Building a Syntactic-Semantic Interface for aSemi-Automatically Generated TAG for Arabic

Cherifa Ben Khelil[1,2], Chiraz Ben OthmaneZribi[1], Denys Duchier[2], and Yannick Parmentier[3]
[1]RIADI-ENSI, Université La Manouba, Tunisia
[2]LIFO, Université d'Orléans, France
[3]LORIA-Projet SYNALP, Université de Lorraine, France

Abstract: *Syntactic and semantic resources play an important role for various Natural Language Processing (NLP) tasks by providing information about the correct structural representations of the sentences and their meaning. To date, there is not a wide-coverage electronic grammar for the Arabic language. In this context, we present a new approach for building a Tree Adjoining Grammar (TAG) to represent the syntax and the semantic of modern standard Arabic. This grammar is produced semi-automatically with the eXtensible MetaGrammar (XMG) description language. First the syntax of Arabic is described using the defined Arab-XMG meta-grammar. Then semantic information is added by introducing semantic frame-based dimension into the meta-grammar. This is achieved by exploiting lexical resources such as ArabicVerbNet. Finally, the link between semantic and syntax is established using a syntax-semantic interface that allows the construction of sentence meaning through semantic role labeling. Experiments were performed to check grammar coverage as well as the syntactic-semantic analysis. The results showed that the generated grammar can cover the basic syntactic structures of Arabic sentences and the different phrasal structures with a precision rate of about 92%. Moreover, it confirms the effectiveness of the proposed approach as we were able to parse semantically a set of sentences and build their semantic representations with a precision rate of about 72%.*

Keywords: *TAG, meta-grammar, syntax-semantic interface, semantic frame, semantic role, Arabic language.*

Received February 19, 2018; accepted April 18, 2018

1. Introduction

Natural language allows building 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 for many applications in the field of Natural Language Processing applications (NLP) such as machine translation systems, human-machine dialogue, and 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.

Consequently, 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 that allows the construction of the semantic representation of a sentence based on the relationship between its syntactic constituents. The link between syntactic and semantic is expressed through rules describing the constituents in the sentence (e.g.,[8, 24, 29, 36]). The semantic representation of the sentence is constructed gradually in parallel to its syntactic structure and according to the used grammatical formalism. This implies that the syntax-semantic interface itself is closely related to the chosen grammar or grammatical formalism.

In this context, having electronic 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) [18] 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 eXtensible Metagrammar (XMG) [9].

The paper is organized as follows. In section 2, we present some related works. Then, in section 3, we introduce briefly the TAG formalism as well as the description of the syntax of Arabic using Arab-XMG meta-grammar. Section 4 describes the semantic integration process into the meta-grammar. Section 5 discusses the phase of semantic role labelingand some cases of semantic ambiguities. Finally, in Section 6, we present the results of the experiments we have carried out to evaluate our approach.

2. Related Work

During the last decade, several syntactic analysis approaches for standard Arabic have been proposed. We can mention for example the following works [2, 3, 4, 5]. However, approaches dealing with semantic are rare or even absent. For example, [16] 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 [1], which propose a Head-Driven Phrase Structure Grammar(HPSG)[33], essentially dealing with the nominal sentences and the works of [27] 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)) [26]. However, to date there is no broad coverage grammar of Arabic.

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 [34]. 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 [30] and French [10], which allowed to simultaneously generating syntactic and semantic analysis of sentences.

Another approach consists in adding semantic representations using the variables of unification (a set of attribute-value pairs that provides morphological, syntactic or semantic information) in the grammar [29].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 [17], an underspecified logic (using labels holes and range constraints) [19], a glue part [12] for English, or a flat semantic [13, 21] for English and [14, 32] for French. Thus, the final semantic representation of a sentence corresponds to the union of these associated semantic formulas.

More recently, the works of [20] propose to introduce another form of semantic representation, which is based on frame semantic. In this approach, each elementary syntactic constructionis 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.

3. Generating A Tree Adjoining Grammar for Arabic

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

3.1. Brief Presentation of the TAG Formalism

TAG [18] is a syntactic formalism that considers 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:

1. Initial tree (having substitution nodes marked with the symbol ↓).
2. Auxiliary tree (having a “foot node” marked with the symbol *).

