Interface syntaxe-sémantique au moyen d'une grammaire d'arbres adjoints pour l'étiquetage sémantique de l'arabe

Cherifa Ben Khelil[1, 2] Chiraz Ben Othmane Zribi[1] Denys Duchier [2] Yannick Parmentier [3]

(1) RIADI, Campus universitaire de la Manouba, 2010, Tunisie
(2) LIFO, Bâtiment IIIA 6 rue Léonard de Vinci, F-45067 Orléans, France
(3) LORIA - Projet SYNALP, Campus Scientifique, BP 23954 506 Vandoeuvre-les-Nancy CEDEX, France
cherifa.bk@gmail.com, chiraz.zribi@ensi-uma.tn, denys.duchier@univ-orleans.fr, yannick.parmentier@loria.fr

RESUME

Dans une grammaire formelle, le lien entre l'information sémantique et sa structure syntaxique correspondante peut être établi en utilisant une interface syntaxe/sémantique qui permettra la construction du sens de la phrase. L'étiquetage de rôles sémantiques aide à réaliser cette tâche en associant automatiquement des rôles sémantiques à chaque argument du prédicat d'une phrase. Dans ce papier, nous présentons une nouvelle approche qui permet la construction d'une telle interface pour une grammaire d'arbres adjoints de l'arabe. Cette grammaire a été générée semi automatiquement à partir d'une méta-grammaire. Nous détaillons le processus d'interfaçage entre le niveau syntaxique et le niveau sémantique moyennant la sémantique des cadres et comment avons-nous procédé à l'étiquetage de rôles sémantiques en utilisant la ressource lexicale ArabicVerbNet.

ABSTRACT

Syntax-semantic interface using Tree-adjoining grammar for Arabic semantic labeling.
In formal grammar, the link between semantic information and its corresponding syntactic structure can be established using a syntax/semantic interface that allows the construction of sentence meaning. Semantic role labeling helps to achieve this task by automatically associating semantic roles with each argument of the predicate of a sentence. In this paper, we present a new approach that allows the construction of such interface for a Tree adjoining grammar for Arabic. This grammar was generated semi automatically from a meta-grammar. We detail the process of interfacing between syntactic and semantic levels through semantic frames and how we proceeded to the semantic roles labeling using the lexical resource ArabicVerbNet.

MOTS-CLES : Grammaire d'arbres adjoints ; méta-grammaire ; interface syntaxe/sémantique ; étiquetage de rôles sémantiques ; cadre sémantique ; langue arabe.

KEYWORDS: Tree adjoining grammar; meta-grammar; syntax/semantic interface; semantic role labeling; semantic frame; Arabic language.

1 Introduction

La construction automatique du sens d'une phrase représente un grand intérêt pour le domaine du Traitement Automatique du Langage Naturel (TALN). Mais pour ce faire, il est souvent utile de

faire correspondre aux composantes syntaxiques de la phrase des représentations sémantiques. L'étiquetage de rôles sémantiques permet de réaliser cette tâche en associant automatiquement des rôles sémantiques à chaque argument du prédicat (par exemple un verbe) d'une phrase. Ces rôles expriment des rôles abstraits que les arguments d'un prédicat peuvent admettre dans un événement ainsi que leur relation probable avec la fonction syntaxique dans cette phrase. L'étiquetage de rôles sémantiques est utile pour diverses applications du domaine TALN tels que les systèmes de traduction automatiques (Liu & Gildea, 2010), les systèmes questions-réponses (Pizzato & Mollá, 2008 ; Maqsud et al., 2014) ou encore les systèmes d'extraction de l'information (Christensen et al., 2010 ; Fader et al., 2011). Plusieurs de ces approches utilisent les ressources PropBank (Kingsbury & Palmer, 2003) et FrameNet (Baker et al., 1998) afin de définir le prédicat, les rôles utilisés lors de l'étiquetage ainsi que l'ensemble de test pour l'apprentissage automatique. En ce qui concerne l'arabe nous pouvons citer les travaux de (Diab et al, 2008) qui utilisent les machines vectorielles (Vapnik, 1998) et (Meguehout et al, 2017) qui se sont basés sur le raisonnement à partir de cas pour réaliser l'étiquetage sémantique.
Dans une grammaire formelle, ce lien entre la sémantique et la syntaxe peut être établi en utilisant une interface syntaxe/sémantique. Cette dernière permet de superviser la construction du sens de la phrase en unifiant les informations sémantiques de ses constituants. Les représentations sémantiques peuvent être sous la forme d'une formule logique des prédicats (Joshi & Vijay-Shanker, 1999 ; Kallmeyer & Romero, 2004 ; Romero & Kallmeyer, 2005), une formule logique sous-spécifiée (Gardent & Kallmeyer, 2003 ; Parmentier, 2007) ou plus récemment sous la forme d'un cadre sémantique (Kallmeyer & Osswald, 2013). A notre connaissance, de tels travaux n'ont pas été menés sur l'arabe.
C'est dans ce contexte que s'inscrit notre travail de recherche qui vise à élaborer une grammaire d'arbres adjoints (TAG) (Joshi et al., 1975) décrivant la syntaxe et la sémantique de l'arabe standard moderne (ASM) en vue d'une analyse syntaxico-sémantique. La grammaire que nous proposons a été produite semi automatiquement grâce au langage de description méta-grammatical XMG (eXtensible MetaGrammar) (Crabbé et al, 2013). À partir de la description méta-grammaticale ArabicXMG nous avons généré ArabTAG V2.0 (Ben Khelil et al, 2016). Ensuite, nous avons étendu cette grammaire en intégrant des informations sémantiques. Notre choix s'est porté sur la sémantique des cadres. Ce choix est motivé par la facilité de l'interfaçage entre le niveau syntaxique et le niveau sémantique.

