ACIT'2017
The International Arab Conference on Information Technology
Yassmine Hammamet, Tunisia
December 22-24, 2017

A new syntactic-semantic interface for ArabTAG an Arabic Tree Adjoining grammar

[1]Cherifa BEN KHELIL, [1]Chiraz Ben Othmane Zribi, [2]Denys Duchier and [3]Yannick Parmentier
[1]RIADI - ENSI, Université La Manouba, Tunisia
[2]Université d'Orléans, France
[3]Université de Lorraine, France

***Abstract*— For a reliable natural language processing (NLP), it is important to know how to link the meaning of a statement or a sentence to its syntactic structure. The link between semantic and syntax can be established using a syntax-semantic interface that allows the construction of sentence meaning. In this paper, we present a new approach, which built a tree adjoining grammar to represent the syntax and the semantic of modern standard Arabic. In the first part, we detail the process that automatically generates this grammar using Arab-XMG meta-grammar. Then we explain how we have established the link between syntax and semantic and how we have introduced the semantic frame -based dimension into the meta-grammar using Arabic Verbnet.**

Keywords*— Tree adjoining grammar; metagrammar; syntax; semantic; syntax-semantic interface ; semantic frame; Arabic language .

I. Introduction

Natural language allows to build a potentially infinite number of meaningful utterances from a finite number of words. Being able to automatically construct the meaning of a sentence represents a great challenge and interest for many applications in the field of natural language processing applications (NLP) (such as machine translation systems, human-machine dialogue, question-answering systems). However, before constructing a representation of the meaning of a sentence or a statement, it is usually essential to produce a representation of its syntactic structure.

Indeed, for a reliable natural language processing, it is crucial to involve morphological, syntactic and semantic data. The semantic information is used to help the detection of relations between various syntactic components. Thus, it is important to have a means to link the sentence meaning to its syntactic structure. This relation can be established using a syntax-semantic interface. This interface allows the construction of semantic representation of the sentence based on the relationship between its syntactic constituents. This link is expressed through rules describing the constituents in the sentence (e.g. [1], [2], [3] and [4]). The semantic representation of the sentence is constructed gradually in parallel to its syntactic structure. According to the used formalism, this approach is implemented in different ways. This implies that the syntax-semantic interface itself is closely related to the chosen grammar or grammatical formalism.

In this context, having digital resources such as grammars should be useful and even indispensable. However, to date there is not a wide-coverage formal grammar for the Arabic language that integrates semantic dimension. In this work, we are interested in producing such grammar describing the syntax and the semantic of modern standard Arabic (MSA). We opted for the tree Adjoining Grammar (TAG) [5] formalism enriched by semantic frames. Our choice was motivated by the power of representation of TAG (simple, complex, combinatorial, shared structures, etc.) and its ability to deal with certain phenomena that are very recurrent in Arabic such as embedding. Our grammar is produced semi-automatically by using a metagrammatical language called XMG eXtensible MetaGrammar) [6].

The paper is organized as follows. In section 2, we present some related works. Then, in section 3, we introduce briefly the TAG formalism. Section 4 gives a concise description of the syntax of Arabic using Arab-XMG meta-grammar. Section 5 describes the semantic integration process into the meta-grammar. Finally, in Section 6, we discuss cases of semantic ambiguities.

II. Related work

During the last decade, several syntactic analysis approaches for standard Arabic have been proposed. We can mention for example the following works [7], [8], [9] and [10]. However, approaches dealing with semantic are rare or even absent. For example, [11] proposed a semantic construction model of Arabic sentences. This approach is based on the use of λ-calculus and considers the structured syntactical categories of the sentence as a guideline for constructing semantic representations in form of logical formulas.

The representation of sentence should be achieved through a representation of its syntactic structure. To our knowledge, few works have constructed a formal grammar of Arabic. For example [12], which propose a HPSG (Head-driven phrase structure grammar) [13], essentially dealing with the nominal sentences and the works of [14] that implement a HPSG grammar fragment of Arabic on a platform known as LKB (Linguistic Knowledge Builder). Concerning tree-adjoining grammars, [15] constructed a TAG by extracting elementary trees from an Arabic Treebank (namely the Penn Arabic TreeBank – PATB) [16]. However, to date there is no broad coverage grammar of Arabic. Moreover, to our knowledge, no work has focused on integration of semantic dimension in the grammar.

Several approaches for other languages, such as English and French, have been proposed in order to ensure the syntax-semantic interface. A grammar formalism, which incorporates semantic information, has been proposed with synchronous tree-adjoining grammars [17]. The idea was to allow coupling between syntactic and semantic trees representations. For a pair of elementary trees, links are defined between a node of the syntactic tree and a node of the semantic tree. The latter can be linked to different nodes in syntactic tree. Furthermore, a single syntactic node may link to more than one semantic node. During derivation, these links are consumed, and two trees are constructed synchronously: a derived syntactic tree and a derived semantic tree. Synchronous tree-adjoining grammars have been successfully used in grammars for English [18] and French [19], which allowed to simultaneously generate syntactic and semantic analysis of sentences.

Another approach consists in adding semantic representations using the variables of unification[1] in the grammar [1]. The idea is to define a syntax-semantic interface allowing the feature structures contained in the terms to be properly unified during the semantic composition. The placement of the semantic variables in the feature structures is done according to a set of rules proper to the adopted approach. The introduced semantic representations correspond to a set of formulas. The latter can be a formula in predicate logic [20], an underspecified logic[2] [21], [22], a glue part [23] for English, or a flat semantic [24], [25] for English and [26], [27] for French. Thus, the final semantic representation of a sentence corresponds to the union of these associated semantic formulas.

More recently, the works of [28] propose to introduce another form of semantic representation, which is based on frame semantic. In this approach, each elementary syntactic construction is associated to a semantic frame. Subsequently, the composition of syntactic building blocks led to the parallel composition of their associated frames. This process is seen as unification.

