IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 3, June, 2023, pp. 611-619
Journal homepage: http://iieta.org/journals/isi

A Comprehensive Context-Free Grammar for the Arabic Language: Including Non-Fundamentalist Phrases

Yasser Yahiaoui[1], Azeddine Bourouis[1], Ayad Q. Al-Dujaili[2], Ahmed I. Abdulkareem[3], Olasumbo O. Agboola[4], Hilary I. Okagbue[4], Adedayo F. Adedotun[4*], Amjad J. Humaidi[3], Ogbu F. Imaga[4], Onuche G. Odekina[4]

[1] Department of Mathematics and Informatics, University Center Salhi Ahmed Naama, (Ctr. Univ. Naama), P.O. Box 66, Naama 45000, Algeria
[2] Electrical Engineering Technical College, Middle Technical University, Baghdad 10001, Iraq
[3] Control and Systems Engineering Department, University of Technology, Baghdad 10066, Iraq
[4] Department of Mathematics, Covenant University, Ota 112212, Ogun State, Nigeria

Corresponding Author Email: adedayo.adedotun@covenantuniversity.edu.ng

https://doi.org/10.18280/isi.280309

Received: 3 November 2022
Accepted: 12 March 2023

Keywords:
Arabic language processing, formal grammar, Subsumption Hierarchical Attribute, syntactic roles, CGF

ABSTRACT

Dixon's assertion regarding the idiosyncratic nature of natural languages initiates an investigation into the unique characteristics of the Arabic language. Contrary to Dixon's viewpoint, some scholars suggest the presence of regularity within Arabic, attributable to its extensive array of syntactic rules and formulations. Yet, the copious volume of terminal vocabulary in Arabic poses significant challenges to grammar development. While annotations have offered partial solutions, they bring forth additional difficulties due to the necessity of retrieving data from the annotated corpora. To mitigate these issues, an innovative study was executed that utilized an annotated taxonomy of syntactic roles, coupled with an examination of both fundamentalist and non-fundamentalist phrases. A codification method was applied to a knowledge base employing the Subsumption Hierarchical Attribute (SHA), enabling the integration of Arabic word classes based on their potential syntactic roles. The SHA acts as an annotation method for deriving a grammar class 02, where classes are coded as terminal vocabulary. Its primary objectives are twofold: to moderate the complexity of the parsing system and to automate the generation of over 1490 distinct possible sentence structures. The study culminated in the development of a novel context-free grammar (CFG) for Arabic, broadening the horizons of language processing techniques.

1. INTRODUCTION

The Arabic language, with its extensive array of rules governing sentence structure, presents a unique opportunity for the development of models capable of generating a myriad of Arabic sentences. Nevertheless, the task is not without its challenges. The rich diversity of lexical elements within the language can contribute to ambiguity, while the presence of non-fundamentalist sentence forms and the absence of vocalization in the majority of written Arabic resources found online only add to the complexity of developing a comprehensive generator model for structured Arabic phrases.

Past research in this field has led to the creation of several grammars, most of which concentrate on regular sentence structures with reasonable success. Yet, the principal challenge remains: to construct a generic generative model utilizing formal grammars that can effectively extract sentences from text, irrespective of their regularity or complexity.

In this academic sphere, some studies have proposed that parsers can produce derivation trees to outline sentence structure [1]. However, these trees often fail to indicate the corresponding syntactic function, a significant issue when dealing with the intricacies of Arabic language rules that demand more sophisticated and context-specific parsing approaches. Other research has focused primarily on verifying the grammatical correctness of given sentences [2], often limiting their experimentation to small datasets of well-structured phrases. Some studies have employed annotation techniques with corpora such as The Penn Arabic TreeBank [3] to generate a Probabilistic Free Context Grammar (PFCG) [4].

The approach proposed in this paper commences with the creation of a hierarchy of potential syntactic roles for Arabic words, a process previously demonstrated in our earlier works. These classes are then codified using the Subsumption Hierarchical Attribute (SHA) representation method, whereby every word is substituted with a vector representing its class. This method transforms the variants of sentence forms into a countable set, thereby facilitating the design of the appropriate grammar class 2.

This proposed grammar is primarily based on classes in lieu of instances of Arabic words, thereby reducing the vocabulary to 83 class names, or 83 representative vectors. This simplification has a positive impact on semantic processing by reducing the complexity of any proposed model and streamlining the syntactic semantic as part of distributed semantics.

The constructed CFG's capacity for generation and recognition enables researchers to enhance effectiveness and efficiency in a wide range of application domains, including automated translations, summarizations, and classifications. The initial grammar was able to handle a substantial number of fundamentalist sentences with no anastrophe (i.e., changes in the order of grammatical elements). The current grammar extends this ability to non-fundamentalist sentences and has been designed to incorporate all existing anastrophe mechanisms. This allows for the processing of all possible sentence types (over 1490) and facilitates evaluation on a larger corpus of sentences, as will be explored in future works. The objective is to determine the precise quantity of sentence forms that can be generated by the proposed grammar class 2.

The research methodology employed in this study aims to scrutinize and analyze both fundamental and non-fundamental structures of well-formed Arabic phrases. The primary objective of this study is the development and evaluation of a context-free grammar (CFG) capable of recognizing and parsing a diverse range of Arabic phrases, regardless of their complexity or structure. This CFG is intended to serve as the groundwork for more sophisticated natural language processing techniques and applications.

2. THEORETICAL FUNDING

The first thing that must be known at this stage is what is the Anastrophe in Arabic sentences and how many types of these mechanisms exist, then identify them to know the elements of their similarity and differentiation.

The second thing is to know what is the influence of the grammatical structures (i.e., well-formed phrases) on the semantics processing. And due to some citations found in linguistics studies that affirm that “syntax is dealing with same questions that the semantic is dealing with …”. It is seen as from the point of view of the existence of its validity and its abstention, and as for the semantics, it is seen as from the point of view of the existence of its validity and its abstention, and as for the science of meanings, it seen as from the point of view of the statement of existence, which outweighs some of them over others [5].

2.1 Fundamentalists phrases

Known in Arabic by ‘ELDJOMAL ELOSSOLYA’ is the sentences having a regular form in which every word has a fundamental position, for example in the verbal sentences the verb take the first position in the sentence and is followed by the subject and this last is followed by 0 to 3 complements relatively to the kind of verb used even it is transitive or non-transitive.

In Arabic language, there are two primary types of sentences: nominal sentences and verbal sentences. Nominal sentences typically have their elements arranged in a predicate-first structure, with the subject following afterwards. Verbal sentences, on the other hand, begin with a verb, which is then followed by the subject and, if applicable, any complements. In general, the order of a verbal sentence is determined by the number of complements it contains - one, two, or three - and follows a specific pattern accordingly.

