I.J. Intelligent Systems and Applications, **2017, 8, 11-24**
Published Online August 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.08.02

Parsing Arabic Nominal Sentences Using Context Free Grammar and Fundamental Rules of Classical Grammar

Nabil Ababou and Azzeddine Mazroui
University Mohammed First, Faculty of Sciences, Oujda, Morocco
E-mail: nabilaababou@gmail.com, azze.mazroui@gmail.com

Rachid Belehbib
University Mohammed First, Faculty of Arts and Humanities, Oujda, Morocco
E-mail: racbel59@hotmail.com

Received: 06 March 2017; Accepted: 06 July 2017; Published: 08 August 2017

Abstract**—**This work falls within the framework of the Arabic natural language processing. We are interested in parsing Arabic texts. Existing parsers generate parse trees that give an idea about the structure of the sentence without considering the syntactic functions specific to the Arabic language. Thus, the results are still insufficient in terms of syntactic information. The system we have developed in this article takes into consideration all these syntactic functions. This system begins with a morphological analysis in the context. Then, it uses a CFG grammar to extract the phrases and ends by exploiting the formalism of unification grammar and traditional grammar to combine these phrases and generate the final sentence structure.

Index Terms—POS tagger, Parser, Arabic phrase, grammar, syntax tree, syntactic functions.

I. INTRODUCTION

Parsing is a fundamental step to the design of several applications in Arabic natural language processing such as spelling and grammar checker, information retrieval, automatic generation of sentences, machine translation, conversion information system and Querying Database [1,2].

Parsing a sentence is usually a tricky task. It is more complex with languages whose morphology and syntax is very rich, as in the case of the Arabic language. This explains the challenges that face the development of automatic systems allowing to carry out a syntactic analysis.

Arabic parsers have been reported in [3,4] All these initiatives use grammars created manually. Recently, Arabic Treebank (ATB) was used to improve the performance of the syntactic analysis since it covers widely the Arabic language [5].

Similarly, approaches based on statistical treatment have been developed [6]. However, these analyzers have adopted techniques used for English and do not take into account the specificities of the Arabic language. Thus, if we consider the outputs of the Stanford parser[1] related to the analysis of the four simple sentences of Table 1, we notice that we have no information about the subject (المبتدأ *Almbtd>*[2]\) or the predicate (الخبر *Alxbr*\) of the first two sentences of the table. The analyzer does not distinguish between the words سعيدا *sEdA*\ (happy) and قادم \qdm\ (coming), while they play two different syntactic roles: predicate for the first and circumstantial phrase (الحال *AlHAl*\) for the second. For the last two examples, the system generates the same tree consisting of a single phrase despite the difference between them. Indeed, the third example is a complete sentence composed of two phrases that are the subject الولد *Alwld*\ (the boy) and the predicate مبتسم *mbtsm*\ (smiling), while the last example is not a complete sentence but only a phrase composed of a noun الولد and its adjective المبتسم *Almbtsm*\ (the smiling).

Table 1. Result the analysis of four examples by the Stanford parser

N	Sentence	Result
1	الولد قادم سعيدا *Alwld qAdm sEydA*\ (The boy is coming happy)	(ROOT (S (NP (DTNN الولد)) (ADJP (JJ قادم) (JJ سعيدا))))
2	الولد سعيدا قادم *Alwld sEydA qAdm* \ (The boy is coming happy)	(ROOT (S (NP (DTNN الولد)) (ADJP (JJ سعيدا) (JJ قادم))))
3	الولد مبتسم *Alwld mbtsm*\ (The boy is smiling)	(ROOT (NP (DTNN الولد) (DTJJ مبتسم)))
4	الولد المبتسم *Alwld Almbts*\ (The smiling boy)	(ROOT (NP (DTNN الولد) (DTJJ المبتسم)))

Unlike the other parsers, which have adopted annotations derived from those introduced by English

[1] https://nlp.stanford.edu/software/lex-parser.html
[2] Buckwalter transliteration http://www.qamus.org/transliteration.htm

treebanks, we have opted for annotations and terminology inspired by classic grammatical analyzes of the Arabic language.

The paper is organized as follows. We recall in the following section the previous works and the different approaches used to build parsers. We give in the third section an overview of the POS tagger Alkhalil [7] used in the first phase of our system. The fourth section is devoted to a description of the adopted method and the evaluation is detailed in the fifth section. We end the paper with a conclusion.

II. State of the Art

Parsers based on machine learning can be grouped into two main categories: rule-based systems [8-10] and systems using statistical approaches [11]. Before presenting the main parsers developed for the Arabic language, we will recall two grammars used by these parsers.

- Constituency grammar: The American linguist Noam Chomsky [12] initiated the phrase structure grammar. In this formalism, the sentence is considered as the juxtaposition of syntactic units, called phrases, themselves decomposable into simpler syntactic units.
- Dependency grammar: This model is based on the theory developed by the works of the French linguist Lucien Tesnière [13,14]. The analysis system takes into account the dependencies between the different elements of the sentence.

We give below an overall idea about the different works in this field.

A. Rule-based Parser

This type of parsers is based on grammatical rules to build the structure of the sentence [9,15]. Thus, Attia's team developed in [16] a parser using XLE environment (Xerox Linguistics Environment). This environment captures the rules of grammar and notations following the Lexical Functional Grammar (LFG grammar). They also provided a description of the main syntactic structures of the Arabic language in the framework of LFG grammars. According to the developers of this analyzer, the parser reaches a coverage of 92%. It should be noted that this parser used annotations imported from Universal Grammar such as 'modifier' and 'specifier', and this is not suited to the traditional grammar. Similarly, Othman et al. developed a chart parser to analyze Arabic sentences by using the formalism of unification-based grammar [8]. The grammar used is implemented in SICStus Prolog 3.10. It is composed of 170 rules divided into 22 groups, each of which is a grammatical category. Nadim’s team [19] implemented a parser based on Context Free Grammar (CFG grammar) to analyze the structures of the Arabic sentences respecting GB theory (Government and Binding) of Chomsky. Finally, Al-Taani et al. developed in [15] a chart parser from top to bottom to analyze simple nominal and verbal Arabic sentences. They used the CFG grammar to represent Arabic grammar. According to their article, the system tested on 36 nominal sentences reached an accuracy of 97.2%, and when tested on 34 verbal sentences the accuracy was equal to 91.2%.

B. Statistical phrasal parsing

These parsers are usually based on Treebank to achieve the training phase [18]. Thus, Kulick‘s team [19] a parser based on the analysis of the PATB (Penn Arabic Treebank) by the use of Bikel analyser [6]. Their evaluation of the system gave an F1-score of 74% for Arabic language. Similarly, a Stanford University team extended the parser developed for English to other languages (Arabic, Chinese, German, French, ...). This parser is constantly improved and is distributed freely on the Stanford University website [20]. Its principle is based on the combination of two models: the phrasal model and the dependency model, and uses the PATB as training corpus. Finally, the Berkeley group from the University of California developed the Berkeley parser [21]. This analyzer can learn other grammars from a treebank. It is freely distributed.

To evaluate these three analyzers (Stanford parser, Bikel parser and Berkeley parser), Green and Manning [5] have experimented them on the PATB. They calculated the accuracy of each parser based on the leaf-ancestor metric [22] instead of Parseval metric [23] The obtained results, which are presented in Table 2, show that the Berkely parser achieves the best accuracy that is in the order of 83.1%.

Table 2. Evaluation of the Three Parsers

Parser	Stanford	Bikel	Berkeley
Accuracy	0.802	0.775	0.831

C. Statistical dependency parsing

Most recent works focused on the dependency grammars that give a representation better suited to languages characterized by a relatively free word order in the sentence, which Arabic language belongs. The majority of these works are based on the MALT Parser system. The latter is used to train dependency syntactic analyzers from an annotated corpus. The system learns to project syntactic and morphosyntactic features on analysis decisions (shift, reduce, creation of dependency arcs). It is a free system implanted in Java and available at http://w3.msi.vxu.se/~nivre/ research / MaltParser.html.

One of the potential benefits of data-driven approaches to natural language is that they can be generalized to new languages provided that the necessary linguistic resources are available. However, it is difficult in practice to realize this passage if the models are applied to a particular language that uses its own linguistic annotations. Thus, several studies have reported an increase in the error rate when applying statistical analyzers developed for English to other languages [24-26].

