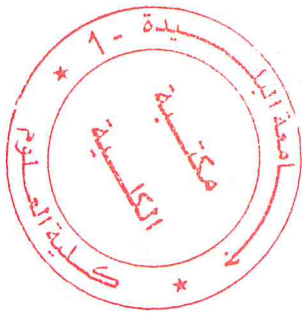
**Ministère de l'enseignement supérieur et de la recherche scientifique**

**UNIVERSITE SAAD DAHLEB DE BLIDA 1**

**Faculté des Sciences**

**Département d'informatique**




**MEMOIRE DE MASTER**

**En Informatique**

**Spécialité : Ingénierie du Logiciel**

**THEME**

# *Mesures de similarité sémantique pour un système d'évaluation automatique des réponses courtes : Application à la langue arabe*

**Réalisé par :**

ATOUB Yasmine

BENAYAD Asma

**Proposé et encadré par :**

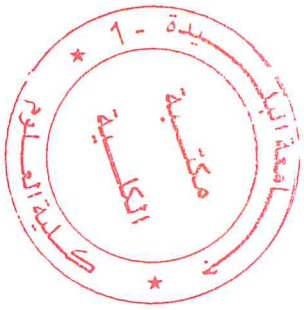
Mme. OUAHRANI Leila

**Composition de jury :**

| | |
|---|---|
| M. BALA Mahfoud | Président |
| M. KAMECHE Abdallah Hicham | Examinateur |

**Soutenu le :**

26/09/2018




# الملخص

في مجال التعليم، احتل التعليم الإلكتروني مكانة مرموقة بفضل خدماته المقدمة و المتمثلة في توفير الوقت والجهد وكذا التكلفة. ومع ذلك، فإن تقييم هذا التعلم ضروري و تشغيله بصفة آلية يعد موضع تقدير. في هذا العمل، نحن مهتمون بالتقييم التلقائي أو الآلي للإجابات القصيرة في اللغة الطبيعية نظرا لكونها تعكس بنزاهة و على نحو محايد درجة إتقان التعلم. نقترح طريقة لتصميم نظام التقييم الأوتوماتيكي للإجابات القصيرة متكيف مع اللغة العربية و القائم على أساس مفهوم نهج الناقلات والتي تسمح ببناء فضاء دلالي للكلمات وبالتالي اقتراح العديد من نماذج حساب التشابه الدلالي باستخدام هذا الفضاء الدلالي. نعتمد تقنيات معالجة اللغة الطبيعية بما في ذلك النهج الجذعي. يتم إجراء عملية التهجين مع قياسات نحوية وكذلك مع قياسات مبنية و مطورة حول تضمين الكلمة مما أعطى نتائج أفضل.

**الكلمات المفتاحية:** التقييم التلقائي للإجابات القصيرة ، جذر الكلمة ، ASAGS، معالجة اللغات الطبيعية ، NLP، قياسات التشابه ، التشابه الدلالي ، فضاء الناقلات ، الفضاء الدلالي ، ناقلات السياق.

# Résumé

Dans le domaine éducatif, l'apprentissage en ligne a marqué sa place grâce à ses services offerts notamment le gain du temps, d'efforts et du coût. Cependant, l'évaluation de cet apprentissage est nécessaire et son automatisation est plus appréciée. Dans ce travail, nous nous intéressons à l'évaluation automatique des réponses courtes en langage naturel car ces dernières reflètent impartialement le degré de la maitrise de l'apprentissage. Nous proposons une méthode de conception d'un système d'évaluation automatique des réponses courtes adapté à la langue arabe basée sur l'approche vectorielle permettant de construire un espace sémantique des mots et de proposer plusieurs modèles de calcul de similarité sémantique en utilisant cet espace sémantique. Nous adoptons des techniques du traitement du langage naturel notamment l'approche par stemming. Une hybridation avec des mesures syntaxiques ainsi qu'avec des mesures développées autour des Word Embedding est faite. Ce qui a donné encore de meilleurs résultats.

**Mots clé :** Evaluation automatique des réponses courtes, ASAGS, Stem, traitement du langage naturel, TALN, mesures de similarité, similarité sémantique, espace vectoriel, espace sémantique, vecteur de contexte.

# Abstract

In the educational field, e-learning has marked its place thanks to its services offered noting the gain of time, effort and cost. However, the evaluation of this learning is necessary and its automation is more appreciated. In this work, we are interested in the automatic evaluation of short answers in natural language because they reflect impartially the degree of mastery of learning. We propose a method of designing a system for automatic evaluation of short answers adapted to the Arabic language based on the vectorial approach allowing to construct a semantic space of the words and to propose several models of calculation of semantic similarity by using this semantic space. We adopt natural language processing techniques including the stemming approach. A hybridization with syntactic measurements as well as with measures developed around Word embedding is made. This gave even better results.

**Keywords:** automatic short answer grading, ASAGS, Stem, natural language processing, NLP, similarity measurements, semantic similarity, vector space, semantic space, context vector

# Remerciements

Nous tenons tout d'abord à remercier **Allah** le tout puissant, qui nous a donné la force, la capacité et surtout la patience pour accomplir ce travail.

Nous tenons à adresser nos plus chaleureux remerciements à Madame « Ouahrani Leila », notre promotrice de ce PFE, pour son aide illimitée et son soutien pendant toute cette période pleine de défis. Ses conseils avisés et sa patience nous ont aidés à surmonter l'hésitation, l'embarras et de rediriger le projet pour avoir de meilleurs résultats. Son œil critique nous a été très précieux pour structurer le travail et pour améliorer la qualité des différentes sections sans oublier sa constante disponibilité à notre égard et de nous avoir donné l'occasion de travailler sur un tel sujet de recherche qui est très riche d'information et de découverte pour nous.

Notre gratitude s'adresse également à l'université de Bouira qui nous a fournit un accès sur un serveur à distance pour accomplir notre travail, ainsi qu'à M. Wael Hassan Gomaa et M. Aly Aly Fahmy pour le partage de leurs dataset.

Nos vifs et chaleureux remerciement et gratitude à nos parents, frères, sœurs ainsi que toute la famille Beriber et Kartout pour leur soutien, encouragement et amour illimité. « Elhamdoullilah de vous avoir à nos cotés, nous vous aimons du fond du cœur ».

Nous tenons à exprimer nos sincères remerciements à nos amies pour leur soutien et encouragement qu'avec, nous avons pu surpasser des périodes de stress. Sans oublier nos collègues Adel, Hamza, Amina et Khadidja pour cette agréable collaboration et ce travail en groupe…Merci à vous.

Nous tenons à remercier toute personne qui a participé de prés ou de loin à l'exécution de ce modeste travail.

# Liste des figures

# Liste des tables

# Liste des abréviations

| | |
|---|---|
| **ASAGS** | **Système d'évaluation automatique des réponses courtes (Automatic Short Answer Grading System)** |
| **RM** | **Réponse Modèle** |
| **RE** | **Réponse Etudiant** |
| **NLP/TALN** | **Traitement automatique du langage naturel** |

# Table des matières

# Chapitre 1 : Contexte et problématique



## 1.1 Introduction générale

L'évaluation est l'étape la plus importante dans la démarche d'un projet dans tous les domaines. Cette importance est due à son efficacité afin de mesurer la qualité du travail fait, mettre des perspectives d'amélioration et les appliquer si nécessaire, valoriser le travail, déterminer les avantages / inconvénients des actions prises pour les adapter / éviter dans le futur ainsi que la prise de décision.

Dans le domaine éducatif, l'évaluation garde son rôle essentiel dans le cursus des étudiants. Néanmoins, cette tâche présente une charge contraignante pour les enseignants en termes de temps, de concentration et de précision. Pour cela les chercheurs se sont focalisés sur le domaine de l'évaluation afin d'automatiser cette tâche et par la suite améliorer la performance de l'apprentissage des étudiants, ainsi que la réduction de la charge des enseignants et enfin intégrer la culture d'évaluation au travail quotidien des apprenants dans un environnement du e-Learning.

Des systèmes d'évaluation automatique des réponses courtes (à travers des examens écrits) sont en pratique dans le domaine de l'enseignement -particulièrement l'enseignement

en ligne-depuis de nombreuses années. Toutefois, cela est confiné autour des questions comme celles à choix multiple (QCM) ou à réponses vrai/faux que les recherches ont jugé comme insuffisantes car elles ne permettent pas de saisir les multiples aspects des connaissances acquises, comme le raisonnement et l'auto-explication. En revanche, les questions à réponses courtes qui nécessitent des réponses données par les examinés en langage naturel ont été jugées plus efficaces pour évaluer les connaissances acquises par les apprenants.

Cependant, la réalisation d'un système d'évaluation automatique pour les réponses courtes noté ASAGS (pour Automatic Short Answer Grading System) qui traite les réponses dans la langue arabe présente un défi en raison de sa complexité vis-à-vis du coté morphologique[1] et sémantique[2] ainsi que la variation linguistique[3] et la nature subjective[4] de l'évaluation.

En outre, le type de question a un impact important sur la réponse. Chaque question nécessite un type précis de réponse. Nous nous intéressons aux questions à réponses ouvertes[5]. La catégorie des réponses ouvertes permet de recueillir des réponses qualitatives et la plupart du temps riches en informations ce qui signifie sa complexité. Les questions à réponses courtes font partie des questions ouvertes qui regroupent aussi les essais :

- **Les essais :**

  Ce type est caractérisé par la longueur et la profondeur des idées présentées. En raison du nombre limité de mots autorisés, les idées dans un court essai devraient être présentées clairement et succinctement. Un court essai devrait être d'environ cinq cents mots. Il est censé répondre à une question ou un argument. S'il s'agit d'un débat, une déclaration de thèse claire devrait être fournie pour montrer la partie prise dans l'argumentation et les arguments attendus à soulever. Les courts essais présentent souvent des opinions ou des opinions individuelles. Les idées sont présentées superficiellement puisque la durée de l'essai est limitée. [6]

---

[1] La morphologie est l'étude de la formation des mots et de leurs variations. Autrement, c'est le regroupement de différents mots à travers leurs parties, comme les suffixes, préfixes, radicaux

[2] La sémantique lexicale est l'étude du sens des mots -ou plutôt des morphèmes- d'une langue donnée.

[3] Une réponse donnée pourrait être articulée de différentes façons qui ont tous le même sens.

[4] Une question peut avoir de multiples réponses possibles.

[5] Une question ouverte est une question pour laquelle il n'y a pas de réponses préétablies proposées au répondant, celui-ci est donc entièrement libre dans sa réponse.[59]

▪ **Les réponses courtes :**

Les réponses courtes sont caractérisées par les critères suivants:

- La question doit exiger une réponse donnée en langage naturel,
- La question nécessite une réponse qui rappelle des connaissances externes au lieu d'exiger que la réponse soit reconnue à l'intérieur de la question elle-même,
- La longueur de la réponse devrait être approximativement entre une phrase et un paragraphe (100 mots au maximum),
- L'évaluation des réponses doit se concentrer sur le contenu plutôt que sur le style d'écriture (style de réponse),
- Une réponse courte doit être précise et objective. Elle doit être relative aux concepts introduits en question.

De plus, les réponses courtes s'appuient fortement sur la similitude sémantique des deux réponses à évaluer, celle de l'étudiant et celle de l'enseignant, constituant le modèle de réponse correcte. Dans ce dernier cas, les humains sont en mesure de juger facilement si un des concepts sont liés les uns aux autres. Par contre, la machine confronte un problème lorsque l'étudiant utilise un mot synonyme au cours de sa réponse (au cas où ils oublient la réponse cible et utilise leur alternative des mots dans la réponse qui seront différents de la réponse modèle). Contrairement aux essais, les réponses courtes nécessitent la présence d'une réponse modèle dans un système d'évaluation automatique.

## 1.2 Problématique

Le concept principal des ASAGS consiste à avoir une réponse modèle/type (RM) donnée par l'enseignant et une réponse de l'étudiant dans un examen sous certaines conditions. Ces réponses doivent être en langage naturel. Ensuite, une valeur de similitude est calculée en appliquant les mesures de similarité sur les deux réponses. Cette valeur est enfin convertie en note sur un barème donné en introduisant une technique de passage de la similarité au score.

En outre, pour conforter la nécessité des ASAGS, il faut prendre en considération les difficultés que les enseignants affrontent au quotidien comme la fatigue, la concentration, la

précision dans la notation (donner 2, 4.75, 3.5,...pour une réponse sur 5 points de barème) ainsi que les formes d'écriture manuscrite (écriture illisible) qui peuvent engendrer une mauvaise évaluation de l'étudiant. En revanche, les recherches s'accordent à dire que tous les types d'approches qui existent jusqu'à maintenant n'arrivent pas encore à donner de meilleurs résultats.

D'autre part, la plupart des ASAGS existants traitent de l'anglais. De plus, l'arabe est une langue assez répandue, parlée par plus de 300 millions de personnes à travers le monde. Du point de vue du langage naturel, la langue arabe se caractérise par une ambiguïté élevée et une morphologie riche et complexe sans oublier le manque considérable de ressources linguistiques : corpus arabes, lexiques et dictionnaires, outils de traitement NLP... Tous ces aspects repoussent et découragent le progrès de la recherche dans la branche de l'évaluation automatique des réponses courtes en langue arabe. C'est pour cela, peu de travaux ont été réalisés dans ce contexte.

## 1.3 Objectifs du travail

Le travail réalisé dans le cadre de ce projet de fin d'étude consiste à :

- ✓ Etudier les diverses techniques et systèmes d'évaluation automatiques des réponses courtes,
- ✓ Etudier plusieurs approches de mesures de similarité dans le contexte de l'évaluation automatique,
- ✓ Construire un système d'évaluation automatique qui est sensé être capable à donner une évaluation (note) approximativement proche à celle donnée par l'expert humain (enseignant),
- ✓ Faire une hybridation avec des mesures syntaxiques sémantiques développées par deux autres PFEs dans le même projet de système d'évaluation automatique.
- ✓ Evaluer la précision (exactitude) du système construit en se basant sur des stratégies d'évaluation utilisant des datasets.

## 1.4 Importance de notre travail

Dans le cadre de ce travail, nous visons à apporter un plus à ce qui a été déjà fait. Autrement dit: réaliser un système d'évaluation automatique des réponses courtes qui introduit le concept de similarité sémantique (le sens des phrases), et particulièrement, un système de traitement de la langue arabe. Pour cela, nous devons adapter des techniques existantes dans le domaine du traitement automatique du langage naturel (NLP).

D'une part, comme mentionné précédemment, la langue arabe souffre du manque de ressources linguistiques (notamment le WordNet[1] arabe qui manque de richesse par rapport à son homologue anglais ainsi que les dictionnaires arabe en ligne). Par conséquent, nous sommes limités à certaines méthodes, mais aussi nous nous adaptons avec le peu d'outils NLP disponibles sur le net bien qu'ils ne sont pas vraiment performants. Ce dernier nous a motivé à réaliser notre propre outil NLP (Natural Language Processing), un plus à d'autres travaux futurs.

D'autre part, les ASAGS réduisent considérablement le besoin d'implication humaine. Cela a un impact important sur le domaine de l'éducation, ce qui facilite ainsi une partie de la charge de l'enseignant, améliore ses performances et détermine si ses objectifs ont été atteints.

Parmi les avantages de l'automatisation du marquage, mentionnons les économies de temps et de coûts, ainsi que la réduction des erreurs et des injustices attribuables aux préjugés humains, à l'épuisement ou au manque de cohérence [1].

En effet, ce travail est généralisable c'est-à-dire qu'on peut l'adapter pour n'importe quelle autre langue qui est dans les mêmes circonstances que celle de l'arabe (il faut juste avoir les données nécessaires que nous évoquerons dans les prochains chapitres).

## 1.5 Champs de notre travail et limites

- Ce travail est affilié au traitement automatique du langage naturel (TALN), qui est un domaine multidisciplinaire impliquant la linguistique, l'apprentissage automatique et

[1] WordNet est une base de données de l'anglais regroupant des unités lexicales selon leurs relations sémantiques et lexicales.

donc l'intelligence artificielle. De plus, nous nous intéressons au domaine éducatif qui constitue la base du savoir et de la moralisation,

- Ce système exige la disponibilité d'une réponse modèle (réponse de référence),
- Les réponses doivent être saisisses par clavier,
- La disponibilité d'un corpus lié au domaine de la recherche ainsi qu'un stemmer de traitement du texte arabe est nécessaire dans notre démarche,
- Le système développé est dédié pour l'évaluation sur la langue arabe.

## 1.6 Structure du mémoire

Ce mémoire est divisé en quatre chapitres principaux :

- Chapitre 1 : Contexte et problématique,
- Chapitre 2 : Etat de l'art traitant des outils d'évaluation automatique des réponses courtes, des mesures de similarité et des travaux connexes,
- Chapitre 3 : Conception du Système d'évaluation automatique des réponses courte et développement de l'approche méthodologique et des modèles de calcul de similarité,
- Chapitre 4 : Résultat expérimentaux et évaluation du système,
- En fin, une conclusion et des perspectives.

# Chapitre 2 : Etat de l'art

La phase élémentaire dans l'initiation d'une étude ou d'une recherche commence par l'état de l'art sur le sujet abordé. Cela consiste à reformuler toutes les connaissances acquises ainsi que les travaux déjà faits sur la même thématique. Cette étape présente la base de prise de décision concernant la méthode et la technique adaptées dans cette recherche.

Dans ce chapitre nous allons entamer le fonctionnement global des ASAGS, les approches d'évaluation automatique de réponses courtes particulièrement celles qui traitent l'arabe, les approches de similarité existantes en général et précisément celles de la similarité sémantique, les outils et ressources NLP existants et disponibles en ligne ainsi que les travaux connexes.

## 2.1 Fonctionnement global des ASAGS

### 2.1.1 Principe

Le principe du fonctionnement d'un ASAGS repose sur un ensemble de processus. Ce système doit avoir la question à poser bien formulée, claire et dans un contexte précis d'un examen réel convenable aux conditions disant ordinaires. De plus, une réponse modèle et sa note données par l'expert du domaine (l'enseignant) sont indispensables comme données de base pour le fonctionnement d'un ASAGS. Ensuite, l'utilisateur/étudiant saisie sa réponse qui doit répondre aux critères de la réponse courte cités précédemment. Pa la suite, le système évalue cette réponse en calculant la similarité entre elle et les données de base en utilisant les outils et ressources NLP, puis il convertie cette similarité en un score (note) selon le barème donnée et enfin renvoie ce résultat à l'utilisateur.

### 2.1.2 Processus principaux des ASAGS

Dans la littérature, S. Burrows et al. [2] ont proposé un pipeline de développement de système ASAG représenté par 6 artefacts (rectangles) et 5 processus (ovales) « Voir Figure 2.1.».

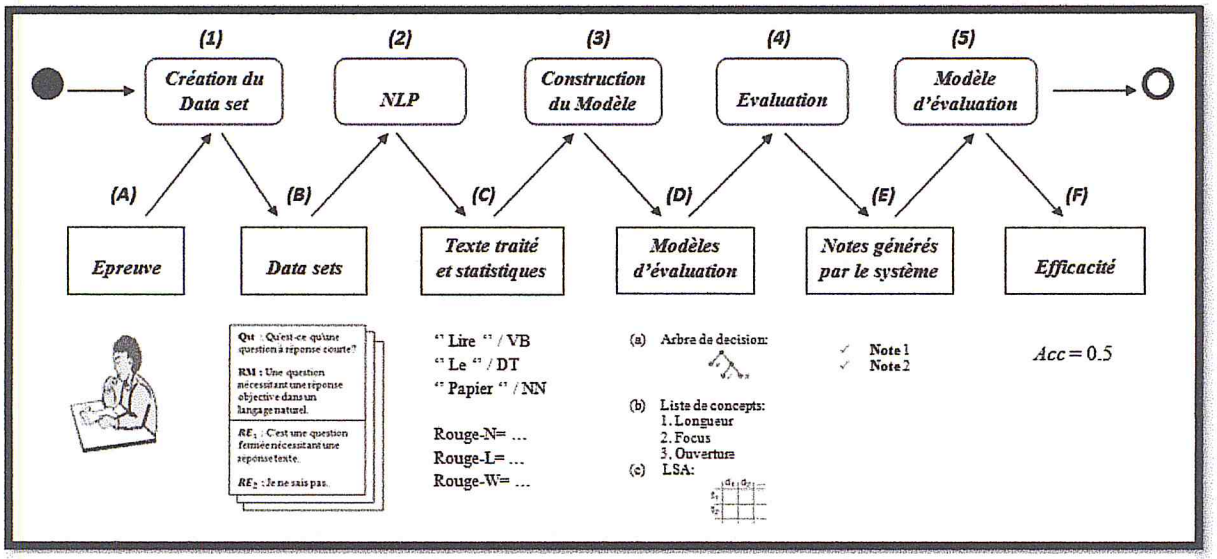



**Figure 2.1 Pipeline de développement de système ASAG [2].**

Les modules du fonctionnement d'un ASAG sont relatifs les uns aux autres :

***1) Création du dataset :***

Cette étape nécessite de faire une épreuve (A) (examen, test, interrogation…) pour plusieurs étudiants (apprenants). Les réponses collectées seront organisées avec leur réponse modèle (réponse type) pour chaque question de l'épreuve donnée par l'examinateur (l'enseignant) ainsi que l'évaluation manuelle donnée par ce dernier (notes manuelles). Toutes ces données représentent le dataset (B).

2) ***Traitement NLP du dataset*** :

Le dataset (B) passe maintenant par une phase de traitement automatique du langage naturel NLP vue que les questions de (A) imposent des réponses en langage naturel. Ce traitement comprend de différentes techniques comme le stemming, la normalisation, étiquetage morpho-syntaxique (POS : Part-Of-Speech Tagging), calcul des pondérations (TF, IDF,…)… Un ensemble de texte traité ainsi que des statistiques (C) résultent.

***3) Construction du modèle d'évaluation :***

Cette étape est l'étape la plus importante dans le fonctionnement d'un ASAG car elle est la base du calcul de similarité entre la réponse étudiant et la réponse modèle.

Les données textuelles traitées sont utilisées afin de les transformer en un modèle (ceci peut exiger des ressources externes). Il existe plusieurs types de représentation du modèle (D) comme le VSM (modèle d'espace vectoriel), les arbres de décisions, …

4) ***L'évaluation automatique du dataset :***

Après avoir construit le modèle d'évaluation, il est maintenant possible de calculer la valeur de similitude entre le couple de réponses (RM, RE). La valeur de similarité varie dans l'intervalle [0, 1]. Par la suite, cette valeur sera interprétée. Pour se faire, un passage au score est effectué pour avoir la note automatique finale (E).

5) ***L'évaluation du modèle construit :***

Dans ce stade, une estimation de la qualité du modèle construit est indispensable. Pour se faire, une valeur de précision ou d'exactitude (F) est calculée comme repère qui reflète cette qualité. Plusieurs méthodes existent déjà pour cet objectif (Coefficient de corrélation de Pearson, l'erreur quadratique RMSE, la précision…).

## 2.2 Une vue sur l'historique des ASAG

Il existe plusieurs approches qui traitent le sujet d'évaluation automatique des réponses courte dont la recherche s'est développée depuis 1996. Steven Burrows et al. dans leur article [2] ont identifié 35 systèmes dans 4 méthodes différentes (représentés dans la Figure 2.2.) :

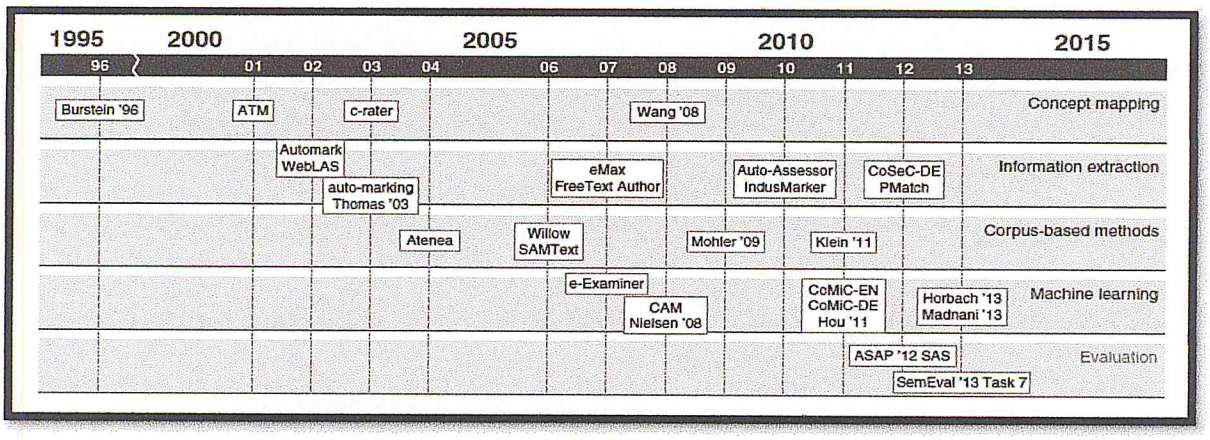



**Figure 2.2 Les ères et les tendances du classement automatique des réponses courtes [2].**

- **Les méthodes basées sur le mappage de concepts (Concept Mapping) :**

Le principe ici est de considérer les réponses des élèves comme constituées de plusieurs concepts et de détecter la présence ou l'absence du concept clé lors du classement. Cela se fait en parcourant les concepts clé (ceux de la réponse modèle) un par un dans la réponse étudiant et pour chaque concept trouvé, incrémenter le score.

Cette méthode s'adapte bien avec deux types de question. L'un demande une solution à un problème plus une justification. L'autre demande plusieurs explications au même problème.

**ATM** : Le marqueur de texte automatique ATM (Automatic texte Marker) [3] décompose les réponses des enseignants et des étudiants dans des listes de concepts minimaux comprenant pas plus de quelques mots chacune, et compte le nombre de concepts en commun pour fournir un score d'évaluation. Chaque concept est essentiellement la plus petite unité possible dans une réponse qui peut être attribuée un poids aux fins de classement. Les pondérations sont additionnées pour produire la note globale.[2]

**C-rater** : L'évaluateur conceptuel (Concept rater ou C-rater) [4] vise à faire correspondre autant de concepts de niveau de phrase que possible entre les réponses de l'enseignant et celles de l'élève. L'appariement est basé sur un ensemble de règles et une représentation canonique des textes utilisant la variation syntaxique, l'anaphore, la variation morphologique, les synonymes et la correction orthographique. Plus précisément, les réponses des enseignants sont entrées dans une phrase distincte pour chaque concept. Cela simplifie l'évaluation puisque seul un concept est considéré à la fois lors du classement. Cette technique évite d'avoir recours à une solution indirecte, telle que diviser la question en plusieurs parties [5] et l'on affirme que cela peut conduire à une plus grande précision [6] . En outre, le format d'entrée en langage naturel est avantageux par rapport à d'autres systèmes qui nécessitent une expertise et l'utilisation d'un langage de balisage [7]. [2]

- **Les méthodes basées sur l'extraction d'information (Information Extraction) :**

Le concept de cette méthode est l'extraction des données à partir des sources non structurées comme les textes libres, ensuite les mettre sous forme structurée (arbres d'analyse...). Les méthodes d'extraction d'informations peuvent être considérées comme une série d'opérations de correspondance de modèle telles que les expressions régulières ou les arbres d'analyse. Dans le cas des systèmes des réponses

courte, chaque réponse est présentée sous forme structuré et par la suite évaluer leurs dépendances.

**Auto-évaluateur**: ou Auto-Assessor [8] se concentre sur le classement (évaluation) des réponses des étudiants à phrase unique sous forme canonique basé sur la correspondance des coordonnées de sac de mots (BOW) et les synonymes avec WordNet [9]. Coordonner l'appariement dans l'ASAG se réfère simplement à des termes individuels correspondant entre les réponses des enseignants et des étudiants. Dans l'auto-évaluateur, chaque mot qui correspond exactement reçoit un point de crédit, les mots liés à partir du WordNet reçoivent un crédit partiel, tandis que le reste ne reçoit pas de crédit.

