Partie III : Extraction des relations

1. Introduction

L'extraction des relations sémantiques entre des termes est une tache très difficile à réaliser, Certains travaux ont tenté d'apporter des réponses à ce problème à partir de l'étude de corpus.

Ce que nous proposons dans cette partie est une méthode hybride d'extraction de relations à partir du texte Coranique. Comme pour les termes, nous allons utiliser d'abord une approche linguistique basée sur les règles JAPE pour extraire des relations entre termes simples, puis nous validerons les résultats obtenus par une approche statistique basée sur l'information mutuelle. Pour les relations entre collocations, nous utiliserons un concordancier pour rechercher des relations entre termes complexes en utilisant des marqueurs, nous sauvegardons les résultats dans une base de données puis nous soumettons des requêtes en imposant certaines contraintes. Les résultats sont aussi filtrés par l'IM.

La première approche se fait à l'aide de l'outil GATE et une grammaire de détection qui couvre les différentes formes syntaxiques d'apparition d'une relation dans le texte. À chaque relation sont associées une ou plusieurs grammaires, l'application de ces dernières sur le texte permet d'identifier une éventuelle expression de cette relation. Etant donné que GATE dispose d'un transducteur permettant l'application des grammaires JAPE sur le texte, nous avons utilisé ce langage pour écrire les règles de détection de relations. Bien que nous utilisons les patrons syntaxiques, nous avons aussi utilisé des marqueurs comme (**عبارة_عن مثل، مثله_كمثل، كـ،**).

Le travail comporte deux parties : L'une traite la recherche de relations entre termes simples, en utilisant des patrons parce que très souvent et spécialement dans la langue arabe deux noms qui se succèdent (le cas du génitif par exemple) exprime une relation sémantique. Exemple (**خالق_السموات**= Créateur des cieux), où on peut déceler une relation (**خُلق_من_طرف**= *créé_par*)

Donc il est très courant que l'on puisse extraire des collocations précédentes des relations sémantiques, mais ce genre de relations, puisqu'elles n'utilisent pas des marqueurs, doivent impérativement être validé par un expert humain**.** Dans ce volet là, nous pouvons trouver des relations nommées comme (**أُنزل_على**) comme

nous pouvons reconnaitre des relations non nommées comme (القواعد_من_البيت) où il est question d'une relation de méronymie qui n'est pas explicite dans le syntagme précédent.

L'autre partie concerne les relations entre collocations, où nous utilisons ici des marqueurs et la liste des collocations extraites précédemment.

Toute fois nous n'avons pu recenser toutes les relations, tant elles sont nombreuses, mais nous donnons la méthode pour les retrouver dans les deux cas.

2. Architecture du système d'extraction de relations

L'architecture du système d'extraction de relation est le même que celui des collocations avec quelque menues modifications.