III. GENERATING A TREE ADJOINING GRAMMAR FOR ARABIC

Before defining the generated grammar, we recall in what follows the TAG formalism.

A. Brief presentation of TAG formalism

Tree Adjoining Grammar (TAG) [5] is a syntactic formalism that takes into account the links between the constituents of the sentence to build grammatical representations. It offers a tree rewriting system whose units are elementary trees. There are two types of elementary trees:

- An initial tree: it is a tree whose leaf nodes are either terminal symbols or non-terminal symbols. The non-terminal symbols are called substitution nodes and are marked with the symbol (↓);
- An auxiliary tree: it has also a leaf node labeled with a non-terminal symbol called "foot node" and is marked with the symbol (*). The foot node and the root of the auxiliary tree must be of the same category.

The two composition operations authorized by TAG are: substitution and adjunction. The resulting tree obtained by the end of these operations is called a derived tree. The substitution operation (Figure. 1) appends a tree α at a frontier node (substitution node) of another tree β. This operation is allowed only if the substitution node and the root node, respectively of β and α, are labeled with an identical symbol.

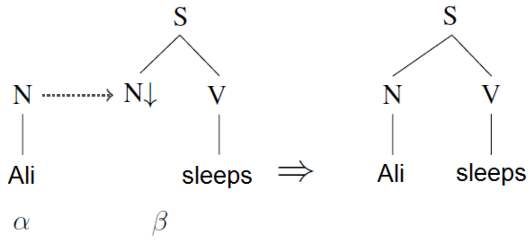

Fig. 1. Substitution in TAG.

The adjunction operation (Figure 2) is more powerful since it allows to insert one tree γ into another tree β on an internal node X. The root of γ replaces the node X located in β. The adjunction is acceptable if the category of the node X is identical to the category of the root node of γ.

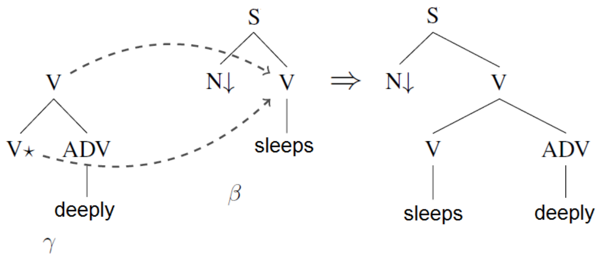

Fig. 2. Adjunction in TAG.

TAG is considered the standard model for mild context-sensitivity [29]. It is slightly more powerful than context-free grammars3, but strictly included in the class of contextual grammars4. It defines an extended domain of locality because the depth of the elementary trees is variable, unlike rewriting rules in context-free grammars whose depth is equal to 1. This means that it has a strong generative power. Also, constructions related to iteration and recursion are modeled by the operation of adjunction. Moreover, from a processing point of view, TAG remains analyzable in polynomial time O (n^6).

We cannot, assert that this formalism is undoubtedly the best to represent Arabic. Nevertheless, its characteristics make it possible to represent frequent phenomena in Arabic such as the embedding. Indeed, the representation of this phenomenon is possible with TAG thanks to adjunction operation. Through adjunction, we can integrate a complete structure in another structure thus making the representation of the embedding clause very natural.

[1] They are a set of attribute-value pairs that provides morphological, syntactic or semantic information. The unification allows to combine the information of two feature structures.

[2] Using labels holes and range constraints.

[3] It is a set of recursive rewriting rules used to generate patterns of strings. The left parts of the rules contain a unique non-terminal, so its derivation does not depend on any context (X→w). This type of grammar makes it possible to define algebraic languages, which are recognizable by a non-deterministic stack automaton.

[4] Rules in contextual grammar require that the right member must be at least as long as the left member (uXv→uwv). The language generated by this type of grammar is recursive and recognizable by a linearly bounded automaton.

B. TAG for Arabic : ArabTAG V 2.0

Our work takes its origins from an existing handcrafted tree-adjoining grammar for Arabic named ArabTAG (Arabic Tree Adjoining Grammar) [30]. This grammar inherits all the basic foundations of TAG. It describes different syntactic components of different levels: sentences, phrases and words, as well as the various information related to them (morphological and syntactic information). ArabTAG has feature structure and is semi-lexicalized. It contains two sets of elementary trees: 35 lexicalized trees and 215 patterns trees. The lexicalized trees have at least one lexical item as a leaf node and are reserved to prepositions, modifiers, conjunctions, demonstrative pronouns On the other hand, the patterns trees represent verbs, nouns, adjectives or any kind of phrases. They include a specific foot node marked with the symbol ◊, which is the anchor. This node will be instantiated with a lexical unit during parsing. [31].

We studied the first version of this grammar and we noted some limitations that can be summarized as follows:

- Minimal coverage of syntactic structures. Structures enriched with supplements (circumstantial complements of time, place, etc.) are not described ;
- The representation of forms of agglutination is not well reflected. These forms should be extended to improve the coverage of the grammar ;
- The lack of semantic information;
- ArabTAG consists of a flat set of elementary trees (that is, without any structure sharing). In particular, it is not organized in a hierarchical way, which does not facilitate grammar extension and maintenance.

Therefore we have proposed a new version ArabTAG V2.0 [32] that takes into account the aspects mentioned above. This new version is rewritten using the XMG (eXtensible MetaGrammar) description language [6]. The latter offers particularly pertinent features to describe elementary trees for Arabic language:

- it is highly expressive, since it defines highly factorized grammar descriptions;
- it is particularly adapted to the description of tree grammars (it has been used to develop several electronic TAG grammars for e.g. French[5], English[6], German[7]) ;
- it is highly extensible and can be configured to describe various levels of language, such as semantic or morphology.