- **Les méthodes basées-corpus (Corpus Based Methods) :**

  Le principe est d'utiliser des propriétés statistiques des corpus, qui sont des ensembles de textes. Ces méthodes peuvent être utiles lors de l'interprétation des synonymes dans les réponses courtes. Alors, afin de limiter les réponses correctes qui peuvent être identifiées, utiliser seulement le vocabulaire des réponses modèles. Ensuite, pour renforcer ce vocabulaire, prendre en considération ses synonymes et effectuer des traductions en autre langue afin d'éviter les difficultés de la langue source. C'est dans ce contexte que se situe notre travail.

- **Les méthodes basées sur l'apprentissage automatique (Machine learning) :**

  Les systèmes d'apprentissage automatique utilisent généralement des données étiquetées qui sont les réponses modèles notées ainsi qu'un certain nombre de mesures extraites du langage naturel (techniques de traitement et similarité) qui sont ensuite combinées afin d'avoir un score en utilisant un modèle (fonction d'estimation) de classification ou de régression. Cela peut être soutenu par une boîte à outils d'apprentissage automatique telle que Weka [10].

  Madnani et al. [11] mettent en place un système de notation des questions de compréhension de lecture sur les niveaux de vie (système nommé Madnani '13). Chaque texte comporte trois paragraphes, et les réponses des élèves requièrent spécifiquement une phrase donnant un résumé global et trois phrases supplémentaires donnant un résumé de chaque paragraphe. L'approche d'apprentissage automatique comprend huit caractéristiques (BLEU, ROUGE, mesures concernant différentes dimensions de la copie de texte, nombre de phrases et nombre de mots de connecteur

de discours couramment utilisés) en tant qu'entrées dans un classificateur de régression logistique.[2]

## 2.3 Les modèles BOW et VSM

Dans la thématique des ASAGS, un modèle est défini comme toute représentation qui permet de mesurer le degré de similitude entre le couple (RM, RE) avec une précision raisonnable entre le score automatique et celui manuel. Fréquemment dans le domaine de l'apprentissage automatique, les données textuelles utilisées sont converties en vecteurs. Le BOW (pour Bag Of Words) est fait dans cet objectif. Le modèle de sac de mots est un moyen d'extraire des caractéristiques du texte pour les utiliser dans des algorithmes de l'apprentissage automatique, plus précisément dans le traitement automatique du langage naturel (TALN).

BOW consiste à extraire les mots uni-gramme (unigram-words) d'un document (du texte) pour créer une liste (un sac) non ordonnée (sans considérer l'ordre des mots ni de la grammaire dans le document source) de ses mots mais en gardant la multiplicité des mots (un mot peut avoir plusieurs occurrences dans la liste). Une autre manière pour se faire est de créer un dictionnaire du vocabulaire du document en gardant la fréquence/occurrence de chaque mot de ce dictionnaire.

| | Terme 1 | Terme 2 | Terme 3 | ... | Terme n |
|---|---|---|---|---|---|
| Document 1 | 1 | 1 | 0 | ... | 0 |
| Document 2 | 1 | 0 | 0 | ... | 1 |
| ... | ... | ... | ... | ... | ... |
| Document n | 0 | 0 | 0 | ... | 1 |

**Figure 2.3. Exemple de représentation VSM.**

Une fois le BOW est construit, une étape de vectorisation est généralement nécessaire. Ceci se fait en construisant une matrice dont les caractéristiques varient selon le domaine. Un exemple simple et connue est la création d'une matrice M de taille D*X où D est l'ensemble de document et X est la taille du dictionnaire du vocabulaire du corpus (l'ensemble des

lexèmes obtenus des documents). Chaque ligne i de la matrice est une représentation d'un document du corpus et chaque colonne j est un mot (terme) du dictionnaire tandis que l'élément M (i,j) représente une valeur binaire indiquant la présence ou l'absence du mot j dans le document i (1 pour présent et 0 pour absent) « Figure 2.3.». D'autres exemples de représentation seront présentés prochainement dans la « section 2.4.2. ».

## 2.4 Les approches de similarité

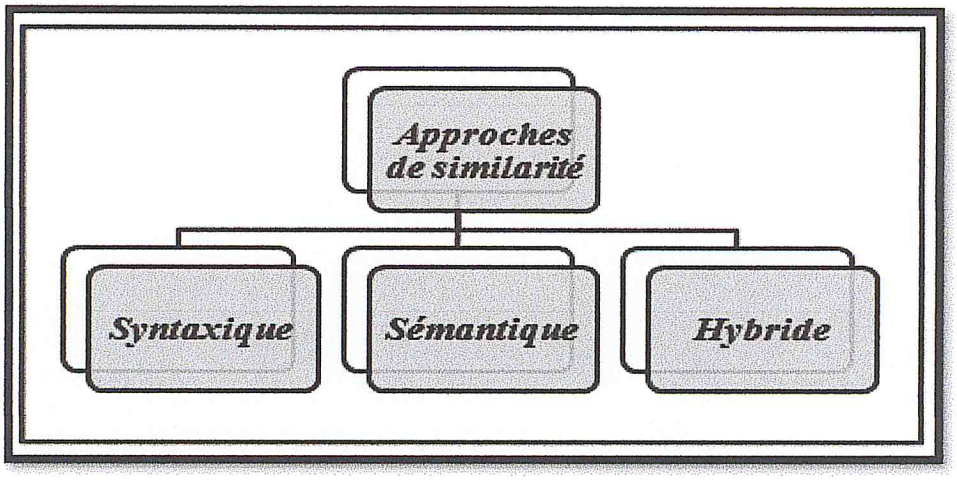



**Figure 2.4. Les approches de similarité.**

La notion de similarité a été très tôt perçue comme un concept clé en intelligence artificielle ainsi qu'elle intervient dans plusieurs de ses domaines : l'apprentissage automatique, la recherche d'information, la détection de fraudes (l'empreinte digitale), la détection du plagiat, la traduction automatique des corpus, le résumé de texte…

Il existe trois catégories principales des approches de similarité (voir Figure 2.4.) :

1. Similarité syntaxique (String-based similarity).
2. Similarité sémantique (Semantic similarity).
3. Similarité hybride (hybrid similarity)

### 2.4.1 Similarité syntaxique

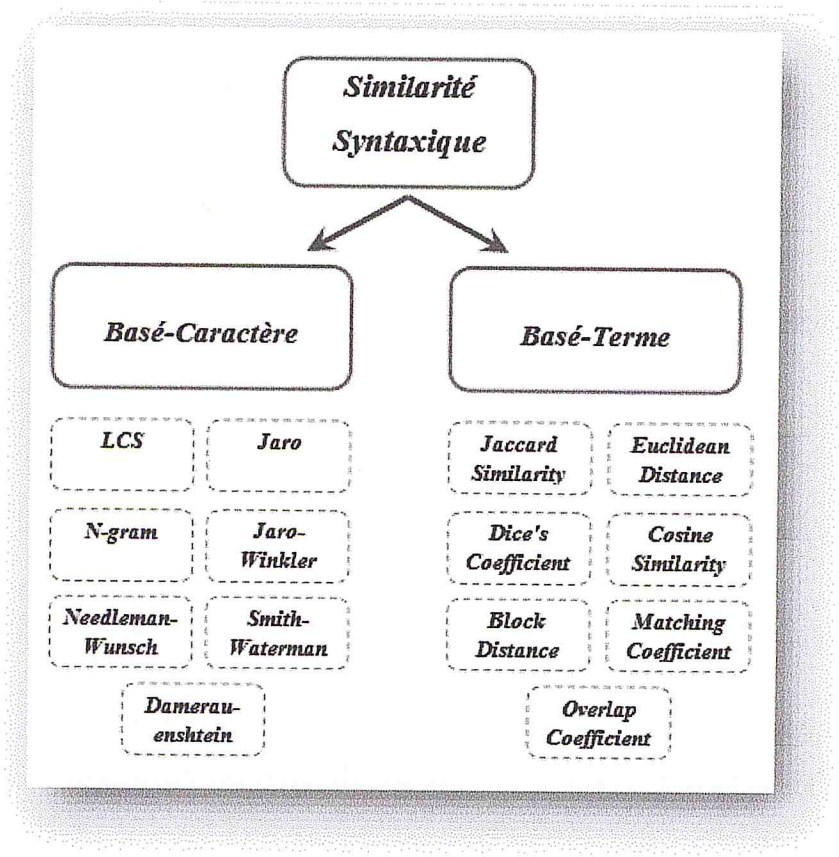



**Figure 2.5. Mesures de similarité syntaxique.**

En mathématiques et en informatique, une mesure permettant de comparer des documents textuels, consiste à comparer des chaines de caractères. C'est une métrique qui mesure la similarité ou la dis-similarité entre deux chaînes de caractères. Par exemple, les chaînes de caractères "voiture" et "voiturier" peuvent être considérées comme similaires, d'autre part, ''voiture'' et ''véhicule'' ne le sont pas. Une telle mesure sur les chaines de caractères fournit une valeur obtenue algorithmiquement [12]. « Figure 2.5. » montre les mesures de similarité syntaxique.

Parmi ses mesures, sept entre elles sont basées sur des caractères tandis que les autres sont des mesures de distance basée sur les termes. Parmi ces mesures nous présentons :

- **Indice de Jaccard:** L'indice de Jaccard (ou coefficient de Jaccard) est le rapport entre le cardinal (la taille) de l'intersection des ensembles considérés et le cardinal de l'union des ensembles [13]. Il permet d'évaluer la similarité entre les ensembles. Soit deux ensembles A et B, l'indice est :

$$J(A,B) = \frac{(A \cap B)}{(A \cup B)}$$

- **Distance euclidienne :** La distance euclidienne calcule la similarité entre deux documents d1 et d2 comme la distance entre leurs représentations vectorielles ramenées à un seul point [12].

$$Sim_{euclidienne} = \sqrt{\sum_{i=1}^{n}(d1_i - d2_i)}$$

Où n est le nombre total de termes représentés, i.e. la taille des vecteurs.

- **Cosinus :** La similarité cosinus est fréquemment utilisée [14] en tant que mesure de ressemblance entre deux documents d1 et d2. Il s'agit de calculer le cosinus de l'angle entre les représentations vectorielles des documents à comparer. La similarité obtenue $simcosinus(d1, d2) \in [0; 1]$ [12].

$$Sim_{cosinus} = \frac{\overrightarrow{d1}.\overrightarrow{d2}}{\|\overrightarrow{d1}\|.\|\overrightarrow{d2}\|}$$

- **Indice de Dice :** L'indice de Dice mesure la similarité entre deux documents d1 et d2 en se basant sur le nombre de termes communs à d1 et d2 [12].

$$sim_{dice} = \frac{2\,Nc}{N1 + N2}$$

Où Nc est le nombre de termes communs à d1 et d2, et N1 (resp. N2) est le nombre de termes de d1 (resp. d2).

- **Coefficient de corrélation de Pearson :** Le coefficient de corrélation de Pearson calcule la similarité entre deux documents d1 et d2 comme le cosinus de l'angle entre leurs représentations vectorielles centrées-réduites. La similarité obtenue $simpearson(d1;\ d2) \in [-1; 1]$ [12].

$$sim_{pearson}(d1, d2) = sim_{cosinus}\,(\,d1 - \overline{d1}\,, d2 - \overline{d2}\,)$$

D'où $\overline{d1}$ (resp.$\overline{d2}$) représente la moyenne de d1 (resp. d2).

### 2.4.2 Similarité sémantique

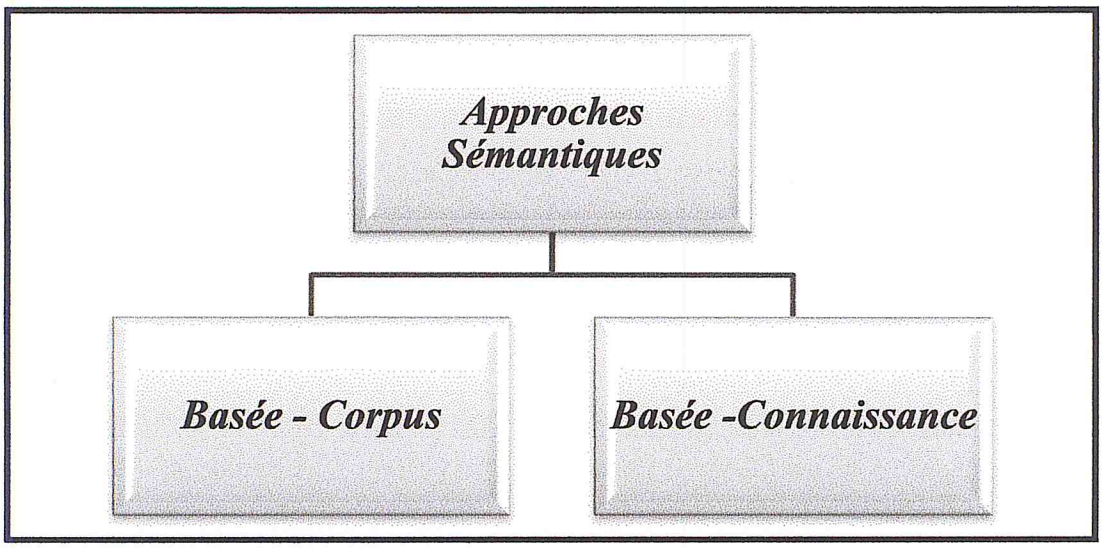



**Figure 2.6. Mesures de similarité sémantique.**

La similarité sémantique se base sur le sens/signification des mots. Deux concepts sont considérés comme sémantiquement similaires s'il y a une synonymie, hyponymie[1] (Figure 2.7.), antonymie, ou troponymie[2] entre eux [12]. Dans cette approche, des ressources NLP sont indispensables.

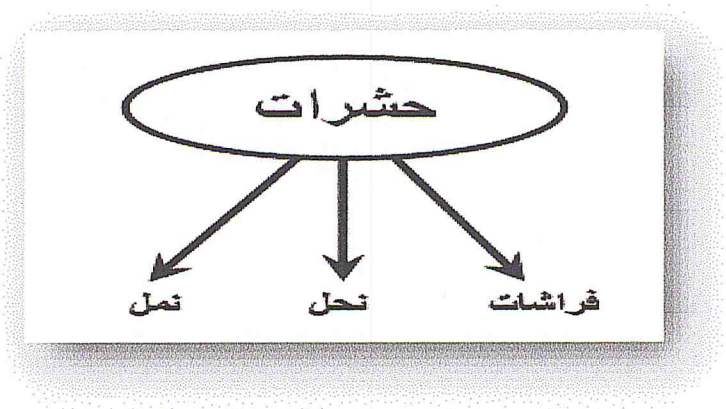



**Figure 2.7. Exemple d'hyponymie.**

Dans la littérature, Nous distinguons deux types de similarité sémantique ( voir Figure 2.6.) :

[1] Relation sémantique hiérarchique d'un lexème à un autre selon laquelle l'extension du premier terme est incluse dans l'extension du second. Le premier terme est dit hyponyme de l'autre. Haut-de-forme est un hyponyme de chapeau et chapeau est un hyponyme de coiffure.

[2] Relation sémantique entre deux verbes, l'un décrivant de manière plus précise l'action de l'autre. Le premier verbe est dit troponyme du second.

### 1) L'approche statistique (corpus-based)

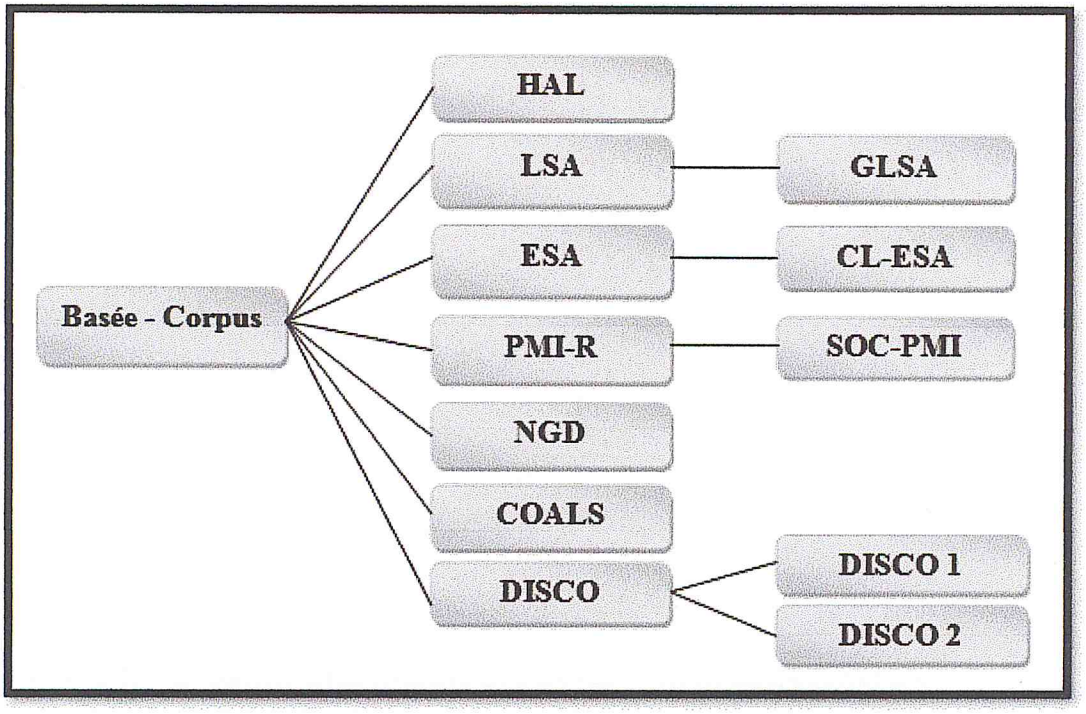



**Figure 2.8. Les mesures de l'approche Basée-Corpus.**

Les mesures basées sur des corpus ne nécessitent pas la compréhension du vocabulaire ou de la grammaire de la langue d'un texte. « Figure 2.8. » montre les mesures basée-corpus.

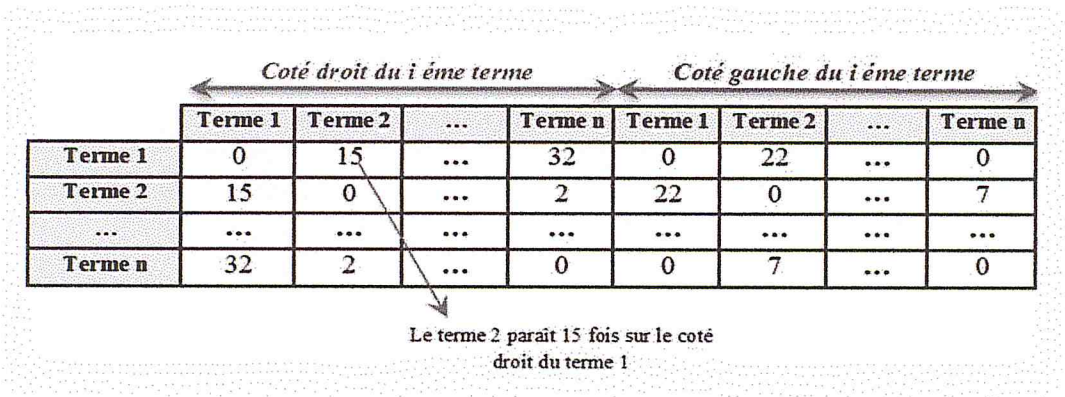
Toutefois, la similarité basée sur le corpus est une mesure de similarité sémantique statistique qui détermine la similarité entre les mots en fonction de l'information obtenue d'un corpus volumineux .Elle consiste a crée un espace sémantique à partir des cooccurrences de mots. Une matrice mot à mot est formée dont chaque élément de la matrice est la force d'association entre le mot représenté par la ligne et le mot représenté par la colonne. Les valeurs matricielles sont accumulées en pondérant la cooccurrence de manière inversement proportionnelle à la distance de focalisation du mot. Les mots voisins les plus proches sont considérés comme reflétant davantage la sémantique du mot cible et sont donc pondérés plus haut c'est-à-dire attribuer une importance [15]. Les mots sont représentés par des vecteurs sémantique d'où la notion d'espace vectoriel VSM (Vector Space Model).

Le modèle VSM vise à convertir l'ensemble des textes du corpus du format textuel en format numérique. Ceci se fait en construisant une matrice des fréquences des termes exhaustifs du corpus (vocabulaire du BOW).

Dans l'approche sémantique, chaque ligne de la matrice représente le vecteur de contexte du terme par rapport à son apparition autour du reste des termes. Cependant, pour calculer la similarité entre deux mots (terme), il faut récupérer leurs vecteurs de contexte et puis appliquer une des mesures syntaxique citées précédemment sur ces deux vecteurs. Les représentations les plus connue dans cette approche sont comme suit :

- **HAL (Hyperspace Analogue to Language) :**

La méthode construit un espace sémantique à partir des cooccurrences de mots obtenues d'un corpus volumineux. Une matrice mot à mot est alors formée dont chaque élément de la matrice est la force d'association entre le mot représenté par la ligne et le mot représenté par la colonne. Au fur et à mesure que le texte est analysé, un mot de mise au point est placé au début d'une fenêtre de dix mots qui enregistre quels mots voisins sont comptés comme co-occurrents. Les valeurs matricielles sont accumulées en pondérant la cooccurrence de manière inversement proportionnelle à la distance du mot de focalisation. Les mots voisins les plus proches sont considérés comme reflétant davantage la sémantique du mot cible et sont donc pondérés plus haut. HAL enregistre également les informations de classement des mots en traitant différemment la cooccurrence selon que le mot voisin est apparu avant ou après le mot de mise au point [16]. « Figure 2.9. » montre un exemple de représentation selon HAL.




| | Terme 1 | Terme 2 | ... | Terme n | Terme 1 | Terme 2 | ... | Terme n |
|---|---|---|---|---|---|---|---|---|
| Terme 1 | 0 | 15 | ... | 32 | 0 | 22 | ... | 0 |
| Terme 2 | 15 | 0 | ... | 2 | 22 | 0 | ... | 7 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Terme n | 32 | 2 | ... | 0 | 0 | 7 | ... | 0 |

**Figure 2.9. Exemple de représentation vectorielle selon HAL.**

- **LSA (Latent Semantic Analysis) :**

| | Document 1 | Document 2 | ... | Document n |
|---|---|---|---|---|
| Terme 1 | 15 | 0 | ... | 50 |
| Terme 2 | 3 | 74 | ... | 5 |
| ... | ... | ... | ... | ... |
| Terme n | 32 | 2 | ... | 12 |

**Figure 2.10. Exemple de représentation vectorielle selon LSA.**

Technique de similitude basée sur le Corpus. LSA suppose que les mots qui ont une signification proche se produiront dans des textes similaires. Une matrice contenant des nombres de mots par paragraphe (les lignes représentent des mots uniques et des colonnes représentent chaque paragraphe) est construite à partir d'un grand texte et une technique mathématique appelée décomposition de valeur singulière (SVD) est utilisée pour réduire le nombre de colonnes tout en préservant la similitude Structure entre les lignes [16]. La différence entre LSA et HAL est que l'espace sémantique construit par HAL se constitue des mots du corpus tandis que celles de LSA contiennent les documents (paragraphes) du corpus. De plus, l'idée d'invention d'HAL est venue à partir de la méthode LSA.

Les mots sont ensuite comparés en prenant le cosinus de l'angle entre les deux vecteurs formés par deux lignes quelconques (calcul de similarité). Un exemple de VSM est présenté dans « Figure 2.10. » selon LSA.

- **COALS (Correlated Occurrence Analogue to Lexical Semantic):**

| | Terme 1 | Terme 2 | ... | Terme n |
|---|---|---|---|---|
| Terme 1 | 0 | 15 | ... | 32 |
| Terme 2 | 15 | 0 | ... | 2 |
| ... | ... | ... | ... | ... |
| Terme n | 32 | 2 | ... | 0 |

**Figure 2.11. Exemple de représentation vectorielle selon COALS.**

La méthode COALS emploie une stratégie de normalisation qui factorise largement la fréquence lexicale. Le processus commence par la compilation d'une table de

cooccurrence de la même manière que dans HAL, sauf que la distinction gauche / droite est ignoré de sorte qu'il n'y a qu'une seule colonne pour chaque mot (en sommant les cooccurrences gauche et droite). Toute fois, COALS emploi aussi une fenêtre à quatre mots voisin comme les mots du même contexte que le mot en question. Il faut noter que plus le corpus est de petite taille, plus la fenêtre doit avoir une taille importante et vice versa. Un exemple de VSM est présenté dans « Figure 2.11. » selon COALS.

### 2) L'approche topologique (knowledge-based)

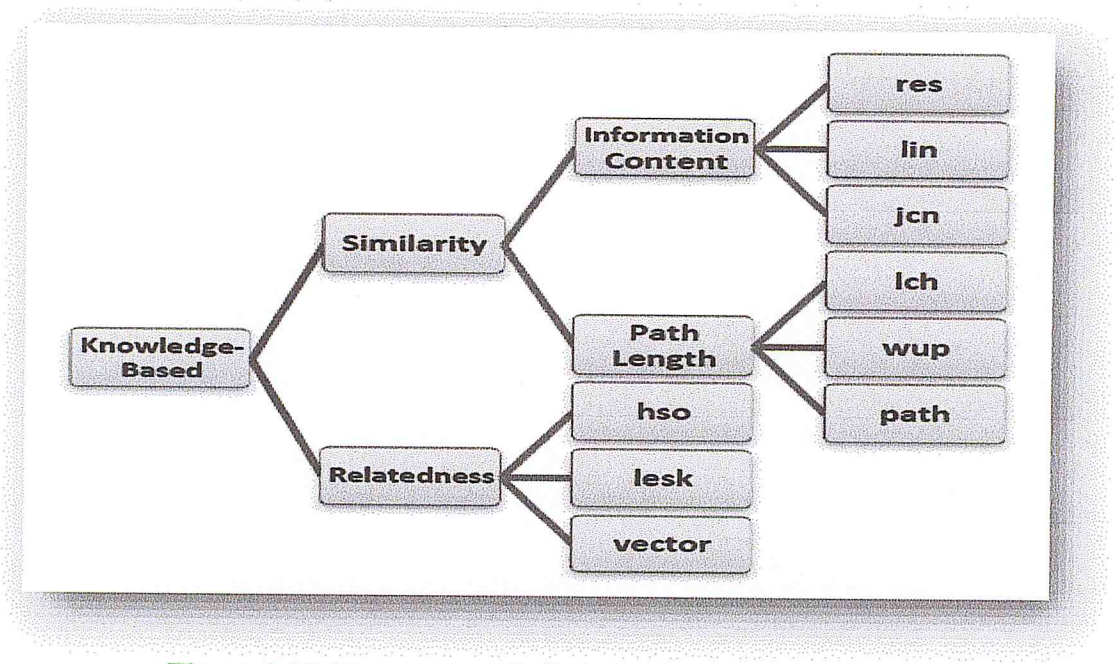



**Figure 2.12. Les mesures de l'approche Basé-Connaissance** [16].

La similarité basée sur la connaissance est l'une des mesures de similarité sémantique qui repose sur l'identification du degré de similitude entre les mots en utilisant des informations dérivées de réseaux sémantiques [16]. Le réseau le plus connue dans le domaine des linguistiques et de mesures de similarité est bien le WordNet. C'est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'années. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise [17]. Quelques mesures de la similarité basé-connaissance sont présentées dans la « Figure 2.12. ».

### 2.4.3 Similarité hybride

Les résultats des recherches faites par les chercheurs concernant la filière du Machine-Learning et plus précisément celle de la linguistique ont prouvé que l'application d'une combinaison de mesures de similarité donne de supérieures valeurs du facteur de corrélation et donc de meilleurs résultats par rapport à l'application d'une seule mesure de similarité.