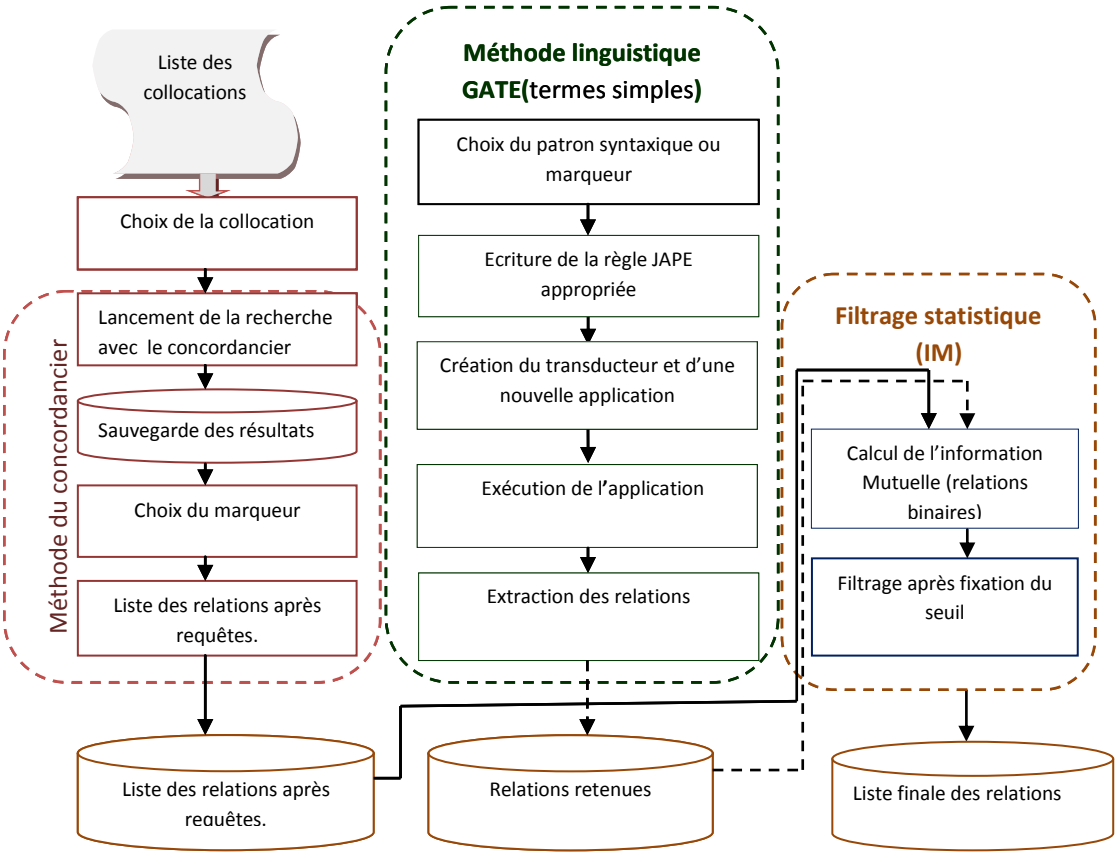

Figure 33: Système d'extraction de relations

3. Description des composants

3.1. Extarction de relations entre termes simples

Pour extraire des relations entre termes simples, nous utilisons GATE, tel que c'est décrit dans *la section 2.3.* Des patrons de relations sont choisis, puis nous écrivons la règle JAPE adéquate, apres avoir créé un nouveau transducteur, nous lui passons la règle comme paramètre, nous chargeons le corpus, ensuite nous créons un nouveau pipeline, constitué de segmenteur et de transducteur. Apres l'execution de l'application le système affiche les relations extraites surlignées.

La premiere relation à laquelle nous nous sommes interessés est la relation de méronymie qui s'exprime entre autre par le patron Nom-مِنْ-Nom, exprimant le fait qu'un terme B est une partie d'un terme A.