With this formalism, we have semi-automatically generated ArabTAG V 2.0 from a reduced description of grammar rules. This compact description of grammatical information corresponds to a meta-grammar (ArabXMG: defined manually using XMG description language). It captures the linguistic generalizations appearing among the trees of the grammar and then makes it possible to generate the corresponding TAG trees. This meta-grammar is automatically compiled into an actual electronic grammar (ArabTAG V2.0) by the XMG compiler.

Tree fragments are described using a tree description logic based on dominance and precedence relations between node variables. These fragments are encapsulated within classes. A class corresponds to the combination of meta-grammatical description with a name making it possible to reuse a given description in various contexts.

XMG allows inheritance between classes. Thus, the new version ArabTAG V2 has been organized into a hierarchical organization. We described the syntax of verbal predicates in Arabic in a concise and modular way. We used the transitivity of the verb as a fundamental criterion for inheritance and we have combined 8 tree fragments together in order to obtain the 3 basic verb families (intransitive, transitive and ditransitive verbs). Each of these families captures the possible syntactic realizations between the different structures of the sentence.

In addition, using XMG allowed us to deal with the semi-free word order of Arabic language. To do this, we avoided imposing precedence constraints between nodes whose order change does not affect the consistency of the sentence.

C. Coverage and validation of ArabTAG V 2.0

So far, we have generated 313 trees from a description made of 28 classes (that is, 28 tree fragments or combination rules) as shown in Figure 3.

The current version of the grammar covers the structures already covered by the first version of ArabTAG, namely verbal phrases (active and passive form), nominal sentences, nominal phrases and prepositional phrases. In addition, it covers elliptical, anaphoric and subordinate structures. It takes into consideration the change of the order of the sentence's components and the agglutinative forms. Coverage has also been extended by adding elementary trees for the representation of additional complements such as circumstantial complement of time, circumstantial place complements and adverbs.

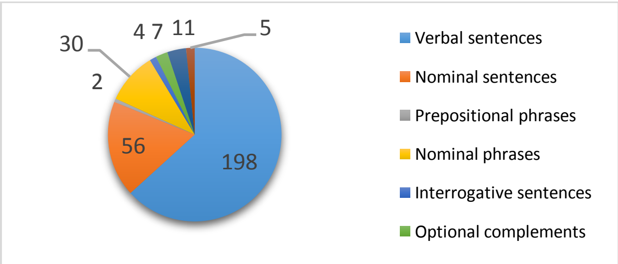

Fig. 3. Current ArabTAG V2.0 coverage .

In order to verify grammar coverage, we set up a development environment while designing ArabTAG with XMG. Moreover, we defined proof of concept syntactic and

[5] https ://sourcesup.renater.fr/xmg/frenchmetagrammar/index.html

[6] http ://homepages.inf.ed.ac.uk/s0896251/XMG-basedXTAG/titlepage.html

[7] http ://www.sfs.uni-tuebingen.de/emmy/res.html

[8] XMG allows combining tree fragments conjunctively or disjunctively

morphological lexicons for Arabic following the 3-layer lexicon architecture (tree templates, lemmas, words) of the XTAG project [33].

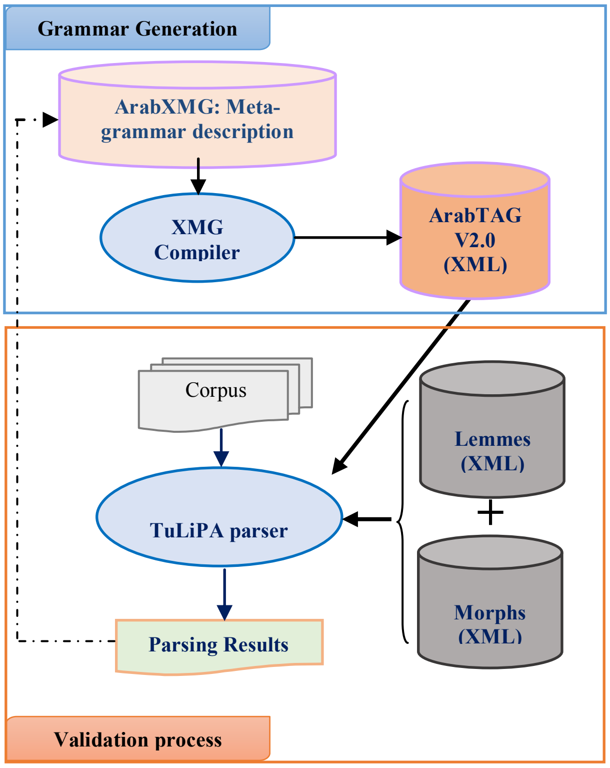

Fig. 4. Validation architecture of ArabTAG V2.0.

The XTAG system consists of three sub-modules:

- A basis of tree schemas classified into families of elementary trees;
- A lemma basis where each lemma is associated with one (or more) family trees;
- A morphological basis in which each flexed form is associated with a lemma and its appropriate morphosyntactic information.

The purpose of this validation (Figure 4) is to evaluate and to reduce both under and over-generation. Each new syntactic phenomena included in ArabTAG V2.0 leads to the extension of a test corpus gathering both grammatical and ungrammatical sentences (associated with the number of expected parses). The TuLiPA parser [34] is then run on the test corpus to check the quality of the grammar. The parsing results help us to fix potential errors and bugs in our metagrammatical description and allow us to check the consistency of the TAG structures when it is extended.

IV. SYNTACTICO-SEMANTIC ANALYSIS FOR ARABIC

In order to integrate semantic information during syntactic analysis, we have extended our meta-grammar and produced a semantic-based TAG grammar. In the literature, we find several approaches concerning adding semantic information in the TAG formalism. These semantic representations can take the form of a formula in predicate logic [20], [35], and [36], or an underspecified semantic representation [22], [24], [26] and [27] or more recently a semantic frame [28]. We decided to use semantic frame as we noted that frame semantic make the interfacing easier between syntax and semantic.

