Chapitre 4 : Le système proposé

1. Introduction

Nous présentons dans ce chapitre une méthode hybride pour l'extraction des termes et des relations, nous avons d'abord procédé à l'extraction de termes simples à l'aide d'une approche statistique basée sur la métrique tf-idf. A la fin, nous avons établi une liste de termes simples.

Pour extraire des termes composés, nous avons utilisé d'abord une approche linguistique en utilisant des patrons prédéfinis, puis nous avons filtré les résultats avec une approche statistique basée sur l'information mutuelle.

Pour l'extraction des relations sémantiques, nous avons suivi le même modèle, d'abord nous appliquons une approche linguistique basée sur les marqueurs permettant de repérer des relations entre termes, puis nous avons validés les résultats par une approche statistique.

Apres avoir défini quelques instances pour des concepts concrets, nous avons procédé à la formalisation des concepts (entités, relations et instances) avec la logique de description et avons donné à la fin un exemple d'opérationnalisation de l'ontologie.

2. Motivation

L'objectif étant de fournir une plateforme pour la construction d'ontologies à partir de textes arabes, le premier corpus auquel nous pouvions penser était le Coran. Notre hésitation ne fit pas long feu. Le corpus était sur le web et un sérieux travail de prétraitement était entrain de se faire par des personnes des quatre coins du monde.

Et si on disposait d'une ontologie du Coran ?

Alors on pouvait l'utiliser dans l'indexation, la recherche d'information, la traduction automatique… (Bien que là c'est un peu trop ambitieux voire même prétentieux !).

Et pourquoi ne serait-elle pas une aide à découvrir de nouvelles interprétations avec toutes les relations conceptuelles dont on pouvait disposer ? L'horizon paraissait infini et le travail gigantesque, mais tout commence par un petit pas. C'est ce premier petit pas que nous avons tenu à faire, pour ouvrir la porte et permettre à d'autres de faire de grands pas de géants !

3. Objectif et Choix du Corpus

L'objectif initial était de fournir une plateforme pour la construction d'ontologies à partir de textes arabes. Pour ce faire il fallait disposer d'un corpus étiqueté sinon, il fallait se résoudre à commencer à faire un travail d'analyse du TALN qui consiste à construire un corpus et fournir les outils nécessaires pour le traiter.

3.1. Choix du premier corpus

Par chance en discutant avec un chercheur John Funk[40], j'ai su qu'il y avait un travail qui se faisait sur la construction du corpus coranique et qui était à l'étape d'étiquetage. Kais Dukes[41] contacté, m'oriente sur le travail qu'il fait, en précisant qu'il est toujours entrain d'affiner l'étiquetage. Donc le choix était fait, reste à choisir une démarche à suivre. La méthodologie adoptée s'inspire de celle proposée par (**Noy & Guiness, 2000**) avec quelque modification selon le besoin. Parce que le travail était gigantesque et nous nous sommes résolu à simplifier certaines étapes comme l'édition et la détermination de tous les attributs et les axiomes et à en rajouter d'autres, comme la formalisation et l'opérationnalisation. Donc les points saillants de notre travail étaient l'extraction des termes et des relations. C'est sur ces deux taches qu'on va se focaliser le plus, sans oublier bien sur la formalisation à la fin.

Le corpus choisi « *The Quranic Arabic Corpus*[42] » ou *The Crescent Corpus*, était en cours de traitement et d'expansion, un travail mené par Dukes et auquel participaient des centaines d'autres personnes des quatre coins du monde. Le travail a suscité l'engouement de tellement de personnes qu'il est devenu plus énorme et ne cesse de s'améliorer. Beaucoup s'y sont intéressés, qui pour

[40] http://www.sekt-project.com/author/adam.funk
[41] http://www.kaisdukes.com/
[42] http://corpus.quran.com

l'utiliser, qui pour y participer, Eric Atwell, Nizar Habash, Ahmed Abdelali et même Tim Buckwalter.