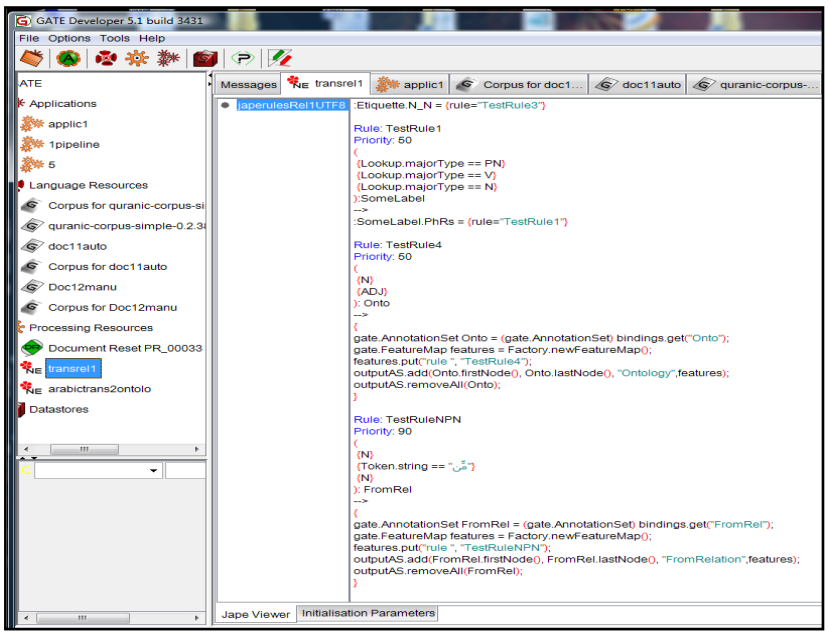

Figure 34: Règle JAPE pour l'extraction d'une relation de méronymie

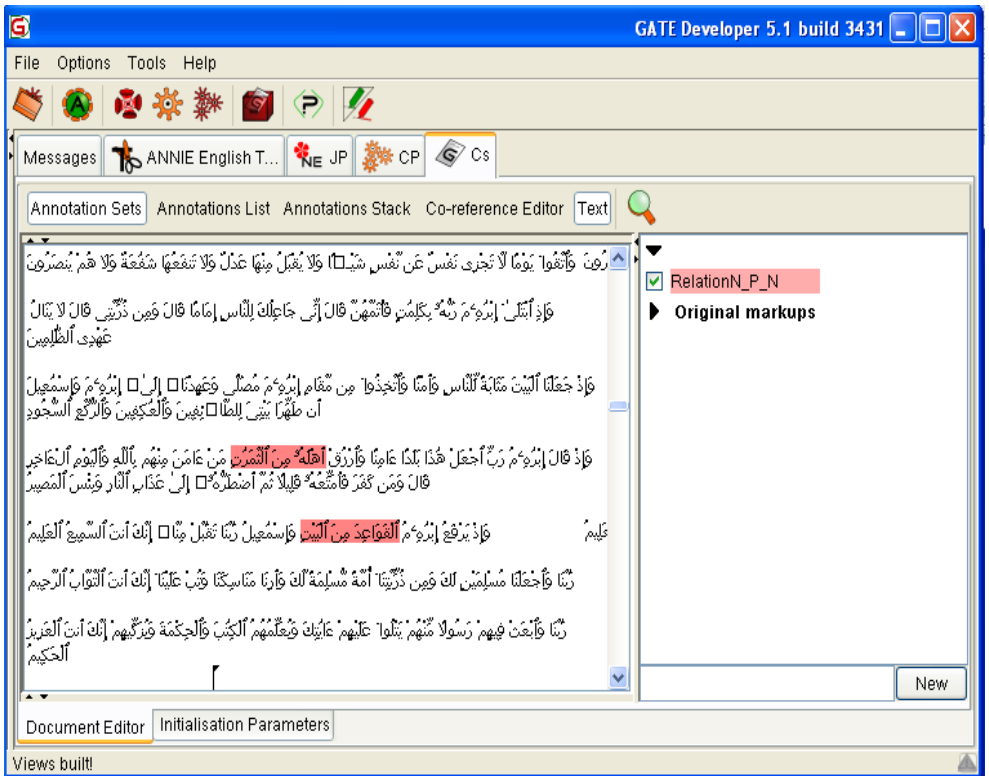

Figure 35: Relation respectant Nom-**منْ**-Nom

Une partie des relations affichées par GATE est donnée par le tableau suivant :

Nom	منْ	Nom
العذاب	مِنَ	بمزحزحه
الثمرات	مِنَ	أهله
البيت	مِنَ	القواعد
الناس	مِنَ	السفهاء
النار	مِنَ	بخرجين
المصلح	مِنَ	المفسد
الغي	مِنَ	الرشد
المشرق	مِنَ	بالشمس

Tableau 12: Tableau des resultats de la relation Nom-**منْ**-Nom (Sourate El-Baqara)

Concernant cette relation, il est clair que vu la richesse de la langue arabe, la préposition « **منْ** » introduite entre deux noms peut avoir des sens différents, cela

peut bien exprimer la méronymie par exemple (**القواعد من البيت**) ou (**السفهاء من الناس**) comme elle peut exprimer l'antonymie (**المفسد من المصلح**) ou (**الرشد منَ الغي**) Elle peut également exprimer une relation de (**يأتي_من**) comme c'est le cas de (**الشمس منَ المشرق**). Quant aux (**بمزحزحه منَ العذاب**) ou (**بخرجين منَ النار**) c'est dû à la qualité de l'étiquetage, les noms étant agglutinés à des particules, ont été étiquetés en tant que nom.