D. Hybrid parser

Other systems try to combine the constituency and dependency parsing in order to improve the analysis results. Thus, the Stanford team [20] implemented classes that combine these two models.

III. Alkhalil POS Tagger

Alkhalil POS Tagger is an Arabic morphosyntactic tagger. It uses a very rich tag set composed of 27 basic tags to which are combined a number of proclitics and enclitics giving a set of 82 tags. The adoption of this tag set have facilitated the analysis of clitics attached to words [7].

This system meets the needs of many applications of Arabic NLP. It is based on the morphological analyzer Alkhalil Morpho Sys [27] and the hidden Markov models. Learning and testing phases were carried out using the Nemlar corpus [28].

This POS Tagger uses annotations to describe phrases composed of words attached to clitics. It also provides the syntactic function of clitics, which will be very useful for the identification of the phrases and their combinations. For example, the phrases لها, له, بهم ,لكم (*lhA*\, *lh*\, *bhm*\, *lkm*\; to her, to his, with them, to you) have all the tag (jarWamajrour جار ومجرور). Similarly, the analysis of the three words ساعده, ساعدا and ساعداه (*sAEdh*\, *sAEdA*\, *sAEdAh*\,; he helps him, they help, they help him) by this POS Tagger gives respectively the tags (VerbPAst + Object: فعل ماض + مفعول به) , (VerbPast + Subject: فعل ماض + فاعل) and (Verbpast + Subject + Object: فعل ماض + فاعل + مفعول به).

IV. Method Description

Our approach is inspired by both the works of Chomsky [12] and those of Sibawayh [29]. These two linguists had given different but not contradictory analyzes. These analyzes are rather complementary and even similar in many parts.

Given the particularities of the Arabic language, we believe it cannot be represented only by a rewrite rule system (CFG grammar, LFG grammar, Generalized phrase structure grammar (GPSG), phrase structure grammar Guided by the Heads (HPSG)). We believe it is necessary to consider, in addition to these grammars, the formalisms of the categorical grammars that resemble the al3amil theory of ancient Arab grammarians [30]. This will allow us to represent the majority of phenomena specific to the Arabic language.

We recall that the origins of the categorical grammars appear in the works of Husserl [31], which has distinguished between categorematic expression and the syncategorematic expressions. Then, several models as those of Ajdukiewicz [32] and of Bar-Hillel [33], which distinguish between basic categories (atomic) and operators categories (functor category), formalized this idea. These express the grammatical link between words in the same sentence.

In addition to these two categories, these models only use two rules of reduction in order to judge whether a sentence is syntactically correct or not.

(1) Right reduction

$$x/y \quad y \quad \rightarrow \quad x$$

(2) Left reduction

$$y \quad y\backslash x \quad \rightarrow \quad x$$

The example below shows how we apply this model to the sentence التلميذ يقرأ الدرس *Altlmy* yqr> Aldrs*\ (the student reads the lesson).

```
الدرس          يقرأ          التلميذ
N            (N\S)/N         N
                      (N\S)
__________________________
              S
```

The functor category (N\S)/N means that the word يقرأ expects a noun phrase to its left and another to its right.

The example below shows that the application of the reduction rules gives the symbol of the basic category "S", which proves that the sentence is correct.

Clearly, these categorical grammars perfectly describe the al3amil theory of classic Arabic grammarians.

Our approach uses both formalism in two juxtaposed phases.

- Phrasal phase: based on the characteristics of the Arabic language, the system uses rewrite rules to create nominal, adjectival and prepositional phrases.
- Categorical phase: the system uses the concepts of the categorical and classical grammars to complete the analysis of the sentence. Functors of our system will be the categories that can act on two arguments: verbs, the verb Kaana and sisters, Inna and sisters, …).

This decomposition allowed us to:

- greatly reduce the number of rewrite rules;
- improve the program complexity;
- use the characteristics of the classical grammar;
- separate the creation stage of nominal, adjectival and prepositional phrases from that identifying the relationship between these phrases and their syntactic functions.

The Arabic language is distinguished from several other languages by the wide flexibility that allows words to change positions without changing their syntactic roles, nor the meaning of the sentence. Thus, the phrases can change their position in the sentence and words can be combined without the need for prepositions (the genitive construction: الإضافة *Al<DAfp*\). This generates

additional difficulties in parsing sentences. Similarly, it partly explains the limitations of approaches that project on the Arabic language the methods developed for the English language.

In the Arabic language, the nominal sentence is composed of two main phrases: the first is the subject of the sentence and is often placed at the beginning of the sentence. The second is the predicate that provides information on the subject and is placed in the most cases after the subject. However, the multitude of different syntagmatic forms that can have the predicate (adverb of time or place, prepositional phrase, verbal sentence, nominal sentence) make the identification of these two entities difficult. In addition, the position before the subject that the predicate can sometimes have makes more delicate the analysis of the sentence.

Unlike the nominal sentence, these phrases still play secondary roles in the construction of verbal sentences.

In the examples below, the phrase في الفصل *fy AlfSl*\ (in the class) in the first sentence, which is verbal, cannot replace the object or the subject of the verb. Therefore, the sentence remains correct without this phrase. However, if we eliminate this phrase in the second sentence, which is nominal, the sentence becomes incomplete.

- دخل أحمد في الفصل *dxl >Hmd fy AlfSl*\ (Ahmed entered the class)
- أحمد في الفصل *>Hmd fy AlfSl* \ (Ahmed is in the class)

Thus, we distinguish between two types of phrases: principal and secondary.

The principal phrase is an indispensable phrase in the sentence structure. The head of this phrase plays one of the following syntactic functions:

- the subject of a nominal sentence (المبتدأ *Almbtd>*\)
- the predicate of a nominal sentence (الخبر *Alxbr*\)
- the subject of a verbal sentence (الفاعل *AlfAEl*\)
- the direct object of a verbal sentence (المفعول به *AlmfEwl bh*\)

Thus, in the sentence دخل مدير مدرسة الحي القريب إلى القاعة *dxl mdyr mdrsp Alhy Alqryb <lY AlqAEp*\ (The principal of the nearby neighborhood school entered the room), the phrase مدير مدرسة الحي القريب is a principal phrase whose head مدير plays the syntactic function subject of a verbal sentence.

The secondary phrase is an additional phrase that enriches the meaning of the sentence. The sentence will therefore remain correct if this phrase is removed. The head of this phrase plays one of the following syntactic functions:

- Appositive (البدل *Albdl*\)'
- Strengthening (التوكيد *Altwkyd*\)
- Circumstantial phrase (الحال *AlHAl*\)
- Excepted, المستثنى (*AlmstvnY*\)
- Adverb of time or place (المفعول فيه *AlmfEwl fyh*\)
- Unrestricted object (المفعول المطلق *AlmfEwl AlmTlq*\)
- Concomitant object (المفعول معه *AlmfEwl mEh*\)
- Causative object (المفعول له *AlmfEwl lh*\)
- Accusative of specification (التمييز *Altmyyz*\)

As we have explained, there are phrases that can play principal roles in nominal sentences and secondary roles in verbal sentences (adverb of time or place, prepositional phrase).

As a result, simple nominal sentence consists of two principal phrases with an unlimited number of secondary phrases (see Fig. 1). Similarly, the number of principal phrases for verbal sentences depend on the nature of the sentence verb (transitive or intransitive).

The three figures below represent the three structures of simple sentences. The dotted arrows represent secondary phrases.

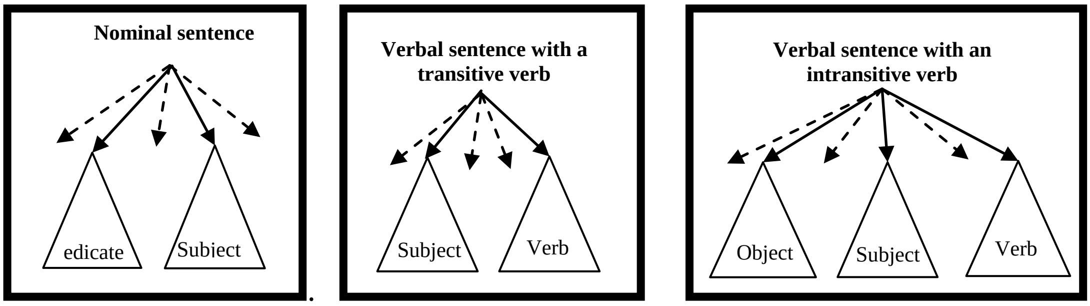
Nominal sentence
edicate
Subject
Verbal sentence with a transitive verb
Subject
Verb
Verbal sentence with an intransitive verb
Object
Subject
Verb

Fig. 1. Structures of three sentences

Note here that the order of the phrases may change. Indeed, the predicate may precede the subject in the nominal sentence and the object can precede the subject in the verbal sentence.