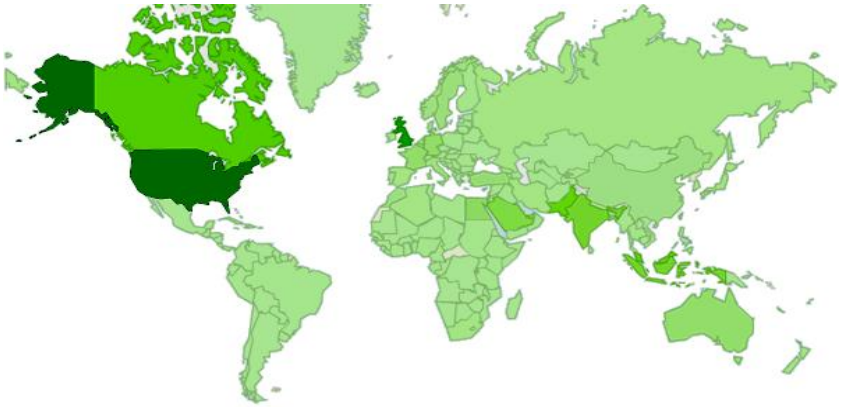

Figure 17: Carte des utilisateurs du Crescent corpus (donné par Google analytics)

Le projet a commencé par un étiquetage automatique des mots du Coran, nombre de personnes du domaine ont participé à des validations manuelles, puis l'étiquetage se raffinait peu à peu et une ontologie en anglais a vu le jour, et plus récemment une multi-interprétation vers l'Anglais.

3.2. Choix du deuxième corpus

Pour les besoins de certaines étapes de notre application, nous avons eu besoin de travailler sur un autre corpus, parce qu'alors nous avions besoin de comparer ou de discriminer les termes de notre corpus par rapport aux autres, nous avons eu recours à celui proposé par Al-Sulaiti[43] (**Atwell & al, 2004**). C'est un corpus construit en XML à l'université de Leeds. Le corpus est divisé en 16 catégories dont la science, le sport, les biographies, les histoires pour enfant etc.

La suite de ce chapitre sera organisée en quatre parties : La première sera consacrée à l'extraction des termes simples, la deuxième traitera l'extraction des termes composés, la troisième abordera l'extraction des relations, quant à la dernière, elle présentera la formalisation des concepts.

[43]http://www.comp.leeds.ac.uk/latifa/

Partie I : Extraction des termes simples

1. Introduction

Cette partie consiste à extraire des termes simples, l'approche utilisée se base sur le calcul de poids.

Ce poids est attribué à tous les mots sans distinction, il se calcule en utilisant la formule tf-idf (*voir section 3.2.1.1*). Nous savons que tf-idf est surtout utilisée dans la recherche d'information pour déterminer la pertinence d'un document par rapport à une requête mais qu'elle possède plusieurs variantes. Le principe de son utilisation dans ce travail est le suivant : Contrairement à la version d'origine, nos documents n'appartiennent pas au même corpus, nous prenons un document du Crescent Corpus, les autres documents sont pris dans quatre catégories différentes du corpus d'Al-Sulaiti (**Atwell & al, 2004**).

L'objectif est le suivant : si un mot *m* considéré est jugé important dans notre document cible, l'est aussi dans les autres documents, c'est qu'il n'est pas représentatif du domaine étudié, puisque les autres documents représentent des catégories différentes comme la politique, les biographies, l'environnement et le sport. Cela veut dire que c'est un mot commun à tous les domaines et donc ne peut être important pour le notre. Par contre si un mot est important dans notre document et ne l'est pas pour les autres et qu'il a un fort poids, il y a de forte chance qu'il soit représentatif du domaine. La liste retenue à la fin peut être validée par un expert humain.

2. La méthode proposée pour l'extraction des termes simples

2.1. Etapes d'extraction

Nous allons travailler sur les deux corpus, nous avons pris les versions en format brut. Chaque sourate est alors considérée comme un document et nous choisissons au hasard des documents des quatre catégories suscitées, à savoir *environnement*, *sport, politique* et *biographies*. Les termes représentatifs du Coran sont distincts de ceux utilisés dans des domaines comme la politique ou le sport, ce choix optimisait la méthode adoptée.

Nous devions choisir pour les besoins de l'approche basée sur tf-idf des sourates dont la taille doit être proche des documents des autres catégories. Cela ne nous a pas empêché de traiter quand même les grandes sourates comme Al-Baqarah, mais nous restons prudents quant au bon fonctionnement de la méthode.