A semantic frame [37] can be defined as a data structure representing a particular situation by associating to its name a set of elements (participants) describing its situational roles (attributes) or relationship roles (semantic roles). These elements can be mandatory or optional. A frame can be evoked by words. When the evocative word is a verb, the analysis is focused on the latter's arguments which represent the frame elements.

For example, the "pursuit" frame may have several evocative words including the verb "طارد" (chase). It has mandatory (AGENT, THEME) and optional (LOCATION, MODE, DURATION, etc.) roles. For the following sentence طارد الشرطي اللص (*The policeman chases the thief*), we want to obtain the final semantic frame shown in Figure 5 at the end of the analysis.

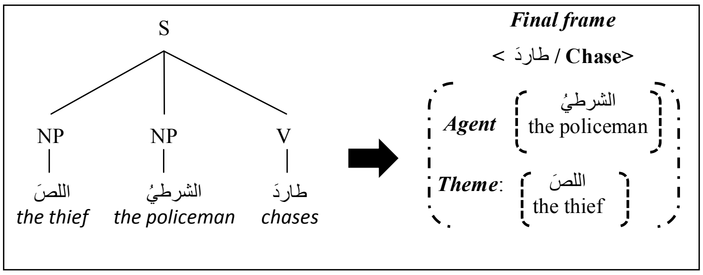

Fig. 5. Example of describing frame "chase" for طارد الشرطيُ اللص / *The policeman chases the thief* [9].

Semantic frames are implemented in the Berkeley University FrameNet project [38] to provide a frame database for English language. This base has been used in several works such as [39] and [40] for the task of semantic role labeling[10]. FrameNet exists for several languages such as French[11], Chinese[12], Spanish[13] and Japanese[14].
Since we do not have such a basis for Arabic, we have thought as [28] to associate each elementary tree with a frame and during the syntactic analysis we build the final frame representing the sentence meaning by unifying these elementary frames. In our case, the governing predicate is the verb. We proceeded as follows:

1. To each elementary tree, we associated an elementary semantic frame;
2. Within the semantic frame of the verb (also called semantic frame of the predicate) we specify the number of arguments and their valences;

[9] unlike English the tree is read from right to left
[10] Semantic role labeling is the task of automatically finding the semantic roles of each argument of each predicate in a sentence.
[11] https://sites.google.com/site/anrasfalda/
[12] http://sccfn.sxu.edu.cn/portal-en/home.aspx
[13] http://spanishfn.org/
[14] http://jfn.st.hc.keio.ac.jp/

3. Within remaining frames, we define the semantic information specific to the lexicon;
4. Each frame is labeled with a value that indicates the attribute of the interface (base-labeled feature structures);
5. The syntactic trees are decorated with interface features, which relies elementary trees and semantic frames (the syntax-semantic interface) and make them accessible for semantic composition;
6. Substitutions and adjunctions trigger the unification of semantic frames according to the label equations.

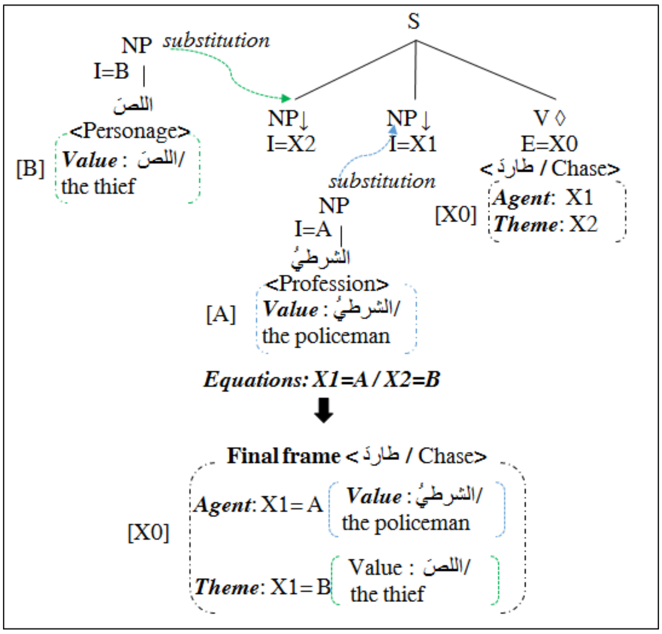

Fig. 6. Frame composition for طارد الشرطيُ اللص / The policeman chases the thief.

Let us consider the previous example of the "pursuit" frame with the verb "طارد" (*chase*). The latter has two arguments matching the syntactic structure v np0 np1 (verb + noun phrase + noun phrase). Two semantic roles are attributed to the arguments: AGENT (the volitional causer of an event) and THEME (the participant most directly affected by an event). The elementary frames (personage and profession) are associated with the elementary trees for the noun phrases. The interface feature I (see Figure 6) assign semantic roles (from semantic frame of the predicate) to syntactic arguments with a base-labeled feature. Substitution operations trigger the equations: [X1 = A] and [X2 = B]. The unification is carried out and leads to the insertion of elementary frames "الشرطي" (*the policeman*) and "اللص" (*the thief*) in the semantic frame of the verb "طارد" (*chase*).

The idea of specifying semantic roles at predicate verb within syntactic structure is based on the linking [15] theory [41] and [42]. According to this theory, the syntactic behavior of a verb can be predicted from its semantic. Indeed, the verb, namely the governing predicate, expresses the semantic of an event and the relationship between its participants.
For example, if a verb-er (such as "a hunter", "a reader", "an eater" etc.) is present in the sentence, it will be in the subject / nominative position. In the Linking theory the verb-er [42] has been granted an autonomous status: they can be called "agents". Thus, if a predicate (the verb) has in its frame an AGENT, the later will assume the grammatical function of subject of an active sentence, THEME the direct object, and so on.