2.2 Non-fundamentalists phrases

Called in Arabic ‘ELDJOUMLA GHIR OUSSOLYA’, found when the positions of structural elements are not on the fundamentalists order.

The sentence necessarily represents a sequence of grammatical units or morphemes, but not every sequence of these morphemes is necessarily a meaningful sentence [6]. But in the Arabic language some of these sequences can be meaningful and reflect secondary ideas and semantics aspect different from the ordinary fundamentalist formulations of the same morphemes.

3. RELATED WORKS

The existing body of literature showcases a multitude of studies utilizing class 2 grammars and parsers to achieve high analysis accuracy. However, the methods and objectives of analysis in each study vary significantly. Often, the pre-processing of text is a prerequisite to ensure optimal automated processing. Several prominent parsers and their unique features are delineated below:

Belguith et al. [7] offers an innovative method for handling Arabic text. This system converts raw text (TXT) into a list of segmented sentences, represented in an XML file format. It utilizes a technique termed 'contextual exploration', which hinges on the detection of stop words and conjunctions as markers of sentence boundaries [8].

Khalifa et al. [9] introduced an approach of text segmentation based on automated learning. This method relies heavily on the syntactic implications of connecting operators, or the rhetorical functions of the Arabic conjunction "و" (waw). Notably, six semantic variations of this conjunction exist [8].

Habash and Rambow [10] employ the use of annotated corpora and tagging techniques to represent Arabic syntax. Their grammar construction is based on comparative analysis between trees derived from elementary forms in corpora of trees. For this purpose, the Penn Arabic TreeBank (PATB) is extensively utilized [8].

Salloum et al. [11] explored the implementation of the Lexical Functional Grammar (LFG) in the context of the Arabic language. The LFG represents a different genre of grammars, placing emphasis on the lexical aspect [8].

Khoufi et al. [12] developed an approach based on the concept of probabilistic CFGs, resulting in an Arabic parser that employs an induced grammar, or PCFG. The procedure first induces the PCFG with the Arabic TreeBank, then implements a parser that aligns forms defined in the corpus with each input phrase [12].

4. BACKGROUND

Here is presented the most important notions that enable us to explain how is built our idea and contribution and what is the result and where it can be used to give maximum of enhancement for the Arabic language (ALP) processing field.

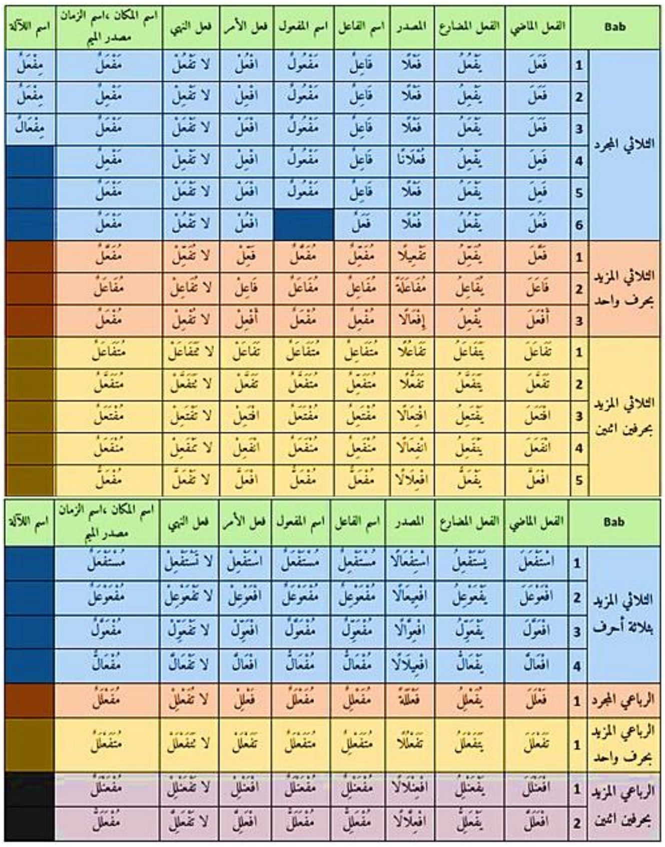
4.1 The word in Arabic

The point of view that yield to syntactic mastering of the word in ALP is the classification. Because of the verry large number of Arabic lexemes; some of the Arabic studies estimate that Arabic language contains 12,302,912 words [13] and this can be explained by the large degree of derivability of roots and the multiplicity of words having the same meaning, for example, see Table 1.

Table 1. Arabic words with similar meaning

Word in English	Word in Arabic	Synonyms or similar	Number
Sword	السيف	الحسام، المهند، الصليت، الصمصام، المشمل، البارقة الزالق....	100
Love	الحب	الهوى، الود، الهيام، العشق، الصبابة،.......	11
Camel	الجمل	البعير،الهيشور،الطبز ، ضائل، الناقة، دهامج،.....	1000

In addition to the large number of Arabic words, another factor contributing to the complexity of the language is the wide range of grammatical derivations that can be generated from a single word root. These derivations can result in a significant number of additional words with distinct meanings that are related to the original root. A visual example of this phenomenon can be seen in Figure 1.



