Parsing Arabic Verbal Sentence Using Grammar Ontology

Khaled M. Almunirawi[1,a]   Rebhi S. Baraka[2,b]

[1]University College of Ability Development, Khan Younis, Gaza, Palestine
[2]Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine
[a]khalediue@gmail.com   [b]rbaraka@iugaza.edu.ps

Abstract—We build a model to parse the Arabic verbal sentence based on Arabic grammar ontology. The ontology conceptualizes the Arabic verbal sentence through the representation of grammar parsing classes, verb properties, and conjunction checking. By populating the ontology with verbal sentences and adding grammar rules, we form a verbal sentence knowledge base. The parsing model is supported by morphological analysis for sentence syntactic analysis and supported by Arabic synonyms extractor for deriving synonyms. We have implemented the model and have provided it with a user interface where the user can enter a sentence to be parsed and obtains the parsing results. The interface has the options to partially or totally add diacritics to the words of the sentence and it has the possibility to remove ambiguity by choosing the most appropriate analysis from lexicon results. To evaluate the model, we have selected a representative set of Arabic verbal sentences from Arabic grammar books that represent all the possibilities of a verbal sentence. We have performed several parsing tests on these sentences with and without diacritics. The results prove the ability of the model to parse the various forms of the verbal sentence. The accuracy increases when the sentence is diacriticized while avoiding free word order and following the Arabic verbal sentence general form.

Index Terms— Arabic Parsing, Arabic Word Net, Arabic Grammar Ontology, Morphological Analysis, Synonym Extraction.

I INTRODUCTION

Parsing is necessary for distinguishing the meaning and understanding the intentions of a sentence and rolling out any ambiguities. It is nearly impossible to misunderstand the meaning if the parsing is done correctly. Parsing Arabic sentence is a difficult task due to the relatively free word order of Arabic, besides the length of the sentence and the omission of diacritics (vowels) in written Arabic. Parsing Arabic sentence is the analysis of an input sentence into its linguistic parts in the form of a parsing tree with syntactic relations among them. This parsing usually contains semantic information [1]. Traditional sentence parsing is performed by understanding the exact (semantic) meaning of a sentence. When the ambiguity in the sentence is resolved, various possible interpretations are reduced and the sentence becomes more obvious. Parsing such a morphologically rich and free word order languages is a challenging task, requiring advanced techniques in NLP for processing the words and requiring the machine to understand the syntactic and semantic analysis of the words. It is nearly impossible to misunderstand the meaning, provided the parsing is done correctly. In case the ambiguity is resolved, the range of possible interpretations will be reduced and the sentence would be parsed.

An ontology, which a semantic web technique, is used to conceptualize language grammar [2]. It presents language entities such words, tenses, verbs, and phrases as ontological class hierarchy with properties, instances and grammar rules.

Then, the ontology can be used for various linguistic tasks such as lexical analysis and sentence parsing.

We develop an ontology-based model for parsing Arabic verbal sentence. The model receives a sentence as an input from the user, identifies if it is verbal by checking the first word of the sentence as it is a verb or not (except if the verb is preceded with verb preposition), starts defining the syntax of each word, and then performs parsing of these words depending on Arabic Grammar Knowledge base. The knowledge base contains mainly the Arabic Grammar Ontology besides five components: Grammar Rules, Verb Properties, Parsing Classes, Conjunction Checker and Word Parser. The ontology contains classes, objects and relations regarding classifying the Arabic sentence. The implemented model allows the user to interact with it and deals with the ontology through OWL-API to extract the parsing result for the given sentence. Based on this motivation, the research aims to:

- Build a grammar ontology for Arabic verbal sentence and represent its characteristics and parsing rules as relations and properties within the ontology. Then populate the ontology with appropriate parsing instances of verbal sentences to form a grammar knowledge base.

- Build the model that uses the ontology and the grammar knowledge base along with needed morphological and synonyms features for parsing a verbal sentence.

To achieve these two objectives, we follow the following research methodology:

1. **Arabic Parsing Domain Review:** Study and analyze current approaches related to Arabic language verbal sentence and its parsing alternatives. Also, study the domain of Arabic grammar and Arabic parsing rules to extract the elements of the ontology including objects, properties, relations, instances.

2. **Data Collection:** The dataset depends on formalized cases and rules applied to parsing Arabic sentence. It includes a collection of verbal sentences with different structures that covers all cases related to the verbal sentence parsing rules.

3. **Parsing Model Development:** The model is based on the knowledge base formed of the ontology, classes, instances, and parts for Conjunction Checker, Parsing Classes, Verb properties and Grammar Rules. The development process includes three main sub-phases:

   – Developing the ontology and the knowledge base, including extending the ontology with needed classes, properties and instances.

   – Building the model mainly including the grammar ontology for representing the Arabic grammar pertaining to the verbal sentence, the Morphological Analyzer for the identification of all syntactic possibilities for each word of the sentence and the Synonyms Extractor for giving the synonym of each word in the sentence.

   – Implementing a prototype of the model realizing the above components where a given verbal sentence can be completely parsed and its parsing results and returned. The implementation will depend on tools such as JAVA programming language, OWL and JENA APIs and other tools and APIs as needed.

4. **Model Evaluation:** Evaluate the accuracy of the model as a whole including the ontology by comparing the model results to the actual results based on a pre-parsed set of sentences representing all possibilities of verbal sentence. This is followed by briefly comparing the model to the XML semantic parser proposed by Al-Rabiah & Al-Salman [3] which is used to resolve parsing problems based on defined parsing-related factors.

The rest of this paper is organized as follows: Section II presents a review of related work. Section III presents the model for Arabic Verbal Sentence Parsing (AVSP). Section IV presents the implementation of the parsing model. Section V presents an evaluation of the model. Finally, Section VI concludes the paper.

II RELATED WORK

Parsing Arabic sentences is one of the challenging Arabic NLP tasks due to the distinct features of Arabic including its morphological richness. Many researchers tried to tackle Arabic parsing problems through various techniques, approaches and algorithms. These approaches can be divided as follows:

A Natural Language Processing Methods

Salloum et al. [4] focused on Arabic parsing issue and built a parser as a part of a machine translation system. The parser have problems regarding the ambiguity, since a huge amount of parse trees are generated due to the ambiguity issues. They defined the Lexical Functional Grammar (LFG) in the Arabic Context, LFG is a linguistic hypothesis of grammar which concerns the nature of the statement structure and generate realistic framework for natural language processing [5]. LFG distinguishes two levels of representation to each sentence of the language. This approach presented these two different formalisms: trees form or constituent structure (called c-structure) and functional structure representing grammatical functions like subject and object and the relation between them as attribute-value matrices.

Arabic language does not use diacritics in the presence of vowels in a sentence and therefore makes the language unclear. In case the ambiguity is resolved, the language would become clearer and the range of possible interpretations is reduced. Syntactic analysis system has been developed for Arabic language including three NLP elements: a lexicon, a morphological analyzer and a syntactic parser. Applying disambiguation approach, the morphology analyzer gives probable readings of the given Arabic word. This would become clearer by following the grammar rules that would be correctly parsed and resolve ambiguity.