The different steps of the system that we have developed are shown in Fig. 2 below.

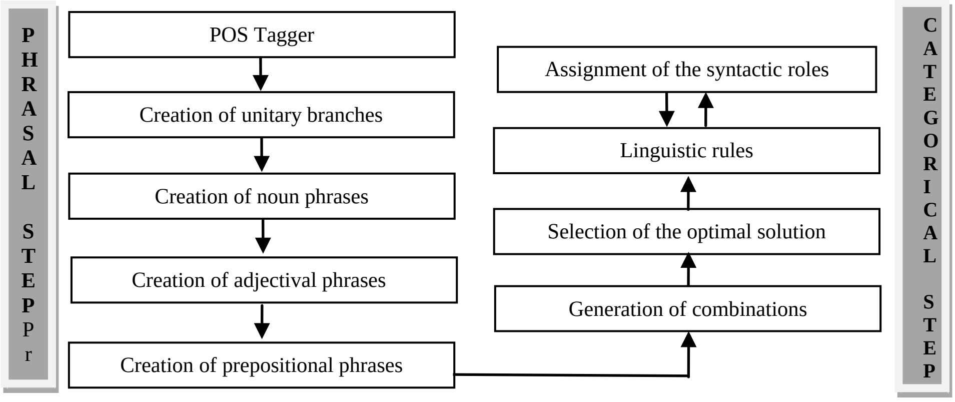
PHRASAL STEP P r
POS Tagger
Creation of unitary branches
Creation of noun phrases
Creation of adjectival phrases
Creation of prepositional phrases
Assignment of the syntactic roles
Linguistic rules
Selection of the optimal solution
Generation of combinations
CATEGORICAL STEP

Fig.2. Steps of the system

We give in the following a detailed description of each of these phases.

A. POS Tagger

The POS Tagger we described above provides us with the most probable tag for each unit of the sentence by exploiting its context.

B. Creation of Unitary Branches

In this part, the system uses the results of the previous phase and some linguistic rules to create unitary branches and judge whether the word may be principal or secondary.

For example of linguistic rules, if the POS Tagger assigns to a word the DSifa tag (definite adjective), the system creates for this word two unitary branches $_{i}[\text{DSifa}]_{i+1}$ and $_{i}[\text{DNo}]_{i+1}$, where (i+1) is the position of the word in the sentence. The first unitary branch will be used in the creation step of adjectival phrases, and the second branch may be a component of a definite annexation (see examples of Table 3).

Table 3. Examples of unitary branches

Sentence	Correct unitary branch	Justification
المجتهد نشيط *Almjthd n$yT*\ (The assiduous is active).	$_{0}[\text{DNo}]_{1}$	The word "المجتهد" is the head of a primary phrase. This is the subject of a nominal sentence. The POS Tagger provides the DSifa tag for this word. However, the system needs a principal phrase $_{0}[\text{DNo}]_{1}$ that plays the role of a subject in order to complete its analysis.
صديق المجتهد نشيط *Sdyq Almjthd n$yT*\ (The friend of the assiduous is active).	$_{1}[\text{DNo}]_{2}$	For this example, the POS Tagger provides the DSifa tag for the word "المجتهد", while the later does not play the role of an adjective. It is part of a genitive construction in which it plays the role of an annexing (مضاف إليه *lDAf <lyi*\). Thus, the system creates a noun phrase.
التلميذ المجتهد نشيط *Altlmy* Almjthd n$yT*\ The assiduous student is active.	$_{1}[\text{DSifa}]_{2}$	The word "المجتهد" is an adjective of an adjectival phrase.

We recall that the unitary branch transfers its nature to all phrases that will accept it as head.

C. Creation of noun phrases

In this part, the system identifies all possible noun phrases based on the genitive construction (الإضافة *Al<DAfp*\). The genitive construction is the juxtaposition of two nouns, so that the second completes the first (كتاب الصديق *ktAb AlSdyq*\ (the friend's book)).The genitive construction can juxtapose more than two nouns (كتاب صديق أخي *ktAb Sdyq >xy*\ (a book of my brother's friend)). In this case, all the nouns except the last are attached neither to the definite article al (ال التعريف *Al AltEryf* \) nor to a pronoun and do not accept nounation (التنوين *Altnwyn*\). We distinguish two types of genitive construction: the definite and the indefinite genitive construction.

An indefinite genitive construction is a noun phrase consisting of a sequence of indefinite words (see example in Fig. 3: تفاصيل حكاية مهاجر *tfASyl HykAyp mhAjr*\ (details of the history of an immigrant)).

It is generated by the following rules:

$$A \longrightarrow N\_N \;\mid N\_No$$

$$N\_N \longrightarrow N\_No + N\_N \;\mid N\_No$$

where

A: indefinite genitive construction

N_No: indefinite word.

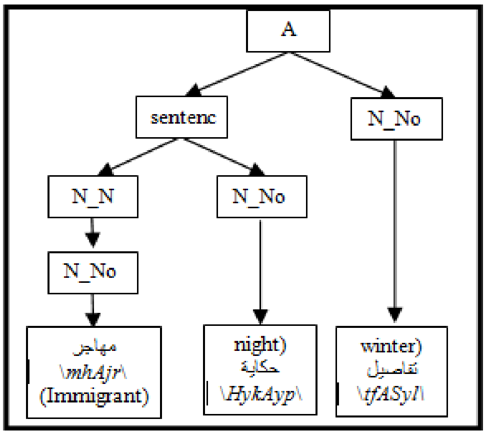

Fig.3. Example of indefinite annexation

Similarly, a definite genitive case is a nominal construction that ends with a definite noun preceded by a series of indefinite words (see example in Fig. 4). It is generated by the following rules:

DA ⟶ D_N | D_No

D_N ⟶ N_No + D_N | D_No | Alam | No_S

where

DA: definite genitive construction

DNo: definite word with the article al

No.S: definite word because it is attached to a pronoun

Alam: proper noun

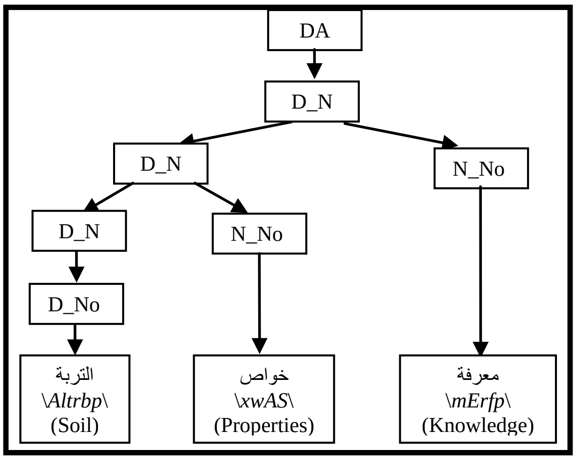

Fig.4. Example of definite annexation

Information on the type of a nominal genitive construction is essential when seeking to combine it with adjectives.

We recall here that the head of the phrase specifies whether this is a principal or secondary phrase.

D. Creation of Adjectival Phrases

For adjectives, we adopted the definition given by the ancient grammarians of the Arabic language, instead of those given in previous works and inspired by English.

The adjectival phrase is a phrase composed of an adjective preceded by a phrase that contains a noun qualified by this adjective.

The adjective is called qualifier (نعت *nEt*\) or describing (صفة *Sft*\). The noun qualified or described by this adjective is called qualified (منعوت *mnEwt*\) or described (موصوف *mwSwf*\).

Based on morphosyntactic information (nounation, the definite article al, ...), we combine in this part the adjectives with noun phrases that precede them directly. The adjectival phrases can be definite or indefinite and are generated by the following rules:

DAdjP ⟶ DA + DSifa

AdjP ⟶ A + Sifa

where

DAdjP: definite adjectival phrase

DSifa: definite adjective

AdjP: indefinite adjectival phrase

Sifa: indefinite adjective

These rules generate branches that indicate that the adjective describes a word of a genitive construction.

