INTERNATIONAL JOURNAL OF SOFT COMPUTING

The Automatic Generation of Arabic Sentences Based on a Minimalist Approach

Chouaib Moukrim, Abderrahim Tragha and El Habib Benlahmer
Faculty of Science Ben M'sik, Laboratory of Information Technologies and Modeling, Hassan II Univercity, Morocco

Key words: Arabic, natural language processing, minimalist grammar generation of sentences, operations, NLP

Corresponding Author:
Chouaib Moukrim
Faculty of Science Ben M'sik, Laboratory of Information Technologies and Modeling, Hassan II Univercity, Morocco

Page No.: 68-76
Volume: 14, Issue 3, 2019
ISSN: 1816-9503
International Journal of Soft Computing

Abstract: The automatic generation of sentences is a domain of Natural Language Processing (NLP); it is placed in the middle of computer science and linguistics. This is a very complex discipline; the aim of this is to create automatically correct sentences from a list of words which can serve as the basis for such various applications such as automatic translation, question-answering systems, correcting syntactic errors and so on. In this study, we present the use of the linguistic approach of Chomsky's minimalist grammar. We begin with the elaboration of the lexicon that constitutes the essential link of the generation. Then, based on this lexicon, we treat the merge and move operations to build a syntactically correct sentence.

INTRODUCTION

There are several grammatical theories such as the Lexical Functional Grammar (LFG) and the Minimalist Program (MP) that offer a sentence structure that takes into account the linguistic characteristics of the lexicon. These two theories are born and developed within a larger framework of Generative Grammar (GG) whose initial postulate is that language belongs to the genetic inheritance of the human species. Language is made up of a number of elements that combine with universal constraints (specific to each language) in order to develop universal grammar (part of human biological knowledge).

The automatic production of text is studied as soon as the research in computational linguistics is born. In the 1950s, several studies focused on the acquisition of language, and Artificial Intelligence (AI) (Turing, 1950; Shannon, 1948), as well as the publication of "syntactic structures" (Chomsky, 1957). After this important period of researches from the 90s- a new approach remained at the core of Chomsky's efforts to construct grammars (Chomsky, 1995) has brought about a revolution in this field.

The Minimalist Grammar (hereinafter referred to as MG) is a formalization of Chomsky's minimalist program that currently covers much of the traditional syntax. It is simple and intuitive to use and slightly context sensitive. Despite the importance of MG, it has not aroused much interest among grammarians in the Arabic language, mainly due to the particularity and richness of this language such as the absence of vowels in most texts, the irregularity of the order of words in the construction of sentences, the agglutination, etc.

In this study, we try to apply this linguistic theory in the Arabic language in order to generate grammatically correct sentences from a list of tagged words.

Aspects of the Arabic language: It is assessed that about 290 million Arabic speakers are native Arabs from

27 countries in the world (Aljasser and Vitevitch, 2018). Moreover, Arabic is the language of the Holy Quran, it is used by nearly 2 billion Muslims (Khan, 2018).

Arabic, like all Semitic languages (Amharic, Aramaic, Maltese and Modern Hebrew) is characterized by the use of certain schemas (word-forming models) that make it possible to obtain words from abstract roots, representing general semantic notions or precise meanings. These roots are generally composed of three consonants which constitute the basic units for the formation of many words derived from this root "فعل". The Arabic language is divided into two main forms: a Classical Arabic (CA) that is totally diacritized, including classical historical texts, ancient literary texts, etc. There is also Modern Standard Arabic (MSA) which is the official language today, including news, official documents, etc. It is uncomplicated compared to the (CA); both forms follow a similar grammatical structure. In general, the words in the Arabic language are classified into three principal classes:

A noun in Arabic is a word that indicates a human, an animal, a place, an object or an abstract idea. Nouns are the most essential part of the vocabulary. The noun is divided into many subdivisions of different consideration, namely:

- Genders (Masculine and feminine)
- Definiteness (indefinite and definite)
- Number (singular, dual and plural)
- Derivation (underived, derived, source of derivation)
- Grammatical case endings, a noun can be nominative when it is the subject, accusative when it is the object of a verb and genitive when it is the object of a preposition

A verb is each word that indicates the existence of an action which is linked with one of the following subdivisions:

- Tense (past, present and future)
- Transitivity (intransitive, transitive)
- Moods (imperfect indicative, subjunctive, jussive and imperative)
- Voice (active, passive)

A particle is among the three classes of the word which has no meaning unless it is assembled with other classes (noun, verb). The particles are divided into operating particle "حروف عاملة" such as "إن-inna" and their sisters and non-operating particle "عاملة حروف غير" as answer particles. It is also divided into specialized for the verbs as exhortation particles "حروف التحضيض" and specialized for the nouns such as prepositions "حروف الجر" and mixed such as coordinating conjunctions "حروف العطف".