Othman et al. [6] stated that since Arabic language does not use diacritics when writing vowels, thus makes the language unclear and slows down its development of Arabic NLP (ANLP). The way to resolve ambiguity will be influenced by certain linguistic constraints while parsing an Arabic sentence. They have developed a syntactic analyzer for the language with three NLP elements: a lexicon, a morphological analyzer and a syntactic parser. When applying a disambiguation approach (based on the parser and analyzer), the morphological analyzer returns all the possible values of the given Arabic word. This would be clearer by applying the grammar rules which ensure correct parsing and resolve ambiguity. They focused on limited categories of ambiguity, rather than examining the performance on small datasets.

Alqrainy et al. [7] developed a simple parser which aimed to check the correctness of a given Arabic sentence through building new Context-Free Grammar (CFG) which makes the Top-Down techniques much more valuable. Many experiments were conducted and the results revealed efficient outcomes while analyzing the nominal and verbal sentences. The system lacks resolving ambiguity issues which gives different meaning for the sentence being parsed.

B Transition Networks Methods

Augmented Transition Networks (ATN) can analyze the sentence structure, however, practically; it is complicated depending on the language features, morphological richness and complexity. A chart-parser approach [8] uses Modern Standard Arabic (MSA) sentences with syntactic constraints to minimize parsing ambiguity using some features in lexical semantics. They are mainly used utilized to solve the structure of ambiguous sentences. Prolog language has been used to implement the developed parser and has the capability to assure syntactic constraints. Recent researches [9] used chart-parser methodology depending on Context Free Grammar (CFG) to parse a simple Arabic sentence. Bataineh & Bataineh [10] employed a methodology that was studying and analyzing the grammar of Arabic language conforming to gender and number. This approach has accuracy of about 85%. The problem with ambiguity leads to bad-formed sentence which can be resolved semantically stressing the need to use semantic-based approaches to give better results.

C Machine Learning Methods

Machine Learning (ML) gives machines the ability to learn without being explicitly programmed. It was used in NLP and parsing. Combined methods [11] using Treebank-based parsers and automatic LFG f-structure annotation methodologies on parsing Arabic. The Arabic Annotation Algorithm (A3) exploits the functional annotations in the Penn Arabic Treebank (ATB) [12] to assign LFG f-structure equations to trees. For parsing, the researches modified Bikel's parser to learn ATB functional tags and merge phrasal categories with functional tags in the training data. Results are low compared with the domain expert. McCord & Cavalli-Sforza [13] used ATB to build Arabic slot grammar (ASG) parser based on same efforts done for European languages using slot grammar (SG).

Al-Emran et al. [14] developed a system that parses MSA sentences through the use of Treebank. It relies on the Arabic Statistical Parser to produce a model through training on the Penn Arabic Treebank and standard Arabic linguistic resource. Not different from previous models, ambiguity was a problem and avoiding it makes sentences meaningless.

D Semantic Approaches

Semantics, in linguistics, is a field concerned with the study of meaning at the levels of words, phrases, and sentences. Some parsers are developed for English language [15]. The Arabic language is a difficult language that may delay the expansion of the tools and applications for semantic web in Arabic [16]. It has many discriminations such as complex morphology, diacritics and short vowels. Al-Salman et al. [17] followed the semantic technology and attempted to reveal the word sense ambiguity, by building a semantic parser using a semantic analyzer. A dependency parsing approach [18] for Modern Standard Arabic (MSA) is used for verbal sentences using data-driven dependency parser. It utilizes the semantic information available in lexical Arabic VerbNet to complement the existing morpho-syntactic information already available in the data. This complementing information is encoded as an additional semantic feature for data driven parsing. They were able to build a dependency parser with accuracy of 71.5% Labeled Attachment Score (LAS), 77.5% Unlabeled Attachment Score (UAS), and 2% increasing in total accuracy compared to the case of not using semantic features.

Other semantic parsing approach such as [3] has collected traditional Arabic grammar rules and presented them into extended Backus Normal Form (EBNF) grammar to serve as a base for other Arabic NLP researches. They presented the architecture of an XML-based semantic parser. The parser was able to reduce the parsing ambiguity by using the lexical, syntactic and semantic feature structures of the unification grammar. It still returns inaccurate parsing results, and a percentage of ambiguity.

E Ontology-Based Approaches

An ontology-based parsing approach [19] analyzes the Turkish sentence and utilized the interlingua representation called Text Meaning Representation (TMR). It represents the relations between events and entities, the semantic properties and pragmatic properties. The core definitions of word senses (without any modifications) were taken from the lexicon. This approach bypassed the syntactic constrains and used the semantics provided from the ontology to achieve high accuracy of parsing results.

A Formalization of Arabic grammar and its entities using ontology is proposed by Elmalki [2]. The language phrases and grammar concepts are conceptualized into Arabic Grammar Ontology. It includes classes for the word, sentence, marks, gender, tense, count, verb and person, beside creating sub-classes. In addition, grammar relations are added as object properties in the ontology and define their domains and ranges from the classes. The ontology is published online and it is available for writing SPARQL queries or Semantic Web Rule Language (SWRL) rules directly.

Recent applied research on representing Arabic language into ontology is due to [20]. The Arabic Ontology is similar to WordNet. Each concept in the ontology (meaning of an Arabic term) is given a Unique Resource Identifier (URI), informally described by a gloss, and lexicalized by one or more of synonymous lemma terms. Some important individuals are included in the ontology, such as individual countries. The Arabic Ontology in the process of linking and integrating all Arabic lexicons. Each meaning in every lexicon will be linked (as much as possible) with a concept in the Arabic Ontology. Based on this, a large linguistic graph integrating Arabic semantics and morphology can be built.
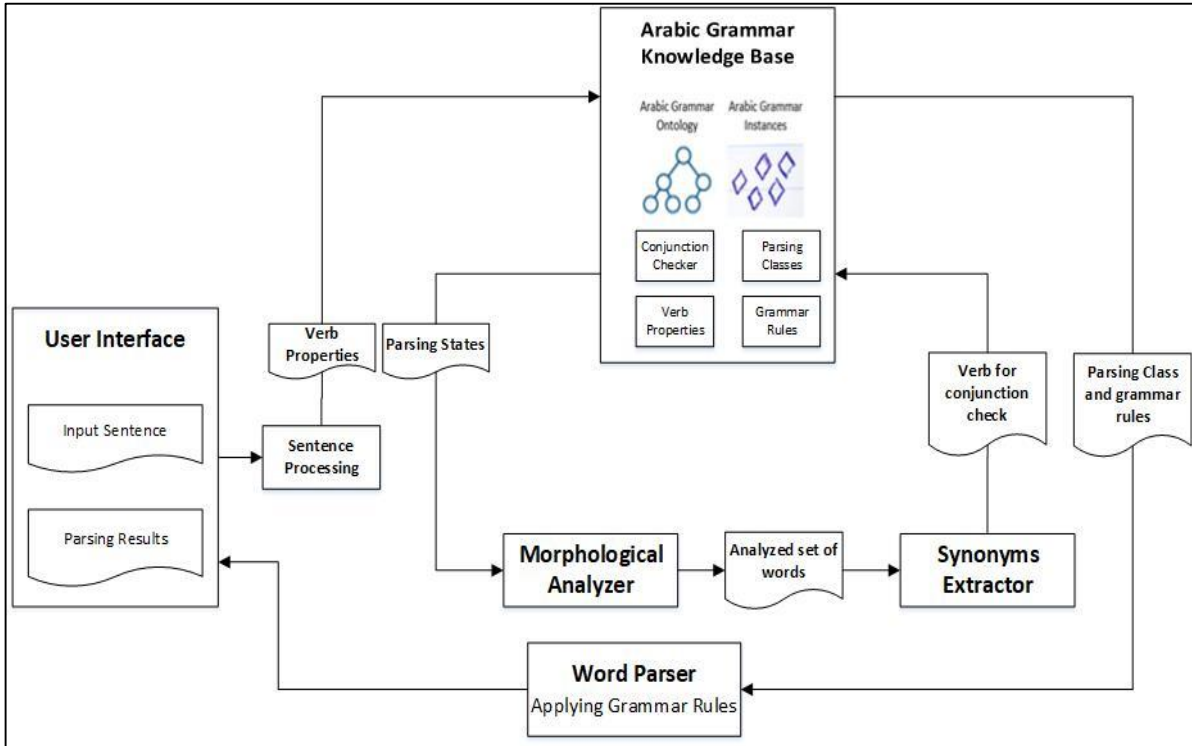
Fig. 1. Structure of the Arabic Verbal Sentence Parsing Model

