ArabTAG: a Tree Adjoining Grammar for Arabic Syntactic Structures
Fériel Ben Fraj, Chiraz Ben Othmane Zribi, Mohamad Ben Ahmed
RIADI Laboratory, National School of Computer Sciences, 2010, University of La Manouba, Tunisia
Feriel.BenFraj@riadi.rnu.tn, cbo@tunet.tn, Mohamed.BenAhmed@riadi.rnu.tn

ABSTRACT
In order to construct a generic grammatical resource for Arabic language, we have chosen to develop an Arabic grammar based on TAG formalism. Our choice is, especially, justified by complementarities that we have noticed between Arabic syntax and this grammatical formalism. This paper consists of two comparative studies. The first is between a set of unification grammars. The second is between the characteristics of Arabic language and those of TAG formalism. These comparisons lead us to defend our choice and to describe the structure and characteristics of ArabTAG: a TAG grammatical model for presenting modern Arabic syntactic structures constructed using XML language. ArabTAG will be useful for Arabic processing applications, especially parsing task.

Keywords: Arabic language, unification grammars, TAG formalism, elementary trees, NLP.

1. INTRODUCTION
The grammar constitutes an important component for different NLP applications. It gives an information source to resolve different linguistic phenomena and ambiguities. Since the construction of the first formal grammar by Chomsky in 1970[1], three formal grammatical paradigms have been consecutively developed. These generations are generative, transformational and unification grammars.

The generative grammars consist of four classes which are regular, context sensitive, contextual and non contextual grammars. All these classes are characterized by their grammatical rules, called rewriting rules. The general form of these rules is: $X \rightarrow YZ$, where $X$, $Y$ and $Z$ can be terminal or non terminal. Nevertheless, all classes differ according to the categories of their specific members $X$, $Y$ and $Z$. This generation has largely served the NLP applications especially its free context class. But, it has multiple limits that we can make a note of the followings. So, generative grammars cannot consider the grammatical marks as the agreement rules when representing a sentence. In addition, they present only the simple syntactic structures and cannot present crossed ones which are very frequent in natural languages.

Chomsky noticed the limits of his first generation. Then he proposed the second generation: the transformational grammar as a remedy. It is an amelioration of rewriting rules by adding syntactic marks to them. This generates an unlimited number of transformational rules[2].

The unification grammar generation tries to resolve this inconvenience by generalizing rules. Unification generation of grammars is the most recent and expressive one. It represents a remedy to the two previous generations. Several types of grammars belong to unification generation. Nevertheless, each generation has its characteristics and advantages but also possess some inconveniences when representing the syntax of a specific language. So, the primordial question that should be asked, when representing a language, is: What kind of grammar is the most suitable for the specificities of a target language?

In this paper, we will concentrate on the unified paradigm. Then, the second section establishes a comparative study between different unification grammars. The comparison invokes their advantages and disadvantages in order to choose the most appropriate for representing Arabic language. The motivations of our choice are described in details in the third section. Whereas, the fourth one presents a description of ArabTAG: a grammatical model that we have constructed for representing modern Arabic syntactic structures.

2. COMPARATIVE STUDY OF UNIFICATION GRAMMARS
According to the previously invoked limits of the two first generations of grammars, we concentrate our study only on unification grammars for a further comparison study.

2.1. INTRODUCTION TO UNIFICATION GRAMMAR?
As rewriting rules for generative grammars, feature structures are for unified grammars. Nevertheless, different kinds of information can be presented by a unified grammar. They enclose linguistic, syntactic as well as semantic knowledge which are organized in pairs of (attribute, value) as it is shown in the following example. The example structure encloses a generalization part (the attributes) and a specification part presented by the linguistic values of the verb “eat”.

Nevertheless, unified grammars differ from each other by the kinds of information manipulated, the manner to present these information or/and the way of unifying feature structures.
Figure 1: Example of feature structure as represented in HPSG.

Then, unification consists of combining two simple feature structures based on their correspondent features if they can be in part superposed. The superposed fragments are identified and so unified. As a result, a new combined structure is constructed [3].

The common features of the different kinds of unification grammars are resumed in the following points. So, they:
- present a flat structure because the sentence is described and analyzed as it is announced in the text;
- are inductive because structures can be developed progressively;
- are declarative as the grammatical structures are directly represented without specifying the procedure used to get them;
- are based on feature structures that can be heterogeneous (lexical, syntactic and semantic features). This helps to disambiguate structures in NLP applications.