A. *Integrating semantic dimension into Arab-XMG metagrammar*

Meta-grammatical factorization offers a fine-grained decomposition of syntactic building blocks, which can be grouped into families. It allows us to separate semantic constructions from the lexicon and to create generalizations across constructions.

We integrated the semantic representations as follows (Figure 7):

1. At the syntactic level of tree families (class), we define the arguments of the predicate (verb). In fact, a family corresponds to a group of unanchored elementary trees for the same category of the predicate verb (intransitive, transitive, ditransitive). In each of these trees, there is a node for every argument of the predicate.
2. At the semantic level, we define the semantic roles of the predicate. The frames we used for semantic are typed feature structures specified within the <frame> dimension of a class.
3. Linking between syntactic and semantic constituents is ensured by the syntax-semantic interface (<iface> dimension). Interfaces correspond to attribute-value matrix defined for each class, allowing one to associate a global name to an identifier or a variable. Thus, it makes it possible to unify variables of the same global name.
4. The elementary semantic frames are defined and stored in a lemma lexicon.

In that resptect, linking is specified in the meta-grammar. Valency information is provided by the unanchored tree and the lexical anchor specifies only its semantics.

The semantic information of the "pursuit" frame is described in Figure 7. The corresponding structure v np0 vnp1 belongs to the transitive verbs family. It admits two arguments (arg0 and arg1). On the semantic side, the two roles associated with the arguments are AGENT and THEME.

[15] Means linking a semantic role with a syntactic structure.

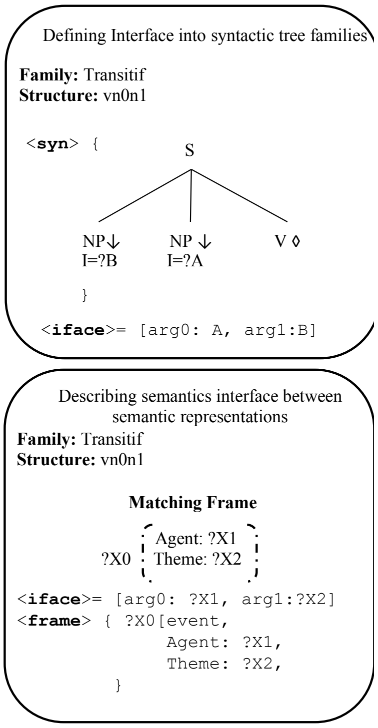

Fig. 7. Describing the syntax/semantic interface between semantic representations and syntactic trees with XMG.

The outcome of the metagrammar compilation is pairs of unanchored elementary trees and predicate frames.

B. Mapping between Arab-XMG and Arabic VerbNet

To avoid manually semantic roles encoding in our metagrammar, we considered using VerbNet.

VerbNet [43] is a lexical resource for English verbs. It is based on the semantico-syntactic classification system of verbs of [41]. Verbs with similar syntactic and semantic behavior are assigned to the same class group. A class group represents a hierarchy established by the semantic relations between its classes. Each class of a verb is described using the following elements:

- Members : a list of verbs belonging to this class or its subclass;
- Roles: these are thematic roles assigned to each member of the verb class. These roles can have a set of restrictions (constraints) on their natures (animation, rental, etc.);
- Frames: define the correspondence between semantic roles and syntactic arguments. For each example of sentence, its syntactic structure and its semantic structure containing semantic predicates and their arguments are defined.

A VerbNet for Arabic called "Arabic VerbNet" [44] was developed. It covers the most used verbs of MSA. The organization of verbs classes is as established by Levin's verb classes using the development procedure of Schuler Kipper but with some adaptations. The current version of Arabic VerbNet has 334 classes, which contain 7672 verbs and 1393 frames.

Classes (Figure 8) provide information about verb root, the deverbal form, the participle of verbs (members) belonging to the same class. Thematic roles are described (possibly with constraints), followed by a set of syntactic descriptions (with an example of sentence) and semantic relations between the arguments of the verb.

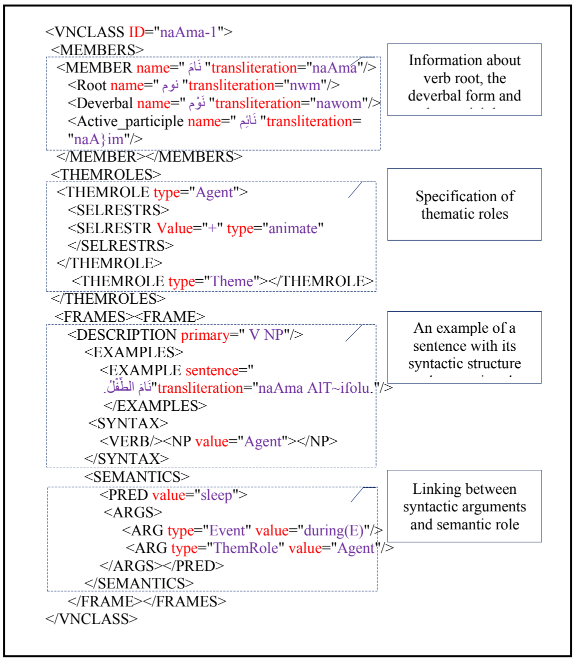

Fig. 8. Extract from naAma.xml (sleep) a class of ArabicVerbNet.

We have reviewed all of Arabic VerbNet classes. First, we have grouped information according to the syntactic structure of the sentence. Each structure group is associated with a family trees defined by our grammar.