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 appends a frontier node with another tree whose top node has the same symbol. The adjunction operation is more powerful since it allows inserting an auxiliary tree into the center of another tree.

TAG is considered the standard model for mild context-sensitivity [38]. It is slightly more powerful than context-free grammars, but strictly included in the class of contextual grammars. 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 is modelled by the operation of adjunction. Moreover, from a processing point of view, TAG remains analysable 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 specific phenomena in Arabic such as embedded structures and crossed dependencies.

3.2. TAG for Arabic: ArabTAG V2.0

Our work takes its origins from an existing handcrafted tree-adjoining grammar for Arabic named Arabic Tree Adjoining Grammar (ArabTAG) [6]. 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.

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 [7] that handles the aspects mentioned above. This new version is rewritten using the XMG description language [9]. With this formalism, we have semi-automatically generated ArabTAG V2.0 from a reduced description of grammar rules (see Figure 2). First, the metagrammatical language XMG is used to define Arabic-XMG. Then, this compact description is automatically compiled into the ArabTAG V2.0 grammar by the XMG2 compiler.

Our choice of applying such grammar production technique for the description of the Arabic language was motivated by many reasons. Indeed, using description languages offers a relatively good control on the grammar being produced. Information is shared among grammatical structures while ensuring a high degree of modularity within the target grammar. This allows the extension of the produced grammar with various levels of description such as morphology or semantics. Moreover, semi-automatic grammar production saves time and decreases costs.

Arabic-XMG is described as (conjunctive and disjunctive) combinations of tree fragments. Such fragments are defined as formulas of a tree description logic based on dominance and precedence relations between node variables. For example, to describe 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. We have combined 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. We refer the reader to [7] for additional information about Arabic-XMG meta-grammar.

3.3. Coverage and Validation of ArabTAG V2.0

Up to now, we have generated 624 trees from a description made of 29 classes (that is, 29 tree fragments or combination rules) as shown in Figure 1.

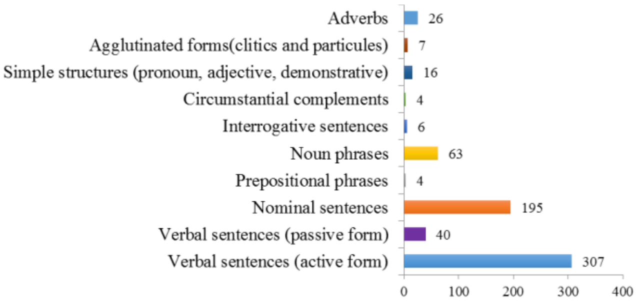

Figure 1. Current ArabTAG V2.0 coverage.

The current version of the grammar covers verbal phrases (active and passive form), nominal sentences and phrasal structures. These latter have several types: noun phrase (مركب اسمي), subordinate phrase (مركب موصولي) and prepositional phrase (مركب حرفي). In addition, it covers elliptical and subordinate structures. It takes into consideration the change of the order of the sentence's components and the agglutinative forms. It contains also elementary trees for the representation of additional complements such as circumstantial complement of time, circumstantial complement of place and adverbs.

In order to verify grammar coverage, we have set up a development environment while designing ArabTAG with XMG. Moreover, we defined proof of concept syntactic and morphological lexicons for Arabic following the 3-layer lexicon architecture (tree templates, lemmas, words) of the XTAG project [39].

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 is to evaluate and to reduce both under and over-generation. Our grammar must be able to recognize valid sentences that cover linguistic phenomena of Arabic (sentences described in schoolbooks, Arabic news, etc.,) and to reject ungrammatical sentences. Each new syntactic phenomena included in ArabTAG V2.0 leads to the extension of a test corpus gathering both grammatical and ungrammatical sentences. This corpus is called corpus of phenomenon since it contains typical sentences and it is constructed gradually.