2.2. COMPARISON

In this section, we present a comparative study that we have elaborate over seven unification grammars: Generalized Phrase Structure Grammar (GPSG) [4], Head-driven Phrase Structure Grammar (HPSG) [5], Lexical Functional Grammar (LFG) [6], Tree Adjoining Grammar (TAG) [7], Dependency Grammar (DG) [8], Categorical Grammar (CG) [9] and Definite Clause Grammar (DCG) [10].

Firstly, let us specify our comparative criteria. These criteria emerge from the characteristics that should be present in a formal grammar to be considered a good formalism for presenting natural languages syntaxs. Then, we consider:
- the context level that specifies if the information in feature structures are less or more contextually inter-related. This level should be neither very high nor low to have a good representative grammar;
- the representational power that specifies the legibility and re-usability of the information presented by the grammar. A good grammar should have a high representational power;
- the combinatory power which consists in the faculty of the grammar to represent crossed and recursive language structures (subordinate structures). As the representational criterion, combinatory one should be sufficiently high to enclose most of the syntactic structures;
- the NLP treatment complexity when using the grammar in an NLP application. Representation complexity should be in compromise with NLP treatment complexity.

According to the table above, we note that GPSG and HPSG have the same features. In fact, GPSG is the predecessor of HPSG. These two grammars use the grammatical rules of Chomsky associated to transformational paradigm to create generalized rule models. GPSG and HPSG do not dissociate between the representational levels (lexical, syntactic and semantic). Furthermore, they are weakly contextual. However, LFG is highly contextual which constitutes a problem for the NLP applications.

<table>
<thead>
<tr>
<th>Criteria Grammars</th>
<th>Contextual Level</th>
<th>Representational power</th>
<th>Combinatory power</th>
<th>NLP treatment complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFG</td>
<td>high</td>
<td>high</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>GPSG</td>
<td>low</td>
<td>middle</td>
<td>low</td>
<td>middle</td>
</tr>
<tr>
<td>HPSG</td>
<td>low</td>
<td>middle (shared structures)</td>
<td>low</td>
<td>middle</td>
</tr>
<tr>
<td>TAG</td>
<td>middle</td>
<td>high (simplicity &amp; shared structures)</td>
<td>high (crossed structures)</td>
<td>low</td>
</tr>
<tr>
<td>DG</td>
<td>low</td>
<td>low</td>
<td>middle</td>
<td>low</td>
</tr>
<tr>
<td>CG</td>
<td>middle</td>
<td>low</td>
<td>middle</td>
<td>low</td>
</tr>
<tr>
<td>DCG</td>
<td>middle</td>
<td>low</td>
<td>low</td>
<td>low</td>
</tr>
</tbody>
</table>

Table 1: Comparative study between unification grammars.

For the three ultimate grammars (DG, CG and DCG), we denote their weak legibility and absence of information sharing between structures. For these three kinds of grammars, the major advantage is the lower complexity when they are used in NLP treatment.

Ultimately, all grammars, except TAG, can not represent crossed structures. Then, the characteristics of TAG motivate us: this formalism seems to be more efficient than the others because it has the best criteria over all the others. So, we consider this formalism in the following section to study its usefulness for representing Arabic language.
3. WHY TAG FORMALISM FOR ARABIC?

3.1. TAG FORMALISM

TAG is initially created by Joshi [7]. It is a unification grammar. It uses feature structures for representing the shared characteristics between language elements. Its feature structure is richer than those of the other grammatical formalisms. In fact, it is not a list of linguistic elements but it consists of organized tree structures. In addition, tree structure does not enclose only the word (or its part-of-speech) information but associate to it some surrounding context. So, two different structures exist. The first ones are initial trees that can be joined to other elementary trees by substitution. The second ones are auxiliary trees that can be associated with other elementary trees by adjunction. The root node of the second kind of elementary trees should have the same category as one of the feat nodes in order to specify the adjunction position. The following figure shows these two kinds of elementary trees and the two operations used for their combination.

![Figure 2: Examples of elementary trees: (a) initial trees, (b) auxiliary tree.](image)

Then, two resulting trees are created: the derived one which is the complete syntactic structure of a sentence whereas the second tree is the derivation one which presents the manner of constructing the derived tree. Nodes of a derivation tree correspond to the elementary structures whereas arcs present the operations used for gluing them.

To resume, we can simply say that TAG uses two operations to combine two elementary structures in order to generate two final trees.