Cet article est organisé de la manière suivante. Dans la section 2, nous détaillons le processus d'intégration de la dimension sémantique dans la méta-grammaire. Ensuite, nous présentons les étapes effectuées pour l'étiquetage de rôles sémantiques. Finalement, dans la section 4, nous exposons les premiers résultats de l'évaluation de cet étiquetage.

2 Intégration de la dimension sémantique dans la méta-grammaire

ArabTAG V2.0 (Ben Khelil et al, 2016) a été générée à partir d'une description méta-grammaticale en utilisant le compilateur XMG2 (Petitjean, 2014). Elle couvre les phrases verbales (forme active et passive), les phrases nominales, les différents types des syntagmes nominaux et les syntagmes prépositionnels. Elle traite aussi les différents phénomènes linguistiques arabes tels que la variation des positions des éléments au sein des composants syntaxiques, les compléments supplémentaires, les règles d'accord et les formes agglutinées. Afin d'étendre notre méta-grammaire et produire une grammaire TAG à portée sémantique, nous avons pensé à associer aux familles des arbres décrites des cadres sémantiques (FillMore, 1982). Nous

nous sommes basés sur la théorie du linking[1] (Levin, 1993 ; Kasper, 2008). Selon cette théorie, le verbe permet d'exprimer dans la plupart des cas la sémantique d'un évènement ainsi que la relation entre ses participants. En effet, l'ensemble des rôles sélectionnés par un prédicat verbal constitue un cadre sémantique. Certains de ces rôles sont obligatoires et déterminent la présence ou non de certaines fonctions grammaticales. Par exemple, lorsque l'acteur du verbe est présent dans une phrase, il est en position sujet (au cas nominatif). Ce genre de composant peut avoir le rôle d'«Agent». Ainsi, la fonction grammaticale permet d'indiquer le rôle à attribuer.

Notre idée consiste à préciser les rôles sémantiques au niveau du prédicat, qui est le verbe, au sein des structures syntaxiques décrites dans notre méta-grammaire. Le cadre de la phrase est ensuite construit au fur et à mesure de l'analyse syntaxique en unifiant les cadres sémantiques élémentaires de ses composants syntaxiques par l'intermédiaire d'une interface syntaxe/sémantique.

2.1 Construction de l'interface syntaxe/sémantique

L'interface syntaxe-sémantique au niveau de notre méta-grammaire est effectuée de la manière suivante (voir figure 1) :

- Au niveau syntaxique des familles de classes décrites par la méta-grammaire, nous avons défini les arguments du prédicat (verbe). Ces familles regroupent les arbres ancrés par un verbe et un nœud (de substitution) pour chaque argument du prédicat.
- Au niveau sémantique, nous avons défini les rôles sémantiques du cadre du prédicat. La dimension <frame> permet de décrire un cadre sémantique à l'aide de structures de traits typées.
- Le lien entre les rôles sémantiques et les constituants syntaxiques est établi à l'aide de l'interface syntaxe/sémantique en utilisant la dimension <iface>. Cette dernière correspond à la définition, pour chaque classe, d'une matrice de traits. Cette matrice permet d'associer un nom global (le trait) à une variable (la valeur du trait) ce qui permettra d'unifier les variables (suite à une opération de substitution ou d'adjonction) du même nom global et faire la correspondance entre les arguments du prédicat et leurs rôles correspondants.
- Les cadres sémantiques élémentaires sont définis aux niveaux du lexique (les lemmes).