In Arabic, the describing is a word that often comes after the described noun. However, in some cases, nouns can be inserted between describing and described. In this case, the described will be identified using the rule which states that the describing must agree in case, number and gender with the described (see examples of Fig. 5). In case these rules fail in eliminating the ambiguity, the system considers the head of the phrase that precedes the describing as described of the latter.

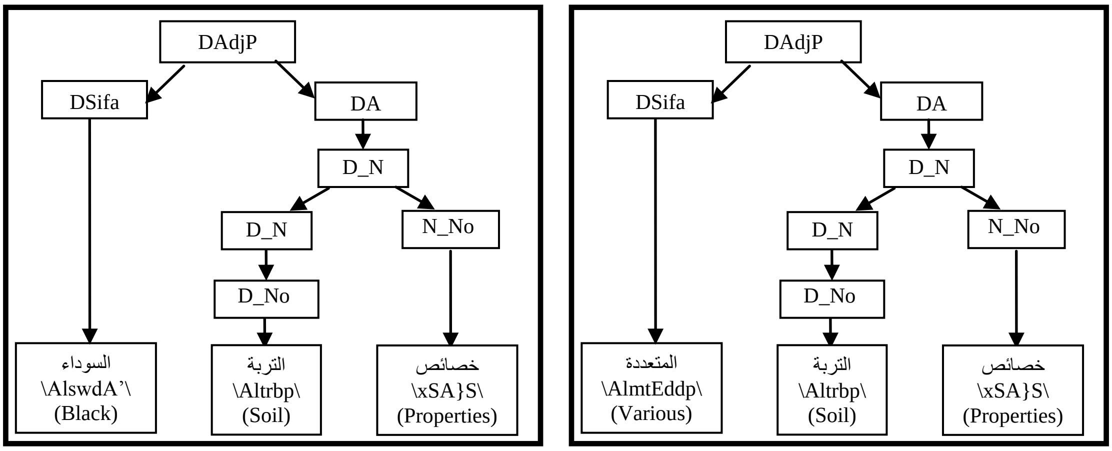

Fig.5. Examples of adjectival phrases

E. Creation of Prepositional Phrases

We combine each particle of the sentence (preposition: حرف الجر *Hrf Aljr*\, conjunction: حرف العطف *Harf AlETf*\, exception particle: أداة استثناء *>dAt AstvnA'*\) with the nominal or adjectival phrase that follows this particle to get a phrase with the syntactic function of the particle.

In nominal sentences, phrases whose head is a preposition or an exception particle can play the role of a principal or secondary phrase (see examples of Table 4)

Table 4. Examples of prepositional phrases

Sentence	Syntactic function
التلميذ في القسم *\Altlmy* fy Alqsm\* The student is in the classroom.	The expression في القسم is a principal phrase. It plays the role of a predicate that is an indispensable element for the sentence decomposition.
التلميذ نشيط في القسم *\Altlmy* n$yT fy Alqsm\* The student is active in the classroom.	The sentence remains complete without the prepositional phrase في القسم, which is secondary. The phrase نشيط plays the role of the predicate in this sentence.

F. Generation of combinations

After identifying phrases, the system searches all possible combinations of the sentence. As it is known that the number of principal phrases in a nominal sentence is often equal to two, and may in exceptional cases reach the value three, we have developed an algorithm that calculates the number of principal phrases in an Arabic sentence previously segmented into phrases. For this, we define for each phrase *S* two parameters *(Ps,Qs)* that we will use later to determine the number of principal phrases in each sentence. These parameters are defined as follows:

- *(Ps,Qs) = (1,0)* when the phrase *S* is always primary.
- *(Ps,Qs) = (1,1)* if the phrase *S* may be primary or secondary depending on its position in the sentence.
- *(Ps,Qs) = (0,0)* if the phrase *S* may in no case be a principal phrase.

The following table gives the values of *(Ps,Qs)* assigned to the unit phrases according to the basic tag proposed by the POS Tagger.

Table 5. Values of (*P*,*Q*) assigned to the unitary branches

Tag	Tag symbol	Tag in Arabic	(P_s,Q_S)	Example		*Tag*	Tag symbol	Tag in Arabic	(P_s,Q_S)	Example
Past tense	VePa	فعل ماض	(0,0)	قال		Preposition	Pr	حرْف جَر	(0,0)	عن
Present tense (Imperfect verb)	VeP	فعل مضارِع	(0,0)	تدور		Coordinating + conjunction	Co	حرف عطف	(0,0)	أو, بل, لكن
Imperative verb	VeIm	فعل أمر	(0,0)	نم		Articles within syntactic effect	Ns	حرف غير عامل	(0,0)	أي, قد, أما
Det+Noun	Dno	اسم معرف	(1,0)	العشب		Vocative particle	Interj	حرف نداء	(0,0)	يا
Noun	N_No	اسم	(1,0)	رجل		Particles that make the imperfect verb in the subjunctive	PaNs	حرفناصِب	(0,0)	أن
Proper noun	Alam	علم	(1,1)	عمر		Prohibition or negative	La	لا الناهية أو النافية	(0,0)	لا
Det+ Adjective	DSifa	صفة معرفة	(1,1)	اليابس		Exception particle	Ex	أداة استثناء	(1,1)	إلا
Adjective	Sifa	صفة	(1,1)	حامل		Particles that make the subject of the nominal sentence in 2nd Arabic syntactic case	PNa	حرف ناسِخ	(0,0)	أن
Demonstrative pronoun	NoDP	اسم إشارة	(0,0)	ذلك		Det+ abbreviation	DAbrv	مختصر معرف	(1,1)	الآر
Relative pronoun	NoRP	اسم موصول	(0,0)	التي		Abbreviation	Abrv	مختصر	(1,1)	بي
Comparative	Compa	اسم تفضيل	(1,1)	أشعث		Particles that make the imperfect verb in the jussive	CJ	حرف جزم	(0,0)	لم
Det +comparative	DCompa	اسم تفضيل معرف	(1,1)	الأحدب		Preposition+Pronoun Compound	PCo	جار ومجرور	(1,1)	له
Time or location adverb	Dz	ظَرف زمان أو مكان	(1,1)	تحت		NounInfinit	NounInfinit	مصدر	(1,1)	ضرب

The unitary branch composed of the noun رجل *\rjl\* (a man) is always principal because all phrases that accept this word as head necessarily play a principal role. However, the phrases that accept the proper noun عمر *\Emr\* (Omar) as head will have the opportunity to play both roles. Indeed, a noun can play the secondary syntactic role of appositive, as it can play a principal role. The example of the following two sentences illustrates these two cases:

- الخليفة عمرعادل *\Alxlyfp Emr EAdl\* (The caliph Omar is impartial)
- عمر عادل *\Emr EAdl* (Omar is impartial)

Similarly, for a sentence *Ph* decomposed into phrases $(S_i)_{1\leq i\leq k}$, we note:

$$(C_P, C_Q) =(\textstyle\sum_{i=1}^{k} P_{S_i}, \sum_{i=1}^{k} Q_{S_i},) \qquad (1)$$

where C_p is the number of phrases suggested to be principal, and of which C_Q phrases may be secondary.

The steps of the following figure will allow the system to specify the nature of this sentence.

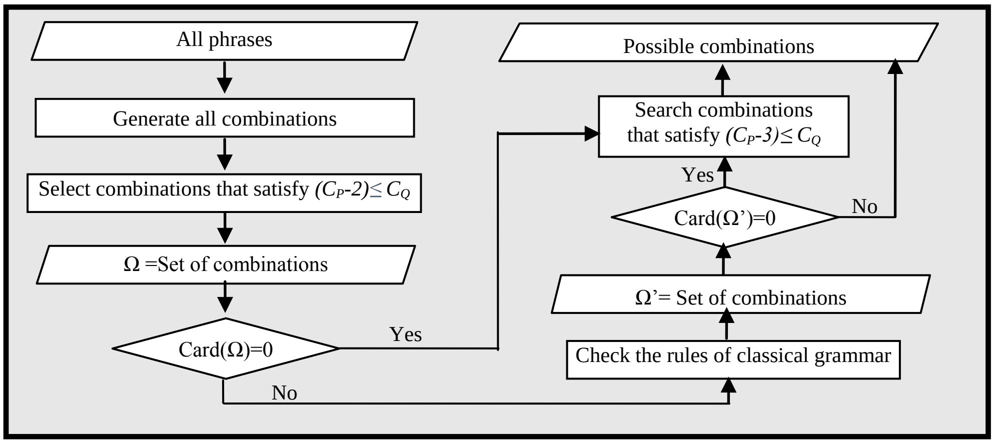
All phrases
Generate all combinations
Select combinations that satisfy $(C_P-2)\leq C_Q$
Ω =Set of combinations
Card(Ω)=0
Yes
No
Search combinations that satisfy $(C_P-3)\leq C_Q$
Possible combinations
Yes
Card(Ω')=0
No
Ω'= Set of combinations
Check the rules of classical grammar

Fig.6. Generation of possible combinations

We will give below some examples to illustrate the different steps of this algorithm.