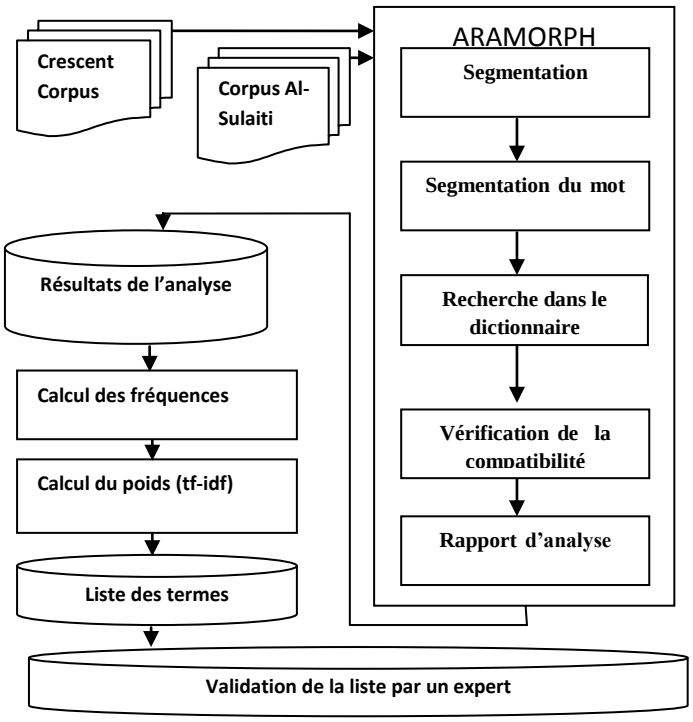

Figure 18: Etapes pour extraire des termes simples

2.2. Description des composants du système

2.2.1. Les corpus

Le Crescent Quranic corpus, représente le texte coranique comme nous l'avons expliqué plus haut, il existe une version sous format.txt et une autre sous format .xml. Il est composé de 114 sourates. Chaque sourate est considérée comme un document constituant le corpus.

2.2.2. Aramorph

Les documents provenant des deux corpus, sont analysés avec Aramorph[44]. AraMorph est un analyseur morphologique développé par Tim Buckwalter pour le compte du LDC[45] . Il prend comme entrée un texte sous format .txt et donne

[44] http://aramorph.sourceforge.net/
[45] http://www.ldc.upenn.edu/

comme sortie une liste de mots et toutes les solutions possibles de ses lemmes, la vocalisation, la morphologie, la catégorie, le glossaire et une analyse statistique qui donne le nombre des lignes, le nombre des mots arabes et le nombre de mots non arabes dans le texte.

Il existe deux versions d'AraMorph, celle en PERL , développé par *Buckwalter* et celle en JAVA, traduite par *Pierrick Brihaye* accessible en ligne[46]. Le projet inclut des classes *Java* permettant l'analyse morphologique de fichiers textuels en arabe et ce, quel que soit leur encodage. A cet effet, il est proposé 3 fichiers de test dans les principaux encodages utilisés pour la langue arabe : UTF-8, ISO-8859-6 et CP1256. Ce projet inclut également des classes compatibles avec l'architecture de Lucene, ce qui permet l'analyse, l'indexation et l'interrogation de documents en arabe.

2.2.2.1. ***Segmentation et segmentation du mot***

L'algorithme d'Aramorph, exécute une segmentation très basique. Les mots arabes sont définis en tant qu'un ou plusieurs caractères arabes contigus. Aramorph travaille sur du texte arabe translittéré.

Cette translittération utilise naturellement le système de translittération de Buckwalter. Ainsi, **كتاب** est translittéré en *ktAb* avant son analyse morphologique (**Banouni & al, 2002**).

Ensuite, AraMorph utilise un algorithme de force brute pour décomposer le mot en une succession de préfixe, radical et préfixe.

2.2.2.2. ***Recherche dans le dictionnaire***

Pour la recherche, il y a trois fichiers de lexiques : dictPrefixes, dictStems, et dictSuffixes. Si les trois composantes (préfixe, radical, suffixe) se trouvent dans leurs tables de hachage, la prochaine étape consiste à déterminer si leurs catégories respectives morphologiques sont compatibles.