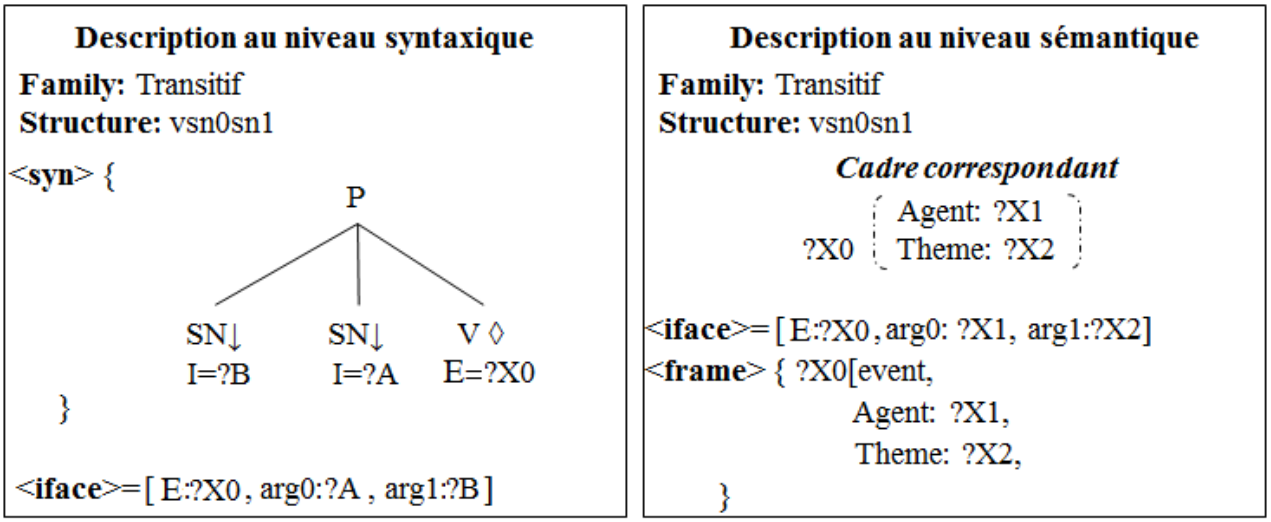

FIGURE 1 : Description de l'interface syntaxe/sémantique dans la méta-grammaire

[1] C'est la mise en relation d'une structure en rôle sémantique avec une structure syntaxique.

2.2 MAPPING entre la méta-grammaire et ArabicVerbNet

L'alimentation automatique de notre méta-grammaire par les rôles sémantiques se fait par l'intermédiaire de la ressource lexicale ArabicVerbNet (Mousser, 2010), version arabe de VerbNet (Kipper, 2008). Celle-ci couvre les verbes les plus utilisés de l'arabe standard moderne[2]. VerbNet est une ressource lexicale pour les verbes anglais qui repose sur le système de classification sémantico-syntaxique des verbes de (Levin, 1993). Les verbes ayant un comportement syntaxique et sémantique similaire sont affectés au même groupe de classes. Chaque classe d'un verbe est décrite au moyen des éléments suivants :

- Les membres : la liste des verbes appartenant à cette classe ou à une sous-classe. Cette liste contient aussi des informations sur la racine verbale, la forme déverbale et le participe de ces verbes.
- Les rôles : ce sont les rôles thématiques attribués à chaque membre du verbe de la classe. Ces rôles peuvent admettre un ensemble de restrictions sur leurs natures (animation, location, etc.).
- Les cadres : ils définissent la correspondance entre les rôles sémantiques et les arguments syntaxiques. Cette correspondance est expliquée à l'aide d'un exemple. En effet, pour chaque exemple de phrase, sa structure syntaxique et les relations sémantiques entre les arguments du prédicat sont définis.