Cette combinaison concerne l'arrangement de deux ou plusieurs mesures de similarité de la même approche ou bien un arrangement de mesures d'approches différentes.

## 2.5 Outils et ressources NLP

Le traitement automatique de la langue naturel est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle. Il vise à créer des outils de traitement de la langue naturelle pour diverses applications.

### 2.5.1 Techniques du traitement linguistique

Les ASAGS fréquemment utilisent les techniques TAL pour faciliter l'intégration du langage humain à la machine. C'est la base et la partie essentielle que les chercheurs utilisent pour accomplir leur projet. [2] présente les différentes techniques trouvées dans la littérature (Figure 5, p21) dont nous présentons quelques unes :

1. Suppression des nombres et chiffres du texte à traiter.
2. Suppression de la ponctuation.
3. Uniformiser la casse (rendre tout le texte en majuscule ou bien en minuscule).
4. La tokenisation signifie la segmentation du texte selon un caractère identifié. (généralement les espaces simples sont les plus fréquemment utilisés).
5. Part-Of-Speach tagging qui est une méthode qui prend en considération le coté grammatical et le contexte (Nom, verbe, adjectif…).
6. Le stemming qui consiste à convertir le mot a sa racine ou bien l'élimination des suffixes, préfixes…
7. La suppression des mots vides ou stop-words (et, ou, avec, pour…).
8. Correction d'orthographe afin d'éviter les ambiguïtés et les erreurs de frappes.

### 2.5.2 Paradigmes de création des ressources

Les recherches et travaux sur la langue arabe qui souffre du manque des ressources (outils NLP, corpus…) affrontent toujours des départs disant stagnants. De plus, l'arabe présente un défi à cause de sa complexité [18] ,qu'on cite brièvement ci-dessous :

- L'absence de la capitalisation rend l'identification des noms propres, acronyme, abréviation, début de phrase difficile.
- L'Arabe est fortement flexionnelle et dérivative, ce qui rend l'analyse morphologique une tâche très complexe.
- Les diacritiques (voyelles) sont, la plupart du temps, omis dans le texte arabe, ce qui rend difficile la déduction du sens du mot et, par conséquent, requiert des règles morphologiques complexes pour définir et analyser le texte.

Pour cela, les chercheurs illustrent trois paradigmes différents pour la création de ressources linguistiques :

- **Ressources linguistiques Crowdsourcing:**

  Le mot Crowdsourcing signifie l'obtention des données (information) en faisant appel aux services d'un certain nombre de personnes, rémunérées ou non, généralement via Internet. Il est utilisé pour produire une ressource rapidement, ne coute pas chère et relativement satisfaisante. Le processus réel est indépendant de la langue, à condition qu'un nombre suffisant d'utilisateurs de la langue soit disponible.

- **Création de ressources à l'aide de la traduction automatique:**

  Traduire un ensemble de données standard à partir de données existants en langue différente. Ceci est relativement facile et rapide mais potentiellement de moindre qualité (notamment une existence considérable de fautes et ambiguïté).

- **Création de ressources linguistiques en utilisant des experts humains:**

  Cela est établit en adoptant un effort manuel des experts qualifiés de la langue concernée pour créer une ressource plus chère en termes du temps et du cout mais effectivement de haute qualité.

### 2.5.3 Ressources NLP nécessaires

Un **corpus** est un ensemble de documents (textes, images, vidéos, etc.) regroupés dans une optique précise. On peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifiques…[19].

La branche de la linguistique qui se préoccupe plus spécifiquement des corpus s'appelle la linguistique de corpus. Elle est liée au développement des systèmes informatiques, en particulier à la constitution de bases de données textuelles.

Il existe deux sources (ensembles) de données :

a) **Corpus d'apprentissage :**

Il sert à retirer un modèle ou un classement à partir d'un nombre suffisant d'information. Concernant la langue arabe, une variété de corpus est disponible en ligne. Wajdi Zaghouani [20] a détaillé dans son article tout les corpus arabe existants pour des finalités d'avancement dans le domaine du traitement du langage naturel.

b) **Dataset (jeux de données):**

Ou encore *le corpus de test* sert à vérifier la qualité de l'apprentissage à partir du corpus d'apprentissage. Autrement dit, après avoir conçue une approche dans n'importe quelle branche du domaine d'informatique, plus précisément celui de la linguistique, on doit évaluer sa performance, précision et cohérence. C'est pour cette raison que les dataset sont conçus.

De plus, on peut rajouter une autre ressource qui est l'ensemble de validation. Elle sert à ajuster les paramètres des données collectées à l'aide du corpus d'apprentissage afin de tester les performances du programme/système créé.

### 2.5.4 Outils NLP

Le traitement automatique du langage naturel (NLP) est la capacité de la machine (programme) à comprendre le langage humain. Autrement, c'est la transformation du langage humain en langage machine. Généralement, le NLP nécessite des sources de données vocabulaire comme lexicon et dictionnaire électronique pour traiter un corpus relatif au

domaine bien spécifié (économie, sport, religion …). Sur le net, peu d'outils sont disponibles, accessibles ou téléchargeable gratuitement. Ci-dessous nous présentons le peu d'outils que nous avons trouvé :

1) **Stemmer**

Le stemmer est un outil qui implémente l'algorithme de stemming. Il peut être défini comme une procédure de réduction de tous les mots qui partagent la même racine à une forme commune notée « forme canonique » [21]. Par exemple: (استعمل- يستعمل -مستعمل ) ont une racine commune : عمل. Ceci est tellement utile, par exemple, pour réduire la taille du dictionnaire des mots dans un corpus ce qui réduit aussi l'espace nécessaire pour le stockage de ces données.

Il existe deux types de stemmer :

- **Le stem léger (Light-Stemming) :** C'est un processus moins complexe, où le stemming est arrêté sur la suppression des préfixes et des suffixes, sans tenter d'identifier la racine du mot.
- **Le stem lourd (Root-Stemming) :** Ce type consiste à supprimer les préfixes et les suffixes bien connus pour extraire la racine de base d'un mot et à identifier le motif en correspondance avec le mot restant.

Les stemmers disponibles en ligne et gratuitement téléchargeables seront prochainement entamés dans la section 3.2.2.

2) **Ghawas** [22]

Le but de cet outil est d'aider principalement les chercheurs de la branche de la linguistique arabe à appliquer des opérations statistiques sur leurs ressources. Il permet relativement : le calcul de la fréquence du mot ainsi que la fréquence relative, la suppression des stop-words (mots vides), le choix des normalisations désirées à partir d'une liste de normalisation prédéfinie… Les résultats de ses tâches peuvent être sauvegardés sous format CSV ou TXT.

3) **Stanford CoreNLP** ([23], [24])

C'est un ensemble d'outils d'analyse et technologiques en langage humain, il traite de différentes tâches complexes comme la reconnaissance des noms propre, baliser la structure

des phrases en termes de phrases et de dépendances syntaxiques (Part-Of speach ), effectuer un stem léger… De même il est conçu pour des finalités NLP et disponible pour plusieurs langues.

## 2.6 Revue sur quelques travaux connexes dans le domaine des ASAGS

De nombreux chercheurs s'intéressent au thème de l'évaluation automatique des réponses courtes et ainsi qu'à la similarité. Chaque approche à ses caractéristiques : langue, domaine, ressources nécessaires.

### 2.6.1 Les ASAG n'utilisant pas nécessairement la langue arabe

El Moatez Billah Nagoudi et al. [25] ont construit un système pour SemEval 2017 Task1: similarité sémantique textuelle (Track1). Le système nommé LIM-LIG propose un modèle d'intégration de mots consacrés à la mesure de la similarité sémantique dans les phrases arabe, L'idée principale est d'exploiter le mot représentations en tant que vecteurs dans un espace multidimensionnel pour capturer la sémantique et les propriétés syntaxiques des mots. Les pondérations IDF et Part-of-Speech tagging sont appliquées sur les phrases examinées à soutenir l'identification des mots qui sont hautement descriptifs dans chaque phrase. Le système LIM-LIG obtient une corrélation de Pearsons de 0.746, qui lui a permis de se classer en deuxième place parmi tous les participants à la tâche STS des couples monolingues arabes organisée dans le cadre de la campagne d'évaluation SemEval 2017.

L'approche LInSTSS proposée par Bojan Furlan et al. [26] présente une méthodologie pour la création d'un système capable à déterminer la similarité sémantique entre deux textes courts donnés. Cette approche est particulièrement appropriée pour une application dans des situations où aucune grande ressource linguistique électronique accessible au public ne peut être trouvée pour la langue désirée ce qui rend LInSTSS indépendante de la langue. LInSTSS est adaptée a la langue serbe afin d'évaluer et analyser les résultats obtenues en appliquant cette approche. Le système a marqué 76.6 d'exactitude et 82.93 de précision , qui sont jugées meilleures en les comparant avec d'autres méthodes STS.

L'outil SimAll présenté par Wael H. Gomaa et Aly A. Fahmi [27] combine les fonctionnalités de différente approche dans un seul outil. Différents types de modules de prétraitement, de méthodes de fusion et de niveaux de similarité sont inclus. Elle prend en charge les langues arabe et anglaise.

Mihalcea et al. [28] ont proposé une méthode pour mesurer la similarité sémantique de deux textes courts (phrases ou paragraphes) en utilisant des mesures basées sur le corpus et fondées sur la connaissance de la similarité des mots. Pour chaque mot du texte, cette méthode identifie la meilleure correspondance du texte opposé et l'inclut ensuite dans la mesure globale de la similarité sémantique. Cette approche donne un score de mesure F élevé, mais elle est exigeante en termes de calcul et nécessite l'utilisation d'un modèle de données de mots. Islam et Inkpen [29] ont amélioré la mesure de similarité en combinant un algorithme d'appariement de chaînes modifié avec une mesure basée sur le corpus de similarité sémantique. La similarité sémantique joue également un rôle important dans la reconnaissance de l'implication textuelle (RTE pour Recognizing Textual Entailment) et partage de nombreuses caractéristiques avec celle-ci.[26]

Mohamed A. Zahran et al. [30] ont comparé les différentes techniques pour construire des représentations spatiales vectorisées pour l'arabe et tester par la suite ces modèles via des évaluations intrinsèques et extrinsèques. L'évaluation intrinsèque évalue la qualité des modèles à l'aide d'un ensemble de données sémantiques et syntaxiques de référence, tandis que l'évaluation extrinsèque évalue la qualité des modèles en fonction de leur impact sur deux applications de traitement du langage naturel: recherche d'information et notation à réponse courte. Enfin, les auteurs cartographient l'espace vectoriel arabe avec l'homologue anglais en utilisant le réseau neuronal de régression d'erreur Cosinus et montrent qu'il surpasse les réseaux neuronaux de régression d'erreur quadratique moyenne standard dans cette tâche.

### 2.6.2 Les enjeux de la langue arabe dans le contexte de l'évaluation automatique [31]

La langue arabe est parlée et écrite par plus de 300 millions de personnes dans plus de vingt pays du monde entier. L'application des tâches de NLP (Natural Language Processing) en général et dans l'évaluation automatique des réponses courtes en particulier est très

difficile en langue arabe. La langue arabe a beaucoup de caractéristiques, qui sont considérées comme des enjeux (défis) à soulever pour l'évaluation automatique :

***Le premier enjeu*** est qu'il existe trois types de langue arabe, connus sous le nom de classique, moderne et familier. L'arabe classique, qui est utilisé dans le Coran, est plus complexe dans sa grammaire et son vocabulaire que l'arabe moderne. Il a un grand nombre de signes diacritiques qui facilitent la prononciation et la détection des mots dans leurs cas grammaticaux. Le deuxième type est l'arabe moderne, tous les signes diacritiques ont été omis pour faciliter et accélérer le processus de lecture et d'écriture. Ce type est considéré comme la langue officielle des pays arabes et est utilisé dans la langue de tous les jours, dans l'éducation et dans les médias. Habituellement, les recherches arabes basées sur l'arabe utilisent l'arabe moderne. En arabe parlé (dit aussi familier), qui est le troisième type, la grammaire et le vocabulaire sont moins sophistiqués par rapport à l'arabe moderne. Cependant, la plupart des gens l'utilisent dans leurs conversations quotidiennes et dans des lettres écrites de manière informelle en raison de sa simplicité. Les arabes font beaucoup d'erreurs dans la grammaire quand ils utilisent l'arabe moderne et ils mélangent entre l'arabe moderne et l'arabe familier.

***Le deuxième enjeu*** est la morphologie arabe. La langue arabe est complexe en raison de la variation morphologique. La forme des lettres change en fonction de leur position dans le mot. De plus, le mot peut être constitué de préfixes, de lemmes et de suffixes dans des combinaisons différentes, ce qui aboutit à une morphologie très compliquée.

***Le troisième enjeu*** est la capitalisation. La langue arabe ne supporte pas la capitalisation de noms propres tels que les noms de pays, les noms de personnes. Considérant que, dans les langues latines, ceux-ci commencent par une lettre majuscule. L'évaluation automatique arabe peut ne pas reconnaître ces entités nommées, ce qui augmente la difficulté de détecter de tels noms dans les réponses en arabe.

***Le dernier défi*** et que nous considérons le plus important est celui lié au manque de ressources linguistiques (outils NLP, Corpus, Datasets, …). Généralement, il y a une limitation sur le nombre de ressources linguistiques arabes, qui sont disponibles gratuitement à des fins de recherche. Plus récemment, un certain nombre de corpus arabes ont été développés; Cependant, peu d'amélioration globale de la situation globale a été observée [32].

Les défis précédents doivent être résolus lors de la construction d'un système pour l'évaluation automatique des réponses courtes. Nous les reprenons dans la discussion des

travaux de similarité utilisant la langue arabe dans la « section 2.6.3 ». Ces travaux ne concernent pas directement l'évaluation automatique des réponses courtes mais nous donnent des indications sur l'utilisation de mesures de similarité dans le contexte de la langue arabe et nous permettent de confirmer ou d'infirmer certains résultats ou constatations.

### 2.6.3 Les travaux sur la similarité de textes utilisant la langue arabe [31]

#### 1) Concernant les similarités syntaxiques :

Pour la langue arabe, de nombreux chercheurs ont utilisé l'algorithme de distance de Levenshtein dont [33] a utilisé pour développer l'outil de vérification orthographique pour les mots arabes. Cependant, Levenshtein ne donne pas de résultats précis lorsqu'il est appliqué sur la langue arabe selon les auteurs.

Le travail de [33] a utilisé une méthode N-gram pour convertir un mot en une suite de N-grammes et l'appliquer dans le contexte des systèmes de recherche textuelle arabes. L'étude indique que l'approche N-gram ne semble pas fournir une approche efficace dans le contexte arabe. [34] a étudié les différentes mesures de similarité syntaxiques dans la recherche d'information arabe et la mesure de similarité Cosinus (appelée souvent Cosine) est la meilleure mesure par rapport à d'autres mesures: coefficient de Dice, coefficient de Jaccard, coefficient de similarité d'inclusion, Mesure du coefficient de chevauchement, mesure de distance euclidienne et mesure de distance de Manhattan.

[35] ont conçu un thésaurus arabe automatique en utilisant la similarité terme-terme. Ils ont comparé la mesure de similarité de Jaccard avec d'autres mesures telles que Cosine et Dice. Les résultats indiquent que les mesures de similarité de Jaccard et de Dice ont la même performance, alors que le Cosinus est légèrement plus efficace que les mesures de Jaccard et de Dice.

#### 2) Similarités sémantiques

L'arabe est une langue mal adaptée pour les approches basées sur les corpus par rapport à l'anglais, car il y a un manque de données, ce qui affecte négativement la recherche sur les approches sémantiques basées sur les corpus en arabe. [36] ont passé en revue quatorze corpus arabes et les ont catégorisés par leur langue cible, objet, date du texte, lieu, domaine de

texte, représentativité, mode de texte, taille. Plusieurs de ces corpus ne fournissent aucune information concernant la période couverte par les textes. De plus, pour tous les corpus, les textes constitutifs ne sont pas classés en fonction de leurs dates ou de la période à laquelle ils appartiennent; il y a donc une limite à l'utilisabilité du corpus et une difficulté à comparer les langues utilisées à différentes périodes, et à observer comment la langue arabe a évolué.

Pour les approches basées sur la connaissance, WordNet est utilisé dans divers domaines tels que la recherche d'information et la similarité sémantique, En raison du succès de WordNet dans les applications en anglais, plusieurs projets sont actuellement menés pour développer WordNet pour d'autres langues. WordNet arabe (AWN) a été développé en utilisant la même méthodologie qu'EuroWordNet. Il se compose de 11 270 synsets et contient 23 496 expressions arabes (mots et multi-mots). Les principales limitations de l'AWN actuel sont un manque d'informations et de concepts par rapport à WordNet en anglais, et quelques relations sémantiques entre les synsets. De nombreux concepts arabes n'ont pas été inclus dans la base de données AWN. Cette limitation constitue un obstacle majeur à l'utilisation d'AWN en tant que source d'approches basées sur la connaissance. AWN pourrait être amélioré et étendu par plusieurs approches différentes, par exemple l'ajout de nouveaux synsets,...Par conséquent, nous pensons que l'approche de similarité sémantique utilisant AWN nécessite des recherches supplémentaires afin d'être plus fiable et plus mûre. Nous la jugeons insuffisante dans le contexte de l'évaluation automatique et c'est pour cette raison que nous l'avons écarté momentanément de nos travaux.

### 2.6.4 **Les travaux connexes à notre recherche utilisant la langue arabe** [31]

Les travaux que nous menons dans le cadre d'une approche hybride permettent de combiner plusieurs approches syntaxiques et sémantiques (particulièrement basés sur le corpus). Dans ce contexte, notre travail est connexe aux travaux menés par Gomaa & al. [37][38]. Les auteurs ont utilisé des mesures de similarité syntaxiques et des mesures basées sur le corpus pour développer leur système de notation à réponse courte. Ils ont testé les mesures sur le dataset (GOMAA dataset) qu'ils ont construit eux-mêmes. Leurs résultats ont montré que les meilleures valeurs de corrélation obtenues en utilisant des mesures syntaxiques ont été obtenues en utilisant respectivement les approches de distance de n-gramme et de distance de Manhattan. Dans la deuxième étape, ils ont mesuré la similarité en utilisant des

mesures de similarité basées sur le corpus [39]: DISCO1 (Calcule la similarité du premier ordre entre deux mots basés sur leurs ensembles de collocation) et DISCO2 (Calcule la similarité du second ordre entre deux mots basés sur leurs ensembles de distribution des mots similaires). Les résultats ont montré que DISCO1 atteint des valeurs de corrélation plus efficaces. Dans la troisième étape, la similarité a été évaluée en combinant des mesures basées sur la syntaxe et le corpus. La meilleure valeur de corrélation a été obtenue en mélangeant n-gramme avec les techniques de similarité DISCO1. En utilisant le SemEval Dataset nous allons avoir une indication sur la généralisation de nos approches dans des domaines connexes en les comparant aux résultats de la compétition 2017 fournis dans [40].

Le travail de [41] et [42] sont tous aussi intéressants en considérant leurs résultats par rapport à une approche basée sur le calcul vectoriel et les word Embedding. [41] ont évalué leur approche sur Gomaa dataset alors que [42] a obtenu le 2ème meilleur score du SemEval 2017 d'où l'intérêt que nous portons pour ces travaux utilisant les mêmes datasets que nous. En considérant ces travaux nous tentons d'améliorer les résultats obtenus dans les différentes étapes (syntaxiques et sémantiques) ensuite atteindre une meilleure hybridation en termes de maximisation du coefficient de Pearson et de minimisation de l'erreur quadratique.

## 2.7 Conclusion

La correction automatique des réponses courtes comme nous venons de voir est un domaine dont les racines sont ancrées et emmêlées avec pleins d'autres domaines tels que le traitement automatique de la langue, le calcul de similarité et l'évaluation automatique. Cet aspect multidisciplinaire offre une multitude d'axes de travail pour traiter le sujet. Dans le prochain chapitre nous exposerons la méthodologie que nous avons suivie dans notre travail et les techniques auxquelles nous avons eu recours.

# Chapitre 3 : Système d'évaluation automatique des réponses courtes



Après avoir présenté une vue globale sur l'existant dans le domaine des ASAGS, nous entamons notre approche étape par étape en détaillant chacune d'elles avec des exemples illustratifs pour mieux visualiser la démarche. Ceci est fait en se basant sur les connaissances acquises de l'étude précédente (chapitre 2).

## 3.1 Méthodologie

Dans ce qui suit, nous allons résumer les grandes lignes (visualisées au niveau de la Figure 3.1.) par les points suivants :

- ✓ **Etape 1** : Création de l'espace sémantique
  - Acquisition du corpus: Dans cette phase, il s'agit de collecter plusieurs corpus arabe, et effectuer une analyse afin de choisir les plus performants en termes de qualité, de volume et de domaine d'étude.
  - Prétraitement du corpus : Pour traiter le corpus, nous appliquons des techniques du traitement du langage naturel dont nous analysons plusieurs stemmer dédiés pour le traitement de la langue arabe et par la suite choisir le plus convenable.
  - Traitement du corpus : Dans cette étape, nous générons notre espace sémantique selon la démarche détaillée ultérieurement.
  - Post traitement du corpus : Ici nous générons des valeurs représentant la spécificité[1] ou l'importance des mots du corpus.

- ✓ **Etape 2** : Modèle de calcul de similarité

  Nous présentons les deux modèles adaptés pour le calcul de similarité sémantique qui sont : le modèle somme-vecteurs, le modèle calcul-matriciel et l'hybridation des deux modèles.

- ✓ **Etape 3** : Passage au score

  Après avoir calculé la similarité, nous passons au score/note automatique en utilisant le classifieur non supervisé k-means.

[1] La spécificité d'un terme est la quantité d'informations spécifiques au domaine contenues dans le terme. Certains termes contiennent une quantité relativement importante d'informations de domaine et d'autres une quantité relativement faible d'informations de domaine [60].

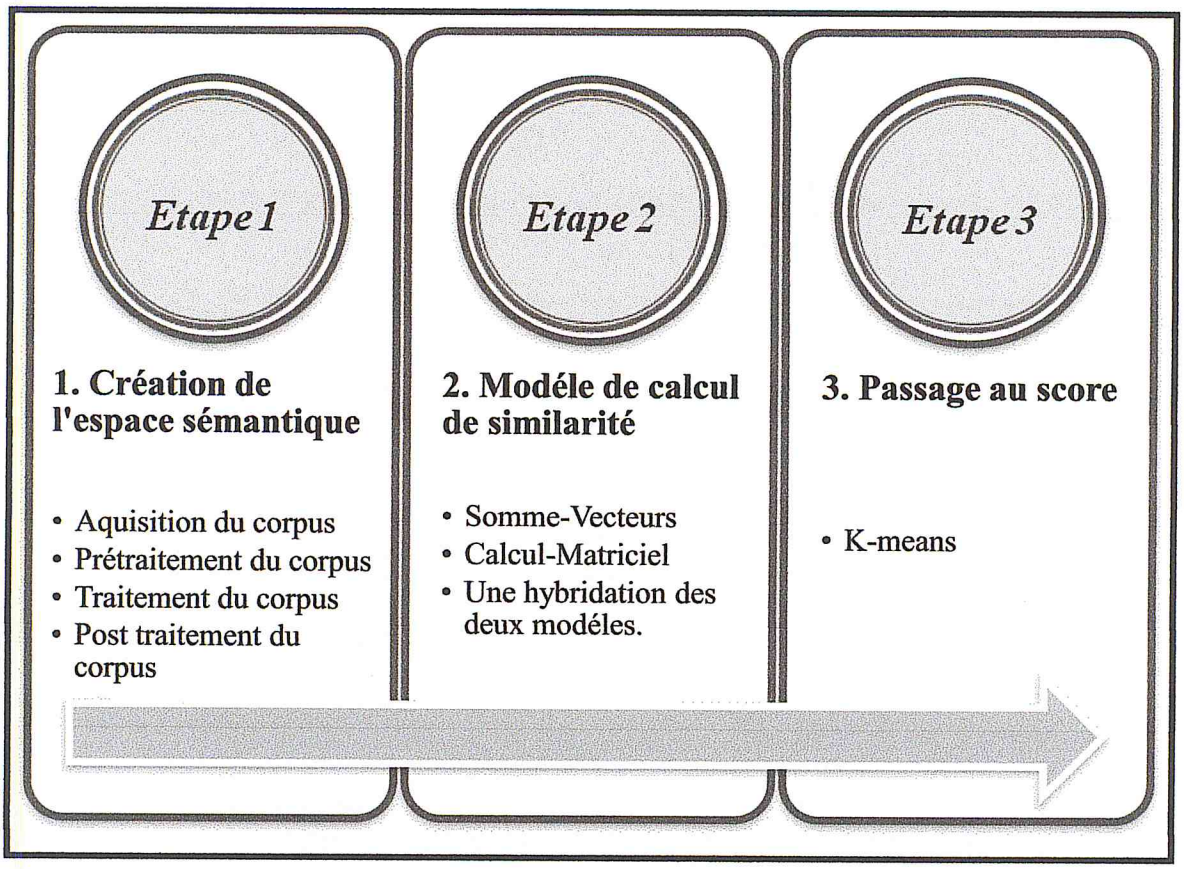



**Figure 3.1. Schéma des étapes principales**

Afin d'arriver à de meilleurs résultats, nous avons intégré des résultats de deux autres thèmes qui se déroulent sur la même thématique :

- L'évaluation automatique des réponses courtes pour le E-learning[1] en utilisant les Word Embeddings: Application à la langue arabe.
- Mesures de similarité syntaxique pour un système d'évaluation automatique des réponses courtes : Application à la langue arabe.

[1] Le e-learning -ou apprentissage en ligne- est une méthode d'apprentissage qui repose sur la mise en disposition des ressources pédagogiques à travers un support électronique.

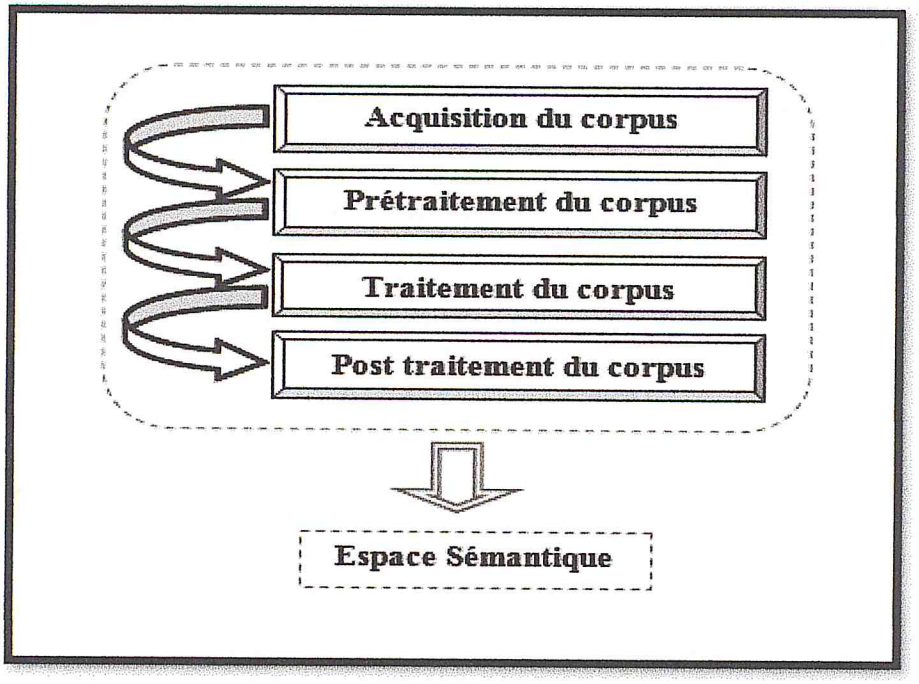



**Figure 3.2. Phases de création de l'espace sémantique**

## 3.2 Construction de l'espace sémantique

Cette partie présente en détail la première étape de la « Figure 3.1. ». Dans la « Figure 3.2. » nous présentons les différentes phases de création de l'espace sémantique Tout d'abord, l'acquisition et l'analyse du corpus afin de choisir celui qui convient. Ensuite, la phase du prétraitement qui consiste à préparer le corpus pour tout usage. La phase du traitement du corpus comprend l'application de l'une des approches de similarité dans le but de construction de l'espace sémantique. En fin, le post traitement du corpus qui prend en considération l'importance des mots dans le corpus.

