Accepted in the 5th International Conference on Arabic Language Processing CITALA, Oujda, Morocco 11/ 2014

Une nouvelle approche pour l'annotation sémantique des corpus Arabes

Mohammed NASRI*, Lahsen ABOUENOUR, Karim BOUZOUBAA
Département d'Informatique
Ecole Mohammadia d'Ingénieurs, Université Mohammed V
Rabat, Maroc
mohammed.nasri@gmail.com, abouenour@yahoo.fr, karim.bouzoubaa@emi.ac.ma

Adil KABBAJ
Département d'Informatique
Institut National des Statistiques et d'Economie Appliquée, INSEA
Rabat, Maroc
akabbaj@insea.ac.ma

***Abstract*— L'annotation de corpus par des métadonnées linguistiques significatives est devenu un sujet de recherche de plus en plus important. En effet, l'information linguistique est considérée comme élément fondamental dans de nombreuses applications de Traitement des Langages Naturels (TLN) telles que la Recherche d'Information ou les systèmes de Question Réponse. Dans cet article, nous présentons une nouvelle approche pour l'annotation de corpus arabes par des métadonnées sémantiques. A la différence des approches d'extraction de l'information sémantique relative à chaque mot séparément (word sense), notre approche se base sur une analyse sémantique profonde des phrases des corpus le résultat de cette analyse est repris en tant que métadonnée à utiliser dans l'annotation de la phrase. Pour ce faire, nous avons développé un processus d'analyse sémantique de texte arabes. Dans ce processus, nous commençons par construire une ontologie Arabe servant de ressource sémantique que nous utilisons en combinaison avec d'autres outils linguistiques (Analyseurs syntaxique et morphologique) afin d'extraire la sémantique de textes arabes. Dans l'objectif d'évaluer cette approche, nous avons mené des expérimentations en appliquant cette nouvelle approche d'analyse sur deux corpus différents. Les résultats obtenus montrent une couverture moyenne d'environ 25% ce qui est encourageant pour l'utilisation de cette approche d'analyse aussi bien dans les applications avancées de TLN que dans l'annotation de corpus par des métadonnées sémantiques.**

Keywords—Annotation sémantique ; Analyse sémantique de l'arabe; Graphes Conceptuels; Plateforme Amine; Ontologie Arabe

I. INTRODUCTION

Le Traitement des Langages Naturels (TLN) a connu un développement intéressant au cours des dernières décennies. Pour les langues latines répandues telles que l'Anglais et le Français, plusieurs systèmes avancés de gestion de connaissances sont développés et sont actuellement à une étape assez avancée. A ce titre, Soriano [1] cite plusieurs travaux dans ce sens notamment dans le domaine de Q/R. S'agissant de la langue Arabe, des recherches intéressantes dans le développement de certains systèmes ont été effectuées, particulièrement dans la morphologie [2, 3], dans les systèmes de la Recherche d'Information (RI) [4] et de Q/R [5, 6, 7], etc. Cependant, ces recherches n'ont pas encore atteint le même niveau d'avancement que celles concernant les langues latines.

Les difficultés rencontrées dans le TLN de l'Arabe sont de plusieurs niveaux: 1) de la langue elle-même, l'Arabe étant une langue à plusieurs particularités, se caractérisant par une morphologie complexe, une syntaxe riche, 2) d'un nombre limité d'outils linguistiques computationnels (analyseurs syntaxiques, sémantique, etc.) et des corpus.

Plusieurs travaux sont menés dans l'objectif de rendre les corpus plus consistants et leurs contenus plus significatifs, en les enrichissant par de l'information linguistique [8, 9, 10, 11]. Ceci peut s'avérer d'une utilité particulière pour les analyseurs automatiques, pour les processus de validation de nouvelles approches ou dans le calcul de niveau de similarité de deux ou plusieurs approches.

Dans cet article, nous présentons une nouvelle approche pour l'annotation de corpus arabes par de l'information sémantique. Contrairement aux approches existantes et qui annotent les corpus par de l'information syntaxique [8, 9] ou morphologique [10, 11], notre approche porte sur l'annotation des phrases des corpus Arabes par de l'information sémantique convoyées par ces phrases Pour ce faire, nous avons développé un processus d'analyse sémantique de texte arabes qui peut être capitalisé aussi bien dans l'extraction de l'information sémantique et l'annotation de corpus que dans la réalisation d'applications avancées pour le TLN telles que les systèmes de RI ou de Q/R.

Notre approche d'analyse sémantique consiste, dans un premier temps (et pour une fois uniquement), à créer une ontologie arabe servant de ressource sémantique pour le processus d'analyse. La construction de cette ontologie s'est basée sur le contenu des deux librairies Arabic WordNet (AWN) [12] et Arabic VerbNet (AVN) [13, 14]. Cette nouvelle ressource présente l'avantage de combiner la haute couverture lexicale et les relations sémantiques entre les mots existants dans AWN avec la représentation formelle des structures syntaxiques et sémantiques correspondant aux verbes en AVN. Ces structures (fournies par AVN) ont été formulées en Graphes Conceptuels (GC) [15] et attachées en tant que situations aux verbes de l'ontologie.

Pour chaque analyse sémantique, l'analyseur commence par une analyse syntaxique et morphologique du texte arabe en utilisant le parseur syntaxique de Stanford [16, 17] et l'analyseur morphologique AlKhalil [2] (que nous utilisons grâce à leur disponibilité et à la disponibilité de leurs

Le travail de recherche de Lahsen Abouenour a été soutenu par la subvention du CODESRIA (Conseil pour le Développement de la Recherche en Sciences Sociales en Afrique), Référence N° SGRT. 38/T13.

documentations), qui étiquette (tag) les constituants du texte avec des rôles thématiques et produit les modèles syntaxiques de la phrase. Ces modèles syntaxiques associés à l'information sémantique fournie par AVN et stockée dans l'ontologie permettent d'extraire l'information sémantique du texte donné. Une formulation préliminaire et basique de cette approche est donnée dans l'article [18].

Le reste de l'article est organisé comme suit: la section 2 présente quelques travaux connexes, la section 3 introduit les outils utilisés dans le développement de ce travail. Les sections 4 et 5 décrivent respectivement l'approche de l'analyse sémantique et la méthodologie proposée pour l'annotation de corpus. La section 6 présente les expérimentations menées. Enfin, dans le section 7 nous tirons les principales conclusions et nous citons quelques perspectives.