Nous avons parcouru toutes les classes d'ArabicVerbNet. Dans un premier temps, les informations ont été regroupées selon la structure syntaxique de la phrase. Le groupement de ces structures respecte les familles d'arbres élémentaires définies par notre grammaire. Ensuite, pour chaque structure syntaxique, nous avons extrait l'ensemble des combinaisons des rôles sémantiques possibles pour construire les cadres sémantiques correspondants. Cet ensemble de cadre sémantique est défini au niveau de la méta-grammaire (avec la dimension <frame>). Au final, le compilateur XMG2 (Petitjean, 2014) compile cette méta-grammaire et génère une grammaire constituée d'un ensemble d'arbres élémentaires associés à leurs cadres prédicats.

3 Étiquetage à base de rôles sémantiques

L'affectation des cadres sémantiques se fait lors de l'analyse syntaxique par l'intermédiaire de l'interface syntaxe/sémantique. Au fur et à mesure qu'une phrase est analysée, son cadre sémantique est construit en unifiant les cadres sémantiques élémentaires de ses constituants et celui du verbe prédicat.

Prenons l'exemple de la phrase suivante (voir figure 2) : طاردَ الشرطيُ اللصَ (*le policier poursuit le voleur*). Le processus de construction du sens de cette phrase est réalisé comme suit : L'arbre syntaxique de la phrase analysée est constitué d'un verbe suivi d'un sujet et d'un objet. Le verbe ancré «طاردَ» (*poursuivre*) admet donc deux arguments. Après avoir effectué l'étiquetage de rôles sémantiques (en consultant ArabicVerbNet), les deux rôles attribués à ces arguments sont : Agent et Theme. Les cadres élémentaires (personnage et profession) sont associés aux arbres élémentaires des syntagmes nominaux. L'élément I qui représente l'interface syntaxe/sémantique permet le

[2] La version actuelle d'ArabicVerbNet comporte 334 classes qui contiennent 7672 verbes et 1393 cadres.

partage des variables de traits des nœuds avec les variables issues des cadres sémantiques. Les opérations de substitution déclenchent les équations d'unification entre ces variables : [X1 = A] et [X2 = B]. L'unification est ainsi opérée et mène à l'insertion des cadres élémentaires de «الشرطيُ» (*le policier*) et «اللصَ» (*le voleur*) dans le cadre sémantique prédicat du verbe «طاردَ» (*poursuivre*). Le cadre final obtenu représente le sens de la phrase.

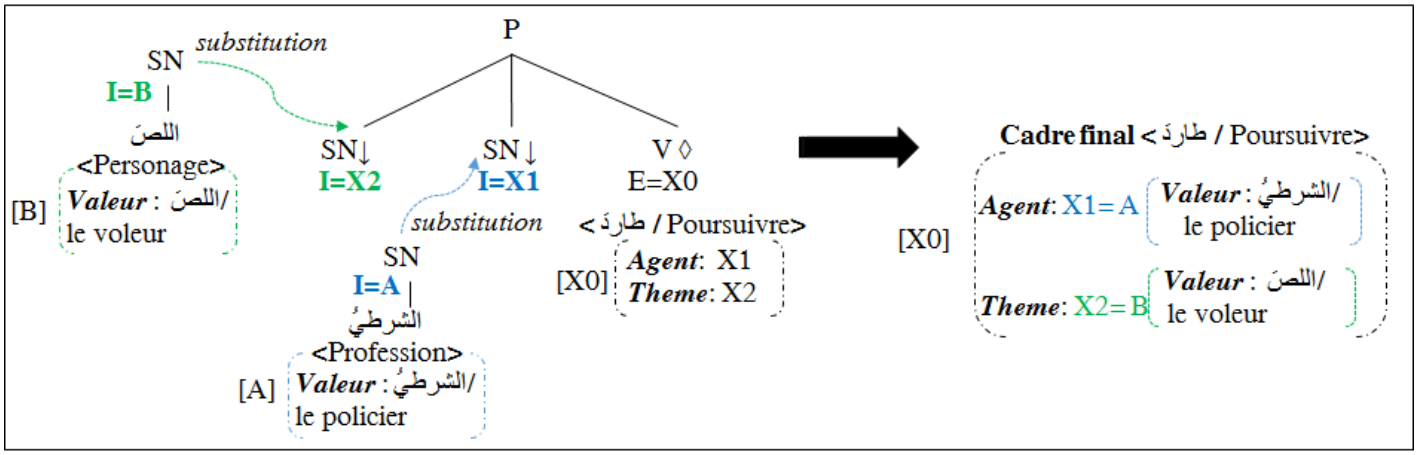

FIGURE 2: Composition du cadre sémantique pour طاردَ الشرطيُ اللصَ / *le policier poursuit le voleur*

