**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE**
**UNIVERSITE SAAD DAHLEB – BLIDA 01**
**FACULTE DES SCIENCES**
**DEPARTEMENT D'INFORMATIQUE**

# MEMOIRE DE MASTER
## Spécialité : Ingénierie des Logiciels

## THEME

**VERS UN CORPUS PARALLELE MONOLINGUE POUR LA LANGUE ARABE A PARTIR DE CORPUS PARALLELES BILINGUES.**

**Présenté par :**
**Mlle. FEKNOUS Hind**
**Mlle. SELHANI Imène**

**Proposé & encadré par :**
**Mme OUAHRANI L.**

**Devant le jury composé de :**

| | |
|---|---|
| **M. HAMMOUDA** | **Président** |
| **M. BENAISSI** | **Examinateur** |

**Soutenu le : 03/10/2021**

# REMERCIEMENTS

Au terme de la rédaction de ce mémoire, c'est un devoir agréable d'exprimer en quelques lignes la reconnaissance que nous devons à tous ceux qui ont contribué de près ou de loin à l'élaboration de ce travail, qu'ils trouvent ici nos vifs respects et notre profonde gratitude.

En tout premier lieu, nous remercions le bon Dieu, tout puissant, de nous avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés.

Nous tenons à exprimer notre plus sincère gratitude au corps professoral et au personnel de l'Université Blida 1 pour la qualité d'enseignement qu'ils nous ont prodigué au cours de nos cinq années passées, nous remercions tout particulièrement notre promotrice Madame L. Ouahrani, qui a fait de grands efforts pour nous guider dans notre travail en nous aiguillant sur des pistes de réflexions riches et porteuses, à tout moment durant toute cette année.

Nous remercions également les membres du jury d'avoir pris le temps de juger ce travail.

Nos remerciements vont également à M. Boussebat Khaled Project Manager à IconSoftware, M.Hedjaz Mohammed Bachir & Mme. Boukhatem Fawzia enseignants de langue arabe pour leur collaboration en effectuant une expertise humaine sur nos jeux de données.

Un énorme merci pour nos parents -la raison pour laquelle nous réalisons ce mémoire de fin d'étude - ainsi que tous les membres des deux familles Feknous et Selhani, pour leur amour, encouragement et leur soutien moral et économique.

Enfin, nous voudrions remercier l'ensemble de nos proches, Hamel Oussama, Lamari Selena, Sba Warda, Kemmoum Amira, Bouskine Redha, Nour Walid, Saidani Smail qui ont toujours été là pour nous soutenir.

## الملخص

يمكن أن يؤدي إنشاء مجموعة أصول نصية موازية أحادية اللغة (عبارة عن مجموعة من البيانات اللغوية عادة ما تكون موجودة في قاعدة بيانات حاسوبية متكونة من جمل مختلفة الصيغة لكنها تحمل تقريبا نفس المعنى) باللغة العربية إلى معالجة مشكلة نقص البيانات العربية وتسهيل العديد من مهام معالجة اللغة التلقائية مثل ترجمة النصوص وأنظمة الأسئلة -الإجابة... نتناول في عملنا تقنيات إنشاء مجموعة أصول نصية موازية أحادية اللغة فنحن مهتمون أكثر بالنهج التي تعتمد على استخراج البيانات من مجموعة أصول نصية موازية ثنائية اللغة.

تم إنشاء مجموعتنا من خلال ترجمة أزواج من الجمل المستخرجة من مجموعات متوازية ثنائية اللغة . من لغات مختلفة إلى اللغة العربية، وقد تمت معالجة هذه الأزواج مسبقًا وتقييمها باستخدام تقنيات مختلفة بما في ذلك تدخل الخبرة البشرية. تم استدعاء نموذج -مصمم في عمل سابق- لتوليد جمل بصيغ مختلفة جديدة لتدريب عدد كبير من الأزواج -الجمل الاصلية والجمل بالصيغ الجديدة والتي تشكل أزواجًا مرشحة -, وكذا من أجل اختبار صحة ومفهومية ومقروئية الأزواج، وأخيرا إعادة صياغة الجمل باللغة العربية.



الكلمات الرئيسية: مجموعة أصول نصية، إعادة الصياغة، إنشاء مجموعة أصول نصية موازية ، المعالجة التلقائية للغة.

# ABSTRACT

The generation of a monolingual parallel corpus - known as paraphrase corpus - in Arabic can remedy the problem of lack of Arabic data and facilitate many NLP (Automatic Language Processing) tasks such as text translation, question systems - answer ... In our work we address techniques for generating paraphrase corpora, we are much more interested in data-driven approaches, more precisely an approach based on data extraction from bilingual parallel corpora. Our corpus was generated by translating pairs of sentences extracted from bilingual parallel corpora of different languages into Arabic, these pairs were preprocessed, aligned and evaluated involving different techniques including the intervention of human expertise. A model of generation of paraphrases named -EDAM- (designed in a previous work) was called to train a considerable number of the pairs - considered as candidate pairs -, and to test and finally generate paraphrases.

Keywords : Corpus, Paraphrase, Parallel corpus, Natural Language Processing (NLP).

# RESUME

La génération d'un corpus parallèle monolingue -dit corpus de paraphrases- en langue arabe peut remédier le problème de manque de données arabes et facilite beaucoup de tâches de TAL (Traitement Automatique de langage) tels que la traduction des textes, les systèmes de question-réponses …

Dans notre travail nous abordons les techniques de génération des corpus de paraphrases, nous nous intéressons beaucoup plus aux approches guidées par les données, plus précisément une approche basée sur l'extraction de données à partir des corpus parallèles bilingues.

Notre corpus a été généré en traduisant des couples de phrases extraits des corpus parallèles bilingues de différentes langues en langue arabe, ces couples ont été prétraités, alignés et évalués en impliquant différentes techniques y compris l'intervention d'une expertise humaine.

Un modèle de génération de paraphrases nommé -EDAM- (conçu dans un travail précédent) a été utilisé pour entrainer un nombre considérable des couples -considérés comme couples candidats-, et pour tester et enfin générer des paraphrases.

Mots clés : corpus, paraphrase, corpus parallèle, Traitement Automatique de langage.

## LISTES DES FIGURES



## Listes des tableaux

# TABLE DES MATIERES

# INTRODUCTION GENERALE

## 1 INTRODUCTION

La définition la plus courante du concept de paraphrase est basée sur le principe d'équivalence sémantique : la paraphrase est une forme de surface alternative dans la même langue, exprimant le même contenu sémantique que la forme originale de la phrase. Deux approches sont utilisées pour la génération de paraphrases : les approches de transformation basées sur la transformation de textes (grammaires, transformations de surfaces de phrases, Théorie sens-texte basée sur l'intention …) et les approches récentes guidées par les données (Data Driven). Cette dernière consiste à utiliser des corpus parallèles (composés de couples de paraphrases) pour entrainer un modèle de génération ou d'identification de paragraphes. Les corpus parallèles qui sont des ensembles de textes accompagnés de leurs traductions dans une autre langue, sont utilisés dans plusieurs tâches notamment la formation (entrainement) de modèles séquences à séquences pour la génération et l'identification de paraphrases, la simplification de texte, la génération automatique de résumés, la traduction de textes, les systèmes de questions/réponses, la recherche d'information, …

Pour la création d'un corpus parallèle monolingue en langue arabe nous optons pour une approche guidée par les données, nous prenons comme sources de données des corpus parallèles bilingues de divers domaines.

### i. Problématique

Nous nous intéressons aux approches guidées par les données appliquées à la langue arabe, cependant nous examinons les questions suivantes :

- Manque de corpus parallèles en langue arabe pour la génération et l'identification de paraphrases, sachant que l'en existe seulement un.
- Généralement les corpus parallèles existants sont des corpus de domaine général (ni spécifique ni multi-domaines) alors que souvent la tâche de paraphrase est liée à un domaine spécifique ou à une multiplicité de domaines partageant un contexte commun.

### ii. Objectifs

L'objectif principal consiste à automatiser la création d'un corpus parallèle (appelé aussi Dataset) pour la langue arabe en se basant sur les corpus parallèles bilingues existants.

La génération d'un corpus parallèle monolingue à partir d'un corpus parallèle bilingue doit passer par les sous objectifs suivants :

1. Explorer les approches de génération de corpus parallèles monolingues ;
2. Etudier les corpus parallèles bilingues disponibles ;
3. Construire des paires de phrases arabes candidates à partir des corpus bilingues préalablement étudiés ;
4. Aligner les paires de phrases candidates ;
5. Faire le prétraitement de paires de phrases candidates ;
6. Concevoir le modèle de génération automatique du corpus arabe ;
7. Entrainer le modèle de génération de paraphrase ;
8. Evaluer la qualité des paraphrases générées par le modèle ;

Pour la réalisation de nos objectifs établis, nous avons divisé le reste du mémoire en trois chapitres distincts.

Dans le chapitre deux, nous présentons les notions de base sur les paraphrases et les corpus de données de différents types ainsi qu'un ensemble d'approches de création des corpus parallèle.

Le chapitre trois porte sur la conception et la création de notre corpus parallèle monolingue pour la langue arabe

Enfin nous présentons dans le dernier chapitre les résultats issus de l'évaluation automatique et qualitative du corpus généré.

# CHAPITRE 1 : ETAT DE L'ART

## 1 INTRODUCTION

La richesse de la langue permet aux humains d'exprimer la même idée de façons très différentes. Cette variabilité d'expressions est une source majeure de difficultés dans la plupart des applications de traitements automatiques des langues. En effet, l'une des méthodes pour résoudre les problèmes engendrés par ce phénomène consiste à acquérir des paraphrases, à savoir un ensemble de phrases exprimant la même idée ou décrivant le même évènement. Dans notre travail nous  nous intéressons à la paraphrase et les corpus de paraphrases.

Les corpus de paraphrases à large échelle sont importants dans de nombreuses applications de TAL -Traitement automatique des langues -.

Dans ce chapitre nous présenterons les notions de base sur les corpus, les approches de création de corpus parallèles monolingues d'une façon générale et les corpus parallèles monolingues de la langue arabe, en citant un ensemble de travaux liés à notre thème, nous parlons en détails du seul et de l'unique corpus en langue arabe pour la génération de paraphrases qui est celui de Raisi présenté dans [1].

## 2 TRAITEMENT AUTOMATIQUE

« Le traitement automatique du langage naturel ou Natural Language Processing (NLP) est aujourd'hui l'un des domaines de recherche les plus actifs.  C'est un domaine à l'intersection du Machine Learning et de la linguistique. Il vise à extraire des informations et du sens à partir du contenu d'un texte, en fournissant plusieurs applications, nous en citons :

- Traduction de texte.
- Correcteur orthographique.
- Résumé automatique d'un contenu.
- Synthèse vocale.
- Classification de textes.
- Analyse d'opinion/sentiment.

- Prédiction du prochain mot sur smartphone.
- Extraction des entités nommées depuis un texte.
- … » [2]

## 3 PARAPHRASE

La paraphrase est une phrase ou une expression qui utilise des mots différents pour exprimer le même sens. Bien que la définition logique de la paraphrase nécessite une équivalence sémantique stricte. [3]

Les phrases ou expressions qui véhiculent le même sens en utilisant des formulations différentes sont appelées paraphrases, elles visent la conservation du sens; la naturalité donc il est nécessaire que la paraphrase soit syntaxiquement correcte, afin qu'elle ait un sens ; et l'adéquation à la tâche, en sachant que la génération automatique de paraphrase n'est pas une activité en soi (contrairement à la traduction par exemple), cette génération est donc toujours associée à une tâche et intégrée dans un processus plus large. Les paraphrases produites doivent être adaptées à l'usage qu'il en sera fait. [3]**.**

Par exemple, considérons les phrases (1) et (2) :

- يجب وضع كل هذا في مكانه الصحيح لضمان تحقيق أفضل النتائج الممكنة للعمل المنجز والمصاريف المتكبدة.
- كل هذا يجب أن يتم تنفيذه لضمان حصولنا على أفضل النتائج الممكنة للعمل الذي تم القيام به وللنفقات التي تم تكبدها.

Bien que l'interprétation stricte du terme "paraphrase" soit plutôt étroite, car elle nécessite exactement le même sens, dans la littérature linguistique, la caractéristique la plus courante de la paraphrase est que le sens des phrases est à peu près le même.

Produire ou générer des paraphrases est le fait de construire des phrases à partir d'autres déjà données afin de simplifier un texte, d'en faire un résumé, d'effectuer une analyse sémantique ou reformuler des recherches web, etc. [3]

Les paraphrases peuvent être classifiées selon deux critères : le niveau de granularité et le niveau d'analyse de la langue.[4]

- **Niveau de granularité**

Le niveau de granularité des paraphrases concerne la taille des séquences. Les séquences à un seul mot sont appelées paraphrases lexicales, les phrases qui partagent la même signification sont appelés des paraphrases phrastiques. Quant aux

fragments de phrases exprimant le même sens sont nommées paraphrases sous phrastiques. [5]

Les définitions de ces trois catégories sont décrites en détail ci-dessous.

- **Paraphrase lexicale**

Ce type concerne les unités lexicales individuelles ayant un sens similaire. Ces unités peuvent avoir une relation de synonymie tel que ( العامل -الأجير) ou bien une relation d'hyperonymie qui présente la notion de spécification/généralisation comme (المهنة - محامي). [5]

- **Paraphrase sous-phrastique**

Contrairement aux paraphrases lexicales, les paraphrases sous phrastiques recouvrent des fragments de texte (groupes de mots) présentant la même signification comme (« لقد حان الوقت لضمان تحقيق العدالة والتعويض عن » – « آن الأوان لإنصافكم وإنصافكم. الظلم »). [5]

- **Paraphrase phrastique**

Les paraphrases phrastiques sont des phrases qui transmettent le même sens en changeant seulement quelques mots et/ou passages comme dans cet exemple (« ابتسمت عندما طلب منها المغفرة » - « قبلت اعتذاره بابتسامة »). [5]

De nos jours, la génération de paraphrases est une tâche très courante dans le domaine de la traduction Du Tal. Elle est particulièrement utilisée pour diverses applications :

- **Niveau d'analyse de la langue**

La différence entre l'interprétation sémantique ou linguistique et l'interprétation pragmatique ou contextuelle est le point de départ de la véritable théorie de la sémantique conditionnelle. Selon cette théorie, comprendre une phrase, c'est savoir dans quelles conditions elle est vraie. Par conséquent, la question n'est pas de savoir si une phrase donnée est vraie ou fausse. [4]

Il est à noter que la différence entre ces deux catégories En linguistique, la paraphrase implique la compréhension de La différence entre deux concepts importants

Une phrase est basée sur les règles structurelles de la syntaxe et selon des critères de grammaticalité bien définis.

L’énoncé est un phénomène variable lié à l’activité langagière. Il se rapporte au contexte et donne du sens Cela dépend de la compréhension et de l'interprétation de ce dernier. [4]

- **Paraphrase sémantique ou linguistique** : La paraphrase linguistique se base sur les correspondances syntaxiques et ou lexicales entre les phrases. On distingue deux types à savoir : les paraphrases lexicales et les paraphrases lexico syntaxique.
- **Paraphrase syntaxique :** afin d’obtenir une paraphrase syntaxique, plusieurs transformations peuvent être utilisées dont la nominalisation, conversion d’un adjectif en syntagme nominal, l’épithétisation, transformation d’une proposition relative en adjectif épithète. Exemple : (« منطقة الصناعة » - « المنطقة الصناعية »). Ce type de paraphrase se base donc sur une règle qui spécifie les conditions du passage d’une phrase à l’autre.
- **Paraphrase lexico-syntaxique :** Dans ce type de paraphrases, en plus des modifications opérées sur le niveau syntaxique, des modifications sur le niveau lexical sont effectuées.

« La constitution de telles paraphrases peut être faite en appliquant une redistribution des arguments sur des actants différents ou bien en appliquant une double négation. Exemple : « لقد رسب في امتحانه » - « لم يجتاز امتحانه ». [4]

- **Paraphrase non-linguistique** Contrairement aux paraphrases linguistiques, les paraphrases non linguistiques se basent sur des phrases comportant la même idée ou référant la même chose sans chercher des correspondances lexicales ou syntaxiques ce qui nécessite l’intervention de l’expert
- **Paraphrase pragmatique :** Deux paraphrases pragmatiques sont des phrases qui réfèrent à la même intention de telle sorte que les phrases sont interprétées de la même façon en se basant sur l’expérience et les connaissances. Exemple : « الجو حار » - « أريد تشغيل التكييف ».
- **Paraphrase référentielle :** Dans ce type de paraphrases, il est nécessaire de connaitre les références de certains termes. Exemple : « مدينة الورود » - « البليدة ». Dans notre cadre, le type de paraphrases à générer dépend du type de paraphrases que comporte le corpus d'entrainement.

# 4 CORPUS

Depuis les années 1980, la linguistique de corpus s'est développée à une vitesse accélérée. Alors que la construction et l'exploitation de corpus de langue anglaise dominent encore la recherche en linguistique de corpus, en particulier des langues européennes typologiquement apparentées comme le français, l'allemand et le portugais et des langues asiatiques comme le chinois, le coréen et le japonais, sont également devenus disponibles et ont notamment ajouté à la diversité des études linguistiques fondées sur des corpus.

Lorsqu'on désigne un corpus impliquant plus d'une langue comme un corpus multilingue, le terme multilingue est utilisé au sens large. Un corpus multilingue, au sens restreint, doit impliquer au moins trois langues tandis que ceux impliquant seulement deux langues sont classiquement appelés corpus bilingue. [7]

Lorsque nous définissons différents types de corpus, nous pouvons utiliser différents critères, par exemple, le nombre de langues concernées, et le contenu ou la forme du corpus. Mais lorsqu'un critère est décidé, le même critère doit être utilisé de manière cohérente. Par exemple, on peut dire qu'un corpus est monolingue, bilingue ou multilingue si l'on prend le nombre de langues concernées comme critère de définition. Mais si nous choisissons de définir les types de corpus par le critère de la forme du corpus, nous devons l'utiliser de manière cohérente. On peut alors dire qu'un corpus est parallèle si le corpus contient des textes sources et des traductions en parallèle, ou qu'il s'agit d'un corpus comparable si ses sous-corpus sont comparables en appliquant la même base de sondage. [7]

## 4.1 CORPUS DE PARAPHRASE

Les corpus de paraphrases sont des collections de paraphrases, qui consistent en des expressions linguistiques avec une formulation différente et (approximativement) la même signification.

## 4.2 Classification des corpus

Nous allons détailler les trois classes de corpus.