III BUILDING THE ARABIC VERBAL SENTENCE PARSING (AVSP) MODEL

The AVSP model consists of five main components as shown in Fig. 1. They include the Arabic Grammar Knowledge Base, the Morphological Analyzer, the Synonym Extractor, the Word Parser, and the User Interface. Next, we describe each of them and then the flow of the model.

A Arabic Grammar Knowledge Base

The Arabic Grammar Knowledge Base consists of the Arabic Grammar Ontology and the Arabic Grammer Instances. The ontology is a conceptualization of the Arabic grammar phrases and concepts in the form of classes and related properties. The ontology is based on [2] and we have extended the it with additional classes and populated it with needed grammer instances that are related to the verbal sentence.

1. Classes: The Arabic Grammar Ontology contains nine classes, some of them contain sub-classes. Fig. 2 shows the classes and their subclasses.

2. Properties: The ontology contains two types of object properties: factor property and functional property. Both of them contain sub-object properties (see Fig. 2).

The ontology does not contain any data properties; since all data types are text. We explain the object factor properties related to our model, as follows:
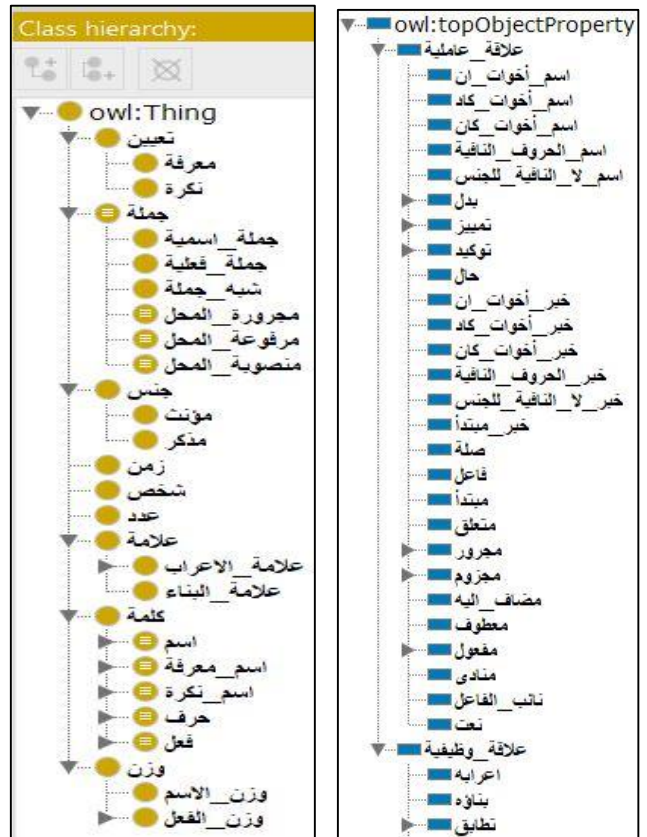

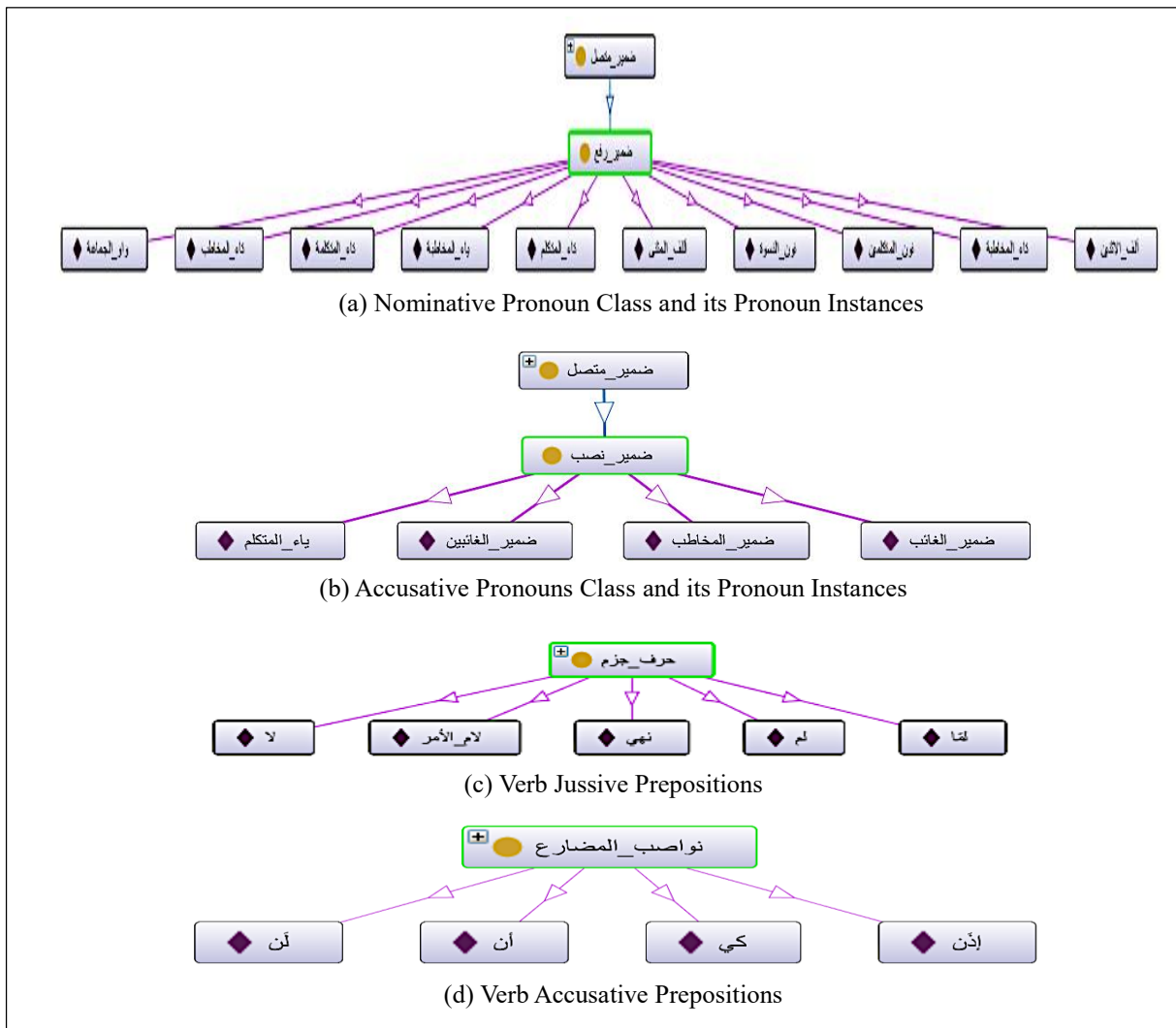**Fig. 2. Arabic Grammar Ontology & its Object Properties**

