Partie II : Extraction des termes composés (Collocations)

1. Introduction

Dans cette partie nous allons présenter une méthode hybride pour l'extraction de termes composés sous forme de collocations. L'objectif étant toujours de construire l'ontologie du Coran. A partir de ce chapitre nous n'allons travailler que sur le Crescent Quranic Corpus.

Les termes complexes sont extraits par une méthode linguistique en utilisant l'outil GATE, que nous avons adapté à l'Arabe et ce, en intégrant de nouvelles règles JAPE respectant les patrons syntaxiques arabes pour l'extraction de collocations. Nous gardons les collocations candidates dans un fichier. Par la suite, cette liste des collocations jugées pertinentes, est filtrée par une méthode statistique basée sur l'information mutuelle.

2. Description des composants du système

La figure26 montre les différentes étapes proposées pour l'extraction de collocations, nous décrivons dans ce qui suit les principaux composants du système.

2.1. La méthode linguistique

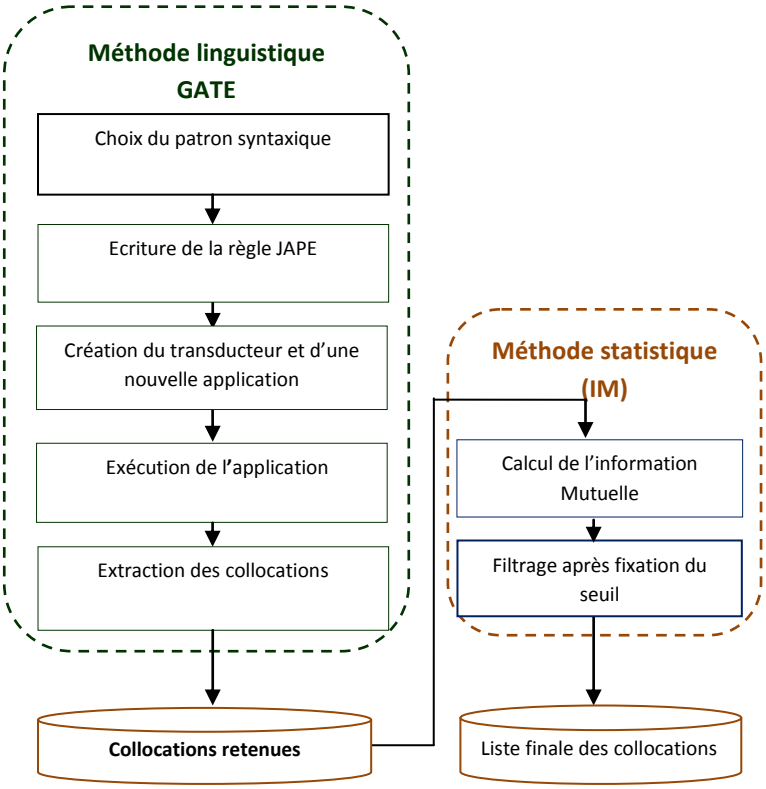


Figure 24: Architecture du système

2.1.1. Le choix des patrons syntaxiques

Les patrons syntaxiques sont le schéma que doit respecter une suite de mots pour qu'elle soit considérée comme une collocation.

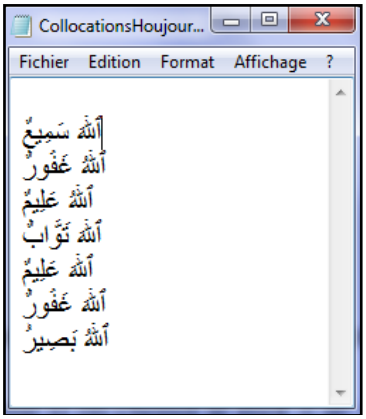
Une collocation étant « *l'emploi d'un terme relativement à d'autres, toutes variantes morphologiques confondues, et sans égard à la classe grammaticale* »[48].

Une collocation est « *la position d'un objet par rapport à d'autres au sein d'un ensemble, d'un mot par rapport à d'autres le long de la chaîne parlée*[49] ».

Une collocation est donc une expression à mots multiples c'est à dire des unités lexicales constituées par plusieurs mots orthographiques tels que *feu rouge* en Français ou « **أجر عظيم**» en Arabe.

