixes, (ii) l'identification des affixes candidats ; (iii) le filtrage lexical, (iv) le contrôle d'association de radicaux de la racine et les

affixes, (v) et la reconnaissance.

4. **Analyse morphologique** : détermine toutes les fonctions morpho-syntaxiques possibles du mot analysé, en trois phases : (i) l'identification de la partie du discours du mot (nom, verbe ou particules), (ii) l'identification des caractéristiques morphologiques (le genre, le nombre, le temps et la personne), et (iii) le filtrage des listes de caractéristiques.

5. **vocalisation et validation** : dépend des deux étapes précédentes. La vocalisation du mot analysé est effectué en fonction des caractéristiques morpho-syntaxiques et en faisant correspondre le mot analysé avec son motif. Le processus de validation sur les opérations de transformation, d'omissions et d'assimilation qui se font sur les mots analysés.

#### 2.5.1.7 Analyseur morphologique AlKhalil

Alkhalil Morpho Sys [BLAA11] est un analyseur morphologique pour l'Arabe standard moderne. Alkhalil traite les textes non voyellés, partiellement et entièrement voyellés. Il est basé sur la modélisation d'un très grand ensemble de règles morphologiques Arabes et sur l'intégration des ressources linguistiques qui sont utiles à l'analyse, tels que :

1. la base de données racine,
2. les motifs morpho-phonologique voyellés associé aux racines,
3. et les listes des enclitiques et proclitiques.

Les résultats de l'analyse des mots Arabes sont présentés dans un tableau qui montre : le stem entièrement voyellé, sa catégorie grammaticale avec des attributs morphosyntaxiques dans les phrases, ses racines possibles associés à des motifs correspondant et ses proclitiques et enclitiques.

Aucune évaluation consistante n'a été signalée à cause de l'indisponibilité d'un corpus de test. Une évaluation a été réalisée pour montrer la capacité du système à analyser des mots, en examinant les résultats de Alkhalil sur un échantillon du Coran, Chapitre 20, qui compte environ 1000 mots. Les sorties d'Alkhalil ont montré qu'environ 13,37% (132 mots sur 987 mots de l'échantillon) n'ont pas d'analyse. La plupart de ces mots non-analysés sont des mots vide ou des noms propres.

### 2.5.2 Approches combinatoires

Dans les approches combinatoires toutes (ou une bonne partie) les combinaisons de lettres générées du mot sont testés ; elles sont comparées à une liste de racines valides. Selon la correspondance avec une racine valide, on extrait le stem et les modèles. Le processus de comparaison est exhaustive, s'il n y a aucune correspondance, d'autres combinaisons seront testés.

#### 2.5.2.1 L'algorithme d'extraction de la racine tri-littérale

Al-Shalabi, Kanaan et Al-Serhan [ASASK03] ont développé un algorithme d'extraction de racine qui n'utilise aucun dictionnaire. Il s'appuie sur l'attribution des pondérations aux lettres d'un mot multiplié par leurs positions, les consonnes ont des poids différent de zéro et des différents poids sont attribuées aux caractères groupés dans le mot سألتمونيها où tous les affixes sont formés par des combinaisons de ces lettres. L'algorithme sélectionne les lettres ayant les plus faibles poids comme des radicaux de la racine.

#### 2.5.2.2 Analyseur de la morphologie Arabe orienté application

L'analyseur de Sabir et al. [SAM09] dépend d'un algorithme qui classe les lettres du mot en lettres appartenant à des affixes ou des lettres sousjacentes. L'algorithme applique les règles qui régissent les relations entre les lettres du mot. L'algorithme ne dépend pas des dictionnaires préenregistrés, juste l'algorithme peut extraire le stem, les affixes et le motif du mot analysé. Les entrées sont soit entièrement voyellés mots, des mots partiellement voyellés ou non voyellés. La sortie est toutes les possibles racines, affixes et les motifs du mot analysé.

Ils rapportent un taux de précision de 97,7% et ils prétendent que l'analyseur est cinq fois plus rapide que n'importe quel analyseur existant. Comme indiqué, l'analyseur peut être intégré dans d'autres applications et des parties de l'analyseur peuvent être réutilisés comme par exemple dans le travail de Sonbul, Ghnaim et Dusouqi (2009) [SD11].

### 2.5.3 Approches non supervisées

Comme il est cité précédemment (voir section 2.3.3), ces approches sont récentes même dans les langues latines, a fortiori la langue Arabe, elles utilisent principalement les techniques de l'apprentissage automatique et les statiques afin d'avoir des modèles et des algorithmes pouvant performer l'analyse morphologique des textes bruts sans avoir besoin de la moindre connaissance linguistique ni l'intervention d'experts.

En effet, le seul travail récent que l'on a pu trouvé comme analyseur non supervisé pour la langue Arabe est le travail de A. Khorsi [Kho12], ci-dessous reporté.

#### Stemming non supervisé pour l'Arabe classique

Le but de Khorsi [Kho12] était d'extraire les radicaux d'une racine du mot analysé d'une manière non supervisée, qui n'appui sur aucune connaissance linguistique préalable, l'entrée de son algorithme est un texte brute. Dans un premier temps, il a travaillé sur l'extraction des stems, dont l'auteur a proposé la définition suivante "*un stem est le plus petit segment du mot qui contient tous les radicaux de sa racine*", la méthode de Khorsi a été appliqué sur l'Arabe classique dont il a utilisé un corpus du Coran, en suite il a testé ses résultats par un lexique de tous les termes du Coran créés à la main, par un groupe des experts. Les étapes principales de son stemmer sont :

1. l'extraction de tous les segments possibles d'un mot,
2. la mesure de dépendance inter lettre pour chaque segment extrait, en utilisant les probabilités conditionnelles,
3. le choix du segment ayant la dépendance maximale comme un stem du mot analysé.

### 2.5.4 Discussion

TABLE 2.4 : Grille récapitulative des stemmers étudiés pour la langue Arabe

| Classe | Auteurs nom d'algo et dates | Connaissances utilisées | Type de texte à traiter et résultat | Précision% |
|---|---|---|---|---|
| Linguistique | Beesley, Kenneth R, Xerox, 1998 | Listes : Motifs, stems et affixes & TAD : machine à états finis | Textes d'entrées : voyellés, partiellement voyellés ou non voyellés & Résultats : Motifs, stems et affixes | non mentionnée |
| | Khoja Shereen et Garshide, 1999 | Listes : Stop words, motifs et racines | Textes d'entrées :voyellés ou non voyellés & Résultats : Racine | 67.32% |
| | Buckwalter Tim, 2002 | Listes : Préfixe, stem, suffixe et tables de compatibilité | Textes d'entrées : non voyellés & Résultats : préfixe, stem et suffixe | 39.30% |
| | Smrž Otakar, ElixirFM, 2007 | Listes : motifs, stem, mots | Textes d'entrées : non voyellés & Résultats : préfixe, stem et suffixe | non mentionnée |
| | Sabir, M. et al. MIDAD, 2009 | Listes : racines et mots irréguliers | Textes d'entrées : voyellés ou non voyellés & Résultats : préfixe, stem et suffixe | non mentionnée |
| | Kammoun, N. et al. MORPH2, 2010 | Listes : Préfixe, stem, suffixe et tables de compatibilité | Textes d'entrées : voyellés ou non voyellés & Résultats : préfixe, stem et suffixe | non mentionnée |
| | Boudlal, A. et al. ALKHALIL, 2010 | Listes : motifs, racines, affixes et un ensemble de règles morphologiques | Textes d'entrées : voyellés, partiellement voyellés ou non voyellés & Résultats : préfixe, stem et suffixe | non mentionnée |
| Combinatoire | Al Shalabi et al. 2003 | Pondération statistique des lettres | Textes d'entrées : non voyellés & Résultats : racine trilittérale | 64.37% |
| | Sonbul, R. et al. 2011 | Pré-catégorisation des lettres | Textes d'entrées : voyellés, partiellement voyellés ou non voyellés & Résultats : le stem, les affixes et le motif | 90% |
| Non supervisée | Khorsi Ahmed, 2012 | Pondération statistique des segments du mots | Textes d'entrées : non voyellés & Résultats : Stems | 90% |

Après avoir recensé quelques travaux (parmi les plus intéressants)dans l'analyse morphologique de la langue Arabe, nous avons remarqué que la majorité des travaux réalisés se basent sur l'utilisation des connaissances linguistiques, que ce soit par l'utilisation des lexiques ou des machine à états finis. En outre, on remarque que les approches combinatoires utilisent aussi des connaissances linguistiques mais un peu allégées par rapport à la première catégorie, pour ceux là la connaissance est représentée par des informations sur les combinaisons des lettres, qui sont obtenues à partir des corpus annotés dont les composants de chaque mot sont déjà étiquetées, ou à l'aide d'une inférence des listes préétablies des affixes et des racines.

Contrairement aux premières méthodes, les algorithmes non supervisés n'utilisent aucune connaissance linguistique préalable pour faire l'analyse, ils manipulent les données en entrée après un apprentissage réalisé sur des textes bruts.

La meilleure précision est encore monopolisés par les techniques basées sur des connaissances linguistiques, malgré la lenteur lors de leurs exécution vu le parcours des lexiques. De plus, elles sont coûteuses en terme de ressources requises. Cependant, les méthodes combinatoires sont plus rapides que les premières approches grâce aux peu de ressources requises ; juste des tables de pondération ou des petits lexiques de validation.

La nouvelle approche vise à réduire les coûts de l'analyse morphologique par l'élimination de l'exploitation de toutes connaissances linguistiques, en la remplaçant par un apprentissage -doit se faire en amont- sur des textes bruts (Corpus). Cet apprentissage offre la connaissance requise pour l'algorithme non supervisé qui fait la segmentation. Ce genre d'algorithme a prouvé ses bonnes performance dans les langues latines, et il est en cours de progression pour la langue Arabe.

Il s'avère que le recours à la dernière catégories a de bons arguments ; réduction des coûts en terme d'espace de stockage surtout pour les systèmes embarqués et en temps d'exécution surtout pour les applications en temps réel, ainsi que ses techniques nous font éviter la tâche laborieuse de préparation des ressources linguistiques et la consultation des experts du domaine.

## 2.6 Conclusion

Nous avons rappelé et défini l'analyse morphologique en tant qu'une étape dans une chaîne de TALN, on a fixé la terminologie de ce domaine, et les opérations fondamentales qu'elle effectue ; elle s'intéresse par la formation (comment générer ?) et l'analyse (comment reconnaitre ?) de tous ce qui en langue naturelle, en donnant quelques applications TALN ayant une relation directe par cette analyse, comme on a démontré l'interaction de la morphologie avec les autres niveaux linguistiques.

Après avoir perçu la morphologie computationnelle, nous avons commencé à explorer les approches existantes dans la littérature, où on les a classé en : basée sur des connaissances linguistiques, combinatoires et des techniques non supervisées, dont on s'est étalé sur l'explication de la dernière classe, où notre travail appliqué en langue Arabe fait partie.

Nous avons consacré la dernière partie de ce chapitre aux spécificités et aux travaux de la langue Arabe, nous l'avons défini et on a donné une description brève de son état actuel, en suite on a démontré les challenges qu'elle impose en tant qu'une langue riche morphologiquement, à la fin de ce chapitre, nous avons recensé les travaux investis dans l'analyse morphologique de la langue Arabe.

En effet, les performances atteintes par les analyseurs morphologiques pour la langue Arabe sont encore peu en nombre et moins performantes par rapport aux autres langues. En profitant de la puissance de l'apprentissage automatique, nous avons fait une tentative de développer un modèle statistique pour la morphologie de cette langue et l'utiliser dans une analyse non supervisée, ce qui est l'objet des deux chapitres suivants.

# Chapitre 3

# Analyse morphologique pour la langue Arabe via les statistiques

Après la description des notions et des concepts liés à la morphologie de la langue Arabe. Il s'est avéré bien clair que le domaine nécessite encore des efforts pour améliorer son état actuel de point de vue précision et ressources requises, notamment dans les applications nécessitant plus de rigueur et d'exactitude. En effet, ces lacunes nous ont poussé à continuer de travailler à ce niveau-là du TALN, dans le but de promouvoir les résultats actuels et offrir un outil d'analyse standard générique, en éliminant toute exploitation d'une ressource ou connaissance linguistique.

La complexité de la langue Arabe est aussi un autre facteur stimulant pour notre recherche dont on a voulu contrecarrer la nature non concaténative de cette langue.

Nous décrivons dans ce chapitre le modèle statistique conçu pour la morphologie de l'Arabe (*SMM*), ensuite, nous exposons une application non supervisée pour l'extraction des racines tri-littérales en exploitant le modèle *SMM*.

## 3.1 Le modèle statistique morphologique *SMM*

L'objectif des modèles de langue dans la recherche d'information *RI* et la reconnaissance de la parole *ASR* est de capturer les régularités linguistiques par l'estimation de la probabilité d'avoir un mot après avoir observé telle séquence de mots, et juger si elle est valide ou non, afin de fournir un jeu d'information adéquat à la requête pour la première (*RI*), et de rendre une reconnaissance cohérente à la deuxième (*ASR*).

Par analogie à cette modélisation statistique ayant les mots comme échelle ; unités évènementielles, notre modèle entreprend à une échelle plus fine, les lettres comme des unités afin d'étudier la structure morphologique et prédire la segmentation correcte d'un mot et ainsi répondre aux besoins de plusieurs applications(voir Figure 3.1) [LH13].

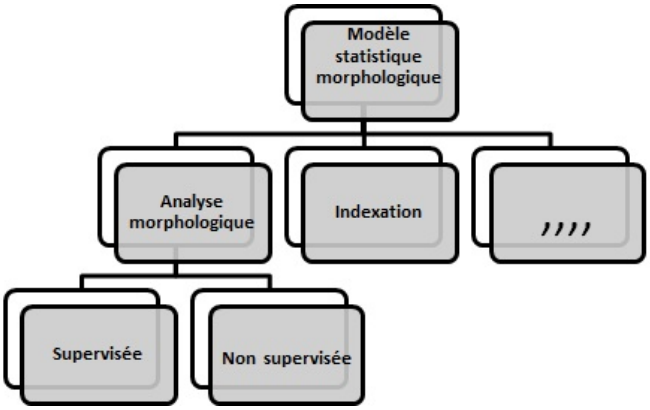


FIGURE 3.1 : Applications du SMM

Le modèle proposé est basé sur les probabilités conditionnelles, qui sont définies par le théorème de Bayes :

$$P(A/B) = P(A,B)/P(B) \tag{3.1}$$

Il définit la probabilité d'avoir l'événement $A$ sachant l'événement $B$ par le quotient des probabilités d'avoir les deux événements $AB$ en même temps sur l'apparition de l'événement $B$. Nous projetons ces règles au phénomène à modéliser ; la morphologie d'une langue naturelle, et plus précisément la morphologie des mots en langue Arabe. Les symboles $AetB$ présentent les évènements d'apparition des lettres dans les mots en respectant l'ordre.

**Exemple**

$$P(\text{س / در}) = P(\text{درس})/P(\text{در})$$

Pour développer un tel modèle statistique, il nous fallait répondre à ces deux questions :

1. Quelle méthode de découpage utiliser ?
2. Quelles structure et paramètres considérer pour faciliter l'usage du modèle ?

### 3.1.1 Quelle méthode de découpage utiliser ?

La segmentation d'un mot $w$ selon la structure conventionnelle :

$$\langle Prefixe \rangle^{\star} \langle Stem \rangle \langle Suffixe \rangle^{\star}$$

utilisée dans les différentes méthodes d'analyse morphologique supervisées ne peut être utilisée vu que la connaissance linguistique sur les préfixes et les suffixes est absente dans le contexte d'une analyse non supervisée.

En effet, une segmentation aléatoire peut mener à des résultats erronés et les segments qui seront obtenus peuvent être de simples morphes sans sémantique. Cette situation est illustrée par l'exemple suivant.

**Exemple**

ال \ منظ \ مة
وال\ ديه
م\ طرها

Par conséquence, il est judicieux d'explorer les méthodes de segmentation par l'utilisation des sous chaines ou des sous séquences.

#### 3.1.1.1 Découpage par sous chaines

Le découpage du mot en sous chaines qui fait appel à la méthode standard de facto *les N-grammes* génère une distribution de probabilité des lettres ; par la prédiction de la prochaine lettre à partir des $(N-1)$ lettres antécédentes.

Cette méthode, largement utilisée, a montré de bonnes performances dans l'analyse morphologique pour les langues Indo-européennes qui sont par définition des langues concaténatives dérivables [MS99]. Par contre, pour les langues sémitiques -non concaténatives dérivables- elle a montré moins de performances [Khr06]. Ceci est justifié par le fait que les stems peuvent contenir des infixes. Par exemple la racine du mot **مجاهدة** ne peut être obtenue par une simple troncature du préfixe **م** et du suffixe **ة**. De ce fait, nous avons fait appel à un découpage par sous séquences.