### 4.2.1 Corpus Quasi-Comparable

« Un corpus quasi-comparable contient des documents bilingues non parallèle et non alignés. Ces documents peuvent porter sur le même sujet ou des sujets très différents. Dans un tel corpus, un petit nombre de phrases bilingues sont des traductions les unes des autres, tandis que d'autres sont des paraphrases bilingues ». [8]

### 4.2.2 Corpus Comparable

« Un corpus comparable peut être défini comme un corpus contenant des composants qui sont collectés en utilisant la même base de sondage et un équilibre et une représentativité similaire, par exemple les mêmes proportions des textes du même genre dans le même domaine dans une gamme de différentes langues dans le même période d'échantillonnage. Cependant, les sous-corpus d'un corpus comparable ne sont pas des traductions les uns des autres. Au contraire, leur comparabilité réside dans leur même base de sondage et un équilibre similaire ». [7]

### 4.2.3 Corpus Parallèle

« Un corpus parallèle peut être défini comme un corpus qui contient des textes sources et leurs traductions. Les corpus parallèles peuvent être bilingues ou multilingues. Ils peuvent être unidirectionnels (par exemple de l'anglais vers le chinois ou du chinois vers l'anglais uniquement), bidirectionnels (par exemple contenant à la fois des textes sources anglais avec leurs traductions chinoises ainsi que des textes sources chinois avec leurs traductions anglaises) ou multidirectionnels. (Par exemple le même texte avec les versions anglaise, française et allemande) ». [7]

Un corpus parallèle peut fournir une grande aide pour les chercheurs des domaines, tels que la traduction automatique statistique, la recherche d'informations multilingue et la lexicographie bilingue. Les phrases des corpus parallèles alignés sont une ressource linguistique majeure pour la traduction automatique statistique. Nous citons quelques domaines d'application :

- **Développement d'outils de calcul**

Les corpus parallèles peuvent engendrer plusieurs outils de calculs tel que :

Le système de comptage de fréquences, le tokenizer de mots, l'analyseur de phrases, etc. dépendent de la disponibilité des ressources linguistiques.

Les mémoires de traduction, les systèmes de traduction automatique, les protocoles bilingues et les traitements de texte nécessitent des corpus parallèles.

- **Etude de caractéristiques linguistiques**

Les mots de la langue peuvent être utilisés dans différents contextes et avoir des significations différentes. Apprendre un mot avec des voisins aide à lever l'ambiguïté et à comprendre sa signification. Par conséquent, les textes multilingues sont utiles pour étudier les changements d'utilisation des mots d'une langue à une autre, qui peuvent être utilisés pour l'extraction des connaissances et l'étude des modèles d'utilisation des langues. Cela peut également aider à étudier la distribution de fréquence dans le texte original et le texte traduit. Ceux-ci sont utiles pour les éducateurs et les linguistes qui éduquent les langues.

### 4.3 Comparaison entre les corpus comparables et parallèles

Les corpus parallèles et comparables sont censés être utilisés à des fins différentes, les deux sont également conçus avec des objectifs différents. Pour un corpus comparable, la base de sondage est essentielle. Les composants représentant les langues concernées doivent correspondre entre eux en termes de proportion, de genre, de domaine et de période d'échantillonnage. Pour un corpus parallèle, la base de sondage n'est pas pertinente, car tous les composants du corpus sont des traductions exactes les uns des autres. Cependant, cela ne signifie pas que la construction de corpus parallèles est plus facile. Pour qu'un corpus parallèle soit utile, une étape essentielle est d'aligner les textes sources et leurs traductions, c'est-à-dire faire le lien entre les deux, au niveau de la phrase ou du mot. [7]

## 5 APPROCHES DE CREATION DE CORPUS PARALLELES

Le corpus parallèle est un corpus bilingue, qui contient le texte dans la langue source et la traduction équivalente dans la langue cible.

En termes de charge de travail et de temps de travail, la création manuelle d'un grand corpus parallèle peut être très coûteuse. Par conséquent, nous avons besoin d'une méthode pour créer automatiquement et efficacement un corpus parallèle.

### 5.1 Extraction de données parallèles

Peu importe son type qu'il soit parallèle ou comparable, un corpus lui-même peut être une source de données pour la génération d'un autre corpus parallèle ce qui introduit que les sources potentielles de phrases susceptibles de produire des phrases parallèles peuvent décider les approches de création d'un corpus parallèle, Nous pouvons également extraire des données parallèles à partir de sources Web.

La figure 1 donne une vision sur les approches de créations existantes.

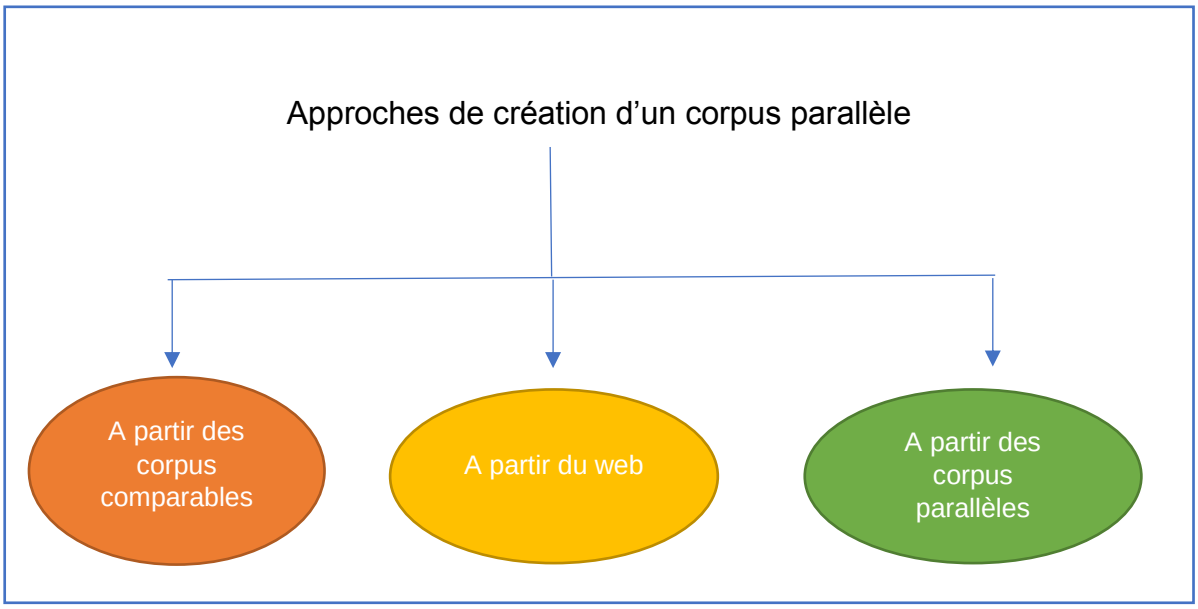


Figure 1 Approche de création des corpus parallèles

#### 5.1.1 Création d'un corpus parallèle à partir d'un corpus comparable

Le corpus comparable est composé de documents bilingues, qui ne sont pas alignés sur des phrases, mais des traductions approximatives les uns des autres. Dans le corpus comparable, les phrases ne sont pas vraiment traduites les unes dans les autres, mais traitent à peu près les mêmes informations, elles peuvent donc contenir des phrases parallèles. Ce type de corpus comparable peut être utilisé pour trouver des phrases parallèles ou des fragments parallèles. Un corpus comparable et un corpus non parallèle peuvent être largement utilisés dans toutes les paires de langues.[1]

- **Extraction De Phrases Parallèles**

Un système d'extraction de phrases parallèles est typiquement divisé en deux tâches. En premier lieu, un alignement au niveau des documents est réalisé pour associer les

paires de documents pertinents. Ensuite, les paires de phrases parallèles sont détectées dans ces paires de documents. Certains travaux se sont concentrés sur la première tâche en proposant des systèmes conçus pour créer des corpus comparables de haute qualité, alors que plusieurs autres font de l'extraction de phrases à partir de ces corpus comparables pour générer de nouveaux corpus parallèles.

Extraire des phrases parallèles à partir de corpus comparables est un vaste domaine de recherche en soi. Elle est restée une tâche difficile même après plus d'une décennie. La plupart des techniques utilisées dans la création du corpus en question sont mises en œuvre sur un ensemble de paires de documents bilingues qui sont alignées sur des sujets, c'est-à-dire que chaque paire de documents a un contenu sur des sujets à peu près similaires, de sorte qu'ils ont tendance à transmettre beaucoup d'informations qui se chevauchent. Cela améliore les chances de trouver de bonnes phrases parallèles tout en réduisant également l'espace de recherche.

Wikipédia est un bon exemple de corpus comparable qui fournit facilement une telle paire de documents signés par le biais de son "Interwiki" liens. [8]

Après avoir identifié la source de données parallèles (Wikipédia, articles, thèses de doctorat…), la prochaine étape consiste à trouver des paires de documents similaires en alignant les documents. Une fois les documents alignés, il devient possible d'extraire des phrases parallèles par paires de documents similaires.

Nous citons comme exemple le corpus parallèle monolingue d'anglais pour la simplification du texte est obtenu à partir d'un corpus comparable contenant des textes complexes et simples présenté dans [9]. La similitude entre les phrases obtenues à partir de toutes les combinaisons de phrases complexes et simples est calculée en utilisant l'alignement entre les plongements de mots, puis les paires de phrases avec une similitude dépassant un certain seuil sont sélectionnées comme entrées dans le corpus parallèle.

### 5.1.2 Le Web comme source de corpus

« L'utilisation du Web comme base pour la constitution de ressources textuelles est très récente. Ces dernières années sont le témoin sur les travaux tentant d'exploiter ce type de données. Dans une perspective de traduction automatique Resnik [10] a

étudié la possibilité d'utiliser les sites Internet proposant les informations en plusieurs langues pour constituer des corpus parallèles bilingues.

Ghani et al [11] ont exposé l'idée de construction de corpus, à partir du web, par interrogation automatique de moteurs de recherche. Ils ont exploité cette idée pour la constitution de corpus de langues minoritaires.

Dans une tout autre approche Issac et al [12] ont mis au point un logiciel pour la constitution d'un corpus de phrases dans le but d'étudier le comportement des noms prédicatifs marquant la localisation et le déplacement, afin de mesurer si l'introduction des prépositions dans les requêtes en recherche d'information permet d'améliorer la précision.

En 2004, Baroni et Bernardini [13] introduisent le « BootCAT Toolkit. Un ensemble d'outils permettant la construction itérative de corpus par interrogation automatique de Google, et l'extraction de terminologie. Bien qu'il soit dédié à la mise au point de corpus spécialisés, cet outil fut utilisé par Baroni et Ueyama [14] pour la constitution de corpus généralisés.

Les documents du web ont également été utilisés par nombre de chercheurs pour adresser le problème du manque de données en modélisation statistique du langage ». [15]

Nous citons ci-dessous un aperçu de quelques travaux de constructions et d'utilisation de corpus parallèle monolingue :

- Le corpus de Wikipédia en anglais simple (SEW) [16] est généré en alignant Wikipedia anglais simple et Wikipedia anglais et contient des sections alignées sur les articles et les phrases. La section alignée sur les phrases contient 167 686 paires de phrases alignées [17].
- Le turc appartient à la famille des langues altaïques de l'Asie occidentale et centrale qui sont nommées sous les langues turques. Le turc est une langue hautement fléchie et agglutinante. L'utilisation productive des affixes est très typique, soit pour changer le sens ou l'accentuation d'un mot. Le turc et l'anglais utilisent le même alphabet latin, à l'exception de quelques lettres. Pour la création de TuPC des phrases ont été extraites et couplées à partir de sites

Web d'actualités quotidiennes. Les paires candidates ont ensuite été annotées à la main en examinant leur contexte et ont été notées en fonction de la similitude sémantique. [18]

### 5.1.3 Corpus parallèle monolingue à partir d'un corpus parallèle bilingue

Le corpus parallèle monolingue est utile pour de nombreuses tâches, y compris la formulation de modèles de séquence à séquence et d'autres tâches de formulation telles que la synthèse, la traduction automatique et la simplification de texte. Son utilisation ne se limite pas à la formation de modèles à des tâches génératives, mais également à des tâches analytiques (telles que la récupération d'informations et la réponse à des questions).

Nous citons ci-dessous un aperçu de quelques travaux de construction et d'utilisation de corpus parallèle monolingue :

- Le corpus Newsela (Actualités) [19] est un corpus monolingue parallèle d'anglais et contient 1 911 articles de presse qui ont été réécrits manuellement jusqu'à cinq fois avec un niveau de complexité décroissant.
- Le corpus de Raisi [1] est un corpus monolingue parallèle d'arabe et contient 100.000 couples de phrases.

# 6 CONSTRUCTION DE CORPUS PARALLELES

Après la collection et le prétraitement de données parallèles, les étapes suivantes sont nécessaires pour la création du corpus.

## 6.1 Traduction de données

Il existe aujourd'hui différentes méthodes de traduction. Dans le passé, la traduction était encore principalement une activité humaine, mais avec l'émergence des nouvelles technologies (ordinateurs, Internet et autres supports informatiques), nous avons aujourd'hui trois méthodes de traduction possibles.

- **Traduction automatique**

La traduction automatique (TA), ou machine translation (MT) en anglais, est un sous-domaine du traitement automatique du langage (TAL) intéressés par la traduction du langage naturel (texte ou parole) à une autre langue à l'aide d'un logiciel. Cette Tâches

complexes impliquant une analyse linguistique détaillée de la langue source et La génération de contenu linguistique dans la langue cible nécessite plusieurs techniques de TAL, la TA est maintenant une industrie Une réponse mondiale majeure à la société, au gouvernement, aux entreprises et à la société armée.
En effet, il existe de nombreux systèmes de Traduction Automatique, dont certains sont d'ores et déjà opérationnels comme Google translate, SYSTRAN, TAUMMETEO, MÉTAL ou les systèmes des sociétés américaines ALPS Systems ou Weidner ; cependant ils ont tous recours, après la phase de traitement automatique, à des traducteurs humains qui assurent une révision du texte produit. [1]

- **Traduction humaine**

La traduction humaine est une traduction effectuée par des humains sans l'aide ou l'intervention d'une machine. C'est la méthode de traduction traditionnelle.

Le professionnel de la traduction utilise les fiches, les dictionnaires, le stylo et les ressources référencées tout au long du processus de travail pour traiter ses documents à traduire du début à la fin. Le traducteur effectue son travail du début à la fin, sans utiliser d'outils informatiques. De plus, le traducteur doit être un généraliste qui connait un peu de tout.

- **Traduction Assistée par Ordinateur (TAO)**

La traduction assistée par ordinateur est une méthode de traduction dans laquelle les traducteurs utilisent des outils informatiques pour la traduction. Il fait sa traduction en personne, mais utilise des logiciels de traduction et des sites Web comme support et référentiel. Ce qu'il faut donc noter ici, c'est que les ordinateurs jouent un rôle auxiliaire dans la traduction manuelle. Dans la traduction assistée par ordinateur, il est vrai que des personnes font la traduction, mais il existe un support informatique pour simplifier la tâche. Idéalement, un logiciel de TAO peut traduire des documents d'une langue appelée langue source vers une autre langue appelée langue cible. Par exemple, traduire des documents anglais (langue source) en documents français (langue cible)

Enfin, disant que la traduction humaine (faite par un traducteur professionnel) reste la méthode de traduction la plus fiable. Pour les documents longs, ils peuvent être

traduits à l'aide d'un ordinateur (TAO) mais faire une traduction purement automatique présente beaucoup de pièges et de défauts.

### 6.2 Filtrage DES CORPUS

Les paires de paraphrases peuvent être identifiées en calculant diverses mesures de similarité / distance de phrases entre les deux phrases d'une paire.

- **MESURES DE SIMILARITE**

Le calcul de similarité est une tâche de base du traitement automatique des données, qu'il s'agisse de texte ou non, et a toujours reçu une attention particulière de la communauté scientifique. Si l'on se limite aux données textuelles, le besoin est avant tout venu de besoin pour la recherche d'information.[20]

En effet, d'une part, c'est la capacité de mesurer la proximité entre documents et d'autre part de pouvoir identifier, et trier, la liste des documents les plus pertinents à offrir en réponse à une requête dans un moteur de recherches.[20]

Il existe trois catégories principales des approches de similarité :

- Similarité syntaxique.
- Similarité sémantique.
- Similarité hybride.

- **Mesures de similarité syntaxiques**

Une mesure de similarité syntaxique permet de comparer des documents textuels en se basant sur les chaines de caractères qui les composent. Par exemple, les chaînes de caractères "voiture" et "voiturier" peuvent être considérées comme similaires alors que "voiture" et "automobile" pourront être considérées comme dissimilaires Dans ce chapitre, nous présentons les mesures de similarité syntaxique les plus utilisées.

- **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, antonymie, ou troponymie2 entre eux (Exemples : MEDECIN-CHIRURGIEN, SOMBRE-CLAIR).

## 6.3 METRIQUES D'EVALUATION AUTOMATIQUE DE PARAPHRASES

Il y a plusieurs problèmes inhérents à l'évaluation automatiques des paraphrases. Premièrement, il est difficile de fournir une liste exhaustive de définitions pour une phrase donnée. Lin et Pantel [21] illustrent les difficultés rencontrées par les personnes lors de la génération de listes de paraphrases, soulignant qu'ils ont raté de nombreux exemples de génération de système qui se sont avérés plus tard corrects. Si la liste de paraphrases de référence est incomplète, alors L'utiliser pour calculer la précision entraînera des chiffres inexacts. Si le système génère des interprétations correctes qui ne figurent pas dans la liste de référence, la précision sera incorrectement faible. De plus, le rappel est incertain, car il n'y a aucun moyen de savoir combien de paraphrases correctes existent.

Il existe d'autres obstacles à l'évaluation automatique des paraphrases. Même la proposition une liste raisonnablement exhaustive de paraphrases pour une phrase, l'acceptabilité de chaque paraphrase sera différente en fonction du contexte de la phrase originale [22]. Alors que les paraphrases lexicales et phrastiques peuvent être évaluées en les comparant à une liste de paraphrases connues (peut-être personnalisées pour des contextes particuliers), cela ne peut pas être naturellement fait pour les paraphrases structurelles qui peuvent transformer des phrases entières.[23]

Ces problèmes amènent les chercheurs à utiliser des mesures de traduction automatique tel que les METEOR, TER, BLEU, WER et GLEU, et ce, pour mesurer la qualité des paraphrases. Nous présentons ces métriques dans ce qui suit.