Nous nous intéressons ici aux collocations formées de deux unités lexicales ou trois et respectant les schémas suivants :

- Nom-NomPropre (**رسول الله**)
- NomPropre-Nom(**الله العليم**)
- NomPropre-Adjectif (**الله عليم**) nous considérons que les adjectifs tels que (**عليم**) deviennent des noms, lorsqu'ils sont déterminés (**العليم**).
- Nom-Nom (**صوت النبي**)
- Adjectif-Nom (**سميع الدعاء**)
- Nom-Adjectif(**زرابي مبثوثة**)
- Nom-Préposition-Nom(**نور على نور**)
- Nom-Verbe (**الملائكة يشهدون**)
- Verbe-Nom (**يتّبع الرسول**)



[48] http://www.cnrtl.fr
[49] http://www.larousse.fr

Figure 25: Exemple de collocations sous forme (NomPropre-Adjectif)

Si le patron choisi est (Nom-NomPropre), le mot « الله » est alors le deuxième selon l'ordre de droite à gauche.

Figure 26: Exemple de collocations sous forme (Nom-NomPropre)

Des collocations respectant le schéma (Verbe-NomPropre) avec le NomPropre choisi toujours :الله . Nous obtenons :

Figure 27: Exemple de collocations sous forme (Verbe-NomPropre)

Il est à noter certains problèmes auxquels nous nous sommes confrontés, comme les particules agglutinées aux noms qui se sont affichées comme Nom-Nom ou Nom-Verbe et que nous avons refusé parce qu'alors le corpus était étiqueté ainsi.

2.2. GATE

GATE, comme nous l'avons présenté dans la *section 3.1.1* du chapitre 3, définit tout en termes de composants: des unités réutilisables spécialisées dans des taches spécifiques.

2.2.1. CREOLE (Collection of **RE**usable **O**bjects for **L**anguage **E**ngineering**).**

Nous pouvons définir dans CREOLE trois types de composants :

2.2.1.1. ***Ressources langagières*** **(**LRs : *language resources*)

Il s'agit d'un certain nombre de données linguistiques tels que des documents, des corpus ou des ontologies. A l'heure actuelle toutes les LRs sont basées sur le texte mais le modèle peut être étendu pour manipuler des données multimédias.

2.2.1.2. ***Ressources de traitement*** **(**PRs : processing resources)

Ce sont des ressources de caractère algorithmique tels que les segmenteurs, les étiqueteurs, les analyseurs etc. Dans la majorité des cas les PRs sont utilisées pour traiter les données fournies par les LRs.

2.2.1.3. ***Ressources de visualisation*** (VRs : visual resources)

Ce sont des composants graphiques affichés par l'interface utilisateur et permettant la visualisation et l'édition d'autres types de ressources.

2.2.2. ANNIE (A Nearly-New Information Extraction system)

GATE comporte un système d'extraction d'information, ANNIE, pour système quasi nouveau pour l'extraction d'information, lui-même formé de modules parmi lesquels un analyseur lexical, un gazetteer (lexique géographique), un segmenteur de phrases (avec désambiguïsation pour l'anglais), un étiqueteur (mais pas pour l'Arabe), un module d'extraction d'entités nommées et un module de détection de coréférences (**Cunningham & al, 2006**).

2.2.3. JAPE (Java Annotation Pattern Engine)

JAPE est un langage dérivé de CPSL (Common Pattern Specification Language) (**Plamondon, 2004**). Il fournit des transducteurs à états finis basés sur des expressions régulières, la grammaire JAPE consiste en un ensemble de phases, chacune d'elle est un ensemble de patron/règle. Les phases sont exécutées séquentiellement constituant une cascade de transducteurs à états finis pour les annotations. La partie gauche (LHS) de la règle contient le patron de l'annotation pouvant contenir des opérateurs d'expressions régulières (*****, **?**, **+**). La partie droite