Example 1: مدير مدرسة الحي رجل عظيم بأعماله *\mdyr mdrsp Alhy rjl Edym b>EmAlh* (The director of the neighborhood school is a great man thanks to his work).

By analyzing this sentence, the system provides in the beginning the following different unitary branches:

Table 6. Possible unitary branches

Sentence		بأعماله	عظيم	رجل	الحي	مدرسة	مدير
Results of the POS Tagger	Unitary branche	$_{5}$[Pr.No.S]$_{6}$	$_{4}$[Sifa]$_{5}$	$_{3}$[N_No]$_{4}$	$_{2}$[DNo]$_{3}$	$_{1}$[N_No]$_{2}$	$_{0}$[Sifa]$_{1}$
	(P_s,Q_s)	(1,1)	(1,1)	(1,0)	(1,0)	(1,0)	(1,0)
additional opportunities following the linguistic rules	Unitary branche		$_{4}$[N_No]$_{5}$				$_{0}$[N_No]$_{1}$
	(P_s,Q_s)		(1,0)				(1,1)

Since the Arabic sentences do not begin with a qualifier, the SIFA tag of the word مدير will not be taken into account. Using the rules for generating nominal, adjectival and prepositional phrases (see sections c, d and e), we get the following phrases:

Table 7. Possible phrases

Sentence	بأعماله	عظيم	رجل	الحي	مدرسة	مدير
Noun phrases		رجل عظيم $_{3}$[A]$_{5}$, (1,0)		مدرسة الحي $_{1}$[DA]$_{3}$, (1,0)		
					مدير مدرسة $_{0}$[A]$_{2}$, (1,0)	
				مدير مدرسة الحي $_{0}$[DA]$_{3}$, (1,0)		
adjectival phrases		رجل عظيم $_{3}$[AdjP]$_{5}$, (1,0)				
prepositional phrases	بأعماله $_{5}$[Pr.No.S]$_{6}$, (1,1)					

where (i+1) and k in $_{i}[\ ]_{k}$ denote respectively the position of the first word and the last word of the phrase.

We present in the following table all possible combinations:

Table 8. Possible combinations of phrases

Sentence	بأعماله	عظيم	رجل	الحي	مدرسة	مدير	(C_p,C_Q)
Generation of combinations	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[No]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	الحي ${}_2$[DNo]${}_3$, (1,0)	مدرسة ${}_1$[No]${}_2$, (1,0)	مدير ${}_0$[No]${}_1$, (1,0)	(6,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[Sifa]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	الحي ${}_2$[DNo]${}_3$, (1,0)	مدرسة ${}_1$[No]${}_2$, (1,0)	مدير ${}_0$[No]${}_1$, (1,0)	(6,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[A]${}_5$, (1,0)		الحي ${}_2$[DNo]${}_3$, (1,0)	مدرسة ${}_1$[No]${}_2$, (1,0)	مدير ${}_0$[No]${}_1$, (1,0)	(5,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[AdjP]${}_5$, (1,0)		الحي ${}_2$[DNo]${}_3$, (1,0)	مدرسة ${}_1$[No]${}_2$, (1,0)	مدير ${}_0$[No]${}_1$, (1,0)	(5,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[No]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	مدرسة الحي ${}_1$[DA]${}_3$, (1,0)		مدير ${}_0$[No]${}_1$, (1,0)	(5,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[Sifa]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	مدرسة الحي ${}_1$[DA]${}_3$, (1,0)		مدير ${}_0$[No]${}_1$,(1,0)	(5,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[[A]${}_5$, (1,0)		مدرسة الحي ${}_1$[DA]${}_3$, (1,0)		مدير ${}_0$[No]${}_1$, (1,0)	(4,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[[AdjP]${}_5$, (1,0)		مدرسة الحي ${}_1$[DA]${}_3$, (1,0)		مدير ${}_0$[No]${}_1$, (1,0)	(4,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[No]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	الحي ${}_2$[DNo]${}_3$, (1,0)	مدير مدرسة ${}_0$[A]${}_2$, (1,0)		(5,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[Sifa]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	الحي ${}_2$[DNo]${}_3$, (1,0)	مدير مدرسة ${}_0$[A]${}_2$, (1,0)		(5,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[A]${}_5$, (1,0)		الحي ${}_2$[DNo]${}_3$, (1,0)	مدير مدرسة ${}_0$[A]${}_2$, (1,0)		(4,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[AdjP]${}_5$, (1,0)		الحي ${}_2$[DNo]${}_3$, (1,0)	مدير مدرسة ${}_0$[A]${}_2$, (1,0)		(4,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[No]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	مدير مدرسة الحي ${}_0$[DA]${}_3$, (1,0)			(4,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	عظيم ${}_4$[Sifa]${}_5$, (1,0)	رجل ${}_3$[No]${}_4$, (1,0)	مدير مدرسة الحي ${}_0$[DA]${}_3$, (1,0)			(4,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$[A]${}_5$, (1,0)		مدير مدرسة الحي ${}_0$[DA]${}_3$, (1,0)			(3,1)
	بأعماله ${}_5$[Pr.No.S]${}_6$, (1,1)	رجل عظيم ${}_3$ [AdjP]${}_5$, (1,0)		مدير مدرسة الحي ${}_0$[DA]${}_3$, (1,0)			(3,1)

As the last two combinations are the only ones to check the condition $C_P\text{-}2\leq C_Q$ (therefore $Card(\Omega) \neq 0$), they will be the only solutions returned by the system. Each of these two potential solutions has two main phrases (the first one is the subject and the second is the predicate), and the third is necessarily secondary.

G. Select the optimal solution

The system uses several priority criteria to choose the best combination among the proposals generated during the previous phase. For example, we quote:

- the system favors combinations with a reduced number of phrases;
- the system favors the prepositional phrase (جار ومجرور \jAr w mjrwr\) on the appositive phrase to be a principal;
- the system favors the adjectival phrase on the noun phrase when the described agrees with the adjective in number, gender and type (defined or indefinite).

Example 2: الولد محمد في القسم *Alwld mHmd fy Alqsm*\ (The boy Mohammed is in the classroom).

For this example, the system generates a single combination.

Table 9. Possible combinations of phrases

Sentence	في القسم	محمد	الولد	(C_P,C_Q)
Suggested combinations	${}_2$[Pr.Dno]${}_3$, (1,1)	${}_1$[Alam]${}_2$, (1,1)	${}_0$[DNo]${}_1$, (1,0)	(3,1)

The condition $C_P\text{-}2\leq C_Q$ is satisfied and the sentence already contains the phrase الولد definitively considered principal. The second principal phrase will be في القسم and not محمد, because the system favors the prepositional phrase on the appositive phrase to be principal.

Example 3: كان للأوقاف أو الحبوس الإسلامية دور جليل في الحضارة الإسلامية *kAn ll>wqAf >w AlHbws Al<slAmyp dwr jlyl fy AlHDArp Al<slAmyp*\ (The awqaf or the Islamic Hobous had a crucial role in the Islamic civilization).