### 3.2.1 Acquisition du corpus

Sur le net, il existe une variété de corpus (gratuit ou commercialisé) dont chacun d'eux est dédié pour une certaine finalité. Toutefois, nous avons eu l'occasion d'analyser et de traiter cinq corpus disponibles gratuitement en ligne consacrés à la recherche dont la description ci-dessous :

1) **CNN Arabic Corpus (2010)[1]**

Ce corpus a été collecté du site web CNN Arabic ***cnnarabic.com,*** il comprend 5070 documents texte. Chaque document texte appartient à l'une des six catégories (Business 836, Divertissement 474, Nouvelle de Moyen-Orient 1462, Science et technologie 526, Sports 762, Nouvelle du monde 1010). Le corpus contient 2.241.348 (2.2M) mots et 144.460 mots-clés de district après le retrait des mots vides [43].

2) **BBC Arabic Corpus (2010)[1]**

BBC Arabic corpus a été collecté du site web BBC Arabic bbcarabic.com, le corpus comprend 4 763 documents texte. Chaque document texte appartient à l'une des sept catégories (Nouvelles du Moyen-Orient 2356, Nouvelles du monde 1489, Business et Economie 296, Sports 219, Presse internationale 49, Science et technologie 232, Art et culture 122). Le corpus contient 1.860.786 mots (1.8M) et 106.733 mots-clés de district après le retrait des mots vides [43].

3) **Osac Arabic Corpus (2010)[1]**

Le corpus arabe OSAC est collecté à partir de plusieurs sites Web (aljazeera.net, fatafeat.com, alkawn.net…). Il comprend 22 429 documents texte. Chaque document texte appartient à l'une des 10 catégories suivantes: économie, histoire, divertissements, éducation et famille, religion et fatwas, sports, santé, astronomie, droit, histoires, recettes de cuisine. Le corpus contient environ 18.183.511 (18M) mots et 449.600 mots-clés de district après le retrait des mots vides[43].

4) **Al Khaleej Corpus (2004)[2][44]**

Ce corpus a été réalisé pour des expériences sur thème de l'identification des sujets pour la langue arabe. Il a été extrait de milliers d'articles téléchargés à partir d'un journal en ligne. Le corpus se constitue de quatre catégories (Nouvelles Internationales 953, Nouvelles Locales 2398, Sports 1430, Economie 909) et contient plus de 5000 articles qui correspondent à près de 3 millions de mots. La ponctuation a été intentionnellement omise [45].

[1] https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/

[2] https://sourceforge.net/projects/arabiccorpus/files/khaleej-2004corpus/

**5) Al watan Corpus (2004)[1][46]**

Le corpus contient environ 20000 articles sur les six catégories suivantes: Culture 2782, religion 3860, économie 3468, nouvelles locales 3596, nouvelles internationales 2935 et sports 4550. Dans ce corpus, la ponctuation a été omise intentionnellement afin de la rendre utile pour la modélisation langagière [45].

CNN Arabic Corpus et BBC Arabic corpus sont convertis au format « txt » tandis que les deux autres corpus sont au format « html ».

### 3.2.2 Prétraitement du corpus

Pour cette tâche, nous travaillons sur tous les corpus décrits précédemment. Elle se compose de deux autres sous-tâches:

i. **Techniques TAL adoptées:**

Ceci est fait par les étapes suivantes :

- Convertir Al khaleej corpus et Al watan corpus du format « html » au format « txt ».
- Réduire les successions des espaces en un espace simple.
- Supprimer les lettres non arabe (A, B…Z, a, b,…z).
- Suppression des diacritiques.
- Suppression des numeros.
- Normaliser les mots en remplaçant :
  - « أ , إ , ؤ , آ par ا ».
  - « ىء , يء par ئ ».
  - « ة par ه ».
  - « ى par ي ».
- Application d'une étape de stemming.
- Tokeniser chaque document et alors chaque corpus en considérant un espace simple comme séparateur.

[1] https://sourceforge.net/projects/arabiccorpus/files/watan-2004corpus/

### ii. L'approche par Stemming adoptée:

Puisque notre travail est inclus dans un autre travail plus général [31], nous considérons un mécanisme basé sur le stemming afin d'en analyser l'impact sur l'évaluation automatique des questions à réponses courtes en langue arabe.

En effet, il est très difficile de mettre en œuvre les mécanismes d'évaluation automatique pour la langue arabe, en raison de sa nature complexe, étant très flexionnelle et ambigüe en l'absence de signes diacritiques. Il n'y a eu que peu de tentatives de recherche sur ce sujet, et jusqu'à présent, aucun d'entre eux n'a été en mesure de fournir un système d'évaluation automatique entièrement fonctionnel. Les techniques de stemming ont été exploitées en combinaison avec les mesures de similarités développées. Un algorithme de stemming peut être défini comme la procédure de réduction de tous les mots qui partagent la même racine à une forme commune [21].

Pour toutes les approches de similarités développées nous avons considérer les deux cas suivants :

- Une technique de stemming lourde ( Heavy Stemming) est appliquée aux réponses à comparer. Le streaming lourd, également appelé « Root-Stemming » (Stemming à la racine), consiste à supprimer les préfixes et les suffixes bien connus pour extraire la racine réelle d'un mot et à identifier le motif en correspondance avec le mot restant.
- Une technique de stemming légère (Light Stemming) est appliquée aux réponses à comparer. Le streaming léger est un processus moins complexe, où le stemming est arrêté sur la suppression des préfixes et des suffixes, sans tenter d'identifier la racine réelle du mot.

Le stemming consiste en général à réaliser les actions suivantes pour chaque couple de réponses à comparer :

- Suppression des nombres des deux réponses.
- Suppression des signes diacritiques des deux réponses.
- Suppression de la ponctuation et lettres spéciaux.
- Suppression des mots vides (stopwords). Une liste de mots vides est disponible dans la base de données (هما هي, هو, اذا, ان, و, في, ...).

- Enlever le (ال : AL), et ses Dérivés, (وبال, لبال, فال, ال, لل, كال بال, وال, تال, فبال, ...)
- Suppression du préfixe si la longueur du mot est supérieure à 3.
- Suppression du suffixe, si la longueur du mot est supérieure à 3. Une liste de préfixes, suffixes est disponible et utilisée par le programme du stemmer. Cette liste est différente selon que le stemming est lourd ou léger.

Pour la mise en œuvre de l'approche du stemming dans notre travail, nous avons recherché parmi plusieurs stemmers existants dans la langue arabe (disponibles en ligne ou téléchargeables). Nous avons testé les différents stemmers sur beaucoup de couples de réponses. Ci-dessous une liste des stemmers trouvés et testés :

- Khoja stemmer[1] [47].
- Light 10 stemmer[2] .
- ISRI stemmer[3] .
- Tashaphyne stemmer[4] [48].
- Motaz stemmer[5].
- Assam stemmer[6].

Le site web [49] rend disponibles les cinq premiers stemmers (présentés ci-dessus). Il donne la possibilité d'insérer les données (input) par clavier ou de format d'un fichier « txt ». De plus, il donne le choix de sauvegarder sous format « excel », « csv » ou « xml ».

Nous avons testés les six stemmer sur les 61 réponses modèles du dataset de Gomaa [37][38]. « Tableau 3.1 » représente les six stemmers testés sur la $50^{éme}$ réponse modèle du dataset de Gomaa comme un échantillon.

Il est à noté que comparé à l'anglais, les quelques stemmers qui existent ne présentent pas de documentation disponible et ne présentent pas une évaluation de la précision des résultats obtenus. L'avis d'un expert en langue arabe nous a été difficile de procurer et par

[1] http://zeus.cs.pacificu.edu/shereen/research.htm#stemming

[2] http://arabic.emi.ac.ma:8080/SafarWeb_V2/faces/safar/morphology/stemmer.xhtml

[3] http://www.nltk.org/_modules/nltk/stem/isri.html

[4] https://pypi.org/project/Tashaphyne/

[5] https://github.com/motazsaad/arabic-light-stemmer

[6] http://www.arabicstemmer.com/

conséquent nous nous sommes basés sur l'appréciation de l'équipe pour évaluer les résultats obtenus et choisir d'utiliser les deux stemmers suivants dans la suite du travail :

- Khodja Stemmer [47] pour un lourd stemming.
- Tashaphyne stemmer [48] pour un léger stemming.

**Tableau 3.1. Echantillon des tests sur les six stemmers**

| | Mot | Khoja | Light 10 | ISRI | Tashaphyne | Motaz | Assem |
|---|---|---|---|---|---|---|---|
| لتوافر الإضاءة الكافية فوق هذا العمق حيث يستطيع الضوء النفاذ حتى عمق 200 م مما يسمح للنباتات بالقيام بعملية البناء الضوئي. | لتوافر | وفر | وافر | وفر | وافر | لتوافر | توافر |
| | الإضاءة | اضاءة | اضا | ضءة | إضاء | اضاء | اضاء |
| | الكافية | كفأ | كاف | كفي | كاف | كاف | كاف |
| | فوق | فوق | فوق | فوق | ق | فوق | فوق |
| | هذا | هذا | هذ | هذا | هذ | هذا | هذا |
| | العمق | عمق | عمق | عمق | عمق | عمق | عمق |
| | حيث | حيث | حيث | حيث | حيث | حيث | حيث |
| | يستطيع | طوع | يستطيع | تطع | طيع | يستطيع | استطيع |
| | الضوء | ضوأ | ض | لضء | ضوء | ضوء | ضوء |
| | النفاذ | نفذ | نفاذ | نفذ | نفاذ | نفاذ | نفاذ |
| | حتى | حتى | ح | حتى | حتى | حت | حتي |
| | عمق | عمق | عمق | عمق | عمق | عمق | |
| | 200 | | 200 | 200 | 200 | 200 | 200 |
| | م | م | م | م | م | م | م |
| | مما | مما | مم | مما | | مما | مما |
| | يسمح | سمح | يسمح | سمح | سمح | يسمح | يسمح |
| | للنباتات | نبت | نبات | نبت | نباتا | نبات | نبات |
| | بالقيام | قوم | قيام | قيم | قيام | قيام | قيام |
| | بعملية | عمل | بعمل | عمل | عمل | بعمل | عمل |
| | البناء | بني | بنا | بنء | بناء | بناء | بناء |
| | الضوئى. | ضوأ | ضوا | ضوئى. | ضوئى. | ضوئي. | ضوبي |

Cependant, la liste des mots vides prise en charge par khoja stemmer est tellement réduite. Par conséquent, nous avons réalisé une nouvelle liste et par la suite, remplacé celle originale de khoja stemmer par la nouvelle. La liste réalisée est une combinaison de trois

autres listes de mots vides : la liste originale de khoja, la liste de Zerrouki[1] et la liste de Mohamed Taher Alrefaie[2].

L'approche par stemming a été adoptée aussi bien pour les réponses que pour les corpus.

Un stemmer doit avoir une valeur de précision ou erreur qui signifie l'estimation de son exactitude. La plupart des stemmers n'ont pas de documentation disponible en ligne ce qui n'est pas pratique et rend la tâche plus difficile (choisir le stemmer approprié).

### 3.2.3 Traitement du corpus

Dans notre démarche, l'approche de similarité statistique « Corpus-Based » est adoptée à cause du manque de ressources arabe comme les dictionnaires et les lexicons dont l'approche topologique « knowledge-based » impose. Nous allons adopter le modèle BOW qui ne prend pas en considération l'ordre des mots. Notre approche se base sur le concept disant que les mots qui sont sémantiquement liés se trouvent dans le même contexte. Pour cela cette approche repose sur la notion du voisinage (mots voisins). Cependant la présence d'un corpus qui sera manipulé à l'aide du modèle BOW est indispensable. En outre, une fenêtre avec une taille prédéfinie est importante pour concrétiser la notion du voisinage.

Nous avons utilisé le corpus « BBC+CNN », le corpus « CNN » et le corpus « Al khaleej » ainsi qu'une fenêtre de taille 4. La « Figure 3.4. » montre le processus effectué afin de générer nos trois espaces sémantiques à l'aide des trois corpus. Pour mieux voir les choses, nous détaillons ce processus ci-dessous :

a) Les corpus sont acquis et prétraités selon l'étape du « Prétraitement du corpus ».
b) **Construction de la matrice des cooccurrences :** Ou matrice de fréquences/poids. Cette étape signifie la transformation des données textuelles au sein du corpus en données numériques sous forme d'une matrice. Chaque case représente la somme des poids de l'apparence du $i^{éme}$ terme (mot) avec le $j^{éme}$ terme. Cette somme est calculée par rapport à l'emplacement des termes du corpus autour du $i^{éme}$ terme (effectuer le même calcul pour chaque occurrence du $i^{éme}$ terme). Ici nous entamons

[1] https://sourceforge.net/projects/arabicstopwords/files/latest/download

[2] https://github.com/mohataher/arabic-stop-words/blob/master/list.txt

la notion de voisinage sous forme de ce que l'on appelle une **taille de fenêtre**. Concrètement, si deux mots d'un texte sont en **voisinage** de taille inférieure à celle de la fenêtre, ils sont comptés comme **co-occurrents**. Dans notre approche nous fixons la taille de la fenêtre à 4 c'est-à-dire considérer que les quatre voisins adjacents au $i^{éme}$ terme (ou qui se trouve autours ce terme) des deux cotés (gauche et droite) et ignorer le reste des mots. Pour illustrer ce fonctionnement, nous prenons le texte suivant comme une partie du corpus :

« شركات امن المعلومات رصدت ارتفاعا في عدد المكالمات الوهمية الصادرة عن اجهزة الهاتف النقال الذكية »

Nous considérons le terme « ارتفاعا » comme le $i^{éme}$ terme. Le partitionnement des poids est présenté dans la « Figure 3.3. ». Les voisins du terme « ارتفاعا » auront un poids de 4, les voisins des voisins du terme auront un poids de 3 et ainsi de suite.

| 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|
| الوهمية | المكالمات | عدد | في | ارتفاعا | رصدت | المعلومات | امن | شركات |

**Figure 3.3. Exemple du fonctionnement de la fenêtre de taille 4**

c) **Construction de la matrice des corrélations :**

a. Après avoir construit la matrice des cooccurrences, nous appliquons une stratégie de normalisation. Nous allons alors construire une nouvelle matrice en appliquant la formule ci-dessous prise de l'algorithme COALS pour chaque case/élément de la matrice des cooccurrences.

$$\begin{cases} w'_{a,b} = \dfrac{Tw_{a,b} - \sum_j w_{a,j} \cdot \sum_i w_{i,b}}{\left(\sum_j w_{a,j} \cdot (T - \sum_j w_{a,j}) \cdot \sum_i w_{i,b} \cdot (T - \sum_i w_{i,b})\right)^{1/2}} \\ T = \sum_i \sum_j w_{i,j} \end{cases}$$

D'où :

- **a** et **b** sont les deux termes de la matrice de cooccurrences (ligne et colonne).
- $w_{a,b}$ est l'élément de la matrice de cooccurrences du terme a et b.

- $i$ est l'indice des lignes tandis que $\mathbf{j}$ est celui des colonnes.
- $\sum_j w_{a,j}$ est la somme des colonnes de la ligne du terme a.
- $\sum_i w_{i,b}$ est la somme des lignes de la colonne du terme b.
- $T = \sum_i \sum_j w_{i,j}$ est la somme de tous les éléments de la matrice de cooccurrences.

Pour un large corpus, les valeurs de corrélations sont petites, alors, il est rare que la valeur de corrélation dépasse le **0.01.** De plus, la majorité des corrélations sont négatives (81.8%) [15].

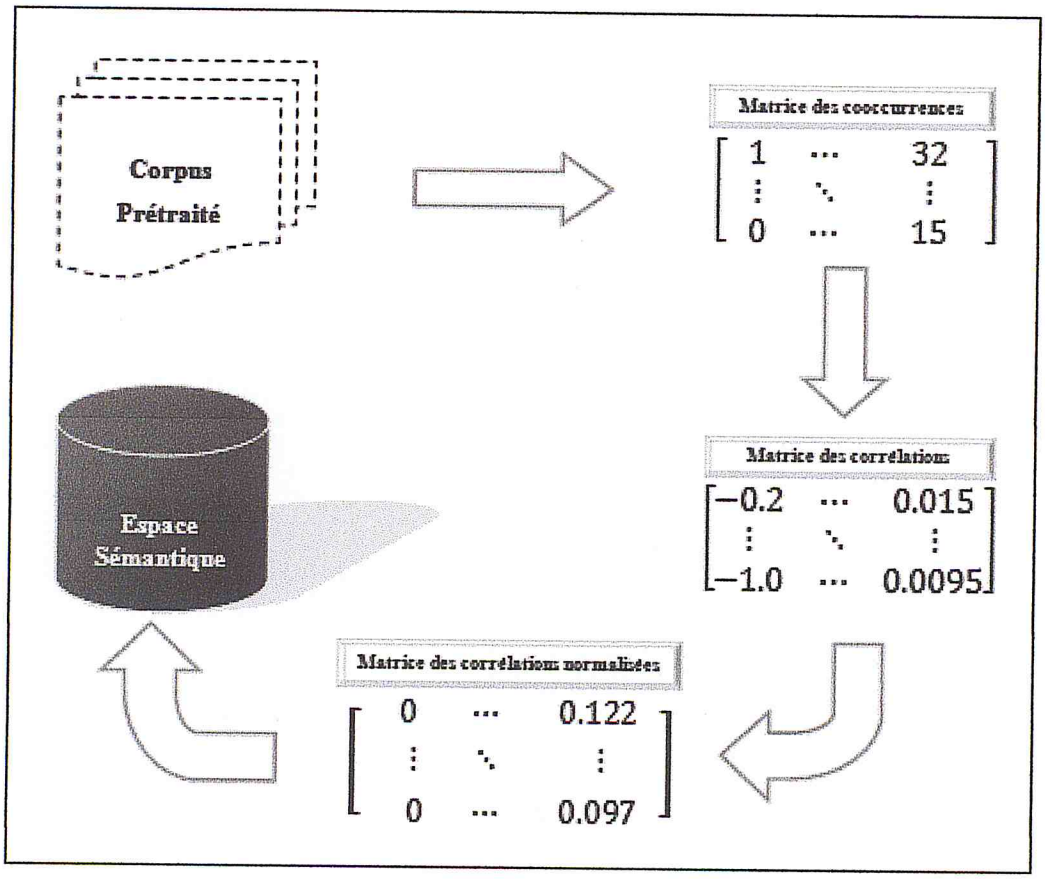



**Figure 3.4. Exemple illustrant le traitement d'un corpus.**

d) **Construction de la matrice de corrélation normalisée :**

Les corrélations négatives transportent très peu d'information, alors, nous effectuons encore une autre normalisation. Les valeurs négatives sont normalisées à 0

tandis que les valeurs positives prennent leurs racine carré afin d'amplifier l'importance des nombreuses petites valeurs par rapport aux grandes valeurs [15].

Cette étape représente la dernière étape de création de l'espace sémantique ou proprement dit : les vecteurs du contexte de chaque mot du corpus en entrée. Il est à noter que l'espace sémantique est généré qu'une seule fois. Il est donc stocké pour une utilisation ultérieure.

➢ **Exemple illustratif:**

Nous considérons le texte dans la « Figure 3.5. » comme un corpus. Dans cet exemple, nous montrons la procédure sans stem et avec stem.

شركات امن المعلومات رصدت ارتفاعا في عدد المكالمات الوهمية الصادرة عن اجهزة الهاتف النقال الذكية

**Figure 3.5. Exemple de corpus.**

✓ **Cas sans stem**

1) Construction de la matrice des cooccurrences :

| | شركات | امن | المعلومات | رصدت | ارتفاعا | في | عدد | المكالمات | الوهمية | الصادرة | عن | اجهزة | الهاتف | النقال | الذكية |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| شركات | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| امن | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| المعلومات | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| رصدت | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ارتفاعا | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| في | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| عدد | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 |
| المكالمات | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 |
| الوهمية | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 |
| الصادرة | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 |
| عن | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 |
| اجهزة | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 |
| الهاتف | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 |
| النقال | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 |
| الذكية | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 |

**Figure 3.6. Matrice des cooccurrences dans le cas sans stem**

2) Matrice des corrélations :

| | شركات | امن | المعلومات | رصدت | ارتفاعا | في | عدد | المكالمات | الوهمية | الصادرة | عن | اجهزة | الهاتف | النقال | الذكية |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| شركات | -0.04 | 0.307 | 0.19 | 0.098 | 0.017 | -0.058 | -0.058 | -0.058 | -0.058 | -0.058 | -0.058 | -0.056 | -0.053 | -0.048 | -0.04 |
| امن | 0.307 | -0.057 | 0.213 | 0.129 | 0.059 | -0.005 | -0.069 | -0.069 | -0.069 | -0.069 | -0.069 | -0.067 | -0.063 | -0.057 | -0.048 |
| المعلومات | 0.19 | 0.213 | -0.07 | 0.165 | 0.099 | 0.04 | -0.018 | -0.076 | -0.076 | -0.076 | -0.076 | -0.074 | -0.07 | -0.063 | -0.053 |
| رصدت | 0.098 | 0.129 | 0.165 | -0.079 | 0.141 | 0.085 | 0.03 | -0.026 | -0.081 | -0.081 | -0.081 | -0.079 | -0.074 | -0.067 | -0.056 |
| ارتفاعا | 0.017 | 0.059 | 0.099 | 0.141 | -0.083 | 0.133 | 0.079 | 0.025 | -0.029 | -0.083 | -0.083 | -0.081 | -0.076 | -0.069 | -0.058 |
| في | -0.058 | -0.005 | 0.04 | 0.085 | 0.133 | -0.083 | 0.133 | 0.079 | 0.025 | -0.029 | -0.083 | -0.081 | -0.076 | -0.069 | -0.058 |
| عدد | -0.058 | -0.069 | -0.018 | 0.03 | 0.079 | 0.133 | -0.083 | 0.133 | 0.079 | 0.025 | -0.029 | -0.081 | -0.076 | -0.069 | -0.058 |
| المكالمات | -0.058 | -0.069 | -0.076 | -0.026 | 0.025 | 0.079 | 0.133 | -0.083 | 0.133 | 0.079 | 0.025 | -0.026 | -0.076 | -0.069 | -0.058 |
| الوهمية | -0.058 | -0.069 | -0.076 | -0.081 | -0.029 | 0.025 | 0.079 | 0.133 | -0.083 | 0.133 | 0.079 | 0.03 | -0.018 | -0.069 | -0.058 |
| الصادرة | -0.058 | -0.069 | -0.076 | -0.081 | -0.083 | -0.029 | 0.025 | 0.079 | 0.133 | -0.083 | 0.133 | 0.085 | 0.04 | -0.005 | -0.058 |
| عن | -0.058 | -0.069 | -0.076 | -0.081 | -0.083 | -0.083 | -0.029 | 0.025 | 0.079 | 0.133 | -0.083 | 0.141 | 0.099 | 0.059 | 0.017 |
| اجهزة | -0.056 | -0.067 | -0.074 | -0.079 | -0.081 | -0.081 | -0.081 | -0.026 | 0.03 | 0.085 | 0.141 | -0.079 | 0.165 | 0.129 | 0.098 |
| الهاتف | -0.053 | -0.063 | -0.07 | -0.074 | -0.076 | -0.076 | -0.076 | -0.076 | -0.018 | 0.04 | 0.099 | 0.165 | -0.07 | 0.213 | 0.19 |
| النقال | -0.048 | -0.057 | -0.063 | -0.067 | -0.069 | -0.069 | -0.069 | -0.069 | -0.069 | -0.005 | 0.059 | 0.129 | 0.213 | -0.057 | 0.307 |
| الذكية | -0.04 | -0.048 | -0.053 | -0.056 | -0.058 | -0.058 | -0.058 | -0.058 | -0.058 | -0.058 | 0.017 | 0.098 | 0.19 | 0.307 | -0.04 |

**Figure 3.7. Matrice des corrélations dans le cas sans stem.**

3) Construction de la matrice des corrélations normalisées :

| | شركات | امن | المعلومات | رصدت | ارتفاعا | في | عدد | المكالمات | الوهمية | الصادرة | عن | اجهزة | الهاتف | النقال | الذكية |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| شركات | 0 | 0.554 | 0.436 | 0.313 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| امن | 0.554 | 0 | 0.462 | 0.359 | 0.243 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| المعلومات | 0.436 | 0.462 | 0 | 0.406 | 0.315 | 0.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| رصدت | 0.313 | 0.359 | 0.406 | 0 | 0.375 | 0.292 | 0.173 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ارتفاعا | 0.13 | 0.243 | 0.315 | 0.375 | 0 | 0.365 | 0.281 | 0.158 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| في | 0 | 0 | 0.2 | 0.292 | 0.365 | 0 | 0.365 | 0.281 | 0.158 | 0 | 0 | 0 | 0 | 0 | 0 |
| عدد | 0 | 0 | 0 | 0.173 | 0.281 | 0.365 | 0 | 0.365 | 0.281 | 0.158 | 0 | 0 | 0 | 0 | 0 |
| المكالمات | 0 | 0 | 0 | 0 | 0.158 | 0.281 | 0.365 | 0 | 0.365 | 0.281 | 0.158 | 0 | 0 | 0 | 0 |
| الوهمية | 0 | 0 | 0 | 0 | 0 | 0.158 | 0.281 | 0.365 | 0 | 0.365 | 0.281 | 0.173 | 0 | 0 | 0 |
| الصادرة | 0 | 0 | 0 | 0 | 0 | 0 | 0.158 | 0.281 | 0.365 | 0 | 0.365 | 0.292 | 0.2 | 0 | 0 |
| عن | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.158 | 0.281 | 0.365 | 0 | 0.375 | 0.315 | 0.243 | 0.13 |
| اجهزة | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.173 | 0.292 | 0.375 | 0 | 0.406 | 0.359 | 0.313 |
| الهاتف | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.2 | 0.315 | 0.406 | 0 | 0.462 | 0.436 |
| النقال | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.243 | 0.359 | 0.462 | 0 | 0.554 |
| الذكية | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.13 | 0.313 | 0.436 | 0.554 | 0 |

**Figure 3.8. Matrice des corrélations normalisées dans le cas sans stem**

✓ **Cas avec stem :**

1) Construction de la matrice des cooccurrences :

| | شرك | امن | علم | رصد | رفع | عدد | كلم | وهم | صدر | جهز | هتف | نقل | ذكي |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| شرك | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| امن | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| علم | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| رصد | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| رفع | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 |
| عدد | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 | 0 |
| كلم | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 | 0 |
| وهم | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 | 0 |
| صدر | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 | 1 |
| جهز | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 | 2 |
| هتف | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 | 3 |
| نقل | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 | 4 |
| ذكي | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 0 |