(a) Nominative Pronoun Class and its Pronoun Instances

(b) Accusative Pronouns Class and its Pronoun Instances

(c) Verb Jussive Prepositions

(d) Verb Accusative Prepositions

Fig. 3. Extending the Grammar Ontology with Pronouns, Prepositions and their Instances

1. Subject (فاعل): a relation that connects the active verb to a noun in one direction, with restricted domain and range as defined in the characteristics of the property.

2. Subject deputy (نائب الفاعل): a relation that connects the passive verb to a noun in one direction with the subject is omitted. It has restricted domain and range as defined in the characteristics of the property.

3. Object (مفعول به): a relation that connects the active verb to a noun in one direction. It has restricted domain and range as defined in the characteristics of the property.

To make the ontology more compatible with our model and to obtain more accurate results with better quality, we have extended the ontology capabilities by adding some classes, and instances. These additions include two classes with their instances to the pronoun class. These classes are called nominative pronoun (ضمير رفع) and accusative pronoun (ضمير نصب) as illustrated in Fig. 3.

The instances in the class nominative pronoun (ضمير رفع) are pronouns that are parsed in a nominative state as a subject. The instances in accusative pronoun (ضمير نصب) class are pronouns parsed in an accusative state as an object. In addition, we added verb prepositions as instances to their respective classes Verb Jussive and Verb Accusative Pronoun as shown in Fig. 3(c)-(d).

These additions of classes, instances, and properties enrich the ontology and form a grammatical knowledge base. They enable the model to deal with verbal sentences containing such properties. They are used define the parsing state of a word and recognize the related parsing cases through the ontology properties to return the accurate parsing result. The defined classes and instances would make better conceptualization for the personal pronouns and prepositions which improve the parsing result.

B The Morphological Analyzer

The Morphological Analyzer is mainly based and depends on Alkhalil Morpho Sys [21] to identify all possible syntactic features associated with a given word. For each possibility, the analyzer returns the voweled word (with diacritics), prefix, stem, type, pattern (weight), root, Part of Speech (POS) and suffix.

If the given word is diacritized, the possibilities are reduced to the minimum number, and the analyzer excludes the words that do not match the added diacritics. In our model, we use the prefix, type, POS and suffix to form the final parsing result.

C The Synonym Extractor

The Synonym Extractor is a lexical resource for Modern Standard Arabic (MSA) based on Arabic WordNet (AWN) and follows the design of Princeton WordNet (PWN). It extracts meanings and alternative synonyms of a given word in the context of a given sentence. For instance, the extracted synonyms of the word "ذهب" (went or gold) are as follows.

ذَهَبَ,غَادَرَ,نَزَحَ,رَحَلَ,تَرَكَ,خَرَجَ

ذَهَبَ,أَعْطَىَ,مُنِحَ

ذَهَبٌ,العُنْصُر التاسِع والسَبْعُون

D The Word Parser

The Word Parser applies the grammar rules based on the given parsing class returned from the knowledge base, particularly from the ontology, corresponding to the words of the sentence to be parsed. It also deals with prefixes and suffixes of a verb. Therefore, it determines the words' final parsing result. Inputs to the Word Parser, which are needed to produce the parsing results, are the word type, the parsing class, and the prefixes and suffixes of each word. The output of the Word Parser is the final parsing results of each word.

E The User Interface

The User Interface accepts a sentence to be parsed from the user, and displays the results after performing the needed parsing steps based on the Arabic Grammar Ontology and the grammar rules. The interface needs to be interactive and able to show the analysis of the sentence depending on the Morphological Analyzer. It also shows the synonyms of the analyzed word based on the Synonyms Extractor. Next, the model flow explains in details how the model works and how its components interact.

F Flow of the Model

The flow of the model starts when a verbal sentence is given to be parsed, then the Sentence Processing stage performs two operations:

1. Checks if the sentence is verbal, i.e., if the sentence starts with a verb, or starts with a preposition which precedes the verb such as accusative or jussive preposition. If the first word includes other probabilities than a verb, then the model takes the verb form, and eliminates all none verbal forms of the first word, which are subject, via-subject and object properties.

2. The model refers to the ontology to extract the object properties that have the verb as its range, and reduces the rest of the sentence's words syntactic results as follows:

    a. If the verb is active, then reduce the second word's syntactic results to the domain of the subject property.

    b. If the verb is passive, then reduce the second word's syntactic results to the domain of the via-subject property.

    c. If the verb is active and transitive, reduce the third word's syntactic results to the domain of the object property.

The results of AWN are passed to the tokenization process, in which each word is tokenized from relevant prefixes and suffixes. Theses prefixes and suffixes (if available) are passed to the ontology to extract its class. The results are either one or both of the following possibilities:

1. The prefixes of a verb are prepositions (accusative or jussive).

2. The suffixes of a verb are personal pronouns with their two types: nominative and accusative. The verb can be connected to two suffixes (both nominative and accusative) if the verb is active and transitive.

At this stage, the words are completely processed in terms of synonym extraction, conjunction checking, and their corresponding parsing classes.

Finally, these words are passed to the Word Parser which in turn applies the Arabic grammar parsing rules on each corresponding parsing class retuned from the grammar knowledge base, and pass the parsing results to the User Interface.

The Sentence Processing stage results in a set of reduced parsing states for each word in the sentence. These states are passed to the morphological analysis stage which uses AlKhalil Morpho Sys to analyze the words set with reduced parsing states morphologically. This set is passed to the Synonyms Extractor and Morphological Analyzer stage. In this stage, the model performs the synonym extraction and makes the syntactic analysis of analyzed words using AWN. The results are the syntactically-analyzed word with their synonyms, so that the model can have more information about the word's synonyms and alternate uses.