The minimalist grammar: Grammar is a set of rules that describes how lexical units can be joined to give a meaningful sentence. Different grammatical formalisms are available for natural language processing (here, in after referred to as NLP) with different perspectives. The majority of these formalisms work on the consequences for the theory of the lexicon; however, the lexical representation is distinctive for each of them. Grammatical theories now become mathematically precise in their description. Research in linguistics and NLP has led to the conclusion that lexicon plays a very important role in grammar (Stabler, 2004). The properties associated with lexical elements are important and play a central role in grammar. This idea has led to the lexicalisation of grammar that is fundamental to computational linguistics. Lexicalisation has brought a number of terms in the linguistic domain such as lexicalised grammar, lexical semantics, lexical conceptual structure, lexical functional grammar and so on.

There are several formalisms of grammar in NLP that have been designed without taking into account lexical data, yet grammarians have deduced that the lexicon plays a primordial role. Certainly, the version of Chomsky's syntax that is called the minimalist theory (Chomsky, 1995a, b) where the rules of grammar are minimal compared to the lexicon which is the most important source of information. The lexicon contains a maximum of information, leaving less room for the grammatical rules. One of the basic assumptions of minimalist syntax is that the human brain has a language ability "the language faculty" which is an autonomous system exclusively dedicated to language (Chomsky, 1995a).

Chomsky (1995) initially called this genetic faculty "Language Acquisition Device", it is a human brain mechanism that allows a child to learn his native language naturally and quickly by stages such as the recognition of sounds, then the phonological development, next the first words and the development of the vocabulary, finally the first sentences and beginning of the grammar. According to Chomsky Universal Grammar (UG) is considered a characterization of the pre-linguistic primary state of the child (Chomsky, 1962).

The main questions guiding minimalist syntax have become the theory of Principles and Parameters (P&P) and have been in existence for about 30 years. According to the theory of P&P, there are universal principles (a set of laws) that underlie the basic architecture of any linguistic system and parameters that govern the variations that this architecture could manifest.

Chomsky has eliminated much of his transformational generative theory. By abandoning the notion of deep structure and surface structure, he published his book the minimalist program (Chomsky, 1995b). The Minimalist Program is the latest version of transformational generative grammar inspired by Noam

Chomsky whose goal is to simplify the theory of syntax and the knowledge of grammatical rules. The attempt at simplification leads to the reduction of the number of levels of representation.

MATERIALS AND METHODS

The components of the minimalist grammar

The syntactic features: The MGs consist of a lexicon and structural construction operations that apply to lexical items and trees resulting from these applications. The elements of the lexicon are constituted of syntactic and non-syntactic features (e.g., phonological and semantic features) (Adger, 2003).

It is necessary to characterize these lexical items. Indeed, Collins and Stabler (2016) define a lexical item in the following way (SEM-F, SYN-F and PHON-F are universal sets of semantic, syntactic and phonological features, respectively), these three features are grouped in substantive elements. Consider the diagram given in Fig. 1 for the substantive element "a book-كتاب":

The description of the substantive element above, the phonological feature matrix (PHON-F) applies only to the component that processes the phonetic characteristics, among others, the pronunciation of the word "كتاب", which for example the phonetic representation [kitab]. It could not be represented [kitabf] where we simply ignore [f].

Semantic Feature Matrix (SEM-F) plays no role in the grammatical process because it concerns purely semantic properties, for example, the word "كتاب" is non-human.

The features related to syntax (SYN-F) determine the morphological form of the objects. These features can be defined as the characteristics that play a role in the grammatical process (morphological or syntactic). They also include categorical features such as noun, verb, etc. In addition, they include the characteristics of the number (singular/dual/plural), the gender (masculine/feminine) and the person (which play a role in the syntax of the subject-verb agreement). Grammatical characteristics also include case endings such as nominative or accusative cases and so on. We are interested in the syntactic characteristics that are basic categories (Table 1), namely:

The selection features: The process of constructing a sentence starts with the selection of a set of elements chosen from the lexicon. These elements are "syntactic head" which can be selected by other linguistic elements. This is encoded by syntactic features, called selectors. Table 2 gives some examples.