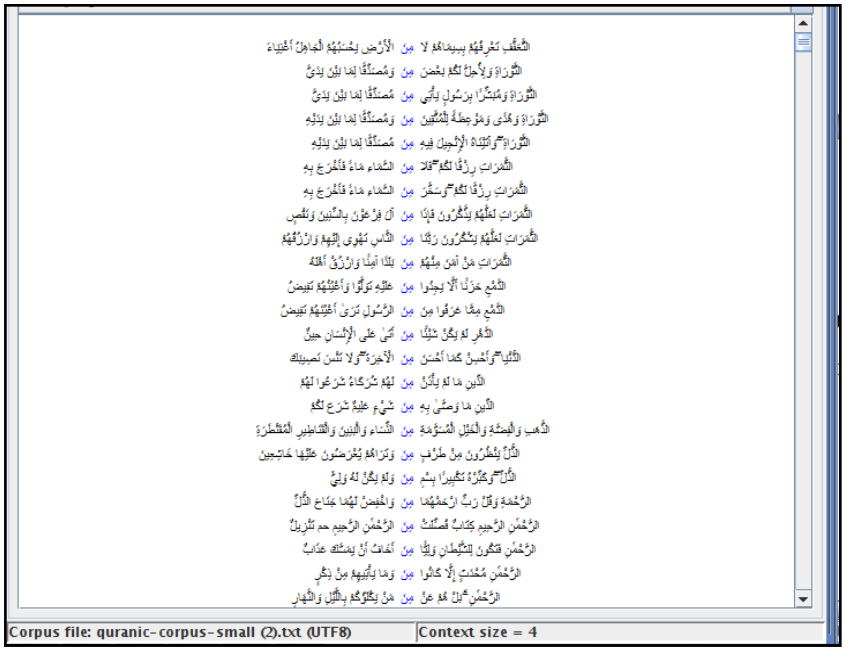

Figure 36: Liste des pharases contenant "**من**"

Il faut noter que les resultats sont affichés de gauche à droite.

3.2. Extraction de relation entre collocations

En plus des relations entre termes simples nous avons tenté de travailler sur les relations entre les collocations extraites précédemment. L'utilisation des patrons dans ce cas aurait relevé d'énormes difficultés compte tenu de la structure qu'il fallait choisir. Donc l'utilisation de GATE s'est avérée inappropriée avec des collocations. Nous avons pensé à revenir vers les concordancier.

Le principe est le suivant :

a) On choisit une collocation parmi celles extraites préalablement,
b) On choisit le marqueur de la relation ciblée
c) On lance la recherche avec la collocation suivie de la préposition de la relation On sauvegarde les résultats dans une base de données

d) On soumet des requêtes en imposant une contrainte sur le contexte gauche
e) On revient à l'étape a) jusqu'à épuisement
f) On évalue les résultats.

La collocation choisie est soumise au concordancier (voir sa définition dans la *section 2.2.5* du chapitre3). Le concordancier choisi pour ce travail est aConCorde[52]. aConcorde est un outil multilingue pour chercher des concordances, il est doté des fonctions de base d'un concordancier. Développé à l'origine pour la langue arabe, il possède également une interface pour l'Anglais. Ecrit en JAVA, il peut être exécuté sur n'importe quelle plateforme où est installée Java Runtime Environment (**Roberts & al, 2005**). Etant donné un corpus chargé, aConCorde affiche tous ses mots par ordre alphabétique, ainsi que la fréquence de chaque mot. Lorsqu'un mot est choisi, aConCorde affiche alors toutes ses occurrences et ses contextes gauche et droit c'est-à-dire les mots le précédant et ceux lui succédant. Ces contextes se comptent généralement en nombre de caractères ou en nombre de mots.