2.2.2.3. ***Vérification de la compatibilité***

Le Format de fichier des tables est extrêmement simple. On a trois tableaux de compatibilité : tableAB, tableAC, et tableBC. Pour chacune des trois composantes (préfixe, radical, suffixe) compatibles il faut vérifier la demande :

[46] http://www.nongnu.org/aramorph

- A est une catégorie *préfixe* compatible avec la catégorie *radical* B (la paire existe dans la table de hachage AB)
- A est une catégorie *préfixe* compatible avec la catégorie *suffixe* C (la paire existe dans la table de hachage AC)
- B est une catégorie *radical* compatible avec la catégorie *suffixe* C? (la paire existe dans la table de hachage BC?)

Si les trois paires se trouvent dans leurs tables respectives, les trois composantes sont compatibles et le mot est valide.

2.2.2.4. ***Résultat d'Analyse***

Exemple pour le mot **كتاب**

كتاب est translitéré en *ktAb* avant son analyse morphologique

```
Processing token :      كتاب
Transliteration :       ktAb
SOLUTION #2
Lemma :         kut~Ab
Vocalized as :  kut~Ab
Morphology :
        prefix : Pref-0
        stem : N
        suffix : Suff-0
Grammatical category :
        stem : kut~Ab   NOUN
Glossed as :
        stem : kuttab (village school)/Quran school
SOLUTION #1
Lemma :         kitAb
Vocalized as :  kitAb
Morphology :
```