The selection of the functional elements needed to verify features depends on the nature of the substantive elements selected in the lexicon. Certainly, the substantive categories are selected by functional categories. For example, if a verb is inflected for the past, the computational system must select the corresponding functional line in order to check the tense of the verb. Successful verification means that the relevant feature of the functional category is deleted.

كتاب-book

$$\begin{bmatrix} SEM-F \\ +concrete \\ +count \\ -animate \\ -human \\ . \end{bmatrix} \begin{bmatrix} PHON-F \\ [\text{kitab}] \\ . \\ . \\ . \end{bmatrix} \begin{bmatrix} SYN-F \\ N \\ =V \\ . \\ . \\ . \end{bmatrix}$$

Fig. 1: Schematic lexical input for "كتاب"

Table 1: The syntactic features

Features	Designation
D	Determinant
V	Verb
N	Noun
P	Particle
T	Tense

Table 2: The selection features

Selectors	Designation
=D	Selector of a determinant
=V	Selector of a verb
=N	Selector of a noun
=P	Selector of a particle
=T	Selector of a tense
…	…

The operations: The construction operations of the MG tree are merging and moving. They use the syntactic features to generate a well-formed sentence for structural trees.

The MG defines a precise formalization of the basic ideas of Chomsky's minimalist program. In order to simplify the notation, we can also talk about the (unique) feature of a tree which is the first syntactic feature of a tree's header list.

The merge operation: The most basic operation of grammar is therefore, the mechanism governing the selection. In an MG, this mechanism is the merge operation (Collins and Stabler, 2016) that uses two syntactic objects α and β and connects them into a single syntactic object γ:

$$\text{Merge}(\alpha, \beta) = \{\alpha, \beta\} \quad (1)$$

For any syntactic object α, β where α is a non-empty selection list $L = \{A_1, ..., A_n\}$ such that $A_1, ..., A_n$ are selection features and β is a categorical feature:

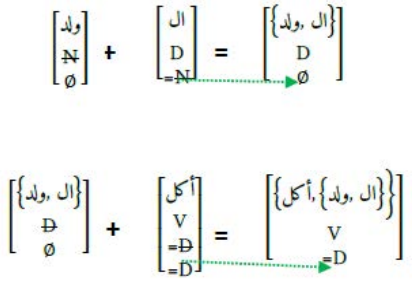


Fig. 2: An example of two merge operations

- α is the header of the tree
- α = {γ, {α−L, β}} γ is the projection of α
- if n >1 then L = {A_2, ..., A_n}, else L = ∅
- $\gamma = \left[\begin{smallmatrix} \text{CAT}[\text{cat}(\alpha)] \\ \text{SEL}[L] \end{smallmatrix}\right]$

Selection is the most basic syntactical relationship. The capture of dependencies between morphemes, words and sentences is the central task of any syntactic theory, the difference between these units is the basis for linguists to pose selection features. The merge operation adds trees (lexical elements) using categories and selectors, as shown in Fig. 2.

The features of the noun "ولد" are:
- Category: N
- Selector: ∅ (it has no selectors)

The features of the determinant "ال" are:
- Category: D
- Selector: = N (it selects a noun)

The features of the verb "أكل" are:
- Category: V
- Selector: = D (it selects a defined noun)
- Selector: = D (it selects another defined noun)

In this example the first merge operation (1) is between the name "ولد" with the determinant "ال", as a result, we have obtained "الولد" a defined noun with a new category "D" without any selector, we say: a projection relation of the determinant on the noun.

The second merge operation (2): the verb "أكل" selects "الولد" as a defined noun, producing the merged tree of the sentence "أكل الولد". The verb "V" is projected immediately on the defined noun "D". Therefore, the verb is on the header of the tree describing a verbal sentence. After merging, the features of the two trees are removed (Fig. 3).