After analysis, the only solutions satisfying the condition *($C_P\text{-}2\leq C_Q$)* are presented below:

Table 10. Possible combinations of phrases satisfying the condition C_P-$2 \leq C_Q$

Sentence	الإسلامية	الحضارة	في	جليل	دور	الإسلامية	الحبوس	و	للأوقاف	كان	(C_p,C_Q)
Suggested combinations	$_7$[Pr.DNo]$_{10}$, (1,1)			$_6$[N_No]$_7$, (1,0)	$_5$[N_No]$_6$, (1,0)	$_2$[Pr.DNo]$_5$, (0,0)			$_1$[Pr.DNo]$_2$, (1,1)	KAnSistersPast (0,0)	(4,2)
	$_7$[Pr.DNo]$_{10}$, (1,1)			$_5$[A]$_7$, (1,0)		$_2$[Pr.DNo]$_5$, (0,0)			$_1$[Pr.DNo]$_2$, (1,1)	KAnSistersPast (0,0)	(3,2)
	$_7$[Pr.DNo]$_{10}$, (1,1)			$_6$[Sifa]$_7$, (1,1)	$_5$[N_No]$_6$, (1,0)	$_2$[Pr.DNo]$_5$, (0,0)			$_1$[Pr.DNo]$_2$, (1,1)	KAnSistersPast (0,0)	(4,3)
	$_7$[Pr.DNo]$_{10}$, (1,1)			$_5$[AdjP]$_7$, (1,0)		$_2$[Pr.DNo]$_5$, (0,0)			$_1$[Pr.DNo]$_2$, (1,1)	KAnSistersPast (0,0)	(3,2)

For this example, the system chooses those whose number of phrases is the smallest, so that for *(C_P,C_Q)=(3.2)*.

H. Assignment of syntactic roles

Once the principal phrases are selected, the system will assign them their syntactic functions (subject or predicate). The secondary phrases keep the syntactic function of their heads, which were allocated during the creation of unitary branches, and will be attached to their fathers according to the syntactic functions.

Moreover, since the Arabic language allows the predicate to precede the subject in some cases, we then set a priority order for that a phrase plays the role of a subject. This order was inferred from the rules of classical Arabic grammar and is presented in the table below.

Table 11. Role priority

Phrase head	Priority
Pronoun (الضمير \AlDmyr\\) ; Demonstrative pronoun (اسم إشارة *Asm <$Arp*\\)	1
Proper noun (اسم علم *Asm Elm*\\)	2
definite noun with the definite article Al (اسم معرف بال *Asm mErf bAl*\\)	3
definite adjective (صفة معرفة *Sfp mErfp*\\)	4
indefinite noun (اسم نكرة *Asm nkrp*\\)	5
indefinite adjective (صفة نكرة \\ *Sfp nkrp*\\)	6
Asma’ sadara (أسماء الصدارة \\ *Asm’ AlSdArp*\\)	7
Adverb of time or place (ظرف \Drf\\) Prepositional phrase (جار ومجرور *jAr wa mjrwr*\\)	8

Thus, if we consider the sentence "في القسم معلم *fY Alqsm mElm*\\ (a teacher is in class), then the system gives us the following result:

Table 12. Possible combinations of phrases

Sentence	معلم	القسم	في	(C_P,C_S)
Suggested combinations	$_2$[N_No]$_3$, (1,0)	$_0$[Pr.Dno]$_2$, (1,1)		(2,1)

The priority of the phrase معلم is equal to 5 (because it is an indeterminate noun), while that of في القسم is equal to 8 (because it is a prepositional phrase). Consequently, the subject will be the phrase معلم and the phrase في القسم will be the predicate.

It remains to note that even after the detection of syntactic roles of the phrases, the system may reject the outcome if the rules of Arabic grammar are not respected. In this case, the system searches a new combination with three principal phrases instead of two. This is especially true when one of the noun phrases of the sentence is attached to a pronoun. The following three examples illustrate this phenomenon.

Example 1: الشتاء ليله طويل *Al$tA’ lylh Twyl*\\ (The night of winter is long)

By analyzing this sentence, the system provides a unique combination of phrases:

Table 13. Possible unitary branches

Sentence		طويل	ليله	الشتاء
Results of the POS Tagger	Unitary branch	$_2$[Sifa]$_3$	$_1$[DNo]$_2$	$_0$[DNo]$_1$
	(P_s,Q_s)	(1,1)	(1,0)	(1,0)
additional opportunities following the linguistic rules	Unitary branch	$_2$[N_No]$_3$		
	(P_s,Q_s)	(1,0)		

The word طويل cannot have the tag Sifa because this word is indefinite, whereas the one that precedes it is defined (defined by annexation because it is attached to a pronoun). Moreover, since the only combination that remains does not verify the property C_P-$2 \leq C_Q$ (i.e. *Card(Ω)=0*), the system automatically switches to combinations composed of three principal phrases (i.e. C_P-$3 \leq C_Q$). Thus, the solution proposed by the system will consist of three principal phrases and no secondary phrase. Using the role priority of Table 11, the first phrase will be a subject and the phrase attached to the pronoun a second subject. The last phrase is the predicate of the second subject and the sentence composed of the last two phrases is the predicate of the first subject. We present in the following figure the results of analysis in the form of a tree.

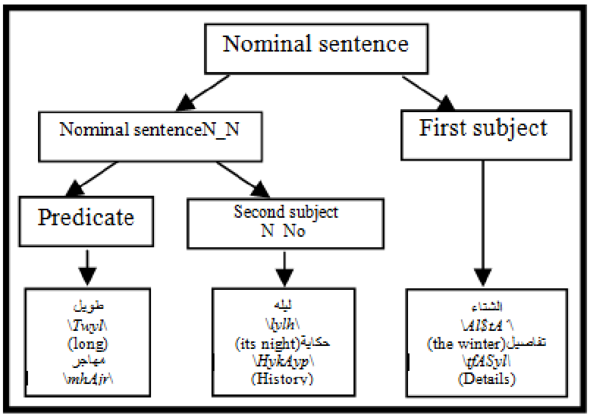
Nominal sentence
Nominal sentenceN_N
First subject
Predicate
Second subject N No
طويل \Twyl\ (long) مهاجر \mhAjr\
ليله \lylh\ (its night)حكاية \HykAyp\ (History)
الشتاء \Al$tA ’\ (the winter)تفاصيل \tfASyl\ (Details)

Fig.7. Analysis of the sentence ‘الشتاء ليله طويل’

Example 2: الشتاء طويل ليله *\Al$tA' Twyl lylh* (The winter, its night is long)

The first stage of the system generates the following result:

Table 14. Possible unitary branches

Sentence		ليله	طويل	الشتاء
Results of the POS Tagger	Unitary branch	$_2$[DNo]$_3$	$_1$[Sifa]$_2$	$_0$[DNo]$_1$
	(P$_s$,Q$_s$)	(1,0)	(1,1)	(1,0)
additional opportunities following the linguistic rules	Unitary branch		$_1$[N_No]$_2$	
	(P$_s$,Q$_s$)		(1,0)	

As the word طويل cannot have the Sifa tag because this word is indefinite whereas the one that precedes it is defined, the unique combination verifying the property $C_P\text{-}2 \leq C_Q$ in this stage is:

Table 15. Possible combinations of phrases

Sentence	ليله	طويل	الشتاء	*(C$_p$,C$_Q$)*
Suggested combination	$_1$[DA]$_3$, (1,0)		$_0$[DNo]$_1$, (1,0)	(2,0)

As a predicate phrase with a derived head used for a description should not be attached to a pronoun that refers to the subject's head, it results the elimination of this combination. Thus, since *Card(Ω)=0* the system restarts the search for three phrases. The only possible combination is shown below:

Table 16. Possible combinations of phrases

Sentence	ليله	طويل	الشتاء	*(C$_p$,C$_S$)*
Suggested combination	$_2$[DA]$_3$, (1,0)	$_1$[No]$_2$, (1,0)	$_0$[DNo]$_1$, (1,0)	(3,0)

The role priority of Table 11 provides the following tree.

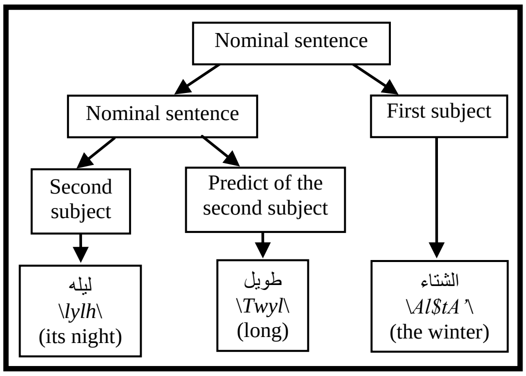

Fig.8. Analysis of the sentence "الشتاء طويل ليله"

Remark: the presence of the pronoun that refers to the subject's head requires the conversion of a combination of three principal phrases in a nested structure of two noun phrases. Otherwise, the combination may be transformed into other forms as shown in the following example.