II. TRAVAUX CONNEXES

L'annotation sémantique de texte arabe fût déjà l'objet de quelques travaux de recherche. Dans son article [19], Alrahabi introduit son approche d'annotation sémantique des phrases d'énonciations dans le texte Arabe. L'auteur se base sur la localisation d'éléments (mots) lexicaux dans le texte Arabe et ne procède pas par une analyse sémantique profonde du texte.

Le projet OntoNotes [20], quant à lui, fournit des corpus riches pour les trois langues Anglaise, Chinoise et Arabe. Ce projet annote les corpus par de l'information syntaxique et morphologique (Arabic Treebank) et par de l'information sémantique relative à chaque mot (Word sense). Ce projet offre des corpus importants dans le sens où chacun est accompagné de l'information syntaxique, morphologique et sémantique.

Une autre approche d'analyse sémantique très similaire à la nôtre a été introduite par Sameh Alansary et al., [21] dans le cadre de développement d'un analyseur linguistique de la langue Arabe appelé IAN (Interactive ANalyzer). Il s'agit d'un système d'analyse de la langue Arabe qui représente les phrases en langage naturel à des constituants morphologiques, puis des arbres syntaxiques et enfin des réseaux sémantiques au format UNL [22]. Cette approche a été adopté par son auteur dans le cadre de l'annotation de textes Arabe par de l'information sémantique.

Premièrement, ces approches sont basées sur l'analyse syntaxique (tagging) pour l'extraction de la connaissance. Par conséquent, le résultat est composé des mots du texte et les balises renvoyées par l'analyseur syntaxique, ce qui fournit une sémantique superficielle. En effet, si le texte du document "T1" et le texte recherché (ou la question) "T2" utilisent différemment deux synonymes pour le même verbe, alors l'opération de recherche échoue, puisque les résultats d'analyse (GC ou UNL) du texte "T1" et "T2" ne contiennent pas les mêmes mots. Dans notre approche, ce problème ne se pose pas puisque les verbes ne sont pas présents dans leur forme dans le CG représentant la sémantique, mais sont remplacés par des valeurs de prédicat (extraites de la librairie AVN) qui sont les mêmes pour les verbes synonymes.

Deuxièmement, notre approche diffère par le fait qu'en plus de l'approche d'analyse, nous proposons une ontologie Arabe composée du contenu d'AVN et d'AWN, de telle ontologie permet de stocker des informations sémantiques supplémentaires telles que les définitions, les canons, les règles et les situations et peut être séparément utilisée dans diverses applications du TLN incluant la RI ou les systèmes de Q/R.

Enfin et à notre connaissance, notre approche et celle de Sameh et al., sont les seules conçues pour la langue arabe (caractérisé par ses particularités difficiles telles que sa morphologie complexe, sa forte inflexion, etc.) et qui intègre une analyse morphologique de mots dans le texte traité. Cela permet l'extraction de l'information avec une précision plus élevée.

III. OUTILS ET FORMALISMES UTILISÉS

Pour l'aboutissement de ce travail, nous avons adopté le formalisme de Graphe Conceptuel [15] pour la représentation sémantique de la connaissance et nous avons intégré et utilisé certains outils/plateformes que nous présentons brièvement dans les sous sections suivantes.

A. *Le formalisme de Graphe Conceptuel*

Introduit par John Sowa [15], la théorie des Graphes Conceptuels (GC) offre un formalisme puissant pour la formulation de connaissances et la rédaction des spécifications. Il permet d'exprimer l'information dans une forme (logiquement) précise, (humainement) lisible et (informatiquement) traitable et manipulable. La figure 1 montre un exemple de GC pour la phrase « رأى الطفل القمر » (L'enfant a vu la lune).

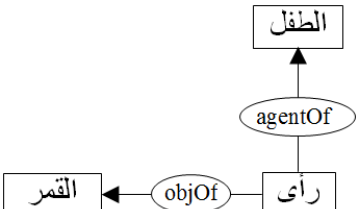


Fig. 1. Exemple de GC

Ce GC peut être lu comme: « **l'agent de** l'action رأى **(voir)** est **الطفل (l'enfant)** et son **objet** est **القمر (la lune)** ».

Les avantages des GCs ne se limitent pas uniquement à leur capacité de représentation de connaissances. En effet, les opérations qui peuvent être effectuées par combinaisons de deux (2) ou plusieurs GCs telles que la projection et la jointure sont à la base de la résolution de divers problèmes dans divers champs de recherche tels que l'extraction d'information, les raisonnements complexes ou encore les systèmes de Q/R.

Depuis leur introduction, les GCs ont connu un développement important et sont actuellement adoptés dans différents champs de recherche tels que le traitement des langages naturels [23], les bases de données sémantiques [24], les systèmes de base de connaissances [25], les systèmes d'informations [26], les systèmes multi-agents [27], l'écriture des spécifications, etc. Les GCs sont adoptés par Sonetto [28] qui a été considéré par Sowa [15] l'un des plus grands systèmes industriels adoptant ce formalisme. Pour ces raisons, nous avons adopté ce formalisme dans notre travail pour exprimer la sémantique extraite par notre processus d'analyse.

B. *La base de données léxicale Arabic WordNet*

La base de données lexicale Arabic WordNet (AWN) [12] est une ressource lexicale gratuite pour l'Arabe Standard Moderne (ASM). Il est basé sur la conception et le contenu de Princeton WordNet (PWN) [29] et permet la traduction « bidirectionnelle » sur le plan lexical avec des dizaines d'autres langues.

C. *La ressource Arabic VerbNet*

Arabic VerbNet (AVN) [13, 14] est l'un des premiers lexiques de verbes, qui classent les verbes les plus utilisés de l'ASM. Ce lexique dans son état actuel couvre 336 classes qui contiennent 7744 verbes et 1399 frames fournissant des informations sur les verbes ainsi que des frames de description syntaxique et sémantique de verbes. Chaque classe AVN définit une liste de membres (des verbes arabes), une liste des rôles thématiques possibles et une liste de frames syntaxiques.

D. *Le parseur syntaxique de Stanford*

Développé par Dan Klein et Christopher Manning, le parseur de Stanford [16, 17] est un parseur syntaxique du langage naturel. Dans sa version actuelle, ce parseur traite l'Anglais, le Chinois et l'Arabe. En résultat de l'analyse, ce parseur génère une structure Treebank pour la phrase analysée ainsi que les dépendances entre ses composants.

E. *L'analyseur morphologique AlKhalil*

Alkhalil [2] est un analyseur morphosyntaxique des mots de l'ASM. Il traite les textes non voyellés ainsi que celles partiellement ou totalement voyellés. Son approche est basée sur la modélisation d'un très grand nombre de règles morphologiques arabes ainsi que sur l'intégration des ressources linguistiques utiles pour l'analyse.