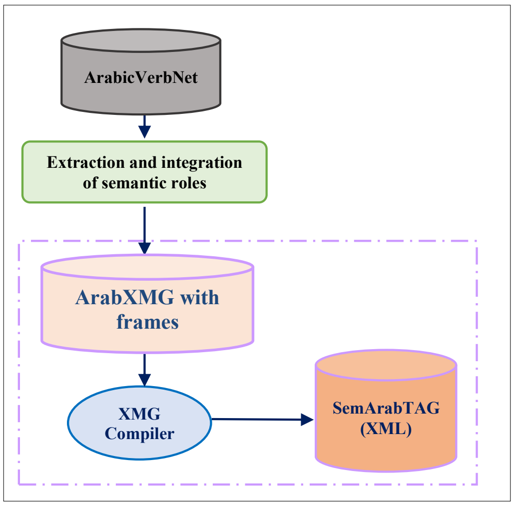

Fig. 9. Mapping between ArabicVerbNet and the metagrammar.

Then, for each syntactic structure, we extract the set of its corresponding semantic roles and frames. These semantic frames are described within our meta-grammar and every argument of the predicate is labeled with a semantic role. Finally, the XMG compiler generates our grammar as described previously.

V. SEMANTIC ROLE LABELING AND AMBIGUITY RESOLUTION

A syntactic structure can correspond to many semantic frames. These frames have different meanings that may give rise to several possible interpretations. For example, a subject can be Agent or Actor depending on the contexts. Furthermore, many verbs allow their semantic roles to be realized in various syntactic positions. For instance, verbs like "أعطَى " (*give*) can realize the THEME and GOAL arguments in two different ways:

1. أعطَى عليّ الكتابَ إلى فاطمة / Ali gives the book to Fatima: AGENT {Ali}+ THEME{the book}+ GOAL{to Fatima}.
2. أعطَى عليّ فاطمة الكتابَ / Ali gives Fatima the book: AGENT {Ali}+ GOAL{Fatima}+ THEME {the book}.

These multiple argument structure realizations are called verb alternations or diathesis alternations. So during the analysis, we have taken into consideration the following criteria, to resolve semantic ambiguities:

- The phrase type of the constituent: some semantic roles tend to appear as NPs, others as PP, and so on.
- The governing predicate: the base-labeled feature are defined according to a particular verb.
- The named entity type of the constituent: if it is a proper noun of persons, locations, organizations etc.
- The voice of the clause: active and passive sentences have different linking of semantic roles.
- The selectional restriction: constraints that a verb imposes to its argument roles.

We carried out a statistical study on Arabic VerbNet. We noted that for a given structure we have a large number of semantic frames. However, knowing the class of the verb allows us to considerably restrict this number. For example, for the syntactic structure (v np0 np1 pp0) Arabic VerbNet admits 70 possible semantic frames, but with the verb class, this number will be lowered on average to 4 possible semantic frames.

Let us consider the following examples to explain how we can resolve the problem of semantic ambiguity using the above information in Arabic VerbNet.

Since the same syntactic structure may have several possibilities of semantic roles, we began by restricting the field of analysis to the verb class. In fact, the verb is the governing predicate, and its class can reduce the ambiguity. Let us analyze the two following sentences whose syntactic structure is (v np0 np1):

1. بَدَأَ الأستاذُ الدرسَ / *the teacher started the course*.
2. شَرَحَ الأستاذُ الدرسَ / *the teacher explained the course.*

This two sentences show *"*الأستاذُ*"* (*the teacher*) as the subject and *"*الدرسَ*"* (*the course*) as the direct object. Based on the governing predicate's class, we apply the semantic role labels to these arguments:

- For sentence (1), the verb class of "بَدَأَ" (*start*) admits these two roles: AGENT and THEME: AGENT{ *the teacher* }+ THEME {*the course*}.
- For sentence (2), the verb class of "شَرَحَ" (*explain*) admits these two roles: AGENT and TOPIC: AGENT{ *the teacher* }+ TOPIC {*the course*}.

This distinction is explained by the fact that the verb class specifies semantic roles. Indeed, the semantic role labels THEME is central to an event or state that does not have control over the way the event occurs and is not structurally changed by the event. This is the case of verbs such as "بَدَأَ" (*start*). However, TOPIC is a type of THEME that is specific to verbs of communication such as "شَرَحَ" (*explain*), "دَردشَ" (*chat*), "تَحدثَ" (*converse*), etc.

Particle can be used to restrict the choice of the corresponding frame. Let us analyze the following sentences whose verb is " نبحَ " (*bark*):

3. نبحَ الكلبُ على الهرِّ / *The dog barks on the cat*.
4. نبحَ الكلبُ منَ الخوفِ / *The dog barks out of fear.*

These two sentences have the same syntactic structure (v np0 pp0). This type of structure has 56 possible semantic frames. However, the verb class of " نبحَ " (*bark*), animal_sounds-1, allows us to reduce significantly this number to 3 for the structure (v np0 pp0):

a) AGENT +{*particle:* على }+ RECIPIENT : the particule "على" (*on*) indicates that the semantic role of the object is a RECIPIENT.
b) AGENT+ {*particle:*من } + CAUSE: the particle "من" (*of*) requires that the semantic role is a CAUSE.
c) LOCATION+{*particle:* بِ}+ AGENT: the particle "بِ " (*with*) indicates that the semantic role of the object is an AGENT.

Using particle restriction, we reduce the choice of the corresponding frame for the two previous sentences (3) and (4). Sentence (3) contains the particle "على" while sentence (4) contains the particle "من". We obtain the following correct semantic correspondences:

3. نبحَ الكلبُ على الهرِّ / *The dog barks on the cat* : - a) AGENT{*the dog*}+ part { على }+ RECIPIENT{*the cat*}.
4. نبحَ الكلبُ منَ الخوفِ / *The dog barks out of fear* : -b) AGENT {*the dog*}+ part { منَ }+ CAUSE{*fear*}.