Example 3: إعدادك الدرس مفيد *\<EdAdk Aldrs mfyd * (Your preparation of the lesson is useful)

The analysis of this sentence generates in the beginning the following combination:

Table 17. Possible unitary branches

Sentence		مفيد	الدرس	إعدادك
Results of the POS Tagger	Unitary branche	$_2$[Sifa]$_3$	$_1$[DNo]$_2$	$_0$[DNo]$_1$
	(P$_s$,Q$_s$)	(1,1)	(1,0)	(1,0)
additional opportunities following the linguistic rules	Unitary branche	$_2$[N_No]$_3$		
	(P$_s$,Q$_s$)	(1,0)		

The word مفيد cannot have the tag Sifa for the same reasons mentioned in the two previous examples.

The unique combination is composed of three principal phrases that contain no pronoun referring to the subject's head. However, the predicate cannot be a nominal sentence as in the case of the two preceding examples. Thus, the system links one of the last two phrases with the one that has a head that can play the role of a verb (المصدر *\AlmSdr* (verbal noun), اسم الفاعل *\Asm AlfAEl* (active participle), اسم المفعول *\Asm AlmfEwl* (passive participle), الصفة المشبهة باسم الفاعل *\AlSfp Alm$bhp bAsm AlfAEl*). Therefore, the head of one of the three phrases plays the role of a verb and the phrase that follows it plays the role of a verbal subject or a direct object. In this example, the verbal noun إعدادك plays the role of the verb as shown in the following tree:

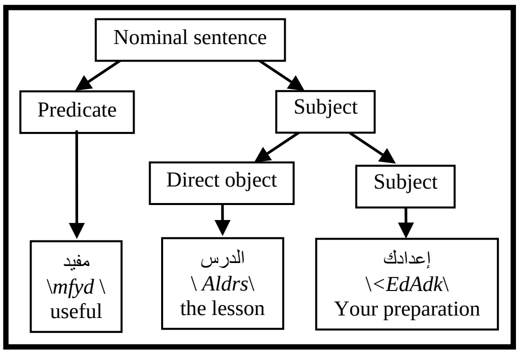

Fig.9. Analysis of the sentence إعدادك الدرس مفيد

V. Testing and Evaluation

To evaluate our system, we have used the Parseval measure [23] usually utilized for components. We calculate the recall and the accuracy of the components, where a component is assumed correct if its type and the types of its left and right neighbouring components are correct.

Accuracy = number of correct components divided by the total number of components produced by the analysis.

Recall = number of correct components divided by the total number of components of the correct analysis.

F-score: combination of recall and precision

$$\text{F} - \text{score} = \frac{2\,(\text{accuracy} \times \text{recall})}{(\text{accuracy} + \text{recall})} \qquad (2)$$

We tested our system on a corpus of 200 nominal sentences from the following three books:

- الأربعين نووية *\Al>rbEyn nwwyp*
- البيئة والحفاظ عليها من منظور إسلامي *\Alby}p wAlHfAZ ElyhA mn mnZwr <slAmy*
- الدستور المغربي *\Aldstwr Almgrby*

These sentences were chosen by browsing the three documents from right to left and keeping only the first 200 nominal sentences. They are composed of 1940 words, and the average number of words by sentence is therefore equal to 9.7. We give in Table 18 some examples of these sentences.

Table 18. Some examples of the selected sentences

Sentences	Book
• إنما الأعمال بالنيات. • كانت هجرته إلى الله ورسوله. • الْمَسْؤُولُ عَنْها أَعْلَمَ مِنَ السَّائِلِ.	الأربعين نووية
• الإسلام خاتم الرسالات الربانية إلي البشر. • كل إفساد مباشر للخصائص العضوية أو الحرارية أو البيولوجية والإشعاعية محرم شرعا. • من أهم مصادر تلوث المياه في العصر الحاضر المواد المشعة • كانت التفجيرات النووية والمفاعلات الذرية ودفن مخلفات المواد المشعة والمواد المشعة المستعملة في الأغراض الطبية والصناعي . وفي توليد الطاقة مصادر أساسية لتلوث المياه بالمواد المشعة.	البيئة والحفاظ عليها من منظور إسلامي
• النظام الدستوري للمملكة قائم على أساس فصل السلط، مع توازنها وتعاونها والديمقراطية المواطنة والتشاركية، وعلى مبادئ الحكامة الجيدة، وربط المسؤولية بالمحاسبة. • الأمة في حياتها العامة مستندة على ثوابت جامعة، متمثلة في الدين الإسلامي السمح، والوحدة الوطنية متعددة الروافد، والملكية الدستورية، والاختيار الديمقراطي. • الأمازيغية أيضا لغة رسمية للدولة، باعتبارها رصيدا مشتركا لجميع المغاربة، بدون استثناء.	الدستور المغربي

After analyzing this corpus, we found that some of the errors are due to the verb tag generated by the POS tagger. Indeed, the POS Tag has given the verb tag for 11 sentences among the 200 analyzed sentences. Since we are interested in the analysis of nominal sentences in this work, we will only give the results of the sentence analysis where the verb tag does not appear among the analysis outputs.

Table 19. Evaluation results

Indicator	value
Accuracy	0.9795%
Recall	0.9750%
f-score	0.9773%
Labeled accuracy	0.9642%
Labeled accuracy	0.9597%
Labeled f-score	0.9619%

The values of the f-score calculated according to the preceding equation are given in Table 19. The analysis of these results shows that the use of classical grammar has a significant influence on the correct identification of dependencies between words. It should also be noted that the majority of failures are consequences of the erroneous results of the pre-processing phases (segmentation and POS tagging).

These performances also show that the approach adopted made it possible to achieve cooperation between the different phases of analysis. Indeed, the analyzer relies on the results of constructing non-atomic phrases to reduce the search space by providing not only atomic objects but also complex information. This greatly reduces the number of ambiguities before assigning syntactic roles.

Thus, our simple parser, which is composed of several layers (unitary branches, phrases, syntactic relations), yielded good results. The simplicity of its structure also makes it possible to add easily and rapidly in one of its layers new techniques, formalisms, priorities or statistical calculations without influencing those used in the other layers. This structure also allowed the system to correct the analysis outputs by acting only on the local results of a layer. These combined factors explain the good results obtained by our system, and encourage us to extend this approach to verbal sentences.

In addition, to the good results obtained compared with those of the state of the art, our system uses tags and trees own to Arabic language, and not deducted of treebanks that follow the English annotations.

Note that a large part of the errors committed by the POS Tag is due to the generation of Sifa and Dsifa tags instead of No and DNo ttags. We solved this problem in the phase of creating unitary branches by systematically adding the No tag (respectively DNo tag) when the POS Tag generates the Sifa tag (respectively DSifa tag).

Nevertheless, it is difficult to compare our results with those of other systems that use different evaluation measures and different test sets. This makes performance comparison a very delicate task. However, we will position our system in relation to the MADAMIRA system [34] by analyzing some example.

The MADAMIRA system aims to identify the superficial syntactic structure of a sentence. It is interested only in simple nominal groups, each consisting of a single noun or pronoun, including its potential immediate adjectival groups, determinants, ... The system does not specify the internal structure of phrases or their syntactic functions. We present in the table below the analysis results of the MADAMIRA system applied to some examples.

Table 20. Some examples of the selected sentences

Sentences	Analysis results			
شيخ القرية الحكيم	ال + قرية ال+ حكيم Noun Phrase			شيخ Noun Phrase
شيخ القرية العظيمة	ال + قرية ال+ عظيمة Noun Phrase			شيخ Noun Phrase
في معرفة خواص التربة	ال+ تربة Noun Phrase	خواص Noun Phrase	معرفة Noun Phrase	في Prepositional Phrase

We note here that this system does not address the phenomenon of genitive construction, while it creates phrases such of (word + adjective). This does not comply with the characteristics of the Arabic language that gives priority to the annexation, then comes the description with adjectives. We also note that neglecting this phenomenon affects the accuracy of the identification of the adjectival phrases as shown in the examples above.

To confirm this, we analyze the two sentences we have already presented in the previous sections by the MADAMIRA system. We illustrate in Table 21 the obtained results.