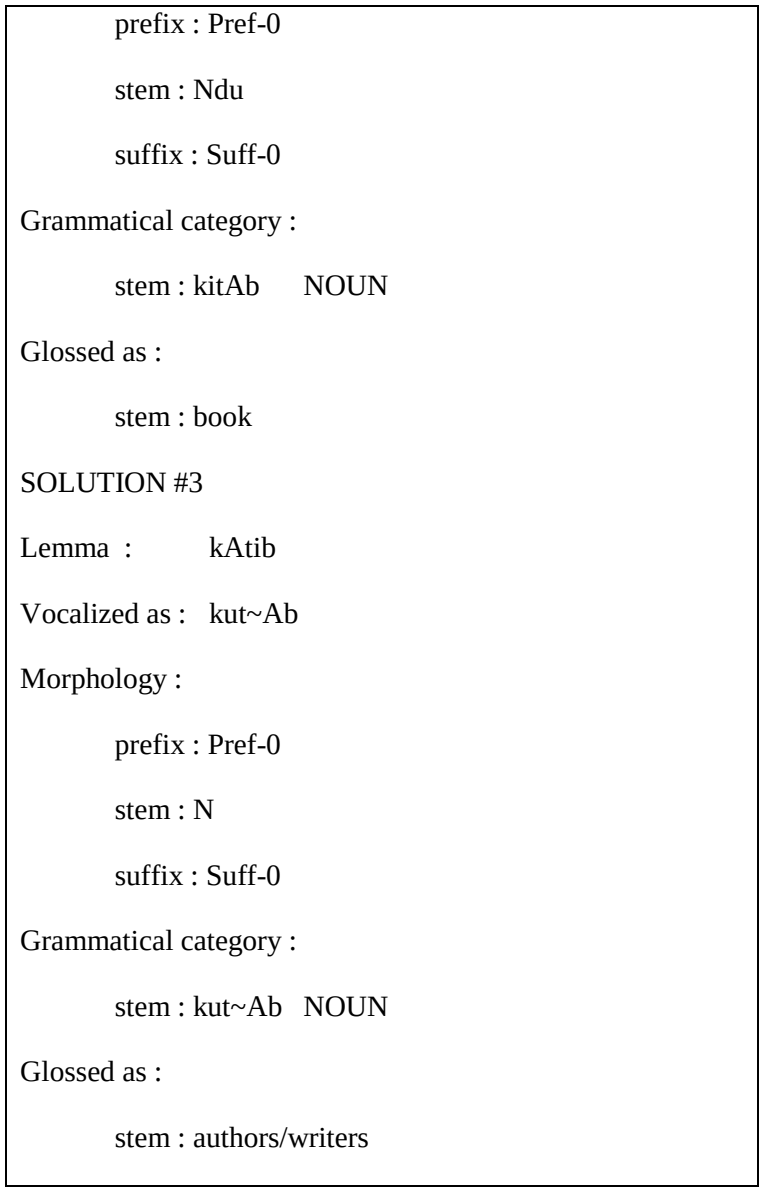

Figure 19: Exemple d'analyse avec Aramorph **(Lanani, 2009)**

2.2.3. Calcul de fréquence des mots

D'après le résultat d'analyse du texte coranique par l'analyseur AraMorph, on fait l'indexation du mot par rapport à ses lemmes[47]. Pour obtenir leurs fréquences dans la sourate du Coran avec le numéro du verset.

Nous suivons les étapes suivantes :

[47] Le lemme est une entrée dans le dictionnaire

- Les *verbes* sont ramenés à la 3e personne du singulier, de l'accompli actif, sauf dans le cas de certains verbes figés n'ayant qu'une conjugaison partielle, exemple : (**كتب**)

- Les *noms variables* sont ramenés à la forme du nominatif singulier. (masculin pour les noms qualificatifs), exemple : (**كاتب**)

- Les *noms invariables* à leur forme classique vocalisée (masculin singulier pour les pronoms), exemple : (**أنا**) ;

- Les *particules* sont ramenées à leur forme classique vocalisée.

- Les lemmes *composés* sont gardés comme composés si le composé lui-même fait l'objet d'une entrée dans les dictionnaires classiques.

A la fin nous obtenons le mot, sa fréquence et le nom de la sourate dans laquelle il apparait.

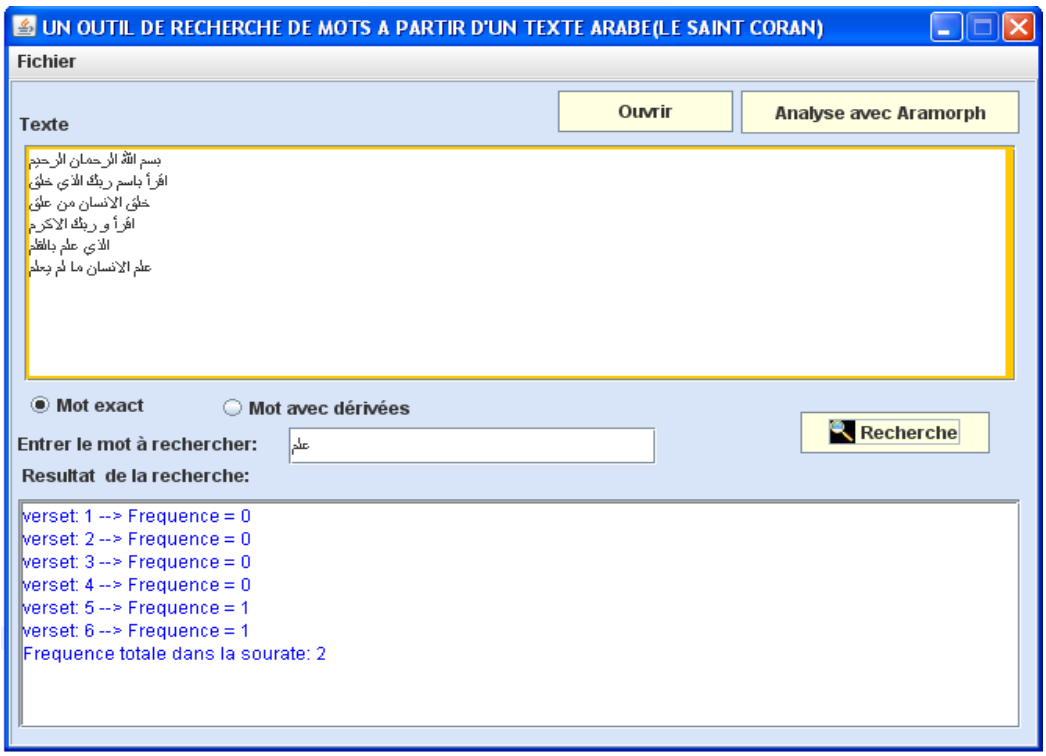

Figure 20: Recherche d'un mot dans le Coran à l'aide d'Aramorph **(Lounis, 2009)**

2.2.4. Calcul du poids des termes par la formule tf-idf

A l'aide de la fréquence nous allons procéder au calcul du poids de chaque mot avec la formule tf-idf.