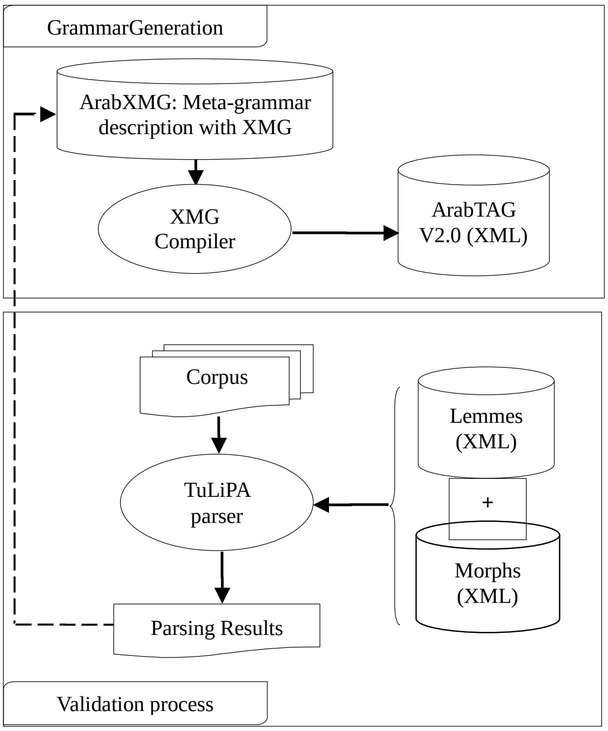

Figure 2.Validation architecture of ArabTAG V2.0.

We used the TuLiPA parser [31] on this 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 defined TAG structures when it is extended.

The corpus of phenomenon is made of 212 examples of phrases and sentences (150 grammatical sentences and 62 ungrammatical sentences). It contains 134 verbal sentences, 45 nominal sentences, 32 nominal phrases and 1 prepositional phrases. Ungrammatical clauses were mainly added to check if the grammar could return syntactic configurations with incorrect agreement. The following table summarizes the different phenomena covered by our grammar:

Table 1. Phenomena covered by the corpus.

Phenomenon	**Number of sentences/phrases**
Active forms	123
Adverbial object	6
Agglutination forms	26
Agreement rules	25
Circumstantial complement	9
Ditransitive verbs	67
Elliptical subject	17
Embedded structures	11
Free word order	44
Interrogative Sentences	10
Intransitive verbs	29
Passive forms	11
Transitive verbs	38

4. 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. We decided to use semantic frame as we noted that frame semantic make the interfacing easier between syntax and semantic.

Semantic frames have been implemented in the Berkeley University FrameNet project [11] to provide a frame database for English language. This base has been used in several works such as [37] and [35] for the task of semantic role labelling that consists in automatically finding the semantic roles of each argument of each predicate in a sentence. FrameNet exists for several languages such as French, Chinese, Spanish and Japanese.

Since we do not have such a basis for Arabic, we have thought as [20] to associate each elementary tree with an elementary frame and during the syntactic analysis we build the final frame representing the sentence meaning by unifying these elementary frames. 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.
3. Within remaining frames, we define the semantic information specific to the lexicon.
4. Each frame is labelled 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.

Let us consider an example of the “pursuit” frame with the verb “طارد”(*chase*). We want to obtain the final semantic frame for the following sentence “ طارد الشرطي اللص” (*The policeman chases the thief*). The verb “طارد”(*chase*) 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).

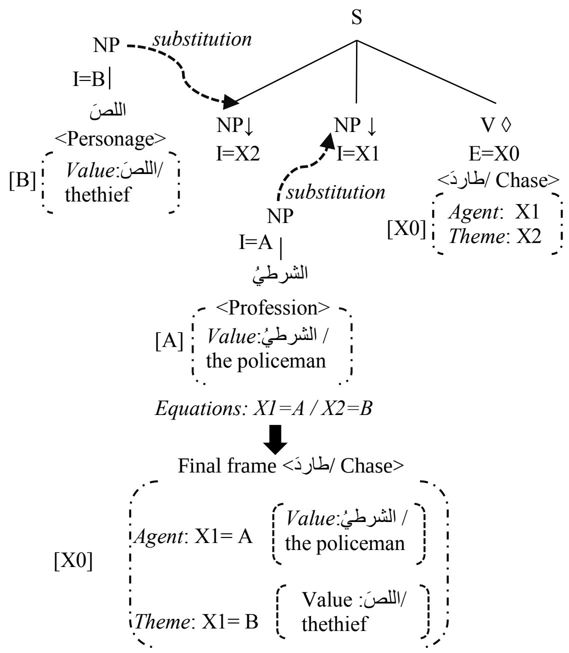

Figure 3. Frame composition for “طاردَ الشرطيُ اللصَ” / The policeman chases the thief.