**Figure 3.9. Matrice des cooccurrences dans le cas avec stem.**

2) Matrice des corrélations :

| | شرك | امن | علم | رصد | رفع | عدد | كلم | وهم | صدر | جهز | هتف | نقل | ذكي |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| شرك | -0.048 | 0.301 | 0.182 | 0.088 | 0.007 | -0.069 | -0.069 | -0.069 | -0.069 | -0.067 | -0.063 | -0.057 | -0.048 |
| امن | 0.301 | -0.068 | 0.204 | 0.119 | 0.047 | -0.018 | -0.082 | -0.082 | -0.082 | -0.08 | -0.075 | -0.068 | -0.057 |
| علم | 0.182 | 0.204 | -0.084 | 0.153 | 0.086 | 0.027 | -0.032 | -0.092 | -0.092 | -0.089 | -0.084 | -0.075 | -0.063 |
| رصد | 0.088 | 0.119 | 0.153 | -0.095 | 0.128 | 0.072 | 0.015 | -0.041 | -0.097 | -0.095 | -0.089 | -0.08 | -0.067 |
| رفع | 0.007 | 0.047 | 0.086 | 0.128 | -0.1 | 0.12 | 0.065 | 0.01 | -0.045 | -0.097 | -0.092 | -0.082 | 0.069 |
| عدد | -0.069 | -0.018 | 0.027 | 0.072 | 0.12 | -0.1 | 0.12 | 0.065 | 0.01 | -0.041 | -0.092 | -0.082 | 0.069 |
| كلم | -0.069 | -0.082 | -0.032 | 0.015 | 0.065 | 0.12 | -0.1 | 0.12 | 0.065 | 0.015 | -0.032 | -0.082 | -0.069 |
| وهم | -0.069 | -0.082 | -0.092 | -0.041 | 0.01 | 0.065 | 0.12 | -0.1 | 0.12 | 0.072 | 0.027 | -0.018 | -0.069 |
| صدر | -0.069 | -0.082 | -0.092 | -0.097 | -0.045 | 0.01 | 0.065 | 0.12 | -0.1 | 0.128 | 0.086 | 0.047 | 0.007 |
| جهز | -0.067 | -0.08 | -0.089 | -0.095 | -0.097 | -0.041 | 0.015 | 0.072 | 0.128 | -0.095 | 0.153 | 0.119 | 0.088 |
| هتف | -0.063 | -0.075 | -0.084 | -0.089 | -0.092 | -0.092 | -0.032 | 0.027 | 0.086 | 0.153 | -0.084 | 0.204 | 0.182 |
| نقل | -0.057 | -0.068 | -0.075 | -0.08 | -0.082 | -0.082 | -0.082 | -0.018 | 0.047 | 0.119 | 0.204 | -0.068 | 0.301 |
| ذكي | -0.048 | -0.057 | -0.063 | -0.067 | -0.069 | -0.069 | -0.069 | -0.069 | 0.007 | 0.088 | 0.182 | 0.301 | -0.048 |

**Figure 3.10. Matrice des corrélations dans le cas avec stem.**

3) Construction de la matrice des corrélations normalisées :

| | شرك | امن | علم | رصد | رفع | عدد | كلم | وهم | صدر | جهز | هتف | نقل | ذكي |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| شرك | 0 | 0.549 | 0.427 | 0.297 | 0.084 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| امن | 0.549 | 0 | 0.452 | 0.345 | 0.217 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| علم | 0.427 | 0.452 | 0 | 0.391 | 0.293 | 0.164 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| رصد | 0.297 | 0.345 | 0.391 | 0 | 0.358 | 0.268 | 0.122 | 0 | 0 | 0 | 0 | 0 | 0 |
| رفع | 0.084 | 0.217 | 0.293 | 0.358 | 0 | 0.346 | 0.255 | 0.1 | 0 | 0 | 0 | 0 | 0 |
| عدد | 0 | 0 | 0.164 | 0.268 | 0.346 | 0 | 0.346 | 0.255 | 0.1 | 0 | 0 | 0 | 0 |
| كلم | 0 | 0 | 0 | 0.122 | 0.255 | 0.346 | 0 | 0.346 | 0.255 | 0.122 | 0 | 0 | 0 |
| وهم | 0 | 0 | 0 | 0 | 0.1 | 0.255 | 0.346 | 0 | 0.346 | 0.268 | 0.164 | 0 | 0 |
| صدر | 0 | 0 | 0 | 0 | 0 | 0.1 | 0.255 | 0.346 | 0 | 0.358 | 0.293 | 0.217 | 0.084 |
| جهز | 0 | 0 | 0 | 0 | 0 | 0 | 0.122 | 0.268 | 0.358 | 0 | 0.391 | 0.345 | 0.297 |
| هتف | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.164 | 0.293 | 0.391 | 0 | 0.452 | 0.427 |
| نقل | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.217 | 0.345 | 0.452 | 0 | 0.549 |
| ذكي | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.084 | 0.297 | 0.427 | 0.549 | 0 |

**Figure 3.11. Matrice des corrélations normalisées dans le cas avec stem.**

### 3.2.4 Post traitement du corpus

En effet, la similarité sémantique est fondée sur l'idée disant que les termes similaires se trouvent fréquemment dans le même contexte. Néanmoins, dans un contexte, les termes n'ont pas tous la même importance. Par conséquent, il est plus judicieux de pondérer l'impact des termes avant l'application d'une mesure de similarité. La fréquence d'apparition d'un terme dans un document est un bon indicateur de l'importance de ce terme.

Une fonction de pondération attribue à chaque terme de chaque document une valeur. Cette valeur (ou poids) est calculée en tenant compte de deux grands critères [50]:

- ✓ La force (capacité) locale du terme dans le document (c'est-à-dire mesurer l'importance du terme dans le document dans lequel il parait).
- ✓ La force globale (c'est-à-dire mesurer l'importance du terme dans tout le corpus).

Plus un terme est présent (fréquent) dans un document, plus sa force locale est importante et plus ce terme est présent dans le corpus, plus sa force globale est élevée.

1. ***TF (Term-Frequency):*** Ou la fréquence du terme dans un document.

   Cette pondération repose sur le calcul de la fréquence du terme dans le document(le nombre de fois que le terme apparait dans le document). Plus un terme est fréquent dans un document plus il est important dans la description de ce document.

$$TF = \begin{cases} \text{Freq}(t,d) \\ TF_{log} = -\log\left(\frac{TF_{count}}{n}\right) \\ TF_{min-max} = \frac{TF_{log}}{\max(TF_{log})} \\ \frac{\text{Freq}(t,d)}{\sum_{\forall t' \in d} \text{Freq}(t',d)} \end{cases}$$

D'où :

- t : terme et d : document
- $TF_{count}$ : le nombre d'occurrence du mot dans le corpus et n est le nombre total de mot dans le corpus
- $TF_{log}$=le nombre de fois que le mot apparait dans le corpus [26]

Supposons que nous avons un corpus contenant n document (D1, D2,…, Dn) et un terme t. la valeur TF(t, D1) est le nombre d'occurrence du terme t dans le document D1 seulement.

2. **IDF (Inverse-Document-Frequency):** Ou la fréquence inverse du document

C'est une mesure de l'importance du terme dans l'ensemble du corpus. La fréquence du document (DF) est le nombre de documents du corpus dont un terme apparaît tandis que la fréquence inverse du document (IDF) est le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme. Cette norme vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants (par exemple les stop-words ont généralement une fréquence élevée alors qu'ils ne sont pas vraiment important par rapport aux autres mots moins fréquent qu'eux).

$$IDF_i = \log \frac{|D|}{\left|\{d_j : t_j \in d_j\}\right|} \qquad (1)$$

D'où :

- $|D|$ : nombre total de documents dans le corpus ;
- $\left|\{d_j : t_j \in d_j\}\right|$ : nombre de documents où le terme $t_j$ apparaît [3].

3. **TFIDF :**

Cette méthode représente l'importance du terme dans le document relativement dans le corpus de document. Elle s'obtient en multipliant les deux mesures TF et IDF:

$$TFIDF = TF * IDF \qquad (2)$$

Dans notre approche, les méthodes de pondération sont utilisées pour distinguer l'importance des termes les uns des autres ou du vocabulaire des deux documents pour lesquels nous mesurons leur similitude. Dans notre thématique, le couple de réponses (modèle, étudiant) représente les deux documents. Cependant, ce qui est certain, est que ce vocabulaire est très petit pour le prendre comme document de source pour le calcul statistique des pondérations (vocabulaire non significatif). Par conséquent, nous nous somme basé sur le concept disant qu'un corpus de la langue en question (l'arabe dans notre cas) est un bon représentant des termes de la langue en générale. Par la suite, nous avons généré les différentes pondérations (TF, IDF, TF IDF) des mots (BOW) de chacun des cinq corpus prétraités (cnn, bbc+cnn, khaleej, Watan, Osac) les résultats obtenues seront présentés dans le prochain chapitre.

Les formules adoptées au sein de ce travail sont :

- La formule $TF_{min-max}$ . pour le calcul des TF,
- La formule (1) pour le calcul des IDF,
- La formule (2) pour le calcul des TFIDF.

Après avoir construit notre espace sémantique, nous sauvegardons ce dernier dans un fichier « .txt ». Chaque ligne de ce fichier représente un vecteur d'un mot. Les éléments d'un vecteur du mot sont séparés par des espaces simples. Chaque élément est sous format chaine de caractères « String ».

## 3.3 Modèles du calcul de similarité sémantique entre deux réponses courtes

Après avoir construit l'espace sémantique, nous entamons la notion du calcul de similarité sémantique. Dans notre travail, nous avons considéré deux niveaux du calcul de similarité :

i. Similarité entre phrases : nous avons réalisé ce concept par un modèle que nous nommons « somme vecteurs ».

ii. Similarité mot-à-mot : nous avons concrétisé ce concept par un modèle de calcul matriciel.

iii. Une hybridation des deux modèles précédents.

### 3.3.1 Le modèle somme-vecteurs (SV)

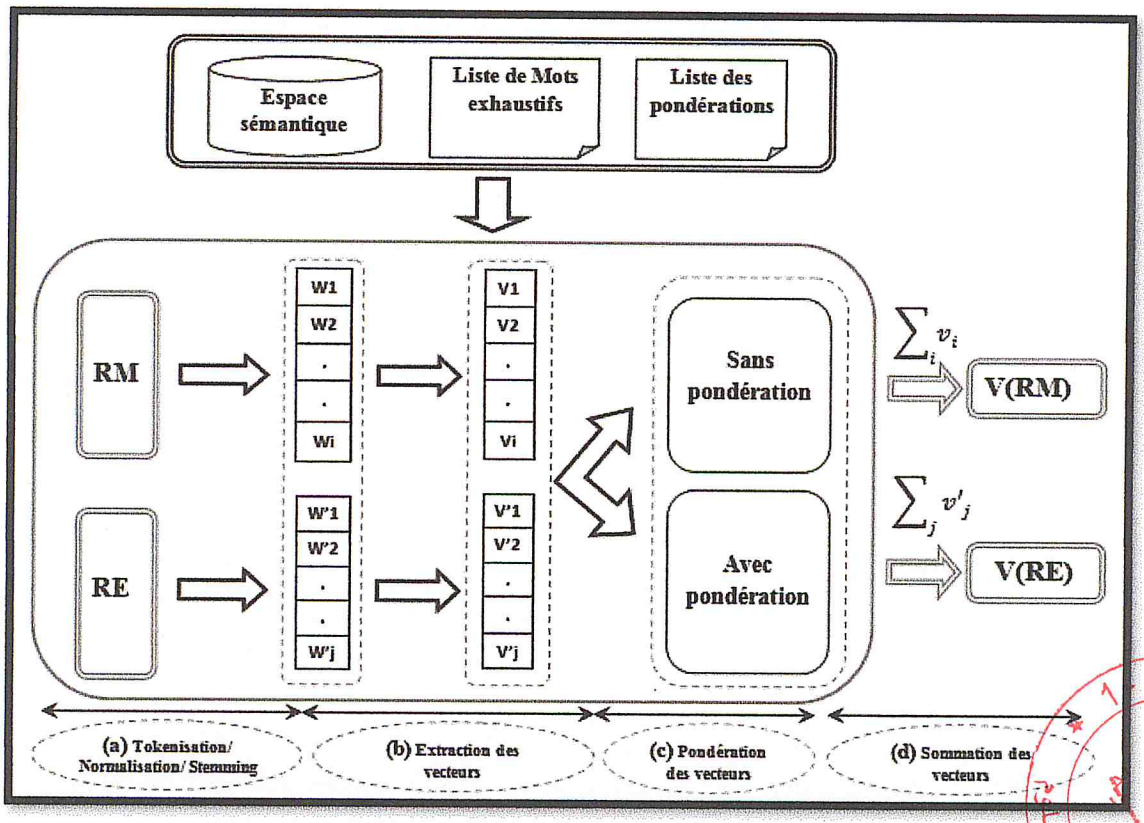



**Figure 3.12. Vue globale sur le fonctionnement du modèle SV.**

Nous détaillons les étapes du fonctionnement de la SV représentées au niveau de la « Figure 3.12.» :

a) Une étape de prétraitement (normalisation, stemming,…) est effectuée sur les deux réponses (RM et RE). Par la suite, chaque réponse est transformée en vecteur en appliquant une tokenisation. Deux vecteurs de mots résultent de cette étape.

b) Maintenant, les mots de chaque réponse sont remplacés par leurs vecteurs de contexte. Cela est fait en récupérant le vecteur de contexte du mot à partir de l'espace sémantique construit et stocké précédemment. Un vecteur de vecteurs résulte à ce stade pour chaque réponse.

c) Ensuite, nous allons récupérer des valeurs de pondération de chaque mot. Ici nous distinguons deux cas : avec une pondération et sans aucune pondération. Dans le cas de pondération, chaque vecteur de contexte du mot i est multiplié par la valeur de pondération qui représente l'importance du mot i.

d) Chaque vecteur de vecteurs de contexte passe par une étape de sommation. Cela est obtenue en sommant les vecteurs de contexte de chaque réponse mot par mot ou proprement dit case par case selon la formule suivante :

$$V(R)=\sum_{j=1}^{i} v_j$$

D'où:

- R est la réponse dont nous calculons son vecteur (réponse modèle/étudiant).
- $i$ est le nombre de termes/mots constituant la réponse R après l'étape du prétraitement.
- $v_j$ est le vecteur de contexte du $j^{éme}$ terme de la réponse R.

Nous arrivons au stade du calcul de la valeur de similarité entre les deux réponses. Nous appliquons la mesure **Cosine** sur les vecteurs V(RM) et V(RE). Une valeur entre 0 et 1 est obtenue et qui représente la valeur de similarité entre les deux réponses.

➢ **Exemple illustratif :**

Nous appliquons ce modèle sur le premier couple de réponses du dataset de Gomaa :

a) Les tâches de cette étape sont représentées dans la « Figure 3.13.».

b) Récupération des vecteurs de contexte des mots depuis l'espace sémantique :

**Vm** = v(درس) + v(نول) + v(جنب) + v(طبع) + v(حدد) + v(حيا) + v(كون) + v(خدم) + v(كون) + v(بيئه)

**Ve** = v(درس) + v(نول) + v(كون) + v(بيئه) + v(خدم) + v(أنس)

- Un exemple d'un vecteur de contexte :

$$v(\text{درس}) = [0.01, 0.12, 0.035, 0, 0, \ldots]$$

| | Réponse Modèle | Réponse Etudiant |
|---|---|---|
| Etat initial | الدراسة التي تتناول جوانب الطبيعة بما يحدده حياة الكائن و كيفية استخدامه لمكونات البيئة . | الدراسة التي تتناول مكونات البيئة و استخدام الإنسان لها. |
| Normalisation | الدراسه التي تتناول جوانب الطبيعه بما يحدده حياه الكائن و كيفيه استخدامه لمكونات البيئه . | الدراسه التي تتناول مكونات البيئه و استخدام الانسان لها. |
| Stemming | درس نول جنب طبع حدد حيا كون خدم كون بيئه | درس نول كون بيئه خدم أنس |
| Tokenisation | Vm= درس \| نول \| جنب \| طبع \| حدد \| حيا \| كون \| خدم \| كون \| بيئه | Ve= درس \| نول \| كون \| بيئه \| خدم \| أنس |

**Figure 3.13. Etape a)**

**c)** Nous supposons que les valeurs des pondérations sont représentées dans la «Figure 3.14.».

| *Terme* | درس | نول | جنب | طبع | حدد | حيا | كون | خدم | بيئه | أنس |
|---|---|---|---|---|---|---|---|---|---|---|
| *Poids* | 0.22 | 0.34 | 0.10 | 0.20 | 0.32 | 0.12 | 0.22 | 0.37 | 0.19 | 0.27 |

**Figure 3.14. Valeurs des pondérations**

- **Vm** = 0.22*v(درس) + 0.34*v(نول) +0.10*v(جنب) + 0.20*v(طبع) + 0.32*v(حدد) + 0.12*v(حيا) + 0.22*v(كون) + 0.37*v(خدم) + 0.22*v(كون) + 0.19*v(بيئه)
- **Ve** = 0.22*v(درس) + 0.34*v(نول) + 0.22*v(كون) + 0.19*v(بيئه) + 0.37*v(خدم) + 0.27*v(أنس)

**d)** Cette tâche est présentée dans la « Figure 3.15.».

Maintenant nous calculons la similarité entre les deux réponses :

- Le résultat sans pondération :

$$Sim(RM, RE) = Sim_{Cosin}(Vm, Ve) = 0.76$$

- Le résultat avec pondération :

$$Sim(RM, RE) = Sim_{Cosin}(Vm, Ve) = 0.79$$

| Réponse Modèle | Réponse Etudiant |
|---|---|
| v (درس) = [0.01 ,0.12 ,0.035,...] | v (درس) = [0.01 ,0.12 ,0.035,...] |
| v (نول) = [0.02 ,0 ,0 , ...] | v (نول) = [0.02 ,0 ,0 , ...] |
| v (جنب) = [0 ,0 ,0.1,...] | v (كون) = [0 ,0 ,0.025,...] |
| v (طبع) = [0 ,0.045 ,0.01,...] | v (بيئه) = [0.004 ,0.002 ,0.1,...] |
| v (حدد) = [0.1 ,0 ,0.015,...] | v (خدم) = [0 ,0.01 ,0,...] |
| v (حيا) = [0 ,0.034 ,0.002,...] | v (أنس) = [0.02 ,0 ,0.1,...] |
| v (كون) = [0 ,0 ,0.025,...] | |
| v (خدم) = [0 ,0.01 ,0,...] | |
| v (كون) = [0 ,0 ,0.025,...] | |
| v (بيئه) = [0.004 ,0.002 ,0.1,...] | |
| **Vm** = [0.134 ,0.12 ,0.312,...] | **Ve** = [0.054 ,0.132 , 0.26,...] |

**Figure 3.15. Etape d)**

### 3.3.2 Le modèle calcul-matriciel (CM)

Contrairement au modèle précédent, nous construisons une matrice dont les termes d'une réponse représentent les lignes de la matrice tandis que les termes de l'autre réponse représentent les colonnes. Par la suite, chaque élément de la matrice signifie la valeur de similarité entre le vecteur de contexte du $i^{éme}$ terme (réponse 1) et celui du $j^{éme}$ terme (réponse 2).Nous nous sommes inspirés du travail de Islam et Inkpen [29] Ce modèle prend en considération l'ordre entre mot. Ci-dessous nous présentons les étapes détaillées :

a) Une étape de prétraitement (normalisation, stemming,…) est effectuée sur les deux réponses RM, RE. Par la suite, chaque réponse est transformée en vecteur en appliquant une tokenisation. Deux vecteurs de mots de taille m et n résultent de cette étape (la même étape que celle du modèle SV).

b) Suppression des termes en commun « ou encore : les mots qui matchent notant $\delta$ le nombre de ces mots» de chaque réponse en gardant trace de l'ordre entre les termes de chaque réponse. Nous aurons les deux réponses REM, RMM de taille (m- $\delta$) *et (n- $\delta$).*

c) Si m- $\delta$ = *0 ou n-* $\delta$, = 0 ou les deux sont nuls, nous passons à l'étape h), sinon nous passons à l'étape d).

d) Construction de la matrice sémantique MSem de taille (m- $\delta$)*(n- $\delta$) dont *la condition* consiste que (**m** ≤ **n**) c'est-à-dire les lignes correspondent aux mots de la réponse (RMM ou REM) qui a une longueur inferieur ou égale par rapport à l'autre (les

colonnes correspondent aux mots de la longue réponse). Chaque élément de MSem présente la corrélation entre les vecteurs de contexte de chaque couple de mots (a, b) tel que a ∈.lignes et b ∈ colonnes, comme le montre la formule suivante (prise de [15]) :

$$MSem(a,b) = \frac{\sum(a_i - \bar{a})(b_i - \bar{b})}{\left(\sum(a_i - \bar{a})^2 \ \sum(b_i - \bar{b})^2\right)^{1/2}}$$

D'où $\bar{a}$ $(resp.\bar{b})$ est la moyenne du vecteur du contexte du mot $a$ $(resp.b)$.

Dans le cas d'une combinaison avec d'autres approches (syntaxique ou word-Embedding), effectuer les tâches e),f). Sinon passer directement à l'étape g) et considérer MSem comme matrice combinée M.

e) Construction de la matrice syntaxique MSyn, ou de la matrice des word-Embedding MWE, ou les deux matrices de taille (m- $\delta$)*(n- $\delta$). Il est à noter que la condition de l'étape d) doit être vérifiée.

f) Construction d'une matrice combinée M de taille (m- $\delta$)*(n- $\delta$) des deux matrices (« MSem,MSyn » ou « MSem,MWE ») ou des trois matrices (MSem,MSyn,MWE) précédentes en effectuant la moyenne comme suit :

$$\begin{cases} M \leftarrow 0.5 * M_1 + 0.5 * M_2 \\ M \leftarrow (M_1 + M_2 + M_3)/3 \end{cases}$$

g) Nous sauvegardons la valeur maximale M (i, j) de la matrice M dans une liste ρ ($\rho \leftarrow \rho \cup M(i,j)$), puis nous supprimons tous les éléments correspondants à sa colonne j et sa ligne i de M. Refaire cette étape jusqu'à la satisfaction de l'une des conditions suivantes :

- *La somme des éléments de la matrice M est nulle.*
- *La matrice M est vide.*

h) Calculer le score de similarité noté So. Considérons le couple de réponse, RM et RE ont respectivement m et n jetons (tokens), c'est-à-dire RM = p1, p2, . . ., pm et RE = r1, r2, . . ., rn et n ≥ m. Sinon, nous changeons RM et RE. Nous comptons le nombre de pi (étant δ) pour lequel pi = rj, pour tout p ∈ RM et pour tout r∈ RE. Autrement dit, il y a des jetons δ dans RM exactement correspond à RE, où δ ≤ m. Nous enlevons tous les jetons δ de RM et les place dans X et RE dans Y, dans le même ordre que dans les réponses. Donc, X = {x1, x2,. . . , xδ} et Y = {y1, y2,. . . , yô}. Nous remplaçons X en assignant un numéro d'index unique pour chaque jeton dans X, de 1 à

δ, c'est-à-dire X = {1, 2,. . . , δ}. Sur la base de ces numéros d'index uniques pour chaque jeton dans X, nous remplaçons également Y où X = Y. La formule suivante est utilisée pour mesurer la similitude de l'ordre des mots communs du couple de réponses:

$$So = 1 - \frac{|x_1 - y_1| + |x_2 - y_2| + \cdots + |x_\delta - y_\delta|}{|x_1 - x_\delta| + |x_2 - x_{\delta-1}| + \cdots + |x_\delta - x_1|}$$

i) Nous sommons tous les éléments de ρ et nous y ajoutons δ (1 - wf + wf So) pour obtenir un score total, où So est le score de similarité d'ordre de mots communs et wf est le poids d'ordre des mots communs où wf ∈ [0, 0.5]. Nous multiplions ce score total par la moyenne harmonique réciproque de m et n pour obtenir un score de similarité équilibré entre 0 et 1, inclusivement.

$$S(P,R) = \frac{(\delta(1-wf+wfS_0)+\sum_{i=1}^{|\rho|}\rho i)\times(m+n)}{2nm} \qquad (11)$$

- L'équation (11) est simplifier en équation (12) si :
  - L'importance de l'information syntaxique est ignorée en mettant wf à 0.
  - La valeur de similarité d'ordre de mots communs So=1.

$$S(P,R) = \frac{(\delta+\sum_{i=1}^{|\rho|}\rho i)\times(m+n)}{2mn} \qquad (12)$$

Enfin, pour mesurer la similarité d'une paire de mots, ce modèle est ensuite utilisé. Nous déterminons la similarité sémantique et syntaxique/WordEmbedding et les combinons en un score de similarité. Ces scores sont ensuite utilisés pour calculer la similarité globale de deux segments de texte. Le score de similarité pour chaque paire de mots est pondéré en tenant compte de la spécificité de ses mots. Afin d'améliorer les résultats de notre modèle, nous avons exprimé la spécificité des mots en utilisant la pondération (présenté dans la section 3.2.4) de fréquence terme normalisé, qui donne un poids plus élevé aux paires avec des mots plus spécifiques.

1) **Exemple illustratif**

Ci-dessous (dans Tableau 3.2.) le couple du dataset de Gomaa (Le deuxième couple de réponses) sur lequel nous appliquons l'étape a) et b):

**Tableau 3.2. Etape a) et b) du modèle CM.**

| Etape | | Réponse Modèle | Réponse étudiant |
|---|---|---|---|
| a) | **Initiale** | الدراسة التى تتناول جوانب الطبيعة بما يحدده حياة الكائن و كيفية استخدامه لمكونات البيئة | هو العلم الذى يتناول كل ما له علاقة بالأرض من حيث مكوناتها وحركتها و تاريخها و الظواهر التى تحدث عليها |
| | **Prétraitement** | درس نول جنب طبع حدد حيا كون خدم كون بيئه | علم نول علق أرض كون حرك أرخ ظهر حدث |
| b) | **Suppression des δ** | درس جنب طبع حدد حيا خدم كون بيئه | علم علق أرض حرك أرخ ظهر حدث |
| | **La liste δ** | نول ,كون | |

e) Voici la matrice générée avec le modèle CM dans le cas d'une matrice sémantique :

**M1=**

| | درس | جنب | طبع | حدد | حيا | خدم | كون | بيئة |
|---|---|---|---|---|---|---|---|---|
| علم | 0.017 | 0.015 | 0.013 | 0.001 | 0.007 | 0.027 | − 0.002 | 0. |
| علق | 0.039 | 0.011 | 0.047 | 0.023 | 0.027 | 0.016 | 0.023 | 0. |
| ارض | 0.027 | 0.014 | 0.034 | 0. | 0.051 | 0.015 | 0.045 | 0. |
| حرك | 0.011 | 0.064 | 0.013 | 0.035 | 0.041 | 0.040 | 0.027 | 0. |
| ارخ | 0.026 | 0.038 | 0.047 | 0.027 | 0.055 | 0.017 | 0.023 | 0. |
| ظهر | 0.086 | 0.021 | 0.046 | 0.002 | 0.071 | 0.050 | 0.023 | 0. |
| حدث | 0.009 | 0.013 | 0.013 | 0.019 | 0.008 | 0.016 | − 0.002 | 0. |

Par la suite, nous appliquons une itération pour récupérer le maximum de la matrice et supprimer la ligne et la colonne de cet élément :

**M2=**

| | جنب | طبع | حدد | حيا | خدم | كون | بينة |
|---|---|---|---|---|---|---|---|
| علم | 0.015 | 0.013 | 0.001 | 0.007 | 0.027 | − 0.002 | 0. |
| علق | 0.011 | 0.047 | 0.023 | 0.027 | 0.016 | 0.023 | 0. |
| ارض | 0.014 | 0.034 | 0. | 0.051 | 0.015 | 0.045 | 0. |
| حرك | 0.064 | 0.013 | 0.035 | 0.041 | 0.040 | 0.027 | 0. |
| ارخ | 0.038 | 0.047 | 0.027 | 0.055 | 0.017 | 0.023 | 0. |
| حدث | 0.013 | 0.013 | 0.019 | 0.008 | 0.016 | − 0.002 | 0. |

**M3=**

| | طبع | حدد | حيا | خدم | كون | بينة |
|---|---|---|---|---|---|---|
| علم | 0.013 | 0.001 | 0.007 | 0.027 | − 0.002 | 0. |
| علق | 0.047 | 0.023 | 0.027 | 0.016 | 0.023 | 0. |
| ارض | 0.034 | 0. | 0.051 | 0.015 | 0.045 | 0. |
| ارخ | 0.047 | 0.027 | 0.055 | 0.017 | 0.023 | 0. |
| حدث | 0.013 | 0.019 | 0.008 | 0.016 | − 0.002 | 0. |

**M4=**

| | طبع | حدد | خدم | كون | بينة |
|---|---|---|---|---|---|
| علم | | | | | |
| علق | | | | | |
| ارض | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| حدث | 0.013 | 0.001 | 0.027 | − 0.002 | 0 |
| | 0.047 | 0.023 | 0.016 | 0.023 | 0. |
| | 0.034 | 0. | 0.015 | 0.045 | 0. |
| | 0.013 | 0.019 | 0.016 | − 0.002 | 0. |

**M5=**

| | حدد | خدم | كون | بيئة |
|---|---|---|---|---|
| علم | 0.001 | 0.027 | − 0.002 | 0. |
| ارض | 0. | 0.015 | 0.045 | 0. |
| حدث | 0.019 | 0.016 | − 0.002 | 0 |

**M6=**

| | حدد | خدم | بيئة |
|---|---|---|---|
| علم | 0.001 | 0.027 | 0. |
| حدث | 0.019 | 0.016 | 0. |

**M7=**

| | حدد | بيئة |
|---|---|---|
| حدث | 0.019 | 0. |

➢ <u>Le calcul de similarité :</u>

La liste des max : [0.086 ,0.064 ,0.055 ,0.047 ,0.045 ,0.019]

Application de la formule (12) détaillée précédemment.