Aramorph est ajouté dans la bibliothèque du projet, après l'analyse, on calcule pour chaque mot i dans un document j son poids w_{ij} par la formule suivante:

$$w_{ij} = tf_{ij} \times idf_i$$

Avec

$$idf_i = log \frac{N}{n_i}$$

Figure 21: Formule du tf-idf (**Bechet, 2009).**

$\boldsymbol{tf_{ij}}$: représente la fréquence du mot i dans le document j, calculée dans l'étape précédente.

$\boldsymbol{n_i}$ *:* est le nombre de documents dans lesquels apparaît le terme *i*.

$\boldsymbol{N}$ *:* Le nombre total des documents

Une fois les poids calculés, on classe les mots dans un ordre décroissant, on fixe un seuil d'une manière expérimentale (dans notre cas la valeur tournait autour de 0,6). On retient les mots dont le poids est égal ou supérieur au seuil. Cette liste représente les termes candidats et elle est soumise à un expert pour une validation éventuelle.

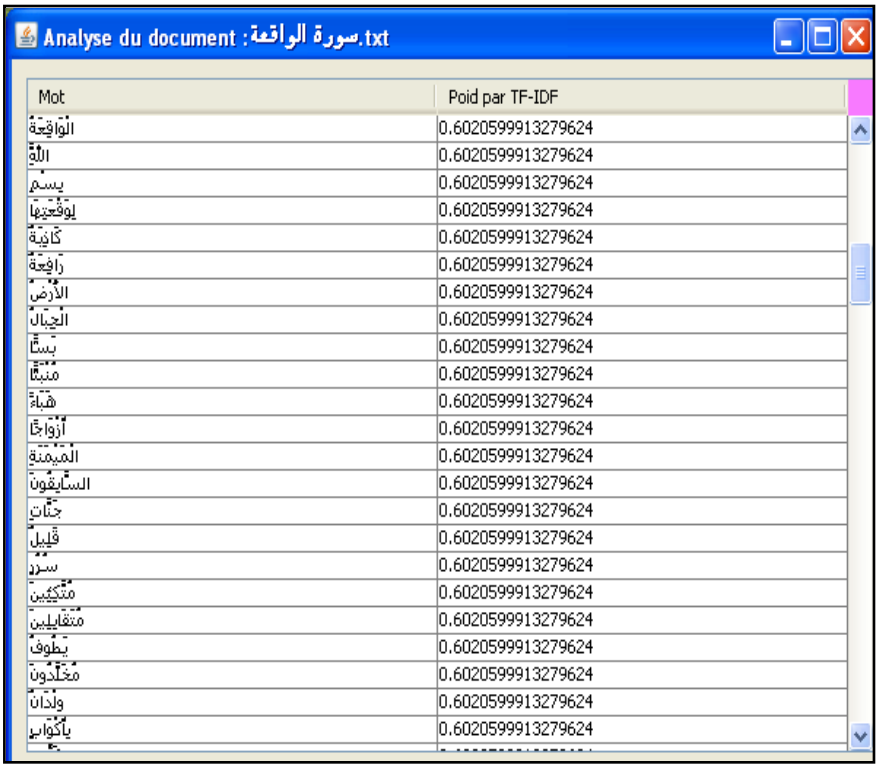

Figure 22: Liste des mots avec leur pondération tf-idf

Cette liste peut être validée par un expert du domaine qui décidera si un terme candidat est bien représentatif du domaine ou pas.