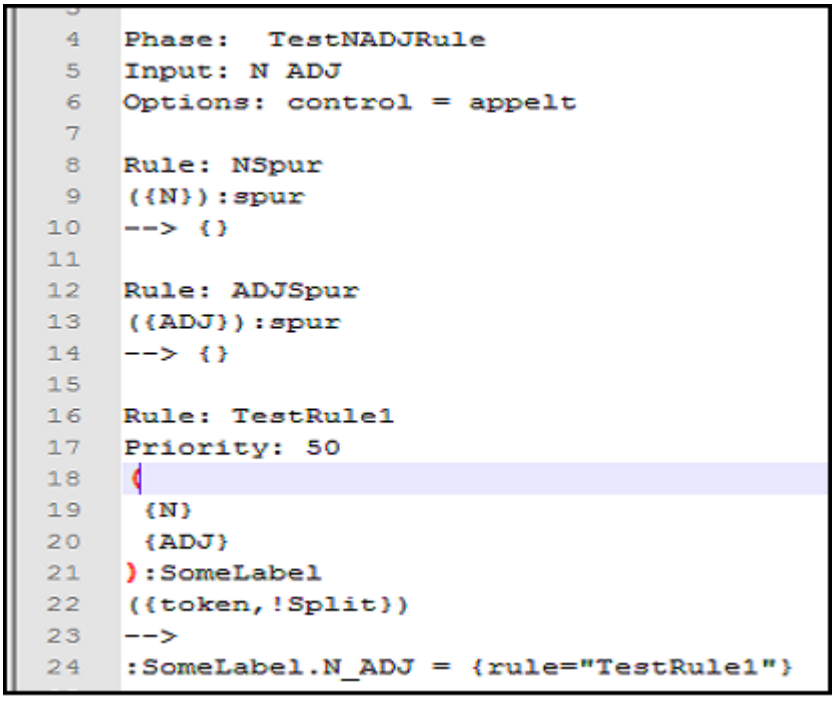
de la règle (RHS) donne le label de l'annotation, attaché au patron : l'action à entreprendre si le patron est détecté **(Thakker & al, 2009**).

2.3. Extraction de collocations avec GATE

Notre idée est d'essayer à l'aide de nouvelles règles JAPE d'extraire des collocations binaires ou ternaires (**Zaidi & al, 2010**b), dans le but de comparer les résultats avec la méthode du concordancier.

Nous avons commencé avec des règles simples pour définir de nouvelles annotations, telles que Nom-adjectif, Nom-Adjectif, Nom-Nom, Verbe-Nom, Nom-Préposition-Nom, à prendre de droite à gauche dans le sens de la langue arabe.

La règle suivante permet de reconnaitre dans un texte des mots avec une étiquette *Nom* suivi d'un *adjectif*, pour donner en sortie la collocation formée du N-ADJ, de la même façon nous avons écrit les règles permettant de reconnaitre les autres types de collocations.



```
4  Phase:  TestNADJRule
5  Input: N ADJ
6  Options: control = appelt
7
8  Rule: NSpur
9  ({N}):spur
10 --> {}
11
12 Rule: ADJSpur
13 ({ADJ}):spur
14 --> {}
15
16 Rule: TestRule1
17 Priority: 50
18 (
19  {N}
20  {ADJ}
21 ):SomeLabel
22 ({token,!Split})
23 -->
24 :SomeLabel.N_ADJ = {rule="TestRule1"}
```

Figure 28: Règle JAPE pour l'extraction de collocation (Nom-Adjectif)

Après la création d'un nouveau transducteur à état fini, nous lui passons la nouvelle règle comme paramètre et le corpus à traiter, le moteur va chercher dans

le texte tous les tokens dont l'étiquette est *Nom*, suivi d'un token dont l'étiquette est *Adjectif*.