Similarité = 0.25

### 3.3.3 Hybridation

Dans cette phase la, nous visons a concrétiser la notion d'hybridation entre les mesures afin d'avoir de meilleurs résultats. Pour ce faire, nous allons adopter une combinaison interne ainsi qu'une combinaison externe.

Pour la combinaison interne, nous combinons les résultats des deux modèles de calcul précédents en deux manières :

1) **La moyenne :**

Dans ce type, nous prenons la moyenne des deux valeurs de similarité (calculé par SV et CM) entre les deux réponses. Les résultats seront présentés dans le chapitre suivant.

2) **Le maximum :**

Contrairement a la moyenne, dans ce cas, nous prenons la valeur maximum entre les deux valeurs de similarité. De même, les résultats dans le prochain chapitre.

Pour la combinaison externe, nous allons maintenant combiner nos modèles de base avec les méthodes de deux autres binômes travaillant sur la même thématique :

- Hennich Adel Nassim et Hannoufi Mohamed Hamza [51].
- Abdallah Amina et Guerroudja khadidja dont le thème est « Mesures de similarité syntaxique pour un système d'évaluation automatique des réponses courtes : Application à la langue arabe ».

## 3.4 Passage au score

Après avoir mesuré la valeur de similarité entre les deux réponses, nous allons effectuer un passage au score de la valeur de similarité vers un note ou score selon le barème donné. Pour ce faire, nous adoptons la technique suivante :

**Classifieur k-means**[52]**(Ou k-moyenne)**

C'est une technique de classification non supervisée. Elle consiste à avoir un ensemble de données comme entrée (input) et elle donne en sortie ces données classifiées dans K classes. C'est un algorithme très populaire et facile dans sa mise en œuvre. L'inconvénient de cet algorithme réside dans l'exigence de la détermination du nombre de classe comme entrée. Néanmoins, dans notre approche, ceci présente un avantage car nous fixons dés le départ le nombre des classes selon le barème donnée.

Dans le prochain chapitre, les couples de réponses dataset utilisés dans l'évaluation de notre approche sont noté sure 5 points. Nous fixons K à 11 pour avoir 11 classes de note (0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5) afin de prendre en compte la subjectivité du processus d'évaluation.

## 3.5 Conclusion

Au cours de ce chapitre, nous venons de présenter les principes de notre système. Ce dernier est un ensemble de plusieurs modules intégrés pour effectuer une évaluation automatique de réponses courtes en langue arabe. Ce processus nécessite un prétraitement du corpus dont il sera utilisé pour créer notre VSM « espace sémantique » ainsi que pour la génération des pondérations dans le but de tester son impact. Le processus suivant s'en charge du calcul de similarité qui sera convertis en score par la suite. Pour se faire, nous nous sommes inspirés d'autres travaux pour élaborer les modèles proposés (SV, CM, Hybridation).

Après avoir conçue notre système, des métriques d'évaluation ainsi que des dataset seront utilisés pour déterminer le degré de son efficacité et sa fiabilité. Les résultats des tests ainsi que ceux des hybridations seront présentés dans le chapitre suivant.

# Chapitre 4 : Résultats expérimentaux et évaluation



Dans ce chapitre, nous montrons les expériences et les résultats des tests réalisés afin d'estimer la qualité de notre système et sa performance comparée aux autres systèmes qui existent, ainsi nous discutons les résultats obtenus.

Le « Tableau 4.1. » présente la liste des abréviations employée au cours de ce chapitre.

**Tableau 4.1. Liste des abréviations**

| Abréviation | Expression |
|---|---|
| DS | Dataset |
| CM | Modèle Calcul Matriciel |
| SV | Modèle Somme-Vecteurs |
| WE | Word-Embeddings |

## 4.1 Démarche expérimentale

### 4.1.1 Ressources matérielles et logicielles utilisées lors du développement

Nous avons rencontré un problème matériel dans la phase principale de notre approche « Traitement du corpus ». Ce dernier n'a pas abouti sa fin à cause de l'insuffisance de la mémoire de nos PC portables (4 Go de RAM). Pour cela, nous avons eu la chance de travailler sur un serveur à distance « Voir Figure 4.1.» fournit par l'université de Bouira et c'est la configuration minimale requise pour notre travail.

**HP 470065-652 ProLiant ML350p Gen8 Intel Xeon**

**Processeur :** E5-2620 / 2 GHz
**RAM:** 16 Go
**OS :** Linux

**Figure 4.1. Références du serveur à distance**

Nous travaillons avec le langage de programmation *Python 3.6* sur nos machines personnelles et *Python 3.4* sur le serveur à distance ainsi que le langage *Java*.

Pour les interfaces, planifiées en java, nous avons utilisé l'environnement de développement *Netbeans 8.1* et *PyCharm Community 2018.1* pour *Python*.

### 4.1.2 Outils développés

Notre application englobe un système d'évaluation automatique des réponses courtes ainsi que d'autres outils du traitement du langage naturel que nous avons dû développer à cause du manque d'outils de traitement pour la langue arabe. Ce système se compose de différents modules intégrés entre eux présentés ci dessous :

- **Outils de Normalisation et stemming** :

  Dans cette partie, nous donnons la possibilité à l'utilisateur de percevoir le changement effectué sur un texte donné en entrée. L'utilisateur peut choisir l'opération du stemming en spécifiant le stemmer (khoja ou Tashaphyne), ou bien l'opération de la normalisation en spécifiant les paramètres à effectuer parmi une liste de paramètres (suppression des numéros, suppression des diacritiques…). En fin du traitement, l'utilisateur a la possibilité de sauvegarder le résultat « voir Figure 4.2. ».

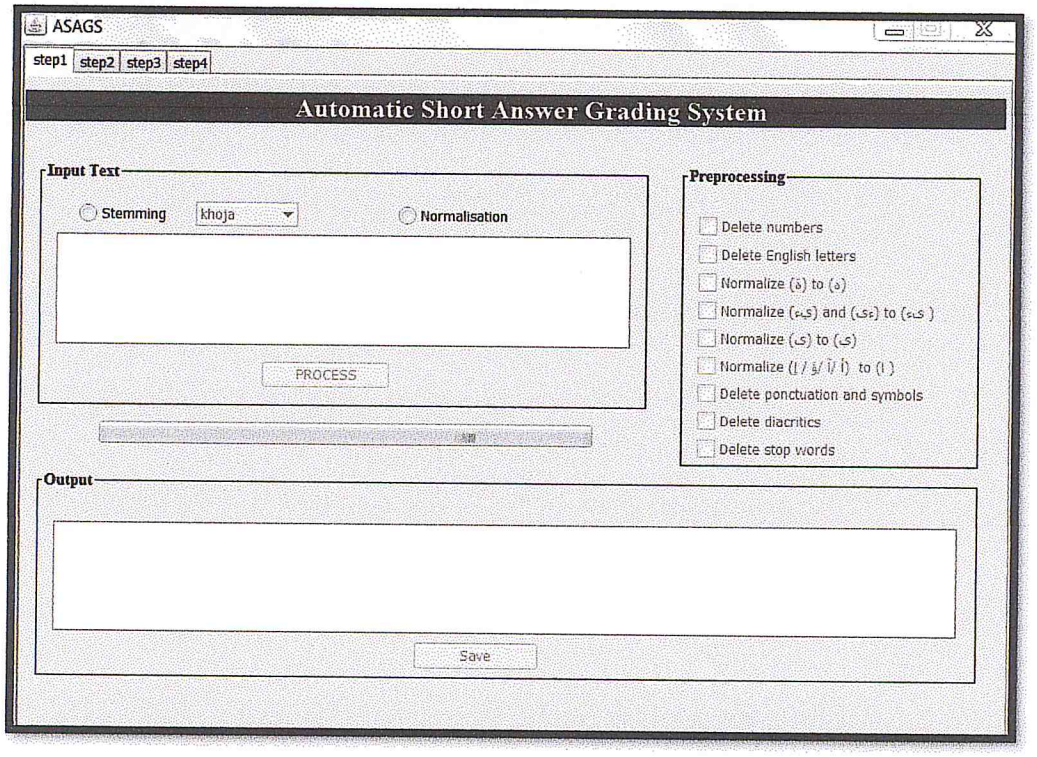



**Figure 4.2. Outils de normalisation et stemming**

- **Outil NLP**

Cette partie permet de faire différentes traitements sur un corpus choisi par l'utilisateur :

1) La normalisation,
2) Le stemming (khoja et tashapyne),
3) Les pondérations (TF, IDF, TF-IDF,),
4) Fréquence des mots et leur nombre total,
5) Liste des mots exhaustifs.

L'utilisateur doit donner en entrés le corpus sous forme d'un dossier contenant que des fichiers texte, il doit davantage indiquer le chemin du dossier de sortie, et choisir un ou bien plusieurs traitement « voir Figure 4.3.».

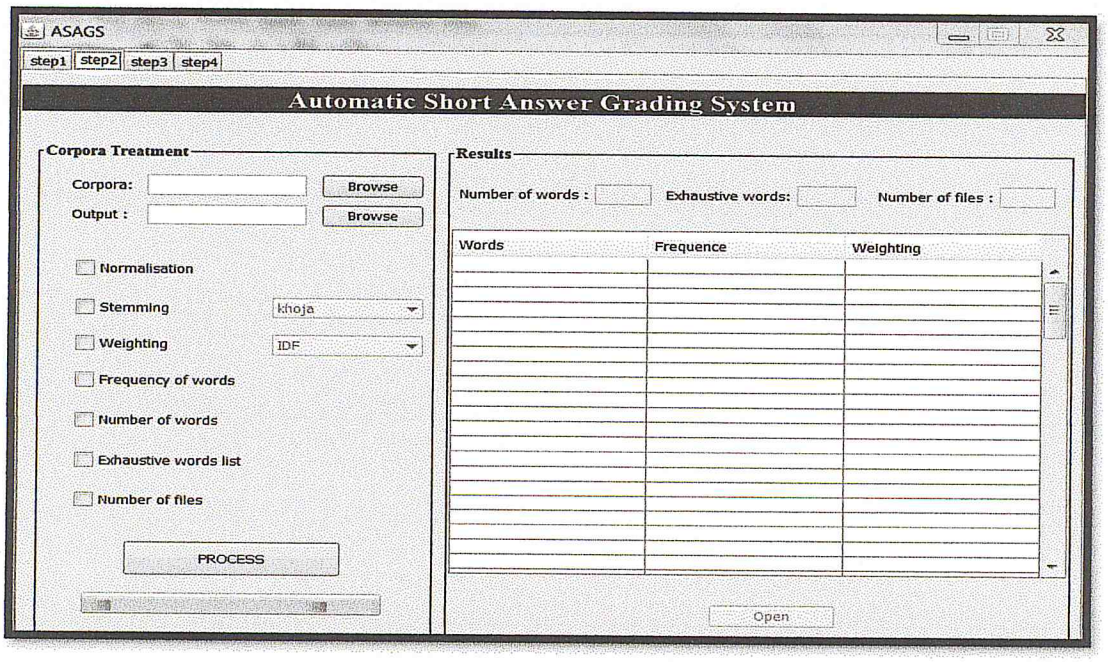



**Figure 4.3. Outil NLP**

- **Outil de création de l'espace sémantique**

Cet outil permet de calculer l'espace sémantique d'un corpus donné par l'utilisateur en entrée et il a le choix de travailler avec une liste des mots exhaustifs

externe ou bien avec la liste généré automatiquement. L'utilisateur doit donner en entrée le chemin du dossier de sortie où les résultats vont être sauvegardés « voir Figure 4.4.».

- **Outils d'évaluation automatique des réponses courte**

Ce module nous permet de calculer la similarité entre deux réponses courtes ainsi qu'un DS avec nos modèles et leurs combinaisons détaillées dans le chapitre précédent. Il permet aux utilisateurs de choisir de faire le calcul avec des ressources internes (Espace sémantique et pondération déjà calculés) ou bien de choisir des ressources externes « voir Figure 4.5.».

Du coté DS, l'utilisateur a la possibilité d'évaluer un DS de son choix c'est-à dire généré les notes automatiques, il doit avoir en entrée deux fichiers texte, l'un présente les réponses modèle et l'autre les réponses des étudiant en choisissant le barème ainsi il doit choisir un dossier de sortie. Tandis que pour le coté des deux réponses, l'utilisateur doit saisir deux réponses (RM et RE) et choisir le reste des paramètres. Le résultat s'affiche directement dans l'interface.

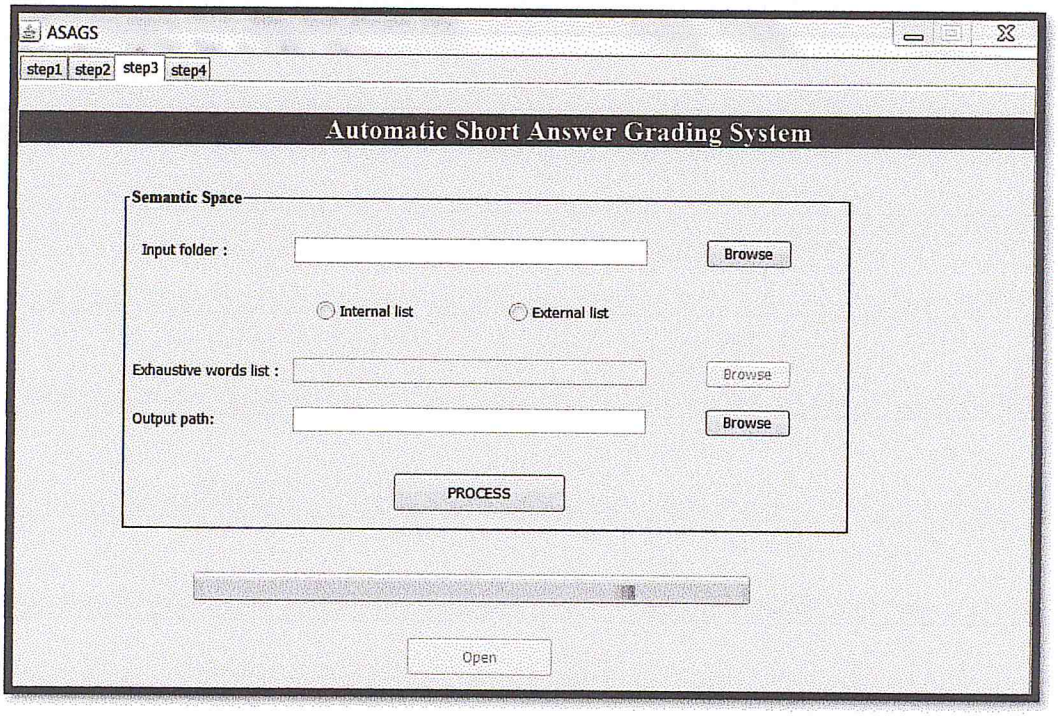



**Figure 4.4. Outil de création de l'espace sémantique**

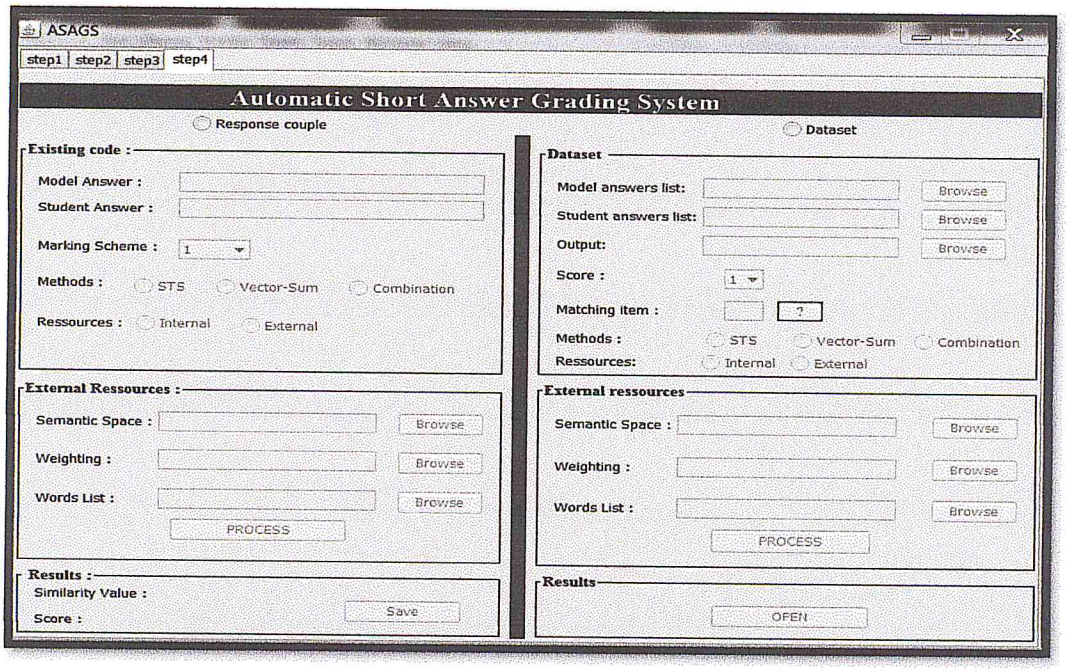



**Figure 4.5. Outil d'évaluation automatique des réponses courtes**

## 4.2 Jeux de données (Datasets) et métriques d'évaluation

### 4.2.1 Datasets[31]

L'évaluation assistée par ordinateur est caractérisée par des progrès isolés avec peu de capacités à comparer les approches et à s'appuyer sur le travail des autres chercheurs particulièrement quand nous considérons la langue arabe. Il n'existe pas à ce jour des ensembles de données publiquement disponibles pour comparer efficacement deux systèmes côte à côte.

En ce qui concerne la langue arabe il existe un seul DS ([37], [38]) largement cité dans l'évaluation des ASAGS en langue arabe et que les auteurs ont accepté de nous transmettre. Dans toute la suite nous allons considérer ce DS et l'identifier par « Gomaa DS ».

Nous avons donc effectué le test de nos différentes approches sur ce DS dans le but de comparer nos résultats par rapport à d'autre travaux ayant utilisé ce même DS.

### 1) Gomaa DS ([37], [38])

Les questions présentées dans le DS couvrent un chapitre du programme d'études égyptien officiel pour le cours de sciences de l'environnement (ES), qui représente 25% du programme global. L'ensemble de données contient 61 questions, 10 réponses pour chacune, avec un nombre total de 610 réponses. La longueur moyenne de la réponse d'un étudiant est de 2,2 phrases, 20 mots ou 103 caractères. L'ensemble de données contient une collection de réponses et notes des élèves, notées par deux annotateurs experts humains qui ont donné des notes entre 0 et 5 et obtenu un coefficient de corrélation de Pearson (r) et une erreur quadratique moyenne (RMSE) de **0,86 et 0,69**, respectivement entre les deux annotateurs. Dans toute évaluation par rapport à ce DS, l'idéal est d'approcher le plus possible ces valeurs. Nous disposons de la version XML du data-set qui nous a été fournie par les auteurs. Le « Tableau 4.2. » représente des exemples de questions, des réponses modèles et des réponses courtes fournies par deux étudiants, et des notes attribuées manuellement par deux experts humains.

**Tableau 4.2. Un échantillon du Gomaa DS**

| ID Question | Type Question | Question | Réponse Modèle (RM) | Réponse Etudiant (RE) | Note Manuelle |
|---|---|---|---|---|---|
| 1 | عرف المصطلحات العلمية | الإيكولوجيا | الدراسة التى تتناول جوانب الطبيعة بما يحدده حياة الكائن و كيفية استخدامه لمكونات البيئة. | تعنى علم دراسة مكان المعيشة أو يقصد بها العلاقات المتبادلة بين الأحياء و البيئة. | 1.5 |
| | | | | الدراسة التى تهتم بالبيئة من حيث مكوناتها و كيفية استخدام الكائنات الحية لهذه المكونات. | 4.5 |
| 22 | إشرح | القدرات الجسمية و الفسيولوجية للحيوانات البحرية التى تعيش فى الأعماق | تتحمل الضغط الزائد للماء، تتحمل البرودة الشديدة فى الأعماق، تستطيع العيش فى الظلام الدامس. | تتميز حيوانات الأعماق فى البحار بهيكل غضروفى و خصائص فسيولوجية لتتحمل الضغط المرتفع و الظلام الدامس و البرد الشديد. | 5 |
| | | | | تتحمل الحيوانات البحرية البرودة الشديدة و الظلام نتيجة تكوين أجسامها. | 3.5 |
| 29 | ما النتائج المترتبة على | تعرض السلاحف الصحراوية لدرجة حرارة منخفضة | تلجأ إلى عملية البيات الشتوى. | تقوم بعملية البيات الشتوى. | 5 |
| | | | | تقوم بالهجرة إلى المناطق التى يوجد بها درجات حرارة مناسبة لها. | 0.5 |

### 2) SemEval Datasets

Afin d'évaluer l'applicabilité et la généralisation des techniques utilisées dans notre système à d'autres domaines connexes, nous avons utilisé des ensembles de données supplémentaires qui ont été largement utilisés dans le domaine de la similarité du texte, de l'implication textuelle et de la paraphrase dans le cadre du « Semantic Evaluation(SemEval) workshop for Semantic Textual Similarity(STS) » ; une compétition qui se déroule chaque année depuis 2012. Nous avons profité du SEMEval 2017(composé de 6 tracks) [40] qui a introduit dans son « Track 1 », dédié aux couples de textes courts « arabe- arabe », plusieurs DS de tests en langue arabe. Nous avons choisi parmi les DS deux DS à savoir :

Le **STS 250 SemEval 2017** : DS d'évaluation des travaux en compétition dans le track 1.

Le **MSRvid 368 SemEval 2017** : DS proposé pour le training des données du Track 1 et que nous avons exploité pour l'évaluation des approches (noté STS 368 dans la suite).

STS est l'évaluation de paires de phrases en fonction de leur degré de similarité sémantique. La tâche implique de produire des scores de similarité à valeur réelle pour les paires de phrases. La performance est mesurée par la corrélation de Pearson des scores de machine avec des jugements humains. L'échelle ordinale guide l'annotation humaine, allant de 0 pour un chevauchement sans signification à 5 pour l'équivalence de sens. Les valeurs intermédiaires reflètent des niveaux interprétables de recouvrement partiel de sens. Les données arabes sont produites en traduisant un sous-ensemble des données anglaises et en transférant les scores de similarité. Le corpus SNLI (Stanford Natural Language Inference) [53] est la principale source de données des deux DS. Les phrases sont traduites indépendamment de leurs paires. La traduction en arabe est assurée par le CMU-Qatar par des arabophones natifs avec de solides compétences en anglais. Cinq annotations humaines sont collectées par paire. Les scores d'or font la moyenne des cinq annotations individuelles.

**Tableau 4.3. Description des deux datasets STS 250 AR et MSRvid 368 AR**

| Année | DS | Nombre de paires | Source |
|---|---|---|---|
| 2017 | STS 250 AR | 250 | SNLI |
| 2017 | MSRvid 368 AR | 368 | Vidéo (speech) |

**Tableau 4.4. Echantillon du DS STS-250 et STS-368**

| DS | ID_pair | Réponse 1 | Réponse 2 | Note_Manuelle |
|---|---|---|---|---|
| STS-250 | 224 | هناك فتاة تركب الدراجة عبر الحديقة. | صبي يركب دراجة ركوب في الحديقة. | 3.2 |
| | 172 | مهرج على وشك أن يتم دهسه من قبل ثور. | راعي البقر على وشك أن يترنح من على الثور. | 1.8 |
| | 26 | يحاول كلب الإمساك بقطة. | قطة تطارد كلبا في الخارج. | 2.6 |
| STS-368 | 2 | شخص ما يقطع بعجلة الفطر بالسكين. | الرجل يقطّع بسرعة بعض الفطر باستعمال السكين. | 3.75 |
| | 269 | طفل صغير يشرب الماء من الكوب. | هرة صغيرة تشرب الحليب من وعاء. | 0.8 |
| | 325 | رجل يفرم فرما ناعما مادة خضراء. | شخص يفرم عشبة. | 3 |

### 4.2.2 Métriques d'Evaluation [31]

L'évaluation d'un système implémenté ou d'une approche proposée est indispensable pour estimer le succès d'une recherche. Il devient primordial d'accorder un rôle central aux métriques d'évaluation qui consiste à comparer un résultat produit avec des résultats corrects attendus. L'analyse de plusieurs situations d'évaluation dans notre cas, illustre l'importance d'un choix cohérent des métriques et de l'utilisation conjointe de plusieurs métriques. En essayant d'analyser les résultats de ce travail, nous avons été confrontés à la détermination de la métrique à utiliser pour évaluer les scores obtenus par rapport aux scores manuels fournis. Notre décision de choix de métriques a été influencée par les DS et les travaux connexes qui ont utilisé ces mêmes DS. La corrélation de Pearson ([54]) est la métrique la plus fréquemment utilisée par les recherches dans ce domaine. C'est le cas aussi des différents DS utilisés dans ce travail. Bien qu'elle ne soit pas citée et utilisée dans la majorité des travaux connexes, nous avons choisi d'inclure conjointement au coefficient de Pearson, l'erreur quadratique moyenne (Root Mean Squared Error (RMSE)[55]) pour quantifier la différence (ou le décalage) entre le résultat(score) obtenu par le système et celui obtenu par l'expert humain.