اسم الآلة	اسم المكان ،اسم الزمان مصدر الميم	فعل النهي	فعل الأمر	اسم المفعول	اسم الفاعل	المصدر	الفعل المضارع	الفعل الماضي	Bab	
مِفْعَلٌ	مَفْعَلٌ	لا تَفْعُلْ	افْعُلْ	مَفْعُولٌ	فَاعِلٌ	فَعْلًا	يَفْعُلُ	فَعَلَ	1	الثلاثي المجرد
مِفْعَلٌ	مَفْعِلٌ	لا تَفْعِلْ	افْعِلْ	مَفْعُولٌ	فَاعِلٌ	فَعْلًا	يَفْعِلُ	فَعَلَ	2	
مِفْعَالٌ	مَفْعَلٌ	لا تَفْعَلْ	افْعَلْ	مَفْعُولٌ	فَاعِلٌ	فَعْلًا	يَفْعَلُ	فَعَلَ	3	
	مَفْعِلٌ	لا تَفْعِلْ	افْعِلْ	مَفْعُولٌ	فَاعِلٌ	فَعَلَانًا	يَفْعِلُ	فَعِلَ	4	
	مَفْعَلٌ	لا تَفْعَلْ	افْعَلْ	مَفْعُولٌ	فَاعِلٌ	فَعْلًا	يَفْعَلُ	فَعِلَ	5	
	مَفْعَلٌ	لا تَفْعُلْ	افْعُلْ		فَعِلٌ	فَعْلًا	يَفْعُلُ	فَعُلَ	6	
	مُفَعَّلٌ	لا تُفَعِّلْ	فَعِّلْ	مُفَعَّلٌ	مُفَعِّلٌ	تَفْعِيلًا	يُفَعِّلُ	فَعَّلَ	1	الثلاثي المزيد بحرف واحد
	مُفَاعَلٌ	لا تُفَاعِلْ	فَاعِلْ	مُفَاعَلٌ	مُفَاعِلٌ	مُفَاعَلَةً	يُفَاعِلُ	فَاعَلَ	2	
	مُفْعَلٌ	لا تُفْعِلْ	أَفْعِلْ	مُفْعَلٌ	مُفْعِلٌ	إِفْعَالًا	يُفْعِلُ	أَفْعَلَ	3	
	مُتَفَاعَلٌ	لا تَتَفَاعَلْ	تَفَاعَلْ	مُتَفَاعَلٌ	مُتَفَاعِلٌ	تَفَاعُلًا	يَتَفَاعَلُ	تَفَاعَلَ	1	الثلاثي المزيد بحرفين اثنين
	مُتَفَعَّلٌ	لا تَتَفَعَّلْ	تَفَعَّلْ	مُتَفَعَّلٌ	مُتَفَعِّلٌ	تَفَعُّلًا	يَتَفَعَّلُ	تَفَعَّلَ	2	
	مُفْتَعَلٌ	لا تَفْتَعِلْ	افْتَعِلْ	مُفْتَعَلٌ	مُفْتَعِلٌ	افْتِعَالًا	يَفْتَعِلُ	افْتَعَلَ	3	
	مُنْفَعَلٌ	لا تَنْفَعِلْ	انْفَعِلْ	مُنْفَعَلٌ	مُنْفَعِلٌ	انْفِعَالًا	يَنْفَعِلُ	انْفَعَلَ	4	
	مُفْعَلٌّ	لا تَفْعَلَّ	افْعَلَّ	مُفْعَلٌّ	مُفْعَلٌّ	افْعِلَالًا	يَفْعَلُّ	افْعَلَّ	5	
اسم الآلة	اسم المكان ،اسم الزمان مصدر الميم	فعل النهي	فعل الأمر	اسم المفعول	اسم الفاعل	المصدر	الفعل المضارع	الفعل الماضي	Bab	
	مُسْتَفْعَلٌ	لا تَسْتَفْعِلْ	اسْتَفْعِلْ	مُسْتَفْعَلٌ	مُسْتَفْعِلٌ	اسْتِفْعَالًا	يَسْتَفْعِلُ	اسْتَفْعَلَ	1	الثلاثي المزيد بثلاثة أحرف
	مُفْعَوْعَلٌ	لا تَفْعَوْعِلْ	افْعَوْعِلْ	مُفْعَوْعَلٌ	مُفْعَوْعِلٌ	افْعِيعَالًا	يَفْعَوْعِلُ	افْعَوْعَلَ	2	
	مُفْعَوَّلٌ	لا تَفْعَوِّلْ	افْعَوِّلْ	مُفْعَوَّلٌ	مُفْعَوِّلٌ	افْعِوَّالًا	يَفْعَوِّلُ	افْعَوَّلَ	3	
	مُفْعَالٌّ	لا تَفْعَالَّ	افْعَالَّ	مُفْعَالٌّ	مُفْعَالٌّ	افْعِيلَالًا	يَفْعَالُّ	افْعَالَّ	4	
	مُفَعْلَلٌ	لا تُفَعْلِلْ	فَعْلِلْ	مُفَعْلَلٌ	مُفَعْلِلٌ	فَعْلَلَةً	يُفَعْلِلُ	فَعْلَلَ	1	الرباعي المجرد
	مُتَفَعْلَلٌ	لا تَتَفَعْلَلْ	تَفَعْلَلْ	مُتَفَعْلَلٌ	مُتَفَعْلِلٌ	تَفَعْلُلًا	يَتَفَعْلَلُ	تَفَعْلَلَ	1	الرباعي المزيد بحرف واحد
	مُفْعَنْلَلٌ	لا تَفْعَنْلِلْ	افْعَنْلِلْ	مُفْعَنْلَلٌ	مُفْعَنْلِلٌ	افْعِنْلَالًا	يَفْعَنْلِلُ	افْعَنْلَلَ	1	الرباعي المزيد بحرفين اثنين
	مُفْعَلَلٌّ	لا تَفْعَلِلَّ	افْعَلِلَّ	مُفْعَلَلٌّ	مُفْعَلِلٌّ	افْعِلَّالًا	يَفْعَلِلُّ	افْعَلَلَّ	2	

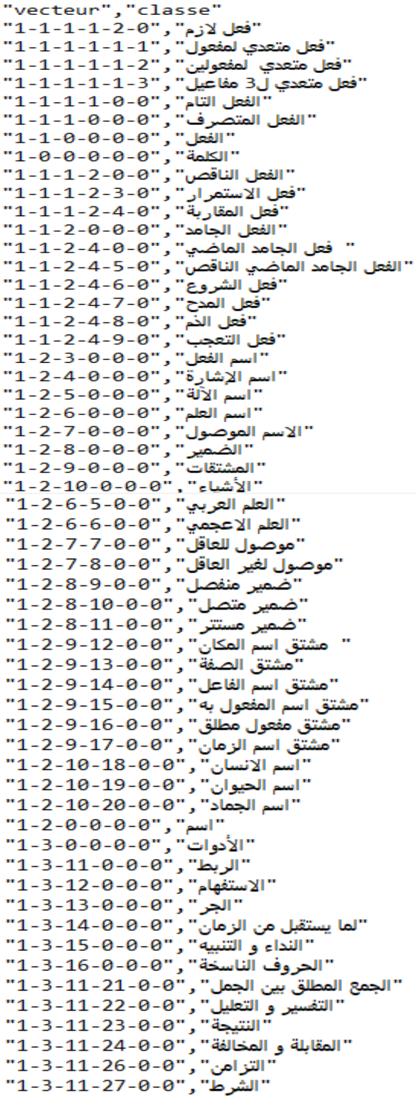
Figure 1. Root ف ع ل and their possible inflexions

From this point of view, a way must be found to reach a controllable number of entrances so that it becomes possible to deal with grammatical syntactic rules as a countable group, and then become representable in a model that allows it to be processed for several purposes such as classifying sentences and their parsing, and this is what prompts a return to a classification. The word in terms of the grammatical role that it can play within the sentence, and we have already done that in a research paper published in 2013 entitle, A meta description logics knowledge base for Arabic language processing [7]. This work was followed by work in order to find organized rules applicable to all possible cases of fundamentalist sentences. Published in 2019 at the Scientific Conference IHSED 2019 intitled, Proposed representation for Arabic text segmentation based on syntactic roles categorization [14]. In this work, the new is the integration of the non-fundamental sentences on the latest proposed grammar class 02.