#### 3.1.1.2 Découpage par sous séquences

Une sous-séquence est une suite de lettres présentes dans la séquence initiale, en respectant l'ordre. Contrairement à une sous chaine, les éléments d'une sous séquence ne sont pas forcément consécutifs ou concaténés dans le mot initial. Un tel découpage est plus général que le précédent. En effet, il supporte la structure des mots concaténatives dérivables, ayant seulement des affixes au début et à la fin des stems. De plus, il supporte la structure des mots non concaténatives des langues sémitiques, comportant des infixes.

**Exemple**

Pour le mot **مجاهد** les sous séquences de longueurs 3 contiennent la racine tri-littérale

| | |
|---|---|
| Mot brut | **مجاهد** |
| Sous chaines | **مجا \ جاه \ اهد** |
| Sous séquences | **مجه \ جهد** |

**Définition**

Une sous-séquence peut être définie formellement par : *Soit* $w = w_1 w_2 \dots w_l$ *un mot de l lettres, une* ***n-m-sous-sequence*** *est définie comme une sous-chaine de n lettres avec m sauts.*

**Exemple**

1. $w$= "école", toutes les $3-1-sous-sequence$ du mot "w" sont écl, éol, cle, coe.
2. $w$= **الكتاب**, toutes les $3-1-sous-sequence$ du mot "w" sont **التاكت ،لكت ، لتا ،كاب ،كتب** .

Par ce découpage, on peut obtenir la racine correcte "**كتب**".

## 3.1.2 Paramétrage et structure du modèle

Dans notre étude, nous nous sommes contentés d'utiliser des $3-1-sous-sequences$. Ce double choix de ne considérer que les sous-séquences de taille 3 et un seul infixe de taille 1 est justifié par les faits suivants :

1. 84% des mots en langue Arabe sont dérivés des racines de trois lettres [EZ01].
2. Les mots ayant une seule lettre comme lettre infixe représentent le cas le plus fréquent [EZ01]. En effet, nous avons omis le traitement des cas des mots multi-infixes tels que les mots **مفاهيم، حاسوب** qui contiennent deux lettres infixes non contiguës, et le mot **عواقب** qui contient deux lettres infixes contiguës.

Nous avons aussi considéré l'emplacement ou la position de la première lettre de la sous-séquence dans le mot, comme une dimension supplémentaire au modèle, le nom de cette dimension est $W_{pos}$, les valeurs qu'elle peut prendre sont :

- **Début** : la première lettre de la sous-séquence coïncide avec le premier caractère du mot.
- **Fin** : la dernière lettre de la sous-séquence coïncide avec le dernier caractère du mot.
- **Mid** : le reste de toutes les sous-séquences du mot.

En fait, le rajout de cette troisième dimension est justifié par l'observation attentive de la segmentation conventionnelle et qui nous a mené à l'idée intuitive suivante :

"*Plus le morphème est au centre du mot, plus il est susceptible d'être la racine et lorsqu'il est proche des extrêmes, il présente un éventuel affixe*".

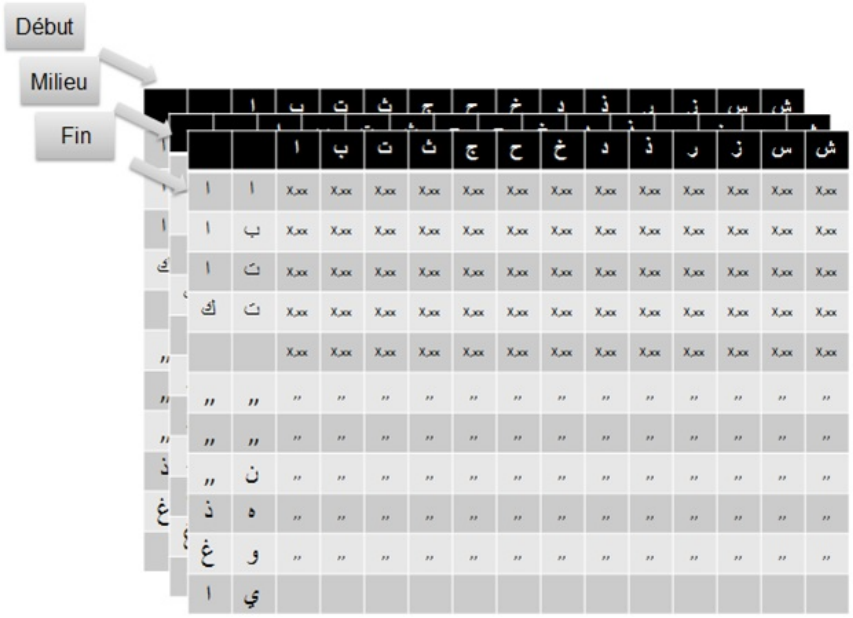


FIGURE 3.2 : Structure du Modèle Statistique de Mot : $SMM$

La figure 3.2 illustre une éventuelle structure du modèle. Où une entrée $SMM[Hist, Pred, W_{pos}]$ dans cette table à trois dimensions est la probabilité de rencontrer la lettre $Pred$ sachant l'historique $Hist = Hist_1\ Hist_2$ à la position $W_{pos}$ dans le mot.
Nous avons considéré que la lettre $Pred$ dans les $3-1-sous-sequences\ s$ peut prendre tout emplacement (comme l'indique la figure 3.3) :

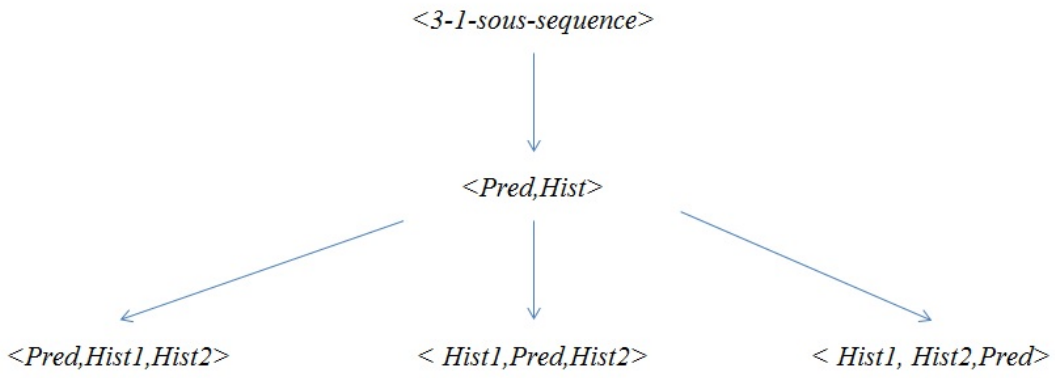


FIGURE 3.3 : Différentes positions de la lettre à prédire

**Exemple**

| | | | | |
|---|---|---|---|---|
| Post-lettre à prédire | $SMM_{Wpos}\,[Hist_1\ Hist_2\ Pred\ W_{pos}]$ | $=$ | $SMM_{Wpos}$ | [س / ر د] |
| In-lettre à prédire | $SMM_{Wpos}\,[Hist_1\ Pred\ Hist_2\ W_{pos}]$ | $=$ | $SMM_{Wpos}$ | [ر / س د] |
| Pré-lettre à prédire | $SMM_{Wpos}\,[Pred\ Hist_1\ Hist_2\ W_{pos}]$ | $=$ | $SMM_{Wpos}$ | [د / س ر] |

Les lignes sont interprétées respectivement comme suit :

1. La probabilité de rencontrer la lettre س après avoir lu les lettres د et ر consécutivement.
2. La probabilité de rencontrer la lettre ر après avoir lu les lettres د puis س en sautant une lettre.
3. La probabilité de rencontrer la lettre د avant de lire des lettres ر et س consécutivement.

Rappelons que l'on ne peut juger la validité et mesurer les performances de ce modèle ainsi conçu que par son exploitation dans une application d'analyse morphologique. Pour cette raison, nous avons conçu un algorithme d'extraction de racine tri-littérale. La description de ce dernier fait l'objet de la section suivante.

## 3.2 Analyseur morphologique non supervisé

En général, 84% des mots en langue Arabe sont dérivés d'une racine de trois lettres nommée racine tri-littérale **جذر ثلاثي** [EZ01].

Compte tenu de ce fait, la structure conventionnelle d'un mot $w$ en langue Arabe peut être définie par l'expression régulière suivante :

$$w = \langle Prefixe \rangle^{\star}\, r_1\, \langle Infixe \rangle^{\star}\, r_2\, \langle Infixe \rangle^{\star}\, r_3\, \langle Suffixe \rangle^{\star} \tag{3.2}$$

Où $R = r_1 r_2 r_3$ représente la racine tri-littérale dont le mot $w$ est dérivé, $r_1$, $r_2$ *et* $r_3$ sont les lettres radicales constituant cette racine, $Prefixe$ et $Suffixe$ sont les éventuels affixes et $Infixe$ sont les éventuels infixes.

Le but de notre analyseur morphologique est de capter cette structure d'une façon non supervisée via le modèle statistique conçu $SMM$.

Le principe de notre algorithme est basé sur l'idée intuitive : exploiter la corrélation inter-lettres. En effet, pour extraire la racine tri-littérale d'un mot $w$, dans une première étape, on crée l'ensemble des sous séquences de trois lettres $S_{seq}$. Dans cette segmentation on essayera d'obéir à la structure infixée de la langue Arabe (Expression 3.2). Dans une seconde étape, on utilise une fonction $Score$ qui affecte à chaque sous séquence $seq$ de $S_{seq}$ un poids. Cette fonction instrumentera le modèle $SMM$. Ainsi, la racine prédite du mot $w$ correspondra à la sous séquence ayant le meilleur score.

### 3.2.1 Segmentation d'un mot $w$

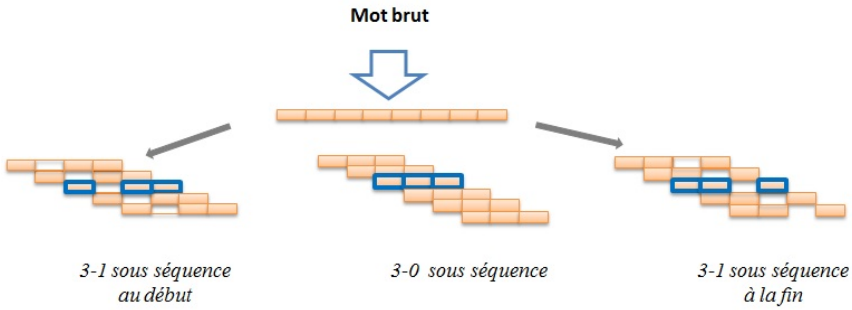


FIGURE 3.4 : Principe de segmentation

La figure 3.4 illustre l'opération de segmentation d'un mot $w$ en vu de capter sa racine. Le nombre de sous séquences de trois lettres d'un mot $w$ de longueur $l$ peut être très grand, pour cela nous avons extrait seulement trois différents types des sous séquences. En effet, la racine tri-littérale d'un mot $w$, peut être éventuellement un élément de l'un des sous ensembles suivants :

**3-1 sous séquences** : ensemble des sous séquences de trois lettres avec une lettre infixe après la première lettre radicale $r_1 =$ ف.

**3-0 sous séquences** : ensemble des sous chaines de trois lettres, où l'on considère que les racines éventuelles ne contiennent pas d'infixes.

**3-1 sous séquences** : ensemble des sous séquences de trois lettres avec une lettre infixe avant la dernière lettre radicale $r_3 =$ ل.

Pour chaque sous séquence $seq$, on garde l'information sur sa position dans le mot $seq_{Wpos}$. L'ensemble $S_{seq}$ sera l'union des ces trois sous-ensembles.

$$|S_{seq}| \leq 3\,(|W| - 3) + 1$$

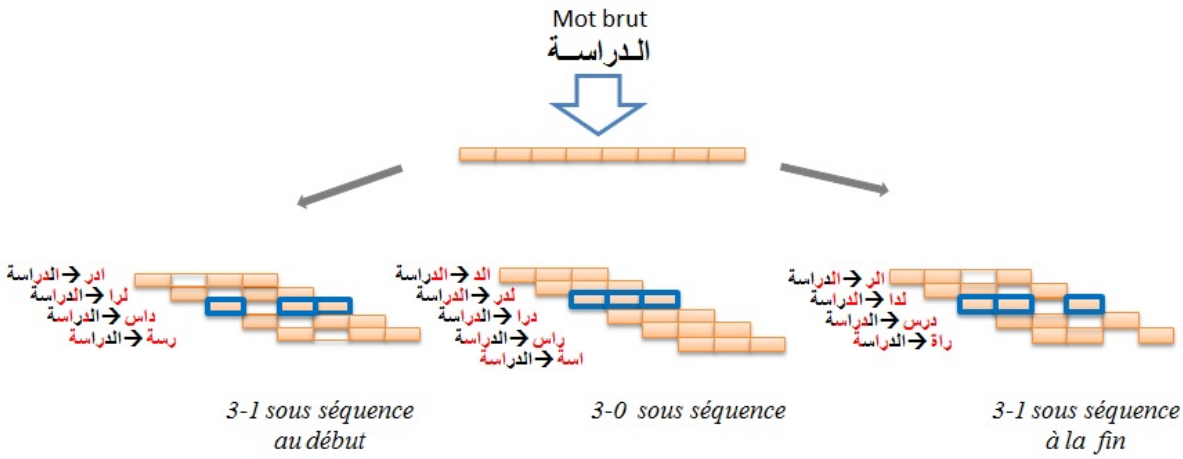


FIGURE 3.5 : Exemple de segmentation du mot **الدراسة**

La figure 3.5 illustre une exemple de segmentation avec le mot **الدراسة**, où les sous ensembles engendrés sont :

***3-1 sous séquences au début*** {(**رسة** , 4),(**داس** , 3),(**لرا** , 2),(**ادر** , 1)}

***3-0 sous séquences*** {(**اسة** , 5),(**راس** , 4),(**درا** , 3),(**لدر** , 2),(**الد** , 1)}

***3-1 sous séquences à la fin*** { (**راة** , 4),(**درس** , 3),(**لدا** , 2),(**الر** , 1)}

L'ensemble $S_{seq}$ pour le mot **الدراسة** est : {(**رسة** , 4),(**داس** , 3),(**لرا** , 2),(**ادر** , 1)(**اسة** , 5),(**راس** , 4),(**درا** , 3),(**لدر** , 2),(**الد** , 1),(**راة** , 4),(**درس** , 3),(**لدا** , 2),(**الر** , 1)}

### 3.2.2 Fonction objectif : *Score*

Afin de calculer la fonction *Score*, nous avons mesuré la corrélation inter-lettres de chaque sous séquence. Nous avons aussi raffiné cette dépendance par l'information sur la position du début de la sous séquence dans le mot.

Le dernier raffinement vise à renforcer la chance de la sous séquence ayant les valeurs les plus élevées, et ceci par la comparaison de corrélations d'un seul morphème dans des emplacements différents, pour en déduire le plus stable.

#### 3.2.2.1 Dépendance inter-lettres

Pour chaque sous séquence résultante de la première étape, on a estimé son degré de dépendance, qui est obtenu en utilisant le modèle SMM. En effet, pour chaque sous séquence dans l'ensemble $S_{seq}$, on procède comme suit :

1. Subdiviser la sous séquence prise $seq = s_1 s_2 s_3$ en deux partie $\langle Hist\ Pred \rangle$, en prenant toutes les combinaisons possibles (voir Figure 3.4), pour la faire correspondre à une entrée du modèle SMM.
2. Chercher et conserver la valeur de dépendance $Dep\langle seq \rangle$ entre les deux parties constituant la sous séquence $seq$ dans $S_{seq}$, en prenant tous les cas possibles (Pre, In et Post-lettre) suivant la formule 3.3.

$$Dep\langle seq \rangle = SMM[s_1\ s_2,\ s_3] + SMM[s_1\ s_3,\ s_2] + SMM[s_2\ s_3,\ s_1] \tag{3.3}$$

#### 3.2.2.2 Pondération

Après l'extraction et la mesure des dépendances des sous séquences extraites du mot, on peut induire les sous séquences les plus dépendantes, qui peuvent être d'éventuels morphèmes corrects, de celles les plus faibles qui peuvent être des segments intermédiaires ; entre deux morphèmes corrects comme le segment **منظ** qui présente un morphe entre les deux morphèmes **نظم** et **م** le mot **منظمة**, mais ce n'est pas suffisant encore pour juger que le découpage est cohérent.

**Facteur de position FP**

Maintenant, on introduit le facteur de position, en croyant l'idée intuitive suivante "*Plus le morphème est au centre du mot, plus il est susceptible d'être la racine, et lorsqu'il est proche aux extrêmes, il présente un éventuel affixe*".

D'abord, on vérifie la longueur du mot $w$ est supérieur à quatre, pour voir s'il mérite cette pondération, car plus le mot est court plus cette opération n'aura pas d'effet. L'algorithme 1 ci-dessous montre comment on procède à la pondération selon la position de commencement de toutes sous séquence $seq_{wpos}$ dans le mot $w$. Le coefficient de pondération (0.25) utilisé a été désigné empiriquement.