Based on the flow of the model and the roles of each component presented above, next we present the implementation of the model.
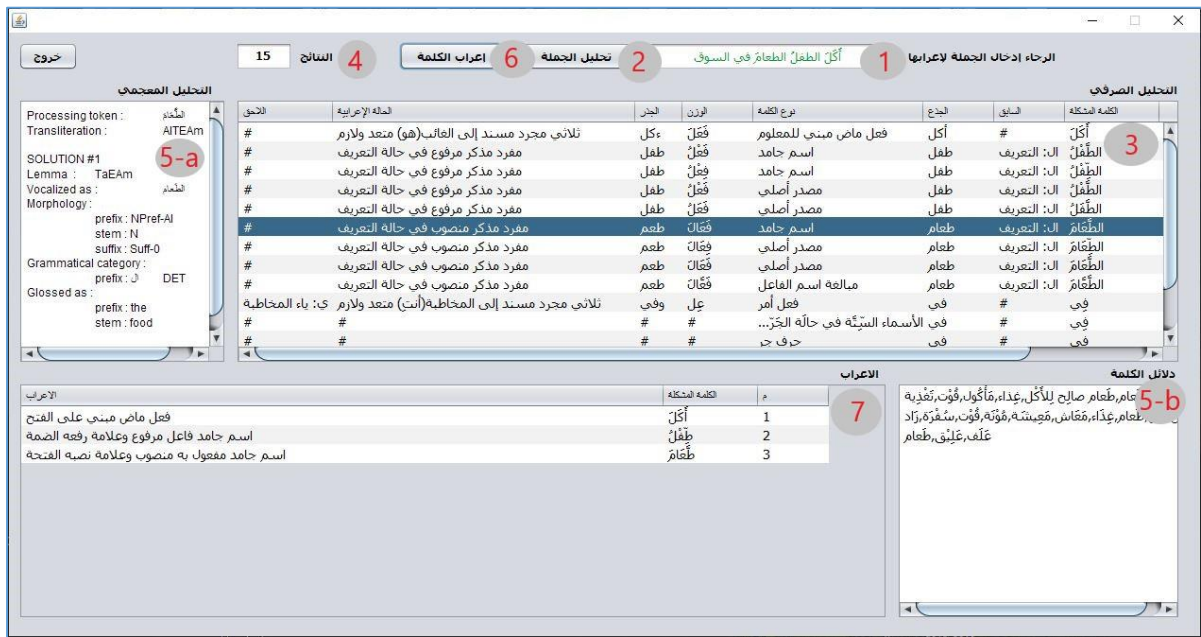
Fig. 4. The User Interface of the AVSP Model

IV IMPLEMENTING THE AVSP MODEL

In the implementation of the AVSP model, we have implemented each component with all of its required functionalities and interactions with the other components as defined in flow of the model. To achieve this, we have employed some tools and APIs and we have dealt with some issues imposed by the diversity of the used tools. The Arabic Grammar Ontology including our extension is encoded in OWL using Protégé ontology framework. As mentioned in the ontology component of the model, our extension includes classes related to accusative and nominative personal pronouns and prepositions as required by the model to deal with all different forms of the verbal sentence.

The ontology is, then, populated with the needed instances for each class such as all kinds of accusative and nominative pronouns as well as all kinds of prepositions related to the verbal sentence. These changes to the ontology required checking the consistency and the integrity of the ontology in terms of the class hierarchy, object (word) properties, type values, and instances belonging to their corresponding classes. OWL API is used to implement, manipulate and serialize the Arabic Grammar Ontology. It enabled us to access each class in the ontology with its instances, object (verb) properties, property values and Arabic grammar rules embedded into the ontology forming the grammar knowledge base (the ontology and its instances).

The Morphological Analyzer in the model is implemented based on AlKhalil Morph System rather than implementing our own or using other morphological systems. This is because we did not want to reinvent the wheel and it is not our purpose to implement another lexicon. Furthermore, it is up to date, open source, comprehensive where it can effectively deal with any word in a given sentence. It has a Java interface that simplified interacting with it in the model through some programming modifications and refactoring to capture the morphological analysis of the words and return them as part of the parsing model.

The Synonym Extractor is implemented based on AWN which is an open source and comprehensive in dealing with any word in a given sentence. Its Java interface simplified interfacing with the model through some programming modifications and capture the synonyms of the words and returns them as part of the parsing model.

The implementation of the model is represented by the user interface which interacts with the previously described components of the model. It accepts a sentence to be parsed from the user, and displays the results after performing the needed parsing steps. The interface shows the analysis of the sentence depending on Morphological Analyzer. For each word in the analysis, Synonyms Extractor shows the synonym set and translation. The interface refers to the ontology to examine the desired rules and to process the needed steps to return the parsing result, as described in the flow of the model. Fig. 4 shows the parts of the user interface labeled 1 to 7 as they illustrate the different interactions, and results of the model applied to parsing the sentence "أكل الطفل الطعام في السوق" (The child eat the food in the market).

This implementation is tested to be correct by ensuring that each component returns the expected output/result based on the model and the final parsing result corresponds to the parsing performed by a human expert. This is verified through a parsing example but more details are presented in the evaluation of the model (see Section V).
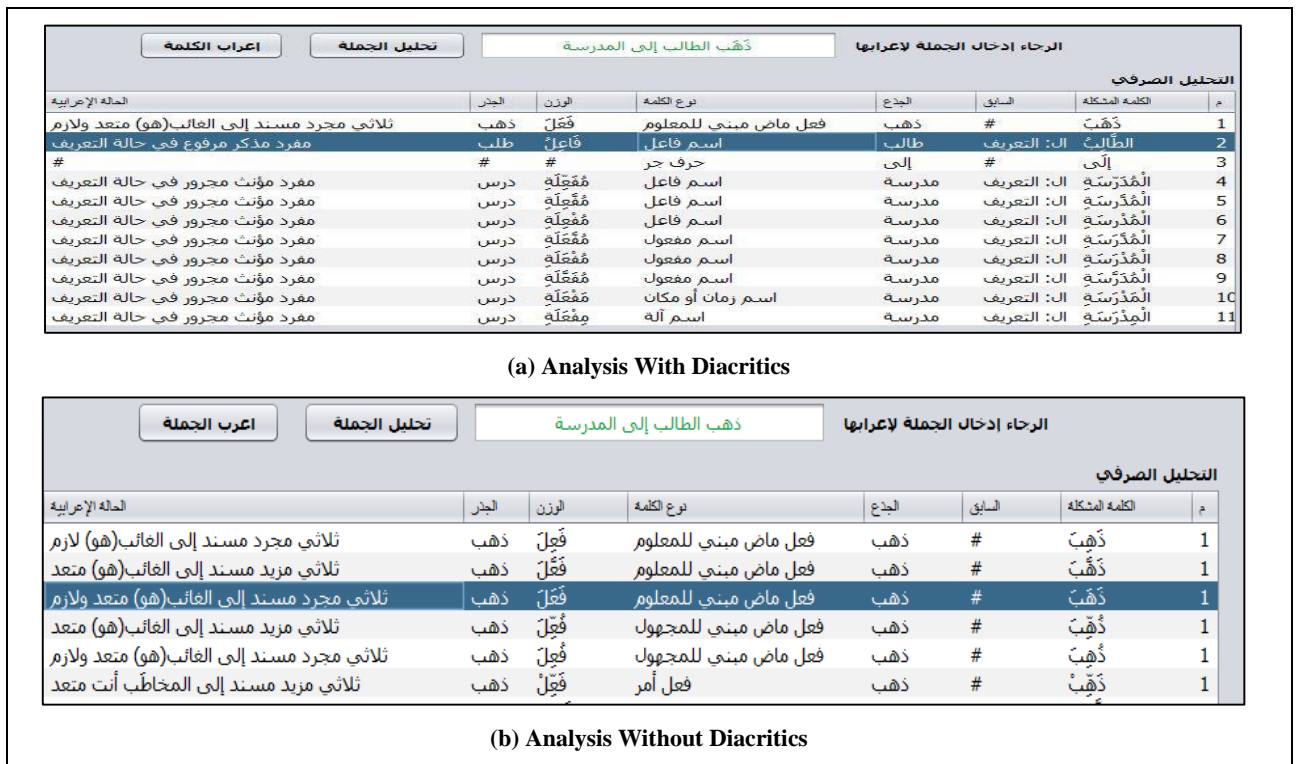
(a) Analysis With Diacritics