Les concordances sont alors sauvegardées dans une base de données simple, On choisit la syntaxe d'un marqueur et on soumet une requête respectant ce marqueur, puis on recueille les résultats retournés dans une liste qu'on sauvegarde. Un exemple de relation sur lequel nous avons expérimenté cette approche est le suivant :

Nom-Adjectif-مِنَ-Nom-Adjectif

[52] http://www.andy-roberts.net/coding/aconcorde

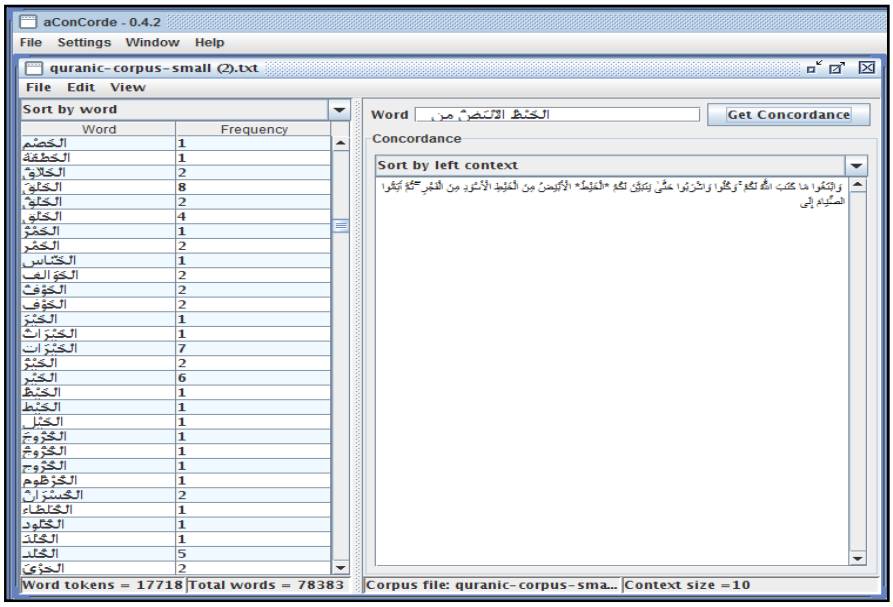

Figure 37: Exemple d'utilisation de aConCorde

La recherche de relations entre collocations s'est avérée plus difficile que nous le pensions. Les limites du concordancier et la complexité de la structure des phrases du Coran ont fait que nous n'avons pu obtenir de bons résultats. Le concordancier affichant le même mot défini et indéfini comme deux mots différents, le calcul de la fréquence se trouve biaisé et l'analyse ne peut donner lieu à des résultats fiables. Nous donnerons des perspectives de solutions concernant ce problème dans la conclusion générale.

3.3. Filtrage statistique

Une fois les relations entre termes simples et collocations extraites (nous n'avons obtenu que peu de résultats pour les collocations), nous procédons au calcul de l'information mutuelle pour valider les résultats obtenus.

Nous avons utilisées cette mesure parce qu'elle est l'une des mesures les plus simples pour déterminer la force d'association entre deux ou plusieurs mots et permet de quantifier l'information partagée par des couples de mots ou termes et de repérer les groupes de mots qui apparaissent ensemble plus fréquemment. Cette mesure se base sur l'hypothèse que l'emploi de deux termes en cooccurrence est l'expression d'une relation sémantique entre ces termes.

$$\mathrm{IM}(\mathrm{x},\mathrm{y}) = \frac{\mathrm{N}f(\mathrm{x},\mathrm{y})}{f(\mathrm{x})\,f(\mathrm{y})}$$

Équation 5: Formule de l'information mutuelle

f(x) et *f*(y) sont les nombres d'occurrences de mots x et y dans un corpus de taille N, et *f*(x, y) est la probabilité de les observer simultanément.

Le x et le y ici pouvant designer un terme simple ou composé considéré comme une seule entité et dont la fréquence est déduite de l'étape précédente lors du calcul de la fréquence du terme simple ou de la collocation.