F. *La plateforme Amine*

Développée par Adil Kabbaj [30, 27] pendant une vingtaine d'années, la plateforme Amine est un Environnement de Développement Intégré (IDE), se caractérisant par son architecture multicouche et adapté à la programmation symbolique, la programmation de systèmes et d'agents intelligents. Amine permet le développement de nombreux types de systèmes intelligents tels que les systèmes de base de connaissances, les applications basées sur le concept d'ontologie, les applications du TLN, les applications basées sur le raisonnement, etc.[1]

Amine implémente le formalisme de GC. Elle permet non seulement la création et la manipulation des GCs, mais aussi leur utilisation dans de nombreux processus basés sur les ontologies tels que la RI, le raisonnement, etc.

Amine est utilisée dans notre travail en tant que plateforme implémentant les ontologies ainsi que la manipulation des GCs.

IV. L'APPROCHE D'ANALYSE SÉMANTIQUE

Comme mentionné ci-dessus, notre approche commence par la construction de l'ontologie arabe qui s'appuie sur le contenu d'AWN et AVN. Cette ontologie combine les couvertures lexicales et les relations sémantiques entre les mots existants dans la libraire AWN avec la représentation formelle des structures syntaxiques et sémantiques correspondant aux verbes d'AVN.

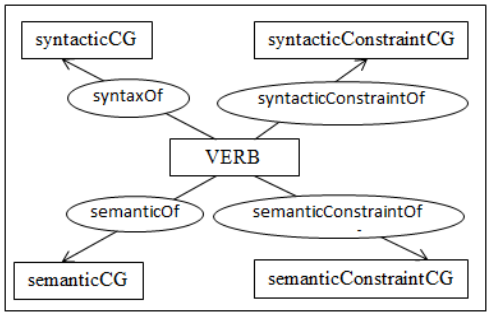
Dans cet article, nous présentons uniquement notre approche d'analyse sémantique et nous ne détaillons pas le processus de construction de l'ontologie. En effet, cette construction fût l'objet d'un travail précédemment publié [31] et n'est pas reprise dans cet article.

Rappelons que cette ontologie se constitue d'une hiérarchie de mots extraits d'AWN et des modèles syntaxico-sémantique sous forme de GCs attachés en tant que situations[2] aux concepts de l'ontologie correspondants aux verbes d'AVN.

Le modèle de ces situations est affiché dans la figure 2 et se compose généralement de 4 sous GCs : le GC syntaxique (syntacticCG), le GC des contraintes syntaxiques (syntacticConstraintCG), le GC des contraintes sémantiques (semanticConstraintCG) et le GC sémantique (semanticCG) respectivement liés au verbe par 4 relations: syntaxOf, syntacticConstraintOf, semanticConstraintOf et semanticOf.

Le premier sous GC (syntacticCG) convoie la structure syntaxique sous la forme d'une succession d'unités syntaxiques (voir l'exemple de la figure 3), si cette succession ne correspond pas à sa structure syntaxique d'une phrase donnée, cette situation n'est pas sélectionnée par le processus d'analyse. Le deuxième quant à lui, se compose des contraintes syntaxiques (exemple, le deuxième NP (Noun Phrase ou Phrase nominale) doit commencer par « أنّ » ou par « ما ») à appliquer aux unités syntaxiques de la phrase pour éliminer les situations qui ne correspondent pas. Le troisième se compose des contraintes sémantiques appliquées pour filtrer les situations (précédemment sélectionnées selon la structure et les contraintes syntaxiques) se basant sur des contraintes sémantiques sur les constituants de la phrase. Enfin le quatrième sous GC représente le modèle sémantique qui est inféré si la situation est retenue.

Ces situations peuvent donc être interprétées comme des correspondances entre le modèle syntaxique d'utilisation du verbe et l'information sémantique qu'il véhicule. Autrement dit, avec de telles structures, une fois le modèle syntaxique d'une phrase donnée est déterminé, sa sémantique peut être inférée. Il s'agit essentiellement de l'approche adoptée par l'analyseur sémantique que nous décrivons dans les sous sections suivantes.



[1] Pour plus d'information, prière de se rendre sur le site web de la plateforme Amine: http://amine-platform.sourceforge.net/

[2] Les situations sont des structures conceptuelles pouvant être attachées aux concepts de notre ontologie, pour plus d'information, voir la documentation de la plateforme Amine : http://amine-platform.sourceforge.net/component/kernel/ontology.htm

Fig. 2. Le modèle général des situations

La figure 3 montre un exemple de situation du verbe «رأى».

```
[super : *c1 ] -
   -syntaxOf->[cg : [np : *c2 ] -
                    -followedBy->[np : *c3 ],
                    <-followedBy-[super : ?c1 ]
                ],
   -syntacticConstraintOf->[list :"[?c3(Oan~a_comp)]" ],
   -semanticConstraintOf->[list :"[?c2(animate)]" ],
   -semanticOf->[cg : [event : *p1 ] -
                         -duringOf->[cg : [perceive : *p2 ] -
                                          -experiencerOf->[np : ?c2 ],
                                          -stimulusOf->[np : ?c3 ]
                                    ],
                         -inReactionTo->[np : ?c3 ]
                   ]
```

Fig. 3. Exemple de situation du verbe "رأى"

Notre approche d'analyse sémantique est basée sur six étapes décrites dans les sous sections suivantes et illustrées dans la figure 4 en utilisant l'exemple de la phrase arabe suivante: " رأى الطفل القمر "

A. *Construction des modèles syntaxiques de la phrase*

Dans cette première phase, le processus extrait le modèle syntaxique correspondant à la phrase à analyser et le transforme en GC, à comparer, dans la prochaine étape, aux sous GC syntaxiques des situations sélectionnées dans la phase B.

L'objectif de cette étape est donc d'extraire les modèles syntaxiques (s'il y en a plusieurs) de la phrase et de les transformer en GCs. A partir de la phrase de l'exemple (رأى الطفل القمر), nous cherchons à extraire la structure: «V NP NP » et à générer la GC suivant:

[verb] -followedBy->[NP] -followedBy->[NP]

Cette étape est divisée en deux sous étapes: 1) l'extraction du modèle syntaxique et 2) la construction du GC syntaxique correspondant. L'extraction du modèle syntaxique commence par l'analyse syntaxique de la phrase en utilisant l'analyseur de Stanford, qui fournit une structure Treebank. Dans ce qui suit, la structure Treebank donné par le parseur de Stanford pour la phrase de l'exemple:

```
(ROOT
  (S
    (VP (VBD راى)
      (NP (DTNN الطفل ))
      (NP (DTNN القمر)))))
```

Cette structure subit un traitement intermédiaire spécifique qui conduit à l'extraction du verbe principal ainsi que les autres unités syntaxiques (NP) et enfin à la détermination de la structure syntaxique désirée « V NP NP », tout en gardant en mémoire que V correspond au verbe "راى" et les autres NPs correspondent respectivement aux mots "الطفل" et "القمر".