We notice that there remains a selector in the header of the tree "= D", certainly this verb is among the normal

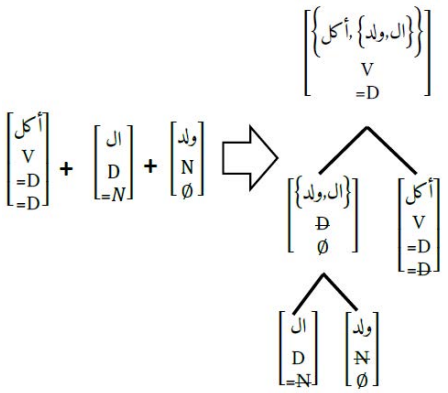


Fig. 3: Syntactic tree of two merge operations

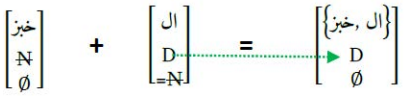


Fig. 4: The third merge operation

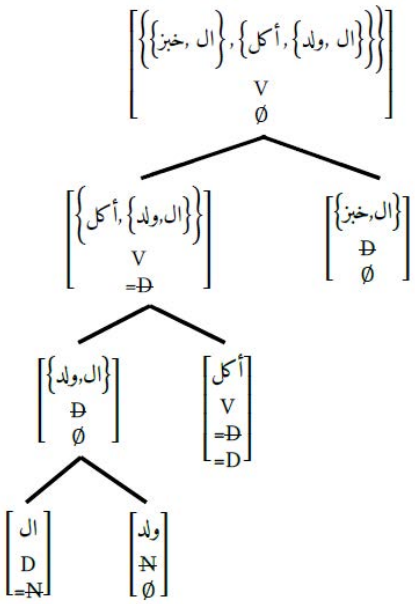


Fig. 5: The result of merging "أكل الولد الخبز"

verbs in the Arabic language which requires a subject and a complement, here, the sentence is incomplete and it lacks a complement (defined noun).

We can apply this approach to correcting syntactical errors when a sentence is incomplete, so, we add a definite noun to complete the sentence: The merging of the phrase "أكل الولد الخبز- 'The boy ate the bread'" becomes as follows (Fig. 4 and 5):

Table 3: The different lexical features

The words/ The features	يضرب Hits	الولد The boy	الكرة The ball
The header features	[present]	[D]	[+Acc]
The selection features	[=D]	⌀	⌀
Complementary features	[-Acc]	⌀	⌀

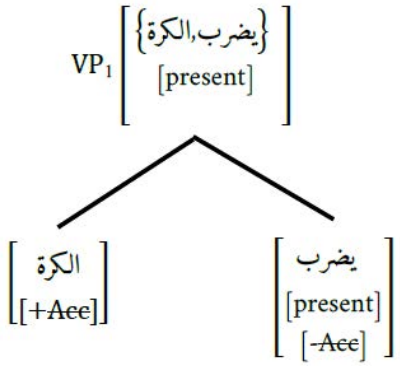


Fig. 6: The PV_1 merger

The move operation: In the minimalist syntax, the move operation is driven by the morphology. This derivation operation proposed by the MP consists of feature checking, it is made necessary by the condition of full interpretation which governs the phonetic and semantic levels and which requires that any feature present at either of these two interfaces be interpretable.

Sometimes there are elements that contain only non-interpretable features; in this case, they can be checked if they come into contact with lexical elements that contain the same lexical elements with the same functional features. The constituents are moving into functional categories so that the non-interpretable features of these functional categories can be erased by the association of the interpretable features that correspond to them. The moving mechanism is illustrated by the following example:

"يضرب الولد الكرة"
The boy hits the ball

The derivation of this sentence begins with the selection of substantive elements, the verb "يضرب" as well as the nouns "الولدُ" (subject) and "الكرةُ" (object), each fully inflected by its particular morphological features (tense and agreement). These elements have the following characteristics (Table 3). The noun "الكرة" is merged with the verb "يضرب" to form PV1 (the verbal phrase). The noun "الولد" is then merged with PV_1 to form PV_2 (Fig. 6 and 7).