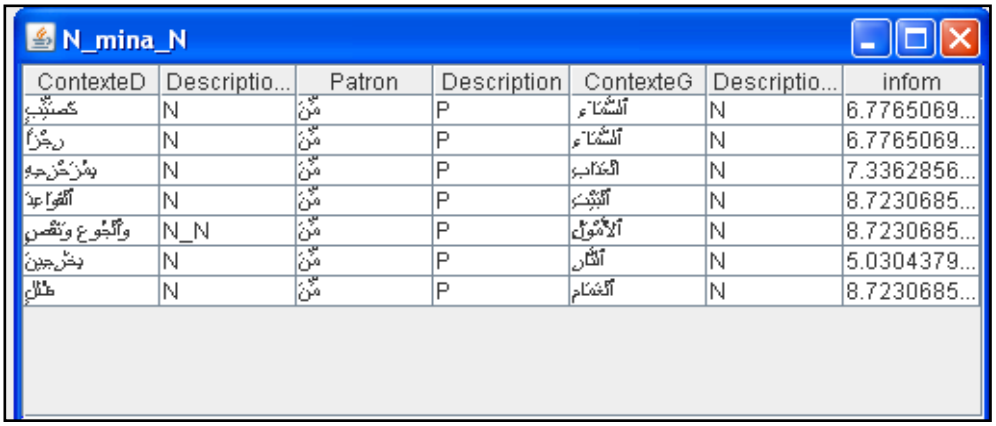

Figure 38: Calcul de l'IM pour le filtrage

Après le calcul de l'IM, un seuil est fixé, l'expérience montre que la valeur de 0.6 est raisonnable comme seuil entre des relations acceptables et celles non acceptables. Les relations dont l'IM est supérieur au seuil sont retenues, les autres sont sauvegardées dans un autre fichier. Un expert pourra toujours valider manuellement les relations gardées et puiser dans le second fichier pour récupérer d'autres relations.

4. Discussion et évaluation

Nous avons essayé de déterminer les patrons syntaxiques pouvant représenter les relations de subsomption, de méronymie, antonymie, comparaison et exception en tenant compte de la nature du texte coranique et aussi d'un texte ordinaire pour que la règle soit valable quelque soit le texte traité.

L'exemple du patron de la relation de subsomption peut aussi être de la forme « Nom-Nom » comme « **وَاذْكُرْ فِي الْكِتَابِ إِسْمَاعِيلَ إِنَّهُ كَانَ صَادِقَ الْوَعْدِ وَكَانَ رَسُولًا نَبِيًّا**» **(**مريم 54 **)**

ou «Nom-**من**-Nom » comme dans «**أولو العزم -من- الرسل** », « *Nom-عبارة_عن-Nom* ». Où *Nom* représente une étiquette dans le corpus traité. Bien que ce dernier marqueur n'est pas présent dans le coran, il est le marqueur type d'une relation de subsomption dans des textes d'ordre général.

Nous avons essayé des relations de comparaison avec le marqueur « **كـ** ou **مثلهم كمثل**», le résultat fut une longue liste mais complètement inexploitable du fait de l'ordre des composant de la phrase qui ne sont pas toujours du type *comparé- outil de comparaison- comparant.*

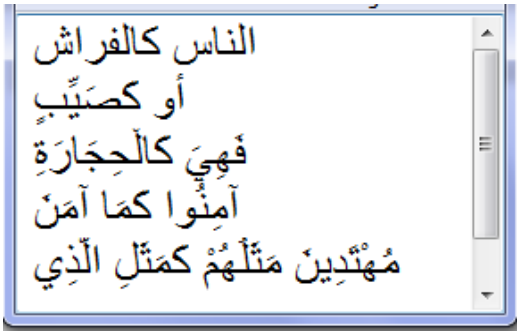

Figure 39: Exemple de phrase contenant des marqueurs de comparaison

La structure complexe du texte coranique rend difficile l'extraction de telles relations. Cela nécessite une désambigüisation et un traitement de la coréférence.

Exemple du verset : (**مثل الذين ينفقون أموالهم في سبيل الله كمثل حبة**), nous remarquons que dans certains cas, le premier nom peut être exprimé par une longue phrase et que dans d'autres cas un mot peut exprimer aussi toute une phrase ou que la première partie (le comparé) n'est pas immédiatement suivie par la préposition et le second nom (le comparant), entre les deux, nous pouvons trouver une ou deux phrases insérées, composées de verbes, adverbes, particules et conjonctions, c'est l'une des caractéristiques de la langue arabe où l'ordre des composants de la phrase n'est pas unique. Le même problème apparait dans les relations d'exception (Nom-**إلا**-Nom). La précision était la plus basse du fait que cette relation est généralement utilisée entre deux verbes, comme c'est le cas du verset (**أتبع إلا ما يوحى**) et les noms retournés ne sont pas le domaine et le rang de la relation. Les deux relations donnent Approximativement le même résultat. Autre

problème dû au fait que dans la langue arabe, il ya des verbes transitifs qui ont besoin de deux ou trois compléments, ce qui conduit quelque fois à une fausse annotation.