Ces constituants nécessitent souvent une analyse morphologique supplémentaire afin d'en extraire leurs lemmes. En effet, 1) les verbes sont stockés dans l'ontologie dans leur forme infinitive (sous forme de lemme). En langue Arabe, cette forme est similaire au temps du passé et à la troisième personne du singulier. Ces verbes figurent très rarement sous cette forme dans les textes arabes, ils sont souvent conjugués ou attachés à des proclitiques ou à des enclitiques, ce qui rend leur identification et leur localisation dans l'ontologie quasi impossible, 2) certains mots/verbes portent confusion parce qu'ils contiennent des lettres pouvant être considérées comme des proclitiques ou des enclitiques, c'est le cas par exemple du verbe « وفـر » (WFR) qui signifie « économiser » ou « et s'est enfuit » selon que la première lettre «و» fasse partie du verbe ou soit une proclitique. De telles confusions sont relevées durant l'analyse morphologique, qui donne toutes les décompositions possibles d'un mot donnée. Dans notre approche, si les constituants disposent de plusieurs décompositions, nous générons autant de structures syntaxiques que nécessaire pour couvrir toutes ces décompositions. Voici un exemple concernant l'analyse de « وفـر الرجل المال », nous avons souligné avant que « وفـر » peut désigner « وفـر » (économiser) ou « و » (et) suivi de « فـر » (s'est enfui), alors que « الرجل » peut désigner lui aussi « الرَجُلُ » (l'homme) ou « الرِّجْلُ » (le pied), quatre combinaisons sont alors possibles : « وَفـَرَّ الرَّجُلُ المال «, » فـَرَّ الرِّجْلُ المال «, » وَفـَرَّ الرِّجْلُ المال«, » فـَرَّ الرَّجُلُ المال«».

A partir de chaque structure, nous générons automatiquement le GC correspondant, que nous appelons GC syntaxique. Les GCs syntaxiques générés pour l'exemple du modèle syntaxique précédant sont tous comme suit:

[verb:*c1]-followedBy->[NP:*c2]-followedBy->[NP:*c3]

Ce GC peut être interprété comme suit: le verbe référencé par c1 est suivi d'un premier NP référencé par c2 qui est, lui aussi, suivi par un second NP référencé par c3. Ces références (c1, c2 et c3) sont gardées en mémoire du processus pour les étapes suivantes.

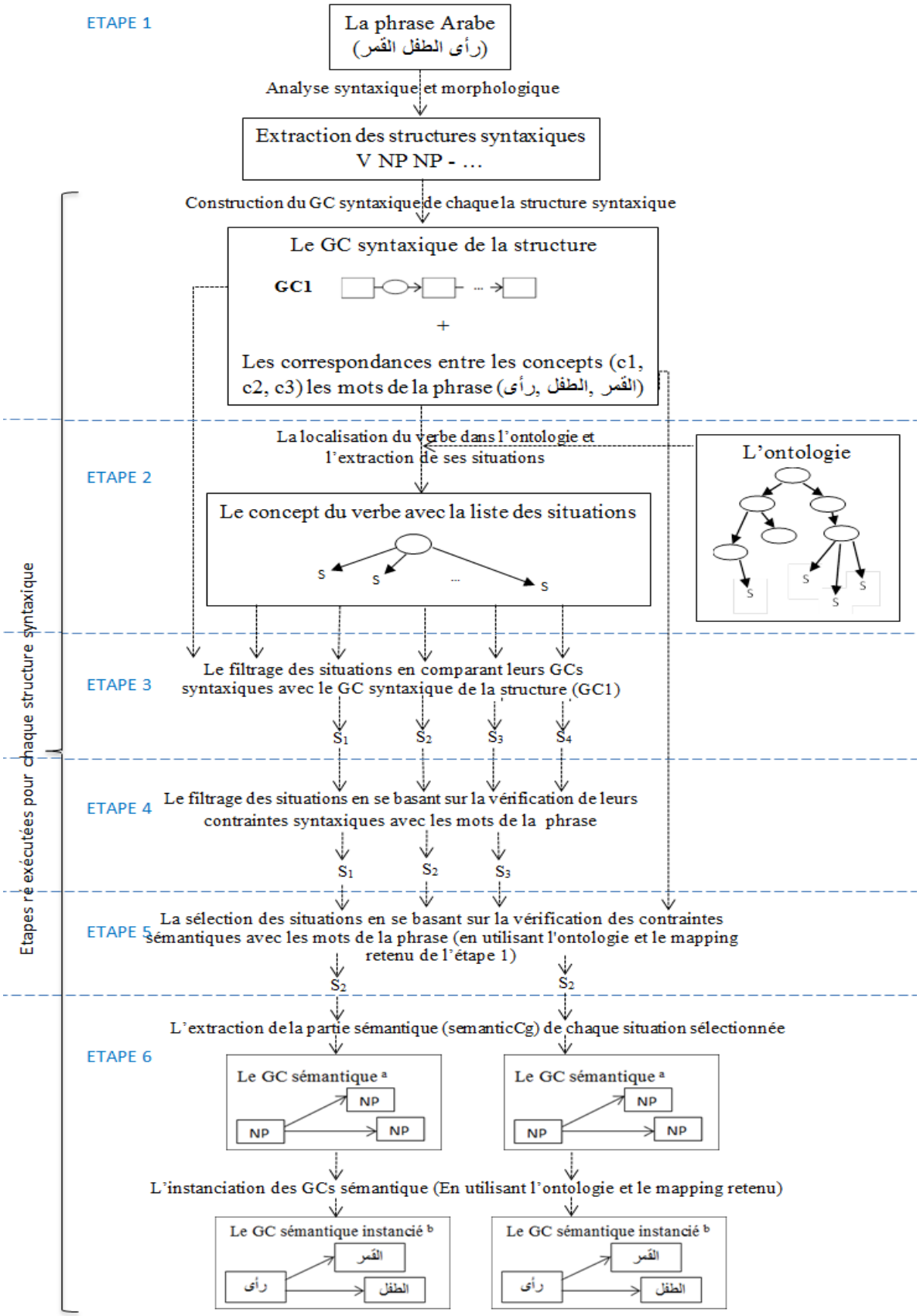


Fig. 4. Les étapes d'analyse sémantique.