**Algorithme 1** Facteur de position ($seq$ , $w$)

**Début**
1. Milieu = $|w|$ *div* 2 ;
2. Facteur = 1 ;
3. **Si** ($|w|$ >= 4) **alors**
4. **Si** ($seq_{wpos} < Milieu$) **alors**
5. Facteur =$0.25 \cdot seq_{wpos}$ ;
6. **Sinon**
7. Facteur =$0.25 \cdot (|w| - seq_{wpos} + 1)$ ;
8. **Finsi**
9. **Finsi**
10. retourne **Facteur** ;
11. **Fin**

Il est à noter qu'il ne faut pas confondre entre $wpos$ et $W - pos$ ; $wpos$ désigne la position du début de la sous séquence dans le mot $w$, cependant la $W - pos$ est la troisième dimension du modèle SMM.

#### Rapport de stabilité

Même si l'avantage du facteur de position. Il reste toujours des situations trompeuses, comme par exemple dans le cas d'inexistence de préfixe. Donc, une dernière tentative intervient afin de distinguer les dépendances inter-lettres des sous séquences selon leurs positions. Rappelons qu'on a calculé la dépendance d'une sous séquence séparément dans les trois positions définies précédemment dans le modèle (début, milieu et fin).

En exploitant la dimension de la position $Wpos$ dans le modèle SMM, la dépendance $Dep\langle seq\rangle$ d'une sous séquence $seq = s_1 s_2 s_3$ sera mesurée de la façon suivante :

$$\begin{aligned} Dep\langle seq\rangle &= SMM_{Dbt}[s_1\ s_2,\ s_3] + SMM_{Dbt}[s_1\ s_3,\ s_2] + SMM_{Dbt}[s_2\ s_3,\ s_1] \\ &+ SMM_{Mid}[s_1\ s_2,\ s_3] + SMM_{Mid}[s_1\ s_3,\ s_2] + SMM_{Mid}[s_2\ s_3,\ s_1] \\ &+ SMM_{Fin}[s_1\ s_2,\ s_3] + SMM_{Fin}[s_1\ s_3,\ s_2] + SMM_{Fin}[s_2\ s_3,\ s_1] \end{aligned}$$

### 3.2.3 Algorithme récapitulatif

La figure 3.6 schématise toutes les étapes de l'analyseur morphologique non supervisé proposé. Par la suite on récapitule par un pseudo code de l'algorithme principal 2 qui fait appel à toutes les fonctions mentionnées ci-dessus.

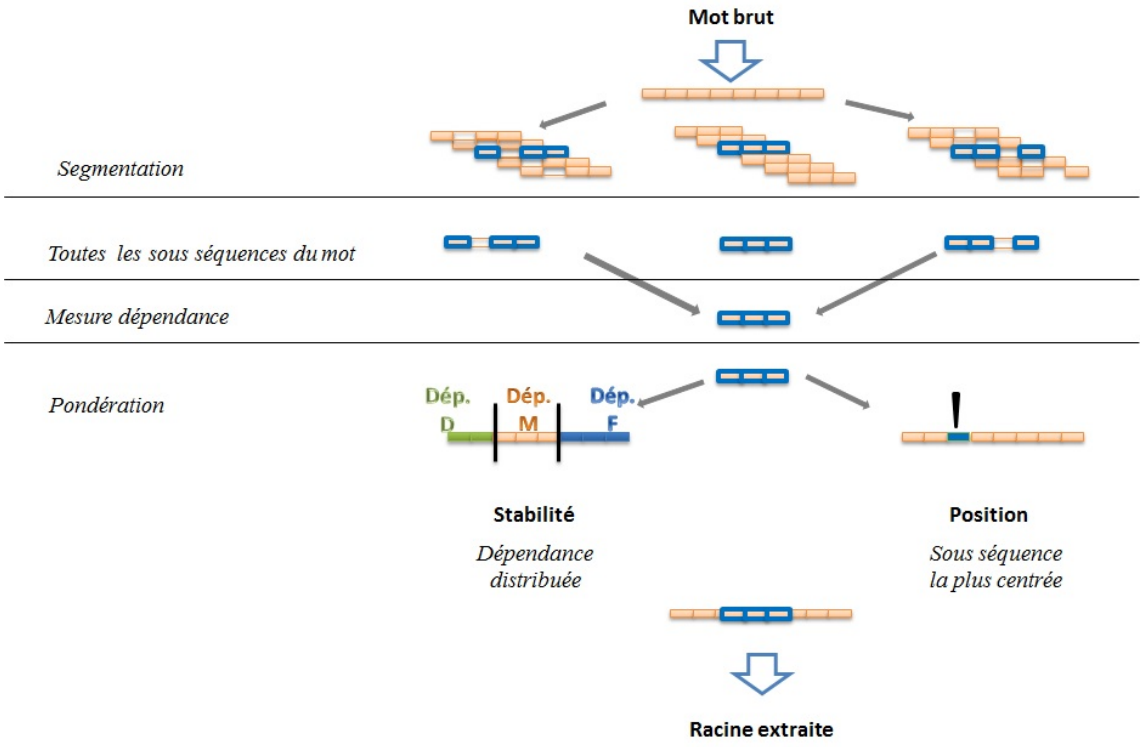


FIGURE 3.6 : Les différentes étapes de la segmentation non supervisée

**Algorithme 2** Analyseur morphologique non supervisé

**Entrée**
$w$ : le mot à analyser ;
$SMM$ : le modèle statistique à exploiter ;
**Début**

/*La segmentation du mot $w$ */

1. $S_{seq} = \{3-1 sous-sequence_i / i = 1 \ldots (|w|-3)\}$
U $\{3-0\ sous-sequence_i /\ i = 1 \ldots (|w|-2)\}$ ;

2. **Pour** $toutesseq_i = s_1 s_2 s_3 \in S_{seq}$ **faire**

/* La mesure de dépendance de la sous séquence $seq_i$ */

3. $$\begin{aligned} Dep\langle seq_i\rangle &= SMM_{Dbt}[s_1\ s_2,\ s_3] + SMM_{Dbt}[s_1\ s_3,\ s_2] + SMM_{Dbt}[s_2\ s_3,\ s_1] \\ &+ SMM_{Mid}[s_1\ s_2,\ s_3] + SMM_{Mid}[s_1\ s_3,\ s_2] + SMM_{Mid}[s_2\ s_3,\ s_1] \\ &+ SMM_{Fin}[s_1\ s_2,\ s_3] + SMM_{Fin}[s_1\ s_3,\ s_2] + SMM_{Fin}[s_2\ s_3,\ s_1] ; \end{aligned}$$

/* La pondération de la sous séquence $seq_i$ */

4. $Dep\langle seq_i\rangle = Dep\langle seq_i\rangle \cdot Facteur\ de\ position(seq_i\ ,\ w)$;
5. **Fin**

/* Racine retournée */

6. Retourner $seq$ ayant $\arg\max_{seq\ \in S_{seq}}(\ Dep\langle seq\rangle)$ ;
7.**Fin**

**Exemple** Les dépendances mesurées à partir du modèle SMM du mot **الدراسة**.

| Sous séquences du mot **الدراسة** | Dépendances | Position | Facteur de position | Dépendances pondérées |
|---|---|---|---|---|
| **الد** | 0.004 | 1 | 0.25 | 0.001 |
| **الر** | 0.006 | 1 | 0.25 | 0.001 |
| **ادر** | 0.10 | 1 | 0.25 | 0.002 |
| **لدر** | 0.053 | 2 | 0.50 | 0.026 |
| **لدا** | 0.024 | 2 | 0.50 | 0.012 |
| **لرا** | 0.014 | 2 | 0.50 | 0.007 |
| **درا** | 0.015 | 3 | 1.0 | 0.015 |
| **درس** | **0.083** | **3** | **1.0** | **0.083** |
| **داس** | 0.002 | 3 | 1.0 | 0.002 |

## 3.3 Conclusion

Dans ce chapitre nous avons présenté notre contribution ; la conception d'un modèle statistique pour la morphologie de la langue Arabe SMM et son exploitation dans la conception d'un analyseur morphologique non supervisé.

On a commencé par la définition des notions et des concepts requis pour notre modélisation (le modèle SMM). Ensuite nous avons mentionné le modèle en décrivant ses dimensions et les informations qu'il porte (dépendances inter-lettres), puis nous avons exposé l'analyseur morphologique proposé ; nous avons annoncé les étapes à suivre et les procédures qu'il requiert.

Nous donnerons dans le chapitre suivant une description des étapes de réalisation du modèle SMM et l'analyseur morphologique non supervisée, ainsi qu'une discussion des résultats obtenus.

# Chapitre 4

# Réalisation et expérimentation

Dans ce chapitre nous allons décrire comment on a procédé pour réaliser le modèle statistique et l'analyseur non supervisé de la langue. Nous commençons par ce que l'on requiert pour faire l'apprentissage du modèle SMM ; nous montrons le corpus utilisé en donnant ses caractéristiques, la taille, le nombre de documents et de mots et les différentes sources de ces documents. Par la suite, on entame l'implémentation en donnant l'architecture en trois couches adoptée, le paramétrage empirique et le benchmark utilisé pour l'évaluation.

## 4.1 Corpus utilisé

En terme de corpus, là aussi, les chercheurs de la langue Arabe disposent de très peu de corpus dédiés à comparer avec les autres langues. Le corpus qu'on a utilisé est une collection de textes bruts prise du Web, il a été créé dans le cadre du projet OSAC "*open Source Arabic Corpora*" réalisé par M.k. Saad [1] qui a regroupé trois collections de documents Web depuis plusieurs sources :

**BBC Arabic corpus** depuis le site web bbcarabic.com, qui contient 4,763 documents texte, formant sept catégories (science et technologie, sport, culture, information du monde et information du moyen orient ...), e recueil contient 1,860,789 mots,

**CNN Arabic corpus** amené depuis cnnarabic.com, il contient 5,070 documents texte, chacun d'eux appartient à une de six catégories prédéfinies (science et technologie, sport, culture, information du monde et information du moyen orient ...), le corpus rassemble 2,241,348 mots,

**Open Source Arabic Corpora-OSAC-** c'est la cueillette la plus grande et la plus diversifiée, elle a été collectée depuis des sources différentes, plusieurs sites Web mentionnés dans le tableau ci-dessous 4.1 chaque catégorie et ses sources, ce dernier corpus regroupe 22,429 documents textes repartis en dix 10 catégories (Economie, Histoire, amusions, Education, Famille, Religieux, Sport, santé, Astronomie, Loi, récit, Recette),

1. https://sites.google.com/site/motazsite/arabic/osac

TABLE 4.1 : Les sources des différents catégories des textes du corpus OSAC

| Catégorie | Nombre des docs | Sources |
|---|---|---|
| Économie | 3102 | – bbcarabic.com<br>– cnnarabic.com<br>– aljazeera.net<br>– khaleej.com<br>– banquecentrale.gov.sy |
| Histoire | 3233 | – تاريخ الحكام www.hukam.net<br>– moqatel.com<br>– التاريخ altareekh.com<br>– تاريخ الاسلام islamichistory.net |
| Education et famille | 3608 | – صيد الفوائد saaid.net<br>– نصائح للسعادة الأسرية naseh.net<br>– المربي almurabbi.com |
| Religieux | 3171 | – CCA corpus - EASC corpus<br>– moqatel.com<br>– شبكة الفتاوى الشرعية islamic-fatwa.com<br>– صيد الفوائد saaid.net |
| Sport | 249 | – bbcarabic.com<br>– cnnarabic.com<br>– khaleej.com |
| Santé | 2296 | – العيادة الاكترونية dr-ashraf.com<br>– CCA corpus - EASC corpus - W corpus<br>– صحة الطفل kids.jo<br>– العلاج البديل العربي arabaltmed.com |
| Astronomie | 557 | – الفلك العربي arabastronomy.com<br>– الكون نت alkawn.net<br>– بوابة الفلك المغربية bawabatalfalak.com<br>– الفلك موسوعة النابلسي nabulsi.com<br>– www.alkoon.alnomrosi.net |
| Droit | 944 | – القانون الليبي lawoflibya.com<br>– قانون كوم qnoun.com |
| Récit | 726 | – CCA corpus<br>– قصص الأطفال kids.jo<br>– صيد الفوائد saaid.net |
| Cuisine | 2373 | – aklaat.com<br>– fatafeat.com |
| Total | 22429 | |

## 4.2 Architecture de l'outil

La figure 4.1 l'architecture de notre implémentation du SMM et toutes application pouvant l'utiliser.

Nous avons adopté une architecture de trois couches (voir Figure 4.1) pour implémenter notre analyseur morphologique, pour qu'il soit indépendant du corpus d'apprentissage et paramétrable, pour qu'on puisse introduire le facteur de position et le rapport de stabilité cités auparavant (voir section précédentes 3.2.2), ainsi qu'elle donne la main à activer ou désactiver la prise en compte des infixes ou pas, ce qui nous laisse attirer l'attention que le contrôle de ces deux premières parties, offre la propriété de l'indépendance de la langue à analyser.

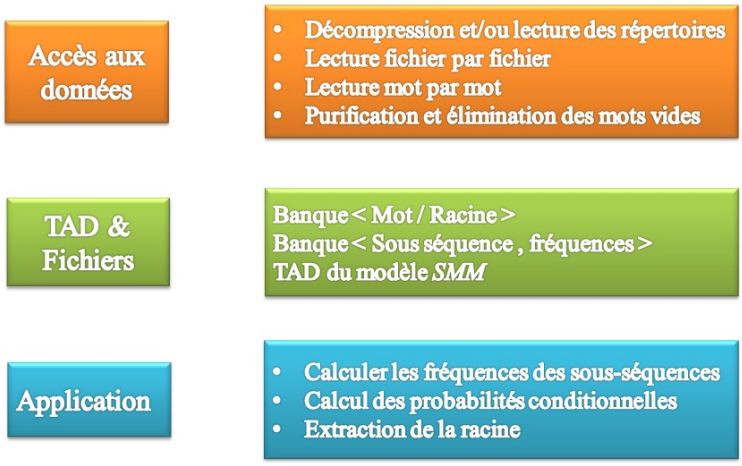


FIGURE 4.1 : L'architecture en couche de l'application

Ce développement en couches est prémédité. En effet, cette architecture vise à rendre le paramétrage de n'importe quelle application indépendante des opérations d'apprentissage ainsi de la génération du modèle.

### 4.2.1 Lecture du corpus

Comme le montre la figure 4.2 Le corpus est lu fichier par fichier. Comme ces corpus sont souvent à l'état complètement bruts avec des codages différents, avec des entêtes balisées (résidus de leurs pages web sources) on devait purifier chaque fichier en éliminant tous les caractères spéciaux et non imprimable, la ponctuation ou toute lettre d'une autre langue étrangère. On ne laisse que les séparateurs simples. Ensuite on crée un tableau (voir annexe) contenant tous les mots rencontrés avec leurs fréquences calculées d'une manière incrémentale à l'exploration de chaque nouveau document.

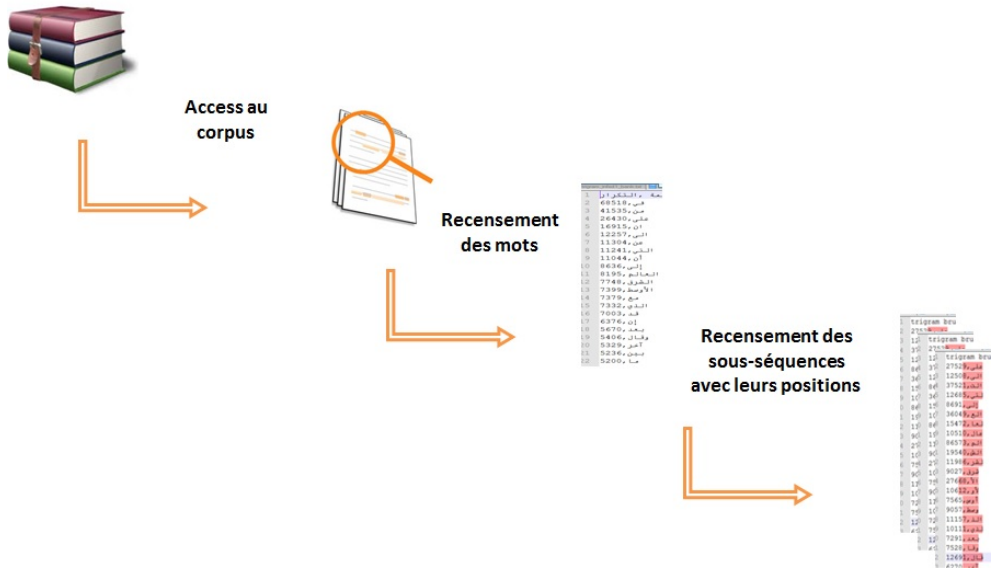


FIGURE 4.2 : Processus de lecture du corpus et génération du SMM

### 4.2.2 Données générée pour l'application

Cette couche a été conçue pour générer les fichiers permanents qui contiennent :
- la banque des mots avec leurs fréquences.
- les valeurs calculées et sauvegardées dans des tables : 1-grammes, 2-grammes et 3-grammes.
- et le modèle statistique.