We can also refer to the nature of the semantic roles and their constraints in order to obtain the correct semantic representation. Let us take, for example, the following two sentences with the verb "أحب" (*love*) :

5. يحبُ عليّ فاطمةً / *Ali loves Fatima* : (EXPERIENCER { *Ali* }+THEME { *Fatima* }).
6. يحبُ الكتابُ فاطمةً / *The book loves Fatima*: (EXPERIENCER {*the book*}+THEME { *Fatima* }).

These sentences are syntactically correct and have the same syntactic structure (v np0 np1) as well as the same semantic frame. However, by reviewing the constraints specified for the semantic roles within the verb class of the "أحب" (*love*), we noticed that the EXPERIENCER (subject) must be animated. Therefore, the first sentence is semantically correct while the second is not, since the subject "الكتاب" is a non-animated object.

From these findings, we can conclude that several information can help to remove the ambiguity of semantic analysis during semantic role labeling. We point to the verb class, the properties of the role (Example: animated agent) and the use of certain particles for prepositional phrases. All this information represents constraints that will be integrated in order to filter the correct semantic frames during the syntactic-semantic parsing.

However, in some cases, ambiguity cannot be removed during the semantic analysis. The reason is the insufficiency of semantic information that indicates the context of the sentence. For example, in this phrase: "أبلغنا الدليل " (the guide informed us), the subject "الدليل"(*the guide*) may refer to a tour guide or a book/directory. In the first case, it will have the role AGENT while in the second it will be an INSTRUMENT. The correct meaning can only be understood by knowing the context of the sentence. In this case, semantic ambiguity can only be resolved at a higher level: pragmatic analysis.

VI. Conclusion and Future Work

In this paper, we introduced a new approach to build Tree adjoining grammar representing syntax and semantic of Arabic. This grammar is rewritten using the metagrammatical description language XMG. This extensible language provides a compact representation of grammatical information and offers a mechanism for combining elementary fragments of information. Thus, it allows information sharing between grammatical structures and can be configured to describe different levels of the language such as semantic descriptions. The idea is to associate semantic frames with the defined families of elementary trees within our metagrammar. Semantic frames are effective in modeling and manipulating semantic knowledge. Semantic roles are extracted from the Arabic VerbNet resource and then integrated in our metagrammar. These roles are defined as generalizations of arguments of the predicate (verb) in order to capture regularities in semantic interpretation of syntactic representations. This allowed us to establish a correspondence between semantic arguments and syntactic arguments. The syntax-semantic interface then corresponds to the definition of links between nodes of arguments of the predicate and their possible semantic roles. This interface allows elementary frames to be correctly unified during the semantic composition.

On the syntactic side, our generated grammar covers verbal sentences (active and passive form), nominal sentences, nominal phrases and prepositional phrases. It deals with the free word order of elements within the syntactic components, the additional complements and the agglutinative forms. On the semantic side, we started analyzing some verbal sentences and we are fixing constraints to ameliorate the semantic role labeling phase.

In the near future, we aim to evaluate our grammar using a larger corpus of trees. We are exploiting the available annotated corpora as well as resources (such as VerbNet examples) to possibly construct a significant syntactic-semantic test corpus.

References

[1] J. Nerbonne, "A Feature-Based Syntax/Semantics Interface", Research Report RR-92- 42, Saarbrücken: DFKI, 1992.

[2] R.D. Van Valin, Jr, "Exploring the Syntax-Semantics Interface", Cambridge University Press,2005.

[3] B. Levin and M. Rappaport Hovav , "Argument Realization", Cambridge University Press, Cambridge, 2005.

[4] F.R. Chaumartin and S. Kahane, "Une approche paresseuse de l'analyse sémantique ou comment construire une interface syntaxe-sémantique à partir d'exemples", In Actes de TALN, Montréal, Canada, 2010. 146, 171

[5] A. Joshi, L. Levy and M. Takahashi, "Tree adjunct grammars. Journal of Computer and System Sciences," 10(1), 136 – 163, 1975.

[6] B. Crabbé, D. Duchier, C. Gardent, J. Le Roux and Y. Parmentier, " XMG :eXtensible MetaGrammar," Computational Linguistics, 39(3), 591–629, 2013.

[7] C. Aloulou, "Analyse syntaxique de l'Arabe: Le système MASPAR," RÉCITAL 2003, Batz-sur-Mer, 11-14, juin 2003.

[8] F. Ben Fraj, C. Ben Othmane Zribi, and M. Ben Ahmed, "Grammaire TAG pour l'Analyse Syntaxique de Textes en Arabe comme un Problème de Classification," In Proceedings, the 9th International Business Information Management conference, Arabic Information Processing session, Marrakech- Morocco, January 2008.

[9] A. Barhoumi, "Analyse syntaxique de la langue arabe Analyse syntaxique basée sur une méthode d'apprentissage automatique," Editions universitaires européennes, 2015.

[10] Y. Bahou, L. Hadrich Belguith, C. Aloulou and A. Ben Hamadou, "Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés", 15ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, 25-27 janvier 2006 à Tours – France 2008.

[11] B. Haddad and M. Yaseen, " A Compositional Approach Towards Semantic Representation and Construction of ARABIC", P. Blache et al. (Eds.): LACL 2005, LNAI 3492, pp. 147 – 161, 2005.

[12] A. Abdelkader, K. Haddar, A. Ben Hamadou, "Étude et analyse de la phrase nominale arabe en HPSG," TALN 2006, Leuven, 10-13, avril 2006.

[13] C. Pollard, I. Sag, "Head-Drive Phrase Structure Grammar", CSLI series, University of Chicago Press, Chicago, 1994.