4.2 Word classes

This classification is driven by the need to determine the grammatical category of a word and its expected position in the sentence. It is based on the three main classes of words: nouns, verbs, and particles, which are further divided into subcategories in a hierarchical form using a dependency graph based on the subsumption relationship.

The application of Subsumption Hierarchical Attribute [15] yields to create a list of SHA vectors to replace every category by her representation as in Figure 2 in which is presented 82 classes with their representative SHA by this way is defined the terminal vocabulary of the grammar.



```
"vecteur","classe"
"1-1-1-1-2-0","فعل لازم"
"1-1-1-1-1-1","فعل متعدي لمفعول"
"1-1-1-1-1-2","فعل متعدي  لمفعولين"
"1-1-1-1-1-3","فعل متعدي لـ3 مفاعيل"
"1-1-1-1-0-0","الفعل التام"
"1-1-1-0-0-0","الفعل المتصرف"
"1-1-0-0-0-0","الفعل"
"1-0-0-0-0-0","الكلمة"
"1-1-1-2-0-0","الفعل الناقص"
"1-1-1-2-3-0","فعل الاستمرار"
"1-1-1-2-4-0","فعل المقاربة"
"1-1-2-0-0-0","الفعل الجامد"
"1-1-2-4-0-0","فعل الجامد الماضي  "
"1-1-2-4-5-0","الفعل الجامد الماضي الناقص"
"1-1-2-4-6-0","فعل الشروع"
"1-1-2-4-7-0","فعل المدح"
"1-1-2-4-8-0","فعل الذم"
"1-1-2-4-9-0","فعل التعجب"
"1-2-3-0-0-0","اسم الفعل"
"1-2-4-0-0-0","اسم الإشارة"
"1-2-5-0-0-0","اسم الآلة"
"1-2-6-0-0-0","اسم العلم"
"1-2-7-0-0-0","الاسم الموصول"
"1-2-8-0-0-0","الضمير"
"1-2-9-0-0-0","المشتقات"
"1-2-10-0-0-0","الأشياء"
 "1-2-6-5-0-0","العلم العربي"
 "1-2-6-6-0-0","العلم الاعجمي"
 "1-2-7-7-0-0","موصول للعاقل"
 "1-2-7-8-0-0","موصول لغير العاقل"
 "1-2-8-9-0-0","ضمير منفصل"
 "1-2-8-10-0-0","ضمير متصل"
 "1-2-8-11-0-0","ضمير مستتر"
 "1-2-9-12-0-0","مشتق اسم المكان  "
 "1-2-9-13-0-0","مشتق الصفة"
 "1-2-9-14-0-0","مشتق اسم الفاعل"
 "1-2-9-15-0-0","مشتق اسم المفعول به"
 "1-2-9-16-0-0","مشتق مفعول مطلق"
 "1-2-9-17-0-0","مشتق اسم الزمان"
 "1-2-10-18-0-0","اسم الانسان"
 "1-2-10-19-0-0","اسم الحيوان"
 "1-2-10-20-0-0","اسم الجماد"
 "1-2-0-0-0-0","اسم"
 "1-3-0-0-0-0","الأدوات"
 "1-3-11-0-0-0","الربط"
 "1-3-12-0-0-0","الاستفهام"
 "1-3-13-0-0-0","الجر"
 "1-3-14-0-0-0","لما يستقبل من الزمان"
 "1-3-15-0-0-0","النداء و التنبيه"
 "1-3-16-0-0-0","الحروف الناسخة"
 "1-3-11-21-0-0","الجمع المطلق بين الجمل"
 "1-3-11-22-0-0","التفسير و التعليل"
 "1-3-11-23-0-0","النتيجة"
 "1-3-11-24-0-0","المقابلة و المخالفة"
 "1-3-11-26-0-0","التزامن"
 "1-3-11-27-0-0","الشرط"
```

Figure 2. List of SHA vectors representing the syntactic roles classes

4.3 Syntactic formulation

For the construction of context free grammar (CFG) is necessary to master knowledges on sentences in Arabic language. These lasts are classified like fundamentalist's sentences and not fundamentalist's sentences. When formulation is regular it concerns the first kind [13]. The fundamental phrases are essentially composed by a predicate and predicated part (Figure 3). Everyone can be composed of one or more words. That yields to have complex predicated part or complex predicate part is in following Tables 2 and 3. Combining these cases leads to find all possible forms of fundamental verbal phrases. And by the same way nominal sentences possible figuration (see next table). The fundamentalist phrases represent de regular form of the Arabic sentences. It’s why the non-fundamentalist are limited in some known cases for the nominal and verbal phrases theses one will be described in the contribution section.