Table 21. Analysis results of two sentences by the Madamira system

Sentences	Analysis results			
كان للأوقاف أو الحبوس الإسلامية دور جليل في الحضارة الإسلامية	دور Noun Phrase	ال+ اوقاف او ال+ حبوس ال+ اسلامية Noun Phrase	ل+ Prepositional Phrase	كان Verb Phrase
		ال+حضارة ال+اسلامية Noun Phrase	في Prepositional Phrase	جليل Noun Phrase
مدير مدرسة الحي رجل عظيم بأعماله	رجل عظيم Noun Phrase	ال+ حي Noun Phrase	مدرسة Noun Phrase	مدير Noun Phrase
		ه Noun Phrase	أعمال Noun Phrase	ب Prepositional Phrase

It is clear that the MADAMIRA system encountered many problems to identify the nominal phrases based on genitive construction, and to recognize the adjectival phrases.

VI. Conclusion and perspectives

We presented in this paper a new method for parsing the Arabic language based on the CFG grammar and rules of ancient grammarians such as the number of principal phrases in the sentence and the nature of secondary phrases. This combination between the two schools has enabled us to greatly reduce the number of generation rules and to significantly decrease the number of ambiguities. Moreover, the results obtained by our parser are very satisfactory.

As future work, we plan to expand this research on verbal sentences, and integrate it later in a hybrid method based on a statistical model using different learning techniques on a treebank that respects the characteristics of the Arabic language.

References

[1] H. Bais, M. Machkour and L. Koutti, "A Model of a Generic Natural Language Interface for Querying Database", *International Journal of Intelligent Systems and Applications* (IJISA), vol. 8, no. 2, pp. 35-44, 2016. DOI: 10.5815/ijisa.2016.02.05

[2] J. Avinash, O. Agrawal and Kakde, G, "A Semantic Analysis of Natural Language Queries Using Domain Ontology for Information Access from Database", *International Journal of Intelligent Systems and Applications* (IJISA), vol. 5, no. 12, pp. 81-90, 2013. DOI: 10.5815/ijisa.2013.12.07

[3] Z. Žabokrtský and O. Smrž, "Arabic Syntactic Trees: From Constituency to Dependency", in *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, Stroudsburg, PA, USA, 2003, pp. 183–186.

[4] N. Khoufi and M. Boudokhane, "Statistical-based System for Morphological Annotation of Arabic Texts", in *RANLP*, 2013, pp. 100–106.

[5] S. and C. D. Manning, "Better Arabic parsing: Baselines, evaluations, and analysis", in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 394–402.

[6] D. M. Bikel, "On the parameter space of generative lexicalized statistical parsing models", University of Pennsylvania, 2004.

[7] N. Ababou and A. Mazroui, "A hybrid Arabic POS tagging for simple and compound morphosyntactic tags", *Int J Speech Technol*, pp. 1–14, Oct. 2015.

[8] E. Othman, K. Shaalan, and A. Rafea, "A chart parser for analyzing modern standard Arabic sentence", in *Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages*, 2003, pp. 37–44.

[9] S. Alqrainy, H. Muaidi, and M. S. Alkoffash, "Context-free grammar analysis for Arabic sentences", *International Journal of Computer Applications*, vol. 53, no. 3, 2012.

[10] E. Al-Daoud and A. Basata, "A framework to automate the parsing of Arabic language sentences", *Int. Arab J. Inf. Technol.*, vol. 6, no. 2, pp. 191–195, 2009.

[11] L. Tounsi and J. Van Genabith, "Arabic parsing using grammar transforms," 2010.

[12] N. Chomsky, *Syntactic structures*, 14. printing. The Hague: Mouton, 1957.

[13] L. Tesnière, *Esquisse d'une syntaxe structurale*. Paris: C. Klincksieck, 1953.

[14] L. Tesnière, *Eléments de syntaxe structurale*. Librairie C. Klincksieck, 1959.

[15] A. T. Al-Taani, M. M. Msallam, and S. A. Wedian, "A top-down chart parser for analyzing arabic sentences", *Int. Arab J. Inf. Technol.*, vol. 9, no. 2, pp. 109–116, 2012.

[16] M. A. Attia, "Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation", University of Manchester, 2008.

[17] O. Nadim, T. Abeer, Moubaiddin Asma, and Hammo Bassam1, "Formal description of Arabic syntactic structure in the framework of the government and binding theory", *Computación y Sistemas*, vol. 18, no. 3, pp. 611–625, 2014.

[18] N. Khoufi, C. Aloulou, and L. H. Belguith, "Parsing Arabic using induced probabilistic context free grammar", *International Journal of Speech Technology*, Sep. 2015.

[19] S. Kulick and R. Gabbard, "Parsing the Arabic Treebank: Analysis and Improvements", 2006.

[20] "The Stanford Natural Language Processing Group", [Online Available]: http://nlp.stanford.edu/software/lex-parser.html. [Accessed: 10-Apr-2016].

[21] S Petrov. Coarse-to-Fine Natural Language Processing. University of California-Berkeley, 2009.

[22] G. Sampson and A. Babarczy, "A test of the leaf-ancestor metric for parse accuracy", *Natural Language Engineering*, vol. 9, no. 04, pp. 365–380, 2003.

[23] E. Black, Meeting of interest group on evaluation of broad-coverage grammars of English. LINGUIST List 3.587. 1992.

[24] Nivre and Johan Hall, "The CoNLL 2007 shared task on dependency parsing", in *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, 2007, pp. 915–932.

[25] Y. Marton, N. Habash, and O. Rambow, "Dependency parsing of Modern Standard Arabic with lexical and inflectional features", *Computational Linguistics*, vol. 39, no. 1, pp. 161–194, 2013.

[26] D. Chen and C. D. Manning, "A Fast and Accurate Dependency Parser using Neural Networks", in *EMNLP*, 2014, pp. 740–750.

[27] M. Boudchiche, A. Mazroui, M. O. A. O. 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*, 2016.

[28] M. Attia, M. Yaseen, and K. Choukri, *Specifications of the Arabic Written Corpus produced within the NEMLAR project*. Technical report, NEMLAR, Center for Sprogteknologi, 2005.

[29] A. Siraf, explanation of Sibawayh's Al-Kitab. *Dar Al kotob AlMasriy*a, Egypt (2000).

[30] A. Mustafa, *binding theory in Arabic grammar and Study of grammatical structure. Damascus University Journal for Literature and Humanities. 18, 41 (2002).*

[31] E. Husserl, Introduction to the Logical Investigations: A Draft of a Preface to the Logical Investigations (1913). Springer Science & Business Media, 2012.

[32] K. Ajdukiewicz, "Die Syntaktische Konnexitat. Studia Philosophica 1: 1-27; translated as 'Syntactic Connecxion'in S. McCall", *Polish Logic*, 1935.

[33] Y. Bar-Hillel, "A quasi-arithmetical notation for syntactic description", *Language*, vol. 29, no. 1, pp. 47–58, 1953.

[34] A. Pasha *et al.*, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic", in *LREC*, 2014, pp. 1094–1101.

Authors' Profiles

Nabil ababou received the Master in computer science from Mohamed First University, Morocco in 2012. He is currently working toward the Ph.D. degree in Arabic natural language processing within the Laboratory Researches in Computer Science. His research interests are especially in Arabic syntactic analysis (parsing).

Azzeddine Mazroui is a Full Professor in Mohammed First University, Morocco. He received the "Doctorat d'Etat" in numerical analysis at the Mohammed First University, 2000, and PhD in probability and statistics at the Pierre & Marie Curie University France, 1993. He is the director of the Natural Language Processing team of the Laboratory Researches in Computer Science (http://oujda-nlp-team.net/?lang=en). His main research interests include the areas of NLP and image processing.

Rachid Belahbib is a Full Professor in Mohammed First University, Morocco. He received the "Doctorat d'Etat" in Arabic Linguistic, University Mohammed First, 1993. He is member of the Natural Language Processing team of the Laboratory Researches in Computer Science, and Deputy executive director and Regional coordinator at the Doha Historical Dictionary of Arabic. His main research interests include the area of Arabic NLP.

How to cite this paper: Nabil Ababou, Azzeddine Mazroui, Rachid Belehbib,"Parsing Arabic Nominal Sentences Using Context Free Grammar and Fundamental Rules of Classical Grammar", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.8, pp.11-24, 2017. DOI: 10.5815/ijisa.2017.08.02