**1) Coefficient de Pearson (CP)**

En statistiques, étudier la corrélation entre deux ou plusieurs variables statistiques numériques, c'est étudier l'intensité de la liaison ("proportionnalité") qui peut exister entre ces

variables. La mesure de la corrélation linéaire entre les deux se fait alors par le calcul du coefficient de corrélation linéaire, noté CP. Ce coefficient est égal au rapport de leur covariance et du produit non nul de leurs écarts types. Le coefficient de corrélation est compris entre -1 et 1 :

**Tableau 4.5. Signification des valeurs de corrélation de pearson**

| Corrélation | Négative | Positive |
|---|---|---|
| **Faible** | de -0,5 à 0,0 | de 0,0 à 0,5 |
| **Forte** | de −1,0 à −0,5 | de 0,5 à 1,0 |

Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte ; nous employons simplement l'expression « fortement corrélées » pour qualifier les deux variables. Une corrélation égale à 0 signifie que les variables ne sont pas corrélées linéairement. Le coefficient de corrélation est multiplié par 100 pour exprimer un pourcentage de corrélation. Dans notre cas les variables statistiques à considérer sont celles définies dans deux vecteurs l'un contenant les valeurs de scores entre les couples de réponses du DS (réponse de l'étudiant, réponse modèle de l'enseignant) calculés automatiquement, le deuxième vecteur contient les scores, pour les mêmes couples de réponses, calculées par l'expert humain. L'objectif dans notre travail revient à maximiser ce coefficient.

2) **Erreur quadratique RMSE (Root Mean Squared Error (RMSE)**[55])

L'erreur quadratique moyenne permet de quantifier une mesure synthétique de l'erreur globale commise. Pour calculer l'erreur quadratique moyenne RMSE, les erreurs individuelles sont tout d'abord élevées au carré, puis additionnées les unes aux autres. Nous divisons ensuite le résultat obtenu par le nombre total d'erreurs individuelles, puis nous en prenons la racine carrée.

L'erreur quadratique est probablement le critère quantitatif le plus utilisé pour comparer valeurs calculées (ici les scores ou notes automatiques) et valeurs observées (scores manuels attribués par l'expert humain. C'est cette fonction que nous tentons de minimiser dans le cadre de ce travail.

*En conclusion, l'évaluation de nos approches correspond à trouver la meilleure minimisation de l'erreur quadratique avec une maximisation du coefficient de corrélation.*

## 4.3 Résultats et discussion

Dans la suite, nous présentons les résultats obtenus dans plusieurs perspectives :

- Dimensionnalité de l'espace sémantique et impact du stemming,
- Les résultats des deux modèles de similarité proposés et de leurs variantes (en termes de pondération),
- Hybridation des deux modèles,
- Hybridation avec les mesures syntaxiques et avec celles développées pour les WE,
- Evaluation par rapport aux résultats obtenus par les travaux connexes.

### 4.3.1 Dimensionnalité de l'espace sémantique et impact du stemming

Généralement, les vecteurs de mots avec plus de quelques centaines de dimensions sont peu pratiques. En mathématiques, la méthode algébrique de décomposition en valeur singulière (Singular Value Decomposition (SVD) [56]) est un outil important pour factoriser des matrices rectangulaires complexes afin de réduire la taille. Cependant, cette approche est très coûteuse en termes de consommation de mémoire et peut être impraticable en particulier pour les grands corpus, où les tailles d'espace initiales peuvent être importantes. Idéalement, pour notre système proposé, l'algorithme SVD serait calculé à l'aide des vecteurs matriciels complets de l'espace sémantique construit. Cependant, ceci est difficile en termes de calcul et est inutile. De bons résultats peuvent être obtenus en utilisant plusieurs milliers de mots parmi les plus fréquents. La dimensionnalité de l'espace sémantique peut être réduite en augmentant la limite des mots peu fréquents après le retrait de Stopwords. Nous utilisons également le stemming lourd pour minimiser les classes de mots et ensuite la dimensionnalité.

Dans le « Tableau 4.6. », nous explorons 3 espaces sémantiques construits avec des dimensions allant jusqu'à 24230. En augmentant la dimensionnalité, le système de base calculé avec le modèle de la somme vectorielle sans pondération en utilisant Gomaa DS, présente le meilleur résultat avec la dimension 16752. Au-dessus, il y a un petit changement de performance. Les performances diminuent lentement à mesure que nous réduisons les vecteurs. Cela confirme bien que, dans la pratique, des performances à peu près équivalentes sont obtenues en utilisant la dimensionnalité de 14 000 à 100 000 [15]. Cette constatation présente un double avantage pour le système proposé. Tout d'abord, encourager l'utilisation de corpus de domaines spécifiques car, autour d'une certaine dimension, les résultats sont comparable. Il y a juste pour construire le domaine de corpus spécifique correspondant au cours ou aux connaissances à évaluer en utilisant le système ASAGS. Deuxièmement, il est plus facile de construire (ou de trouver) un corpus (pas nécessairement de taille gigantesque) et de ne pas avoir besoin de beaucoup de ressources machine pour l'implémentation du système ASAG.

De plus, les temps de génération des espaces sémantiques comme nous pouvons le voir dans le « Tableau 4.7.» sont assez raisonnables comparés à des valeurs beaucoup plus importants dans des approches de machines learning surtout si nous rappelons que l'espace sémantique n'est généré qu'une seule fois.

Pour les mêmes corpus nous avons généré les espaces sémantiques correspondants en appliquant un stemming léger. Nous faisons deux constatations essentielles :

- L'impact de dimensionnalité est bien confirmé puisque les résultats pour des dimensions entre 18630 et 28062 sont restés comparables en corrélation (CP) et en RMSE. Le meilleur résultat (CP=**70.58%**) étant obtenu pour une dimension 18630.
- L'impact de stemming lourd est bien visible puisqu'il ya une baisse de pratiquement 6% en corrélation en utilisant le stemming léger. Ceci peut être expliqué que travailler sur les racines de mots permet mieux dans une approche statistique d'accentuer le calcul de cooccurrences sur la racine d'où un meilleur résultat quand les cooccurrences sont calculées à partir des racines de mots.

Etant comparables pour les espaces sémantiques déjà construits, tous les résultats rapportés dans ce qui va suivre sont calculés en utilisant un stemming lourd.

**Tableau 4.6. Dimensionnalité des espaces sémantiques générés (modèle SV sans pondération)**

| | Stemming Lourd (Khodja) | | | Stemming Léger (Tashaphine) | | |
|---|---|---|---|---|---|---|
| **Corpus** | **Dimension des vecteurs** | **CP(%)** | **RMSE** | **Dimension des vecteurs** | **CP(%)** | **RMSE** |
| **Khaleej** | 14908 | 75.13 | 1.13 | **18630** | **70.58** | **1.28** |
| **CNN** | **16752** | **76.07** | **1.11** | 21032 | 69.63 | 1.29 |
| **BBC+CNN** | 24230 | 75.64 | 1.11 | 28062 | 69.81 | 1.26 |

**Tableau 4.7. Temps de génération des espaces sémantiques**

| Corpus | Temps d'exécution |
|---|---|
| **CNN** | 1h 19 min 6s |
| **BBC+CNN** | 3h 28 min 45s |
| **Khaleej** | 1h 19 min 29s |

### 4.3.2 Les résultats des deux modèles de similarité proposés et de leurs variantes

***1) Les résultats du modèle CM :***

Dans les tableaux suivants (« Tableau 4.8. », « Tableau 4.9. », « Tableau 4.10. »), nous présentons les résultats obtenus pour le modèle CM sur les 3 espaces sémantiques.

Nous constatons d'un coté que même si les meilleurs résultats sont obtenus pour l'espace CNN, les résultats pour les autres restent très proches et comparables. Pour les trois corpus, la similarité d'ordre (représentée ici par WF) n'a pas une importance significative dans notre travail puisque les meilleurs résultats sont obtenus avec wf=0.0, wf=0.1 ou Wf=0.2.

**Tableau 4.8. Résultats du modèle CM avec l'espace sémantique CNN**

| | | Espace sémantique Cnn | | | | | |
|---|---|---|---|---|---|---|---|
| **DS** | **WF** | **0.0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
| **GOMAA** | **RMSE** | 1.21 | **1.15** | 1.21 | 1.25 | 1.35 | 1.37 |
| | **CP (%)** | 75.99 | **76.50** | 75.22 | 74.42 | 73.45 | 71.97 |
| **STS 250** | **RMSE** | **1.19** | 1.19 | 1.17 | 1.20 | 1.21 | 1.21 |
| | **CP (%)** | **70.83** | 70.73 | 70.81 | 69.73 | 68.99 | 68.43 |
| **STS 368** | **RMSE** | **0.99** | 0.99 | 0.99 | 1.00 | 1.00 | 1.02 |
| | **CP (%)** | **76.91** | 76.72 | 76.40 | 75.96 | 75.66 | 74.97 |

**Tableau 4.9. Résultats du modèle CM avec l'espace sémantique BBC+CNN**

| | | Espace sémantique BBC+CNN | | | | | |
|---|---|---|---|---|---|---|---|
| **DS** | **WF** | **0.0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
| **GOMAA** | **RMSE** | 1.21 | **1.17** | 1.21 | 1.30 | 1.35 | 1.34 |
| | **CP (%)** | 75.90 | **75.92** | 75.05 | 74.65 | 73.31 | 72.28 |
| **STS 250** | **RMSE** | 1.20 | 1.19 | **1.17** | 1.20 | 1.20 | 1.20 |
| | **CP (%)** | 70.36 | 70.34 | **70.70** | 69.73 | 68.99 | 68.78 |
| **STS 368** | **RMSE** | **1.00** | 1.00 | 1.00 | 1.01 | 1.01 | 1.02 |
| | **CP (%)** | **76.50** | 76.31 | 76.03 | 75.59 | 75.25 | 74.72 |

**Tableau 4.10. Résultats du modèle CM avec l'espace sémantique Khaleej**

| | | Espace sémantique Khaleej | | | | | |
|---|---|---|---|---|---|---|---|
| **DS** | **WF** | **0.0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
| **GOMAA** | **RMSE** | 1.20 | **1.17** | 1.22 | 1.26 | 1.35 | 1.37 |
| | **CP (%)** | **76.17** | 76.09 | 75.39 | 74.38 | 73.52 | 72.30 |
| **STS 250** | **RMSE** | 1.20 | 1.19 | **1.18** | 1.20 | 1.21 | 1.24 |
| | **CP (%)** | **70.87** | 70.77 | 70.14 | 69.82 | 68.77 | 68.11 |
| **STS 368** | **RMSE** | **1.00** | 1.00 | 1.00 | 1.01 | 1.01 | 1.03 |
| | **CP (%)** | **76.52** | 76.24 | 76.04 | 75.67 | 75.51 | 74.67 |

### 2) *Les résultats du* modèle *SV :*

Dans les tableaux suivants (« Tableau 4.11. », « Tableau 4.12. » « Tableau 4.13. ») nous présentons les résultats obtenus pour le modèle SV avec et sans pondération de fréquences sur les 3 espaces sémantiques. Nous constatons d'un coté que même si les meilleurs résultats sont obtenus pour l'espace CNN, les résultats pour les autres restent très proches et comparables.

Nous constatons essentiellement que la pondération TFmin-max a nettement amélioré les résultats pour les 3 DS. En considérant par exemple l'espace sémantique CNN, la pondération a amélioré la corrélation de Pearson de +2,27, +3,85, +10,95 pour les 3 DS respectivement. D'où l'importance de considérer l'importance des mots (exprimée ici par les poids TFs) dans le corpus.

**Tableau 4.11. Résultats du modèle SV avec l'espace sémantique CNN**

| | | Sans pondération | Avec TF pondération |
|---|---|---|---|
| **GOMAA** | **RMSE** | 1.11 | **1.04** |
| | **CP (%)** | 76.07 | **78.33** |
| **STS 250** | **RMSE** | 1.31 | 1.25 |
| | **CP (%)** | 63.03 | 66.88 |
| **STS 368** | **RMSE** | 1.33 | 1.12 |
| | **CP (%)** | 62.86 | 73.77 |

**Tableau 4.12. Résultats du modèle SV avec corpus BBC+CNN**

| | | Sans pondération | Avec TF pondération |
|---|---|---|---|
| **GOMAA** | **RMSE** | 1.11 | 1.03 |
| | **CP (%)** | 75.64 | 77.97 |
| **STS 250** | **RMSE** | 1.32 | **1.19** |
| | **CP (%)** | 62.81 | **68.01** |
| **STS 368** | **RMSE** | 1.34 | 1.17 |
| | **CP (%)** | 62.72 | 73.18 |

**Tableau 4.13. Résultats du modèle SV avec l'espace sémantique Khaleej**

| | | Sans pondération | Avec TF pondération |
|---|---|---|---|
| GOMAA | RMSE | 1.13 | 1.05 |
| | CP (%) | 75.13 | 77.50 |
| STS 250 | RMSE | 1.32 | 1.23 |
| | CP (%) | 62.10 | 66.43 |
| STS 368 | RMSE | 1.33 | **1.06** |
| | CP (%) | 63.55 | **74.77** |

### 3) Combinaison des modèles (CM et SV) :

Comme le montre le « Tableau 4.14. », l'hybridation des deux modèles de similarité a permis d'améliorer nettement les résultats obtenus séparément et ceci pour les 3 DS. Remarquons aussi que la combinaison par la moyenne donne un meilleur résultat dans les différents cas.

**Tableau 4.14. Résultats de la combinaison des modèle s CM et SV**

| | | CP (%) | RMSE |
|---|---|---|---|
| GOMAA | Max | 77.44 | 1.05 |
| | Moy | **78.62** | **1.09** |
| STS 250 | Max | 68.52 | 1.23 |
| | Moy | **70.46** | **1.19** |
| STS 368 | Max | 75.08 | 1.07 |
| | Moy | **77.17** | **0.98** |

## 4.3.3 Hybridation avec les mesures syntaxiques et avec celles des WE

Les résultats obtenus par les deux modèles de similarité proposés ont été combinées avec ceux développés dans deux autres approches dans le même projet : les modèles syntaxiques et ceux utilisant les WE.

*1) Combinaison avec les mesures syntaxiques :*

Le modèle de calcul matriciel a été développé pour une mesure syntaxique où la similarité entre deux mots est calculée dans une matrice syntaxique en appliquant l'algorithme LCS ([57][58]). Une combinaison des deux matrices CM sémantique et CM syntaxique a été faite ainsi qu'avec la mesure syntaxique DICE. Les résultats sont reportés dans le « Tableau 4.15. ».

**Tableau 4.15. Résultats de la combinaison du modèle CM avec les mesures syntaxiques**

| | | WF=0.1 (CM sémantique, CM Syntaxique) | WF=0.0 (CM sémantique, DICE) |
|---|---|---|---|
| GOMAA | RMSE | 1.15 | 1.09 |
| | CP (%) | 76.61 | 80,78 |
| STS 250 | RMSE | 1.18 | 1.12 |
| | CP (%) | 71.23 | 71,55 |

La combinaison du modèle SV avec la mesure syntaxique a donné les meilleurs résultats qui sont reportés dans le « Tableau 4.16. » en pondérant avec les TFmin-max.

**Tableau 4.16. Résultats de la combinaison des modèles SV et Dice syntaxique**

| | | Sans Pondération | TF Pondération |
|---|---|---|---|
| GOMAA | RMSE | 1.03 | 0.98 |
| | CP (%) | 80,61 | 81,49 |
| STS 250 | RMSE | 1.16 | 1.11 |
| | CP (%) | 69,42 | 70.87 |

Dans les deux cas, une amélioration importante de la corrélation et de l'erreur quadratique est enregistrée en combinant avec les mesures syntaxiques particulièrement la

mesure DICE qui prend en considération le nombre de mots communs entre les deux réponses à comparer.

2) ***Combinaison avec le modèle WE :***

Nous avons combiné nos résultats (des modèles CM et SV) avec ceux obtenus pour les modèles utilisant les WE (les modèles CBOW et SkipGram). Plusieurs combinaisons ont été effectuées en utilisant les 3 espaces sémantiques. Nous ne présentons dans le « Tableau 4.17. » que celles ayant apporté une amélioration significative des résultats obtenus déjà par notre approche.

**Tableau 4.17. Résultats des combinaison WE**

| Modèle | DS (CP // RMSE) | |
|---|---|---|
| **Combinaison WE** | **GOMAA** | **STS-250** |
| CM COMB (0.5) | 75.74% // 1.22 | **72.20% // 1.13** |
| CM Sémantique + Best WE | 77.84% // 1.12 | 71.83% // 1.09 |
| SV+Best WE | **79.46% // 1.03** | 70.41% // 1.14 |
| (CM Sémantique+SV)+Best WE | 78.83% // 1.08 | 71.27% // 1.13 |

### 4.3.4 Récapitulation des résultats et discussion

Le « Tableau 4.18. » résume les meilleurs résultats obtenus pour notre approche et les différentes combinaisons :

**Tableau 4.18. Récapitulatif des résultats**

| Meilleurs résultats | | | |
|---|---|---|---|
| **Modèle** | **DS ( CP // RMSE )** | | |
| | **GOMAA** | **STS 250** | **STS 368** |
| ***Combine SEM*** | | | |
| CM SEM | 76.50% // 1.15 | 70.87% // 1.20 | 76.91% // 0.99 |
| SV | 78.33% // 1.04 | 68.01% // 1.21 | 74.77% // 1.16 |
| SV+CM-SEM (moy ) | 78.62% // 1.09 | 70.46% // 1.19 | 77.17% // 0.98 |
| SV+CM-SEM (max ) | 77.44% // 1.05 | 68.52% // 1.23 | 75.08% // 1.07 |
| ***Combinaison Syntaxique*** | | | |
| CM COMB (0.5) | 76.61% // 1.15 | 71.23% // 1.18 | **77.86% // 0.98** |

| | | | |
|---|---|---|---|
| CM-SEM +Dice | 80.78% // 1.09 | 71.69% // 1.12 | - |
| SV+Dice | **81,49% // 0.98** | 70.87% // 1.11 | - |
| SV+CM+Dice | 80.11% // 1.11 | 71.24% // 1.13 | |
| ***Combinaison WE*** | | | |
| CM COMB (0.5) | 75.74% // 1.22 | **72.20% // 1.13** | - |
| CM-SEM + BestWE | 77.84% // 1.12 | 71.83% // 1.09 | - |
| SV+BestWE | 79.46% // 1.03 | 70.41% // 1.14 | - |
| COMB+BestWE | 78.83% // 1.08 | 71.27% // 1.13 | - |

**1) Evaluation par rapport aux résultats obtenus par les travaux connexes.**

Dans le « Tableau 4.19. » nous comparons les résultats obtenus par rapport à ceux des travaux connexes sur Gomaa DS.

**Tableau 4.19. Evaluation par rapport aux travaux connexes sur Gomaa DS**

| | CP (%) | RMSE |
|---|---|---|
| **IAA (Arabic and English DS ) Manual scores** | 86.00 | 0.69 |
| **Notre approche combinée** | 81,49 | 0.98 |
| **(Gomaa's system 2014) Arabic DS** [37] | 73.00 | 1.07 |
| **(Gomaa'system 2014) Manual English translated DS** [37] | 83.00 | 0.75 |
| **(Zahran WE System 2015) (Arabic )** [30] | 82.00 | 0.95 |
| **(Vectorized – Arabic System WE 2016 )** [41] | 84.00 | 0.89 |

Comparés aux résultats rapportés dans [37], [30] et [41] sur le Gomaa DS, les résultats obtenus par le système final proposé (avec un modèle de similarité combiné) sont très intéressants. Le système proposé surpasse largement le système Gomaa pour le DS arabe natif (Pearson + 8.49%, RMSE +0.1), ce qui est important pour le système proposé. La traduction anglaise de l'ensemble de données a permis au système Gomaa d'améliorer ses performances en utilisant WordNet pour combiner la similarité fondée sur les connaissances. Le système proposé a enregistré presque le même résultat que ceux de [30] et se rapproche étroitement (-2,5 de moins que) des résultats de [41] qui ont atteint 84% en utilisant une combinaison importante de représentations vectorielles multidimensionnelles avec une multitude de mesures syntaxiques basées sur la connaissance. Les résultats obtenus par le système proposé

sont très encourageants, d'autant plus que nous n'avons pas introduit dans le calcul de similarité une combinaison de diverses mesures syntaxiques et / ou sémantiques sophistiquées existant dans la littérature. Nous visons un système simple, efficace et opérationnel.

Pour évaluer la capacité du système proposé à se généraliser, nous l'évaluons par rapport aux travaux traitant de la similarité des paires de textes courts arabe-arabe du SEMEval 2017. Le « Tableau 4.20. » présente les résultats obtenus sur le DS SemEval générique. Le système proposé a obtenu 11,75% de plus que le système de base SEMEVAL, mais il a atteint -2,43% de moins que le LIM-LIG [25], le deuxième score qui utilise une approche vectorisée (similaire à la nôtre). Considérant ces résultats comme acceptables, nous pouvons en déduire que le système proposé peut bien se généraliser.

**Tableau 4.20. Evaluation sur le DS SEMEval 2017**

| | CP (%) | RMSE |
|---|---|---|
| **SEMEVAL 2017 baseline** [40] | *60.45* | - |
| **SEMEval 2017 2nd score Track 1 LIM-LIG** [25] | 74.63 | - |
| **Système proposé (Arabic)** | ***72.20*** | ***1.03*** |

**2) Discussion**

Au terme de cette synthèse expérimentale, nous aboutissons à plusieurs constatations que nous discutons à travers les points suivants :

1) Une amélioration des résultats est bien marquée avec les combinaisons (interne ou externe). Ce qui confirme bien que les modèles hybrides donnent de meilleurs résultats (Le best de nos résultats est obtenue avec les combinaisons externes).
2) Le meilleur résultat avec le DS de Gomaa est obtenu avec la combinaison du modèle SV et la méthode Dice (81.49%). Ceci est prévisible puisque il y a plus de chance dans ce DS de trouver plusieurs mots en commun dans la réponse de l'étudiant et celle de la réponse modèle.

3) Le meilleur résultat avec le data-set de STS 250 est obtenu avec la combinaison du modèle CM sémantique et CM WE (72.20%). Ce qui permet de penser que la généralisation de l'approche pour une autre tâche est faisable très correctement.
4) L'aspect d'ordre n'a pas d'importance dans notre approche car les meilleurs résultats ont été obtenus avec un WF qui ne dépasse pas 0.2.
5) Les pondérations ont améliorée nettement les résultats avec les trois DS dans le modèle SV et toutes les combinaisons. Tandis que ces pondérations n'ont aucun impact dans le modèle CM.
6) Le DS de Gomaa est le plus indicatif pour notre approche. Il a été crée dans le contexte d'une évaluation automatique des réponses courtes (l'objet même de notre thématique). Les datasets du SemEval ont été considérés à titre indicatif pour mesurer le degré de généralisation de l'approche pour d'autres domaines.

# Conclusion et perspectives

L'approche statistique est la solution la plus appropriée pour les langues qui manque de ressources linguistique. L'évaluation automatique est un sujet d'actualité qui n'est pas vraiment facile mais il a beaucoup de bénéfice pour l'avancement de la technologie. Pour cela nous avons axé notre recherche sur la création de l'espace sémantique dans le contexte de la langue arabe.

Notre thème est très vaste et lié avec de nombreux domaines, en conséquence nous avons présenté dans notre mémoire, un état de l'art sur les systèmes d'évaluation, le traitement de la langue arabe, les approches de similarité, ainsi que les enjeux de la langue arabe dans le contexte de l'évaluation automatique.

Par la suite, nous avons proposé notre approche après une large recherche et comparaison des approches existantes, en résultats un système d'évaluation automatique a été développé qui se compose de plusieurs modules. Le système est combiné avec d'autres approches syntaxiques et Word Embeddings.

Nous avons évalué notre système avec trois dateset, et nous avons obtenue de meilleurs résultats par rapport aux systèmes des travaux connexes sur le dataset de Gomaa arabe natif. Pour les datasets Sem-Eval, nos résultats sont comparables cela nous déduit que notre système peut être généralisable à d'autres tâches liées à la similarité de mots et de phrases. Sachant que les stemmers et les corpus sont les seuls ressources dépendantes de la langue utilisées dans notre approche, notre approche peut être appliquée sur diverses langues existantes notamment celles qui partagent les mêmes difficultés que l'arabe. Ceci exige, bien sure, une adaptation selon la langue en question : Disponibilité des ressources et outils (corpus et stemmer) ainsi que le prétraitement du corpus approprié à cette langue.

En perspectives, nous proposons de considérer entre autres les aspects suivants :

- Ajouter une étape de correction grammaticale automatique au début de l'étape du traitement du texte afin de corriger les erreurs de saisies et par la suite éviter les ambiguïtés.

- Identifier les entités nommées[1] ainsi que les collocations[2] et les types de mots (verbe, adjectif, nom…).
- Prendre en considération le rôle important des diacritiques[3] dans l'arabe car deux mots peuvent avoir la même morphologie mais pas le même sens en introduisant les diacritiques. Pa exemple : جِمَال et جَمَال , le premier mot signifie « chameaux » tandis que l'autre signifie « beauté ».
- Le système est valide seulement pour une seule RM c'est-à-dire qu'il ne supporte pas une variété de RM. Nous supposons que l'ajout de cette fonctionnalité au système donnera encore mieux.
- Elaborer plus de justification à l'utilisateur/étudiant en retournant son score. Ceci peut être établit à l'aide des commentaires à propos de ses fautes et en retournant la RM.

[1] Se sont les noms propres par exemple : noms des pays, ville, personne…etc.

[2] Considérer les mots constituant de deux mots attachés comme un seul token. Par exemple : محامي المجلس (avocat du conseil) qui représente un seul token et محامي (avocat) sont deux termes différents. Avocat signifie un métier de manière générale tandis que l'autre signifie le même métier aussi mais plus spécifique (troponymie).

[3] Voyelles en français ou التشكيل en arabe.

# Bibliographie


[1] M. H. Abu Mugasib et R. S. Baraka, « An Ontology-Based Automated Scoring System for Short Answer Questions », 2015.

[2] S. Burrows, I. Gurevych, et B. Stein, *The eras and trends of automatic short answer grading*, vol. 25, n° 1. 2015.

[3] D. Callear, J. Jerrams-Smith, V. Soh, D. J. Jerrams-smith, et H. P. Ae, « CAA of Short Non-MCQ Answers », in *In Proceedings of the 5th International CAA conference*, 2001.

[4] C. Leacock et M. Chodorow, « C-rater: Automated Scoring of Short-Answer Questions », *Comput. Hum.*, vol. 37, n° 4, p. 389‑405, 2003.

[5] S. Jordan, « Investigating the Use of Short Free Text Questions in Online Assessment », *Final Proj. report, Cent. Open Learn. Math. Sci. Comput. Technol. Open Univ. Milt. Keynes, United Kingdom*, 2009.

[6] J. Sukkarieh et J. Blackmore, « c-rater: Automatic Content Scoring for Short Constructed Responses. », *FLAIRS Conf.*, p. 290‑295, 2009.

[7] J. Z. Sukkarieh et S. Stoyanchev, « Automating Model Building in C-rater », in *Proceedings of the 2009 Workshop on Applied Textual Inference*, p. 61‑69, 2009.

[8] L. Cutrone, M. Chang, et Kinshuk, « Auto-Assessor: Computerized Assessment System for Marking Student's Short-Answers Automatically », in *2011 IEEE International Conference on Technology for Education*, p. 81‑88, 2011.

[9] T. Pedersen, S. Patwardhan, et J. Michelizzi, « WordNet::Similarity: Measuring the Relatedness of Concepts », in *Demonstration Papers at HLT-NAACL 2004*, p. 38‑41, 2004.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten, « The WEKA Data Mining Software: An Update », *SIGKDD Explor. Newsl.*, vol. 11, n° 1, p. 10‑18, 2009.

[11] N. Madnani, J. Burstein, J. Sabatini, et T. O'Reilly, « Automated scoring of a summary writing task designed to measure reading comprehension », *Naacl/Hlt 2013*, p. 163, 2013.

[12] E. Negre, « Comparaison de textes: quelques approches... », 2013.

[13] « Indice et distance de Jaccard », 2018. [En ligne]. Disponible sur: https://fr.wikipedia.org/wiki/Indice_et_distance_de_Jaccard. [Consulté le: 17-juin-2018].