The elementary frames (personage and profession) are associated with the elementary trees for the noun phrases. The interface feature I (see Figure 3) 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 theory [22, 25]. According to this theory, the syntactic behaviour of a verb can be predicted from its semantic. For example, if the actor of the verb is present in the sentence, it will be in the subject / nominative position. In the Linking theory this actor has been granted an autonomous status: they can be called “agents”. Thus, if the governing 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.

4.1. 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 representation as follows (Figure 4):

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 (they include the anchor node marked with the symbol ◊) 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.

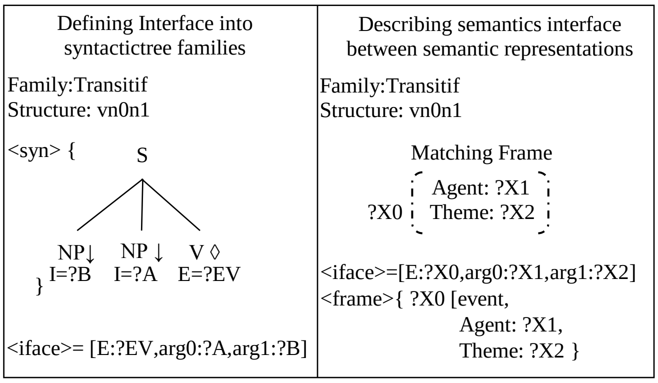

Figure 4. Describing the syntax/semantic interface between semantic representations and syntactic trees with XMG.

The semantic information of the “pursuit” frame is also described in Figure 4. 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.

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

4.2. Mapping between Arab-XMG and ArabicVerbNet

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

VerbNet [23] is a lexical resource for English verbs. It is based on the semantico-syntactic classification system of verbs of [28]. Verbs with similar syntactic and semantic behaviour 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” [28] 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 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.

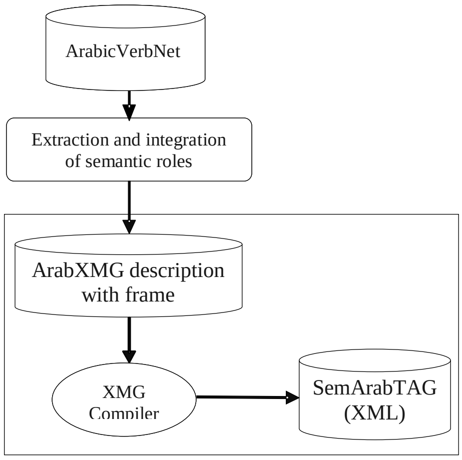
ArabicVerbNet
Extraction and integration of semantic roles
ArabXMG description with frame
XMG Compiler
SemArabTAG (XML)

Figure 5. Mapping between Arabic VerbNet and the metagrammar.

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

5. 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 context. 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:

- “أعطَى عليٌّ الكتابَ إلى فاطمة” / Ali gives the book to Fatima: AGENT {“عليٌّ”/ Ali} + THEME {“الكتابُ”/ the book} + GOAL {“فاطمةً”/ Fatima}.
- “أعطَى عليٌّ فاطمة الكتابَ” / Ali gives Fatima the book: AGENT {“عليٌّ”/ Ali} +GOAL {“فاطمةً”/ Fatima} + THEME {“الكتابُ”/ 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 features 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 preposition: when a semantic role appears as PP, the preposition can indicate its meaning and would restrict the choice of the corresponding frame.
- The selectional restriction: constraints that a verb imposes to its argument roles as animated, human, concrete, etc.

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

Preposition 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" and the same verb. 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 particle "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.