[a, b] les GCs ne sont que des exemples et ne correspondent pas au vraies parties sémantiques ni au parties sémantique sinstanciées des situations sélectionnées.

Pour chaque GC syntaxique généré, nous gardons en mémoire les correspondances entres c1, c2, c3 et les constituants correspondants.

Une fois que les modèles syntaxiques de la phrase sont extraits et que leurs GCs syntaxiques sont générés, ces derniers subissent, un par un, le reste du processus.

B. *Identification du verbe principal et extraction des situations de l'ontologie*

L'objectif de cette deuxième étape est d'identifier le verbe principal de la phrase, de le localiser dans l'ontologie et d'extraire toutes ses situations. A partir du résultat de la phase précédente, nous extrayons le lemme du verbe principal, nous le localisons dans l'ontologie et nous extrayons toutes ses situations, qui correspondent sémantiquement aux cas d'utilisation de ce verbe (correspondances syntaxique-sémantique). Si le verbe n'existe pas dans l'ontologie ou ne dispose pas de situations, le GC syntaxique est rejeté et le processus passe au GC syntaxique suivant. Dans le cas du verbe "رأى", nous extrayons sept (7) situations que nous illustrons dans le tableau 1.

TABLE 1. LISTE DES SITUATIONS DU VERBE "رأى"

Sit.	*GC de la situation*
Sit. 0	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]], -syntacticConstraintOf->[list :"[?c3(kayofa_extract)]"], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]
Sit. 1	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]], -syntacticConstraintOf->[list :"[?c3(maA_extract)]"], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]
Sit. 2	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3]-followedBy->[np: *c4], <-followedBy-[super : ?c1]], -syntacticConstraintOf->[list :"[?c4(sentential)]"], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]
Sit. 3	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]], -syntacticConstraintOf->[list :"[?c3(Oan~a_comp)]"], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]
Sit. 4	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3]-followedBy->[prep :"**في**"]- -followedBy->[np: *c4], <-followedBy-[super : ?c1]], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]
Sit. 5	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3]-followedBy->[np: *c4], <-followedBy-[super : ?c1]], -syntacticConstraintOf->[list :"[?c4(participle)]"], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]
Sit. 6	[super : *c1] - -syntaxOf->[cg : [np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]], -semanticConstraintOf->[list :"[?c2(animate)]"], -semanticOf->[cg : [event : *p1] - -duringOf->[cg : [perceive : *p2] - -experiencerOf->[np : ?c2], -stimulusOf->[np : ?c3]], -inReactionTo->[np : ?c3]]

C. *Vérification syntaxique*

Maintenant que le GC correspondant au modèle syntaxique de la phrase est généré, il est comparé à chaque GC syntaxique des situations extraites de l'ontologie. Si la structure des deux GCs est la même (même succession d'unités syntaxiques), la situation est sélectionnée, sinon elle est rejetée.

Chaque situation du verbe "رأى" va donc subir une vérification entre son sous GC syntaxique et le GC syntaxique de la phrase. Le tableau 2 montre la vérification syntaxique des syntacticCg des situations avec la syntacticCG de la phrase d'exemple.

TABLE 2. LA VÉRIFICATION DU MODÈLE SYNTAXIQUE DES SITUATIONS DU VERBE "رأى" AVEC LE SYNTACTICCG DE LA PHRASE EXEMPLE

Situation	Syntactic CG	Résultat
Sit. 0	[np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]	Succès
Sit. 1	[np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]	Succès
Sit. 2	[np : *c2] - -followedBy->[np : *c3]-followedBy->[np: *c4], <-followedBy-[super : ?c1]	Echec
Sit. 3	[np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]	Succès
Sit. 4	[np : *c2] - -followedBy->[np : *c3]-followedBy->[prep :"في"]-followedBy->[np: *c4], <-followedBy-[super : ?c1]	Echec
Sit. 5	[np : *c2] - -followedBy->[np : *c3]-followedBy->[np: *c4], <-followedBy-[super : ?c1]	Echec
Sit. 6	[np : *c2] - -followedBy->[np : *c3], <-followedBy-[super : ?c1]	Succès

Appliqué aux situations du verbe, cette étape va rejeter les situations 2, 4 et 5 et conserve les situations 0, 1, 3 et 6.

Il se peut qu'aucune situation ne soit sélectionnée, dans ce cas, le processus ignore ce verbe ainsi que ces situations et passe au traitement du GC syntaxique suivant.

D. *Vérification des contraintes syntaxiques*

Cette étape consiste à éliminer les situations qui ont des restrictions syntaxiques ne s'appliquant pas à la structure de la phrase. Ceci est réalisé par un filtrage sur la base de contraintes syntaxiques extraites du sous GC syntacticConstraintCg de chaque situation, appliquées à l'ensemble des constituants de la phrase. Ainsi, pour chaque situation, nous extrayons les restrictions contenues dans ce sous GC. Si ces restrictions correspondent aux constituants de la phrase, la situation est sélectionnée, sinon, elle est rejetée. Le tableau 3 affiche les syntacticConstraintCgs des situations du verbe "رأى" ainsi que les résultats de vérification avec la phrase exemple.

TABLE 3. LES SYNTACTICCONSTRAINTCG DES SITUATIONS SÉLECTIONNÉES ET LE RÉSULTAT DE VALIDATION AVEC LA PHRASE EXEMPLE

Sit.	syntacticConstraintCg	Signification	Résultat
Sit. 0	[?c3(kayofa_extract)]	c3 commence par kayofa (كيف)	Echoue
Sit. 1	[?c3(maA_extract)]	c3 commence par maA (ما)	Echoue
Sit. 3	[?c3(Oan~a_comp)]	c3 commence par Oan~a (أنّ)	Echoue
Sit. 6	*empty*	Pas de restriction syntaxique sur c3	Réussit

Les contraintes syntaxiques des situations 0 et 1 désignent respectivement que l'unité syntaxique référencée par c3 est une phrase subordonnée jointe à la phrase principale via le mot « كيف - how » ou « ما – what ». La contrainte de la situation 3 précise quant à elle que c3 référence une phrase complément de « أن - that » et la situation 6 ne porte aucune contrainte syntaxique. Ainsi, il est évident que la phrase ne valide aucune des contraintes syntaxiques des situations 0, 1 et 2 qui seront donc rejetées et la seule situation à retenir est la situation numéro 6.