[14] R. A. Baeza-Yates et B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[15] D. L. T. Rohde, L. M. Gonnerman, et D. C. Plaut, « An Improved Method for Deriving Word Meaning from Lexical », *Cogn. Psychol.*, vol. 7, p. 573‑605, 2004.

[16] W. H.Gomaa et A. A. Fahmy, « A Survey of Text Similarity Approaches », *Int. J. Comput. Appl.*, vol. 68, n° 13, p. 13‑18, 2013.

[17] « WordNet — Wikipédia », 2018. [En ligne]. Disponible sur: https://fr.wikipedia.org/wiki/WordNet. [Consulté le: 17-juin-2018].

[18] M. El-Haj, U. Kruschwitz, et C. Fox, « Creating language resources for under-resourced languages: methodologies, and experiments with Arabic », *Lang. Resour. Eval.*, vol. 49, n° 3, p. 549‑580, 2015.

[19] « Corpus — Wikipédia », 2018. [En ligne]. Disponible sur: https://fr.wikipedia.org/wiki/Corpus. [Consulté le: 17-juin-2018].

[20] W. Zaghouani, « Critical Survey of the Freely Available Arabic Corpora », *Proc. Work. Free. Arab. Corpora Corpora Process. Tools Work. Program.*, p. 1‑8, 2017.

[21] J. B. Lovins, « Development of a stemming algorithm », *Mech. Transl. Comput. Linguist.*, vol. 11, n° June, p. 22‑31, 1968.

[22] « Khawas - Browse Files at SourceForge.net ». [En ligne]. Disponible sur: https://sourceforge.net/projects/kacst-acptool/files/. [Consulté le: 06-oct-2018].

[23] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, et D. McClosky, « The

Stanford CoreNLP Natural Language Processing Toolkit », *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr.*, p. 55‑60, 2014.

[24] « Stanford CoreNLP – Natural language software | Stanford CoreNLP ». [En ligne]. Disponible sur: https://stanfordnlp.github.io/CoreNLP/. [Consulté le: 06-oct-2018].

[25] D. S. El Moatez Billah Nagoudi, Jérémy Ferrero, « LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting », in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, nº June, p. 134‑138, 2017.

[26] B. Furlan, V. Batanović, et B. Nikolić, « Semantic similarity of short text in languages with a deficient natural language processing support », *Decis. Support Syst.*, vol. 55, nº 3, p. 710‑719, 2013.

[27] W. H. Gomaa et A. A. Fahmy, « SimAll : A flexible tool for text similarity SimAll : A flexible tool for text similarity », nº December, 2017.

[28] R. Mihalcea, C. Corley, et C. Strapparava, « Corpus-based and knowledge-based measures of text semantic similarity », *Proc. 21st Natl. Conf. Artif. Intell.*, vol. 1, p. 775‑780, 2006.

[29] A. Islam et D. Inkpen, « Semantic text similarity using corpus-based word similarity and string similarity », *ACM Trans. Knowl. Discov. Data*, vol. 2, nº 2, p. 1‑25, 2008.

[30] M. Zahran, A. Magooda, A. Mahgoub, H. Raafat, M. Rashwan, et A. Atyia, « Word Representations in Vector Space and their Applications for Arabic ». 2015.

[31] L. Ouahrani, « String similarity for Arabic short answer grading », *Intern. report, LIMPAF/118, LIMPAF Lab. Bouira Univ.*, 2018.

[32] M. M. A. Alqahtani et E. Atwell, « A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus », *Corpus Linguist. 2015*, nº July, p. 7‑9, 2015.

[33] K. Shaalan, M. Attia, P. Pecina, Y. Samih, et J. van Genabith, « Arabic Word Generation and Modelling for Spell Checking », *Proc. Eight Int. Conf. Lang. Resour. Eval.*, p. 719‑725, 2012.

[34] S. I. Hajeer, « Comparison on the Effectiveness of Different Statistical Similarity Measures », vol. 53, nº 8, p. 14-19, 2012.

[35] H. Khafajeh *et al.*, « Automatic Query Expansion for Arabic Text Retrieval Based on Association and », *Inf. Retr. Boston.*, nº October, 2010.

[36] A. O. Al-Thubaity, « A 700M+ Arabic corpus: KACST Arabic corpus design and construction », *Lang. Resour. Eval.*, vol. 49, nº 3, p. 721-751, 2015.

[37] W. H. Gomaa et A. A. Fahmy, « Arabic Short Answer Scoring with Effective Feedback for Students », *Int. J. Comput. Appl.*, vol. 86, nº 2, p. 35-41, 2014.

[38] W. H. Gomaa et A. A. Fahmy, « Automatic scoring for answers to Arabic test questions », *Comput. Speech Lang.*, vol. 28, nº 4, p. 833-857, 2013.

[39] P. Kolb, « Disco: A multilingual database of distributionally similar words », *Proc. KONVENS-2008, Berlin*, nº 2003, p. 37-44, 2008.

[40] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, et L. Specia, « SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation », *Proc. 11th Int. Work. Semant. Eval. (SemEval-2017), pages 1–14, Vancouver, Canada, August 3 - 4*, p. 1-14, 2017.

[41] A. Magooda, M. A. Zahran, M. Rashwan, H. Raafat, et M. B. Fayek, « Vector Based Techniques for Short Answer Grading », *Proc. Twenty-Ninth Int. Florida Artif. Intell. Res. Soc. Conf. Val. Spain*, p. 238-243, 2016.

[42] E. M. Billah Nagoudi et D. Schwab, « Semantic Similarity of Arabic Sentences with Word Embeddings », *Proc. Third Arab. Nat. Lang. Process. Work.*, p. 18-24, 2017.

[43] M. Saad et W. Ashour, « OSAC: Open Source Arabic Corpora », *6th Int. Conf. Electr. Comput. Syst. (EECS'10), Nov 25-26, 2010, Lefke, Cyprus.*, nº May 2014, p. 118-123, 2010.

[44] M. Abbas et K. Smaïli, « Comparison of Topic Identification methods for Arabic Language », in *International Conference on Recent Advances in Natural Language Processing - RANLP 2005*, nº 14-17, 2005.

[45] M. Abbas, « Corpus arabe - Mourad Abbas ». [En ligne]. Disponible sur: https://sites.google.com/site/mouradabbas9/corpora. [Consulté le: 19-juin-2018].

[46] M. Abbas, K. Smaili, et D. Berkani, « Evaluation of topic identification methods on arabic corpora », *J. Digit. Inf. Manag.*, vol. 9, n° 5, p. 185‑192, 2011.

[47] S. Khoja et R. Garside, « Stemming Arabic text », 1999.

[48] T. Zerrouki, « Tashaphyne, Arabic light stemmer ». 2012.

[49] « SAFAR Web v2 ». [En ligne]. Disponible sur: http://arabic.emi.ac.ma:8080/SafarWeb_V2/faces/safar/morphology/stemmer.xhtml. [Consulté le: 21-juin-2018].

[50] C. Tambellini et C. Berrut, « Pondération des données incertaines dans les systèmes de recherche d ' informations : une première approche expérimentale . », p. 247‑261, 2006.

[51] H. M. HANNOUFI et N. A. HENNICHE, « Les Word-Embedding pour l'évalutaion automatique des réponses courtes en apprentissage en ligne : Application à la langue arabe. », Mémoire de master, Université Saad Dahleb Blida1, Juin 2018.

[52] J. Macqueen, « Some methods for classification and analysis of multivariate observations », *Proc. Fifth Berkeley Symp. Math. Stat. Probab.*, vol. 1, n° 233, p. 281‑297, 1967.

[53] S. R. Bowman, G. Angeli, C. Potts, et C. D. Manning, « A large annotated corpus for learning natural language inference », 2015.

[54] J. Cohen, *Statistical power analysis for the behavioral sciences (2nd ed.)*. 1988.

[55] W. Greene, « Econométrie, Paris, Pearson Education », *5e éd. (ISBN 978-2-7440-7097-6)*, 2005.

[56] M. Berry, Z. Drmac, et E. Jessup, « Matrices, Vector Spaces, and Information Retrieval », *SIAM Rev.*, vol. 41, n° 2, p. 335‑362, 1999.

[57] J. W. Hunt et M. D. MacIlroy, *An algorithm for differential file comparison*. Bell Laboratories, 1976.

[58] L. Allison et T. I. Dix, « A bit-string longest-common-subsequence algorithm », *Inf. Process. Lett.*, vol. 23, n° 5, p. 305‑310, 1986.

[59] « Définition : Question ouverte » Définitions marketing ». [En ligne]. Disponible sur: https://www.definitions-marketing.com/definition/question-ouverte/. [Consulté le: 15-sept-2018].

[60] P. Ryu et K. Choi, « Measuring the specificity of terms for automatic hierarchy construction », *Proc. ECAI Work. Ontol. Learn. Popul.*, 2004.

# Annexe

## A. Liste des stop-words :

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| هلّا | كان | امسى | عدة | س | سَاءَ | أضحى | ثمَّ | يوليو | والتي | لَسْنَ | بشكل |
| حاء | بدلا | مازال | جوان | ألفى | أولئك | اما | برس | خمسين | وقد | سنوات | اثنين |
| سنتيم | التى | لكنَّ | بعد | كأيّن | شمال | اربعون | كأنّما | الاولى | ماذا | تسعمئة | شين |
| كيت | كي | ثمَّ | هكذا | أبريل | مهما | خاء | اللتين | مثل | ولم | حيَّ | اي |
| جيم | صهْ | تحت | إلّا | أمامك | شخصا | أنّى | قد | طَق | هَؤلاء | راء | لدي |
| نا | أهلا | مِن | دولار | إليكم | أنتم | أبو | ظل | لَيْسَا | ذواتي | مايو | علم |
| نوفمبر | اعادة | هل | سبع | انبرى | نفسه | لايزال | عدَّ | لَسْتِ | كم | مقابل | ّأيّان |
| عيانا | شتانَ | سبعين | فاء | أوشك | ثمانين | ساءما | بسّ | لكنه | بل | ذ | اللذين |
| يومئذ | هَاتِي | ما | إذ | معاذ | عشر | لات | مليون | هَيْهَاتَ | أخ | عَمَّا | ك |
| إلي | ذَيْنِ | مليار | سيما | كلم | بهما | لَيْتَ | إحدى | عندما | م | خمس | حمو |
| لازال | عن | خمسمائة | لمّا | وهذا | د | هَذِي | صبرا | كيفما | ستكون | اذ | أل |
| ثان | أعطى | سنة | قاطبة | إليكَ | شيكل | ان | خامس | وهي | لن | أقل | إنَّ |
| بنا | هَاتانِ | تلكم | كاف | تسعمائة | صار | بان | فانه | همزة | كى | لعلَّ | حاليا |
| والذي | انقلب | هناك | بهم | صفر | ثلاثمئة | فى | عنه | يناير | لَسْتُ | الي | ماانفك |
| فيم | فلان | وليس | حول | فوق | أربعمئة | آهاً | ميم | ثاء | ثمانون | حيثما | لوما |
| أغسطس | إزاء | نبَّا | ترك | واضاف | ين | كانت | الاخيرة | إذا | كل | آناء | بات |
| لهم | أمد | الان | منذ | بمن | مذ | سين | هلمَّ | ة | بلى | وحتى | لام |
| عل | هذا | ومنذ | أحد | عنها | ساء | عام | عليك | علق | نَّ | بطآن | هَاتِه |
| الثانية | سرا | غير | هيت | ياء | ويكأنّ | نِعْمَ | مه | عشرة | فكان | فرادى | هَذَيْنِ |
| خال | فمنهم | ليت | اذا | طاء | فبراير | وجد | نحن | أولاء | الف | إمّا | باسم |
| مكانَك | واحد | ذينك | تم | ي | نفس | وعن | واوضح | وكانت | أقبل | غين | سبعون |
| أب | لم | بي | و | بك | فيفري | تلك | ب | من | منها | ث | الماضي |
| لَسْتُنَّ | أجمع | مايزال | ز | وا | بَلْهَ | ليرة | أفٍّ | مافتئ | ذلك | أفريل | رُبَّ |
| ذيت | احد | أصلا | اتخذ | لكن | نيف | أجل | إى | ها | هؤلاء | اليها | واهاً |
| خاصة | بما | تينك | أنَّ | أمس | ع | اكد | وفي | هن | خلال | برح | استحال |
| اثر | لو | سرعان | هَذا | ى | بكم | سوى | فيما | أوّهْ | صراحة | خلا | حزيران |
| جميع | أمام | لك | هَاتَيْنِ | مائة | لنا | أين | ستين | ذِي | ق | مكانكنّ | لعل |
| خمسون | تموز | عاشر | ثمنمئة | ستون | هي | فلس | نَخْ | تسعة | حوالى | كأيّ | إليه |
| ولا | سبعمئة | تعسا | عاد | بد | ايار | سَاءَمَا | ثم | حسب | ل | آهِ | أنَّ |
| ذواتا | بن | فمنذ | جدا | جويلية | لهما | اللواتي | اللاتي | لولا | يوم | قطّ | ريث |
| ثلاثون | خبَّر | فيها | يمكن | رابع | سبعة | إنّما | كيف | هنالك | كَأَنَّ | علي | حبيب |
| ايام | عدم | هيا | ظنَّ | اثنان | أن | هم | جعل | شباط | معه | ر | أربعة |
| إن | ابتدأ | الا | قلما | طفق | حار | اثني | أبّ | إيانا | وأبو | حادي | لكم |
| وهذه | الألاء | ض | حَذارِ | اخلولق | لئن | لبيك | رُبَّ | ء | علَّ | أصبح | حمدا |

| آنفا | أفعله | اجل | ذال | وأن | آض | زعم | بعض | لأن | عسى | االا | او |
|---|---|---|---|---|---|---|---|---|---|---|---|
| أيلول | عدد | الاول | وَيْ | خلف | ولكن | ضحوة | وهب | رويدك | ا?ى | ح | أيا |
| تانِ | وُشْكَانَ | ثالث | نِعِمّا | في | ظنَّ | لوكالة | اللذان | بهذا | ستة | منه | خ |
| إياه | حَبَّذَا | ديسمبر | لعلَّ | وتلك | نيسان | تلقاء | إياهن | إلَيْكَ | ذا | ارتدّ | جمعة |
| عين | ابين | بكما | أيّ | آمينَ | مكانكما | ذانك | آب | ضمن | دون | له | لقاء |
| ن | اللتي | ثمّة | اف | غادر | لَيْسُوا | واو | ألا | اليه | قوة | أنتِ | عليه |
| أيّان | كلتا | حمٌ | ذانِ | وكان | كلَّا | ثلاثين | إنَّ | دال | ثلاثمائة | خلافا | أيار |
| صباح | باء | كسا | ذِه | ظاء | وُشْكَانَ | لي | أفٍّ | دواليك | تَيْنِ | خمسمئة | لَسْتُم |
| لكنَّما | ريال | قاف | الذين | ثلاث | أمامكَ | ثمان | فمنها | هَذِه | كأن | درهم | شَتَّانَ |
| ستمئة | حيَّ | إما | حم | آي | ظ | اخرى | جانفي | لكي | لاسيما | تخذ | يفعلون |
| بئس | اصبح | سبحان | ليس | ست | عما | إنَّ | حاشا | اضحى | تسعين | إياكن | كذا |
| أنت | اليوم | يورو | هَيْهات | فهو | بَسْ | جير | صدقا | كليهما | كلا | سبت | دينار |
| إليكما | لكنَّ | هلم | ج | ف | جلل | يونيو | كذلك | طرا | راح | كلّما | إياها |
| حَتَّى | وراءَك | سبتمبر | لهذا | وبين | الحالي | فإن | سادس | بخ | آه | سابع | أفٍّ |
| أنتن | كلها | هذه | صهٍ | كما | حرى | ثلاثاء | الثاني | كلَّا | وكذلك | تفعلان | ثمّ |
| ذو | لَيْسَ | انفك | أو | تعلَّم | اربعة | إلى | جنيه | مَنْ | غداة | خميس | عليها |
| غدا | أولئكم | ثلاثة | زاي | ذلكن | واضافت | ؤ | متى | مكانكم | عَدَسْ | اللذي | ص |
| تكون | هاؤم | أنا | اثنا | تلكما | هما | آل | بها | هيّا | تِي | حتى | و6 |
| ضاد | لها | التي | وقالت | تاء | مليم | صاد | سمعا | لَيْسَتْ | كثيرا | كن | وثي |
| بعدا | لدن | إي | اكثر | نهاية | حقا | إياهم | تِه | لكما | ومع | للامم | نون |
| قال | إذما | وله | أخو | ألف | فان | ظلّ | اللتيا | عند | بِئْسَ | ذلكما | هو |
| فقط | إياكما | أمسى | لَعَلَّ | غالبا | أ | أيضا | هَذانِ | فو | تانِك | رجع | الآن |
| بؤسا | تشرين | الى | لَيْسَتَا | مرّة | لقد | يوان | تحوّل | اول | بين | عشرون | أفعل |
| فضلا | إياك | كانون | ممن | ا | غ | لَسْتُمَا | عاما | أول | وعلى | ورد | أربع |
| تجاه | هنا | حبذا | يفعلان | أربعاء | فيه | انه | زيارة | إذاً | لعمر | كأين | حاي |
| أنتما | لَكِنَّ | اعلنت | أكتوبر | طالما | درى | إياي | ستمائة | يكون | كلاكما | حين | ذات |
| تفعلون | كرب | ومن | وهو | اربعين | ت | ثمانمئة | أخٍّ | ذلكم | اللائي | ايضا | الذى |
| يمين | ثاني | ثُمَّ | مساء | الوقت | أنَّ | هللة | هبّ | أولالك | آهٍ | اطار | حيث |
| بماذا | عدا | خمسة | كاد | ء | أم | نعم | ط | ثمانية | ذاك | إليكنّ | هاء |
| هاهنا | وان | مابرح | بغتة | االتى | ذوا | أينما | أخذ | ثماني | مع | امس | أى |
| الذاتي | بضع | دونك | إياهما | الذي | به | أما | لذلك | سوف | ولايزال | أكثر | لا |
| قرش | أربعمائة | تفعلين | تارة | لَسْتَ | مادام | شبه | طاق | الألى | شَتَّانَ | عوض | |
| إذن | بسبب | سحقا | لَسْنَا | كليكما | عجبا | بيد | مارس | غدًا | مئة | مما | |
| عشرين | حجا | هَجْ | فقد | ماي | سبعمائة | كِخ | لهن | تسع | وقف | ثامن | |
| بهن | السابق | على | يلي | أبدا | لدى | أفٍّ | إيهٍ | قبل | ه | ئ | |
| اللاواتي | انها | لكيلا | أَيُّ | أنه | علَّ | آ | امام | حدَث | بكن | ءَ | |
| مئتان | لما | أوت | نحو | بإن | كأنّ | تاسع | مئتين | عامة | تسعون | إياكم | |
| أنشأ | أي | المقبل | أمّا | آذار | ضد | هيهات | ش | كلاهما | اللتان | هاكَ | |

## B. Nombre des mots non-trouvés par khoja stem :

| | Corpus CNN | Corpus BBC+CNN | Corpus Khaleej |
|---|---|---|---|
| DS Gomaa | 78 | 67 | 72 |
| DS STS 250 | 34 | 29 | 37 |
| DS STS 368 | 42 | 37 | 45 |

## C. Liste des mots non-trouvés par Khoja stem :

| | Corpus cnn | Corpus bbc+cnn | Corpus khaleej |
|---|---|---|---|
| DS Gomaa | كاكلات', 'واحه', 'غضروفي', '{ ,"تتحل', 'بيئي', 'غضروفيه', 'يتممد ايكولوجي', 'كغذ', 'فيتجه', ' ,"فيكثر', 'ملاءمه', 'لتمتص', 'فيوفر كلورقيل', 'ونوفمبر', 'جرثم', ' 'منتظما', 'فيظل', 'سلالسل', 'فتظل', ميته', 'ضوءفالاشعه', ' ,"فيطفو 'فيحدث', 'فيزيايئه', 'ثعالب', 'تتعد', ماءالثدييات', 'للبيئه', ' ,"تتصف 'تستخلص', 'جرائيم', 'ازدباد', نيتروجيه', 'اوكسينات', ' ,"فسيولوجيه 'ييما', 'نادي', 'لغذاء', 'فينحني', قيقل', 'بيئه', ' ,"كعازل', 'وهكذا 'هاما', 'لتت', 'كلوروفيل', ,"حويصلات', 'للغطاء', 'يرابيع فيتوفر', 'جائرللحيوانات', 'بسسب', ' ,"صعر', 'فتزداد', 'بيئات', 'اغذائها تنعدم', 'بهياكل', 'فيزيد', 'يمتص', ' ,"نتيروجين', 'بلانكتون', 'فيتمدد تتغذي', 'موجهو', 'ماءيلعب', ' 'تراكيب', 'تناقلم', 'استضاءه', 'بيئيه', فتمتص', 'بلاستيدات', ' ,"حوصل "فتادي', 'فتحتل} | كاكلات', 'واحه', 'تتحل', '{ ,"بيئي', 'غضروفيه', 'يتممد ايكولوجي', 'كغذ', 'فيتجه', ' 'كطبقه', 'فيكثر', 'لتمتص', جرثم', ' ,"كلورقيل', 'ونوفمبر 'فيظل', 'سلالسل', 'فتظل', 'فيطفو', 'ميته', 'ضوءفالاشعه', فيزيايئه', 'فيوفر', ' ,"فيحدث 'تتصف', 'ماءالثدييات', 'للبيئه', ازدباد', ' ,"تستخلص', 'جرائيم 'فسيولوجيه', 'نيتروجيه', 'اوكسينات', 'ييما', 'نادي', فينحني', 'كعازل', ' ,"لغذاء 'وهكذا', 'قيقل', 'بيئه', 'هاما', حويصلات', ' ,"لتت', 'كلوروفيل 'للغطاء', 'يرابيع', 'فيتوفر', 'جائرللحيوانات', 'بسسب', فتزداد', 'بيئات', ' ,"صعر 'غذائها', 'فيزيد', 'يمتص', فيتمدد', ' ,"نتيروجين', 'بلانكتون 'تتغذي', 'موجهو', 'ماءيلعب', ,"استضاءه', 'بيئيه', 'فتمتص {'بلاستيدات', 'فتادي' | كاكلات', 'واحه', 'تتحل', '{ ,"بيئي', 'غضروفيه', 'يتممد لعناصرها', 'كغذ', 'فيتجه', ' 'كطبقه', 'فيكثر', 'لتمتص', ايكولوجيه', 'ونوفمبر', ' ,"كلورقيل 'جرثم', 'سلالسل', 'بكتريا', ,"طحالب', 'خنافس', 'فتظل فيطفو', 'ميته', 'ضوءفالاشعه', ' 'فيحدث', 'فيزيايئه', 'ثعالب', ماءالثدييات', 'للبيئه', ' ,"فيوفر 'تستخلص', 'جرائيم', 'ازدباد', نيتروجيه', 'ييما', ' ,"اوكسينات 'نادي', 'لغذاء', 'فينحني', 'كعازل', بيئه', 'كهف', ' ,"وهكذا', 'قيقل 'هاما', 'لتت', 'كلوروفيل', 'للغطاء', يرابيع', ' ,"جائرللحيوانات 'فوسفور', 'فيتوفر', 'بسسب', ,"صعر', 'فتزداد', 'بيئات', 'ذبل بهياكل', 'يمتص', 'بلانكتون', ' 'فيتمدد', 'تتغذي', 'موجهو', استضاءه', 'بيئيه', ' ,"ماءيلعب 'حوصل', 'فتمتص', 'بلاستيدات', "فتادي', 'فتحتل"} |
| DS STS 250 | قطاريتطلع', 'تسيرعلي', 'تنانير', '{ يقفزعلي', ' ,"تستمتع', 'قيلوله 'لكاميره', 'شتويا', 'ترتري', 'تماثيل', تنزهما', ' ,"زعر', 'الكاميرا', 'زخارف 'يهما', 'قيثاره', 'سراويل', 'قزح', تتصتع', ' ,"يراها', 'أحد', 'يقفزالي 'ياتي', 'قيثارته', 'غيرمطلي', 'شقبله للسباحهحالا', ' ,"ويجيلز', 'يغفو 'دايتونا', 'فمم', 'ترامبولين', 'غيتاره, | قطاريتطلع', 'تسيرعلي', '{ ,"تستمتع', 'قيلوله', 'يقفزعلي لكاميره', 'شتويا', 'ترتري', ' 'لكاميرا', 'زخارف', 'تنزهما', يراها', 'أحد', ' ,"يهما', 'قيثاره 'يقفزالي', 'تتصتع', 'ياتي', ,"قيثارته', 'غيرمطلي', 'شقبله ويجيلز', 'يغفو', 'للسباحهحالا', ' 'دايتونا', 'فمم', 'ترامبولين', | ذقن', 'قطاريتطلع', 'تسيرعلي', '{ ,"تنانير', 'قيلوله', 'يقفزعلي لكاميره', 'شتويا', 'ترتري', 'زعر', ' ,"لكاميرا', 'ترتدين', 'تنزهما', 'يهما قيثاره', 'سراويل', 'يراها', 'أحد', ' 'رنح', 'دغل', 'يقفزالي', 'رمث', ياتي', 'قيثارته', ' ,"تتصتع 'اصطحبت', 'غيرمطلي', 'شقبله للسباحهحالا', ' ,"ويجيلز', 'يغفو |

| | | | |
|---|---|---|---|
| | {"وقيثاره | {'وقيثاره' ,"غيتاره | 'دايتونا', 'فمم', 'ترامبولين', 'جوق', {'وقيثاره' ,"غيتاره |
| DS STS 368 | فلفل', 'بيتزا', 'ارنبا', 'تقشرحبه', ' جمبري', ' ,"ميكرفون', 'بوليود 'تشيلو', 'كزبر', 'طماطم', 'بمكياج', تبل', ' ,"ارنب', 'تورتيا', 'بروكولي 'خيارالي', 'ضفدع', 'زنجبيل', 'ماكياجا', 'تقشربرتقاله', 'ثنن', زنجبيلا', 'غوريلا', ' ,"صندوق 'سرخ', 'قيثار', 'اناء', 'قيثاره', بلاستيكيا', 'أحد', ' ,"تظيف', 'طماطما 'زعانف', 'صقف', 'جايم', 'مضادا', ,"ماكياج', 'باذنجان | فلفل', 'بيتزا', 'ارنبا', ' {"تقشرحبه', 'ميكرفون', 'بوليود تشيلو', 'طماطم', 'بمكياج', ' 'ارنب', 'تورتيا', 'بروكولي', زنجبيل', ' ,"تبل', 'خيارالي 'ماكياجا', 'تقشربرتقاله', 'صندوق', 'زنجبيلا', 'غوريلا', قيثار', 'اناء', 'قيثاره', ' ,"سرخ 'تظيف', 'طماطما', 'بلاستيكيا', جايم', 'مضادا', ' ,"أحد', 'صقف 'ماكياج', 'باذنجان', 'قيثارته', {'نقانق', "الريهانا', 'يرقصعلي | فلفل', 'بيتزا', 'ارنبا', 'تقشرحبه', ' {' ارنب', ' ,"بوليود', 'كزبر', 'صلص 'بمكياج', 'تورتيا', 'بروكولي', ,"تبل', 'خيارالي', 'ضفدع', 'غيتار زنجبيل', 'بروكلي', 'ماكياجا', ' 'تقشربرتقاله', 'واصطدم', غوريلا', ' ,"صندوق', 'زنجبيلا 'سرخ', 'ميكروفون', 'قيثار', ,"بفاس', 'قيثاره', 'تظيف', 'طماطما بلاستيكيا', 'أحد', 'زعانف', ' 'صقف', 'خنزير', 'جايم', 'مضادا', قيثارته', 'صنبور', ' ,"باذنجان 'الريهانا', 'يرقصعلي', 'نقانق', {"نشف', 'غيتارا |