The verification of the features in the derivation (Fig. 8) implies that the specifying feature [= D] of the verb "يضرب" is checked in relation to the features of the header [D] of "الولد". The header feature of "الولد" plays

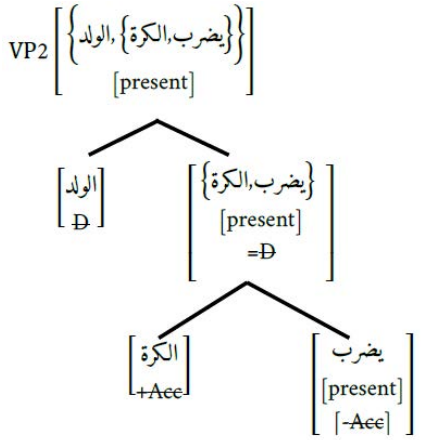


Fig. 7: The PV_2 merger

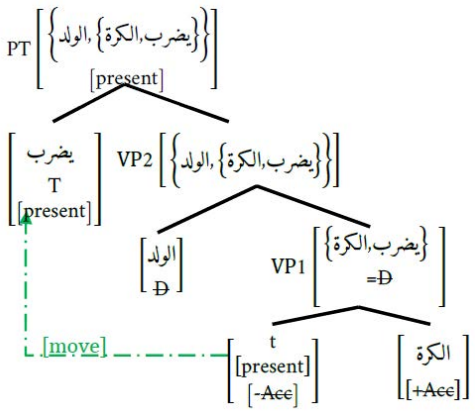


Fig. 8: The PT move

a role in the semantic part. However, the specifying feature [= D] of the verb "يضرب" does not play any role in the semantic interpretation because it just indicates that the verb must correspond to its subject. Therefore, the specifying feature [= D] of the verb "يضرب" is deleted.

The complementary feature [-Acc] of the verb "يضربُ" is checked in relation to the header feature [+Acc] of the noun "الكرةُ" which is defined by the accusative case ending. A perfect match found and both [+Acc] and [-Acc] features are removed.

The only feature that has yet to be checked is the feature [present] of the verb "يضرب". To perform this check, the VP2 is merged with the T header, carrying the tense feature which gives the phrase category PT (a phrase defined by a tense T). The verb "يضرب" is then moved to this check position where its tense feature is checked.

Merging and moving are universal operations which should give uniformity in languages and especially in the Arabic language. This uniformity exists only in the semantic part. In fact, languages vary in their structure.

The lexicon: Accepting that the organization of the lexicon is an essential step in the whole process of generating sentences. Indeed, the lexicalisation of grammar specifies a basic role to the lexicon.

In order to achieve this objective, it seems necessary to translate the minimalist grammar data in terms of a structure in the form of a quadruplet MG = (Σ, LA, GF, OP). Such us MG denotes the minimalist grammar, Σ is the vocabulary, GF is a set of grammatical features, LA is a finite lexical array and OP operation. The word x in the MG can be expressed by the following equation:

$$(\forall x \in \sum) \text{for } i \geq 1$$

- $x = [x^1,\ldots,xi]$ such x1, …, I∈LA
- $x^i \in GF = \{\langle \text{Sem-F, Syn-F, Phon-F}\rangle, i\rangle\}$ where "i" is its index
- $x_i = \{\text{merge, move}\}$

A lexicon is a finite collection of lexical objects which is composed of either syntactic, semantic or phonological objects such as SEM⊆SEM-F, SYN⊆SYN-F and PHON∈PHON-F. Example: Let the word (أكل-ate)∈Σ

أكل [أَكَلَ، أُكِلَ، أُكْلٌ]
أكل[1]←⟨Sem-F[i]={[-human];[-animate];… }, Syn-F[i] = {N; case[nomin]; gender[m]; nbr[p]; …},Phon-F[i]={}⟩
أكل[2]←⟨Sem-F[i]={ tense[passive voice]; gender[m];… }, Syn-F[i]={V; =D; …}, Phon-F[i]={}⟩
أكل[3]←⟨Sem-F[i]={ tense[past]; gender[m];… }, Syn-F[i] = {V; =D; …}, Phon-F[i]={}⟩

In the Arabic language, ambiguity exists between the principal grammatical categories. Furthermore, a non-vocalized word can appear as a verb, noun or particle (e.g., أَنْ، أَنَّ، إِنَّ and أَنْ).

Unfortunately, there are no previous works (that, we have known up to the present) on the conception of a lexicon (in all languages) containing all the inputs that are syntactic, semantic and phonological. For Arabic which is a Semitic language, its words are lexically very close to each other with an editing error. Indeed, the average number of neighbouring forms is 26.5% for Arabic, 3.5% for French and 3% for English (Zribi and Ahmed, 2003).

These clues allow us to arrive at two deductions, the first is that the risk of making an error will be greater in Arabic than in other languages, the second is that the size of the list of lexical objects will be very large. We have used the class diagram to model the minimalist grammar of the Arabic language (Appendix A).