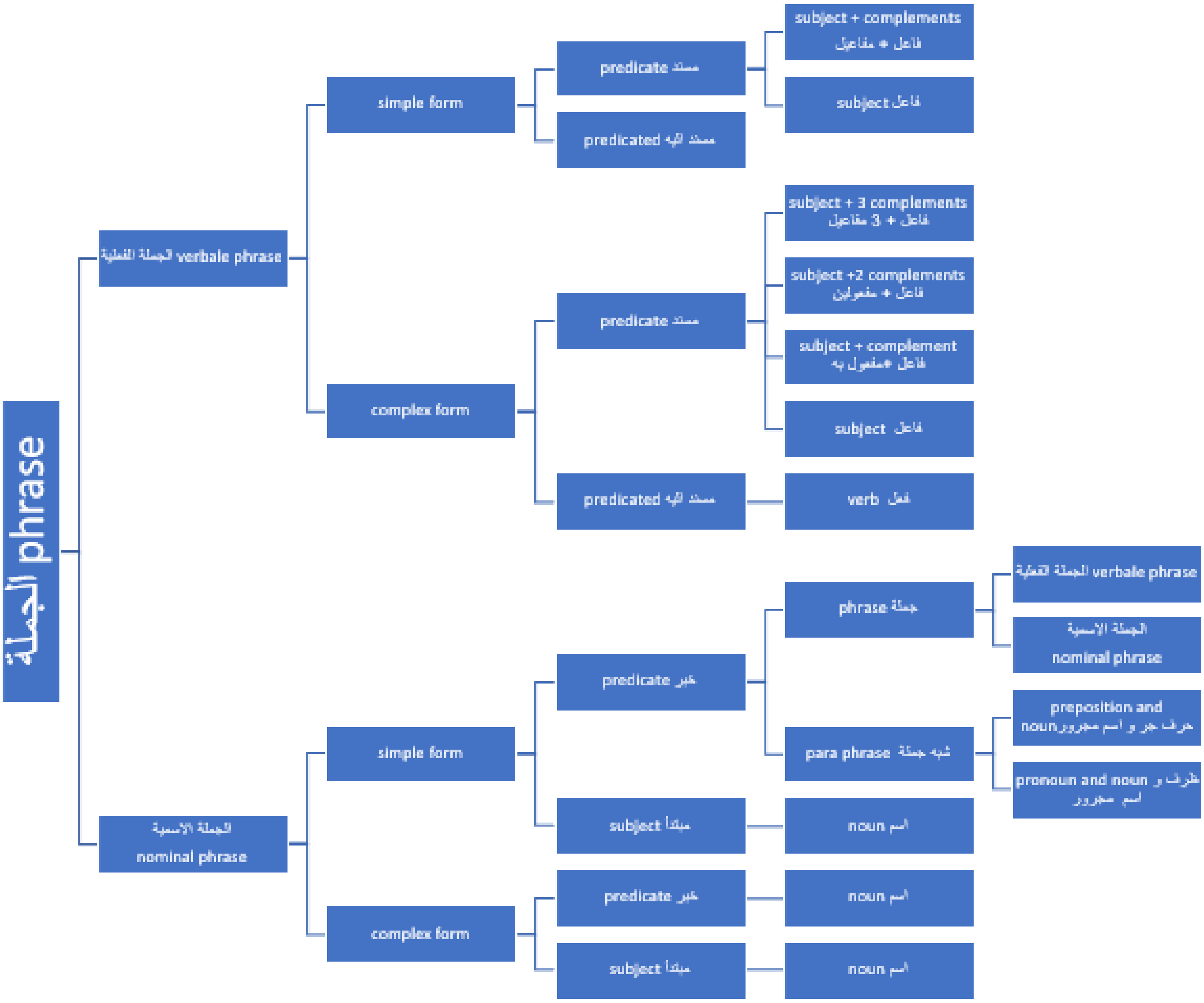


Figure 3. Possible structures of fundamental phrases

Table 2. Components and syntax of verbal fundamental phrase

Phrase الجملة		
The predicated		**The predicate**
Complement	***Subject***	
Explicit noun	Explicit noun	Verb in active voice
Connected pronoun	Hidden pronoun	
Separated pronoun	Connected pronoun	
An authentic source	Linked noun	
Phrase	Signal name	
Complement replacing subject		Verb with unknown subject (Passive voice)
noun Connected pronoun Separated pronoun An authentic source hidden pronoun		

Table 3. Components and syntax of nominal fundamental phrases

Phrase الجملة								
The predicate			***The predicated***					
Phrase	Para-phrase	noun	noun	Signal noun	Linked noun	conditional proposition	Separated pronoun	Authentic source

5. MOTIVATION

The use of formal grammar is a very interesting technique that yields to solve some important problematics of the naturel language processing. Like in the following [10]:

1. In order to overcome the challenge of ambiguity in Arabic

sentences, various linguistic constraints must be taken into account. To address this issue, heuristics can be employed to ensure proper formation, and disambiguation methods can be based on parsing analysis. Morphological analysis can also be a useful tool, as it provides all possible interpretations of a given Arabic word. By adhering to proper grammar rules, a correct parse can be achieved, which helps in resolving ambiguity [16].

2. Definite Clause Grammar [10]: the Arabic Language is different and particular comparing to most of languages [17] and it is declared that Arabic language; the use of the grammar in Arabic is presented only in descriptive form. Various efforts are attempted to formalize the Arabic sentences [18-20] such as LFG model [21], dependency grammar and functional grammar. Nevertheless, this issue is under discussions and researchers now days. A formal description of Arabic syntax has been developed by El-Shishiny [17] in Definite Clause Grammar form. It has been developed in Prolog and implemented in syntactic analyser. For the aim of better understanding of the diversity of syntactic structures the non-terminal alphabet is hold. This influence on the possibility of defining Clause grammar to understand the context.

3. Parsing Arabic Sentences: generic Arabic parser system development is difficult and very hard task comparing to other language. the nature of the Arabic morphology which is very rich. And the roughness of their syntax. Several efforts have been resulted in different models and different methodologies for the development of Arabic parsing [10]. As argued before, developing parsers for Arabic was not done on a large scale. Many scholars in Arabic NLP systems focused on morphological analysis [22-24] discussed the problem of implementing a morphological analyser for inflected Arabic vocabulary [25] hybridity of morphological analysing and the syntactic analysing systems makes efficient chart parser system creates. This parser is capable to satisfy the syntactic constrains of the Arabic phrases. And decrease the parsing inexactness using features related to the words or vocabulary of Arabic [10]. From that it is clear that FG give important help to the parsing system and have a good impact on the semantic aspect. this yields to improve the quality and performance of NLP system.

4. There are many challenges and issues that can be addressed through the use of formal analysis and generative grammar, making it highly desirable to have a regular or Para-regular formulation to enhance the ability of automated systems to comprehend natural language.

The context free grammars are well mastered formulas that can give good result for the Arabic language processing but it is faced by the delicacies of the large lexical richness. For that is proposed in this work to replace every word (lexeme) by his possible syntactic role. This yields to reduce the terminal vocabulary to 83 class of role and enable to define an efficient CFG generator and analyser of grammatical correct Arabic sentences. Note that this work is done in early works but without taking count of non-fundamentalists structures which is treated here.

This present CFG has as strength some novel aspects that are resumed in:

- ✓ The limitation of the terminal vocabulary where is replaced the large lexical corpus by small corpus of 83 classes, that enable a definition of finite state machine to deal with the syntactic analysing after studying the existing possible sentences structures. Because of the countability of the set of possible roles and by consequence the correct sentences structures.
- ✓ The integration of the processing of the non-fundamentalist phrases because which will be discussed in following sections.
- ✓ The limitation of rules number existing in the presented CFG which gives large number of generated structures.

The CFG will be presented in contribution section with detail and the modifications occurred to take account about the non-fundamentalists' phrases.