E. *Vérification sémantique*

Après l'étape de vérification syntaxique, les situations sélectionnées sont soumises à une vérification sémantique, autrement dit, une vérification basée sur les types des constituants de la phrase. Ceci se fait en appliquant les contraintes véhiculées par le GC représentant les contraintes sémantiques « semanticConstraintCG » de chaque situation (sélectionnée) aux composants du GC syntaxique généré (i.e. de la phrase). Par exemple, une contrainte pour la NP (c2) qui suit immédiatement le verbe est d'être « animate » ou « organisation ». Pour assurer cette vérification, le processus localise le mot (correspondant à ce NP) dans l'ontologie et vérifie que le concept correspondant est une spécialisation (un sous type) du concept en question (« animate » ou « organisation » par exemple).

F. *Extraction du GC sémantique*

Maintenant que les bonnes situations sont choisies[3] (après la vérification du modèle syntaxique, les contraintes syntaxiques et les contraintes sémantiques), il est temps d'extraire de chacune le modèle sémantique (le sous GC sémantique) et de l'instancier. Nous commençons donc par 1) extraire la partie sémantique du GC, puis nous 2) remplaçons les références dans ce GC par les concepts correspondants et enfin nous 3) instancions ce GC en remplaçant les concepts NP dans le GC par les GCs correspondant à la sémantique de ces NPs dans la phrase. Dans notre exemple, chaque NP est restreint à un seul mot. Dans ce cas, nous substituons simplement chaque NP par le concept correspondant au mot dans la phrase en se basant sur le mapping retenu de l'étape 3 entre les références des concepts (c1, c2 et c3) et les mots de la phrase (رأى, الطفل, القمر).

Le CG sémantique extrait dans notre exemple est:

```
[event : *p1 ] -
      -duringOf->[cg : [perceive : *p2 ] -
                              -experiencerOf->[np : ?c2 ],
                              -stimulusOf->[np : ?c3 ]
                             ],
      -inReactionTo->[np : ?c3 ]
```

Après son instanciation, ce GC devient:

```
[event : *p1 ] -
      -duringOf->[cg : [perceive : *p2 ] -
                        -experiencerOf->[ طفل_id9256 :
*c2 ],
                        -stimulusOf->[ قمر: *c3 ]
                             ],
      -inReactionTo->[ قمر: ?c3 ]
```

[3] Dans la plupart des cas, une seule situation est sélectionnée, mais ce n'est pas général. Isolé de son contexte, une phrase peut avoir plus qu'un sens.

Ce qui signifie que le concept «طفـل» (enfant) « perçoit » le concept «قمر » (lune) en réaction de ce dernier («قمر »).

Durant le processus d'instanciation du GC, si un mot de la phrase réfère à une Entité Nommée (EN) (individu, endroit ou objet particulier), le processus localise l'individu correspondant dans l'ontologie et construit ainsi le concept adéquat. L'image est un peu différente quand la phrase contient une EN contenant plusieurs mots, ce genre d'ENs peuvent ne pas être identifiées au moment de l'analyse syntaxique et donc ne pas être reconnues en tant qu'ENs mais comme une succession de mots, c'est le cas par exemple de l'EN « الولاياتالمتحدة الامريكية». Dans de telles situations, ces ENs devront être identifiées et remplacées par des abréviations ou des acronymes afin d'avoir une analyse syntaxique correcte. Le traitement de ce genre d'ENs n'entre pas dans le périmètre de cet article. Néanmoins, pour une amélioration de la représentation en GC, leur intégration est à envisager.

V. METHODOLOGIE D'ANNOTATION DE CORPUS

Dans le but d'évaluer la fiabilité de notre approche d'analyse sémantique et de mesurer sa couverture par rapport aux corpus à annoter par de l'information sémantique, nous l'avons appliqué sur deux corpus arabes différents :

- Le premier est le corpus d'AVN qui a été construit par le premier auteur en se basant sur les exemples d'illustration fournis par la ressource AVN. Ce corpus contient 1344 phrases.
- Le deuxième est le corpus annoté OntoNotes [20], ce corpus rassemble différents genres de textes (news, diffusion, blogs, groupes de discussion, conversations téléphoniques, etc.) en trois langues (Anglais, Chinois, Arabe) annotés par des informations structurelles (syntaxe) et sémantiques de surface (sens des mots). OntoNotes s'appuie sur deux ressources : Penn Treebank pour la syntaxe et Penn PropBank pour la structure des arguments de prédicat. OntoNotes comprend environ 1,5 millions de mots anglais, 800 milles chinois et 300 milles mots arabes. Plus de détails sont fournis dans [32].

L'annotation de corpus par de l'information sémantique se base sur trois étapes, pour chaque phrase du corpus, il faut 1) extraire les sémantiques possibles, 2) générer les graphes sémantiques équivalents et 3) annoter la phrase par ces graphes sémantiques.

Les deux premières étapes sont garanties par l'analyseur sémantique présenté dans la section III et une fois le graphe sémantique est généré, il peut faire l'objet d'une ancre à utiliser pour annoter la phrase en question. L'utilisation d'une ancre s'avère une bonne solution lorsque les métadonnées utilisées ont un format textuel qui, injectées dans le texte d'origine, ne perturbe pas la lisibilité ni la lecture. Ce n'est malheureusement pas le cas lorsque ces métadonnées sont formulées en termes de GC. Il est préférable, dans ce cas, que les corpus contenant des phrases Arabes séparés, tel que le corpus d'AVN, soient annotés en insérant le GC immédiatement après la phrase en question. Alors que pour les corpus composés de texte entiers, tels qu'OntoNotes, il serait préférable de rajouter une couche spécifique pour l'annotation sémantique, contenant uniquement les phrases, pour lesquelles l'analyse sémantique a réussi, suivies du GC résultat de leurs analyses.

VI. EXPERIMENTATIONS

Les premières expérimentations que nous avons menées ont concerné uniquement les phrases se constituant de 2 et 10 mots. Les phrases de plus de 10 mots ont une structure compliquée et seront considérés dans des expérimentations futures. Dans ces expérimentations, nous nous sommes focalisé sur les 3 indicateurs suivants:

- Le nombre de phrases verbales **V** dans le corpus,
- Le nombre de phrases pour lesquelles le verbe principal est couvert par l'ontologie Arabe et dispose bien de situations **VC** et
- Le nombre de phrases analysées avec succès et pour lesquelles un GC sémantique a été généré **VCAn**.