RESULTS AND DISCUSSION

In this study, we present the result of an experiment for the automatic generation of a syntactically correct sentence from the following tokens (Table 4). We have developed the merge algorithm as follows (Appendix B). Figure 9 shows the result of this algorithm.

We have put all lexical elements into one-dimensional arrays with their grammatical and selector features. In a first step, the algorithm looks for the element that does not have any selector feature; in this example, there are two, namely: "ولد'boy'" and "خبز" 'bread'". These two elements can be selected by the two determinants "ال". Then the verb "أكل" selects each category of the head is "D". Finally, we get the following sentence:

{{أكل ,{ال ,ولد}} ,{ال خبز}

We notice that the sentence can be generated otherwise if the input "خبز 'bread'" comes before "ولد 'boy'", indeed the sentence generated could well become:

أكل الخبز الولد-The boy the ate bread the buy

Even if the sentence is syntactically correct, it is wrong at the semantic level. In order to remedy this

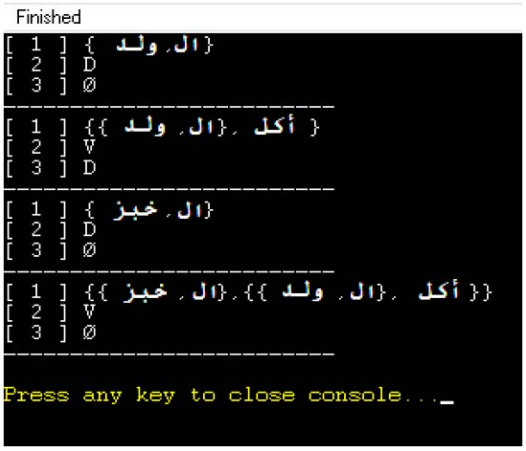


Fig. 9: The result of the merging algorithm

Table 4: Lexical inputs

Lexicon words	The syntactic features	The selection features
أكل	[V]	[=D] [=D]
ال	[D]	[=N]
ولد	[N]	∅
ال	[D]	[=N]
خبز	[N]	∅

problem, other features that have semantic roles ([+ human], [+ animated], [+ abstract], etc.) are necessary for interpretation need to be added to the lexicon.

CONCLUSION

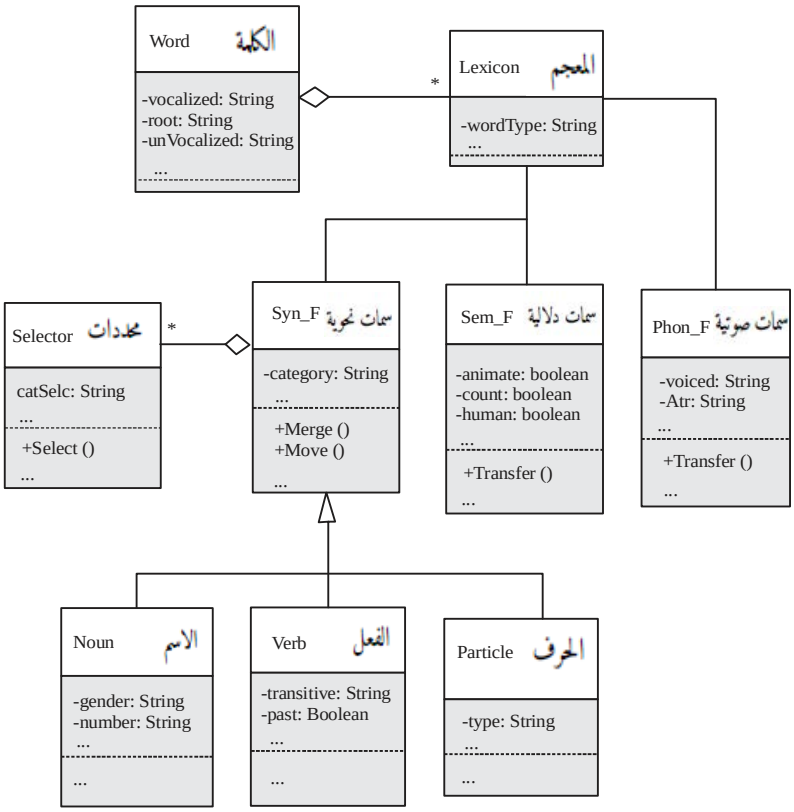
The evolution of minimalist grammar is in favour of the main role of development of a lexicon which provides all the necessary information may either be phonetic, semantic or syntactic. Indeed, the use of this approach in the Arabic language requires an important work in the elaboration of a dictionary containing all the lexical input relating to grammatical, selection or complementary features, which opens up a major research project in Arabic Natural Language Processing (ANLP).