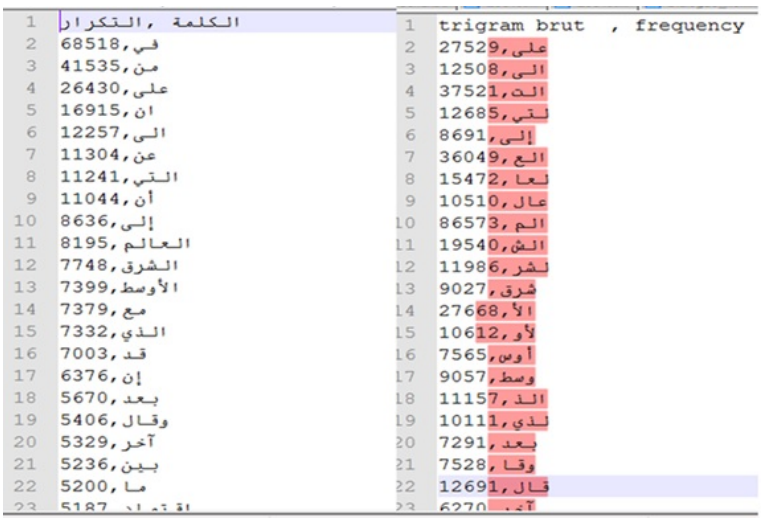


FIGURE 4.3 : Fichiers à utiliser par le modèle SMM

Ces fichiers sont en formats texte (txt) ou Excel (csv) (voir Figure 4.3), mais pour l'analyse d'un mot donné, il y a la possibilité de voir les résultats détaillés de la liste des morphes

avec dépendance jusqu'à la racine d'une façon instantanée. Comme on peut générer un fichier de la forme « mot, racine » dans le cas de l'analyse d'un jeu de mot, comme lors de l'expérimentation et validation du modèle.

### 4.2.3 Moteur d'inférence

La deuxième couche se compose de toutes les fonctions d'analyse. Elle traite essentiellement le fichier résultant du premier module. On commence par (i)l'extraction des sous séquences $N$-grammes (où n = 1, 2 et 3) et on les sauvegarde dans des banques de la forme « $N$-gramme, fréquence », ensuite, (ii)on épluche ces tables dans la génération du modèle statistique pour la morphologie (voir Figure 4.3), qui est sous forme matricielle à deux entrées ayant une interprétation naturelle. (iii)On exploite les probabilités obtenues par le modèle pour mesurer les dépendances des sous séquences générées par la troncature du mot à analyser. A la fin, et après le listing de tous les morphes avec leurs dépendances, on applique les deux paramètres : facteurs de position et rapport de stabilité, pour en déduire le morphe ayant le meilleur score.

## 4.3 Paramétrage et analyse

Comme la plupart des tâches d'apprentissage automatique. la segmentation morphologique, utilise l'indice de la précision pour mesurer la performance du système, elle peut être définie par le nombre des racines proposées par le système coïncidant avec celles correctes dans le jeu de teste, pour indiquer la validité des morphèmes proposés.

### 4.3.1 Comment évaluer un analyseur morphologique

On veut dire par l'évaluation, la mesure de degré de précision des résultats obtenus par l'algorithme ou la méthode à évaluer. Pour se faire, il nous faut une référence ou une métrique standard qu'on utilise pour comparer les résultats. Dans le TALN, on se réfère toujours à la réalité lorsqu'on est entrain soit de générer, soit de comprendre le langage naturel. Cette réalité est présentée par la production humaine d'un échantillon (le plus grand possible), de ce qu'on veut automatiser.

C'est difficile d'obtenir ce type d'échantillon souvent désigné dans la littérature "*un gold standard*", car l'apprentissage automatique de la morphologie est pénible car il exige la résolution de beaucoup de questions, tels que l'ambiguïté des mots, de la complexité morphologique des langues. Lorsque toutes ces questions sont examinées, on constate que l'élaboration d'un gold standard est aussi une tâche laborieuse. Les Gold standards sont des petits lexiques construits à la main par des experts en linguistique, contenant un jeu d'essai répondant pertinemment :

1. Au type du problème à résoudre : à quel niveau est faite cette analyse, la morphologie ou la syntaxe, segmentation (préfixe / stem / suffixe) ou une remontée à la racine tri-littérale étiquetage des partie de discours ....

2. A la nature du corpus utilisé par l'approche, diversifié son format et codage de son texte, la langue familière ou formelle ...
3. A un format précis et clair :

   de la structure de données qui le supporte qui doit inclure des informations concernant la morphologie (racine, motif, quelques affixes ...) et les parties du discours pour chaque mot sous une forme tabulaire.

   Du format de ses fichiers où il doit être stocké sous un format standard ex. XML, HTML et le codage aussi doit être reconnu et répandu l'Unicode UTF8 par exemple.
4. A une taille qui doit être relativement importante pour couvrir le plus de mots possible et le plus d'informations et de situations éventuelles qu'un mot peut prendre. La taille est mesurée par le nombre des mots qu'il contient.

La figure 4.4 montre le gold standard utilisé par la compétition morphochallenge (voir section 2.3.3)

| | | | |
|---|---|---|---|
| بِسْمِ | سم | None | ب+Prep , سم+Noun+Triptotic+Sg+Masc+Gen , |
| اللَّهِ | None | None | الله+Noun+ProperName+Gen+Def , |
| الرَّحْمَـنِ | رحم | فَعلان | رَحمان+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def , |
| الرَّحِيمِ | رحم | فَعِيل | رَحِيم+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def , |
| بسم | سم | None | ب+Prep , سم+Noun+Triptotic+Sg+Masc+Gen , |
| الله | None | None | الله+Noun+ProperName+Gen+Def , |
| الرحمـن | رحم | فعلان | رحمان+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def , |
| الرحيم | رحم | فعيل | رحيم+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def , |
| bisomi | sm | None | b+Prep , sm+Noun+Triptotic+Sg+Masc+Gen , |
| All~hi | None | None | llaah+Noun+ProperName+Gen+Def , |
| Alr~aHom_ani | rHm | faElaAn | raHmaan+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def |
| Alr~aHiymi | rHm | faEiyl | raHiim+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def , |
| bsm | sm | None | b+Prep , sm+Noun+Triptotic+Sg+Masc+Gen , |
| Allh | None | None | llAh+Noun+ProperName+Gen+Def , |
| AlrHm_n | rHm | fElAn | rHmAn+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def |
| AlrHym | rHm | fEyl | rHym+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def , |

FIGURE 4.4 : Le gold standard utilisé dans la compétition morphochallenge

### 4.3.2 Notre benchmark

Afin d'évaluer l'efficacité de notre racineur, nous avons adopté un gold standard similaire de Sawalha et Atwel [SA08].

Ce choix est doublement justifié. Premièrement, ceci nous servira pour vérifier et valider le modèle à l'aide d'un jeu de tests fait main un "*handmade*". Deuxièmement, ce gold standard nous servira pour comparer les performances de notre racineur avec trois autres algorithmes bien répandus dans la littérature de l'analyse morphologique à savoir : Shereen Khoja stemmer [Kho01], Tim Buckwalter Morphological analyzer 2004 [Buc02]

et Tri-literal Root Extraction Algorithm 2003 [ASASK03].

En effet, notre gold standard est tiré du corpus de l'Arabe contemporaine (CCA Sulaiti et Atwel 2006) [ASA06]. Il contient plus de 1000 mots tirés aléatoirement. Puis à chaque mot, on rajoute sa racine tri-littérale. Ce rajout est fait à la main à l'aide d'expert de la langue Arabe. Nous avons aussi veillé à l'inspection et la vérification de la correction de ses racines via le dictionnaire معجم العين d'*ELFARAHIDI* [KH96].

### 4.3.3 Influence des différents paramètres

Pour mesurer l'influence des différentes améliorations et paramètres de l'approche non supervisée, nous avons évalué la méthode en insérant ses paramètres étape par étape et mesurer l'influence sur la précision.

En effet la précision de notre algorithme a évolué comme suit :

1. Calcul des fréquences des sous séquences, Précision → 30 %
2. Introduction de facteur de position, Précision → 46 %
3. L'utilisation du rapport de stabilité, Précision → 65 %

La figure 4.5 schématise cette évaluation.

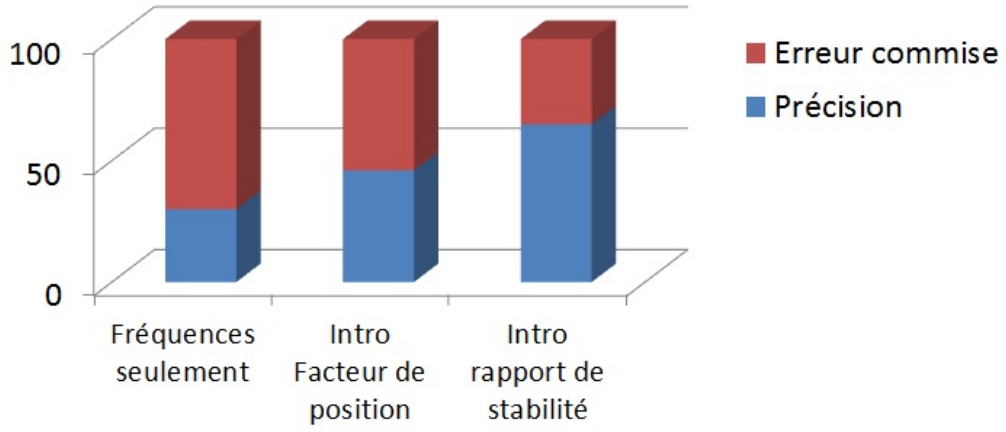


FIGURE 4.5 : L'influence du facteur de position et le rapport de stabilité sur la précision du racineur

### 4.3.4 Comparaison avec d'autres analyseurs morphologiques

Comme nous l'avons mentionné précédemment dans la description du benchmark, nous avons confronté notre méthode avec trois autres algorithmes, dans le but d'étudier la faisabilité de l'application de l'analyse morphologique non supervisée en langue Arabe. Les auteurs(Sawalha Atwel [SA08]) ont justifié leur choix de ces algorithmes, par leur accessibilité par rapport aux autres.

Nous tenons à attirer l'attention une seconde fois sur l'absence d'un standard en linguistique Arabe permettant l'évaluation de toute invention dans ce domaine particulier. Une autre idée nous stimule à créer un standard consistant pour l'analyse morphologique en langue Arabe, et faire une enquête sur son état actuel en recensant les travaux entrepris dans ce domaines d'ici une décennie en derrière.

1. Shereen Khoja stemmer [Kho01],
2. Tim Buckwalter Morphological analyzer 2004 [Buc02]
3. et Tri-literal Root Extraction Algorithm 2003 [ASASK03].
4. Notre Approche

Nous tenons aussi à rapporté que les deux premiers stemmers sont complètement supervisés et le dernier racineur d'AL Shalabi fait partie des analyseurs morphologiques combinatoires qui utilisent des connaissances linguistiques un peu réduites par rapport aux premiers algorithmes.

TABLE 4.2 : Racineurs à comparer

| | Précision % | Erreur commise % |
|---|---|---|
| Khoja | 67,32 | 32,68 |
| Buckwalter | 39,3 | 60,7 |
| Al Shalabi et al | 64,37 | 35,63 |
| Notre approche | 65 | 35 |

Le tableau 4.2 montre la précision des analyseurs reconnus à comparer avec notre racineur.

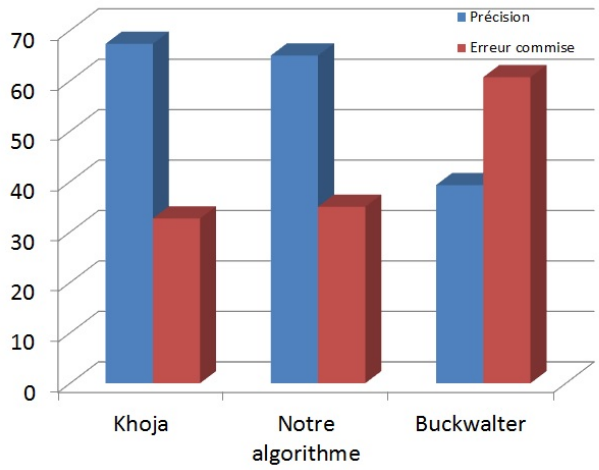


FIGURE 4.6 : Comparaison de l'analyseur développé avec les stemmers supervisés connus

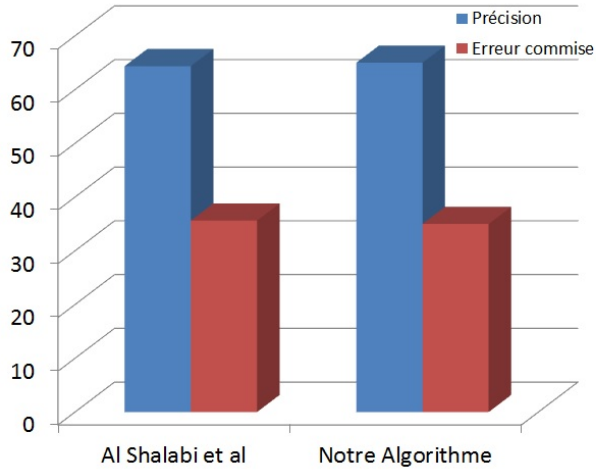


FIGURE 4.7 : Comparaison de l'analyseur développé avec le stemmer combinatoire d'Al Shalabi

Comme la figure 4.6 montre, la performance de notre algorithme est meilleure que celle de l'analyseur de Buckwalter et très compétitif à l'un des algorithmes supervisés pionniers (Khoja2001). En effet, tenant compte de la gourmandise en terme des ressources linguistiques requises par ces approches supervisées, la technique proposée est jugée comme un analyseur morphologique allégé présentant une bonne performance interprétée par une précision de 65%.

Maintenant comparer avec la troisième approche, le stemmer combinatoire d'Al Shalabi et al, la figure 4.7 montre que les résultats obtenus par notre racineur sont pratiquement meilleurs. Il convient de mentionner que certains indices ont été utilisés par cet analyseur. En effet, il utilise une liste prédéfinie de lettres qui pourraient faire partie des affixes de la langue Arabe et qui sont réunis dans le mot سألتمونيها. Ce qui lui écarté d'être un racineur non supervisé, chose qui fait qui que nos résultats sont meilleurs.

On remarque aussi que les approches combinatoires pour l'analyse morphologique sont restreintes à l'extraction de la racine et n'atteignent pas des analyses profondes.

Cependant et contrairement aux méthodes précédentes, notre approche non supervisée utilise un modèle générique de mot et elle peut s'étaler pour discerner une segmentation profondes des mots en langue Arabe. En effet, elle peut caractériser les fragments apparaissant fréquemment au début ou à la fin des mots et les juger comme affixes comme elle peut déceler les infixes après l'extraction des lettres radicales d'un mot et en déduire le motif (الوزن).

Les arguments cités au dessus laissent la voix de l'analyse non supervisée de la mor-

phologie mérite d'être explorée.

## 4.4 Conclusion

Dans ce chapitre, nous avons évalué notre contribution par l'évaluation des performances de l'analyseur morphologique proposé. Cette évaluation est faite à l'aide d'un benchmark de 1000 mots, qui a été créé aléatoirement depuis un corpus contemporain, auquel les racines ont été rajoutées à la main avec l'aide des experts en linguistique Arabe.

Après l'implémentation,l'expérimentation et le paramétrage empirique de la méthode, les résultats obtenus ont montré que l'algorithme d'extraction des racines non supervisé conçu assure une bonne précision qui a atteint 65 %.

Nous avons aussi évalué les performances de notre analyseur, en le comparant avec d'autres analyseurs morphologiques bien reconnu dans le domaine (Khoja, Buckwalter et Al-shalabi).

L'approche non supervisée proposée a donné des résultats compétitifs surtout en terme de ressources requises par rapport aux algorithmes supervisés dont notre analyseur n'exige aucune ressource après l'apprentissage du modèle via le corpus qui ne se fait qu'une seule fois. Concernant la comparaison avec le raccineur combinatoire d'Al Shalabi et al. ou pseudo supervisé, notre algorithme a donné de bon resultats pour l'Arabe standard moderne.

# Conclusion générale

Plusieurs applications sont couvertes par le domaine du TALN depuis son émergence aux années cinquante du siècle précédent jusqu'à nos jours, chose qui lui a permis de devenir une discipline à part de l'informatique et la linguistique. Comme toutes discipline, le TALN est discriminée par des axes de recherche propres comme il partage d'autres intérêts avec d'autres disciplines ; exemple de ces intérêts la reconnaissance de la parole, la traduction automatique etc.