Une structure syntaxique peut avoir plusieurs cadres sémantiques correspondants. Ces cadres représentent différents sens susceptibles d'occasionner plusieurs interprétations possibles. Par exemple : Un sujet peut être «Agent» ou «Acteur» selon les contextes. Nous avons exploité d'avantage les classes d'ArabicVerbNet et nous avons établi un ensemble de contraintes afin d'optimiser la tache de l'étiquetage de rôles sémantiques au moment de l'analyse sémantique :

- ***La classe du verbe :*** les cadres sémantiques pour un verbe sont définis en fonction sa classe.
- ***Le type de la préposition pour les syntagmes prépositionnels :*** certains rôles sémantiques ont tendance à apparaître comme des syntagmes prépositionnels. Dans ce cas, la préposition peut indiquer le sens de ce syntagme et ainsi intervenir pour restreindre le choix du cadre correspondant. Considérons l'exemple des deux phrases suivantes :
  (1) نبحَ الكلبُ على الهرّ/ *Le chien aboie sur le chat*
  (2) نبحَ الكلبُ منَ الخوفِ/ *Le chien aboie de peur*
  Selon la classe verbale de «نبح» (*aboyer*), animal_sounds-1 définie dans ArabicVerbNet, nous avons trois combinaisons possibles de cadres sémantiques pour la structure de ces deux phrases :
  a) Agent+ {prep (على)} +Recipient : la préposition «على» (*sur*) indique que le rôle sémantique de l'objet est Recipient.
  b) Agent+ {prep (من)} +Cause : la préposition «من» (*de*) exige que le rôle sémantique de l'objet soit Cause.
  c) Location+ {prep (بـ)} +Agent : la préposition «بـ» (*avec*) indique que le rôle sémantique de l'objet est Agent.
  Après avoir filtrer ces résultats en tenant compte de la contrainte sur la préposition, nous obtenons les correspondances sémantiques suivantes :
  1) -a) Agent {الكلبُ/ *le chien*} + {prep (على/ *sur*)} +Recipient {الهرّ/ *le chat*}.
  2) -b) Agent {الكلبُ/ *le chien*} + {prep (من/ *de*)} +Cause {الخوفِ/ *peur*}.
- ***Les contraintes :*** un verbe peut imposer un ensemble de restrictions à ses rôles d'argument. Par exemple, en exigeant qu'un rôle soit humain et /ou animé, etc. Soient les deux phrases suivantes avec leurs interprétations sémantiques :
  (3) يحبُ عليٌ فاطمةً / *Ali aime Fatima*: Experiencer {عليٌ/ *Ali*} +Theme {فاطمةً/ *Fatima*}.

(4) يحبُ الكتابُ فاطمةً / *Le livre aime Fatima*: Experiencer {الكتابُ/ *le livre*} + Theme {فاطمةً/ *Fatima*}.

Le prédicat est le verbe «أحب» (*aimer*). Bien que les deux phrases soient syntaxiquement correctes la deuxième est sémantiquement incorrecte. Le sujet «الكتابُ» (*livre*) ne peut pas éprouver des sentiments envers un humain. Lors de l'analyse sémantique, nous faisons intervenir les contraintes spécifiées pour les rôles sémantiques au niveau de la classe du verbe «أحب» (*aimer*). Après avoir examiné cette classe, nous avons remarqué que l' «Experiencer» doit être animé et humain. Par conséquent, nous pouvons confirmer que la première phrase est sémantiquement correcte alors que la deuxième ne l'est pas, puisque son sujet ne satisfait pas cette contrainte.

4 Expérimentations

Afin d'évaluer notre grammaire dans sa tâche d'analyse syntaxico-sémantique, nous avons défini un corpus de test de 500 phrases (347 phrases verbales et 153 phrases nominales) extraites à partir d'un livre scolaire tunisien (niveau 8ième année[3]). Ce choix est dû à l'indisponibilité des corpus annotés d'information syntaxico-sémantique pour l'arabe standard moderne (Ben Khelil, 2017). Nous avons développé un outil afin d'effectuer cette analyse. Cet outil permet dans un premier lieu de faire un étiquetage morphosyntaxique des éléments de la phrase, suivit par l'analyse syntaxico-sémantique suivant les étapes expliquées dans la section 3.