### 6.3.1 METEOR

Le Météor est une mesure d'évaluation de la traduction automatique, initialement développée et publiée en 2004, a été conçue dans le but explicite de produire des scores au niveau de la phrase qui correspondent bien aux jugements humains de la qualité de la traduction au niveau de la phrase. Plusieurs décisions de conception clés ont été intégrées dans Météor à l'appui de cet objectif. Contrairement à IBM Bleu, qui utilise uniquement des fonctionnalités basées sur la précision, Météor utilise et met l'accent sur le rappel en plus de la précision, une propriété qui a été confirmée par plusieurs métriques comme étant critique pour

une corrélation élevée avec les jugements humains. Météore aborde également le problème de la variabilité des traductions de référence en utilisant une correspondance de mots flexible, permettant de prendre en compte les variantes morphologiques et les synonymes comme des correspondances légitimes. De plus, les ingrédients caractéristiques dans Météore sont paramétrés, permettant le réglage des paramètres libres de la métrique à la recherche de valeurs qui aboutissent à une corrélation optimale avec les jugements humains. Les paramètres optimaux peuvent être réglés séparément pour différents types de jugements humains et pour différentes langues.[24]

### 6.3.2 BLEU

IBM Bleu métrique [25] a été la métrique automatique la plus largement utilisée ces dernières années. Bleu est rapide, facile à exécuter et peut être utilisé comme fonction cible dans les méthodes d'entraînement à l'optimisation des paramètres couramment utilisées dans les systèmes de TA statistique de pointe [26]. Cependant, bien que populaire, des faiblesses ont été notées dans Bleu ces dernières années, notamment le manque de scores fiables au niveau de la peine. Météor, ainsi que d'autres métriques telles que GTM, TER [27] et CDER [28], ont été développés dans le but de remédier spécifiquement à cela et à d'autres faiblesses identifiées dans Bleu. [24]

Son principe de base consiste à compter le nombre des n-grammes (uni-gramme, bi-grammes, trigrammes et quadrigrammes) de la phrase candidate présents dans les phrases de références.

Ce nombre est exprimé par le calcul de la précision standard P = n /c tel que :

(n): Nombre de n-grammes de la phrase candidate présents dans les références.

(c): Nombre de n-grammes de la phrase candidate.

### 6.3.3 WER

La métrique WER (Word Rate Error) est utilisée pour mesurer la performance des systèmes qui traitent la langue naturelle, tel que les systèmes de traduction automatique [29].

Cette métrique, dérivée de la distance de Levenshtein, travaille sur le niveau des mots, elle est très utilisée pour faire des comparaisons entre systèmes. Elle compare entre deux phrases, référence et hypothèse.

### 6.3.4 GLEU

La métrique GLEU (Google-BLEU) est une variante de la métrique BLEU, utilisée spécialement pour mesurer taux de corrections d'erreurs grammaticales des n-grammes générés avec l'ensemble des phrases références. Il a été prouvé que les résultats générés par cette métrique sont proches de ceux relatifs aux humains [30] [31].

## 7 CORPUS PARALLELE EN LANGUE ARABE

Nous présentons un corpus monolingue parallèle de phrases complètes en arabe, généré automatiquement à partir de la traduction d'un corpus bilingue parallèle, pour cela nous présentons quelques notions sur la langue arabe.

### 7.1 Définition de la langue arabe :

L'alphabet arabe comparé à son monôme latin est plus difficile à traiter car il contient plus de caractéristiques, qui compliquent la tâche grammaticale.

De par ses caractéristiques morphophonologiques, morpho-syntaxiques et parfois même lexicales, la langue arabe, tout comme les langues de cette même famille sémitique, l'arabe repose essentiellement sur la syntaxe morphologique dite non enchaînant de ses morphèmes, dans laquelle il existe plusieurs modèles théoriques de description (structuralisme, fonctionnalisme, générativisme). En ce sens, c'est encore un langage assez complexe qu'il faut gérer dans le domaine du traitement automatique du langage.

En règle générale, la langue arabe diffère du français et de l'anglais principalement sur le plan de sa structure syntaxique, de ses temps et modes. C'est une langue du type beaucoup plus VSO (Verbe-Sujet-Objet) que SVO(Sujet-Verbe-Objet).

Elle fonctionne avec trois temps, le perfectif, l'imperfectif et le participe et cinq modes, l'indicatif, l'impératif, le subjonctif, le jussif et l'énergétique. [32]

- **Le lexique arabe**

Le lexique de la langue arabe comprend trois catégories grammaticales de mots : verbe, nom et particule.

a. **Verbe :** Unité lexicale référant à un état ou une action exprimant un sens dépendant du temps comme :

Travailler → عمل , partir → ذهب [23]

Nous pouvons classer les verbes arabes selon plusieurs critères [33] :

- Selon le critère de temps, il existe trois types : l'accompli, inaccompli, impératif.
- Selon leur sens et leur transitivité de sujet au complément aux deux types : Intransitive, transitive.
- Selon leurs modes : la voix passive et la voix active.
- Selon le nombre des consonnes de la racine, la majorité des verbes (à peu près de 85%) sont formés sur 3 lettres et le reste entre les racines de 4 et 5 lettres. Ces racines peuvent donner plusieurs schèmes avec des transformations morphologiques.
- Selon leur conjugaison il existe : le conjugué et le non conjugué (ou bien invariant).

**b. Nom:** Toute unité lexicale référant à un sens indépendant du temps [23] , regroupent: Les adjectifs ; féminin et masculin ; les noms démérites, les noms prolongé ainsi que les noms réduits ; les noms communs et les noms propres ; les pronoms et leurs types (connectés et séparés) ; les pronoms relatifs ; les pronoms démonstratifs ; les noms d'interrogations ; les noms déterminés et non-déterminés ; les noms de périphrases ; les noms du verbe ; les noms de voix ; les semblables des verbes de noms [24].

**c. Particule** : Entité invariable exprimant un sens dépendant de compréhension. La langue arabe contient un nombre limité ne dépasse pas 80 éléments, ils se nomment en arabe les particules de sens حروف المعاني)) par contre l'alphabet arabe se nomment les particules déconstruction (حروف المباني) [24].

Nous étudions en détails la création de l'unique corpus parallèle arabe.

Dans [1] Raisi a conçu le premier corpus parallèle monolingue arabe en novembre 2018, spécifiquement pour la formation de modèles séquence à séquence, il s'agit d'un corpus généré automatiquement à partir de la traduction de corpus parallèles bilingues. Sa construction est représentée en utilisant des phrases appariées des deux autres langues d'un corpus bilingue parallèle et en les traduisant en arabe en parallèle, pour cela, [1] a choisi un corpus parallèle bilingue puis a utilisé de puissants outils de traduction, l'API google, est utilisé pour générer des paires de phrases français-arabe anglais-arabe. Finalement, elle fait correspondre la sortie pour construire un corpus parallèle de corpus Arabe. Enfin, il effectue une évaluation grammaticale des résultats

obtenus et élimine toute paire de phrases qui ont généralement le même contenu dans l'étape de reconnaissance de la paraphrase.

### 7.2 Analyse des résultats de création du corpus arabe

- Un corpus monolingue parallèle d'arabe contenant 100000 paires de phrases.
- Aucune paire de phrases ne contient des phrases identiques.
- La longueur moyenne des phrases arabes traduites à partir de la source anglaise est de 20 mots.
- La longueur moyenne des phrases arabes traduites à partir de la source française est de 19 mots.
- Les phrases obtenues à partir de la traduction du français sont plus courtes.
- Les phrases obtenues à partir de la traduction de l'anglais sont plus longues.

## 8 DIFFICULTES ET DEFIS A SOULEVER

Durant notre période de recherche en langue arabe nous avons constaté qu'il y a très peu de projets de recherche sur les corpus et la génération de paraphrases.

La langue arabe est une langue connue par la richesse de son vocabulaire et en revanche l'absence de voyelles qui nécessite un travail approfondi pour l'étudier. Pour cela la tâche de recherche et de traitement automatique en langue arabe est devenue difficile à gérer en raison du manque de ressources de base, il n'y a pas assez de corpus accessibles au public pour la faciliter. Toutes ces difficultés seront une sorte de motivation pour penser à la génération d'un corpus arabe multi-domaines, pour l'enrichissement des outils de manipulation de cette langue et aussi la recherche dans ce domaine.

## 9 CONCLUSION

Dans ce chapitre nous avons abordé la notion des corpus de données en général et les corpus parallèles monolingue de langue arabe plus précisément, nous avons présenté les approches de création d'un corpus et l'intérêt de sa génération dans multiples domaines.

Dans le chapitre suivant nous allons présenter notre méthodologie pour la génération d'un corpus parallèle monolingue en langue arabe.

# CHAPITRE 2 : CONCEPTION

## 1 INTRODUCTION

La création de notre corpus a commencé par une étape de planification minutieuse où les principes de conception ont été élaborés. Ces principes comprenaient les critères de sélection qui ont servi de base à la collecte des corpus parallèles bilingues déjà existants. La conception du corpus comprend les phases d'attribution, de collecte, de prétraitement, de traduction et de post-traitement des données.

Lors de l'attribution de ces dernières, nous essayons de dresser un plan des données à couvrir et à collecter. Nous préparons une liste en conséquence, une fois qu'un corpus parallèle bilingue approprié a été identifié et que l'autorisation de l'utiliser a été obtenue.

Les couples de phrases des corpus obtenus sont ensuite traduits en langue arabe puis prétraités et édités et enfin évalués automatiquement.

Notre travail vise la génération d'un corpus parallèle riche et volumineux en langue arabe, pour cela nous allons suivre les étapes illustrées dans la figure ci-dessous :

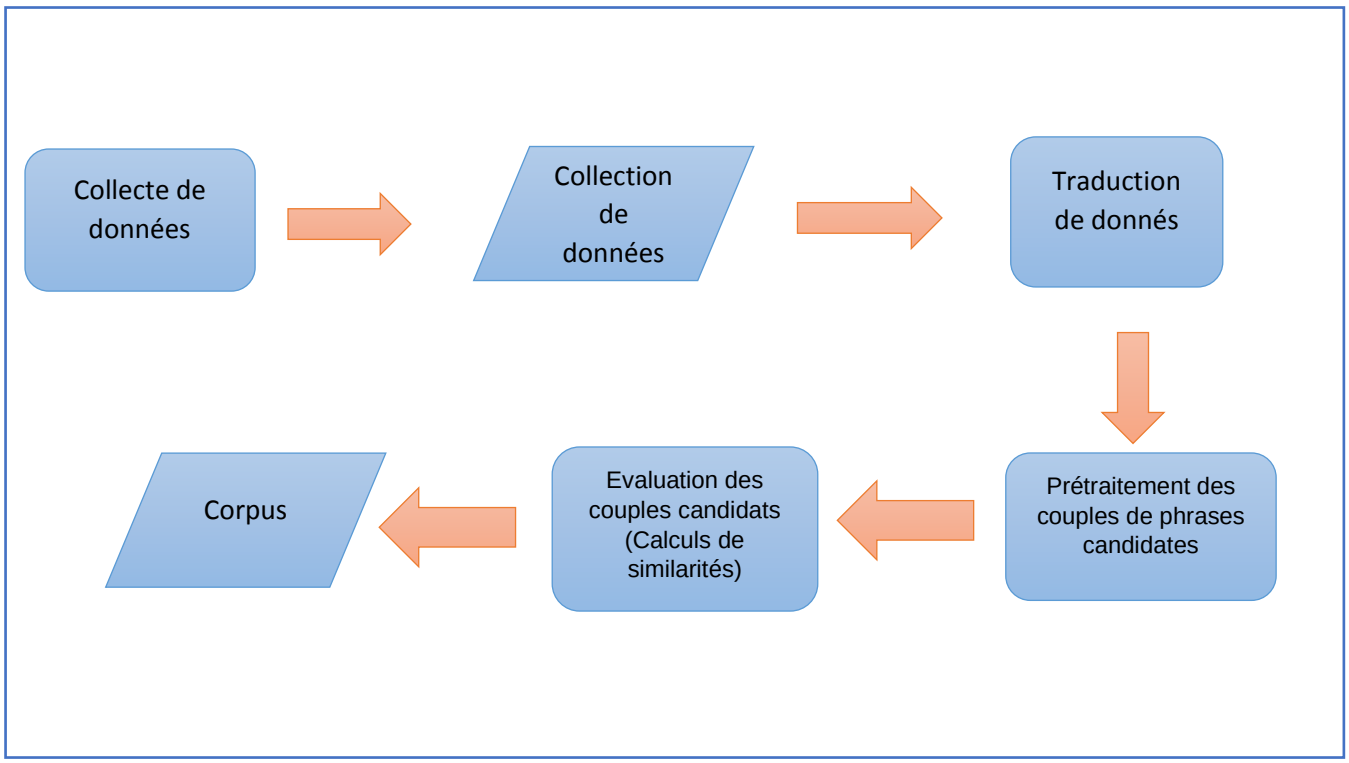


*Figure 2 Etape de conception du corpus*

Pour réussir notre objectif représenté par la génération d'un corpus multi domaines assez riche et volumineux, nous allons entamer notre travail par collecter des corpus parallèles bilingues comportant des couples de phrases dans différents domaines, que nous allons traduire en langue arabe par la suite. Cependant nous allons faire passer ces phrases par une étape de filtrage et de prétraitement, puis à l'évaluation de leur qualité par les techniques des calculs de similarités, nous aurons ensuite des couples de phrases candidats prétraités et prêts pour être évalué automatiquement dans une étape suivante.

Dans la section suivante nous allons expliquer chaque étape.

## 2 COLLECTION DE DONNEES

Les contenus parallèles nécessaires pour construire un corpus parallèle dans n'importe quelle langue sont généralement collectés à partir de textes accessibles au public, principalement à partir du Web, ou bien à partir d'autres corpus qu'ils soient parallèles ou comparables. Cependant, dans ce projet nous avons choisi d'extraire nos données arabes à partir de corpus parallèles existants car de toute évidence, le plus grand avantage des corpus parallèles est que les paires de phrases sont des paraphrases presque par définition ; ils représentent différents rendus du même sens créés par différents traducteurs faisant des choix lexicaux différents. En effet, ils contiennent des paires (ou ensembles) de phrases qui sont soit sémantiquement équivalentes, soit ont un chevauchement sémantique significatif.

Par conséquent, nous avons opté pour la traduction des corpus parallèles bilingues de différents domaines- détaillés sur le tableau ci-dessous.

Tableau 1 Présentation des corpus utilisés

| Corpus Parallèle | EUROPARL Français-Anglais | JRC | EUROPARL Italien-Français | MIZAN | SPC |
|---|---|---|---|---|---|
| Taille Initiale | 2.1 millions | 4.4 millions | 1.9 million | 1 million | 2.228 |
| Nombres de couples utilisés | 648.000 | 480.000 | 260.000 | 240.000 | 2.228 |
| Langue | Français-Anglais | Italien-Romain | Italien-Français | Anglais-Persan | Anglais-Chinois |
| Type du corpus | Unidirectionnel | Unidirectionnel | Unidirectionnel | Unidirectionnel | Unidirectionnel |
| Domaine | Parlement Européen | Acquis communautaires de l'union européenne | Parlement Européen | Chefs-d'œuvre de la littérature | Règlements des sociétés |

- **EUROPARL**

Europarl est un corpus parallèle créé à partir des travaux du Parlement européen dans les langues officielles de l'UE. Il comprend 21 langues européennes : roman (français, italien, espagnol, portugais, roumain), germanique (anglais, néerlandais, allemand, danois, suédois), slavik (bulgare, tchèque, polonais, slovaque, slovène), finni-ougrienne (finnois, Hongrois, estonien), balte (letton, lituanien) et grec. Le corpus a été élargi à plusieurs reprises avec une taille finale d'environ 60 millions de mots par langue. Les textes datent de la période janvier 2007 - novembre 2011.

  - La version 1, c'est la version originale, contient des données depuis avril 1996 jusqu'à décembre 2001.

- La version 2 ajoute les données de janvier 2002 jusqu'à septembre 2003.
- La version 3 ajoute les données d'octobre 2003 jusqu'à octobre 2006.
- La version 5 ajoute les données de novembre 2007 jusqu'à octobre 2009.
- La version 6 ajoute les données de novembre 2009 jusqu'à décembre 2010.
- La version 7 ajoute les données de Janvier 2011 jusqu'à novembre 2011.
- 3/ ce corpus contient plusieurs corpus parallèles bilingue :
    1. Un corpus parallèle Allemand-Anglais 04/1996-11/2011, il contient 44,548,491mots en German et 47,818,827 mots en anglais.
    2. 1.Un corpus parallèle Espagnol-Anglais 04/1996-11/2011, il contient 51,575,748 en espagnol 49,093,806 mots en anglais.
    3. 1.Un corpus parallèle Français-Anglais 04/1996-11/2011, il contient 51,388,643 en français et 50,196,035 en anglais.
    4. 1.Un corpus parallèle Italien-Anglais 04/1996-11/2011, il contient 47,402,927 mots en italien et 49,666,692 mots en anglais.

Toutes les version d'europarl sont disponible sur : https://www.statmt.org/europarl/

- **MIZAN**

Le contenu parallèle requis pour la construction du corpus est souvent collecté à partir de textes accessibles principalement depuis le Web. Cependant, de nombreuses langues, y compris le persan, manquent de corpus parallèles suffisamment applicables pour bénéficier de SMT.

Ainsi, rechercher n'importe quel texte anglais disponible qui a un équivalent persan à n'importe quel degré, et utiliser des chefs-d'œuvre littéraires libres de droits était la solution adopter pour la création de Mizan le corpus parallèle persan-anglais aligné manuellement contenant 1 million de paires de phrases avec 25 millions de termes, plus précisément 1 021 596 paires de phrases persan-anglais uniques, est publié dans deux fichiers encodés en Unicode.

Chaque fichier contient des phrases dans une langue, chaque ligne de fichiers représente une phrase et les phrases sont comptées en numéros de ligne. [34]

Mizan est disponible sur : https://github.com/omidkashefi/Mizan/

- **SPC**

Il s'agit d'une collection de corpus parallèles rassemblés par Hercules Dalianis et son groupe de recherche pour la construction de dictionnaires bilingues.

Il contient 4 langages (afrikaans - English - Grec - chinois). Ce corpus est disponible sur : https://opus.nlpl.eu/SPC.php

- **JRC-Acquis** [35]

JRC-Acquis Le corpus JRC-Acquis ou (Joint Research Centre) est une ressource considérable de textes parallèles. Il s'agit des acquis communautaires7 de l'union européenne (UE). La version 3.0 de ce corpus8 est disponible en 22 langues officielles de l'union européenne. Le JRC-Acquis comprend 4,4 millions de documents alignés (dans toutes les langues).

Ce corpus est disponible sur : https://opus.nlpl.eu/JRC-Acquis.php

Pour avoir un corpus multi domaine riche en vocabulaire et de bonne qualité, nous avons commencé par extraire des couples de phrases à partir de cinq corpus volumineux de divers domaines et langues. Les data set sont disponibles sous format txt et tmx, nous les avons pris sous leur format txt pour faciliter le travail de traduction, à savoir que chaque corpus comprend deux txt séparés le premier contient les phrases en langue initiale et le deuxième contient les mêmes phrases en langue de traduction (deuxième langue).

Nous les avons bien choisis en considérant leur large taille, et les langues donnant de meilleurs résultats dans l'étape suivante qui est la traduction en langue arabe.

## 3 TRADUCTION DES COUPLES DE PHRASES EN ARABES

Plusieurs méthodes ont été testées pour sortir de cette phase avec un maximum de couples de phrases arabes correctes et pertinentes et bien traduites servant de couples candidates.