الرجاء إدخال الجملة لإعرابها — ذَهَب الطالب إلى المدرسة — | تحليل الجملة | إعراب الكلمة |

التحليل الصرفي

م	الكلمة المشكلة	السابق	الجذع	نوع الكلمة	الوزن	الجذر	الحالة الإعرابية
1	ذَهَبَ	#	ذهب	فعل ماض مبني للمعلوم	فَعَلَ	ذهب	ثلاثي مجرد مسند إلى الغائب(هو) متعد ولازم
2	الطّالبُ	ال: التعريف	طالب	اسم فاعل	فَاعِل	طلب	مفرد مذكر مرفوع في حالة التعريف
3	إلَى	#	إلى	حرف جر	#	#	#
4	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم فاعل	مُفَعِّلَة	درس	مفرد مؤنث مجرور في حالة التعريف
5	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم فاعل	مُفَعِّلَة	درس	مفرد مؤنث مجرور في حالة التعريف
6	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم فاعل	مُفَعَّلَة	درس	مفرد مؤنث مجرور في حالة التعريف
7	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم مفعول	مُفَعَّلَة	درس	مفرد مؤنث مجرور في حالة التعريف
8	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم مفعول	مُفَعَّلَة	درس	مفرد مؤنث مجرور في حالة التعريف
9	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم مفعول	مُفَعَّلَة	درس	مفرد مؤنث مجرور في حالة التعريف
10	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم زمان أو مكان	مَفْعَلَة	درس	مفرد مؤنث مجرور في حالة التعريف
11	الْمَدْرَسَةِ	ال: التعريف	مدرسة	اسم آلة	مِفْعَلَة	درس	مفرد مؤنث مجرور في حالة التعريف

(b) Analysis Without Diacritics

الرجاء إدخال الجملة لإعرابها — ذهب الطالب إلى المدرسة — | تحليل الجملة | اعرب الجملة |

التحليل الصرفي

م	الكلمة المشكلة	السابق	الجذع	نوع الكلمة	الوزن	الجذر	الحالة الإعرابية
1	ذَهَبَ	#	ذهب	فعل ماض مبني للمعلوم	فَعَلَ	ذهب	ثلاثي مجرد مسند إلى الغائب(هو) لازم
1	ذَهَّبَ	#	ذهب	فعل ماض مبني للمعلوم	فَعَّلَ	ذهب	ثلاثي مزيد مسند إلى الغائب(هو) متعد
1	ذَهَبَ	#	ذهب	فعل ماض مبني للمعلوم	فَعَلَ	ذهب	ثلاثي مجرد مسند إلى الغائب(هو) متعد ولازم
1	ذُهِّبَ	#	ذهب	فعل ماض مبني للمجهول	فُعِّلَ	ذهب	ثلاثي مزيد مسند إلى الغائب(هو) متعد
1	ذُهِبَ	#	ذهب	فعل ماض مبني للمجهول	فُعِلَ	ذهب	ثلاثي مجرد مسند إلى الغائب(هو) متعد ولازم
1	ذَهِّبْ	#	ذهب	فعل أمر	فَعِّلْ	ذهب	ثلاثي مزيد مسند إلى المخاطب أنت متعد

Fig. 5. Analysis of the Word "ذهب" (Went)

**Example: Parsing the sentence: "ذهب الطالب إلى المدرسة"
(The student went to school)**

1. The word "ذهب" holds two meanings: the past active transitive verb "went" and the noun "gold". The model uses the lexicon to reflect the sentence word's categories in Arabic and returns the syntactic and lexical feature of each category. The lexicon tells us that the word "ذهب" is a verb and returns the following syntactic feature: (tense: past, transitivity: no). In addition, the lexicon returns another type of the word "ذهب" which is a noun, meaning "gold" and this result is excluded since it is a noun and the model is limited to parse the Arabic verbal sentence only. Fig. 5 shows all available and returned options for the verb status of the word "ذهب" that have five forms with different meanings.

2. If the user has added a diacritic on a letter (the diacritics can be on any letter in the word, including the last letter) in the word, the lexicon returns less options. Then the parsing ambiguity decreases as shown in Figure 5(a). Here we add a diacritic on the word "ذهب" such that it became "ذَهَبَ" the returned morphological analysis is only one as shown. Next, the model connects to the ontology and checks the appropriate related object property to the active verb, which is (subject) "فاعل", and gets its domain and range for the subject property. The results will be used to limit the lexicon results for the words after the verb.

3. Since the verb is not transitive "لازم" then the next word must be the subject, and it must be nominative "مرفوع" as Arabic grammar rules states.

4. Therefore, the model limits all the successive word options to the nominative possibilities as shown in Figure 5(b) for the word "الطالب". If the verb was transitive, then the second word after the verb is considered accusative "منصوب". Therefore, limit its options to the accusative possibilities.

5. Now, the results of the domain and range for the object properties has decreased the lexicon options, and the parsing process starts when the user selects the desired word from the lexicon results. The selected word enters other text processing level to check the suffixes and prefixes for the word. This is because some of them affects the parsing status and the parsing mark. The prefixes are the allowed propositions, grammatically, to proceed the verb, including accusative "النصب" and jussive "الجزم" prepositions. The suffixes include checking two types of personal pronouns that are connected to the verb, i.e. nominative "الرفع" and accusative "النصب" pronouns. After checking the domains and ranges, besides the verb transitivity, the nominative suffix is parsed as a subject and the accusative suffix is parsed as an object.

6. The main objects in the sentence, i.e., the verb "ذهب" and subject "الطالب" are parsed, and the rest of the sentence is called semi-sentence, the proposition's parsing is fixed. the succeeding word is parsed as genitive noun.