In addition, TAG is called a structural grammar and is mildly contextual since its structures are not very contextual inter-related.

3.2. ARABIC LANGUAGE SPECIFICITIES

Arabic language is a Semitic language. It has some specific features that let it more ambiguous than other natural languages. The ambiguity has several reasons: lexical, syntactic and semantic. In fact, a word in Arabic can have more than one lexical and grammatical interpretation [11]. The variety is caused by the:

- vocalic phenomena because the vocalic signs play an important role in the comprehension of the Arabic words. The absence of these signs, which is frequent in modern Arabic, increases the lexical and grammatical ambiguities.
- agglutination characteristic since Arabic is an agglutinative language. It allows tying up clitics to simple forms in order to have more complex forms. Then, a sentence can correspond to just one agglutinative form as for the example: "فزّرناهم" *(and we visited them)* where all sentence’s components (verb, subject and object) are enclosed in the same textual form.
- multiplicity of the sentences’ structures given that the order of the sentence’s components in Arabic language is not already fixed but flexible. Then, for the three basic components Subject, Verb and Object, we can have different combinations (VSO, VOS, SVO) that are all syntactically corrects [12] [13].
- recursive structures as in the Arabic sentences, there is an abundance use of recursive structures as in the sentence: "الكلاب هي التي أعفته نباحها الذي لم يتوقف طوال الليل" *(the dogs are those that aroused him with their barking that do not stop during the night)*. Consequently, the sentence length is not limited.

3.3. ARABIC LANGUAGE AND TAG

After looking at the characteristics of TAG and those of Arabic language, we denote the complementarities mentioned in the table 2.

<table>
<thead>
<tr>
<th>Arabic language</th>
<th>Tree Adjoining Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>- accepts more than only one order of components in a sentence.</td>
<td>- is mildly contextual which makes it suitable to represent natural languages with flexible order as Bouillon declared in [15].</td>
</tr>
<tr>
<td>- presents an abundance of crossed and recursive structures</td>
<td>- possesses a strong combinatorial power that allows the presentation of the crossed structures without constraints on the sentences length.</td>
</tr>
<tr>
<td>- is grammatical and vocalic ambiguous and agglutinative.</td>
<td>- is rich by its heterogeneous feature structures, then it can be useful to disambiguate structures in an NLP application.</td>
</tr>
</tbody>
</table>

Consequently, we assume that TAG formalism admits a representative and combinatorial powers that contribute to a good representation of Arabic syntactic structures. Then, we consider that TAG formalism is theoretically...
very appropriate for Arabic language, especially in NLP applications.

4. ARABTAG: A GRAMMATICAL MODEL FOR ARABIC

4.1. MOTIVATIONS

The construction of ArabTAG is essentially motivated by an Arabic parsing problem’s study. In fact, we proposed a parsing approach [14] that consists in using machine learning models to incrementally construct complete syntactic trees for Arabic sentences. The parser should learn its knowledge from an Arabic Treebank. Then, according to the already mentioned complementarities between Arabic language and TAG formalism we choose this one to construct a grammatical model that will serve to build a TAG Arabic Treebank.

Furthermore, this Treebank is used to train a combinatory pattern recognition model to predict simultaneously the most appropriate elementary tree and the conjunction operation for each word in a target sentence. Then, each prediction is directly attached to the syntactic tree part constructed until the previous word. Consequently, syntactic tree will be constructed incrementally.

4.2. DESCRIPTION

Generally, the previous works on constructing an Arabic tree adjoining grammar try to extract derived trees and their correspondent elementary trees from already constructed Treebank. However, our work consists in transforming a collection of Arabic grammatical rules into a tree adjoining formalism without resorting to any textual resources. It consists in constructing a set of elementary trees that represents the basic Arabic syntactic structures and lexicon. In fact, we are based on some books that describe Arabic grammar as those of Kouloughli [17] and Aloulou [18] in order to collect grammatical rules. We concentrated on basic rules which present the essential structures of sentences. These components are the Subject, Object and Verb for a verbal sentence and the Subject and Predicate for a nominal sentence. Our model is baptized ArabTAG and it is coded using XML in order to profit from its ability to represent and use linguistic phenomena.

ArabTAG is a semi-lexicalized grammar that builds feature structures from a collection of grammatical enriched rules. In fact, we have associated to free context rules some syntactic and semantic information to help its transformation in a unification model. The following example gives an idea about the form of the grammatical rules which are used to be coded.