ACKNOWLEDGEMENT

I would like to express the very thanks to my PhD supervisor, Professor Abderrahim Tragha form Hassan II University who gave me the opportunity to do such research as well as all the members of the laboratory of Information Technologies and Modeling (LTIM).

APPENDIX

Appendix A: The class diagram of the Arabic minimalist grammar



Appendix B: Merging algorithm of the sentence "أكل الولد الخبز"

```
INPUT: w1←["أكل","V","D","D"] /* Array of the word "أكل" */
       w2←["ال","D","N"]
       w3←["ولد","N","ø"]
       w4←["ال","D","N"]
       w5←["خبز","N","ø"]
       s←[word1,word2,word3,word4,word5] /* Array s : "sentence" a 2D array */
       i←1,j←1
```

```
INPUT: w1←["أكل","V","D","D"] /* Array of the word "أكل" */
       w2←["ال","D","N"]
       w3←["ولد","N","ø"]
       w4←["ال","D","N"]
       w5←["خبز","N","ø"]
       s←[word1,word2,word3,word4,word5] /* Array s : "sentence" a 2D array */
       i←1,j←1
BEGIN
  FORALL the size(s)>1 DO
    IF (s[j][3]="ø") THEN /*we seek the words that have no selector */
     FORALL the size(s)>i DO
      IF(s[j][2]=s[i][3])THEN /*if selector = category */
        IF(size(s[i])=4)THEN /*the phrase size = 4 words */
         s[i][1]←"{"+s[j][1]+","+s[i][1]+"}" /* merge */
         s[i][3]←s[i][4]
         DESTROY s[i][4]
         ELSE  /* the phrase size = 3 words */
              s[i][1]←"{"+s[j][1]+","+s[i][1]+"}"
              s[i][3]←"ø"
        ENDIF
        FOR k = 1 TO size(s[i]) DO
           WRITE "[",k,"]",s[i][k]
        ENDFOR
        WRITE "-------------------------"
        DESTROY s[j]
      ENDIF
      i←i+1
     ENDFOR
     j←0
    ENDIF
    j<-j+1
  ENDFOR
END
```

REFERENCES

Adger, D., 2003. Core Syntax: A Minimalist Approach. Oxford University Press, Oxford, UK., ISBN:9780199243709, Pages: 440.

Aljasser, F. and M.S. Vitevitch, 2018. A Web-based interface to calculate phonotactic probability for words and nonwords in Modern Standard Arabic. Behav. Res. Methods, 50: 313-322.

Chomsky, N., 1957. Syntactic Structures. Mouton Publ., Paris, France.

Chomsky, N., 1962. Explanatory Models in Linguistics. In: Logic, Methodology and Philosophy of Science, Nagel, E., P. Suppes and A. Tarski (Eds.). Stanford University Press, Stanford, California, USA., pp: 528-550.

Chomsky, N., 1995a. Language and nature. Mind, 104: 1-61.

Chomsky, N., 1995b. The Minimalist Program. MIT Press, Cambridge, Massachusetts, USA., ISBN-13: 9780262531283, Pages: 420.

Collins, C. and E. Stabler, 2016. A formalization of minimalist syntax. Syntax, 19: 43-78.

Khan, M.I., 2018. Revival of bio medical research in the Muslim world. Int. J. Human Health Sci., 2: 5-7.

Shannon, C.E., 1948. A mathematical theory of communications. Bell Syst. Tech. J., 27: 379-423.

Stabler, E.P., 2004. Varieties of crossing dependencies: Structure dependence and mild context sensitivity. Cognit. Sci., 28: 699-720.

Turing, A.M., 1950. Computing machinery and intelligence. Mind, 59: 433-460.

Zribi, C.B.O. and M.B. Ahmed, 2003. Efficient automatic correction of misspelled Arabic words based on contextual information. Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, September 3-5, 2003, Springer, Berlin, Heidelberg, Germany, ISBN: 978-3-540-40803-1, pp: 770-777.