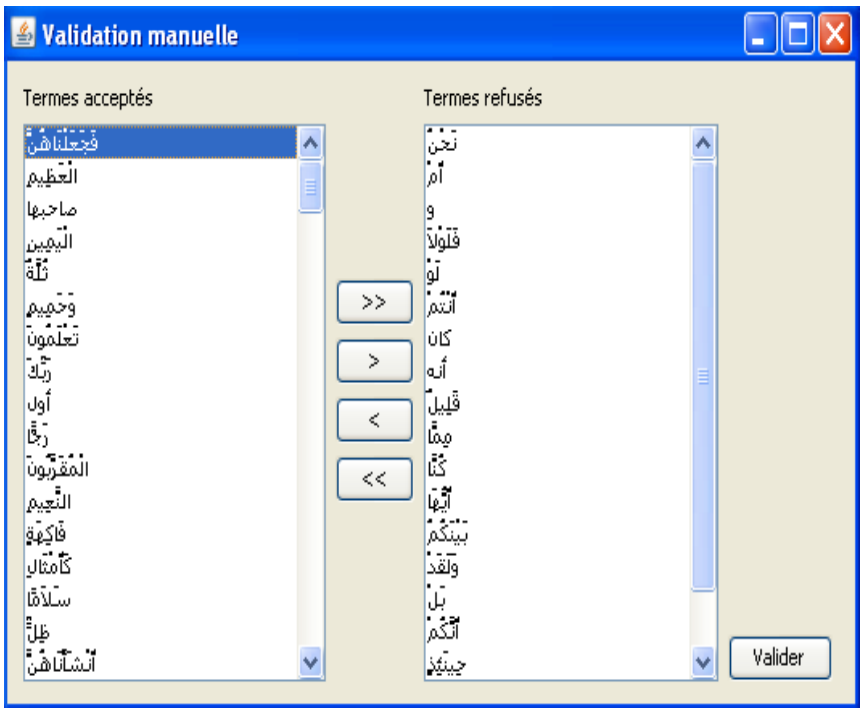

Figure 23: Validation manuelle de la part d'un expert

L'approche que nous avons privilégiée est une approche semi-automatique (**Zaidi & al., 2012a**), parce que si les méthodes statistiques sont réputées pour être robustes et faciles à mettre en œuvre, le résultat sans une validation manuelle est loin d'être convaincant du fait de l'absence quasi-totale des connaissances linguistiques.

3. Evaluation

Les mesures utilisées généralement, pour juger la justesse de l'extraction des termes sont la précision et le rappel.

La précision permet d'évaluer le nombre correct de termes extraits alors que le rappel permet d'évaluer la proportion des termes corrects qui n'ont pas été extraits. Les formules suivantes sont données en supposant qu'il existe une liste de référence. Dans le cas échéant un expert peut remplacer la liste.

$$\text{Précision} = \frac{\textbf{nombre de termes extraits et qui sont presents dans la liste de reference}}{\textbf{le nombre de termes extraits}}$$

$$\text{Rappel} = \frac{\textbf{nombre de termes extraits et qui sont presents dans la liste de reference}}{\textbf{nombre de termes de la liste de reference}}$$

La mesure de ces deux quantités a donné les résultats du tableau suivant :

Précision	Bruit	Rappel	Silence	F.Mesure
0.88	0.11	0.92	0.05	0.89

Tableau 9 : Précision et Rappel de l'approche adoptéele travail sur des sourates de

Bien que ces résultats soient représentatifs, toutefois ils restent à vérifier pour les longues sourates, car nous avons effectué l'analyse sur des sourates de moyenne longueur, telles que El-Insan, El-Waqiaa, El-Houjourat etc..

La méthode basée sur tf-idf ne fonctionne pas à son optimum si les documents ne sont pas homogènes, c'est-à-dire approximativement de même taille, les résultats peuvent devenir aléatoires. Outre cela nous avons remarqué que dès que nous chargeons des fichiers volumineux le système devient vite très lent jusqu'à se bloquer carrément.

4. Conclusion

Dans cette partie, nous avons présenté une méthode statistique basée sur le calcul de poids en fonction de la formule de tf-idf, pour l'extraction de termes simples du Coran. L'opération s'est faite après avoir analysé les documents avec Aramorph. Le système fournit à la fin une liste de termes candidat que nous pouvons comparer à une liste de référence ou alors la soumettre à un expert humain pour validation. Toutefois nous avons tenu à conserver les mots au dessous du seuil fixé, dans le cas où l'expert désirerait les rajouter à la liste finale.