[14] M.F. Mammeri and N. Bouhassain, "Implémentation d'un fragment de grammaire HPSG de l'arabe sur la plate-forme LKB," 3rd International Conference on Arabic Language Processing (CITALA'09), May 4-5, Rabat, Morocco, 2009.

[15] N. Habbash, O. Rambow, "Extracting a Tree Adjoining Grammar From the Penn ArabicTreebank," In Actes de JEP-TALN 2004, Traitement automatique de l'Arabe. Fès : 446- 454, 2004.

[16] M. Maamouri, A. Bies, T. Buckwalter and W. Mekki, "The penn arabic treebank :Building a large-scale annotated arabic corpus," NEMLAR International Conference on Arabic Language Resources and Tools, 2004.

[17] S.M. Schieber and Y. Schabes, " Synchronous tree-adjoining grammars", In Proceedings of the 13th international conference on computational linguistics, volume 3, pages 253–258, Helsinki, 1990.

[18] R. Nesson and S. Shiebers, “Simpler TAG semantics through synchronization”, In Formal Grammars, Malaga,2006.

[19] L. Danlos , “D-STAG : un formalisme pour le discours basé sur les TAG synchrones”, TALN 2007, Toulouse, 5–8 juin 2007.

[20] A. Joshi and K. Vijay-Shanker, “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, 1999.

[21] L. Kallmeyer and A. Joshi, “Factoring Predicate Argument and Scope Semantics : Underspecified Semantics with LTAG,” In Proceedings of 12th Amsterdam Colloquium, pages 169–174, December 1999.

[22] L. Kallmeyer and A. Joshi, “Factoring Predicate Argument and Scope Semantics : Underspecified Semantics with LTAG,” Research on Language and Computation, volume 1 :1-2, pages 3–58, 2003.

[23] A. Frank and J. Van Genabith, “GlueTag Linear Logic based Semantics for LTAG and what it teaches us about LFG and LTAG”, In Proceedings of LFG01, Hong Kong, 2001.

[24] C. Gardent and L. Kallmeyer, “Semantic construction in FTAG,” In Proceedings of the European chapter of the Association for Computational Linguistics (EACL’03), Budapest, 2003.

[25] L. Kallmeyer and M. Romero, “Scope and Situation Binding in LTAG using Semantic Unification”, Research on Language and Computation 6:3-52, 2008.

[26] C. Gardent , “Integrating a unification-based semantics in a large scale Lexicalised Tree Adjoining Grammar for French”,Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 249–256 Manchester, aout 2008.

[27] Y. Parmentier, “SemTAG : une plate-forme pour le calcul sémantique à partir de Grammaires d’Arbres Adjoints,” Ph.D thesis, université Henri Poincaré – Nancy 1, 2007.

[28] L. Kallmeyer and R. Osswald, “Syntax-Driven Semantic Frame Composition in Lexicalized Tree Adjoining Grammars,” Journal of Language Modelling Vol i2, pp. 1–63, 2013.

[29] D. J. Weir, “Characterizing Mildly Context-Sensitive Grammar Formalisms,” PhD thesis,Université de Pennsylvanie, États-Unis, 1988.

[30] F. Ben Fraj, “Construction d’une grammaire d’arbres adjoints pour la langue arabe,” In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles, Montpellier, France : Association pour le Traitement Automatique des Langues, 2011.

[31] C. Ben Othmane Zribi, F. ben fraj and M. Ben Ahmed , “Combining classifiers for supertagging Arabic texts,” Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2010. . 10.1109/NLPKE.2010.5587841, 2010.

[32] anonymous reference

[33] XTAG Research Group, “A lexicalized tree adjoining grammar for english,” Technical Report IRCS-01-03, IRCS, University of Pennsylvania, 2001.

[34] Y. Parmentier, L. Kallmeyer, T. Lichte, W. Maier and J. Dellert, “ TuLiPA : A Syntax-Semantics Parsing Environment for Mildly Context-Sensitive Formalisms,” In 9th International Workshop on Tree-Adjoining Grammar and Related Formalisms (TAG+9), p. 121–128,Tübingen, Germany, 2008.

[35] L. Kallmeyer and M. Romero, “LTAG Semantics with Semantic Unification,” In Proceedings of TAG+7, Vancouver, pages 155–162, May 2004.

[36] M. Romero and L. Kallmeyer, “Scope and Situation Binding in LTAG using Semantic Unification,” In Proceedings of the Sixth International Workshop on Computational Semantics IWCS-6, Tilburg, January 2005.

[37] C. Fillmore (1982), “Frame Semantics,” In H. P. CO, Ed., Linguistics in the morning calm, p. 111–137, 1982.

[38] C. Fillmore, C. Johnson, and M.R. Petruck, “Background to FrameNet,” International Journal of Lexicography, 16(3):235–250, 2003.

[39] G.Venturi G, “Semantic annotation of Italian legal texts: A FrameNet-based approach,” In: Mirjam Fried; Nikiforidou Kiki (eds.), Advances in Frame Semantics, Benjamins Current Topics, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 51-84, vol. 3(1), 2013.

[40] J. Trione, F. Béchet, B. Favre and A. Nasr, “Rapid FrameNet annotation of spoken conversation transcripts,” Joint ACL-ISO Workshop on Interoperable Semantic Annotation ,2015.

[41] B. Levin, “English Verb Classes and Alternations A Preliminary Investigation,” Chicago: University of Chicago Press, 1993.

[42] S. Kasper, “A comparison of thematic role theories,” Master Homework. Marburg University, 2008.

[43] K. Kipper, A. Korhonen, N. Ryant and M. Palmer, “A large-scale classification of English verbs Lang,” Resour. Eval. J., 42 (2008), pp. 21-40, 2008.

[44] J. Mousser, “A large coverage verb taxonomy for Arabic,” In: Seventh Conference on International Language Resources and Evaluation (LREC’10), Valetta, Malta, pp. 2675–2681, 2010.