V EVALUATING THE AVSP MODEL

The evaluation of the model aims to prove its correctness and show its accuracy in parsing a verbal sentence in its different forms. This requires the results of the Morphological Analyzer which is used in the model to return the true parsing for the word based on its properties and conjunctions within the sentence. The evaluation also requires the Synonyms Extractor to return synonyms of the selected word. The model's accuracy is evaluated using a set of representative sentences with different parsing rules as specified in Arabic grammar references. Parsing results are compared with the results obtained from these references. The representative sentences are classified into six categories presented as follows:

1. Verb-Subject-Object (VSO) Sentences.
2. Verb-Object-Subject (VOS) Sentences.
3. Sentences start with a verb having nominative conjunction.
4. Sentences start with a verb having accusative conjunction.
5. Sentences start with a verb having both nominative and accusative conjunctions.
6. Sentences start with a verb preceded by a verb proposition.

For verbal sentences shown in Table 1, the AVSP model checks the properties of the verb through Verb Properties and returns active/transitive verb.

TABLE 1

Verb-Subject-Object (VSO) Sentences

#	Sentence	Sentence with Diacritics
1	درس الطالب الدرس	دَرَس الطالبُ الدرسَ
2	درس الطالب المحاضرتين	دَرَس الطالب المحاضرتين
3	درس الطالبان المحاضرتين	دَرَس الطالبان المحاضرتين
4	غادر الركاب الحافلات	غادَر الركاب الحافلات
5	حضر المشرفون المناقشة	حَضَر المشرفون المناقشة
6	درست الطالبة الدرس	درسَتْ الطالبة الدرس
7	دعا الطالب الله	دعَا الطالب الله

The Conjunction Checker checks the verb and reports that it does not contain conjunctions, except in the sixth sentence where the conjunction does not affect the parsing result.

The Parsing Classes of the model checks the diacritics on the end of the words and returns nominative class for the second word and accusative class for the third word. Then Grammar Rules part applies Arabic grammar rules on the words. The verbs are checked for vowel/unvowel, and the words are checked for Single/Dual/Plural form and applies the rules.

Because the sentences follow the VSO form, the model returns accurate parsing results after applying the parsing rules, based on adding diacritics to the words. This is not expected if the sentences follow VOS form, as shown in Table 2. This sentence follow the VOS form, so the user must add diacritics to the words to be correctly parsed. The Parsing Classes part checks the diacritics on the words and adds the correct parsing class to the word depending on the diacritics added on it. If the

user did not add diacritics to the sentences, then the AVSP model returns false results.

TABLE 2

Verb-Object-Subject (VOS) Sentences

Sentence	Sentence with Diacritics
درس الدرس الطالب	درس الدرسَ الطالبُ

For sentences shown in Table 3, the Conjunction Checker part checks the verb and returns a conjunction that belongs to nominative pronouns type. So, it parses the pronouns as Subject "فاعل". The Parsing Classes part returns accusative class for the third word. Then the Grammar Rules part applies Arabic grammar rules on the words and parses the rest of the sentence and returns the parsing result.

The Parsing Classes part returns nominative class for the third word. Now, the Grammar Rules part applies Arabic grammar rules on the words and parses the rest of the sentence.

TABLE 3

Verb with Nominative Conjunction Sentences

#	Sentence	Sentence with Diacritics
1	درست الدرس	درسْتُ الدرس
2	درست الدرس	درسْتَ الدرس
3	درست الدرس	درسْتِ الدرس
4	درسا الدرس	درَسا الدرس
5	درسوا الدرس	درَسُوا الدرس
6	درسنا الدرس	درَسْنا الدرس

For the sentence shown in Table 4, the Conjunction Checker checks the verb and returns a conjunction which belongs to accusative pronouns. It parses the pronoun as Object "مفعول به".

TABLE 4

Verb with Accusative Conjunction Sentences

Sentence	Sentence with Diacritics
درسها المدرس	درّسهاً المدرس

For the sentence shown in Table 5, the Conjunction Checker checks the verb and reports that it contains two conjunctions.

TABLE 5

Verb with Nominative and Accusative Conjunction

Sentence	Sentence with Diacritics
ضربته	ضربتُّه

The first one belongs to nominative pronouns type, and the second one belongs to accusative pronouns type So it parses the first pronouns as Subject "فاعل", and parses the second pronouns as Object "مفعول به".

For sentences shown in Table 6, the Conjunction Checker checks the verb and finds that it is preceded with verb preposition that belongs to jussive type, then the AVSP model parses

the verb as Jussive "مجزوم" and continue parsing the sentence as explained in the above cases.

TABLE 6

Verb with Jussive Conjunction Sentences

#	Sentence	Sentence with Diacritics
1	لما يحضر الطالب المحاضرة	لمَّا يحضرْ الطالب المحاضرة
2	ليتقن العامل العمل	ليتقنْ العامل العمل

Next, we address the cases where the parsing results might be inaccurate or a sentence is falsely parsed:

- *Verbal Sentence preceded with a preposition:* The model checks if the sentence starts with a verb, or starts with a preposition that precedes the verb. The lexicon does not recognize some words, while other words are analyzed in many forms such as the word "أكل" (eat) where it can be analyzed as a verb or as a noun "كل" (all) preceded with "أ" which is a question preposition. Our model cannot recognize such sentences which is out of the model scope.

- *Word order:* Arabic sentence formalization tends to use the VSO form. It can use VOS which is uncommon but can be used in cases where the user wants to insure the importance of the object. If another form is used i.e., VOS the model will give false parsing results in parsing unless the user adds diacritics to the subject and object. If diacritics are added to the end of a word, the model can recognize the word's proper POS in the syntactic analysis stage.

- *Diacritics:* adding diacritics on the words, partially or totally, reduces the results of syntactic analysis and eliminates the ambiguity of the sentence. In most cases, the parsing process looks for the diacritics added to the last letter of the word. If the user did not add them, the parsing process returns many and inaccurate results. The word "درست" has three meanings: 'I studied', 'She studied' and 'He studied'. Without diacritics on the last letter, ambiguity appears resulting in these three cases. If the user added diacritics to the word to be "درستُ" or "درسَت" or "درستْ", this will clarify the meaning. The diacritics are necessary to remove ambiguity at least on the last letter of the word.

- *Word order and diacritics:* parsing a sentence like " درس الدرسَ الطالبُ" (The student studied the lesson) without diacritics results in true parsing for invalid sentence. The user means other semantics of the sentence where he wants to ensure that the lesson is studied by the student. If the user did not add diacritics to the sentence, the model returns regular parsing result. This is called "تقديم المفعول به" (forwarding the object) in Arabic. Id diacritics are added to the words, the model returns different parsing results, and the word 'الدرس' will be parsed as accusative object (مفعول به منصوب) and the word 'الطالب' will be parsed as nominative subject (فاعل مرفوع). This case has been taken into consideration when the model parses words entered by the user with diacritic.

Compared to the XML-based semantic parser of Al-Rabiah & Al-Salman [3], our AVSP model uses OWL (which is more expressive than XML) as a formalism to represent Arabic grammar rules while their parser uses XML only to represent the grammar rules. AVSP depends on the syntactic analysis using morphological analyzers to reduce the word's possibilities. Then, it interacts with the Arabic grammar ontology to perform parsing. Their parser reduces the parsing ambiguity using lexical, syntactic and semantic feature structures of the unification grammar. These feature structure makes their parser less modular and less flexible than AVSP. AVSP achieves more accurate parsing results than the XML-based parser (taking into consideration the conditions and parameters we mentioned in Section V) because it handles the free word order issues within the verbal sentence and processes words with diacritics which are not taken into consideration by the XML-based parser.