Vu le nombre énorme de couples des phrases que nous avons à traduire -plus d'un million et demi couple de phrases-, nous avons cherché la méthode la plus fiable et rapide, nous avons tout d'abord essayé avec des codes python mais le processus

prenait beaucoup de temps pour traduire un mille couples, cependant nous avons opté pour un logiciel de division de corpus (G-Split) puis découper nos corpus en fichier (. Txt) de 9 mo pour pouvoir les traduire en utilisant l'API de Google.

La figure 3 donne une vision plus claire sur le processus de traduction.

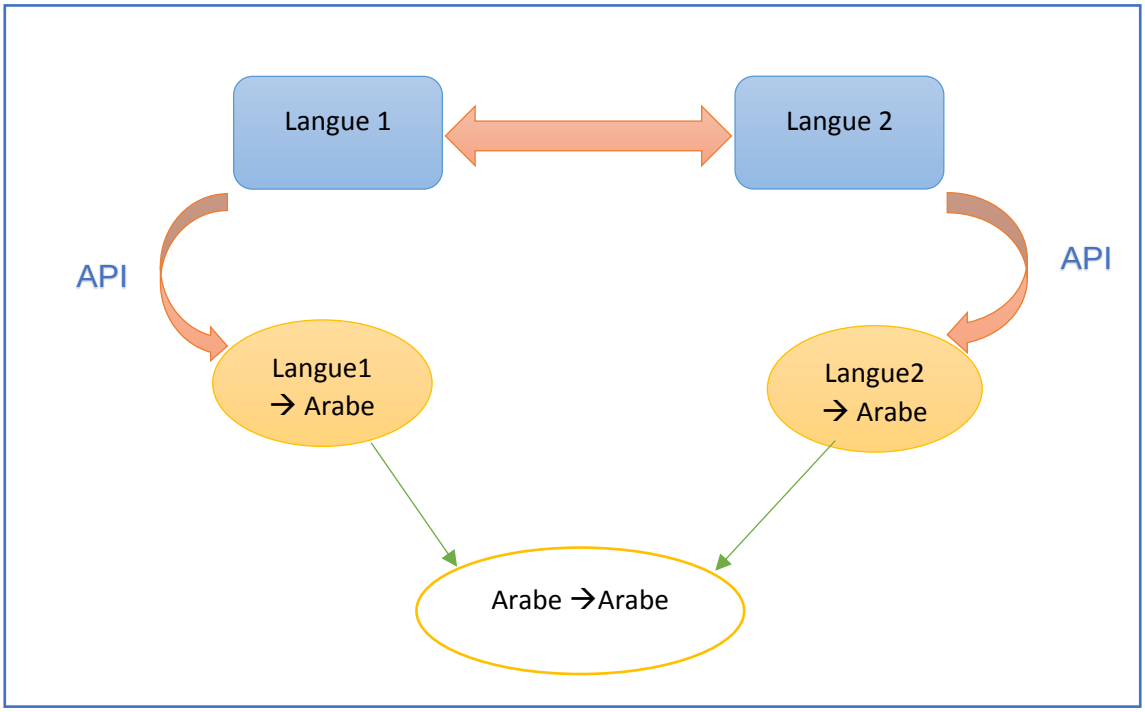


Figure 3 Aperçu d'alignement entre langue pour la génération du corpus arabe.

En fin de cette phase nous obtiendrons deux fichiers txt le premier servira d'ensemble de phrases initiales et le deuxième servira d'ensemble de paraphrases que nous regroupons par la suite dans un seul txt.

Pour s'assurer de la bonne qualité des couples de phrases résultantes de cette phase un appel à une étape de filtrage et d'évaluation est très nécessaire.

## 4 PRETRAITEMENT ET FILTRAGE DU CORPUS

Le prétraitement est dédié pour le nettoyage et l'édition du corpus obtenu de la phase précédente, dans le but de générer des couples de phrases de format plus adapté pour son utilisation dans les prochaines étapes.

La figure suivante montre en détails le processus de prétraitement du corpus.

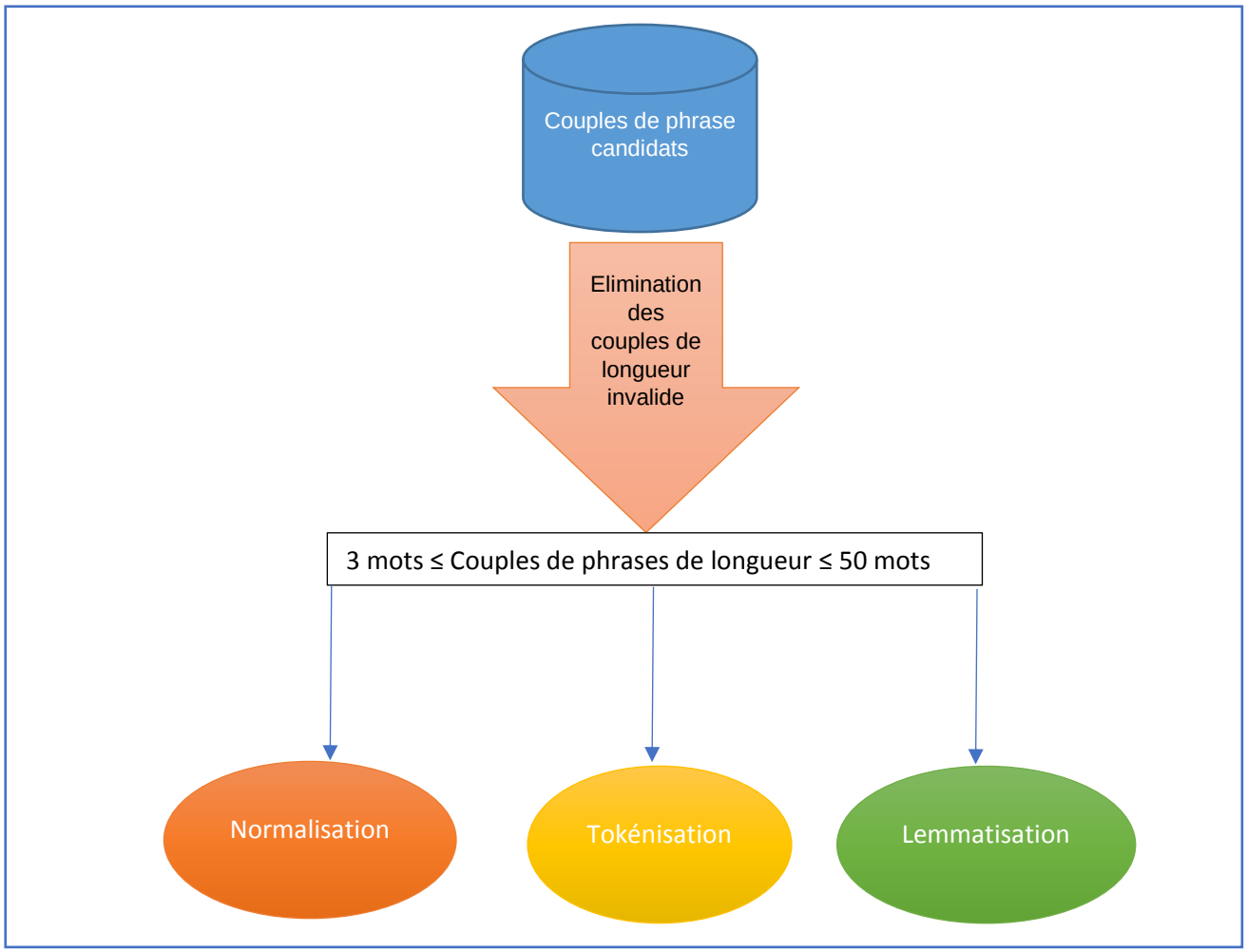


Figure 4 Prétraitement du corpus

Nous détaillons cette phase étape par étape dans ce qui suit.

### 4.1 Elimination des phrases longues

Pour générer des paires de phrases candidates, nous avons d'abord supprimé toute phrase jugée trop courte (moins de cinq mots) ou trop longue (plus de cinquante mots). Ce critère est adapté de la méthodologie adoptée pour un modèle d'entrainement crée dans un travail précédent -nous détaillons cette partie dans le chapitre suivant-.

### 4.2 Normalisation

Pour garantir la bonne qualité des couples de phrases de notre data set nous devons en premier lieu le faire passer par un processus de génération d'une forme canonique des mots pour maximiser la correspondance entre les termes de requête et un autre ensemble de mots dans le document, ce processus s'appelle « la normalisation ».

Dans sa forme simple, la normalisation prétraite les mots en une seule forme, mais elle est très légère. Cela se fait généralement en plusieurs étapes de prétraitement afin de rendre les différentes formes d'une lettre spécifique sous la forme d'une seule représentation Unicode, par exemple, en remplaçant la lettre arabe non pointillée par une lettre finale en pointillé, lorsque cette lettre apparaît à la fin d'un mot arabe. [36]

Plusieurs étapes de nettoyage et de normalisation du corpus combiné ont été appliquer pour texte collecté nous notons :

- Nettoyage des caractères bruyants, des balises et suppression des signes diacritiques.
- Normalisation des caractères arabes : la normalisation de ( ا à أ, إ, آ) et ( ه à ة).
- StopWords : une base de données qui contient 841mots qui sont considérés comme des mots d'arrêt pour la langue arabe comme : « ... على, هو, ي ه»

  **Exemple :**

Sentence=' (DE) قواعد تحديث حول الأبيض الكتاب مع ، والسادة السيدات ، والسادة السيدات ، الرئيس السيد المعنية الدوائر في وحيوية مكثفة مناقشة المفوضية أثارت ، الأوروبية المنافسة.'

Après la normalisation

Sentence=' DE اثارت الاوروبيه المنافسه قواعد تحديث حول الابيض الكتاب مع والساده السيدات الرئيس السيد المعنيه الدوائر في وحيويه مكثفه مناقشه المفوضيه'

Après l'élimination des stopwords

Sentence=' DE المفوضيه اثارت الاوروبيه المنافسه قواعد تحديث الابيض الكتاب والساده السيدات الرئيس السيد المعنيه الدوائر وحيويه مكثفه مناقشه'

### 4.3 Tokenisation

La tokenisation est une méthode pour diviser un morceau de texte – de phrase dans notre cas- en unités plus petites appelées jetons. Les jetons ici peuvent être des mots, des caractères ou des sous-mots. Par conséquent, la tokenisation peut être grossièrement divisée en 3 types : la tokenisation des mots, des caractères et des sous-mots.

**Exemple :**

استراتيجية الاتحاد الأوروبي لمؤتمر نيروبي لتغير المناخ

استراتيجية، الاتحاد، الأوروبي، لمؤتمر، نيروبي، لتغير، المناخ

### 4.4 Lemmatisation

L'arabe est une langue très fléchie et a une structure morphologique complexe. Certaines des applications de le traitement du langage naturel arabe nécessite les bases forme du mot (racine ou radical) pour être la plus efficace [37], par conséquent, nous devons appliqué une procédure sur les phrases arabes candidates ramenant un mot portant des marques de flexion (par exemple, la forme conjuguée d'un verbe) à sa forme de référence (dite lemme), quelle que soit la forme sous laquelle le mot apparaît dans un texte. La lemmatisation -ou le stemming- sert ainsi à la reconnaissance morphologique des mots d'un texte.

« Les formes radicales du même mot sont généralement problématiques pour l'analyse des données textuelles, car elles ont une orthographe différente et une signification similaire, donc la radicalisation est un processus de transformation d'un mot en sa racine (forme normalisée) ».[38]

- Stemming lourd -**root stemming-** : Cela inclut généralement la suppression des préfixes et des suffixes connus. Il vise à retourner la racine réelle du mot, il inclut implicitement le stemming léger.
- Stemming léger -**light stemming-** : C'est le type le moins compliqué et implique de supprimer l'affixe sans essayer de retourner la racine du mot.

## 5 CALCUL DE SIMILARITE

La notion de similarité entre textes est très souvent utilisée dans les applications du traitement de la langue destinée à l'exploitation de collections de documents de grande taille. Par exemple, en recherche documentaire, les documents pertinents retournées par le moteur de recherche peuvent être d´définis comme les plus proches de la requête selon une certaine mesure de similarité ; de même, le regroupement incrémental de documents en classes en fonction de leurs similarités peut permettre

une structuration automatique de bases de données textuelles à l'aide de techniques de classification automatique non supervisée.[39]

Deux types de similarités sont sollicitées dans ce travail et détaillés dans la figure 5.

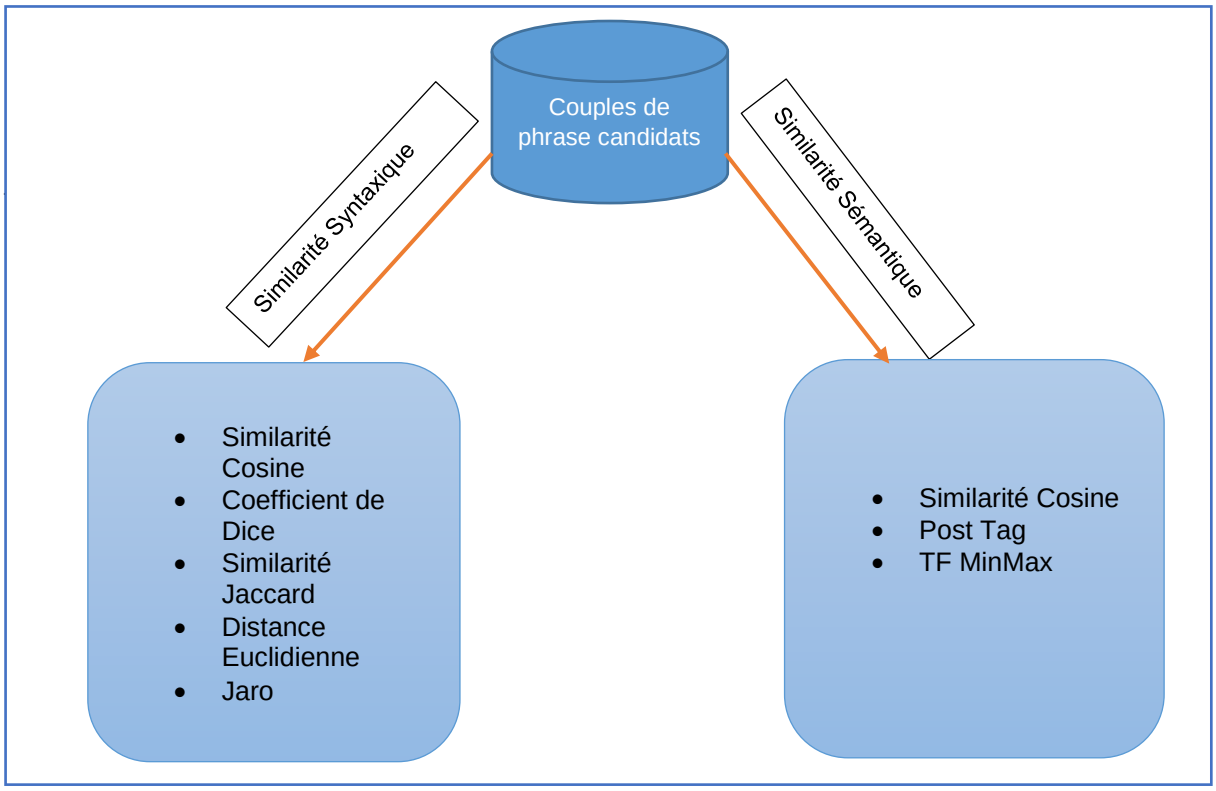


Figure 5 Calculs de similarité

### 5.1 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 dissimilarité entre deux chaînes de caractères. Par exemple, les chaînes de caractères "طالع" et "المطالعة" peuvent être considérées comme similaires alors que "طالع" et "قرأ" pourront être considérées comme très différents.

Une telle mesure sur les chaines de caractères fournit une valeur obtenue algorithmiquement. Parmi de telles mesures de similarité, citons par exemple, la distance de Levenshtein (ou distance d'édition), le coefficient de Dice, l'indice de Jaccard, la distance euclidienne, le cosinus, ... [40]

Dans ce chapitre, nous présentons les mesures de similarité syntaxique que nous avons utilisées dans notre projet.

- **Avantages de l’approche syntaxique**

Les techniques basées sur l'approche syntaxique ne laissent pas de place aux exceptions, elles sont donc facilement automatisables.

- **Inconvénients de l’approche syntaxique**

En appliquant ces approches nous remarquons que les relations syntaxiques sont ignorées. Dans le contexte de notre étude, les relations syntaxiques peuvent influer sur la pertinence. Par conséquent, il faudrait trouver un moyen d'incorporer des techniques d'analyse de variation du texte, De même, les rôles sémantiques sont ignorés. Par exemple, dans "الرئيس الجزائري يستقبل الرئيس التونسي" et " الرئيس التونسي يستقبل الرئيس الجزائري", seule la forme verbale change. Cela peut engendrer des problèmes de pertinence. Une proposition serait peut-être d'analyser les classes verbales.

- **Similarité Cosinus**

La similarité cosinus, ou similarité cosinus, est une technique heuristique pour la mesure de similarité entre deux vecteurs effectués en calculant le cosinus de l'angle entre eux, généralement utilisé pour la comparaison des textes datamining et d’analyse de texte.

$$sim_{cosinus}(d_1, d_2) = \frac{\overrightarrow{d_1}.\overrightarrow{d_2}}{\|\overrightarrow{d_1}\|\|\overrightarrow{d_2}\|}$$

- **Coefficient de Dice**

Ce coefficient est utilisé en statistique pour déterminer la similarité entre deux échantillons. La formule ci-dessous résume la méthode de calcul :

$$S = \frac{2 * |\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|}$$

- **Similarité de Jaccard**

La similarité de Jaccard également appelée indice Jaccard ou coefficient de similarité Jaccard, est un terme inventé par Paul Jaccard en 1901, mesurant les similitudes entre les ensembles. Étant donné deux ensembles, A et B, la similarité Jaccard est définie comme la taille de l’intersection de l'ensemble A et de l'ensemble B (le nombre d'éléments communs) sur la taille de l'union de l'ensemble A et de l'ensemble B (le nombre d'éléments uniques).

La similarité Jaccard est calculée à l'aide de la formule suivante :

$$(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

La bibliothèque contient à la fois des procédures et des fonctions pour calculer la similarité entre des ensembles de données. La fonction est mieux utilisée lors du calcul de la similarité entre un petit nombre d'ensembles. Les procédures parallélisent le calcul et sont donc plus appropriées pour calculer des similitudes sur des ensembles de données plus importants.