6. CONTRIBUTION

First thing to describe in this section is the cases in where predicate and subjects take new orders deferent than the ordinary syntactic formulation. These cases are:

1. Nominal sentence with subject is unknown noun and the predicate is a paraphrase.

2. Nominal sentence with a subject in form of متصلا ضمير أً Connected pronoun that reflect on a part of the predicate.

3. Nominal sentence with a predicate in form of question mark.

4. Nominal sentence with in the event that the predicate is confined to the subject (exception case no one… except).

5. Verbal sentence in which a subject in form of متصلاضمير ً أ Connected pronoun that reflect on complement.

6. Verbal sentence in which the subject is confined to the verb (exception case no one… except).

7. Verbal sentence in which the subject is an Explicit noun and the complement ضمير أ متصلاً Connected pronoun.

These cases can be considered in the CFG generator of fundamentalist sentences created in our early works. At this stage, it's necessary to look for integration of these non-fundamentalist cases for reuse of the CFG. The study must be on existence of ambiguous cases caused of intersections of firsts. Her is the grammar [13].

6.1 Context-free grammar

First the grammar has as terminal vocabulary the 82 SHA vectors and non-terminal vocabulary which are defined in generative rules and for beginning axiom '*Djoumla'*.

G[Vn, Vt, Djoumla, R]

Vn: the non-terminal vocabulary.

Vt: terminal vocabulary that contains {(1-1-1-1-1-1), (1-1-1-1-1-2), (1-1-1-1-1-3), (1-1-1-1-2-1), …… (1-3-20-0-0-0), …}

R: *Djoumla* → الجملة

- الجملـــــــــــــــــــــة← جملـــــة_فعلية / جملـــــة_إسمية / جملة الاستفهام / جملة الشرط / النسخ / النداء
- جملـــــة_فعلية ←فعل_متعدي1 فاعل مفعول / فعل_متعدي2 فاعل مفعول مفعول/ فعل_متعدي 3 فاعل مفعول مفعول مفعول/فعل_لازم فاعل/ فعل_لازم فاعل شبه_جملة/فعل_مبني_للمجهول_متعدي1 نائب_فاعل/فعل_مبني_للمجهول_متعدي2 نائب_فاعل مفعول / فعل_مبني_للمجهول_متعدي3 نائب_فاعل مفعول مفعول / فعل_أمر فاعل/ فعل_أمر فاعل مفعول / فعل_أمر شبه_جملة / فعل_المدح_الذم فاعل المخصوص _المدح_الذم/ الفعل_الناقص المبتدأ الخبر / فعل_الاستمرار المبتدأ الخبر /فعل_المقاربة المبتدأ الخبر / فعل_الرجاء المبتدأ الخبر/ فعل_الشروع المبتدأ الخبر
- فاعـــــــــــــــــل←اسم_ًظاهرا/ ضمير_متصل / اسم_موصول/ اسم_إشارة / ً

ضمير_مستتر/ مصدر__ مؤوّل

- مفعـــــــــــــــول←ً اسم_ظاهرا/ ضمير_متصل / ضِمُ يرـــم ْنفَ الجملـــــــــــــة صِ لُ َ / مصدر_مؤَّوّل/الجملـــــــــــــة
- نائب_فاعـــــل←ً اسم_ظاهرا/ ضمير_متصل / ضِمير_منفَصل
- ◌َمصدر_مؤوّل←حرف__مصدري فعل_مضارع
- المخصوص _المدح_الذم←اسم ظاهر /اسم__موصول
- جملـة_إسمية←المبتدأ الخبر / شبه جملة إسم/ أداة استفهام مبتدأT
- جملة_الاستفهام ←حرف_ الاستفهام جملـة_إسمية /حرف_ الاستفهام جملــــة_فعلية
- جملة الشرط← أداة الشرط جملة فعلية جملة فعلية
- المبتـــــــــــــــدأ←ً اسم_ظاهرا/ مصدرا_مؤولاً/ اسم_إشارة/ اسم__موصول / ضميراً_منفصل
- الخبـــــــــــــــــر←اسم_ظاهرا/ شبه_جملة/ جملــــة_إسمية / جملــــة_فعلية
- شبه_جملة←حرف__جرمجرور/ ظرف مجرور
- مجرور←َ اسم_ظاهرا / مصدر__مؤولً / ضمير__متصل
- النسخ←حرف__ناسخ المبتدأالخبر / حرف__ناسخ شبه__جملة المبتدأ

النداء←حرف__نداء اسم_ظاهرا / حرف__نداء اسم_ظاهرا اسم_ظاهرا

To consider the non-fundamentalist cases; it is occurred some modifications on the proposed grammar [14] that yields to generation of all non-fundamentalist sentences. This gives new derivations and more generative rules indicated in the following (modifications are mentioned in red and underlined):

- الجملــــــــــــــــة← جملــــة_فعلية / جملــــة_إسمية / جملة الاستفهام / جملة الشرط / النسخ / النداء/ أداة النفي Z
- **جملــــة_فعلية** ←فعل_متعديR1 / فعل__متعدي2 فاعل مفعول مفعول/ فعل__متعدي 3 فاعل مفعول مفعول مفعول/فعل__لازم فاعل/ فعل__لازم فاعل شبه__جملة/فعل__مبني__للمجهول__متعدي1 نائب__فاعل/فعل__مبني__للمجهول__متعدي2 نائب__فاعل مفعول / فعل__مبني__للمجهول__متعدي3 نائب__فاعل مفعول مفعول / فعل__أمر فاعل/ فعل__أمر فاعل مفعول / فعل__أمر شبه__جملة / فعل__المدح__الذم فاعل المخصوص __المدح__الذم/ الفعل__الناقص المبتدأ الخبر / فعل__الاستمرار المبتدأ الخبر /فعل__المقاربة المبتدأ الخبر / فعل__الرجاء المبتدأ الخبر/ فعل__الشروع المبتدأ الخبر
- جملـة_إسمية←المبتدأ الخبر / النسخ / النداء/شبه جملة إسم/ T أداة استفهام مبتدأ
- **جملة الشرط**← أداة الشرط جملة فعلية جملة فعلية
- **Z** ← خبر أداة الاستثناء مبتدأ / فعل متعدي 1 مفعول أداة استثناء فاعل / جملة
- **T** ← ضمير متصل / Ɛ
- **R** ← فاعل مفعول / مفعول فاعل ضمير متصل / فعل ضمير متصل فاعل

6.2 Parsing system implementation based on CFG

In this section is presented how can be implemented a parser based on the context free grammar. The following schema shows the components of the analysis system and their disposition (Figure 4).

The parsing implementation is based on two major parts of processing and it focus on annotated corpus in which is found the SHA codification. These parts are:

1. The annotated corpus in which is found the codification of the possible class of syntactic role for every word.

2. The codification process that replaces every word by his SHA of syntactic role class.

3. The parser that uses the CFG to check the correctness of the phrase and gives the syntactic tree. Using nouns of syntactic roles classes names.