Ce qui caractérise le TALN est l'étude et la maîtrise de la forme et du contenu de la langue naturelle d'une façon automatique afin qu'elle soit exploitable par les autres applications. Cette étude particulière présente la recherche fondamentale de la linguistique informatique. La forme de la langue est l'ensemble de ses constituants élémentaires (lettres, mots, phrases, paragraphe) et le contenu est le sens à transférer. La forme s'étudie aux niveaux linguistiques : Lexical (forme de texte ; stopword, séparateurs ...), morphologie (forme des mots), Syntaxe (forme des phrases) et le contenu s'étudie au niveau sémantique (sens des mots) et niveau pragmatique (sens contextualisé).

Dans ce travail,nous nous sommes intéressé à l'analyse au niveau morphologique qui traite essentiellement la forme des mots. Il est prouvé par plusieurs travaux que la remontée du mot brut au mot segmenté, à sa racine et aux affixes ajoutés influe considérablement sur la performance de l'analyse aux niveaux supérieurs ; sémantique et pragmatique. Nous avons entamé notre étude par survoler le domaine de la linguistique informatique, ensuite on s'est étalé sur les concepts de l'analyse morphologique, puis on a cité les approches adoptées pour l'étude de la morphologie en donnant quelques exemples pour chaque technique.

Nous avons exposé et expliqué dans ce document les spécificités de la langue Arabe. Nous avons démontré les difficultés qu'elle impose aux concepteurs d'analyseurs morphologiques. Nous avons terminé la partie consacré à la langue Arabe par un état de l'art des travaux faits vis-à-vis de cette issue.

Apercevant cet état de l'art, nous avons constaté que très peu de travaux ont considéré la conception d'analyseurs morphologiques non supervisés pour la langue Arabe. Et ceux malgré leurs performances prouvées pour les autres langues.

Dans ce mémoire, nous avons proposé un modèle statistique pour la morphologie en langue Arabe SMM. Ce modèle profite des atouts de (i)l'apprentissage automatique et (ii)l'analyse statistique, pour capter les régularités de la morphologie de la langue Arabe. Désirant que ce modèle soit le plus général possible, nous l'avons alimenté à partir d'un corpus contemporain OSAC, contenant des documents Web bruts de taille globale de 18 millions de mot.

Afin d'évaluer les performances de ce modèle, nous avons conçu un analyseur morphologique non supervisé pour la langue Arabe standard moderne. Cet analyseur remonte du mot brut à la racine tri-littérale en utilisant le modèle SMM et sans aucune autre connaissance linguistique.

L'évaluation des performances de l'analyseur morphologique proposé est faite à l'aide d'un benchmark de 1000 mots annotés par les racines, qui a été créé aléatoirement depuis un corpus contemporain CCA, les racines ont été rajoutées à la main avec l'aide d'experts en linguistique Arabe.

Dans cette étude expérimentale, nous avons réalisé 65% de précision. Afin de comparer ces performances avec d'autres analyseurs, deux lots d'expériences ont été effectués. D'une part, nous avons comparé notre analyseur à un analyseur combinatoire Al Shalabi. Les résultats sont légèrement meilleurs. Ce dernier ne peut être considéré comme analyseur morphologique complètement non supervisé, en tenant compte des indices utilisés par cet analyseur ; une liste prédéfinie de lettres qui pourraient faire partie des affixes. D'autre part nos résultats sont très compétitifs surtout en terme de ressources requises par rapport à l'algorithme supervisés de Khoja et bien meilleur par rapport aux résultats obtenus par Buckwalter.

Les résultats obtenus nous encouragent à poursuivre cette voie d'analyse non supervisée de la morphologie. En effet nous envisageons beaucoup de perspectives, parmi celles qui nous intéressent :

- La création d'une boite à outils open source implémentant les analyseurs ayant les meilleures performances et permettant de rajouter n'importe quel analyseur morphologique pour l'évaluer, l'utiliser ou même l'améliorer.
- L'extension de notre racineur depuis l'extraction des racines tri-littérales seulement à un analyseur morphologique capable de générer une segmentation profonde, dont il a pu capter les sous séquences corrélées aux début et à la fin de mot (Préfixe et suffixe).

# Annexe

Dans cette annexe, nous présentons un aperçu sur les banques de données engendrées à partir de l'inférence du corpus des textes bruts, ainsi que le benchmark utilisé pour l'évaluation de notre algorithme non supervisé d'extraction des racines tri-littérales. Ces banques sont respectivement :

- la table $\langle mot, Frequence \rangle$,
- les tables :

  $\langle N-gramme, FrequenceDebut, FrequenceMilieu, FrequenceFin, FrequenceTotale \rangle$

  où $N = 1, 2\ et\ 3$,

- et le gold standard (benchmark) utilisé pour l'évaluation.

Le tableau suivant présente un fragment de la banque (*Mot, Fréquence*) qu'on a obtenu en traitant le corpus purifié ; c à d. après le pré-traitement.

TABLE 0.3: Un fragment de La banque (*Mot-Fréquence*)

| الكلمة | التكرار | الكلمة | التكرار | الكلمة | التكرار |
|---|---|---|---|---|---|
| حتى | 1516 | العربية | 986 | في | 68518 |
| وهو | 1503 | تم | 984 | من | 41535 |
| بأن | 1500 | مما | 984 | على | 26430 |
| وقالت | 1498 | باكستان | 981 | ان | 16915 |
| السلام | 1477 | اي | 963 | الى | 12257 |
| الا | 1474 | فقد | 959 | عن | 11304 |
| المنطقة | 1461 | قرار | 959 | التي | 11241 |
| الثلاثاء | 1451 | النووي | 958 | أن | 11044 |
| الامريكي | 1447 | البريطاني | 953 | إلى | 8636 |
| مجلس | 1445 | شمال | 952 | العالم | 8195 |
| الاثنين | 1436 | قالت | 951 | الشرق | 7748 |
| اوباما | 1428 | آذار | 937 | الأوسط | 7399 |
| بشأن | 1423 | فيما | 937 | مع | 7379 |
| الخميس | 1420 | المئة | 933 | الذي | 7332 |
| الدول | 1415 | عبر | 926 | قد | 7003 |

*Suite en page suivante*

TABLE 0.3 – *Suite de la page précédente*

| الكلمة | التكرار | الكلمة | التكرار | الكلمة | التكرار |
|---|---|---|---|---|---|
| و | 1412 | عاما | 915 | إن | 6376 |
| الأمريكية | 1410 | دون | 910 | بعد | 5670 |
| ضد | 1397 | الشهر | 907 | وقال | 5406 |
| ومن | 1388 | نوفمبر | 902 | آخر | 5329 |
| باسم | 1382 | المقبل | 900 | بين | 5236 |
| دبي | 1370 | بلاده | 898 | ما | 5200 |
| الجمعة | 1361 | بينما | 898 | اقتصاد | 5187 |
| او | 1361 | وهي | 890 | وأعمال | 5075 |
| منطقة | 1344 | احد | 889 | الرئيسية | 5056 |
| بها | 1334 | ولا | 889 | رياضة | 5037 |
| العاصمة | 1317 | الوقت | 887 | علوم | 5014 |
| هي | 1310 | مليون | 882 | وتكنولوجيا | 5006 |
| فيه | 1303 | سبتمبر | 880 | شارك | 4896 |
| البريطانية | 1298 | محمد | 880 | منوعات | 4882 |
| أي | 1291 | والتي | 880 | راديو | 4776 |
| تلك | 1286 | يكون | 878 | برأيك | 4765 |
| السبت | 1281 | به | 869 | وتلفزيون | 4763 |
| العراقي | 1275 | ولم | 866 | المتحدة | 4674 |
| أنه | 1269 | عليه | 865 | لا | 4522 |
| نحو | 1263 | أحد | 859 | تحديث | 4486 |
| حزب | 1255 | الامن | 859 | هذه | 4384 |
| وزارة | 1250 | قائلا | 859 | هذا | 4129 |
| إنه | 1248 | قتل | 859 | الرئيس | 4092 |
| قوات | 1241 | شهر | 854 | بي | 3992 |
| إسرائيل | 1226 | امام | 847 | لم | 3737 |
| أوباما | 1221 | شركة | 841 | الحكومة | 3681 |
| انها | 1220 | اتفاق | 833 | كما | 3451 |
| القاعدة | 1215 | العمل | 826 | كان | 3440 |
| جنوب | 1209 | شخص | 819 | خلال | 3413 |
| جديدة | 1204 | تقول | 818 | عام | 3390 |
| الاسرائيلية | 1200 | الحدود | 817 | رئيس | 3302 |
| الأمريكي | 1188 | إلا | 814 | وكان | 3073 |
| إيران | 1183 | الأمن | 813 | قبل | 3038 |
| الصين | 1175 | أيضا | 808 | الماضي | 3034 |
| ويقول | 1172 | ليس | 801 | العراق | 2855 |
| الحرب | 1166 | الحركة | 800 | وقد | 2815 |
| افغانستان | 1160 | الانترنت | 797 | ذلك | 2701 |
| عبد | 1160 | الدفاع | 797 | قال | 2649 |
| دولار | 1159 | تكون | 796 | الانتخابات | 2609 |
| الاتحاد | 1148 | الآن | 792 | انه | 2524 |
| وأضاف | 1148 | بيان | 791 | منذ | 2506 |

*Suite en page suivante*

TABLE 0.3 – *Suite de la page précédente*

| الكلمة | التكرار | الكلمة | التكرار | الكلمة | التكرار |
|---|---|---|---|---|---|
| يقول | 1148 | حكومة | 786 | الثاني | 2463 |
| الاسرائيلي | 1144 | لدى | 781 | العام | 2420 |
| لن | 1144 | الايرانية | 776 | سي | 2402 |
| مصر | 1143 | تركيا | 775 | كانت | 2378 |
| وذلك | 1133 | الهجوم | 772 | لكن | 2322 |
| وكالة | 1130 | الفلسطينيين | 764 | الولايات | 2301 |
| واضاف | 1124 | قطاع | 760 | غير | 2220 |
| اكثر | 1123 | حسب | 758 | اسرائيل | 2215 |
| بغداد | 1122 | الفلسطيني | 757 | الخارجية | 2158 |
| عملية | 1122 | عدم | 757 | القوات | 2112 |
| يمكن | 1115 | ايضا | 752 | حول | 2107 |
| الاربعاء | 1109 | السياسية | 751 | عدد | 2063 |
| الدولي | 1105 | بناء | 749 | وكانت | 2032 |
| بشكل | 1103 | كبير | 748 | وفي | 2000 |
| القدس | 1098 | كبيرة | 748 | كل | 1977 |
| أكثر | 1097 | المفاوضات | 743 | الذين | 1920 |
| السعودية | 1093 | العسكرية | 732 | كانون | 1888 |
| يذكر | 1085 | آب | 731 | يوم | 1842 |
| واشنطن | 1084 | منها | 731 | ايران | 1824 |
| وقت | 1084 | الوكالة | 730 | حيث | 1823 |
| حماس | 1083 | اخرى | 727 | هو | 1816 |
| الغربية | 1053 | الايراني | 727 | مقتل | 1814 |
| الفلسطينية | 1052 | المحكمة | 725 | الجيش | 1811 |
| الاحد | 1050 | القاهرة | 723 | الوزراء | 1811 |
| اليوم | 1049 | علي | 723 | بعض | 1785 |
| مارس | 1045 | فقط | 721 | هناك | 1765 |
| شخصا | 1044 | وسط | 718 | غزة | 1747 |
| السودان | 1038 | مصادر | 717 | وزير | 1679 |
| مثل | 1034 | البرلمان | 714 | الامريكية | 1673 |
| أفغانستان | 1033 | وان | 714 | الأول | 1665 |
| ولكن | 1026 | المعارضة | 712 | العراقية | 1665 |
| عليها | 1020 | المالكي | 710 | البلاد | 1657 |
| نتنياهو | 1020 | المصري | 710 | الدولية | 1655 |
| له | 1019 | ألف | 709 | فيها | 1652 |
| ثلاثة | 1017 | الامم | 707 | تشرين | 1646 |
| عندما | 1011 | جانب | 704 | الله | 1620 |
| بريطانيا | 1001 | داخل | 704 | بسبب | 1577 |
| لها | 998 | الطريق | 698 | أو | 1574 |
| اليمن | 997 | عدة | 698 | السلطات | 1569 |
| باراك | 997 | الضفة | 697 | الشرطة | 1542 |
| يناير | 997 | لجنة | 694 | مدينة | 1530 |

*Suite en page suivante*

TABLE 0.3 – *Suite de la page précédente*

| الكلمة | التكرار | الكلمة | التكرار | الكلمة | التكرار |
|---|---|---|---|---|---|
| طهران | 994 | الإسرائيلي | 691 | طالبان | 1527 |
| ديسمبر | 990 | أمام | 690 | السابق | 986 |
| حركة | 1518 | تقرير | 690 | أخرى | 689 |

Le tableau suivant présente un fragment de la banque (*3-1-gappy sequence, et leurs fréquences*) qu'on l'a obtenue par une inférence statistique des de la banque précédente.

TABLE 0.4: Un fragment de la banque des *3-1 gappy sequence*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| اعا | 16017 | 15023 | 368 | 626 |
| لال | 14841 | 500 | 1228 | 13113 |
| علم | 9555 | 267 | 5983 | 3305 |
| اشر | 11516 | 11304 | 817 | -605 |
| لرق | 9297 | 43 | 40 | 9214 |
| أأو | 10508 | 10489 | 0 | 19 |
| لوس | 9481 | 90 | 63 | 9328 |
| أسط | 8009 | 604 | 4 | 7401 |
| وال | 11564 | 8976 | 1071 | 1517 |
| اتص | 7997 | 6392 | 0 | 1605 |
| قصا | 6919 | 19 | 24 | 6876 |
| تاد | 10967 | 635 | 17 | 10315 |
| وعم | 5265 | 5265 | 159 | -159 |
| أما | 6191 | 678 | 81 | 5432 |
| عال | 7031 | 120 | 1116 | 5795 |
| ارئ | 10713 | 10502 | 0 | 211 |
| لئي | 10042 | 104 | 0 | 9938 |
| ريس | 14222 | 3782 | 172 | 10268 |
| ئسي | 6297 | 0 | 0 | 6297 |
| يية | 17122 | 0 | 3 | 17119 |
| راض | 5321 | 5094 | 19 | 208 |
| يضة | 5085 | 0 | 19 | 5066 |
| عوم | 6100 | 5080 | 310 | 710 |
| وكن | 6346 | 6331 | 3106 | -3091 |
| تنو | 5253 | 50 | 18 | 5185 |
| كول | 5277 | 51 | 12 | 5214 |
| نلو | 5528 | 36 | 0 | 5492 |
| ووج | 5520 | 179 | 3 | 5338 |

*Suite en page suivante*

TABLE 0.4 – *Suite de la page précédente*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| لجي | 5672 | 162 | 95 | 5415 |
| ويا | 8509 | 396 | 323 | 7790 |
| شرك | 7407 | 5064 | 111 | 2232 |
| موع | 4942 | 4890 | 928 | -876 |
| نعا | 5451 | 65 | 78 | 5308 |
| وات | 10288 | 190 | 84 | 10014 |
| ردي | 4989 | 4797 | 118 | 74 |
| ايو | 9194 | 2579 | 58 | 6557 |
| بأي | 4816 | 4810 | 5 | 1 |
| ريک | 4797 | 9 | 48 | 4740 |
| ولف | 5470 | 5464 | 8 | -2 |
| تفز | 5296 | 156 | 47 | 5093 |
| لزي | 6874 | 104 | 16 | 6754 |
| فيو | 5328 | 7 | 14 | 5307 |
| زون | 5409 | 7 | 9 | 5393 |
| امت | 9832 | 9266 | 919 | -353 |
| لتح | 7142 | 251 | 0 | 6891 |
| محد | 5755 | 387 | 1049 | 4319 |
| تدة | 4887 | 0 | 37 | 4850 |
| تدي | 7920 | 6275 | 108 | 1537 |
| حيث | 5098 | 293 | 0 | 4805 |
| احک | 5289 | 5105 | 0 | 184 |
| لکو | 5723 | 229 | 2 | 5492 |
| حوم | 6504 | 1387 | 8 | 5109 |
| کمة | 5177 | 184 | 85 | 4908 |
| خال | 3831 | 3695 | 0 | 136 |
| وان | 9491 | 5294 | 334 | 3863 |
| اما | 8830 | 6869 | 877 | 1084 |
| لاض | 4064 | 18 | 243 | 3803 |
| مضي | 3693 | 35 | 1 | 3657 |
| اعر | 8561 | 8273 | 9 | 279 |
| لرا | 12753 | 481 | 11 | 12261 |
| عاق | 8919 | 1322 | 18 | 7579 |
| اان | 10774 | 8886 | 0 | 1888 |
| لنت | 7751 | 279 | 2 | 7470 |
| اتخ | 6575 | 2152 | 0 | 4423 |
| نخا | 4243 | 8 | 41 | 4194 |
| تاب | 5226 | 113 | 61 | 5052 |
| خبا | 3723 | 7 | 17 | 3699 |