**Exemple**

Soit :

D1 ={هل تريد أن تحضر لها بعض الملابس}

D2= {هل تريد أن تعطيها بعض الملابس}

Union (D1, D2) = { 'ط', 'س','ع','م' , 'ب' ,'ض', 'ح' , 'ن', 'ا','ر','ي' ,'د' ,'ت' ,'ل' ,'ه' }

Card (union (D1, D2)) = 15

Intersection (D1, D2) = { 'س','ع','م' , 'ب' ,'ض', 'ن' , 'ا','ر','ي' ,'د' ,'ت' ,'ل' ,'ه' }

Card (intersection (D1, D2)) = 13

**J (D1, D2) = 13/15 = 0.86**

- **Performance des measures** [40]

  Huang dans [41] et Strehl et al. dans [42] ont tous les deux montré que les performances de la similarité cosinus, du coeffcient de Jaccard sont très proches et qu'elles sont significativement meilleures que celles de la distance euclidienne. Cependant, Bavi et al. dans [43] fait apparaitre que plus le document est de petite taille, meilleurs sont les résultats obtenus avec la distance euclidienne, tandis qu'ils sont plus mauvais avec la similarité cosinus ou avec le coefficient de Jaccard.

- **Jaro**

L'algorithme Jaro a pour but de comparer deux chaînes lorsque la similitude des caractères dans les chaînes est une priorité, autrement dit c'est une technique basée caractères utilisée pour la détection des doublons.

Si les deux caractères identiques de s1 et s2 ne diffèrent pas de plus de :

$$\left(\frac{\mathrm{MAX}(|S1|,|S2|)}{2}\right)-1$$

Dans leurs chaînes respectives, ils sont considérés comme identiques, en sachant que La distance de Jaro entre chaînes $s_1$ et $s_2$ est définie par :

$$dj=\frac{1}{3}\left(\frac{m}{|\mathbf{s1}|}+\frac{m}{|\mathbf{s2}|}\ \frac{m-t}{m}\right)$$

Tels que : **|si|** est la longueur de la chaîne de caractères **si**, **m** est le nombre de caractères correspondants, et **t** est le nombre de transpositions.

**Exemple**

| | د | ي | ع | س |
|---|---|---|---|---|
| ب | 0 | 0 | 0 | 0 |
| ع | 0 | 0 | 1 | 0 |
| ي | 0 | 1 | 0 | 0 |
| د | 1 | 0 | 0 | 0 |

M= 3, (|s1| = 4 ; |s2| = 4), t =0/2.

**Dj = 1/3 (3/4 + 3/4 + 3/3) = 0.833**

- **Distance euclidienne**

Une technique calculant la similarité entre deux documents *d*1 et *d*2 comme la distance entre leurs représentations vectorielles ramenées à un seul point.

$$Sim(d1,d2)=\left|\left|\overrightarrow{d1}-\overrightarrow{d2}\right|\right|=\sqrt{\sum_{i=1}^{n}(d1i-d2i)^2}$$

Le tableau 2 Représente un exemple de calcul des différents types de similarité syntaxique sur deux couples de phrases candidates de notre corpus.

Tableau 2 Exemple de calculs de similarité syntaxique.

| Phrase | Paraphrase | Euclidien | Cosine | Jaro | Dice | Jaccard |
|---|---|---|---|---|---|---|
| لقد طلبت مناقشة حول هذا الموضوع خلال الأيام القليلة القادمة ، خلال هذه الجلسة الجزئية | لقد رغبت في إجراء مناقشة حول هذا الموضوع في الأيام المقبلة ، خلال هذه الجلسة الجزئي | 0.74 | 0.61 | 0.78 | 0.77 | 0.8 |
| سوف تكون على علم من الصحافة والتلفزيون أنه كان هناك عدد من التفجيرات والقتل في سريلانكا | ربما تكون قد سمعت من الصحافة والتلفزيون عن وقوع عدة تفجيرات وجرائم في سريلانك | 0.7 | 0.68 | 0.74 | 0.67 | 0.81 |

### 5.2 Similarité sémantique

Les mesures de similarité sémantique décrivent un concept dans lequel un ensemble de documents ou de termes se voit attribuer une mesure basée sur la similarité de leur contenu sémantique [40]. Ce sont des fonctions très utilisées dans plusieurs domaines de l'informatique parmi lesquels nous pouvons citer le Traitement
Automatique du Langage Naturel (TALN), la Bio-informatique, la Recherche d'Information...
Elles permettent de déterminer la similarité entre des termes ou concepts qui n'ont aucune ressemblance syntaxique. Leurs utilisations reposent généralement sur une bonne organisation des documents en structure hiérarchique grâce `à l'utilisation de bases de connaissances comme les ontologies.[44]

- **Approches de similarité sémantique**

Dans notre projet nous ne nous intéressons qu'à l'approche statistique plus précisément l'utilisation des word-embeddings.
Parmi les approches statistiques -basées corpus- avec lesquels on va combiner nos approches syntaxiques de notre travail, c'est les approches qui utilisent les **WORD EMBEDDINGS (WE)** :

Le Word Embedding est un ensemble de techniques, ayant pour but le mappage de mots en vecteurs de nombres réels, dans un espace dimensionnel réduit [45]. Cette méthode permet de capturer la similarité sémantique, ainsi que le contexte d'un mot dans un corpus donné.

Il existe plusieurs techniques de Word Embedding. Nous nous intéressons au Word2Vec, plus précisément au Skip-gram [45] .

- **Word2Vec :** Cet algorithme de Word embedding est parmi les plus connus. Il a été développé par une équipe de recherche de Google sous la direction de Tomas Mikolov. Il s'appuie sur un réseau de neurones à deux couches et essaie d'apprendre la représentation vectorielle des mots qui composent le texte, de sorte que les mots partageant un contexte similaire soient représentés par des vecteurs numériques proches.

Word2Vec a de différents paramètres, dont les plus importants sont :

- La dimensionnalité de l'espace vectoriel à construire, c'est à dire le nombre de descripteurs numériques utilisés pour décrire les mots (entre 100 et 1000 en général).
- La taille du contexte d'un mot, c'est à dire le nombre de termes entourant le mot en question (les auteurs suggèrent d'utiliser des contextes de taille 10 avec l'architecture Skip-Gram).

Étant donné que Word2Vec n'est composé que de deux couches, cet algorithme est rapide à entraîner et à exécuter, ce qui est un gros avantage par rapport à d'autre méthodes de Word embedding.

- **Skip-gram « SG » :** dans le modèle Skip-gram, la représentation distribuée du mot d'entrée est utilisée pour prédire le contexte. En effet, la couche d'entrée correspond au mot cible et la couche de sortie correspond au contexte. Il vise la maximisation de l'équation ci-dessous :

$$\frac{1}{v} \sum_{t=1}^{v} \sum_{j=t-c, j\neq t}^{t+c} \log p(mj|mt)$$

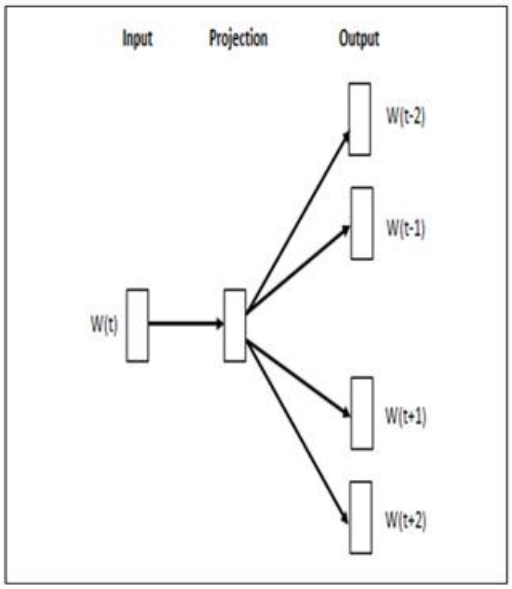


Figure 6 SkipGram

- **Pondération des termes**

C’est l’ajout des poids aux mots de deux manières, la fréquence des mots selon leur apparition dans le corpus spécifique de domaine, ainsi que le type linguistique du mot (verbe, nom, etc.…). [46]

Dans notre travail nous avons recouru à la méthode de pondération **TF-MinMax** qui est généré à partir de corpus du domaine étudié et la partie du baliseur de discours **PosTag**.

Nous utilisons "StanfordCoreNlp" pour la représentation et le passage des mots Ensuite, nous pouvons attribuer des poids à chaque type d'étiquette afin que nous puissions Utilisez-les comme poids dans notre calcul de similarité.

**TF-IDF :**

Le tf-idf permet d'évaluer l'importance des termes contenus dans le document par rapport à la collection. Le poids augmente proportionnellement au nombre de fois où le mot apparaît dans le document. Elle dépend aussi de la fréquence d'apparition des mots dans l'ensemble. Par conséquent, la fréquence de document inverse (idf) est une mesure de l'importance du terme dans l'ensemble de documents. Dans le cas d tf-idf, dont le but est de donner plus de poids aux termes moins fréquents, est considéré comme le plus discriminant.

Il s'agit de calculer le logarithme de l'inverse de la proportion de documents qui contiennent le terme :

**Idfi = log (|D| / | {dj : ti appartient à dj} |)** où :

|D| est le nombre total de documents

| {Dj : ti appartient à dj} | est le nombre de documents où le terme ti apparait.

Finalement, le poids s'obtient en multipliant les deux mesures :

**Tfidf i, j = tf i,j * idfi.**

- **TF-MinMax :**

La notion des **TF- MinMax (**fréquence des termes) est une méthode de pondération utilisé souvent dans le domaine de recherche d'information, elle permet d'évaluer l'importance d'un terme contenu dans un document (un texte) relativement à une collection ou un corpus donné. La pondération **TF-minmax** représente le nombre de fois qu'un mot apparaît dans un document, son poids est calculé en fonction de sa fréquence dans le corpus, il s'augmente proportionnellement au nombre d'occurrences du mot dans le document.

Nous obtenons la normalisation minmax de TF pour chaque mot en utilisant la formule suivante :

$$TF_{min-max} = \frac{TF_{log}}{Max(TF_{log})}$$

Où est le nombre de fois où le mot donné apparaît dans le corpus, et est le nombre de fois où le mot le plus fréquemment utilisé dans le corpus y apparaît.

- **PosTag :**

Part of speech tagging signifie « la partie du baliseur de discours » est un processus qui nous permet d'extraire les informations grammaticales importantes d'une phrase (marquage grammatical d'un mot (ou jetons) dans un texte) telle que le type des mots (verbe, nom, adjectif, etc.). Un mot peut avoir plus d'une partie du discours en fonction du contexte dans lequel il est utilisé. C'est pour cela que cette tâche n'est pas simple.

Le tableau 3 Représente un exemple de calcul des différents types de similarité sémantique sur deux couples de phrases candidates de notre corpus.

Tableau 3 Exemple de calculs de similarité sémantique

| Phrase | Paraphrase | Cosine | Tfmin-max | Postag |
|---|---|---|---|---|
| لقد طلبت مناقشة حول هذا الموضوع خلال الأيام القليلة القادمة ، خلال هذه الجلسة الجزئية | لقد رغبت في إجراء مناقشة حول هذا الموضوع في الأيام المقبلة ، خلال هذه الجلسة الجزئي | 0.96044304 | 0.81541411 | 0.87393501 |
| سوف تكون على علم من الصحافة والتلفزيون أنه كان هناك عدد من التفجيرات والقتل في سريلانكا | ربما تكون قد سمعت من الصحافة والتلفزيون عن وقوع عدة تفجيرات وجرائم في سريلانك | 0.92366268 | 0.81989801 | 0.8189091 |

Dans le but d'une évaluation plus crédible de la qualité de nos couples de phrases candidates nous avons calculer un ensemble de moyenne que nous détaillons dans cette section.

### 5.3 La moyenne

La moyenne est un outil de calcul permettant de résumer une liste de valeurs numériques en un seul nombre réel, indépendamment de l'ordre dans lequel la liste est donnée.

De façon générale, on peut résumer la moyenne comme étant une donnée qui représente les données, il existe également différentes méthodes pour calculer une moyenne. En voici quelques exemples :

#### 5.3.1 La moyenne Arithmétique

La moyenne arithmétique est la *moyenne* ordinaire, c'est-à-dire la somme des valeurs numériques (de la liste) divisée par le *nombre* de ces valeurs numériques.

La « moyenne » se note « $\bar{\mathbf{X}}$ » (x barre) on lira : Si la variable statistique est donnée sous forme d'une série $x_1, x_2, \ldots\ldots, x_n$, la moyenne arithmétique est à la somme des « $x_i$ » divisée par le nombre « n » (« n » étant légal au nombre de « x » de la série).

La moyenne arithmétique est égale au rapport :

$$\bar{\mathbf{X}} = \frac{x\,1 + x\,2 + .. + x\,n}{\mathbf{n}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{xi}$$

En appliquant la moyenne arithmétique sur nos valeurs des calculs de similarités Syntaxique nous obtenons :

« $x_1 = x_{1(Sim\ Jaccard)}$, $x_2 = x_{1(SimCos)}$ , $x_3 = x_{1(simeuclidien)}$ , $x_4 = x_{1(SimJaro)}$ , $x5 = x_{1(SimDice)}$ avec n=5 ».

En appliquant la moyenne arithmétique sur nos valeurs des calculs de similarités Syntaxique nous obtenons :

« $x_1 = x_{1(SimCos)}$, $x_2 = x_{1(SimTfMinMax)}$, $x_3 = x_{1(SimPosTag)}$ avec n=3 ».

### 5.3.2 La moyenne arithmétique pondérée

La moyenne arithmétique pondérée C'est la moyenne qui est la plus connue après la moyenne arithmétique (simple).

Elle est utilisée quand les valeurs n'ont pas toutes la même importance par rapport au résultat final. Dans ce cas, on donne une pondération (généralement en pourcentage). C'est-à-dire une importance à chacune des valeurs. Par ailleurs, la somme des pondérations doit être de 100%.

Chaque valeur numérique est multipliée par son poids. Les résultats obtenus sont additionnés et la somme obtenue est alors divisée par la somme des poids.

$$M = m1.a1 + m2.a2 + \cdots + mn.an$$

Nous avons calculé la moyenne arithmétique pondérée en utilisant deux pondération différentes tels que :

- La première pondération :

« m1=la moyenne arithmétique syntaxique multiplié par la pondération a1=0.2.
« m2=la moyenne arithmétique sémantique multiplié par la pondération a2 =0.8.

- La deuxième pondération :

« m1=la moyenne arithmétique syntaxique multiplié par la pondération a1=0.3.
« m2=la moyenne arithmétique sémantique multiplié par la pondération a2=0.7.

### 5.3.3 La moyenne harmonique :

On se sert de cette moyenne notamment pour les grandeurs quotients. La moyenne harmonique de N valeurs est le nombre dont l'inverse est la moyenne arithmétique des inverses desdites valeurs.

$$\mathbf{X = 2(\frac{n1.x1}{n1 + x1})}$$

Tel que : n1= la moyenne arithmétique syntaxique

X1= la moyenne arithmétique sémantique

Le tableau 4 Représente un exemple de calcul des différents types de similarité sémantique sur deux couples de phrases candidates de notre corpus.

Tableau 4 Exemple de calculs des moyennes

| Phrase | Paraphrase | M ART syntaxique | M ART sémantique | M Arithmétique | M Harmonique | M Pondérée (0.2;0.8) | M Pondérée (0.3;0.7) |
|---|---|---|---|---|---|---|---|
| لقد طلبت مناقشة حول هذا الموضوع خلال الأيام القليلة القادمة ، خلال هذه الجلسة الجزئية | لقد رغبت في إجراء مناقشة حول هذا الموضوع في الأيام المقبلة ، خلال هذه الجلسة الجزئي | 0.74 | 0.88 | 0.81 | 0.80 | 0.85 | 0.84 |
| سوف تكون على علم من الصحافة والتلفزيون أنه كان هناك عدد من التفجيرات والقتل في سريلانكا | ربما تكون قد سمعت من الصحافة والتلفزيون عن وقوع عدة تفجيرات وجرائم في سريلانك | 0.66 | 0.85 | 0.75 | 0.746 | 0.81 | 0.79 |

En considérant la moyenne calculée avec la première pondération (0.2, 0.8) nous avons organisé notre corpus dans six classes :

- Classe 1 contient les couples qui ont une moyenne Pondérée entre [0, 0.3 [;
- Classe2 contient les couples qui ont une moyenne Pondérée entre [0.3, 0.4 [;
- Classe3 contient les couples qui ont une moyenne Pondérée entre [0.4, 0.5 [;
- Classe4 contient les couples qui ont une moyenne Pondérée entre [0.5, 0.6 [;
- Classe5 contient les couples qui ont une moyenne Pondérée entre [0.6, 0.9 [;
- Classe6 contient les couples qui ont une moyenne Pondérée entre [0.9, 1] ;

Le tableau ci-dessous donne une vision sur le partitionnement des couples de phrases dans les DataSet selon les classes

Tableau 5 partitionnement des DataSet

| Data Set | Europarl (fr-an) | Europarl(fr-it) | Mizan | Jrc | Spc |
|---|---|---|---|---|---|
| Classe1 | 0 | 0,01% | 3% | 0% | 0,50% |
| Classe2 | 0,20% | 0,01% | 4% | 0% | 0% |
| Classe3 | 0,30% | 0,08% | 7% | 0,01% | 0,50% |
| Classe4 | 0,50% | 0,4 | 12% | 0,40% | 1% |
| Classe5 | 9% | 44% | 69% | 2% | 48% |
| Classe6 | 91% | 55% | 5% | 97% | 50% |

A noter que les deux premières classes contiennent des couples qui sont complètement différents syntaxiquement et sémantiquement.

- **Example**

Phrase = ' والاستثناءات الغطاء نطاق '

Paraphrase = ' والمسؤولية التأمين مسؤولية من الإعفاء '

Les trois autres classes contiennent des vraies paraphrases.

- **Example**

Pharse = ' المقبولية عدم بشأن 143 بالمادة يتعلق فيما رأيك إبداء أود'

Paraphrase = ' المقبولية بعدم المتعلقة 143 القاعدة بشأن المشورة عن أسألكم أن أود.'

La dernière classe contient des vraies paraphrases et quelque phrase qui sont complétement identique.

- **Example :**

Phrase = ' سرور بكل ذلك سأفعل ، سجني سيد لك شكرا.'

Paraphrase = ' سرور بكل ذلك سأفعل ، سجني سيد لك شكرا.'

## 6 STATISTIQUE DU DATA SET

Le tableau suivant donne des statistiques sur le data set généré.