Les résultats de ces expérimentations sont affichés dans les deux tableaux 4 et 5.

TABLE 4. RÉSULTAT D'EXPÉRIMENTATIONS SUR LE CORPUS D'AVN

Taille de la phrase	**V**	**VC**	***VC % V***	**VCAn**	**VCAn % VC**	**VCAn *% V***
2	138	62	*45%*	**60**	*97%*	***43%***
3	257	128	*50%*	**107**	*84%*	***42%***
4	295	145	*49%*	**114**	*79%*	***39%***
5	303	177	*58%*	**111**	*63%*	***37%***
6	165	82	*99%*	**51**	*62%*	***61%***
7	78	51	*65%*	**25**	*49%*	***32%***
8	47	23	*49%*	**10**	*43%*	***21%***
9	23	14	*61%*	**7**	*50%*	***30%***
10	5	2	*40%*	**1**	*50%*	***20%***
Avg.	**1311**	**684**	***52%***	**486**	***71%***	***37%***

TABLE 5. RÉSULTAT D'EXPÉRIMENTATIONS SUR LE CORPUS D'ONTONOTES

Taille de la phrase	**V**	**VC**	***VC % V***	**VCAn**	**VCAn % VC**	**VCAn *% V***
2	78	36	*46%*	**10**	*28%*	***13%***
3	239	113	*47%*	**55**	*49%*	***23%***
4	287	159	*55%*	**81**	*51%*	***28%***
5	308	157	*51%*	**50**	*32%*	***16%***
6	357	222	*62%*	**56**	*25%*	***16%***
7	401	188	*47%*	**55**	*29%*	***14%***
8	408	277	*68%*	**70**	*25%*	***17%***
9	482	334	*69%*	**164**	*49%*	***34%***
10	472	302	*64%*	**47**	*16%*	***10%***
Avg.	**3032**	**1788**	***59%***	**588**	***33%***	***19%***

Nous remarquons, à partir de ces résultats, que l'ontologie que nous avons construite couvre 52% des principaux verbes des phrases verbales du premier corpus (AVN) et 59 % du second (OntoNotes), ce qui fait une couverture moyenne de

57% [4] des principaux verbes des phrases verbales des deux corpus.

Parmi cette catégorie (les phrases verbales dont le verbe est couvert par l'ontologie), environ 71% des phrases d'AVN et 33% des phrases d'OntoNotes sont analysées avec succès, ce qui fait une moyenne totale d'environ 43%. Par rapport aux nombre total des phrases verbales, environ 37% des phrases d'AVN et 19% de cette d'OntoNotes sont analysées avec succès, donc une moyenne de 25%.

Ces résultats montrent que la présente approche peut être utilisée, en l'améliorant et la fiabilisant encore plus, dans l'annotation sémantique des corpus Arabe tels que OntoNotes.

Pour conclure, nous illustrons dans ce qui suit le résultat de l'analyse sémantique de deux phrases Arabe extraites de ces deux corpus.

Exemple 1 : « تناولنا الطعام[5] (Nous avons mangé la nourriture) »:

GC résultant:

```
[event]-duringOf->[cg : [take_in : *p1 ] -
                        -agentOf->[np : *c2  "نحن" ],
                        -patientOf->[طعام_id8543 : *c3 ]
                ] ,
        -causeOf->[np : ?c2  "نحن"]
```

Explication: Un évènement "take_in" (manger) a eu lieu, don't l'agent et le patient (objet) sont respectivement les deux concepts "نحن" (nous) et "طعام" (la nourriture).

Exemple 2 : « وزع الباحث بيانين[6] (Le chercheur a distribué deux articles) » :

GC résultant:

```
[event : *p1 ] -
  -duringOf->[cg : [transfer_info : *p2 ] -
            -agentOf->[باحث_id2023 : ?c2 ],
            -topicOf->[بيان_id2210 :*c3]-cardinalityOf-
>[cardinality :"=2"]
          ],
  -causeOf->[باحث_id2023 : *c2 ]
```

Explication: Un évènement "transfer_info" (transférer une information) a eu lieu, dont l'agent et le sujet (objet) sont respectivement les deux concepts " باحث_id2023" (chercheur) et "بيان_id2210" (article). Le GC donne une information supplémentaire précisant que la cardinalité du deuxième concept " باحث_id2023" est 2 (بيانين).

VII. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons présenté notre nouvelle approche pour l'annotation de corpus arabes par de l'information sémantique. A l'encontre des approches précédemment citées, la présente approche se caractérise par une analyse sémantique profonde des textes Arabe, basée sur la construction d'une ontologie arabe ainsi qu'un processus d'analyse sémantique (pouvant être utilisés dans d'autres applications de TLN telles que dans la RI ou dans les systèmes de Q/R). L'information sémantique extraite par le processus d'analyse peut être capitalisée pour annoter les phrases des corpus Arabe tels qu'OntoNotes.

Les avantages du système résultent de son approche prometteuse, en voici quelques-uns:

- L'analyse sémantique des textes arabes,
- La représentation de l'information sémantique dans le formalisme de GC qui est un formalisme puissant et prometteur surtout avec une plateforme basée sur les GCs comme Amine,
- L'intégration d'autres ressources (ontologie) et outils (analyseurs) qui, plus ils évoluent, le système devient plus fiable et
- Sa capacité d'être utilisé séparément dans les applications avancées de TLN telles que la RI ou les systèmes de Q/R.

Cependant, cette approche présente certains inconvénients:

- Les résultats d'analyse sont très dépendants de la performance et la fiabilité des outils utilisés, en particulier l'analyseur de Stanford et l'analyseur Alkhalil,
- La sémantique extraite du texte Arabe et utilisée dans l'annotation dépend de la sémantique définie par AVN,
- Cette approche ne prend en considération que les phrases verbales et dont le verbe principal est couvert par AVN. Ces verbes ne représentent que 60,53% des verbes d'AWN et
- La sémantique véhiculées par AVN est formulée en terme d'un ensemble limité de prédicats et ne maintient pas les verbes utilisés dans le texte arabe, ce qui mène à une perte de précision et rend la tâche difficile pour paraphraser l'information sémantique extraite, en particulier dans les systèmes de Q/R lorsque la réponse est communiquée à l'utilisateur final en langue naturelle.

Comme perspectives, nous visons à améliorer notre approche d'analyse par:

- L'élimination des inconvénients ci-dessus,
- L'amélioration de la composante de reconnaissance des entités nommées
- La considération de phrases plus longues et plus complexes et
- L'amélioration de l'approche pour la considération des questions afin de capitaliser ce travail pour le développement de systèmes de Q/R basés sur la sémantique.