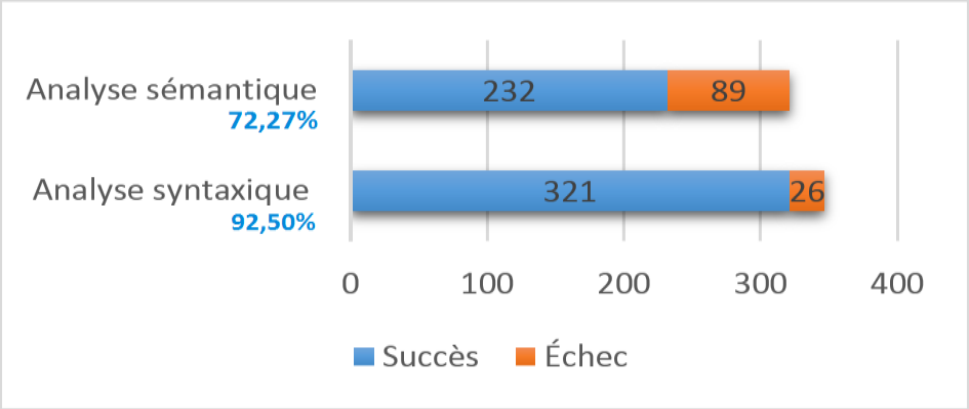

FIGURE 3 : Résultat de l'analyse syntaxico-sémantique des phrases verbales

Parmi les 347 phrases verbales testées, nous avons réussi à analyser syntaxiquement 321 phrases (92,50%) et sémantiquement 232 phrases (72,27%). Nous nous focalisons, dans cet article, à décrire l'évaluation de la partie sémantique. Les causes de l'échec de l'analyse sémantique sont principalement dues à un manque de couverture au niveau d'ArabicVerbNet. Nous avons constaté que 19,93% des verbes des phrases testées ne sont pas définis dans cette ressource. De plus, pour un verbe donné dans ArabicVerbNet, la liste des structures syntaxiques des phrases qu'il peut admettre n'est pas exhaustive. En effet, nous avons mesuré un taux de 5,29% d'échec d'analyse du à l'absence de la structure de la phrase correspondante au moment de l'étiquetage de rôles sémantiques. Nous avons aussi obtenu 2,49% d'échec pour les phrases complexes. L'analyse de ce genre de phrases est plus compliquée vu qu'elles contiennent plusieurs verbes.

[3] Equivalent à la 4ième année au collège en France.

5 Conclusion

Nous avons présenté une nouvelle approche visant à construire une grammaire d'arbres adjoints pour représenter la syntaxe et la sémantique de l'arabe. Nous nous sommes concentrés dans cet article sur le processus d'intégration de l'information sémantique. Notre idée est d'associer aux familles d'arbres élémentaires de la grammaire une sémantique à base de cadres et d'intégrer les rôles sémantiques à partir de la ressource ArabicVerbNet. Ceci a permis d'établir une correspondance entre arguments sémantiques et arguments syntaxiques par l'intermédiaire d'une interface syntaxe/sémantique permettant aux cadres sémantiques élémentaires de s'unifier lors de la composition syntaxique.

Lors de l'étiquetage de rôles sémantiques nous avons constaté que plusieurs informations peuvent aider à lever l'ambigüité sémantique. Nous citons ; la classe du verbe, les propriétés du rôle et aussi l'utilisation de certaines prépositions pour les syntagmes prépositionnels. Bien que les premiers résultats de l'analyse soient encourageants, nous envisageons dans le futur proche d'augmenter notre corpus de test et d'avoir recours à l'apprentissage automatique pour améliorer le taux de réussite et pallier le manque de données au niveau d'ArabicVerbNet.

Références

JOSHI A., LEVY L., TAKAHASHI M. (1975). Tree adjunct grammars. Journal of Computer and System Sciences. 10(1), 136 – 163.

CRABBÉ C., DUCHIER D., GARDENT C., LE ROUX J., PARMENTIER Y. (2013). XMG : eXtensible MetaGrammar. Computational Linguistics. 39(3), 591–629.