Tableau 6 Statistiques du Data Set

| Data Set | Europarl (fr-an) | Europarl (fr-it) | Mizan | Jrc | Spc | Corpus |
|---|---|---|---|---|---|---|
| Nombre de mots | Entre 3 et 50 | Entre 3 et 50 | Entre 3 et 50 | Entre et 50 | Entre 3 et 50 | Entre 3 et 50 |
| Nbr de couples Classe1 | 0 | 25 | 6631 | 0 | 3 | 6659 |
| Nbr de couples Classe2 | 5 | 35 | 9038 | 2 | 0 | 9080 |
| Nbr d couples Classe3 | 104 | 206 | 16352 | 25 | 6 | 16693 |
| Nbr de couples Classe4 | 295 | 813 | 28153 | 73 | 20 | 29354 |
| Nbr de couples Classe5 | 56910 | 107401 | 158627 | 7855 | 942 | 331735 |
| Nbr de couples Classe6 | 589796 | 137137 | 11111 | 345709 | 963 | 1084716 |
| | | | | | Total : | 1478237 |

## 7 Conclusion

Dans ce chapitre nous avons commencé notre création du corpus parallèle monolingue en langue arabe, que nous avons entrepris par la collection de données, ces dernières sont des couples de phrases de différentes langues, nous les avons traduits en langues arabe. Enfin nous avons évalué les couples de phrases arabes résultants.

Dans le chapitre suivant nous allons faire une évaluation qualitative et quantitative du dataset.

# CHAPITRE 3 : EVALUATION QUALITATIVE ET QUANTITATIVE DU CORPUS

## 1 INTRODUCTION

Pour évaluer le travail réalisé nous avons effectué deux évaluations :

- Une évaluation qualitative manuelle intrinsèque sur le corpus construit,
- Une évaluation extrinsèque du corpus par rapport à la tâche de paraphrase.

Nous présentons dans la suite les deux évaluations :

### 1.1 Evaluation qualitative manuelle intrinsèque sur le corpus construit

Pour évaluer la qualité des ensembles de données générés, une expertise humaine est nécessaire.

Nous avons sélectionné un échantillon aléatoire de paires de phrases pour une évaluation humaine qualitative faite par deux experts en langue arabe (voir annexe N° **1).**

Chaque annotateur doit définir la similarité entre les paires sélectionnées dans l'échantillon dans l'une des classes apparaissant la figure 7 :

| | | |
|---|---|---|
| Classe 1 | Les deux phrases sont complètement semblables et donc paraphrases | الجملتان متشابهتان تمامًا |
| Classe 2 | Les deux phrases sont fondamentalement les mêmes (certains détails manquent) | الجملتان متشابهتان أساسًا ( بعض التفاصيل مفقودة) |
| Classe 3 | Les deux phrases partagent quelques détails | الجملتان تشتركان في بعض التفاصيل |
| Classe 4 | Les deux phrases n'ont que le sujet en commun | الجملتان تشتركان في الموضوع فقط |
| Classe 5 | Les deux phrases sont complètement différentes | الجملتان مختلفتان تمامًا |

Figure 7 Analyse des classes

Vue la nature subjective de l'annotation humaine, les deux annotateurs ont une corrélation de 89%.

Pour calculer la corrélation nous avons considéré le pourcentage des mêmes réponses sur l'ensemble de l'échantillon.

L'annotation nous a permis de confirmer les points suivants :

1) A partir du seuil de similarité 0,4 les annotateurs ont considéré la paire dans l'une des deux premières classes que nous considérons des paraphrases.
2) En dessous du seuil, les annotateurs ont annoté dans l'une des 3 dernières classes (3, 4 et 5) que nous ne considérons pas des paraphrases.
3) Nous avons retenu directement à partir du seuil 0,4 des paraphrases.
4) Nous avons confirmé que 299 paires sont considérées comme paraphrases

### 1.2 Evaluation du corpus par rapport à la tâche de paraphrase.

Dans cette section nous allons intégrer les couples de paraphrases issues des processus de traitements et d'évaluations précédents dans un générateur de paraphrases pour constituer de nouvelles paraphrases, nous passons ensuite à l'évaluation automatique des paraphrases générées en utilisant de différentes techniques pour décider de la qualité des paraphrases résultantes.

Le réseau de neurones récurrent qui convertit une séquence de données d'un domaine en entrée vers une nouvelle séquence de données dans un autre domaine en sortie.

Généralement, un modèle séquence à séquence est implémenté en utilisant deux réseaux de neurones récurrents, un premier réseau est un encodeur et le second est un décodeur. On parle ici d'une architecture encodeur-décodeur avec mécanisme d'attention développer dans un travail précèdent. Dans ces modèles, l'entrée et la sortie ne sont pas nécessairement de la même longueur.

La figure 8 montre les étapes de notre travail.

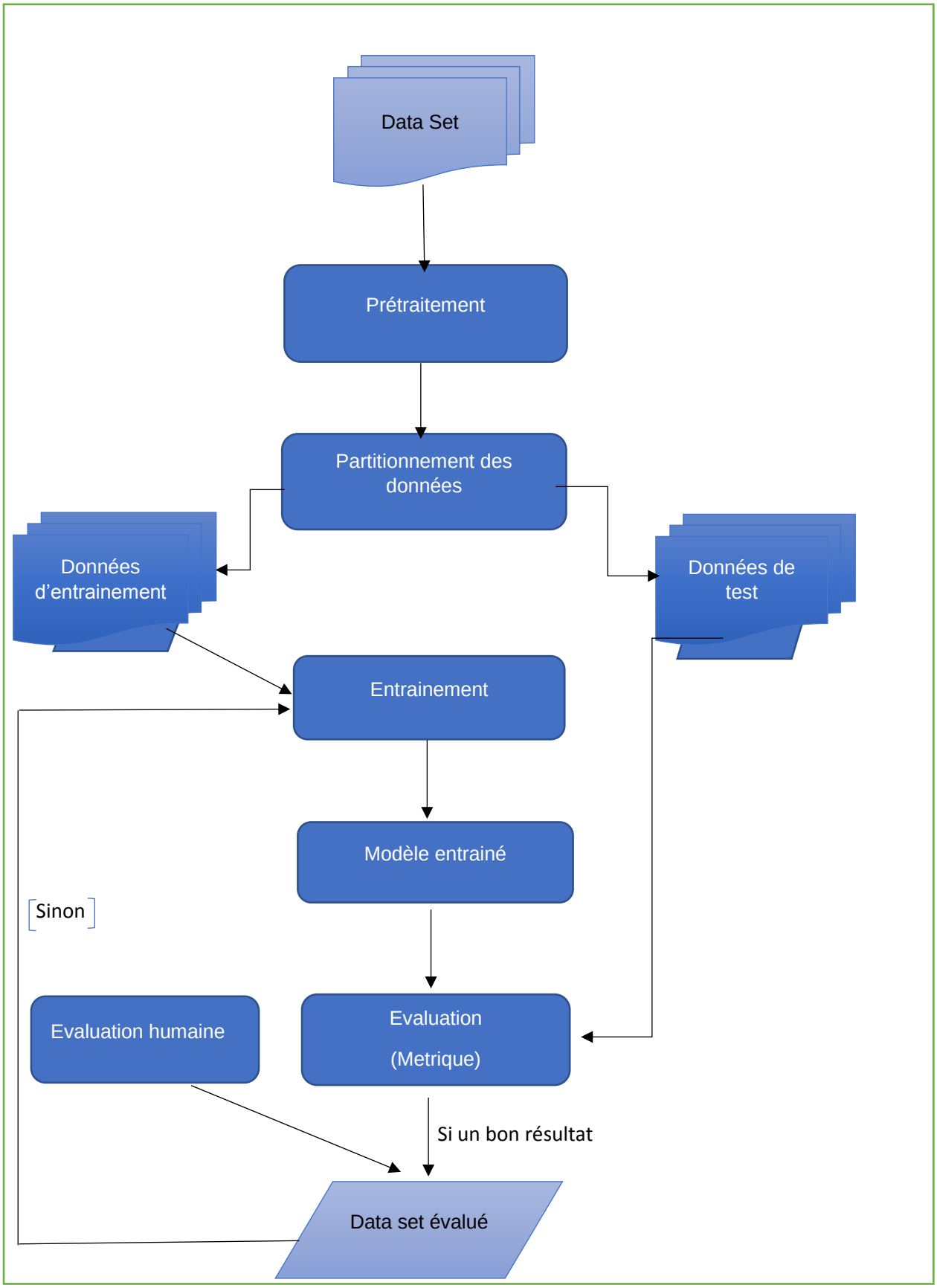


Figure *8* Processus *d'évaluation automatique du corpus.*

### 1.2.1 LECTURE DE DATA SET

Compte tenu du temps nécessaire à l'exécution du programme sur le nombre de paires de phrases que nous avons (un million de phrases) et de la nécessité d'une évaluation fiable, nous avons choisi environ cent mille paires, tirer aléatoirement de notre corpus, nous avons donc eux trois parties contenant des couples de phrases de différentes classes organisées comme suit :

La première partie contient 42 mille couples de phrases, la deuxième contient 35 mille couples tandis que la troisième partie comporte 25 mille couples de phrases candidates.

### 1.2.2 PRETRAITEMENT

Le prétraitement des données est une étape capitale pour le Machine Learning et du Deep Learning. La qualité des données affecte directement la capacité d'apprentissage du modèle. Cette étape comprend la conversion du corpus obtenu de la phase précédente, en format plus adapté et utilisable par le modèle en impliquant les notions de normalisation, tokenisation et lemmatisation expliqué dans le chapitre précédant.

### 1.2.3 PARTITIONNEMENT DE DATA SET

Dans le cadre des processus d'entrainement et de tests des couples de paraphrases candidats, nous avons pris cent milles couples de phrases tels que 80% des couples ont participé dans le processus d'entrainement, et 20% ont été laissé pour la phase de test pour un test plus fiable et crédible.

Lors de l'exécution des codes de générateur de paraphrases sur Google colaboratory, nous nous sommes retrouvées face à un problème de limitation de taille des fichiers d'entrée, ce souci dû aux mises à jour des versions des bibliothèques nous a empêché d'utiliser cent milles couples à la fois, par conséquence nous les avons divisés sur trois parties comme le montre le tableau 7 ci-après.

Tableau 7 Partitionnement du Data Set

| Data Set | Couples d'entrainement | Couples de Test | Total |
|---|---|---|---|
| Partie1 | 33134 | 8033 | 41167 |
| Partie2 | 31739 | 3415 | 35154 |
| Partie3 | 23228 | 2500 | 25728 |

### 1.2.4 ENTRAINEMENT ET TEST

Dans cette étape, nous entamerons les détails de modèles, Il s'agit d'un encodeur-décodeur intégrant le mécanisme d'attention,

Ce modèle est exécuté pour les trois phases d'un Deep Learning à savoir : l'entrainement, la validation et le test.

### 1.2.5 Modèle Encodeur-Décodeur avec mécanisme d'attention

Nous illustrons dans la Figure 9, l'architecture globale de ce modèle. Cette dernière est composée de deux parties distinctes : l'encodeur et le décodeur.

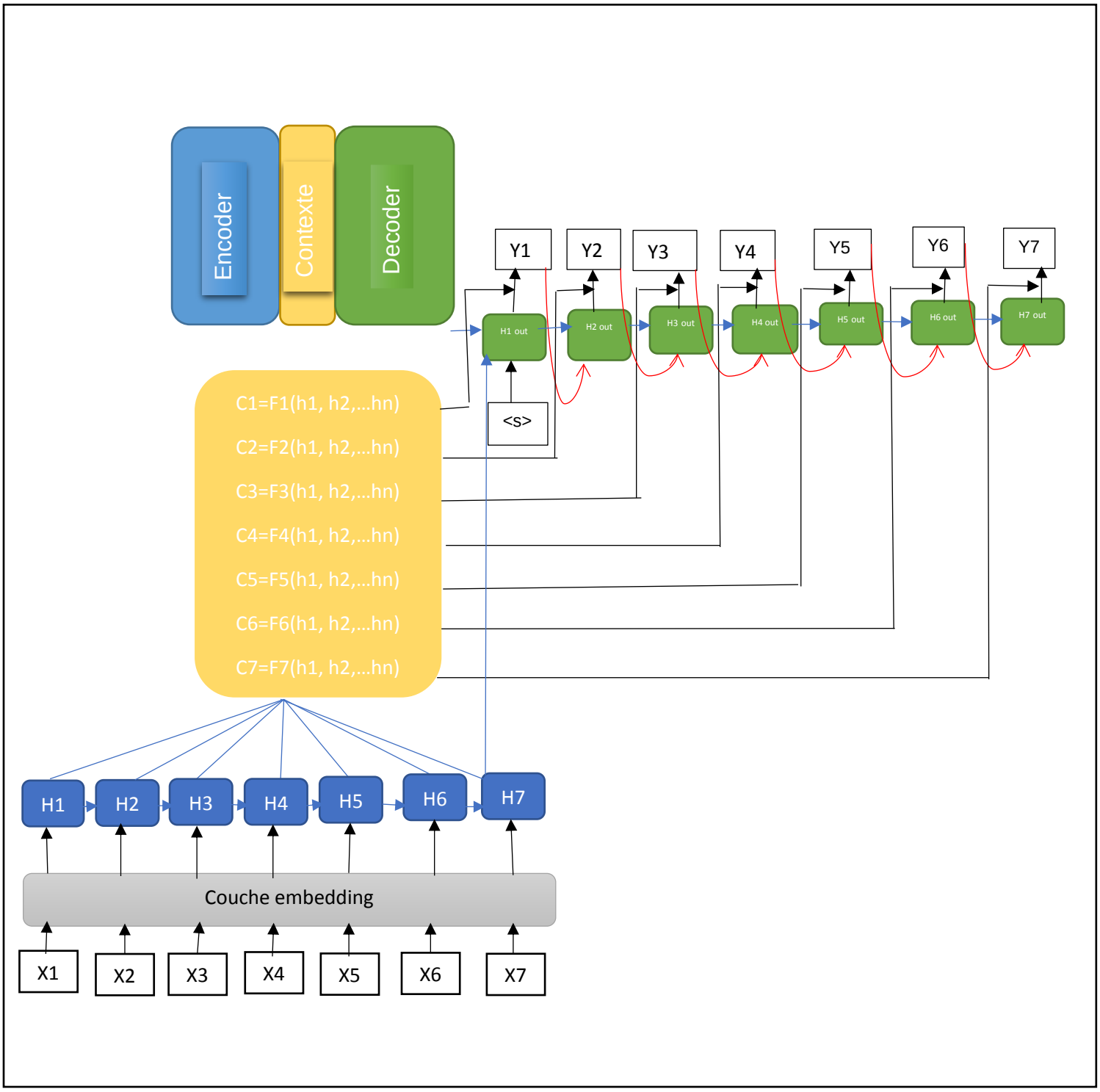


Figure 9 l'architecture globale de l'encodeur-décodeur.

Dans l'architecture codeur décodeur la première séquence est entrée mot à mot
Lorsque la séquence entière est entrée dans l'encodeur celui-ci crée un contexte représenté la séquence d'origine.
Ce contexte est ensuite transmis au réseau décodeur qui l'utilise pour construire la séquence en sortie mot après mot.

- **Encodeur**

L'encodeur est décomposé en deux modules distincts :

L'Embedding et le générateur du vecteur contenant les informations relatives aux entrées ce vecteur est utilisé comme premier état caché du décodeur pour guider le décodeur pour faire des prédictions. Ensuite, nous allons présenter leurs opérations en détail.

Les entrées du réseau sont une représentation vectorielle des mots (one hot encoding, Word embedding…) et non les mots eux même.

- **Couche embedding**

  Avoir effectué tous les prétraitements nécessaires dans la première étape, nous obtenons un ensemble de vecteurs de mots, dont chacun représente une phrase. C'est l'entrée de notre générateur.

  La méthode d'embedding qui généralement utilisé pour réduire la dimension d'un vecteur est d'utiliser le résultat que retourne une dense layer comme embedding, c'est à dire de multiplier une matrice d'embedding W par la représentation « one hot » du mot.

- **Couche GRU**

  Comme mentionné ci-dessus, le RNN souffre de la disparition/explosion des gradients et ne se souvient pas des états pendant très longtemps. Les GRU, sont une application de modules multiplicatifs qui tente de résoudre ces problèmes. C'est un exemple de réseau récurrent avec mémoire (un autre est LSTM). La structure d'une unité GRU est présentée ci-dessous dans la figure 10 [47] .

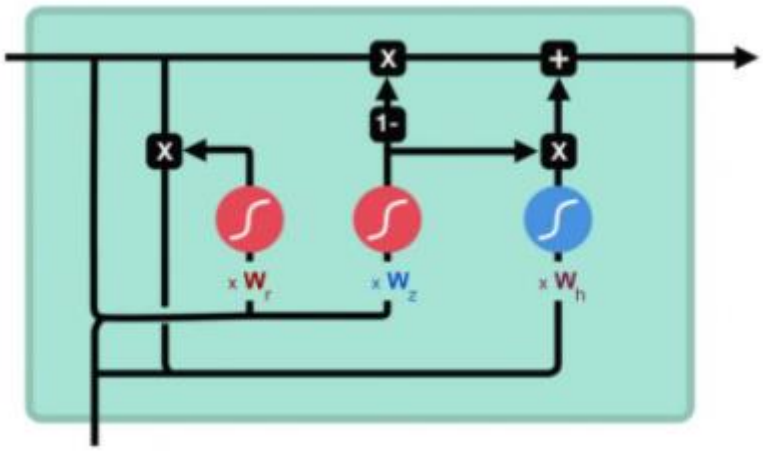


Figure *10 Gated Récurrent Unit*

Cette dernière est composée de

- Wr : pondère l'entrée de la porte de reset (reset gate)
- Wz : pondère l'entrée de la porte de mise à jour (update gate)
- Wh : pondère les données qui vont se combiner pour définir l'état caché courant

- **Porte de reset (reset gate)**

Cette porte sert à contrôler combien d'information passée le réseau doit oublier. L'état caché précédent, concaténé avec les données d'entrée, passe par une sigmoïde (pour ne conserver que les coordonnées pertinentes) puis est multiplié par l'ancien état caché : on n'en conserve donc que les coordonnées importantes (telles qu'elles) de l'état précédent (on a donc perdu une partie de l'état précédent dans cette porte).

- Porte de mise à jour (update gate)

Elle décide des informations à conserver et de celles à oublier.

Les données d'entrées et l'ancien état caché sont concaténés et passent par une fonction sigmoïde dont le rôle est de déterminer quelles sont les composantes importantes.

Les calculs opérés par le GRU sont plus rapides et plus simples. Notons cependant que les capacités/l'efficacité de ce dernier ne sont plus à prouver

Pour générer un mot $y$i, nous devons faire attention à chaque mot de la séquence d'entrée. Il est représenté par le poids au niveau de l'encodeur. Ce dernier génère un score pour chaque état caché, de sorte que l'état caché qui a besoin d'attention aura un score élevé. Après avoir calculé tous les poids d'attention, le vecteur de contexte est calculé selon la formule suivante :

$Vecteur\_contexte = \sum_{i=1}^{n} Pi\ hi$

« $h$i : Etat caché au pas de temps $i$ ;

$P$i: Poids accordé à l'état caché $h$i. »

Le texte d'entrée est traité par l'encodeur pour être codée en un vecteur contexte, le premier vecteur de contexte ($V$i) est transmis avec le dernier état caché de l'encodeur au premier nœud du décodeur ($GRU$i)

- **Gradient**

Pour entraîner un perceptron, c'est-à-dire apprendre les poids de connexion, nous allons chercher à minimiser l'erreur de prédiction sur le jeu d'entraînement. Nous pourrions faire ça de manière explicite.