Example: NS [NP subject, nominative] (intermediate) [N part-of-speech, nominative](anchor])], PP [predicate, genetive](substitution)].

This example describes the composition of a simple Nominal Sentence (NS). It consists of two phrases: a Nominal Phrase (NP) followed by a Prepositional Phrase (PP). The NP has a category and a function (Subject) and consists of a single sub-level which is a Noun (N). This N is the anchor of the elementary tree that can be instantiated by its grammatical category and its form and should be nominative. The PP is the predicate of the sentence that may be genitive and can be substituted in further decompositions. The type of the operation which is used for each component is presented in parentheses.

So, each elementary tree is made of a hierarchy of nodes that represent the different components or sub-components. We have limited the maximum number of sub-levels in an elementary tree to three in order to reduce the complexity of the constructed structures.

The figure 3 presents a three-level elementary tree. It corresponds to the rule written above.

![Figure 3: Example of elementary tree in ArabTAG.](image)

The root and the intermediate nodes correspond to the sentence and/or the phrases. Each elementary tree contains a specific foot node which is the anchor. This node specifies if the elementary tree is lexicalized or not. Each phrasal structure contains a function and unification criteria. The anchor possesses also some instantiation criteria. Semi-lexicalization, instantiation and unification criteria are described in the following sub-sections.

4.3. CHARACTERISTICS

4.3.1. SEMI-LEXICALIZATION

ArabTAG is semi-lexicalized since it is composed of two parts: lexicalized trees and patterns trees. It makes use of words as well as parts-of-speech as anchors of the elementary trees. On the one hand, the lexicalized
trees are those corresponding to prepositions, modifiers, conjunctions, demonstratives, etc. On the other hand, the non-lexicalized part corresponds to patterns of generalized elementary trees representing verbs, nouns, adjectives or any kind of phrases. These elementary trees can be instantiated by information extracted from the sentence.

The choice of the semi-lexicalized TAG variant helps to reduce the probably important number of possible syntactic structures if we use a completely lexicalized version.

### 4.2.2. INFORMATION IN ARABTAG

The construction of ArabTAG is, especially, based on the information generated by an already developed morpho-syntactic analyzer [16]. This analyzer gives for each item (simple or agglutinative form) a set of morpho-syntactic information that includes its decomposition in proclitic/stem/enclitic, where each component is associated with its value and grammatical category. Furthermore, it gives the information about the stem as its lemma, its gender and number if the form is a noun or adjective, its pronoun and transitivity in case of verb. The parts-of-speech, whose number is 288, are refined ones. They include 65 for the clitics and 223 for stems. They are positional categories because they include information as the casual sign (the vocalic sign of the last character in the word). As an example of refined parts-of-speech we note the following: - نعت (accusative adjective).

In addition to morpho-syntactic information, ArabTAG is enriched by semantic ones. Each node in an elementary tree is affected by its role (verb, subject, object, conjunctive phrase, etc) in the whole elementary structure.

### 4.2.3. UNIFICATION AND INSTANTIATION CRITERIA

For representing words as elementary trees, syntactic information generated by the morpho-syntactic analyzer is divided into two types of criteria which are associated to the nodes: unification and instantiation criteria. The unification criteria allow the verification of the correspondence validity of syntactic criteria of the components in the same sentence. Whereas, the instantiation criteria help to specify the correspondent part-of-speech and word to the anchor of the elementary tree since non-lexicalized trees can be affected to several words and/or parts-of-speech. In addition, ArabTAG associates to the nodes of an elementary tree functions that specify their roles in the target fragment. The roles can belong to the following set: {subject, object, verb, complement, etc}.

The unification set of criteria is associated to each node in the elementary tree and is composed of the gender, number, casual sign, enclitic, proclitic, pronoun, transitivity and voice, whereas, the instantiation criteria are affected to each anchor node. They consist of the transitivity for the verbs and the words’ values and the grammatical categories of proclitic, stem and enclitic for all kinds of items.

### 4.2.4. ARABTAG COVERAGE

ArabTAG covers several Arabic regular structures. It consists of:

- two DTDs: the first to detail the structure of the lexicalized elementary trees and the second to specify the patterned elementary trees,
- a set of XML files for representing elementary trees organized by families
- and a dictionary which describes the correspondence between the instantiation criteria and the 288 finite grammatical categories that we use.