VI CONCLUSION

The main contribution of this research is building a parsing model for Arabic verbal sentence that uses a grammar ontology to conceptualize the Arabic grammar rules. The conceptualization includes presenting grammar phrases into classes and defines their domains and ranges based on strong understanding of the domain and relate them together via object properties. Additional classes and instances are defined in order to classify the word's type and then apply the correct parsing roles. The model is supported by morphological analysis and synonym extraction. The morphological analyzer performs the syntactic analysis of each word in the sentence and extracts all of its linguistic features such as a verb with its tense and type, noun and preposition. The synonym extractor extracts word's feature and its synonyms that help the model to remove the ambiguity by choosing the exact meaning of the word.

The implementation of the model has realized and connected those components and has provided a user interface to facilitate using the model. Various text processing and word tokenization steps have been performed between the components to handle and prepare the sentence to be applicable to the Arabic grammar parsing rules. To evaluate the accuracy of the parsing model, we have performed parsing on a number of representative verbal sentences. They include sentences with different verbal forms, different word order, with and without diacritics. The model demonstrated a high parsing accuracy for diacritisized sentences. Diacritics minimized the lexicon results, reduced ambiguity of the parsing model and hence increased its accuracy. Verbal sentences preceded by a preposition or having VOS order returned low accuracy or false results.

There are still several improvements that can be addressed in a future work. The ontology can be extended to include other classes, instances, lexical features, synonyms related to a word. This might simplify the parsing process making it

more direct, more semantic, and more accurate. Extend and add more features to the model so it will be able to handle verbal sentences preceded by a preposition or having VOS order, extended sentences with and without diacritics. In this case the model might cover other types of Arabic sentences like nominal sentences. The implementation of can cover these improvements including a web and mobile versions.

REFERENCES

[1] S. Green and C. Manning, "Better Arabic Parsing: Baselines, Evaluations, and Analysis," in The 23rd Int. Conference on Computational Linguistics, 2010.

[2] T. Elmalki, A Computer Ontology of Arabic Grammar: Toward a Modern Logical and Linguistic Description of the Arabic Language, vol. 1st edition, Al-Nabigha House for Publishing and Distribution, 2015.

[3] M. Al-Rabiah and A. Al-Salman, "An XML-Based Semantic Parser for Traditional Arabic," in *Proc. of the 4th Int. Universal Communication Symposium (IUCS)*, 2010.

[4] S. Salloum, M. Al-Emran and K. Shaalan, "A Survey of Lexical Functional Grammar in the Arabic Context," *Int. Journal of Comm. Network Technology,* vol. 4, no. 3, 2016.

[5] L. Tounsi and J. Van Genabith, "Arabic Parsing Using Grammar Transforms," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Malta, 2010.

[6] E. Othman, K. Shaalan and A. Rafea, "Towards Resolving Ambiguity in Understanding Arabic Sentence," in *Proceedings of the Int. Conference on Arabic Language Resources and Tools (NEMLAR 2004)*, Egypt, 2004.

[7] S. Alqrainy, H. Muaidi and M. Alkoffash, "Context-free Grammar Analysis for Arabic Sentences," *International Journal of Computer Applications,* vol. 53, no. 3, 2012.

[8] E. Othman, K. Shaalan and A. Rafea, "A Chart Parser for Analyzing Modern Standard Arabic Sentence," in *Proc. of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, 2003.

[9] A. Al-Taani, M. Msallam and S. Wedian, "A Top-Down Chart Parser for Analyzing Arabic Sentences," *Int. Arab Journal of Information Technology,* vol. 9, no. 2, 2012.

[10] B. Bataineh and E. Bataineh, "An Efficient Recursive Transition Network Parser for Arabic Language," in *Proceedings of the* World *Congress on Engineering*, 2009.

[11] L. Tounsi, M. Attia and Van Genabith, "Parsing Arabic Using Treebank-Based LFG Resources," in *Proceedings of the LFG09 Conference*, 2009.

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

[13] M. McCord and V. Cavalli-Sforza, "An Arabic Slot Grammar Parser," in *Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, 2007.

[14] M. Al-Emran, S. Zaza and K. Shaalan, "Parsing Modern Standard Arabic Using Treebank Resources," in *Proceedings of International Conference on Information and Communication Technology Research*, 2015.

[15] Y. Wang, J. Berant and P. Liang, "Building a Semantic Parser Overnight," in *Proceedings of the 53rd Annual Meeting of the* Association *for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015.

[16] A. Al-Zoghby, A. Ahmed and T. Hamza, "Arabic Semantic Web Applications: A Survey," *Jr. of Emerging Technologies in Web Intelligence,* vol. 5, no. 1, pp. 52-69, 2013.

[17] A. Al-Salman, Y. Al-Ohali and M. AlRabiah, "An Arabic Semantic Parser and Meaning Analyzer," *Egyptian Computer Science Journal,* vol. 28, no. 3, pp. 8-29, 2006.

[18] H. Elnajjar and R. Baraka, "Improving Dependency Parsing of Verbal Arabic Sentences Using Semantic Features," in *Procdings of the International Conference on Promising* Electronic *Technologies*, 2018.

[19] M. Temizsoy and I. Cicekli, "An Ontology-Based Approach to Parsing Turkish Sentences," in *Machine* Translation *and the Information Soup*, 1998.

[20] M. Jarrar, "Building a Formal Arabic Ontology," in Proc. *of the Experts Meeting on Arabic Ontologies and Semantic Networks*, Alecso, Arab League, Tunisia, 2011.

[21] M. Boudchiche, A. Mazroui, M. Bebah, A. Lakhouaja and A. Boudlal, "AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer," *Journal of King Saud University-Computer* and *Information Sciences,* vol. 29, no. 2, pp. 141-146, 2017.

Khaled M. Almunirawi is a computer engineer working as the head of IT department at the University College of Ability Development. Almunirawi is working as a university lecturer in the field of computer science and programming. He is also working as a Web Applications Developer since 2008 with an interest in Arabic language issues. He obtained his Master degree in Information Technology and BSc in Computer Engineering from the Islamic University of Gaza, Palestine in 2019 and 2007 respectively. His research interests are in the area of semantic web and NLP with application to Arabic language.

Rebhi S. Baraka is an associate professor of Computer Science and ex-dean of the Faculty of Information Technology, the Islamic University of Gaza, Palestine. He obtained his PhD degree in Computer Science from Johannes Kepler University, Austria in 2006. He obtained his MSc degree in Computer Science from De La Salle University, Philippines in 1996. He obtained his BSc degree in Electronics and Communications Engineering from the University of the East, Philippines in 1991. Dr. Baraka has more than 18 years of teaching and research. His research interests include Semantic Web and Ontology Engineering with focus on Arabic related issues, Parallel and Distributed Computing with focus on Cloud Computing, Big Data and Web Services. He is a referee in a number of scientific journals in the above areas. He has joined several research and capacity building projects related to scientific and social issues. He is also a member and organizer of several scientific, academic and social committees and events.