L'entraînement d'un perceptron est donc un processus itératif. Après chaque observation, nous allons ajuster les poids de connexion de sorte à réduire l'erreur de prédiction faite par le perceptron dans son état actuel. Pour cela, nous allons utiliser l'algorithme du gradient : le gradient nous donnant la direction de plus grande variation d'une fonction (dans notre cas, la fonction d'erreur), pour trouver le minimum de cette fonction il faut se déplacer dans la direction opposée au gradient. (Lorsque la fonction est minimisée localement, son gradient est égal à 0.)

Voici un exemple du résultat de la fonction de gradient qu'on a obtenue sur l'un d'entrainement.

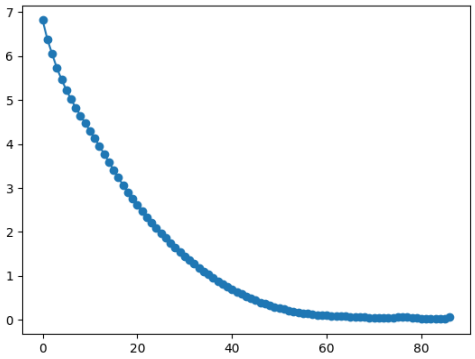


Figure 11 Courbe de gradient

- **Décodeur**

Le décodeur utilise le contexte et le dernier état caché comme état initial, génère le premier mot, puis l'utilise comme entrée pour générer le mot suivant avec le deuxième vecteur de contexte, et ainsi de suite, jusqu'à la fin de la séquence de génération- Il y a deux couches, une couche GRU et une couche Softmax. La couche GRU fonctionne de la même manière que l'encodeur, mais basée sur une seule exception d'E/S Pour la couche SoftMax, il prédit la distribution de probabilité sur l'entier représentant le mot de vocabulaire, et le mot de sortie correspond à l'entier avec la probabilité la plus élevée.

Le dernier état caché du premier réseau RNN est utilisé comme contexte qui représente la phrase d'entrée.

Tableau 8 Exemple de paraphrases générées.

| Phrase origine | Paraphrase | Phrase générée |
|---|---|---|
| إنني اؤيد بشده فكره ان اللجنه ستبحث في العام المقبل عن حل افضل واكثر عمليه واسهل للتحكم | انني اؤيد بشده الفكره القائله بان الهيئه ستسعى في العام القادم الى حل افضل واكثر قابليه للتطبيق ويمكن التحكم فيه بسهوله اكبر | انا اويد تماما ان المفوضيه قد وافق عليه في معاهده زمنيه فعاله والتي من استراتيجيه منظمه التجاره العالميه |
| السيد الرئيس اولا وقبل كل شيء اود ان اشكر مقرر السيده لوكاتيلي الذي قدم ممتازه وتقرير مثير للاهتمام | سيدي الرئيس اود بادئ ذي بدء ان اشكر مقررنا السيده لوكاتيللي التي قدمت تقريرًا ممتازًا وممتعًا | سيدي الرييس اولا وقبل كل شيء اود ان اشكر بحراره المقرر لوكاتيللي على تقريره الممتاز الذي طرحته السيده مان |

- **Interprétation des résultats**

Pour l'évaluation automatique de nos couples de phrases candidats nous avons appliqué un modèle encodeur décodeur avec mécanisme d'attention sur les trois parties de corpus.

En premier lieu nous devons entrainer et tester nos couples de phrases candidates pour cela nous avons utilisé un environnement nommé Google Colaboratory qui est souvent raccourci en "Colab".

Colab est un produit de Google Research, il permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un

environnement particulièrement adapté au machine Learning, à l'analyse de données et à l'éducation. Cet environnement offre la possibilité d'utiliser des accélérateurs graphiques GPU. Pour optimiser le temps d'exécution.

Une fois le modèle est entrainé et testé il passe par une étape d'évaluation en utilisant les métriques définis auparavant dans le chapitre 2. Nous allons discuter cette évaluation.

### 1.2.6 EVALUATION DES PARAPHRASES GENERES PAR LE GENERATEUR -EDAM-

Nous avons utilisé deux techniques :

- Dans cette section nous relevons les résultats obtenus par les métriques d'évaluation automatiques usuelles – que nous avons définis dans le chapitre 1-, nous précisons que nous avons appliqué la méthode « bleu » sur nos trois parties du corpus en raison que seulement les deux métriques « Bleu – Gleu » sont adaptées pour les calculs en langue arabe.

Tableau 9 Résultats des calculs de « Bleu ».

| Data Set | Bleu |
|---|---|
| Partie1 | 0.6746135905242221 |
| Partie2 | 0.6725040521085891 |
| Partie3 | 0.6514574546309663 |

En prenant en considération l'interprétation des score bleu présenté dans le tableau 10 [48], et le score des paraphrases de nos trois parties 67%, 67% et 65% respectivement, nous pouvons déduire que toutes les paraphrases générées sont de qualité souvent meilleure que celle donnée par l'humain.

Tableau 10 Interprétation des scores Bleu [48]

| Score Bleu (%) | Interprétation |
|---|---|
| <10 | Résultat presque inutile. |
| 10 à 19 | L'idée générale est difficilement compréhensible. |
| 20 à 29 | L'idée générale apparait clairement, mais le texte comporte de nombreuses erreurs grammaticales. |
| 30 à 40 | Résultats compréhensibles et correctes. |
| 40 à 50 | Résultat de haute qualité. |
| 50 à 60 | Résultat de très haute qualité, adéquat et fluide. |
| >60 | Qualité souvent meilleure que celle donnée par l'humain. |

Ces résultats seront vérifiés par une deuxième évaluation qualitative humaine faites par des experts sur les paraphrases générées.

- **Evaluation du processus de paraphrase utilisant le Dataset construit**

Pour clôturer notre mission d'évaluation des couples de phrases générées, nous sollicitons des experts humains en langue arabe, pour cette tâche nous avons opté pour une évaluation qualitative en raison que par rapport à l'évaluation quantitative, l'évaluation qualitative a l'avantage d'être plus proche d'une utilisation réelle, mais aussi l'inconvénient d'être difficile à mettre en œuvre, pour des raisons comme le manque de travaux de référence et le problème de la subjectivité des évaluateurs. (Voir annexe 1).

Comme l'évaluation précise des paraphrases est un problème ouvert, l'évaluation automatique ne suffit pas pour évaluer les paraphrases dans une perspective fine, en termes de deux aspects :

- **Relevance (Pertinence en sens) :** exprime la pertinence de la paraphrase générée avec la phrase d'entrée. Ici il est question de noter à quel point la phrase générée préserve le même sens que la phrase originale.

- **Readability (lisibilité en forme) :** la lisibilité de la paraphrase générée en termes de forme, de grammaire sans considérer le sens de la phrase générée.

Pour quantifier ces aspects qui ne sont pas abordés par les métriques d'évaluation automatiques, l'évaluation humaine devient nécessaire pour notre problème à savoir la génération de phrases ayant le même sens qu'une phrase originale.
Nous recueillons donc des jugements humains sur un échantillon de 100 couples de (phrase originale, Phrase générée). Ces couples sont pris de manière aléatoire à partir du jeu de test déjà évalué automatiquement. Les deux aspects Relevance (Pertinence en sens) et Readability (lisibilité en forme) sont vérifiés dans l'évaluation humaine pour chaque couple de l'échantillon. (Voir annexe N° 1**).**

Le travail demandé à l'expert humain consiste à :

- Lire la phrase originale et la phrase générée
- Accorder un score entre 1 & 5 (1 est le pire et 5 est le meilleur) pour l'aspect « Relevance » de la phrase générée par rapport à l'originale
- Accorder un score entre 1 & 5 (1 est le pire et 5 est le meilleur) pour l'aspect « Readability » de la phrase générée.

L'expert humain attribue un score sur une échelle continue de 1 à 5 pour chaque aspect par paraphrase générée, où 1 est le pire et 5 est le meilleur.
Nous avons fait 2 évaluations humaines pour les mêmes couples par deux experts humains volontaires, maitrisant la langue et nous garderons la moyenne des notes obtenues pour chaque couple.

Moyennant la corrélation entre les experts (car l'évaluation humaine reste aussi subjective), nous pouvons ainsi compléter l'évaluation de notre dataset sur ces deux aspects liés à la relevance et la lisibilité.

- **Résultats obtenus :**

Le tableau 11 représente les résultats annoncés par les experts humains

Tableau 11 Résultat Final d'évaluation humaine

| **Dataset** | **Relevance/5** | **Readability/5** |
|---|---|---|
| **Bilingual** | 3.1 | 4 |

Selon les résultats montrés sur le tableau ci-dessus, la pertinence des phrases générées automatiquement par le modèle EDAM est estimée de 62% ce qui nous permet de dire que la sémantique des paraphrases est assez préservée globalement. Cependant, nous pouvons

aussi dire que les phrases générées par le même modèle sont correctes syntaxiquement avec une lisibilité grammaticale moyenne de 80%.

Nous constatons que les annotations des experts arabes sont assez bonnes.

## 2 OUTILS & CHOIX D'IMPLIMENTATION

Pour concrétiser notre travail de recherche, traitant la génération automatique d'un corpus monolingue arabe en utilisant une approche basée sur l'extraction de données à partir des corpus parallèles bilingues, nous avons développé une application desktop, tel que l'interface a été conçue par tkinter qui est une bibliothèque graphique libre d'origine pour le langage python permettant la création des interfaces graphiques, quant à l'ensemble des tâches représentant la partie dynamique de notre application, a été développé en utilisant le langue python, et exécuté sur les deux environnement google colaboratory et pycharm.

Nous avons en premier lieu, une interface ayant deux entrées pour faire appel aux deux fichiers comportant les phrases (dans la première entrée) ainsi que leurs paraphrases (dans la deuxième entrée) sur lesquels nous allons effectuer nos traitements, la seule contrainte sur les fichiers est qu'ils soient bien alignés et sans décalage.

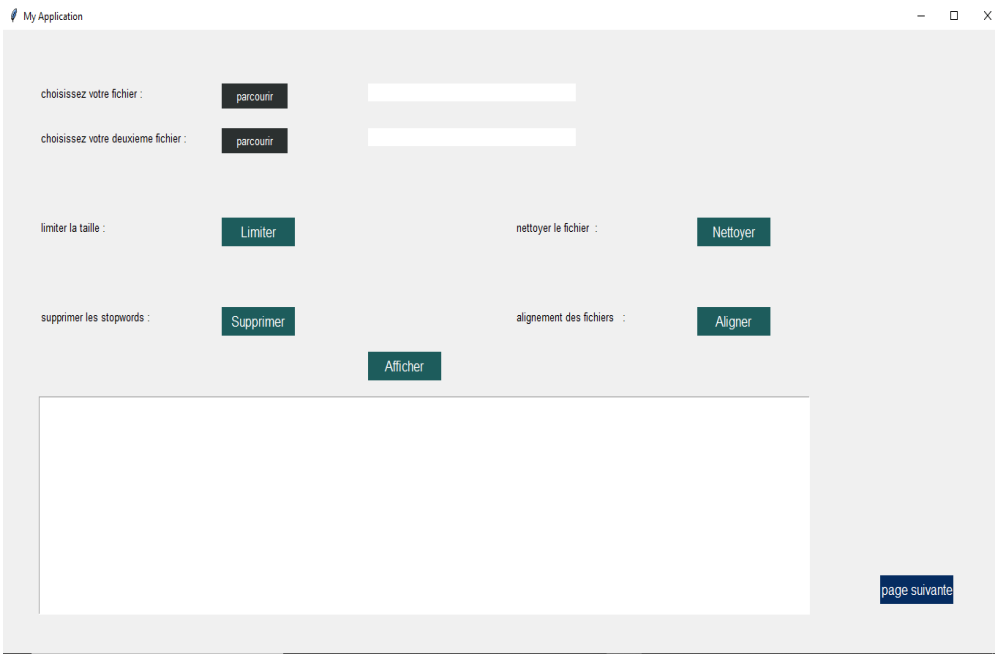


Figure 12 L'interface de prétraitement

En entrant les fichiers du format «. Txt » convenable, l'utilisateur aura un libre droit de choix de traitement (pas forcément en suivant les étapes de génération de corpus), il peut limiter la taille des phrases de ses fichiers (longueur de phrases comprises entre 3 et 50 mots), supprimer les stopwords, nettoyer ou même aligner ses fichiers.

Les résultats des traitements effectués sont affichés en bas de la fenêtre.

La deuxième interface est dédiée pour les calculs de similarités syntaxiques et sémantiques -dont les détails de calculs est expliqué en détails dans le chapitre 3-.

Pour l'ensemble des ces calculs nous nous sommes retrouvées face à un problème de temps énorme que prenaient les codes pour s'exécuter en vue du grand nombre de couples que nous avions à traités (plus d'un million et demi couples),

Pour remédier à ce problème nous avons opter pour l'exécution des codes sur un environnement plus performant qui est le google colaboratory plustot qu'un environnement moins performant tel que pycharm qui exige une certaine qualité de hardware pour pouvoir effectuer les calculs plus rapidement.

Le seul souci avec le google colab était le temps limité d'utilisation du gpu (accélerateur graphique), nous avions seulement douze heures d'utilisation offertes par jour, alors ça nous prenait plusieurs jours pour pouvoir terminer les calculs sur plus d'un million et demi couples candidats.

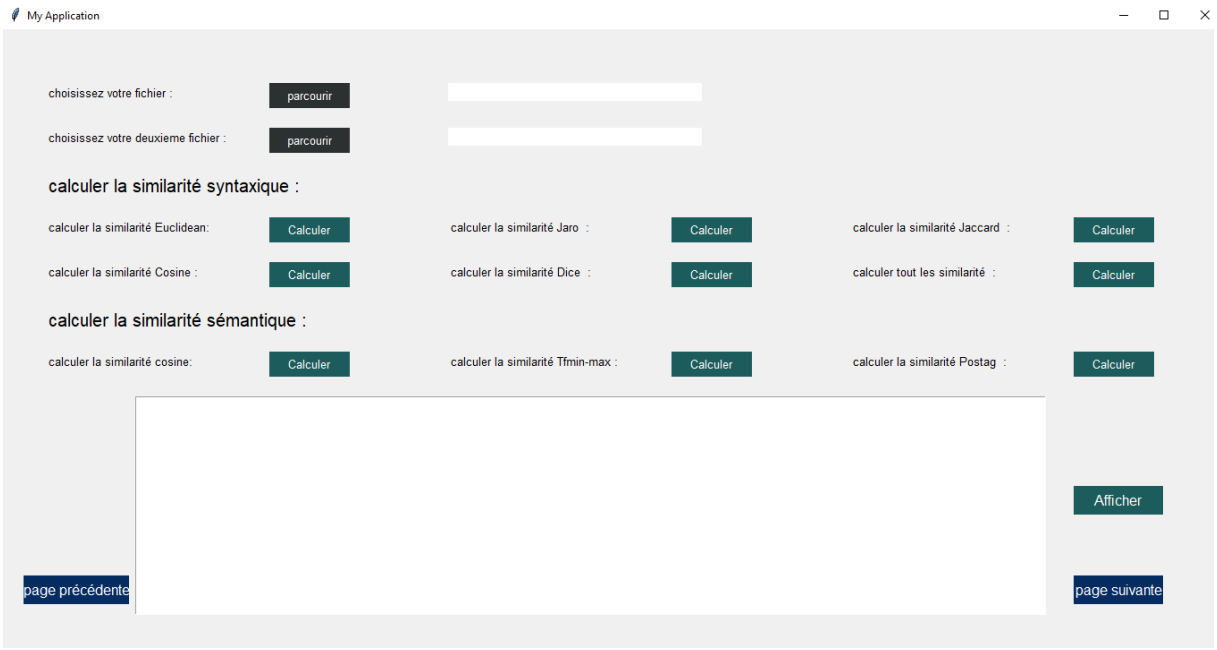


Figure 13 L'interface des calculs de similarité

Dans la dernière fenêtre l'utilisateur est invité pour calculer les moyennes des calculs de similarités -le détail de calculs est expliqué dans le chapitre 3-

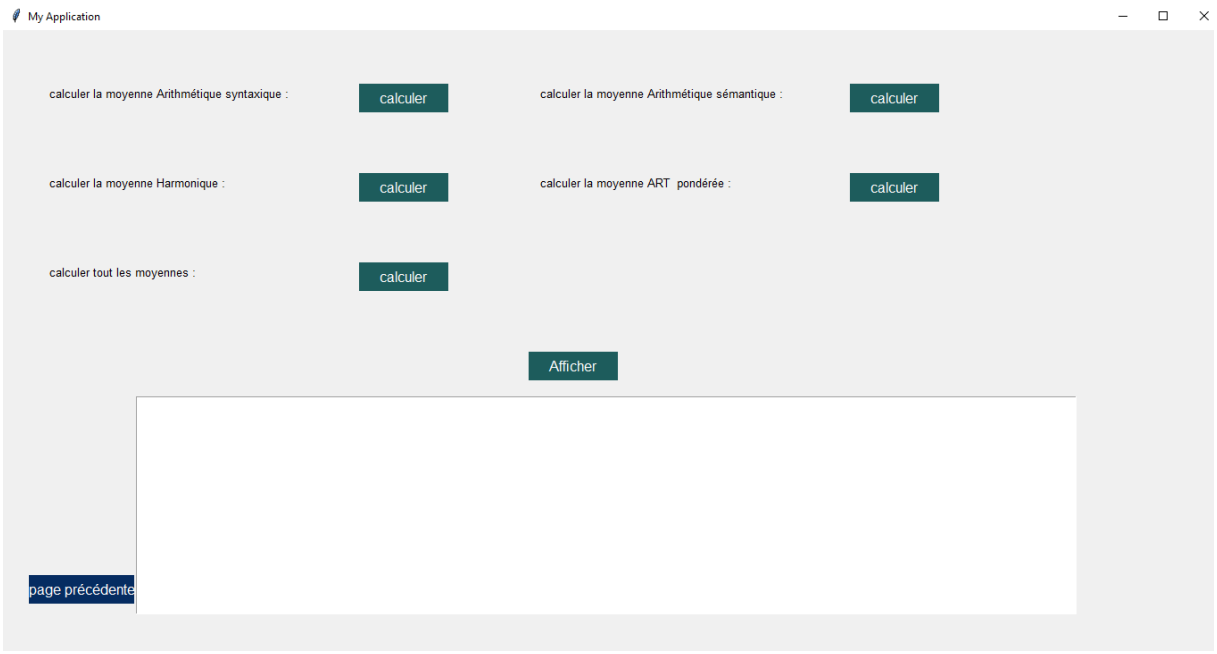


Figure 14 L'interface de calculs des moyennes

Pour l'implémentation de notre application ainsi que la génération de notre corpus nous avions à notre disposition l'ensemble des outils informatique (hardware & software) suivants :

- Hardware : Le système d'exploitation : Windows 10, CPU : Intel, core i5 Mémoire :8GB.
- Software: Python 3.9, Pycharm community Edition 2021.2, G-Ssplite 3.