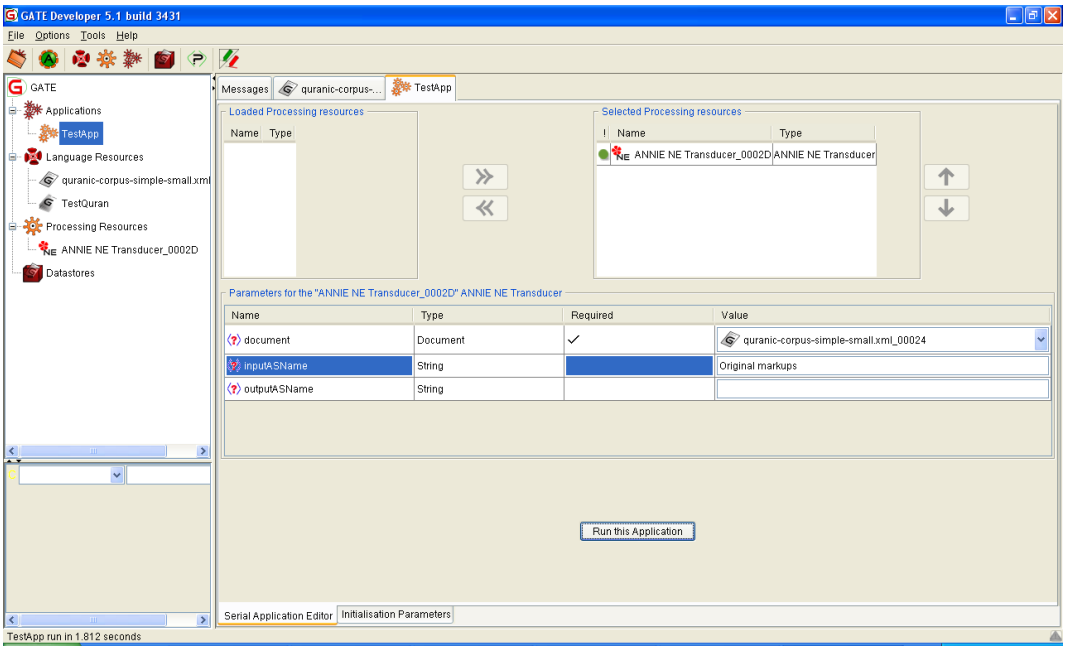


Figure 29: Création de transducteur dans GATE

Dans ce cas le corpus doit être obligatoirement étiqueté, parce que GATE ne dispose pas d'un étiqueteur pour l'Arabe, ce qui aura comme conséquence logique que la précision de l'analyse va dépendre étroitement de la précision de l'étiquetage.

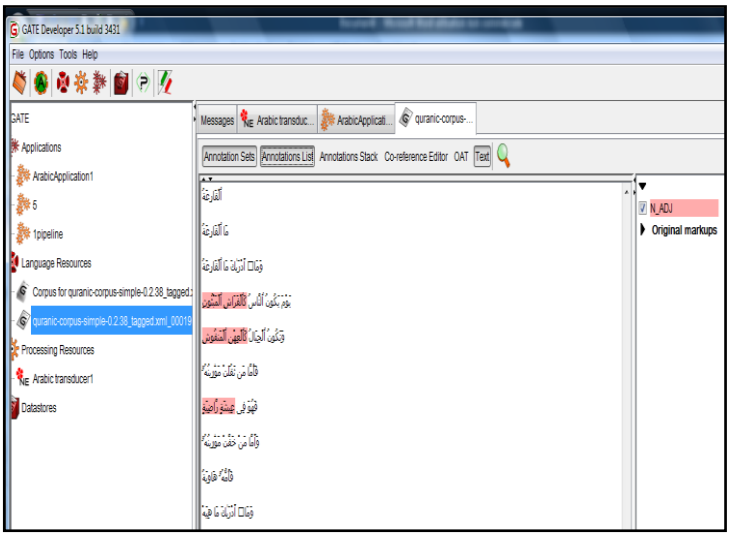


Figure 30: Extraction de collocations (Nom-Adjectif) avec GATE

En passant la règle à *ANNIE NE Transducer*, nous obtenons l'étiquette d'une nouvelle annotation N_ADJ. De la même façon nous pouvons créer d'autres annotations Nom-Nom, Verbe-Nom, Nom-Préposition-Nom etc. La figure suivante montre l'extraction de ce type de collocations.

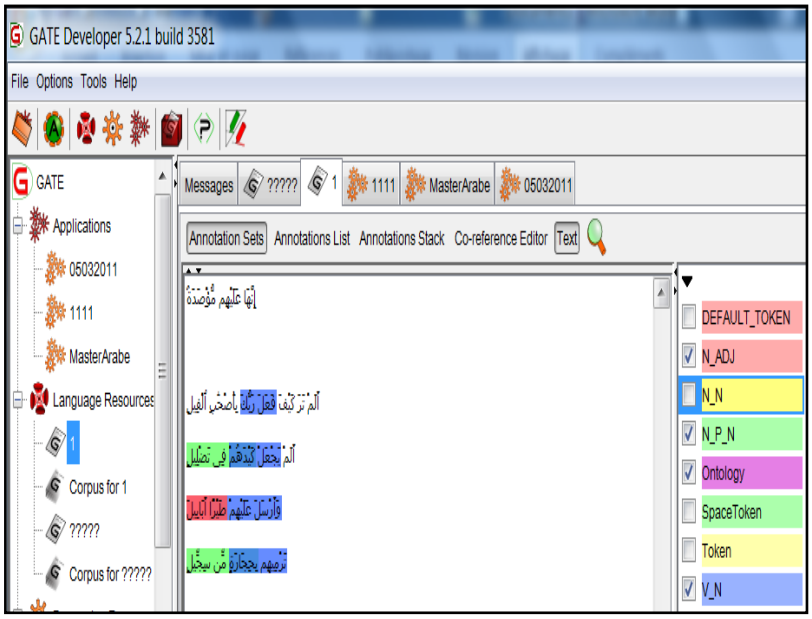


Figure 31: Extraction d'autres types de collocations

Nous avons effectué une évaluation primaire sur les annotations obtenues, à partir des mesures classiques de précision et de rappel sur la base d'un corpus annoté manuellement (Les sourates courtes). Puis, nous l'avons traité automatiquement avec Gate et avons effectué la comparaison à l'aide de *AnnotationDiff*[50] (**Maynard & Aswani, 2009**) qui a permis de donner un rappel de 1, où toutes les collocations ont été retournées, et une précision de 0.5 (environ la moitié a été jugée pertinente).