JOSHI A., VIJAY SHANKER K. (1999). Compositional Semantics with Lexicalized Tree Adjoining Grammar (LTAG) : How Much Underspecification is Necessary . In Proceedings of the Third International Workshop on Computational Semantics, IWCS-03, Tilburg, The Netherlands.

KALLMEYER L., ROMERO M. (2004). LTAG Semantics with Semantic Unification. In Proceedings of TAG+7, Vancouver, pages 155–162.

ROMERO M., KALLMEYER L. (2005). Scope and Situation Binding in LTAG using Semantic Unification. In Proceedings of the Sixth International Workshop on Computational Semantics IWCS-6, Tilburg.

KALLMEYER L., JOSHI A. (2003). Factoring Predicate Argument and Scope Semantics: Underspecified Semantics with LTAG. Research on Language and Computation, volume 1 :1-2, pages 3–58.

GARDENT C., KALLMEYER L. (2003). Semantic construction in FTAG. In Proceedings of the European chapter of the Association for Computational Linguistics (EACL'03), Budapest.

PARMENTIER Y. (2007). SemTAG : une plate-forme pour le calcul sémantique à partir de Grammaires d'Arbres Adjoints. Ph.D thesis, université Henri Poincaré – Nancy 1.

KALLMEYER L., OSSWALD R. (2013). Syntax-Driven Semantic Frame Composition in Lexicalized Tree Adjoining Grammars. Journal of Language Modelling Vol i2, pp. 1–63.

KINGSBURY P., PALMER M. (2003). Propbank: the next level of treebank. In Proceedings of Treebanks and Lexical Theories.

BAKER F., FILLMORE J., LOWE B. (1998). The berkeley FrameNet project. In COLINGACL '98: University of Montréal.

LIU D., GILDEA D. (2010). Semantic role features for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 716–724, Beijing, China.

PIZZATO L.A., MOLLÁ D. (2008). Indexing on semantic roles for question answering. In Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering, pages 74–81, Manchester, UK, August.

MAQSUD U., ARNOLD S., HÜLFENHAUS M., AKBIK A. (2014). Nerdle: Topic-specific question answering using wikia seeds. In COLING (Demos), pages 81–85.

CHRISTENSEN J., MAUSAM ., SODERLAND S., ETZIONI O. (2010). Semantic role labeling for open information extraction. In Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, pages 52–60, Los Angeles, California, June. Association for Computational Linguistics.

FADER A., SODERLAND S., ETZIONI O. (2011). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1535–1545. Association for Computational Linguistics.

DIAD M., MOSCHITTI A., PIGHIN D. (2008). Semantic role labeling systems for Arabic language using kernel methods. In: 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2008: HLT), Columbus, Ohio, USAS.

MEGUEHOUT H., BOUHADADA T., LASKRI M.T. (2017). Semantic role labeling for Arabic language using case-based reasoning approach. Int J Speech Technol (2017) 20: 363. https://doi.org/10.1007/s10772-017-9412-6

VAPNIK N. (1998). Statistical Learning Theory. JohnWiley and Sons.

LEVIN B. (1993). English Verb Classes and Alternations A Preliminary Investigation. Chicago: University of Chicago Press.

KASPER S. (2008). A comparison of thematic role theories," Master Homework. Marburg University.

KIPPER K., KORHONEN A., RYANT N., PALMER M. (2008). A large-scale classification of English verbs Lang. Resour. Eval. J., 42 (2008), pp. 21-40.

MOUSSER J. (2010). A large coverage verb taxonomy for Arabic. In: Seventh Conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta, pp. 2675–2681.

BEN KHELIL C. (2017). Générer une grammaire d'arbres adjoints pour l'arabe à partir d'une méta-grammaire. Présenté sous forme de poster à la 24° édition de la conférence TALN aux 19es Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL 2017) Orléans, France.

BEN KHELIL C., DUCHIER D., PARMENTIER Y., ZRIBI C., BEN FRAJ F. (2016). ArabTAG : from a Handcrafted to a Semi-automatically Generated TAG. In TAG+12: 12th International Workshop on Tree-Adjoining Grammars and Related Formalisms, Düsseldorf, Germany.

PETITJEAN S. (2014). Génération Modulaire de Grammaires Formelles. Ph.D. thesis, Université d'Orléans, France.