## 3 CONCLUSION

Dans ce chapitre, nous avons tout d'abord évalué les couples candidats (phrases-paraphrases) par le biais de l'expertise humaine. Ensuite nous avons évalué les couples de phrases générés en exploitant deux principes : l'évaluation automatique en utilisant la métrique « Bleu », et une évaluation humaine en faisant appel aux connaissances des experts en langue arabe ; enfin nous pouvons dire que les résultats sont satisfaisants pour notre corpus monolingue arabe.

Enfin nous avons présentés notre implémentation de l’application desktop permettant d’effectuer des prétraitements et des alignements sur des couples de paraphrases.

# CONCLUSION GENERALE

L’objectif de ce travail est de générer un corpus parallèle monolingue de langue arabe.

Nous avons commencé par présenter l’état de l’art, qui compotait l’ensemble des concepts fondamentaux de paraphrases, corpus, corpus de paraphrases ainsi que les approches de création d’un corpus parallèle monolingue.

A partir de nos recherches profondes dans ce thème, nous avons choisie l’approche la plus adéquate pour atteindre notre objectif. Nous avons opté par la création à partir des corpus parallèles bilingues, à savoir que cette approche fournit des paraphrases de bonnes qualité vue qu’elles proviennent d’une opération de traduction.

Nous avons effectué une sélection de corpus parallèles bilingues disponibles, et en extraite plus d’un million et demi couples de phrases dans de différentes langues, les traduire en langue arabe en utilisant l’API de google.

Les couples de phrases résultants considérés comme couples candidats sont ensuite passées par une étape de prétraitement dans laquelle nous avons utilisés des techniques de normalisation, tokenisation et de lemmatisation pour concevoir des phrases arabes adapté et utilisable par les modèles de génération de paraphrases.

Une étape de filtrage était nécessaire pour examiner la qualité des phrases concrète, ici nous avons fait appels à un ensemble d’opérations des calculs de similarités syntaxiques et sémantiques ainsi que leur moyenne arithmétique, arithmétiques pondérées, et harmoniques. Ce qui nous a permis de classer nos couples candidats dans des classes représentant le dégrée de similarité entre la phrase et sa paraphrase.

Pour garantir la qualité de nos phrases candidates nous les avons fait passer par un ensemble d’évaluation automatique et humaine.

Tout d’abord nous avons solliciter des annotateurs (experts humains en langue arabes) pour évaluer nos couples (phrases-paraphrases) candidats, ensuite nous avons fait appel à un modèle de génération de paraphrases -EDAM- conçu dans un travail précédant, et qui nous a permis d’entrainer un nombre considérable de couples candidats et de tester seulement nos phrases candidates, lors de cette dernière phase EDAM nous a généré de nouvelles paraphrases.

Dans le cadre de l'évaluation des paraphrases générées nous avons utilisé la métriques « Bleu », et vu la nécessité de l'expertise humaine pour beaucoup de tâche de traitement automatique de langage nous avons encore une fois solliciter les annotateurs humains pour évaluer les paraphrases générées.

L'ensemble des techniques d'évaluation de notre data set a dévoilé des résultats très prometteurs.

## Références Bibliographiques


[1] F. Al-Raisi, W. Lin, and A. Bourai, “A Monolingual Parallel Corpus of Arabic,” *Procedia Comput. Sci.*, vol. 142, pp. 334–338, 2018, doi: 10.1016/j.procs.2018.10.487.

[2] M. Leman, *Lecture notes in artificial intelligence*, vol. 1317. 1997.

[3] A. Gravano, “Turn-taking and affirmative cue words in task-oriented dialogue,” *Diss. Abstr. Int. B Sci. Eng.*, vol. 70, no. 8, p. 4943, 2010, doi: 10.1162/COLI.

[4] H. Bouamor, “Etude De La Paraphrase Sous-Phrastique En Traitement Automatique Des Langues,” p. 180, 2012, [Online]. Available: https://tel.archives-ouvertes.fr/tel-00717702/.

[5] O. HAMEL and S. LAMARI, “LES RESEAUX DE NEURONES POUR LA GENERATION AUTOMATIQUE DE PARAPHRASES,” Saad Dahleb, 2021.

[6] M. D. E. M. Ii and M. L. Selena, “Présenté par : Remerciements.”

[7] T. McEnery and R. Xiao, “Chapter 2. Parallel and Comparable Corpora: What is Happening?,” *Inc. Corpora*, pp. 18–31, 2018, doi: 10.21832/9781853599873-005.

[8] R. C. K. Under, “Extraction of Parallel Corpora from Comparable Corpora,” [Online]. Available: https://www.cfilt.iitb.ac.in/resources/surveys/ComparableCorporaSurvey.pdf.

[9] T. Kajiwara and M. Komachi, “Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings,” *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, pp. 1147–1158, 2016.

[10] P. Resnik, “Parallel strands: A preliminary investigation into mining the web for bilingual text,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1998, vol. 1529, pp. 72–82, doi: 10.1007/3-540-49478-2_7.

[11] R. Ghani, R. Jones, and D. Mladenić, “Mining the web to create minority language corpora,” in *Proceedings of the tenth international conference on*

*Information and knowledge management*, 2001, pp. 279–286, [Online]. Available: https://dl.acm.org/doi/abs/10.1145/502585.502633.

[12] F. Isaac, T. Hamon, C. Fouqueré, L. Bouchard, and L. Emirkanian, “Extraction informatique de données sur le web,” *Rev. DistanceS*, vol. 5, no. 2, pp. 195–210, 2001, [Online]. Available: https://www.researchgate.net/profile/Louisette-Emirkanian-2/publication/267939326_Extraction_informatique_de_donnees_sur_le_web/links/5adf92eaa6fdcc29358fd565/Extraction-informatique-de-donnees-sur-le-web.pdf.

[13] M. Baroni and S. Bernardini, “BootCaT: Bootstrapping Corpora and Terms from the Web.,” in *LREC*, 2004, pp. 1313–1316, [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.3245&rep=rep1&type=pdf.

[14] M. Ueyama and M. Baroni, “Automated construction and evaluation of Japanese web-based reference corpora,” *Proc. Corpus Linguist. 2005*, 2005.

[15] “Constitution d’un corpus de la langue Arabe à partir du Web.” .

[16] W. Coster and D. Kauchak, “SimpleEnglishWikipedia_ ANewTextSimplificationTask.pdf,” *Proc. ofthe 49th Annu. Meet. ofthe Assoc. Comput. Linguist. pages 665–669, Portland, Oregon, June 19-24, 2011.*, pp. 665–669, 2011.

[17] W. Coster and D. Kauchak, “Learning to Simplify Sentences Using Wikipedia,” *Work. Monolingual Text-To-Text Gener. pages 1–9, Proc. 49th Annu. Meet. Assoc. Comput. Linguist.*, no. June, pp. 1–9, 2011.

[18] A. Eyecioglu and B. Keller, “Constructing a Turkish corpus for paraphrase identification and semantic similarity,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9623 LNCS, pp. 588–599, 2018, doi: 10.1007/978-3-319-75477-2_42.

[19] W. Xu, C. Callison-Burch, and C. Napoles, “Problems in Current Text Simplification Research: New Data Can Help,” *Trans. Assoc. Comput. Linguist.*, vol. 3, pp. 283–297, 2015, doi: 10.1162/tacl_a_00139.

[20] D. Buscaldi *et al.*, “Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? To cite this version : HAL Id : hal-02784738 Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? Introduction,” 2020.

[21] P. Pantel, H. Hill, and A. C. A. Th, “Discovery of Inference Rules from Text,” vol. 2, no. 12, p. U.S. Patent No. 7,146,308. Washington, DC: U.S. Pa, 2006.

[22] I. Szpektor, E. Shnarch, and I. Dagan, “Instance-based evaluation of entailment rule acquisition,” *ACL 2007 - Proc. 45th Annu. Meet. Assoc. Comput. Linguist.*, no. June, pp. 456–463, 2007.

[23] C. Callison-Burch, T. Cohn, and M. Lapata, “ParaMetric: An automatic evaluation metric for paraphrasing,” *Coling 2008 - 22nd Int. Conf. Comput. Linguist. Proc. Conf.*, vol. 1, no. August, pp. 97–104, 2008.

[24] A. Lavie and M. J. Denkowski, “The METEOR metric for automatic evaluation of Machine Translation,” *Mach. Transl.*, vol. 23, no. 2–3, pp. 105–115, 2009, doi: 10.1007/s10590-009-9059-4.

[25] G. Wentzel, “Funkenlinien im Röntgenspektrum,” *Ann. Phys.*, vol. 371, no. 23, pp. 437–461, 1922, doi: 10.1002/andp.19223712302.

[26] F. J. Och, “Minimum Error Rate Training,” *Proc. 41st Annu. Meet. Assoc. Comput. Linguist.*, pp. 160–167, 2003.

[27] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” *AMTA 2006 - Proc. 7th Conf. Assoc. Mach. Transl. Am. Visions Futur. Mach. Transl.*, no. August, pp. 223–231, 2006.

[28] G. Leusch, N. Ueffing, and H. Ney, “CDER: Efficient MT evaluation using block movements,” *EACL 2006 - 11th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Conf.*, pp. 241–248, 2006.

[29] D. Klakow and J. Peters, “Testing the correlation of word error rate and perplexity,” *Speech Commun.*, vol. 38, no. 1–2, pp. 19–28, 2002, doi: 10.1016/S0167-6393(01)00041-3.

[30] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, “GLEU Without Tuning,” pp. 0–2, 2016, [Online]. Available: http://arxiv.org/abs/1605.02592.

[31] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, “Ground truth for grammatical error correction metrics,” *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 2, pp. 588–593, 2015.

[32] P. B. Farouk, “La langue arabe et le TAL : étude de cas,” pp. 59–76.

[33] W. Salloum and N. Habash, “Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation,” *Proc. first Work. algorithms Resour. Model. dialects Lang. Var. Assoc. Comput. Linguist.*, pp. 10–21, 2011, [Online]. Available: http://www.mt-archive.info/EMNLP-2011-Salloum.pdf%0Ahttps://dl.acm.org/doi/10.5555/2140533.2140535.

[34] O. Kashefi, “MIZAN: A Large Persian-English Parallel Corpus,” 2018, [Online]. Available: http://arxiv.org/abs/1801.02107.

[35] S. C. U. Oboi, “Alignement de phrases parallèles dans des corpus bruités,” 2013, [Online]. Available: https://www.iro.umontreal.ca/~felipe/new-home/memoires/msc_fethi_lamraoui.pdf?fbclid=IwAR2abBBkkcVvZCR3U3c-lmcGZAt7lXIyST7eX9iNu-dvdN8wDPBH1Z3cTAY.

[36] M. Mustafa, A. S. Eldeen, S. Bani-Ahmad, and A. O. Elfaki, “A Comparative Survey on Arabic Stemming: Approaches and Challenges,” *Intell. Inf. Manag.*, vol. 09, no. 02, pp. 39–67, 2017, doi: 10.4236/iim.2017.92003.

[37] A. M. Goweder, H. A. Alhammi, T. Rashed, and A. Musrati, “A hybrid method for stemming arabic text,” *Second Conf. Arab. Lang.*, pp. 1–7, 2010.

[38] S. Bessou, M. Louail, A. Refoufi, Z. Kadem, and M. Touahria, “Un systeme de lemmatisation pour les applications de TALN,” no. June 2007, 2017.

[39] M. Keywords, “Filtrages syntaxiques de co-occurrences pour la repr ´ esentation vectorielle de documents R ´ esum ´ e - Abstract 1 Introduction 2 Le mod ` ele de repr ´ esentation DSIR,” pp. 24–27, 2002.

[40] E. Negre, “Comparaison de textes: quelques approches...,” 2013.

[41] A. Huang, “Similarity measures for text document clustering,” *New Zeal. Comput. Sci. Res. Student Conf. NZCSRSC 2008 - Proc.*, no. April, pp. 49–56, 2008.

[42] A. Strehl, J. Ghosh, and R. Mooney, “Impact of similarity measurement,” pp. 1–7, [Online]. Available: http://www.ideal.ece.utexas.edu/papers/strehl_aaai00.pdf.

[43] V. Bavi, T. Beirne, N. Bone, J. Mohr, and B. Neal, “Comparison of document similarity metrics,” *Comput. Sci. Dep. West. Washingt. Univ. Inf. Retrieval, Winter*, vol. 2010, 2010.

[44] A. N. Ngom, “Étude des mesures de similarité semantique basees sur les arcs,” in *CORIA 2015 - Conference in Search Infomations and Applications - 12th French Information Retrieval Conference*, 2015, pp. 535–544, [Online]. Available: https://asso-aria.org/coria/2015/RJCRI/rjc_22.pdf?fbclid=IwAR2abBBkkcVvZCR3U3c-lmcGZAt7lXIyST7eX9iNu-dvdN8wDPBH1Z3cTAY.

[45] D. Karani, “Introduction to Word Embedding and Word2Vec,” *Towards Data Science*, 2018. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa (accessed Aug. 25, 2021).

[46] G. Salton and C. Buckley, “Solton-1-29-03.Pdf.” pp. 513–523, 1988, [Online]. Available: https://doi.org/10.1016/0306-4573(88)90021-0.

[47] Lambert.R, “Comprendre le fonctionnement d’un LSTM et d’un GRU en schémas,” 2020. https://penseeartificielle.fr/comprendre-lstm-gru-fonctionnement-schema/?fbclid=IwAR3GzL_-dQ-nuV8qH6TQmtR9IOZ-6hX1vrbNwOAC1xYx_DzFIrViXwQtxMY.

[48] É. des Modèles, “Documentation d’AutoML Translation.” https://cloud.google.com/translate/automl/docs/evaluate?hl=fr&fbclid=IwAR3f_J5uf2u (accessed Sep. 25, 2021).

## ANNEXE 1

1. Echantillon de couples de phrases utilsées pour l'evaluation qualitative

**يرجى وضع علامة X في الخانة الموافقة**

**لتقييمك للتشابه في المعنى بين الجملتين.**

| الجملة الأولى | متشابهة تماما | متشابهة بشكل أساسي (بعض التفاصيل مفقودة) | مشاركة بعض التفاصيل (الكثيرمن التفاصيل مفقودة) | مشاركة الموضوع فقط | مختلفة تماما | ملاحظة اخرى | الجملة الثانية |
|---|---|---|---|---|---|---|---|
| آن الأوان لإنصافكم وإنصافكم. | | | | | | | لقد حان الوقت لضمان تحقيق العدالة والتعويض عن الظلم. |
| إنه شخصية محرجة. | | | | | | | هذا رقم كبير بشكل محرج. |
| إن السياسة العامة للمفوضية جيدة. | | | | | | | إن مجمل مسار عمل اللجنة جيد. |
| المقلدة صفقة كبيرة. | | | | | | | التزوير عمل كبير. |
| أعتقد أنه سيعطينا تمويلًا جيدًا. | | | | | | | أعتقد أن هذا من شأنه أن يمنحنا تمويلًا جيدًا. |
| إنه سؤال صعب. | | | | | | | إنها قضية صعبة. |
| إنك تبذر عاصفة مناهضة للراديكالية وستحصد إعصارًا شعبيًا. | | | | | | | إنك تبذر عاصفة ضد الجذور الشعبية وستحصد إعصارًا على مستوى القاعدة الشعبية. |
| اجعل نقطة صحيحة للغاية. | | | | | | | لقد ابديت وجهة نظر سليمة. |
| في هذه الحالة ، في رأيي ، ستكون قد فشلت في مهمتها وهدفها. ..... | | | | | | | في هذه الحالة ، ستكون ، في رأيي ، قد فشلت في مهمتها والغرض منها. ..... |

2. Echantillon des phrases originales et les paraphrases générées utilisées pour l'évaluation

| | Source phrases (العبارات الأصلية المدخلة) | Generated paraphrases ( العبارات المصاغة الناتجة) | Relevance ( الملاءمة في المعنى) | Readability (المقروئية) |
|---|---|---|---|---|
| 1 | انني اؤيد بشده فكره ان اللجنه ستبحث في العام المقبل عن حل افضل واكثر عمليه واسهل للتحكم | انا اويد تماما ان المفوضيه قد وافق عليه في معاهده زمنيه فعاله والتي من استراتيجيه منظمه التجاره العالميه | | |
| 2 | لذلك انا سعيد جدًا لان المفوضيه تاخذ مبادره البرلمان بجديه كبيره وهذا مكان جيد جدًا للبدء | لذلك انا سعيد للغايه للمجلس ان الهييه ان البرلمان بشكل خاص عن المستقبل | | |
| 3 | ثم يمكنك انت المفوض والمفوضيه ان تفعل ما تريد وان تجعل الامور تعمل اعتبارًا من 1 يناير | لذلك اذا طلبت المفوض بعد ان تقدم لنا رغبتها في البدايه القيام به | | |
| 4 | علاوه على ذلك فان الهيكل المؤسسي السليم هو شرط اساسي لتقديم الخدمات العامه بكفاءه وشفافيه وعادله | ومن ناحيه اخرى نحن نويد بشكل فعال اجراء تقييم علمي مستقل من خلال الاستخدام الصحيح الى اساس علمي مستقل | | |
| 5 | لذلك انا سعيد جدًا لان المفوضيه تاخذ مبادره البرلمان بجديه كبيره وهذا مكان جيد جدًا للبدء | لذلك انا سعيد للغايه للمجلس ان الهييه ان البرلمان بشكل خاص عن المستقبل | | |
| 6 | واشارت اللجنه في ردها على بعض اسئله البرلمانيين الى انها لا تعتبر مواءمه محدده ضروريه في هذه المرحله | تطلب المفوضيه عن المجموعه الاشتراكيه من الضروري مراعاه الاهداف من خلال حقيقه انها غير المرغوب فيها | | |
| 7 | لهذا السبب بالاضافه الى المديرين الاقوياء والمثابرين نحتاج ايضًا الى شروط اطار عمل واضحه | لهذا السبب بالاضافه الى المديرين الاقوياء والمثابرين نحتاج ايضًا الى شروط اطار عمل واضحه | | |

| | | | | |
|---|---|---|---|---|
| | | اذا لم نرغب في هذه المجموعه الاوروبيه قد يعمل دون عوايق ان تجد اموالا اضافيه<br><br>..... | ذا لم يُطلب من المواطنين الاوروبيين ابداء رايهم في هذه القضيه فسيكون ذلك بمثابه ضربه قاسيه للعمليه الديمقراطيه<br><br>..... | 8 |