*Suite en page suivante*

TABLE 0.4 – *Suite de la page précédente*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| اات | 24960 | 2682 | 1 | 22277 |
| اثا | 5073 | 3477 | 70 | 1526 |
| لان | 10462 | 836 | 1277 | 8349 |
| ثني | 3521 | 503 | 11 | 3007 |
| لام | 6009 | 567 | 1223 | 4219 |
| كنت | 4541 | 2423 | 8 | 2110 |
| اول | 3301 | 2534 | 81 | 686 |
| للا | 11852 | 361 | 27 | 11464 |
| واي | 3664 | 474 | 1260 | 1930 |
| ليا | 9597 | 292 | 1932 | 7373 |
| ارا | 15729 | 7446 | 875 | 7408 |
| سائ | 9080 | 5 | 3 | 9072 |
| رئي | 9054 | 0 | 34 | 9020 |
| ايل | 10305 | 58 | 211 | 10036 |
| اخا | 4815 | 4400 | 12 | 403 |
| لار | 6922 | 102 | 640 | 6180 |
| خرج | 3288 | 665 | 366 | 2257 |
| اجي | 4730 | 2078 | 74 | 2578 |
| رية | 7087 | 110 | 1495 | 5482 |
| اقو | 3859 | 3357 | 11 | 491 |
| لوا | 8515 | 608 | 188 | 7719 |
| قات | 4770 | 1484 | 2825 | 461 |
| لين | 9857 | 222 | 3620 | 6015 |
| كنو | 2729 | 2588 | 3 | 138 |
| اون | 4610 | 38 | 95 | 4477 |
| يان | 15181 | 203 | 5 | 14973 |
| متل | 2923 | 2331 | 46 | 546 |
| ليش | 2265 | 7 | 19 | 2239 |
| اوز | 2435 | 2403 | 10 | 22 |
| لزر | 1933 | 26 | 5 | 1902 |
| ورا | 3112 | 739 | 1220 | 1153 |
| زاء | 2364 | 50 | 4 | 2310 |
| هاك | 2064 | 1780 | 6 | 278 |
| وير | 4000 | 3042 | 139 | 819 |
| اام | 9722 | 8919 | 1 | 802 |
| لمر | 8410 | 369 | 138 | 7903 |
| اري | 7702 | 2121 | 330 | 5251 |
| ميك | 8496 | 305 | 1 | 8190 |
| ركي | 7934 | 11 | 251 | 7672 |

*Suite en page suivante*

TABLE 0.4 – *Suite de la page précédente*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| لول | 10553 | 400 | 12 | 10141 |
| رقي | 5110 | 9 | 11 | 5090 |
| اية | 26303 | 4 | 169 | 26130 |
| ابل | 3363 | 3245 | 285 | -167 |
| باد | 3751 | 1360 | 228 | 2163 |
| ادو | 6591 | 6058 | 4812 | -4279 |
| دلي | 3550 | 556 | 70 | 2924 |
| وية | 11451 | 13 | 241 | 11197 |
| فها | 2429 | 1660 | 121 | 648 |
| تري | 8177 | 5089 | 584 | 2504 |
| شين | 2619 | 8 | 9 | 2602 |
| اله | 2438 | 1777 | 2269 | -1608 |
| ببب | 1620 | 1588 | 1 | 31 |
| اسل | 4479 | 4275 | 22 | 182 |
| للط | 2341 | 18 | 6 | 2317 |
| سطا | 2352 | 355 | 8 | 1989 |
| لات | 5468 | 263 | 39 | 5166 |
| لرط | 2168 | 25 | 13 | 2130 |
| شطة | 2137 | 347 | 0 | 1790 |
| مين | 3507 | 1738 | 393 | 1376 |
| دنة | 2408 | 1 | 356 | 2051 |
| طلب | 3594 | 2076 | 159 | 1359 |
| ابا | 7782 | 4038 | 582 | 3162 |
| حكة | 2813 | 1521 | 1232 | 60 |
| قلت | 2719 | 1080 | 74 | 1565 |
| سام | 4038 | 354 | 172 | 3512 |
| امن | 6455 | 6094 | 962 | -601 |
| لنط | 1725 | 62 | 754 | 909 |
| مطق | 3172 | 1409 | 45 | 1718 |
| نقة | 3088 | 1 | 454 | 2633 |
| اثل | 2253 | 2185 | 12 | 56 |
| ثاث | 3684 | 1486 | 287 | 1911 |
| لثا | 1775 | 39 | 2 | 1734 |
| ااء | 1683 | 154 | 0 | 1529 |
| ملس | 2092 | 1521 | 1 | 570 |
| ااث | 1895 | 1689 | 0 | 206 |
| لثن | 1657 | 13 | 41 | 1603 |
| اني | 3799 | 1131 | 227 | 2441 |
| ثين | 2687 | 15 | 983 | 1689 |

*Suite en page suivante*

TABLE 0.4 – *Suite de la page précédente*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| وام | 3225 | 214 | 184 | 2827 |
| بما | 3250 | 272 | 0 | 2978 |
| بأن | 1516 | 1505 | 30 | -19 |
| اخم | 1692 | 1674 | 23 | -5 |
| لمي | 5078 | 110 | 134 | 4834 |
| خيس | 1428 | 8 | 0 | 1420 |
| الأم | 6317 | 6125 | 0 | 192 |
| أري | 4375 | 1246 | 15 | 3114 |
| بسم | 1525 | 1514 | 9 | 2 |
| اجم | 2909 | 2823 | 27 | 59 |
| لمع | 1889 | 197 | 116 | 1576 |
| جعة | 1489 | 25 | 31 | 1433 |
| لاص | 2481 | 44 | 13 | 2424 |
| عصم | 1757 | 282 | 35 | 1440 |
| امة | 4690 | 198 | 2839 | 1653 |
| ابر | 6073 | 5617 | 166 | 290 |
| لري | 6967 | 196 | 252 | 6519 |
| بيط | 4324 | 1554 | 0 | 2770 |
| رطا | 4261 | 0 | 34 | 4227 |
| طني | 8633 | 6 | 2 | 8625 |
| اسب | 1773 | 1713 | 63 | -3 |
| لبت | 1441 | 43 | 709 | 689 |
| وار | 3798 | 1759 | 474 | 1565 |
| زرة | 2447 | 3 | 30 | 2414 |
| إرا | 7158 | 3770 | 2 | 3386 |
| أبا | 2418 | 1581 | 60 | 777 |
| اها | 8982 | 2659 | 152 | 6171 |
| اقا | 6596 | 5352 | 205 | 1039 |
| لاع | 2174 | 147 | 5 | 2022 |
| قعد | 1895 | 260 | 1 | 1634 |
| ادة | 3668 | 1 | 2688 | 979 |
| جوب | 3054 | 1693 | 104 | 1257 |
| جيد | 3795 | 1882 | 3 | 1910 |
| ددة | 2279 | 0 | 39 | 2240 |
| ااس | 9431 | 8661 | 0 | 770 |
| لسر | 4992 | 266 | 374 | 4352 |
| ئلي | 5061 | 0 | 6 | 5055 |
| اصي | 2602 | 2343 | 37 | 222 |
| وقو | 2069 | 2068 | 0 | 1 |

*Suite en page suivante*

TABLE 0.4 – *Suite de la page précédente*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| يو ل | 3858 | 1567 | 6 | 2285 |
| احر | 3772 | 3558 | 94 | 120 |
| لرب | 7269 | 166 | 42 | 7061 |
| اغا | 3279 | 2248 | 20 | 1011 |
| فان | 4678 | 443 | 4 | 4231 |
| غنس | 2432 | 0 | 0 | 2432 |
| است | 5388 | 315 | 316 | 4757 |
| نتا | 2659 | 79 | 239 | 2341 |
| سان | 7836 | 682 | 72 | 7082 |
| دلا | 1937 | 1378 | 51 | 508 |
| احا | 4792 | 2999 | 446 | 1347 |
| وضا | 2997 | 2889 | 105 | 3 |
| أاف | 1481 | 67 | 0 | 1414 |
| و لك | 1398 | 1397 | 161 | -160 |
| كلة | 2595 | 74 | 487 | 2034 |
| االف | 3929 | 2135 | 0 | 1794 |
| اثر | 1531 | 1186 | 185 | 160 |
| بدا | 2073 | 1453 | 218 | 402 |
| غاد | 1326 | 0 | 3 | 1323 |
| علي | 3470 | 1929 | 149 | 1392 |
| مية | 4855 | 2 | 217 | 4636 |
| يكن | 1566 | 1403 | 1559 | -1396 |
| االر | 5723 | 4264 | 1 | 1458 |
| ابع | 2834 | 1236 | 360 | 1238 |
| رعا | 1703 | 2 | 65 | 1636 |
| باء | 3803 | 855 | 1 | 2947 |
| بكل | 1169 | 1119 | 1 | 49 |
| اقد | 2034 | 1885 | 0 | 149 |
| لدس | 1291 | 24 | 10 | 1257 |
| اسع | 1742 | 1595 | 32 | 115 |
| لعو | 2421 | 110 | 207 | 2104 |
| سود | 2116 | 356 | 70 | 1690 |
| عدي | 2268 | 116 | 22 | 2130 |
| يكر | 1234 | 1153 | 52 | 29 |
| وشن | 1254 | 1112 | 9 | 133 |
| انط | 1329 | 100 | 1 | 1228 |
| شطن | 1227 | 0 | 131 | 1096 |
| حاس | 1608 | 1249 | 178 | 181 |
| اغر | 1697 | 1616 | 4 | 77 |

*Suite en page suivante*

TABLE 0.4 – *Suite de la page précédente*

| 3-1-gappy sequence | Freq Totale | Freq Mid | Freq Fin | Freq Debut |
|---|---|---|---|---|
| غبي | 1996 | 455 | 5 | 1536 |
| افل | 3366 | 3105 | 215 | 46 |
| للس | 3283 | 33 | 4 | 3246 |
| فسط | 4107 | 965 | 1 | 3141 |
| لطي | 4270 | 20 | 43 | 4207 |
| سين | 5536 | 316 | 712 | 4508 |
| ااح | 2750 | 2499 | 0 | 251 |
| لحد | 2494 | 330 | 75 | 2089 |
| لوم | 2828 | 35 | 15 | 2778 |
| مرس | 1726 | 1229 | 1 | 496 |
| شصا | 1085 | 1081 | 0 | 4 |
| اسو | 3651 | 3320 | 41 | 290 |
| لود | 2422 | 39 | 182 | 2201 |
| سدا | 2255 | 165 | 11 | 2079 |
| أغا | 2007 | 1206 | 0 | 801 |
| عيه | 2304 | 2216 | 58 | 30 |
| لها | 3928 | 174 | 392 | 3362 |
| ننني | 2861 | 1075 | 6 | 1780 |
| تيا | 1469 | 26 | 321 | 1122 |
| ناه | 1323 | 47 | 90 | 1186 |
| يهو | 1271 | 19 | 0 | 1252 |
| لثة | 1496 | 3 | 207 | 1286 |
| عدم | 1094 | 1012 | 363 | -281 |

Dans ce qui suit, nous présentons le benchmark utilisé pour évaluer notre algorithme d'extraction tri-littérale.