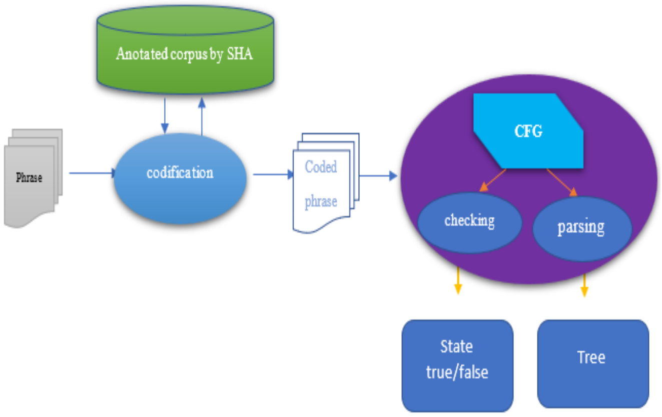


Figure 4. Architecture of the parsing system

6.3 Capabilities of the CFG

To estimate the minimal number of structures that can be generated by a free-context grammar, we can disregard the possibility of creating complex structures within other complex structures. This means that we only consider composition once by synthesis, where one element of the sentence is itself another sentence. By doing so, we can obtain a statistic that does not take this case into account for more than one time. The resulting minimal generation can be seen in Table 4.

Table 4. Result of estimation of possible complex structures

	For once complex struct use at more	For twice complex struct use at more	Total
Verbal sentences	85	85+(5*30) +(5*85) =660	745
Nominal sentences	30	30+(6*30) +(6*85) = 720	750
Total	115	1380	1495

The more complex sentences are very rare to exist in naturel language, even written text or spoken language. It’s why the obtained result is a successful attempt to find a syntactical analyser able to generate a major part of possible sentence structures, to make de parsing and to find the reel role of every morpheme of the phrase and to give a possibility of recognition of a part of her semantic. This can be the power of this proposed CFG.

To know how to get these numbers, it is sufficient to trace all grammatical structures that can be generated. That result from the application of all production rules found in the CFG. This allows it to be counted with one at more execution of recursive calls, based on a Terminal vocabulary consisting of the vectors SHA representing the classes of the word, as previously explained. Figure 5 gives a summarization of possible derivations.

6.4 Use of the CFG

The first use of this grammar can be text segmentation and sentences extraction, this goal when being achieved can have a good impact on lot of kind of processing and especially on automated translation.

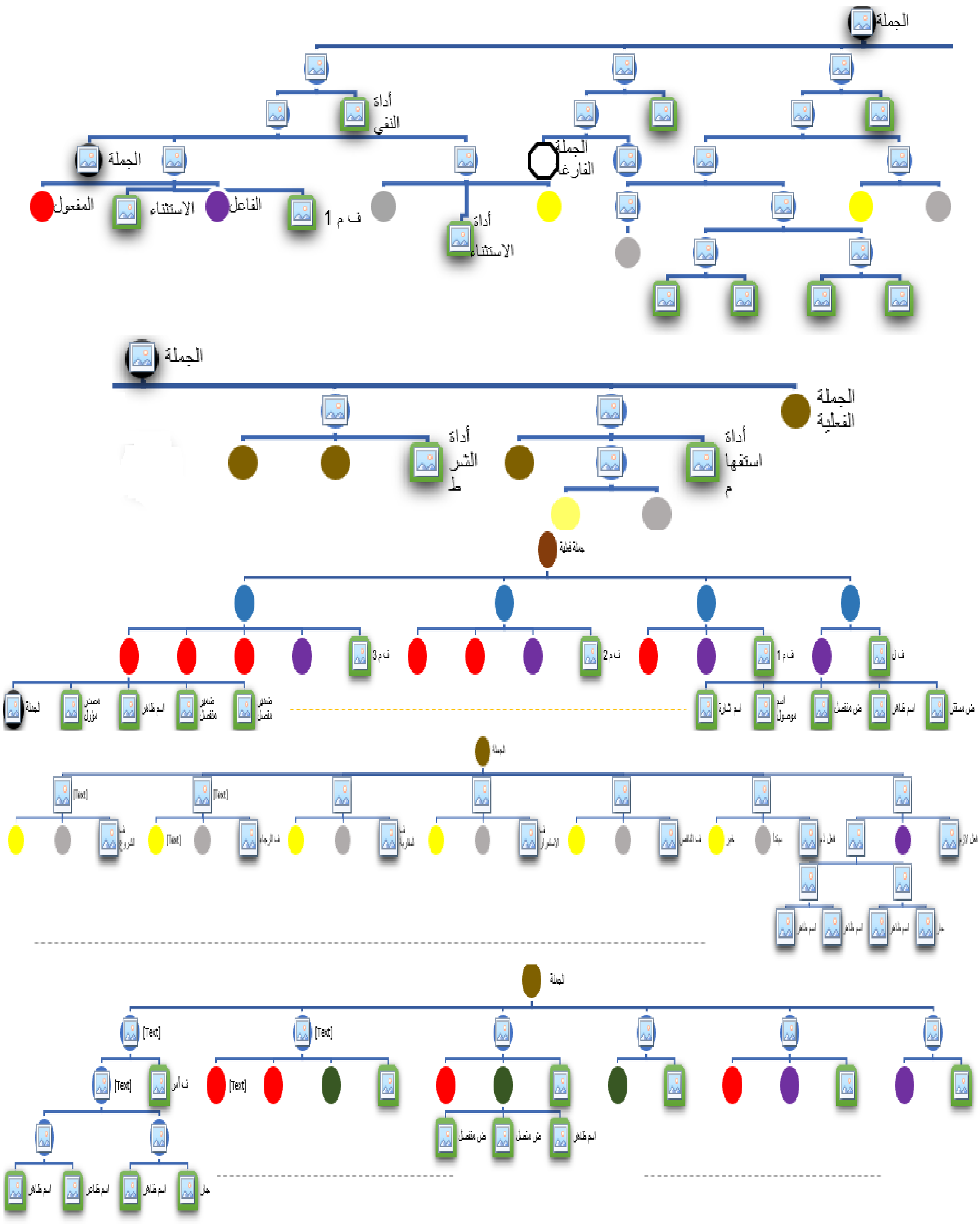


Figure 5. Derivation of syntactic structures derived by the CFG

Parsing, can be performed by the use of this grammar as model of recognition of well-structured sentences and recognition of her parts, and to give good result for the explication of these phrases called in Arabic ‘EL Iêrab’.

The possibility of segmentation of text to phrases can yield to the summarisation, the classification, and disambiguation and so others kind of application.