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

1. "نبحَ الكلبُ على الهرِّ" /The dog barks on the cat: a) AGENT {"الكلبُ"/ the dog} + part {"على"} + RECIPIENT {"الهرِّ"/ the cat}.
2. "نبحَ الكلبُ منَ الخوفِ" / The dog barks out of fear: b) AGENT {"الكلبُ"/ the dog} + part {"منَ"} +CAUSE {"الخوف"/ 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):

3. "يحبُ عليٌّ فاطمةً" / Ali loves Fatima: (EXPERIENCER {"عليٌّ"/ Ali} +THEME {"فاطمةً"/ Fatima}).
4. "يحبُ الكتابُ فاطمةً" / The book loves Fatima: (EXPERIENCER {"الكتابُ"/ the book} +THEME {"فاطمةً"/ 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 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 as pragmatic.

6. Experiments

In order to evaluate our approach, we have built a test corpus of 500 sentences (347 verbal sentences and 153 nominal sentences) selected arbitrarily from a Tunisian schoolbook (eight grade). We have developed a tool to carry out the syntactic-semantic analysis. First, this tool ensures the morpho-syntactic labeling of the input sentence, and then the syntactic-semantic analysis is performed. The assignment of the semantic frames is done during the parsing through the syntax / semantic interface. As a sentence is analyzed, its semantic frame is constructed by unifying the elementary frames of its constituents with semantic frame of the predicate. At the end of this analysis, we obtain the syntactic tree corresponding to the parsed sentence and its final semantic frame.

Since nominal sentences do not have a verb as the governing predicate, we only expose the results obtained for the 347 verbal sentences of our test corpus.

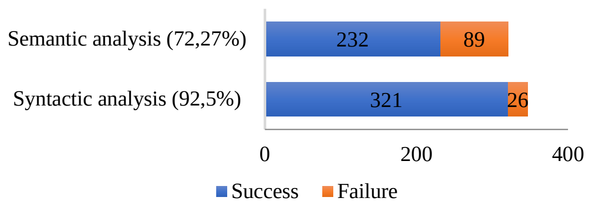

Figure 6. Result of syntactic-semantic analysis of verbal sentences.

We succeeded to parse syntactically 321 sentences representing 92.50% of the tested sentences. This success is illustrated by getting the correct syntactic trees at the end of the analysis. Semantically, we were able to obtain 232 correct semantic frames matching 72.27% of the parsed sentences.

The 27.73% failure rate of the semantic analysis is mainly due to a lack of coverage in ArabicVerbNet: In fact, we found that 19.93% of the verbs of tested sentences are not defined in this resource. Moreover, for a given verb in ArabicVerbNet, the list of syntactic structures of sentences is not exhaustive. Besides, we measured a 5.29% failure rate due to the absence of the corresponding sentence structure during the semantic role labeling. We also got 2.49% failure rate for complex sentences. The analysis of this kind of sentences is more complicated since they contain more than one verb.

Although these first results are encouraging, we aim to evaluate our grammar using a larger corpus and compare our results with results of other approaches.

7. Conclusions

In this paper, we introduced a new approach to build Tree adjoining grammar representing syntax and semantic of Arabic. The idea is to associate semantic frames with the defined families of elementary trees within our metagrammar. Plausible semantic roles are extracted from the Arabic VerbNet resource. 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 define the syntax-semantic interface that corresponds to the definition of links between nodes of arguments of the predicate and their possible semantic roles.

Our generated grammar covers verbal sentences, 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.

We have evaluated our approach using a corpus of 500 sentences. Although these first results are satisfactory, we aim to evaluate our grammar using a larger corpus. Furthermore, we are exploring the possibility to apply machine learning techniques to compensate the lack of data in Arabic VerbNet.

References

[1] Abdelkader A., Haddar K., Ben Hamadou A., “Étude et analyse de la phrase nominale arabe en HPSG,” *Verbum ex machina (TALN vol. 1)*, Presses universitaires de Louvain, 2006.

[2] Aloulou C., “Analyse syntaxique de l’Arabe: Le système MASPAR,” *RÉCITAL 2003*, Batz-sur-Mer, 2003.

[3] Bahou Y., HadrichBelguith L., Aloulou C., and Ben Hamadou A., “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*, Tours, pp. 25-27, 2008.

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

[5] Ben Fraj F., Ben OthmaneZribi C., and Ben Ahmed M., “Grammaire TAG pour l’Analyse Syntaxique de Textes en Arabe comme un Problème de Classification,” *in Proceedings of the 9th International Business Information Management Conference*, Marrakech, pp. 1-8, 2008.

