MoGAR: Morphological Generator for Arabic Language using Rule-based Generation Process

Ahmed Benfatma
Mathematics and Computer Science Department
Ahmed Draia University
POBOX 01000 Adrar, Algeria

Mohamed Amine Cheragui
Mathematics and Computer Science Department
Ahmed Draia University
POBOX 01000 Adrar, Algeria

ABSTRACT
The automatic generation of text (also known as the automatic generation of textual resources) consists in initially producing words and sentences with meaning. Based on parameters generally derived from other phases of processing, such as analyzing the translation process.

The aim of our work is to contribute to the development of the Arabic language processing, by proposing a technique of generation of words (verbs and derivable nouns) and sentences based on the use of variables (features). The latter may have morphological traits (gender, number, voice, etc.), or syntactic traits (structure of Arabic sentence), so the originality of this work lies mainly in the identification of the different features which can Influenced on the process of generation but also to find a kind of cohabitation between these traits to lead to a correct generation.

General Terms
Natural Language Processing (NLP).

Keywords
Arabic language, Morphology, generation, Root, Pattern, derivation, inflexion.

1. INTRODUCTION
Arabic is a rich language morphologically [1], [2]. However, if this wealth is considered by our linguistic ancestors as a source of pride, it can be considered today by researchers in Arabic language processing as the focal point of several sources of problems (segmentation, ambiguity, Agreements, etc.).

Morphology is a domain of language that allows the description of the rules governing the internal structure of words (lexical units), in the grammarians morphology is the study of the forms of words (flexion and derivation), in other words, Morphology is the study of words considered isolation (out of context) under the double aspect of nature and the variations that they can undergo [3]. In the Arabic language, morphological analysis is all the more important as words are strongly agglutinated, that is to say, they are formed in their majority by assembling elementary lexical and grammatical units [4].

When we talk about morphology in natural languages processing, we speak of two opposing variants, namely: the analysis which aims mainly at describing a word through a set of morphological traits (time, voice, Gender, number, ... etc.).

The generation that relies on these same morphological features to give birth to words, that can be used to elaborate sentences in the case of language learning or machine translation.

The purpose of this article is to shed light on one of these variants, namely the morphological generation of Arabic words, or the originality of this work lies in the setting up of a generation tool specific to the Arabic language Adopting a variable-based approach (rule-based approach) in addition to its, our system allows to exploit this generation to elaborate simple sentence for the Arabic language.

This article is organized around seven points, an introduction setting out the working context, and then a state of the art on Arab morphology with an emphasis on the generation process. Then, an overview of the works related to the morphological generation specific to the Arabic language, will come later a detailed description of the general architecture of our generator that we named MoGAR, followed by a presentation of different interfaces of our generator as well as the various tests and results obtained, to finish with a conclusion.

2. ARABIC MORPHOLOGY
By its morphological and syntactic properties, the Arabic language is considered a language difficult to master in the field of automatic language processing [5], [6]. It is thus essential to understand the different mechanisms and concepts (basic lexicon) related to the Arab morphology, in order to develop a system of morphological generation

2.1 Basic concepts
The Arab morphology is based on four basic concepts, namely:

- Root: nominal or verbal is a sequence of consonants considered by linguists to be the basis for a whole family of words attached to that root. (Usually composed of 03 or 04 radical letters).
- Pattern (Schema): is a predefined form that characterizes a class of verbs or nouns, it represents the mold in which the root is made to get the derived words. By taking advantage of the regularity of the Arabic language, the Arab grammarians have thought of representing each class of verbs or nouns by the root "فعل" or "فعلل" which is called pattern. [7]:
- Agglutination: In Arabic a word can mean a whole sentence because of its compound structure which is an agglutination of elements of grammar. In other words, the consonant part of the word is decomposable into five elements: proclitic, prefix, lexical basis, suffix and enclitic.
- Particles: in the Arabic language is an invariable word that accompanies a verb or a noun, and cannot convey any meaning when they are isolated. Among the particles, some are used with verbs, others with nouns and others with the noun and the verb.