7. CONCLUSION

This work addresses an important need in the field of Arabic language processing by shifting focus from the entire lexical system to a mini-lexicon containing only 83 entries. This approach allowed for the creation of a context-free grammar that is capable of generating a significant number of correct Arabic syntactic structures.

These eighty-three entries represent the syntactic categories funded on a taxonomy that mainly depends on the roles that the word can play within the correct sentence. Therefore, through this work searchers were able to present this free-context grammar whose terminal vocabulary is SHA vectors, their limited number is allowing to replace word with the class that belong to him. Thus, what is needed here is to create an embedded lexical corpus by adding the SHA to the definition of words like annotation, so that lieds to replace words with their representations in order to express them within sentences or use it for syntactic analysing of texts.

This research paper dealt with work that falls within a project aimed at creating applications in the field of automatic processing of the Arabic language with a new view that depends on the specificities of the Arabic language, which makes it different from other languages, which has always been a source of problems, as most of the research tried to subject this language to approaches that have proven Its success in processing other languages that we consider to be very different from the Arabic language, which led to the limited results obtained for Arabic.

The work presented in this study has made significant strides in the field of automated Arabic language processing, particularly in syntactic analysis. Several key factors were taken into consideration, including the richness of the lexicon and the strictness of the syntax. Previous works have emphasized the need to adopt derivations and flexions to address the morphological aspect. Furthermore, our study recognizes that the Arabic language is inherently dependent on distributed semantics. Therefore, we focus on approaches that are based on distributed semantics, which involve a continuous semantic treatment throughout all stages, from morphology to the lexicon and then to the syntax.

Future work will explore the potential of using the SHA representation as an embedding method. Additionally, the proposed taxonomy and representation can be leveraged for question and answering systems to accurately identify the different parts of a question and train machine learning models for this purpose.

ACKNOWLEDGMENT

To complete this research, the authors would like to thank Covenant University Centre for Research, Innovation, and Discovery (CUCRID) for their unending support in making this research a reality.

REFERENCES

[1] Ababou, N., Mazroui, A., Belehbib, R. (2017). Parsing Arabic Nominal sentences using context free grammar and fundamental rules of classical grammar. International Journal of Intelligent Systems and Applications, 9(8): 11-24. https://doi.org/10.5815/ijisa.2017.08.02

[2] Alqrainy, S., Muaidi, H., Alkoffash, M.S. (2012). Context-free grammar analysis for Arabic sentences. International Journal of Computer Applications, 53(3): 7-11. https://doi.org/10.5120/8399-2167

[3] Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., Eskander, R. (2014). Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 2348-2354.

[4] Khoufi, N., Aloulou, C., Belguith, L.H. (2015). Arabic probabilistic context free grammar induction from a treebank. Research in Computing Science, 90: 77-86.

[5] El-Kaoub, A.A., Echetoui, A.S. (1993). Al-Kafi in Arabic Rhetoric Sciences. Open University of ELESKANDRIA Egypt, pp. 40.

[6] Hedjazi, M.F. (2007). Introduction to Linguistics Fields and Directions. The Egyptian Saudi edition, pp. 126.

[7] Belguith, L.H., Baccour, L., Ghassan, M. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. In Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles Courts, pp. 451-456.

[8] Yahiaoui, Y., Lehirech, A. (2013). Meta description logics knowledge base for Arabic language processing. In Proceeding, The Third International Conference on Digital Information Processing and Communications (ICDIPC2013) IAU United Arab Emirates. 04/02/2013.

[9] Khalifa, I., Feki, Z., Farawila, A. (2011). Arabic discourse segmentation based on rhetorical methods. International Journal of Electric & Computer Sciences, 11(1): 10-15.

[10] Habash, N., Rambow, O. (2004). Extracting a tree adjoining grammar from the Penn Arabic Treebank. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04), pp. 277-284.

[11] Salloum, S.A., Al-Emran, M., Shaalan, K. (2016). A survey of lexical functional grammar in the Arabic context. International Journal of Computing and Network Technology, 4(3): 141-146. http://dx.doi.org/10.12785/IJCNT/040304

[12] Khoufi, N., Aloulou, C., Belguith, L.H. (2016). Parsing Arabic using induced probabilistic context free grammar. International Journal of Speech Technology, 19(2): 313-323. https://doi.org/10.1007/s10772-015-9300-x

[13] Laila, J. (2021). https://mqaall.com/how-many-words-in-the-arabic-language-and-a-comparison-between-the-number-of-words-in-the-arabic-language-and-other-languages/, accessed on 1 December, 2022.

[14] Yahiaoui, Y., Boudjenane, S., Bendebiche, R. (2019). Proposed Representation for Arabic Text Segmentation based on syntactic roles categorization. Conference paper IHSED, Munich, Germany, pp. 16-18.

[15] Yahiaoui, Y., Lehireche, A., Bouchiha, D. (2016). Proposed representation approach based on description logics formalism. I.J. Intelligent Systems and Applications, 8(5): 1-9. https://doi.org/10.5815/ijisa.2016.05.01

[16] Daimi, K. (2001). Identifying syntactic ambiguities in single-parse Arabic sentence. Computers and the Humanities, 35(3): 333-349. https://doi.org/10.1023/A:1017941320947

[17] El-Shishiny, H. (1990). A formal description of Arabic syntax in definite clause grammar. In COLING 1990 Volume 3: Paper Presented to the 13th International Conference on Computational Linguistics, pp. 345-347.

[18] Bakir, M.J. (1979). Aspects of Clause Structure in Arabic: A Study in Word Order Variation in Literary Arabic. Indiana University.

[19] Al-Khuli, M.A. (1979). A Contrastive Transformational Grammar: Arabic and English (Vol. 10). Brill Archive.

[20] Ayoub, G. (1981). Structure de la phrase verbale en arabe standard. Etudes Arabes Saint-Denis, 1(2): 1-367.

[21] Fehri, A.F. (1981). Complémentation et anaphore en arabe moderne: Une approche lexicale fonctionnele. Doctoral dissertation, Univeristé de Paris III, Paris, France.

[22] Ditters, E. (2001). A formal grammar for the description of sentence structure in modern standard Arabic. In EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects, pp. 31-37.

[23] Jaccarini, A. (2001). A modifiable structural editor of grammars for arabic processing. In the Proceeding of Arabic. NLP Workshop at ACL/EACL.

[24] Rafea, A.A., Shaalan, K.F. (1993). Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network. Software: Practice and Experience, 23(6): 567-588. https://doi.org/10.1002/spe.4380230602

[25] Othman, E., Shaalan, K., Rafea, A. (2003). A chart parser for analyzing modern standard Arabic sentence. In Workshop on Machine Translation for Semitic Languages: Issues and Approaches, pp. 37-44.