[6] Ben Fraj F., “Construction d’une grammaire d’arbres adjoints pour la langue arabe,” *in Proceedings of the Actes de la 18e Conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, 2011.

[7] Ben Khelil C., Duchier D., Parmentier Y., Zribi C., and Ben Fraj F., “ArabTAG: from a Handcrafted to a Semi-automatically Generated TAG,” *in Proceedings of the 12th International Workshop on Tree-Adjoining Grammars and Related Formalisms*, Düsseldorf, 2016.

[8] Chaumartin F. and Kahane S., “Une Approche Paresseuse de L’analyse Sémantique ou Comment Construire une Interface Syntaxe-Sémantique à Partir D’exemples,” *in Proceedings of Actes de TALN*, Montréal, pp. 146-171, 2010.

[9] Crabbé B., Duchier D., Gardent C., Le Roux J., and Parmentier Y., “XMG: eXtensible Metagrammar,” *Computational Linguistics*, vol. 39, no. 3, pp. 591-629, 2013.

[10] Danlos L., “D-STAG: Un Formalisme Pour Le Discours Basé sur les TAG Synchrones,” *Revue TAL*, vol. 50, no.1, pp. 111-143, 2009.

[11] Fillmore C., Johnson C., and Petruck M.R., “Background to FrameNet,” *International Journal of Lexicography*, vol. 16, no. 3, pp. 235-250, 2003.

[12] Frank A., Van Genabith J., Butt M., and King T., “GlueTag Linear Logic based Semantics for LTAG and what it teaches us about LFG and

LTAG," *in Proceedings of LFG01*, Hong Kong, 2001.

[13] Gardent C. and Kallmeyer L., "Semantic Sonstruction in FTAG," *in Proceedings of the European Chapter of the Association for Computational Linguistics*, Budapest, pp. 3-8, 2003.

[14] Gardent C., "Integrating a Unification-Based Semantics in a Large Scale Lexicalised Tree Adjoining Grammar for French," *in Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, pp. 249-256, 2008.

[15] Habbash N. and Rambow O., "Extracting a Tree Adjoining Grammar from the Penn Arabic Treebank," *JournéesD'Étudessur la Parole*, pp. 446-454, 2004.

[16] Haddad B. and Yaseen M., "A Compositional Approach Towards Semantic Representation and Construction of ARABIC," *in Proceedings of Logical Aspects of Computational Linguistics*, Bordeaux, pp. 147-161, 2005.

[17] Joshi A. and Vijay-Shanker K., "Compositional Semantics with Lexicalized Tree Adjoining Grammar (LTAG): How Much Under specification is Necessary?," *Computing Meaning*, Springer, 1999.

[18] Joshi A., Levy L., and Takahashi M., "Tree Adjunct Grammars," *Journal of Computer and System Sciences*, vol. 10, no. 1, pp. 136-163, 1975.

[19] Kallmeyer L. and Joshi A., "Factoring Predicate Argument and Scope Semantics: Underspecified Semantics with LTAG," *Research on Language and Computation*, vol. 1, no. 1-2, pp. 3-58, 2003.

[20] Kallmeyer L. and Osswald R., "Syntax-Driven Semantic Frame Composition in Lexicalized Tree Adjoining Grammars," *Journal of Language Modelling*, vol. 1, no. 2, pp. 1-63, 2013.

[21] Kallmeyer L. and Romero M., "Scope and Situation Binding in LTAG using Semantic Unification," *Research on Language and Computation*, vol. 6, no. 1, pp. 3-52, 2008.

[22] Kasper S., "A comparison of Thematic Role Theories," M.S. Thesis, Marburg University, 2008.

[23] Kipper K., Korhonen A., Ryant N., and Palmer M., "A Large-Scale Classification of English Verbs Lang," *Language Resources and Evaluation*, vol. 42, no. 1, pp. 21-40, 2008.

[24] Levin B. and Hovav M., *Argument Realization*, Cambridge University Press, 2005.

[25] Levin B., *English Verb Classes and Alternations a Preliminary Investigation*, University of Chicago Press, 1993.