TABLE 0.5: le gold standard utilisé pour l'évaluation

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| قرا | قراءة | لجن | اللجنة | جاب | إيجابية |
| وزن | الميزانية | نظم | المنظمة | سهل | للتسهيلات |
| عام | العامة | لذع | لاذعا | وضع | وضعتها |
| جدد | الجديدة | قول | بقوله | اسس | مؤسسة |
| بدا | البداية | ضاع | أضاع | نقد | النقد |
| شار | نشير | لعب | اللاعبون | عرب | العربي |
| نتج | النتائج | جهد | جهود | اهم | لاهتم |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| مال | المالية | اعد | الاستعداد | روج | المروجون |
| عام | للعام | دقق | دقيقة | قطع | باستقطاع |
| سبق | السابق | ضيف | مضيفا | اهم | مهم |
| عبر | تعتبر | سجل | يسجلوا | مال | أموال |
| جاد | جيدة | وصل | واصلوا | دول | الدولة |
| حقق | تحقق | نفس | بنفس | غنى | الغنية |
| فاض | فائض | قنع | واقتنع | نمى | لتنمية |
| بلغ | بلغ | بعد | ابتعدوا | حقق | حقيقية |
| رغم | الرغم | نفس | المنافسة | خلق | وخلق |
| رفع | ارتفاع | أكد | تأكدنا | عمل | عمل |
| بعض | بعض | بعد | ابتعدنا | مكن | يمكن |
| صرف | مصروفات | ذهب | الذهب | قبل | يقبل |
| حكم | الحكومة | سأل | مسألة | على | عليها |
| سبب | بسبب | صعد | الصعود | هجر | المهاجرون |
| طور | التطورات | توج | التتويج | دول | الدول |
| أخر | الأخيرة | علم | علم | لكن | ولكن |
| في | في | غيب | الغيب | حدث | يحدث |
| نطق | المنطقة | حلف | يحالف | نوع | نوع |
| غطي | وتغطية | وفق | التوفيق | كسر | تكسير |
| نفق | نفقات | حقق | ويحقق | حجز | الحواجز |
| جنب | الجوانب | قبل | المقبلة | مهد | وتمهيد |
| أمن | الأمنية | فرص | والفرصة | أرض | الأرض |
| لتي | التي | سنح | سانحة | شرك | الشركات |
| فرض | فرضتها | شور | مشوار | عدد | متعددة |
| ظرف | الظروف | نصر | الانتصارات | جنس | الجنسيات |
| دخل | الداخلية | لقي | ستلتقي | عبر | والعابرة |
| اضف | بالإضافة | يمن | اليمن | تحد | تحد |
| الى | إلى | كشف | انكشف | ثمر | استثماراتها |
| سدد | تسديد | وضع | المتواضع | بلد | بلاد |
| لزم | التزامات | صمد | صمد | جدد | جديدة |
| قول | المقاولين | لعب | لاعبوه | شكل | وبشكل |
| ورد | والموردين | برى | مباراتهم | جعل | يجعلها |
| جدر | تجدر | نزع | وانتزعوا | مرر | باستمرار |
| شار | الإشارة | عدل | تعادلا | ورد | موردا |
| زاد | زيادة | رأف | الرأفة | عمل | للعمالة |
| يرد | الإيرادات | سعو | يسعون | رخص | الرخيصة |
| عام | العام | عاد | لاستعادة | طاق | والطاقة |
| كان | مكنت | ثقة | الثقة | وهب | والمواهب |
| جزء | جزء | نفس | بأنفسهم | خبر | والخبرات |
| الدين | الدين | خوض | وخوض | جهز | الجاهزة |
| هذا | وهذا | عني | بمعنويات | دوم | الدوام |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| شيء | شيء | رفع | مرتفعة | سوق | سوقا |
| ضاف | بالإضافة | ظلم | وظلمت | باح | مباحا |
| حسن | تحسن | قرع | قرعة | نتج | تنتجه |
| أدا | أداء | يمن | اليمني | مود | المواد |
| قصد | الاقتصاد | فعل | فعلا | هلك | الاستهلاكية |
| سعد | السعودي | حوض | يخوض | هرب | الهروب |
| كما | كما | غضن | غضون | ترث | التراث |
| عكس | عكسته | عجز | تعجز | دعى | دعاوى |
| رقم | أرقام | نخب | المنتخبات | بدأ | بدأت |
| نتج | الناتج | علم | العالمية | طرق | تطرق |
| محل | المحلي | عجز | عاجزين | بوب | أبواب |
| جمل | الإجمالي | صدى | التصدي | ذات | الذات |
| نتج | كنتيجة | طرد | وطرد | فرد | الفردية |
| طبع | طبيعية | لعب | لاعبان | حول | محاولة |
| طبق | لتطبيق | عمد | اعتمدوا | دأب | دءوب |
| حزم | حزمة | خشن | الخشونة | وجب | فيجب |
| صلح | الإصلاحات | يقف | لايقافهم | عقد | واعتقادهم |
| قصد | الاقتصادية | أمر | والأمور | كان | تكون |
| فيم | وفيما | سوء | سوءا | ثقف | ثقافية |
| يخص | يخص | نسب | بالنسبة | وطن | وطنية |
| يزن | الميزانية | لقى | لقاء | فرد | فردية |
| رقم | أرقامها | صدر | الصدارة | دعى | الداعين |
| جمع | جميع | صدر | وتتصدر | هجن | التهجين |
| وقع | التوقعات | رتب | الترتيب | خلط | والاختلاط |
| سبق | سبقت | نقط | نقاط | عرق | العريق |
| صدر | صدور | فرق | بفارق | دعى | يدعون |
| حيث | حيث | هدف | الأهداف | وهم | وهم |
| وزن | توازنها | ملك | يملك | بدى | يبدون |
| بين | بيان | فتح | مفتوحة | عجب | عجبهم |
| كون | سيكون | حمل | الاحتمالات | أسف | وأسفهم |
| هنا | هناك | سعى | ستسعى | صرع | صراعات |
| عجز | عجز | حصد | حصد | حفظ | الحفاظ |
| وقع | متوقع | فرد | للانفراد | ارض | أرضهم |
| بلغ | يبلغ | صدر | بالصدارة | ذات | وذاتهم |
| قدر | قدرت | قطع | وقطع | غصب | اغتصبوا |
| بلغ | بمبلغ | شوط | شوط | قدم | يقدمون |
| صرف | والمصروفات | كبر | كبير | وصف | وصفة |
| كان | ويمكن | طرق | طريق | عجب | عجيبة |
| فهم | فهم | حرز | احراز | عاد | إعادة |
| هذا | هذا | لقى | والتقى | صيغ | صياغة |
| عجز | العجز | دور | دورات | ترخ | التاريخ |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| طار | إطار | أتى | ويأتي | جعل | يجعل |
| ساس | سياسة | جدد | الجديد | فرد | والأفراد |
| نفق | الإنفاق | عمل | بعمليات | نسى | تنسى |
| حكم | الحكومي | طور | تطويرية | دفع | يدفعها |
| قدر | قدر | وسع | واسعة | مسك | التمسك |
| قدر | قدرها | مثل | تتمثل | هوي | بهويتها |
| قدر | تقديرات | جدد | تجديد | قدم | القديمة |
| حقق | تحقيقاً | فرع | فروع | ذهب | يذهبون |
| وجه | لتوجه | نشر | المنتشرة | قول | القول |
| نحو | نحو | زمن | وتزامناً | خرع | اختراعه |
| عزز | تعزيز | دخل | إدخال | ثقف | بالثقافة |
| حفظ | للمحافظة | طور | التطور | هوي | والهوية |
| على | على | حسن | والتحسينات | لغة | واللغة |
| عدل | معدل | عمل | لعملائه | دين | والدين |
| نمو | النمو | جدر | الجدير | خلف | تخلف |
| قصد | الاقتصادي | ذكر | بالذكر | جمد | وجمود |
| فتح | وفتح | عمل | عامل | حجر | وتحجر |
| فرص | فرص | بدأ | بدأ | نسي | تناسي |
| وظف | التوظيف | نشط | نشاطه | حدث | حدث |
| شبب | للشباب | صرف | المصرفي | حرب | الحرب |
| اكد | والتأكيد | حفظ | بمحافظة | أهل | الأهلية |
| اهم | أهمية | شرك | الشركة | شمل | الشمال |
| قرر | الاستقرار | تجر | التجارية | جنب | والجنوب |
| امن | الأمني | وظف | ويوظف | ولي | الولايات |
| بلد | للبلاد | وظف | موظفاً | تحد | المتحدة |
| عبر | باعتباره | فرع | فرعاً | سقط | وإسقاط |
| ركز | ركيزة | دخل | داخل | حدث | الأحداث |
| ساس | أساسية | ملك | ويمتلك | سما | والأسماء |
| ىجب | يجب | سهم | المساهمون | نسب | والمناسبات |
| سهل | التساهل | شرك | الشريك | فعل | فعله |
| هن | التهاون | عود | ويعود | شمل | الشماليون |
| برز | وتبرز | بدء | بدء | قدم | وتقدمها |
| لمح | ملامح | شرق | الشرقية | عمر | لعمري |
| رصد | رصد | خبر | الخبر | غلط | مغالطة |
| شرع | للمشاريع | قوم | ويقوم | كبر | كبرى |
| ركز | التركيز | حال | حالياً | حرب | فالحرب |
| قطع | قطاعات | قدم | بتقديم | كان | كانت |
| علم | التعليم | وزع | موزعة | جزء | جزءا |
| درب | والتدريب | عظم | معظم | خاض | مخاض |
| صاح | والصحة | وقع | المواقع | يلد | الميلاد |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| جمل | والجميل | مهم | المهمة | طول | طويل |
| عمد | اعتماد | دعم | بدعم | مدى | المدى |
| برمج | برنامج | شرك | شركة | حرب | حرباً |
| عسكر | العسكري | ورد | الموارد | فرق | فريقين |
| هني | المهني | حسب | للحاسبات | حول | يحاول |
| شاب | شاب | حدد | المحدودة | فرض | فرض |
| عمل | العمل | وقع | وقع | سلب | الأسلوب |
| يعد | يعد | وكل | وكيل | ألف | ألفه |
| ساس | أساساً | عام | عام | حيا | حياته |
| دعم | لدعم | سوق | للتسويق | صلح | المصالح |
| فعل | فعاليات | خدم | والخدمات | شمل | للشمال |
| قطع | القطاع | صمم | لتصميم | صنع | الصناعي |
| خاص | الخاص | حسن | لتحسين | زعم | زعموا |
| اهل | وتأهيل | قدر | قدرات | حرك | حركة |
| وطن | توطين | دار | الإدارة | حرر | لتحرير |
| نحو | ناحية | حصل | حصول | عبد | العبيد |
| اخر | أخرى | شبب | الشباب | حقق | والحقيقة |
| حمل | حمل | قوى | القوى | حدث | الحديثة |
| اشر | مؤشرات | درب | تدريباً | خدم | المستخدمة |
| اهم | مهمة | طور | وتطويراً | زرع | الزراعة |
| لعل | لعل | هني | مهنياً | جعل | جعلت |
| اهم | أهمها | وقع | توقيع | عبئ | عبئا |
| نمو | نمو | حفل | احتفال | قصد | اقتصاديا |
| احل | المحلي | حضر | حضره | جمع | واجتماعيا |
| نسب | بنسبة | نظم | التنظيم | دعى | داعي |
| سعر | بالأسعار | عمد | عميد | وجه | وجهة |
| ثبت | الثابتة | وصل | الاتصالات | نظر | نظر |
| حال | حالة | نسق | التنسيقية | طور | التطور |
| جسد | ويجسد | جرا | اجراءت | وزي | موازياً |
| فعل | فاعِلية | جرس | اجراس | عشر | عشرات |
| مهد | مهدت | جزا | اجزائه | سنن | السنين |
| وضح | وأوضح | جسد | اجسادهم | هدق | بهدف |
| زاد | تزايد | جسم | اجسام | نهض | النهوض |
| دور | دور | جعل | اجعلوها | صعد | الأصعدة |
| عطف | عطفاً | جلا | اجلائهم | بنك | البنك |
| سهم | إسهامه | اجل | اجلنا | فتح | افتتح |
| يضع | يضع | جمل | اجمالى | مدن | مدينة |
| سوق | تسوق | جنح | اجنحتها | ريض | الرياض |
| امن | بأمان | جهض | اجهاض | مضي | الماضي |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| ودع | وداعاً | جهز | اجهزته | أول | الأول |
| خوف | للخوف | جهش | اجهش | ركز | لمركز |
| صبح | أصبحت | جوب | اجوبة | ميز | المتميزة |
| جمع | مجموعة | جيز | اجيز | ملك | بالمملكة |
| اول | أول | حدث | احاديث | عرب | العربية |
| زود | مزود | حاط | احاطوا | شار | يشار |
| خدم | لخدمات | حال | احالات | ركز | مركز |
| تجر | التجارة | حبا | احبائهم | سمي | سمي |
| امن | الآمنة | حبط | احباطا | اسم | الاسم |
| ملك | المملكة | شمل | يشمل | نسب | نسبة |
| تجر | للتجار | شمل | شاملة | رسم | الرسام |
| احل | المحليين | كمل | ومتكاملة | شهر | الشهير |
| شرى | الشراء | قرر | المقررات | عاش | عاش |
| عبر | عبر | دور | والدورات | قرن | القرن |
| عمد | بالاعتماد | صمم | تصميم | تسع | التاسع |
| حمى | بحماية | نسق | نسق | عشر | عشر |
| علم | المعلومات | نطق | نطاق | يلد | الميلاد |
| عمد | معتمد | علم | العالم | عرف | عرف |
| اتى | وتأتي | طبق | تطبيقه | بدع | وإبداعه |
| خطى | الخطوة | عون | بالتعاون | خال | وخياله |
| غير | غير | يقع | ويقع | خصب | الخصب |
| سبق | المسبوقة | نشر | تنتشر | صاغ | الصياغة |
| هدف | بهدف | فرع | فروعها | رسم | رسوماته |
| وفر | توفير | وظف | موظف | تحف | التحف |
| عمل | عمليات | سند | وتستند | ندر | النادرة |
| قال | وقال | لعب | لعبت | خوض | يخوض |
| جمع | المجتمعات | دعم | ودعم | نخب | منتخب |
| ذات | ذات | نتج | بالمنتجات | قدم | القدم |
| جنس | التجانس | بكر | المبتكرة | برا | مباراته |
| عرق | العرقي | كمل | المتكاملة | مسى | مساء |
| سجم | والانسجام | نظم | وأنظمة | برا | لمبارايات |
| وفق | والتوافق | عمل | الأعمال | بطل | بطولة |
| ثقف | الثقافي | شرك | شراكتها | كأس | كأس |
| هدد | مهددة | عزز | فستعزز | خلج | الخليج |
| خطر | بخطر | وقع | موقعها | سدس | السادسة |
| جمد | الجمود | جعل | وتجعلها | عشر | عشرة |
| ثبت | والثبات | نطق | منطقة | مرر | تستمر |
| خلف | والتخلف | شرق | الشرق | جري | الجاري |
| سبل | سبيل | وسط | الأوسط | سبق | ويسبقها |
| روج | الترويج | جلس | المجلس | بحر | البحرين |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| فكر | لأفكارهم | على | الأعلى | امل | الأمل |
| عاد | يعيدون | ثور | للثورة | بقي | المتبقي |
| عرف | تعريف | سلم | الاسلامية | غضب | الغاضبة |
| هوي | الهوية | عرق | العراق | رضى | ترضى |
| حدد | وتحديد | صرح | التصريحات | غير | بغير |
| معن | معاني | دلى | أدلى | هدف | أهداف |
| جمع | اجتماعية | كبر | الأكبر | بدي | بداية |
| هجر | الهجرة | صدم | لصدام | نصر | لانتصارات |
| هجر | والمهاجرين | ذكر | وذكر | قبل | مقبلة |
| جذر | والجذور | نشط | النشاطات | نسي | تنسيها |
| دعى | ويدعون | شهد | ستشهد | مرر | مرارة |
| خرع | اختراع | غير | متغيرات | خسر | الخسارة |
| ارخ | التاريخ | مرس | الممارسات | امر | الامارات |
| صاغ | صياغة | وصف | واصفا | عدل | والتعادل |
| سطر | الأساطير | اثر | للاستئثار | برا | المباراتين |
| ورث | والموروث | سلط | بالسلطة. | أول | الأوليين |
| شكل | بالشكل | حكم | الحكيم | فلت | افلت |
| فسح | يفسح | شأن | شأن | حصل | وحصل |
| فرد | الأفراد | كيف | كيفية | نقط | نقطة |
| ولا | الولاءات | عطي | تعاطي | خسر | خسر |
| بين | بين | عرض | المعارضة | كشف | كشفت |
| كثر | أكثر | جاء | جاءت | وضع | تواضع |
| وطن | وطن | ساق | سياق | سوا | مستواه |
| ثقف | ثقافة | نفس | التنافس | حنك | حنكة |
| جنح | وأجنحة | صغر | الأصغر | لعب | لاعبيه |
| هدف | والهدف | مسك | الإمساك | لعب | اللاعب |
| زعم | يزعم | فتح | بمفاتيح | ميز | تميز |
| وصل | الوصول | ملك | تمتلك | وهب | بموهبته |
| جمع | مجتمع | رصد | رصيد | دفع | دفع |
| هجن | هجين | وقع | الواقعية | شيخ | بالشيخ |
| خلط | مختلط | مرس | مارس | علن | اعلان |
| عرق | الأعراق | بشع | أبشع | قال | استقالته |
| كون | تكون | ظلم | الظلم | راس | رئاسة |
| دين | دينية | لحظ | الملاحظ | نظر | ينتظر |
| لغة | لغوية | صعد | تصعيد | نها | انتهاء |
| ربط | تربط | حمل | لحملته | دور | الدورة |
| انس | الإنسان | دعا | الدعائية | بات | باتت |
| عين | بعينه | هجم | ومهاجمته | سار | سارية |
| فخر | فخورون | وزر | للوزارات | فعل | المفعول |
| وفر | يوفر | أدا | وأدائها | فاز | فازت |
| سعد | سعداء | زمن | تزامن | لقب | باللقب |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| غاي | للغاية | سري | سريان | مرة | مرتين |
| نجز | الإنجاز | شاع | شائعات | أرض | أرضها |
| راد | الرائد | وسط | الأوساط | ساد | استاد |
| ضفا | يضفي | خطر | خطيرة | ناد | نادي |
| بعد | بعداً | ظهر | ظهر | حدد | بالتحديد |
| لبي | يلبي | عرض | استعراضات | فال | تفاءل |
| حاج | الحاجة | بدل | بديل | جهر | الجمهور |
| زاد | المتزايدة | عقد | اعتقاده | قبل | قبل |
| سوق | للمتسوقين | قرب | المقربين | طلق | انطلاق |
| نجح | سينجحون | رشد | لمرشد | أمر | الأمور |
| كبر | كبيرة | عمم | عموما | أخذ | أخذت |
| دور | بدورهم | جرم | كمجرم | نحى | منحى |
| جرا | إجراء | قسم | يتقاسمه | تمم | تماما |
| يوم | اليومية | نصف | مناصفة | علم | العلم |
| كتب | مكاتبهم | عضو | عضواً | حمل | تحمل |
| نزل | منازلهم | حسب | محسوب | رقم | الرقم |
| امن | تامين | طرف | الاطراف | قاس | القياسي |
| وقت | وقت | جوب | الاستجواب | رصد | برصيد |
| ذلك | وذلك | هدر | إهدار | لقب | ألقاب |
| فضل | بفضل | خلف | ومخالفة | عذر | عذر |
| دعم | الدعم | ثرا | لإثراء | لعب | لعب |
| هوي | هوية | نفذ | النافذين | تحت | تحت |
| حمل | حامل | ولد | ولادة | ضغط | الضغط |
| بطق | البطاقة | وجه | مواجهتين | علم | والاعلامي |
| عند | عند | قضا | القضاء | أثر | تأثر |
| عمل | معاملة | جمع | جامعة | رهب | برهبة |
| شعر | وإشعار | سلم | سلام | بدي | البداية |
| تجر | المتجر | ازم | الأزمة | برى | المباراة |
| جنب | يجنب | وقف | الموقف | أول | الأولى |
| وقع | الوقوع | نفذ | تنفيذ | قدم | يقدم |
| شكل | إشكاليات | قاد | قيادة | شيء | شيئا |
| عمل | العملاء | صوت | التصويت | يمت | يمت |
| سعد | ويساعد | لجا | اللاجئين | كرة | كرة |
| نفس | أنفسهم | خلف | خلفية | طور | المتطورة |
| غلب | التغلب | ضحا | ضحايا | صلة | بصلة |
| ظهر | ظاهرة | شار | واشار | برر | تبرير |
| خوف | الخوف | زيد | مزيد | قدم | قدمه |
| برز | برزت | نحر | انتحاري | أدى | أداء |
| سنت | السنوات | بعث | المبعوث | ساء | سيئ |
| مضي | الماضية | فصل | تفاصيل | أمر | الامارات |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| شان | بشأن | عشر | العشرات | فوز | فوزها |
| امن | أمن | وضع | الموضوع | ارخ | تاريخها |
| شخص | الشخصية | حمل | الحملة | نتج | بنتيجة |
| وقت | الوقت | خلف | الخلاف | نقد | الانتقاد |
| ذات | ذاته | عصم | عاصمة | جاء | جاء |
| طاع | يستطيع | حسد | محسود | رفع | أرفع |
| طمن | الاطمئنان | عرف | يعرف | سال | المسؤولين |
| سلم | سلامة | قبل | المستقبل | شيخ | الشيخ |
| جرا | يجرونها | قتل | قتيلا | فهد | الفهد |
| خلل | خلال | كمل | كامل | ناب | نائب |
| خار | اختيار | بلغ | البالغ | رئس | رئيس |
| رمز | رمز | درس | الدراسة | طرق | الطريقة |
| سري | سري | نمي | النامية | غرب | الغربية |
| صبح | يصبح | جمد | تجميد | بشر | البشرية |
| ثاب | بمثابة | حفل | يحتفل | قوم | وتقوم |
| بصم | بصمة | مرر | بمرور | وهب | الموهوبين |
| رقم | رقمية | خمس | خمسة | فاد | للاستفادة |
| امر | الأمر | فتح | بافتتاح | قدر | القدرة |
| شعر | يشعر | صرف | المصرفية | نفس | التنافسية |
| بين | بيانات | ماز | المتميزة | لقي | تلقي |
| حسب | حساباتهم | فرع | الفرع | شعب | الشعوب |
| امن | مأمن | دار | إدارة | وهب | مواهب |
| دخل | الدخلاء | لزم | التزام | قدر | وقدرات |
| وقع | وتتوقع | مرر | المستمر | علم | المتعلمين |
| رجع | تتراجع | قدم | بتقديم | خبر | والخبراء |
| عدل | معدلات | فضل | أفضل | بنا | أبنائها |
| حال | الاحتيال | خدم | الخدمات | روج | ويروج |
| حمل | المحتملة | شرح | شرائح | جرب | لتجربة |
| طبق | بتطبيق | شرك | وشركات | واد | وادي |
| شهد | ويشهد | علق | العلاقة | طور | لتطوير |
| قبل | إقبالاً | متن | المتينة | بحث | الأبحاث |
| زاد | متزايداً | عمل | بعملائه | طبق | وتطبيقاتها |
| حقق | حققت | عدد | عدد | سال | السؤال |
| حمل | حاملو | نشب | تنشب | عاد | المعتادة |
| بطق | بطاقاتها | ناب | أنيابها | حرك | تحرك |
| قرن | مقارنة | بقي | تبقى | دعا | الدعوة |
| وقع | المتوقع | كان | الكيان | عبر | باعتبارها |
| عزز | وتعزيز | علم | العالمي | حدث | تتحدث |
| شجع | وتشجيعهم | مصص | لامتصاص | نسب | المناسبة |
| وهب | ومواهبها | دما | دماء | وفق | وفق |

*Suite en page suivante*

TABLE 0.5 – *Suite de la page précédente*

| الجذر | الكلمة | الجذر | الكلمة | الجذر | الكلمة |
|---|---|---|---|---|---|
| صلح | لصالح | نطق | المناطق | سوي | المستويات |
| احد | المتحدة | فقر | الفقيرة | ثمن | يثمنه |
| دعو | الدعوة | نزف | ونزف | عمل | العاملين |
| طلق | تنطلق | عقل | عقولها | حفل | الاحتفال |

# Bibliographie


[AAAK04] Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1058. Association for Computational Linguistics, 2004.