[4] La moyenne n'est pas calculée sur la base des deux moyennes (52% / 2 + 59% / 2) puisque les deux corpus ne contiennent pas le même nombre de phrases, mais elle est calculée sur la base du nombre total de phrases: (684 + 1788) / (1311 + 3032).

[5] Extrait du corpus d'AVN

[6] Extrait du corpus OntoNotes

REFERENCES

[1] Soriano, J. M. G., y Gómez, M. M., Arnal, E. S., & Rosso, P. "A passage retrieval system for multilingual question answering. In Text, Speech and Dialogue". Springer Berlin Heidelberg. pp. 443-450, January , 2005.

[2] Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Ould Abdallahi Ould Bebah, M., and Shoul, M.: "Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts" In the proceedings of the 11th International Arab Conference on Information Technology. Benghazi, Libya, 2010.

[3] Buckwalter, T. "Buckwalter {Arabic} Morphological Analyzer Version 1.0", 2002.

[4] Abu El-Khair, I. " Arabic information retrieval". *Annual review of information science and technology*, *41*(1), pp. 505-533, 2007.

[5] Hammo, B., Abu-Salem, H., & Lytinen, S. "QARAB: A question answering system to support the Arabic language". In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*. Association for Computational Linguistics. pp. 1-11, July, 2002.

[6] Mohammed, F. A., Nasser, K., & Harb, H. M. "A knowledge based Arabic question answering system (AQAS)". *ACM SIGART Bulletin*, *4*(4), pp. 21-30, 1993

[7] Benajiba, Y., & Rosso, P. "Arabic question answering. Diploma of advanced studies". Technical University of Valencia, Spain. 2007.

[8] Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. "The penn arabic treebank: Building a large-scale annotated arabic corpus". In *NEMLAR conference on Arabic language resources and tools*. pp. 102-109, September, 2004.

[9] Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., & Beška, E. "Prague Arabic dependency treebank: Development in data and tools". In *Proceedings of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools,* pp. 110-117, September, 2004

[10] Dukes, K., & Habash, N. "Morphological Annotation of Quranic Arabic". In *LREC*, May, 2010.

[11] Alkuhlani, S., & Habash, N. "A corpus for modeling morpho-syntactic agreement in Arabic: gender, number and rationality". In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 357-362. Association for Computational Linguistics, June, 2011.

[12] Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., and Alkhalifa, M. "Arabic WordNet and the challenges of Arabic". In the Proceedings of the Arabic NLP/MT Conference, London, UK, 2006.

[13] Mousser, J. "A Large Coverage Verb Taxonomy for Arabic". In the Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 2675- 2681, 2010.

[14] Mousser, J. "Classifying Arabic verbs using sibling classes". In the Proceedings of the 9th International Conference on Computational Semantics, Association for Computational Linguistics, Oxford, UK, pp. 355-359, 2011.

[15] Sowa, J., F. "Conceptual Graphs". In. Van Harmelen, F., Lifschitz, V., and Porter, B (eds.) Handbook of Knowledge Representation. Chapter 5, pp. 213-237. Elsevier, 2008.

[16] Klein, D., and Manning, C., D. "Fast exact inference with a factored model for natural language parsing". In the Proceedings of the 15th annual conference on Neural Information Processing Systems, British Columbia, Canada, pp. 3-10, 2002.

[17] Klein, D., and Manning, C., D. "Accurate unlexicalized parsing". In the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Sapporo Convention Center, Sapporo, Japan, Volume 1, pp. 423-430, 2003.

[18] Nasri, M., Abouenour, L., Kabbaj, A., and Bouzoubaa, K.,: Toward a Semantic Analyzer for Arabic Language. In the Proceedings of the 22nd International Business Information Management Association, Rome, Italy 13-14 November 2013.

[19] Alrahabi, M., Djioua, B., & Desclés, J. P. "Annotation sémantique des énonciations en arabe". *XXIV ème Congrès en INFormatique des Organisations et Systèmes d'Information et de décision*. 2006.

[20] Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M., et al. "OntoNotes Release 5.0". 2012.

[21] Sameh Alansary, Magdy Nagy, NohaAdly , IAN: An Automatic tool for Natural Language Analysis , 12th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt, December 12 - 13 2012

[22] Uchida, Hiroshi, M. Zhu, and T. Della Senta. "UNL: Universal Networking Language–An Electronic Language for Communication, Understanding, and Collaboration." Tokyo: UNU/IAS/UNL Center (1996).

[23] Fargues, Jean, Landau, Marie-Claude, Dugourd, Anne, et al. Conceptual graphs for semantics and knowledge processing. *IBM Journal of Research and Development*, 30(1), pp. 70-79, 1986.

[24] Sowa, John F. Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, 20(4), pp. 336-357, 1976.

[25] Sowa, John F. A conceptual schema for Knowledge-based systems. *ACM SIGART Bulletin*, 16(74), pp. 193-195, 1981.

[26] Sowa, John F. et Zachman, John A . Extending and formalizing the framework for information systems architecture. *IBM systems journal*, 31(3), po. 590-616, 1992.

[27] Kabbaj, A.: Development of intelligent systems and multi-agents systems with amine platform. In Conceptual Structures: Inspiration and Application, the Proceedings of the 14th International Conference on Conceptual Structures, Aalborg, Denmark, Volume 4068, pp. 286-299, 2006.

[28] Sarraf, Qusai et Ellis, Gerard. Business rules in retail: the Tesco. com story. *Business Rules Journal*, 7(6), 2006.

[29] Miller, GA.: WordNet: a lexical database for English. In Communications of the ACM, Volume 38, Issue 11, pp. 39-41, 1995.

[30] Kabbaj, A.: An overview of Amine. In: P. Hitzler and H. Schärfe (eds.) Conceptual Structures in practice, pp. 321-347, 2009.

[31] Abouenour, L., Nasri, M., Bouzoubaa, K., Kabbaj, A., and Rosso, P.: Construction of an ontology for intelligent Arabic QA systems leveraging the Conceptual Graphs representation. In the Journal of Intelligent and Fuzzy Systems (JIFS), DOI: 10.3233/IFS-141248, IOS Press, 2014.

[32] Weischedel, R., Hovy, E., Marcus, M., Palmer M., Belvin, R., Pradhan, S., Ramshaw, L., Xue, N. OntoNotes: A Large Training Corpus for Enhanced Processing. In Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, Eds. Joseph Olive, Caitlin Christianson, and John McCary, Springer, 2011.