The number of the patterns trees is 110 which are decomposed as shown in the figure 4.

In this figure, we describe the partition of the patterns of elementary trees. They essentially correspond to the simple syntactic structures of the Arabic sentences (nominal and verbal), all classes of phrasal structures (NP (Nominal Phrase), VP (Verbal Phrase), PP (Prepositional Phrase)) as well as the different sub-classes of the phrasal structures. For instance, for an Arabic nominal phrase, different sub-cATEGORIES as: Adjectival NP (مركب نعتي (مركب نعتي)), Complement NP (مركب شبه أساسي (مركب أساسي)), Propositional NP (مركب إضافي (مركب إضافي)), exist. ArabTAG presents all these phrases. That’s why the percentage of the elementary trees representing nominal phrases is the most important.

![Figure 4: Current ArabTAG coverage (Patterns part).](image)

For the lexicalized part, we also represent 120 elementary trees that correspond to the particles as the prepositions (حرف الجر (حرف الجر)), conjunctions (الuxtapositions (الuxtapositions)), etc.

Moreover, our model presents different kinds of sentences: active, passive, interrogative, etc. In addition, ArabTAG is able, based on adjunction and substitution operations, to compose complex structures even the complex sentences. Then, it can present complex sentences as (S1) which is a nominal sentence that consists of more than one proposition. It is also able to present mono-word sentence (S2) which consists only of one agglutinative form.

**Examples:**
(S1) كل في ذلك يسحون (All are swimming in an orbit)
(S2) فزرونهم (Then, we visited them)
These two sentences are represented respectively by the two following figures (5 and 6). The first one presents the XML code that corresponds the sentence (S2), whereas, the second presents the different necessary steps to compose the derived tree of the sentence (S1).

In figure 5, we denote that the sentence (S2) is a verbal sentence with a hidden subject (we), whereas the enclitic "هم" (them) plays the role of the sentence complement.

For the figure 6, we have alleviated it by eliminating the information that corresponds to the unification and/or instantiation criteria. Our objective is to gain in figure legibility.

Figure 5: Elementary trees corresponding to the words of the sentence (S2).

4.4. ARABTAG LIMITS

We should mention that the current ArabTAG version cannot describe sentences with circumstantial components. In fact, these components are facultative, numerous and can take any position in the sentence. Then, describing them into our model will heavily increase the number of elementary trees and allow redundancies. We think, then, to introduce other elements to the standard TAG to support these frequently used facultative elements. The modifications of standard TAG will be especially in its operations since we need additional operations that allow inserting circumstances in any place when analyzing the sentence. ArabTAG can also be ameliorated by further syntactic structures and/or other syntactic and semantic information.

In addition, the inference of the agglutinative forms is not well considered in our syntactic structures. In fact, agglutinative forms can take any position in the sentence and have any role. These forms should be taken into consideration to ameliorate the coverage of the developed grammatical model.

Figure 6: Elementary trees corresponding to the words of the sentence (S1).

5. RELATED WORKS

Different researches have been made to develop natural language resources based on TAG formalism. These researches concern several languages: English [19], French [20], Vietnamese [21], etc. Almost of these models are lexicalized versions. Arabic language has also been presented within TAG formalism in the research of Habash [22]. His version was completely lexicalized since he has extracted it from an already analyzed Arabic Treebank. Then, the number of elementary trees generated is very high but they have not, necessarily a good syntactic coverage. In fact, their version is based on the information enclosed in the corpus. Also, the representation is redundant because they manipulated textual forms and not parts-of-speech. However, ArabTAG uses rich grammatical rules. Then, it is a generic grammatical model that can be used in several NLP applications. In fact, it constitutes a first step in a project that has as objective to parse Arabic texts, etc.

6. CONCLUSION AND PERSPECTIVES

ArabTAG allows marrying Arabic syntax with TAG formalism. Our choice of this formalism is not arbitrary but is justified by the complementarities between Arabic language and advantages of TAG. ArabTAG will constitute a primordial point for a classifier parser for Arabic texts described in [14]. In our current work, ArabTAG is used as basis to construct an Arabic Treebank. In this Treebank, each word in a sentence is affected by the appropriate elementary tree and each sentence is coupled to its derivation structure. The
constructed corpus constitutes the training data for our approach to parse Arabic texts. It is an incremental pattern recognition approach that associates to each word in a sentence the suitable elementary tree according to morpho-syntactic, contextual and compositional information of this target word.

REFERENCES