[26] Maamouri M., Bies A., Buckwalter T., and Mekki W., "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus," *in Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, 2004.

[27] Mammeri M. and Bouhassain N., "Implémentation d'un fragment de grammaire HPSG de l'arabesur la plate-forme LKB," *in Proceedings of the 3rd International Conference on Arabic Language Processing*, Rabat, 2009.

[28] Mousser J., "A Large Coverage Verb Taxonomy for Arabic," *in Proceedings of the7th Conference on International Language Resources and Evaluation*, Valetta, pp. 2675-2681, 2010.

[29] Nerbonne J., "A Feature-Based Syntax /Semantics Interface," *Annals of Mathematics and Artificial Intelligence*, vol. 8, no. 1-2, pp. 107-132, 1993.

[30] Nesson R. and Shiebers S., "Simpler TAG Semantics through Synchronization," *in Proceedings of the 11th Conference on Formal Grammer*, Malaga, 2006.

[31] Parmentier Y., Kallmeyer L., Lichte T., Maier W., and Dellert J., "TuLiPA: A Syntax-Semantics Parsing Environment for Mildly Context-Sensitive Formalisms," *in Proceedings of the 9th International Workshop on Tree-Adjoining Grammar and Related Formalisms*, Tübingen, pp. 121-128, 2008.

[32] Parmentier Y., "*Semtag: Une Plate-Forme Pour Le Calcul Sémantique a Partir De Grammaires d'Arbres Adjoints*," Ph.D Thesis, université Henri Poincaré-Nancy 1, 2007.

[33] Pollard C. and Sag I., *Head-Drive Phrase Structure Grammar*, University of Chicago Press, 1994.

[34] Schieber S. and Schabes Y., "Synchronous Tree-Adjoining Grammars," Technical Report, 1990.

[35] Trione J., Béchet F., Favre B., and Nasr A., "Rapid FrameNet Annotation of Spoken Conversation Transcripts," *in Proceedings of the Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, London, 2015.

[36] Van Valin R., *Exploring the Syntax-Semantics Interface*, Cambridge University Press, 2005.

[37] Venturi G., "Semantic Annotation of Italian Legal Texts: a Framenet-Based Approach," *Advances in Frame Semantics*, vol. 3, no. 1, pp. 46-79, 2011.

[38] Weir J., "Characterizing Mildly Context-Sensitive Grammar Formalisms," Ph.D thesis, Université de Pennsylvanie, 1988.

[39] XTAG system "A Lexicalized Tree Adjoining Grammar for English," Technical Report, 2001.

Cherifa Ben Khelil received her Master’s in Software Engineering from Higher Institute of Computer Science Ariana, Tunisia, and she is pursuing her Doctoral degree under joint supervision between the National School of Computer Sciences (ENSI), University of La Manouba in Tunisia and the University of Orleans in France. Her research interests are related to Natural Language Processing in particular grammar generation to represent the syntax and the semantic of Arabic.language.

Chiraz Ben Othmane Zribi is a professor at the National School of Computer Science, University of La Manouba, Tunisia and a researcher at the RIADI-GDL laboratory. She received her PhD in computer science in 1998 from PARIS XI University, France. Her principal research interests are in the area of Arabic language processing. Her recent work has focused on natural language parsing, detection and correction of errors, generation of dictionaries and knowledge retrieval.

Denys Duchier has been Professor of Computer Science at Universitéd 'Orléans, France, since 2006. He received his PhD from Yale University, United States, in 1991. After postdoctoral fellowships at University of Ottawa and University of Vancouver, Canada, he moved in 1996 to Saarland University, Germany, where he worked on the design and implementation of the Oz programming language. His research interests focus on the application of constraints in computational linguistics, and on the design and implementation of programming languages.

Yannick Parmentier is an Associate Professor at Université de Lorraine, France. He got his PhD in Computer Science from Henri Poincaré University in Nancy, France, in 2007. During his PhD, he took part in the design and implementation of the XMG description language and its application to the formal description of French. In 2007-2008, he was a postdoctoral fellow at University of Tübingen, Germany, where he worked on symbolic parsing. From 2009 to 2017, he was an Associate Professor at University of Orléans working on constraint-based approaches in computational linguistics.