Au lieu d'utiliser l'étiquette *Nom*, il aurait mieux valu écrire des patrons avec (Nom de sujet = **اسم فاعل**), (Nom d'objet = **اسم مفعول**) (Nom d'outil= **اسم آلة**) (Adverbe de lieu= **ظرف مكان**) et (Averbe de temps= **ظرف زمان**) etc. La précision aurait été bien meilleure.

La précision moyenne pour ces types de relations était de 0.51, elle variait entre 0.26 pour la relation d'exception et 0.66 la relation de subsomption.

Outre cela le Coran étant un texte écrit en arabe traditionnel, il n'a pas les mêmes caractéristiques que les autres corpus écrits en arabe standard moderne, ceci affecte également la précision. Cependant les résultats obtenus sont comparables à un travail similaire dans un autre langage, en l'occurrence l'Anglais, Roberts (**Roberts & al, 2007**) obtient une F.mesure de 0.7 pour l'extraction d'entités à partir de textes cliniques en Anglais. La majorité de ces problèmes (étiquetage, ambigüité linguistique, arabe traditionnel) sont présents dans toutes les étapes puisque le travail se fait sur le même corpus.

Pour évaluer la méthode utilisée, nous avons procédé au calcul de la précision avant puis après le filtrage. Nous avons noté une amélioration sensible de la précision après avoir appliqué la méthode statistique.

Approche	**Précision**
Linguistique	0.57
Après Filtrage	0.66

Tableau 13: Précision avant et après le filtrage par la méthode statistique

Dans cette précision, nous n'avons pas tenu compte des relations correctes extraites mais avec le patron d'un autre type de relation.

Prenons l'exemple de Nom-**منَ**-Nom (**المفسد من المصلح**) et (**الرشد من الغي**) qui expriment bien une relation d'antonymie, mais elles sont obtenues alors qu'on

cherchait une relation de méronymie. Même exemple pour (**بالشمس من المشرق**) qui exprime quant à elle une relation (**يأتي_من**) dépendante du corpus.

Nom	**منَ**	Nom	
المصلح	**مِنَ**	**المفسد**	Acceptée ?
الغي	**مِنَ**	**الرشد**	Acceptée ?
المشرق	**مِنَ**	**بالشمس**	Acceptée ?

Tableau 14: Relations correctes mais extraites avec le patron d'une autre relation.

Pour améliorer, la précision nous proposons un affinage de l'étiquetage de telle sorte que l'on puisse faire la différence entre nom de sujet, nom d'objet, nom d'outils, adverbe de lieu et adverbe de temps et l'écriture de règles plus complexes prenons en compte ces nouvelles étiquettes.

5. Conclusion

La méthode que nous venons de proposer consiste en l'application d'une méthode linguistique basée sur des grammaires pour l'extraction de termes et de relations, filtrée par une approche statistique utilisant une métrique qui a donné de bons résultats dans ce domaine en l'occurrence l'information mutuelle.

Cette méthode n'est pas encore parfaite, beaucoup de chose reste à faire, beaucoup de zones à explorer, mais elle a au moins le mérite d'avoir défriché un terrain jusque là inexploité. Nous avons mis la lumière sur beaucoup de points d'ombre concernant la construction d'ontologies à partir de textes arabes. Le traitement du Coran nous a permis de découvrir à quel point les programmes existants ne sont pas encore en hauteur d'explorer les merveilles du texte saint. Nous espérons que des travaux continuent dans ce sens pour mettre à la disposition de tout le monde un outil facile permettant de naviguer dans l'interprétation du Coran autant que dans ses mots.