Le taux relativement faible de précision est due essentiellement au fait que l'étiquetage du corpus n'était pas très fin, ce qui a influé sur la précision calculée.

2.4. Le filtrage par calcul de l'information mutuelle

La liste sur laquelle nous travaillons maintenant est celle des collocations obtenues précédemment en utilisant les règles JAPE.

[50] Outil de GATE permettant d'effectuer une comparaison entre deux annotations

Nous avons calculé, pour chaque paire de mots l'information mutuelle (IM). L'IM sert à déterminer si deux mots sont fortement liés ou pas. Elle consiste à calculer le nombre d'occurrence de mots, ensuite calculer la probabilité de ces mots, considérés comme variables, en utilisant les fréquences.

La formule utilisée est la suivante : Etant donnés deux mots désignés par les variables x et y l'information mutuelle se calcule de la façon suivante :

$$IM(x, y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

Où :

P(x) et P(y) : sont respectivement les probabilités d'observation des mots x et y.

P(x, y) : est la probabilité de les observer ensemble.

Si f(x), respectivement f(y), sont les fréquences de x et y alors :

$$p(x) = \frac{f(x)}{N}, \quad p(y) = \frac{f(y)}{N} \quad \text{et} \quad p(x, y) = \frac{f(x,y)}{N}$$

Où N est le nombre total des mots dans le corpus.

D'où la formule finale: $IM(x, y) = \log_2 (\frac{N f(x,y)}{f(x)f(y)})$

Exemple de calcul de l'information mutuelle pour les deux mots **رسولٌ** et **الله**

Mot1	Fréqmot1	Mot2	Fréqmot2	Fréq(mot1,mot2)	IM(mot1,mot2)
رسول	5	الله	27	2	5.22

Tableau 10: Exemple de calcul de l'IM entre deux mots dans la sourate El- Houjourat

N= 353

IM (**ٱلْحَقُّ**, **رَبِّهِمْ**)=log (353*2)/ (5*27))=5.22

Une fois l'IM de chaque couple calculée, nous fixons expérimentalement un seuil pour privilégier les paires ayant une forte cohésion, nous classons les collocations selon leur IM par ordre décroissant et nous retenons celles dont l'IM est supérieure au seuil.