[AHK+03] Lili Aunimo, Oskari Heinonen, Reeta Kuuskoski, Juha Makkonen, Renaud Petit, and Otso Virtanen. Question answering system for incomplete and noisy data. In *Advances in Information Retrieval*, pages 193–206. Springer, 2003.

[AK08] Eleftherios Avramidis and Philipp Koehn. Enriching morphologically poor languages for statistical machine translation. *Proceedings of ACL-08 : HLT*, pages 763–770, 2008.

[Ali87] Abdul Sahib Mehdi Ali. *A linguistic study of the development of scientific vocabulary in standard Arabic*. K. Paul International, 1987.

[AmE69] أبي منصور الجواليقي Abi mansour Eljawalik. المعرب من الكلام الأعجمي على حروف المعجم. دار الكتاب العربي, 1969.

[App87] Douglas E Appelt. Bidirectional grammars and the design of natural language generation systems. In *Proceedings of the 1987 workshop on Theoretical issues in natural language processing*, pages 206–212. Association for Computational Linguistics, 1987.

[ASA06] Latifa Al-Sulaiti and Eric Steven Atwell. The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11(2) :135–171, 2006.

[ASAK04] Imad A Al-Sughaiyer and Ibrahim A Al-Kharashi. Arabic morphological analysis techniques : A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3) :189–213, 2004.

[ASASK03] Hasan Muaidi Al-Serhan, R Al Shalabi, and G Kannan. New approach for extracting arabic roots. In *Information Technology, 2003. ACIT'2003. Proceedings., Arab conference on Information Technology*, volume 2, pages 42–59, 2003.

[Bau03] Laurie Bauer. Natural morphology. *Introducing linguistic morphology*, 2 :253–267, 2003.

[BC06] L Belguith and Nouha Chaaben. Analyse et désambiguïsation morphologiques de textes arabes non voyellés. *Actes de la 13ème confrence sur le Traitement Automatique des Langues Naturelles*, pages 493–501, 2006.

[Bee01] Kenneth R Beesley. Finite-state morphological analysis and generation of arabic at xerox research : Status and plans in 2001. In *ACL Workshop on Arabic Language Processing : Status and Perspective*, volume 1, pages 1–8, 2001.

[BFRB96] Andre Berton, Pablo Fetter, and Peter Regel-Brietzmann. Compound words in large-vocabulary german speech recognition systems. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 1165–1168. IEEE, 1996.

[BH60] Yehoshua Bar-Hillel. The present status of automatic translation of languages. *Readings in Machine Translation*, pages 45–77, 1960.

[BHM06] Paul Baker, Andrew Hardie, and Tony McEnery. *A glossary of corpus linguistics.* Edinburgh University Press, 2006.

[Big06] Brigitte Bigi. Cours taln informatique. Avril 2006.

[BKL04] Matthew W Bilotti, Boris Katz, and Jimmy Lin. What works better for question answering : Stemming or morphological query expansion. In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR*, volume 2004, pages 1–3, 2004.

[BLAA11] A Boudlal, A Lakhouaja, M Azzeddine, and M Abdelouafi. Alkhalil morpho sys1 : A morphosyntactic analysis system for arabic texts. *Proceedings of ACIT'2010*, 2011.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1) :1–8, 2011.

[Bor05] Stefan Bordag. Unsupervised knowledge-free morpheme boundary detection. In *Proceedings of RANLP*, volume 5, 2005.

[Bou12] Houda Bouamor. *Etude de la paraphrase sous-phrastique en traitement automatique des langues.* PhD thesis, Université Paris Sud-Paris XI, 2012.

[Buc02] Tim Buckwalter. Buckwalter {Arabic} morphological analyzer version 1.0. 2002.

[Can11] Burcu Can. *Statistical Models for Unsupervised Learning of Morphology and POS Tagging.* PhD thesis, University of York, 2011.

[CHK+07] Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1) :3, 2007.

[Cho57] Noam Chomsky. *Syntactic Structures.* Mouton, The Hague, 1957.

[DCFVM+07] Iria Da Cunha, Silvia Fernandez, Patricia Velázquez Morales, Jorge Vivaldi, Eric SanJuan, and Juan Torres-Moreno. A new hybrid summarizer based on vector space model, statistical physics and linguistics. *MICAI 2007 : Advances in Artificial Intelligence*, pages 872–882, 2007.

[Déj98] Hervé Déjean. Morphemes as necessary concept for structures discovery from untagged corpora. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 295–298. Association for Computational Linguistics, 1998.

[Del00] Lionel Delafosse. Glossaire taln informatique. 2000.

[Dem07] Vera Demberg. A language-independent unsupervised model for morphological segmentation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 920, 2007.

[DPDS08] Guy De Pauw and Gilles-Maurice De Schryver. Improving the computational morphological analysis of a swahili corpus for lexicographic purposes. *Lexikos*, 18(1), 2008.

[dSBSR83] F. de Saussure, C. Bally, A. Sechehaye, and A. Riedlinger. *Course in general linguistics*. Open Court Classics. OPEN COURT Publishing Company, 1983.

[EM67] الحلمي موسى Elhilmi Moussa. دراسة احصائية لجذور معجم الصحاح باستخدام الكمبيوتر. مطبوعات جامعة الكويت, 1967.

[EZ01] البدراوي، زهران Elbedrawi Zahran. الشركة. لغويات. التصريف الملوكي. المصرية العالمية للنشر، لونجمان, 2001.

[GGJ06] Sharon Goldwater, Tom Griffiths, and Mark Johnson. Interpolating between types and tokens by estimating power-law generators. *Advances in neural information processing systems*, 18 :459, 2006.

[Gol06] John Goldsmith. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(04) :353–371, 2006.

[Gru04] Peter Grunwald. A tutorial introduction to the minimum description length principle. *arXiv math*, 2004.

[Hab10] Nizar Y Habash. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1) :1–187, 2010.

[Har54] Zellig Harris. Distributional structure. *Word*, 10(23) :146–162, 1954.

[Har55] Zellig S Harris. From phoneme to morpheme. *Language*, 31(2) :190–222, 1955.

[Har91] Donna Harman. How effective is suffixing ? *JASIS*, 42(1) :7–15, 1991.

[HB11] Harald Hammarström and Lars Borin. Unsupervised learning of morphology. *Computational Linguistics*, 37(2) :309–350, 2011.

[Hei10] Ilana Heintz. *Arabic Language Modeling with Stem-Derived Morphemes for Automatic Speech Recognition*. PhD thesis, The Ohio State University, 2010.

[HS02] Martin Haspelmath and Andrea D Sims. *Understanding morphology*. Arnold London, 2002.

[HSZ+04] Jan Hajic, Otakar Smrz, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. Prague arabic dependency treebank : Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117, 2004.

[HW74] Margaret A Hafer and Stephen F Weiss. Word segmentation by letter successor varieties. *Information storage and retrieval*, 10(11) :371–385, 1974.

[IK12] Mohamed Hadi Maaloul Lamia Hadrich Belguith Iskandar Keskes, Mohamed Mahdi Boudabous. étude comparative entre trois approches de résumé automatique de documents arabes. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, pages 225–238, Grenoble, France, 2012.

[IV98] Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation : the state of the art. *Computational linguistics*, 24(1) :2–40, 1998.

[JM08] Daniel Jurafsky and James H Martin. Speech and language processing : An introduction to speech recognition. *Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, ISBN*, 10(0131873210) :794–800, 2008.

[Jon97] Karen Spärck Jones. *Readings in information retrieval*. Morgan Kaufmann, 1997.

[JP05] Kalervo Järvelin and Ari Pirkola. Morphological processing in mono-and cross-lingual information retrieval. *Inquiries into Words, Constraints and Contexts*, pages 214–226, 2005.

[Kat06] Stonham J. Katamba, F. *Modern Linguistics Morphology (2nd ed.)*. New York : Palgrave Macmillan, 2006.

[KBH10] Nouha Chaâben Kammoun, Lamia Hadrich Belguith, and Abdelmajid Ben Hamadou. The morph2 new version : A robust morphological analyzer for arabic texts. In *10th International Conference on the Statistical Analysis of Textual Data (JADT 2010), Rome (Italy)*, 2010.

[KH96] كشلي، حكمت Kashli Hikmet. دار الكتاب العين للخليل بن أحمد الفراهيدي. الكتب العلمية، 1996.

[KH07] Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, volume 868, page 876. Prague, 2007.

[Kho01] Shereen Khoja. Apt : Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at NAACL*, pages 20–25, 2001.

[Kho12] Ahmed Khorsi. Effective unsupervised arabic word stemming : Towards an unsupervised radicals extraction, 2012.

[Khr06] Laila Khreisat. Arabic text classification using n-gram frequency statistics a comparative study. In *Conference on Data Mining— DMIN*, volume 6, page 79, 2006.

[Kir01] George Anton Kiraz. *Computational nonlinear morphology : with emphasis on semitic languages*. Cambridge University Press, 2001.

[KKJ05] Kimmo Kettunen, Tuomas Kunttu, and Kalervo Järvelin. To stem or lemmatize a highly inflectional language in a probabilistic ir environment ? *Journal of Documentation*, 61(4) :476–496, 2005.

[Kos83] Kimmo Koskenniemi. Two-level model for morphological analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 683–685, 1983.

[KP06] Samarth Keshava and Emily Pitler. A simpler, intuitive approach to morpheme induction. In *Proceedings of 2nd Pascal Challenges Workshop*, pages 31–35, 2006.

[Kro93] Robert Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202. ACM, 1993.

[KVB+06] Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke. Morphology-based language modeling for conversational arabic speech recognition. *Computer Speech & Language*, 20(4) :589–608, 2006.

[KY05] Katrin Kirchhoff and Mei Yang. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128. Association for Computational Linguistics, 2005.

[Lam04] Andrew Lampert. A quick introduction to question answering, 2004.

[Lan68] Frederick Wilfrid Lancaster. Information retrieval systems ; characteristics, testing and evaluation. 1968.

[Lap00] Eric Laporte. Mots et niveau lexical. *Ingénierie des langues*, pages 25–49, 2000.

[LB56] William N Locke and A Donald Booth. Machine translation of languages. *American Documentation*, 7(2) :135–136, 1956.

[LBC02] Leah S Larkey, Lisa Ballesteros, and Margaret E Connell. Improving stemming for arabic information retrieval : light stemming and co-occurrence analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–282. ACM, 2002.

[LCMY09] Constantine Lignos, Erwin Chan, Mitchell P Marcus, and Charles Yang. A rule-based unsupervised morphology learning framework. In *Working Notes, CLEF 2009 Workshop*. Citeseer, 2009.

[LH13] Abdellah Lakhdari and Cherroun Hadda. Effective unsupervised morphological analysis and modeling : Statistical study for arabic language. In *BOOK OF ABSTRACTS OF THE 23RD MEETING OF COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS : CLIN 2013*, page 85, 2013.

[LL09] Jean-Francois Lavallee and Philippe Langlais. Morphological acquisition by formal analogy. In *Working Notes, CLEF 2009 Workshop*, 2009.

[McD92] D. D. McDonald. Type-driven suppression of redundancy in the generation of inference-rich reports. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation : Proc. of the 6th International Workshop on Natural Language Generation*, pages 73–88. Springer, Berlin, Heidelberg, 1992.

[Mit05] R. Mitkov. *The Oxford Handbook of Computational Linguistics*. Oxford Handbooks. OUP Oxford, 2005.

[MP90] John J McCarthy and Alan S Prince. Foot and word in prosodic morphology : The arabic broken plural. *Natural Language & Linguistic Theory*, 8(2) :209–283, 1990.

[MS99] C.D. Manning and H Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[MT90] P. Miller and T. Torris. Formalismes syntaxiques pour le traitement automatique du langage naturel. 1990.

[NF02] Sylvain Neuvel and Sean A Fulop. Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pages 31–40. Association for Computational Linguistics, 2002.

[Ngu05] N.H. Nguyen. *Dialogue homme-machine : modélisation de multisession*. 2005.

[NK08] Ngan LT Nguyen and Jin-Dong Kim. Exploring domain differences for the design of pronoun resolution systems for biomedical text. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 625–632. Association for Computational Linguistics, 2008.

[Oet60] Anthony G Oettinger. Automatic language translation : Lexical and technical aspects, with particular reference to russian. harvard monographs in applied science, number 8. 1960.

[OvHdJ03] Roeland Ordelman, Arjan van Hessen, and Franciska de Jong. Compound decomposition in dutch large vocabulary speech recognition. In *Proc. Eurospeech*, pages 225–228. Citeseer, 2003.

[Pin99] Bénédicte Pincemin. Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative. *A. Condamines, M.-P. Péry-Woodley & C. Fabre (éds), Atelier Corpus et TAL : pour une réflexion méthodologique*, pages 26–36, 1999.

[Por80] Martin F Porter. An algorithm for suffix stripping. *Program : electronic library and information systems*, 14(3) :130–137, 1980.

[Rey77] A. Rey. *Le lexique : images et modèles*. Linguistique (Paris. 1973). A. Colin, 1977.

[Ris78] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5) :465–471, 1978.

[RS07] Brian Roark and Richard William Sproat. *Computational approaches to morphology and syntax*. Oxford University Press, 2007.

[SA08] Majdi Sawalha and ES Atwell. Comparative evaluation of arabic language morphological analysers and stemmers. In *Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics (Poster Volume)*, pages 107–110. Coling 2008 Organizing Committee, 2008.

[Sal68] Gerard Salton. Automatic information organization and retrieval. 1968.

[SAM09] M. Sabir and A.M Abdul-Mun'im. Midad morphological analyzer for arabic text. In *workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Damascus, Syria*. Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology ( KACST) and Arabic Language Academy, 2009.

[Saw11] Majdi Shaker Salem Sawalha. *Open-source resources and standards for Arabic word structure analysis : Fine grained morphological analysis of Arabic text corpora*. University of Leeds, 2011.

[SB08] Benjamin Snyder and Regina Barzilay. Unsupervised multilingual learning for morphological segmentation. *Proceedings of ACL-08 : HLT*, pages 737–745, 2008.

[SBK+08] Otakar Smrz, Viktor Bielickỳ, Iveta Kourilová, Jakub Krácmar, Jan Hajic, and Petr Zemánek. Prague arabic dependency treebank : a word on the million words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, 2008.

[SD11] Ghnaim N. Sonbul, R. and M. S Dusouqi. An application oriented arabic morphological analyzer. In *workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Damascus, Syria*. Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology ( KACST) and Arabic Language Academy, 2011.

[SE16] شهاب الدين أحمد الخفاجي Shihabeddin Elkhafaji. شفاء الغليل فيما كلام العرب من الدخيل. India, 1916.

[Seb99] F. Sebastiani. A tutorial on automated text categorisation. In *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, pages 7–35. Buenos Aires, AR, 1999.

[Smr07] Otakar Smrž. Elixirfm : implementation of functional arabic morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to*

*Semitic Languages : Common Issues and Resources*, pages 1–8. Association for Computational Linguistics, 2007.

[Wea55] Warren Weaver. Translation. *Machine translation of languages*, 14 :15–23, 1955.

[Yvo10] François Yvon. Une petite introduction au traitement automatique des langues naturelles, 2010.