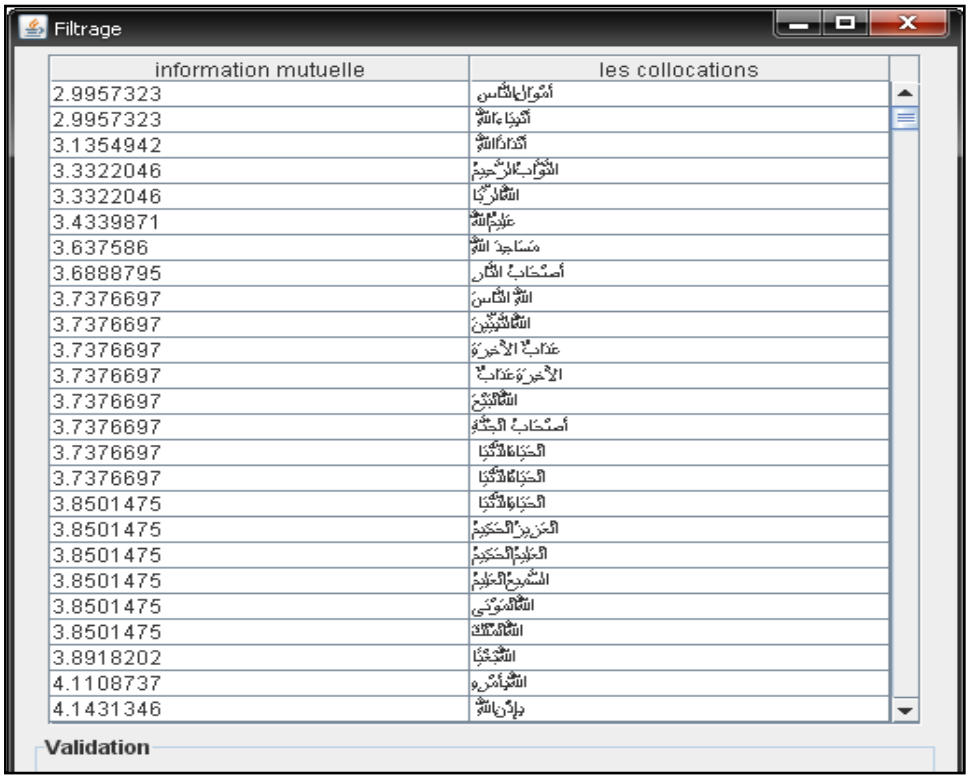


Figure 32: Calcul de l'information mutuelle des collocations

2.5. Evaluation de la méthode

Pour évaluer la pertinence de notre choix à savoir l'hybridation des deux approches, nous avons calculé la précision avant et après le filtrage par l'IM, nous avons utilisé la formule suivant :

$$\text{Précision avant filtrage} = \frac{nombre\ \ de\ collocations\ \ \boldsymbol{correctes}\ \ extraites}{nombre\ \ \boldsymbol{total}\ de\ collocations\ \ extraites\ \ par\ GATE}$$

$$\text{Précision après filtrage} = \frac{nombre\ \ de\ collocations\ \ \boldsymbol{correctes}\ apr\grave{e}s\ filtrage}{nombre\ \ \boldsymbol{total}\ de\ collocations\ \ extraires\ \ apres\ \ filtrage}$$

Le tableau suivant montre les résultats obtenus avant et après hybridation.

Type d'extraction	Précision
Extraction avec GATE (linguistique)	0.54
Extraction après hybridation (statistique)	0.91

Tableau 11:Précision avant et apres hybridation

Le fait marquant que nous avons observé c'est au lieu d'avoir une amélioration avec des valeurs de l'IM supérieure au seuil, comme il est de coutume, nous avons pu remarquer que pour les très faibles valeurs de l'IM comme pour celles supérieures au seuil fixé au début, les collocations étaient en majorité correctes, celle qui étaient comprises entre les deux valeurs dont celle du seuil l'étaient beaucoup moins, nous avons cherché à comprendre ce phénomène et voir s'il est exclusivement réservé au corpus coranique, vu ses caractéristiques propres. Mais pour cela, il fallait refaire tout le travail avec d'autres corpus, c'est ce que nous avons prévu de faire dans de futurs travaux. Cependant nous avons relevé quelques remarques concernant ces résultats, le Coran étant un corpus spécial avec ses spécificités et ses caractéristiques la cooccurrence de deux mots même si elle est très rare ne veut forcement pas dire qu'elles ne forment pas une collocation pertinente.

D'après les résultats affichés au tableau10, nous remarquons que la précision a connu une nette amélioration après le filtrage par la méthode statistique cela explique largement la pertinence de l'hybridation et que le choix que nous avons fait en couplant les deux approches est plus que justifié.

3. Conclusion

Cette partie a traité le problème d'extraction de termes simples et composés sous forme de collocations à partir du corpus coranique.

L'extraction des termes est le point anguleux de la construction d'ontologies. En effet, de la qualité des termes extraits, va dépendre celle des relations à extraire et donc la structure même de l'ontologie. Nous n'avons pas mentionné le fait que parallèlement à GATE, nous avons essayé une autre méthode linguistique basée sur l'utilisation d'un concordancier aConCorde[51], qui consiste à, étant donné un corpus chargé, affiche tous ses mots par ordre alphabétique, ainsi que la fréquence de chaque mot. Lorsqu'un mot est choisi, aConCorde (**Roberts & al, 2005**) 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**.**

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

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. Voyant que les résultats obtenus n'amélioraient en aucune façon ceux issus de GATE, nous avons préféré en faire abstraction. Mais nous le mentionnant ici, parce que nous pensons que c'est le propre même de la recherche que de prendre plusieurs chemins et de découvrir que certains ne mènent pas forcément là où on veut aller. Cependant cette démarche n'est jamais complètement vaine, nous en sortons souvent plus éclairés.

L'opération d'extraction telle qu'abordée ici n'est pas complètement automatique, le recours à un expert est quasi nécessaire à la fin pour valider la liste finale des termes extraits.