2.2 Inflectional morphology

Arabic is an inflectional language. It uses, for the conjugation of the verb and the declination of the noun, indices of appearance, mode, time, person, gender, number and cases, which are generally suffixes and prefixes [4],[8]. Usually, these flexional marks make it possible to distinguish, by way of example:

- Time value: determines the time of the action, characterized by the three aspects (accomplished, unfulfilled and imperative).
- The person value: determines the actor of the fact, it is specified by the three categories of persons (the enunciator, interlocutor and the absent person) masculine or feminine, singular, dual or plural.
- The modal value: The mode denotes the way in which the action expressed by the verb is conceived and presented, it expresses the voice (active or passive).
- Declination of Nouns: Declination is the transformation that can be exemplified by a singular feminine noun, dual or plural, it applies only to the inflectional nounes. The declension of a noun has three cases, it can be: "مرفوع" (marfoue - nominative), "منصوب" (mansoub - accusative) and "مجرور" (majerour - genitive).
- Nature of the nouns: According to its nature a noun can be masculine or feminine.

2.3 Derivational morphology

The derivational morphology studied the formation of new words. It is characterized by the creation of new words having a different meaning by associating prefixes, infixes or suffixes with a given root or changing the deratting of the letters composing it. Arabic words entering into derivation frame are nouns (الأسماء المتمكنة) and verbs (الأفعال المتصرفة), which may be simple or augmented. This mechanism consists in the juxtaposition of the root with the pattern associated according to the need, that is to say to replace the letters of the pattern by the consonants of the root in question while keeping the same vowels and the same letters augmented with The respect of the order of consonants.

Among the derived words are:

- The verbal noun: is an abstract noun formed on the same root as the verb to which it is associated and expresses the same semantic content as it; however, it does not imply any notion of time, aspect, modality, person, or even voice. Semantically, it expresses an action, a state or a process according to the meaning of the verb to which it is associated.
- The active participle: is a noun associated with any verb of action (transitive or intransitive), and which designates the agent of the verb that is to say the one who does the action.
- The passive participle: is a noun associated with any transitive action verb. It refers to the patient who undergoes the action or the result of this action.
- The place (or time) noun: The place noun is supposed to designate the place where the action occurs.
- Instrument noun: is a noun that designates the instrument used to execute the action expressed by the verb.

3. RELATED WORK

In the literature proper to the work on Arab morphology, it is found that most research and applications are oriented towards morphological analysis; however, those on morphological generation are rare [9]. We can cite the works of Beesly [10], [11] its Xerox Arabic generator system, which relies on state machines, but this approach is somewhat limited in that it treats the generation process at a superficial level. We can also talk about the work of Cavalli-Sforza [12] and Habash [13] on the implementation of a rule-based generation process that better adapts to the morphological characteristics of the Arabic dialect takes as the input entity the elementary form of a morphological unit as: the root and using the appropriate morphological traits can generate panoply of words. Lately, we have seen systems combining both analysis and generation, such as: the works of Khaled Shaalan [14] and the MAGEAD system [15], and ALMORGEANA (Arabic Lexeme-base MORphological GEnerator / ANAlyzer)]. It should be noted that in our MOGAR generation tool we adopted a rule-based approach (which we also called variables-based).

4. MOGAR ARCHITECTURE

Our system consists of two phases of treatment. The first deals with the pre-processing of the input term, the second and the generation phase of the various grammatical categories. Figure 1 illustrates the principle of our generation algorithm.

4.1 Pretreatment phase

This phase is necessary for the generation process, word processing, goes through two levels of analysis; the level of validation and the level of identification.

4.1.1 Validation level

Validation is carried out in three chained processing steps. The first step consists in the test of belonging to the letters constituting the Arabic alphabet; the second step is the verification of the formation of the word, that is to say, if the word entered is written according to the Criteria or the laws of writing of the Arabic language. The third step is that of the segmentation of the vocable to generate the radical letters.

4.1.2 Identification Level

This level has as objective the identification of the validated root, it consists in determining whether it is: trilateral <Tr> Primitive <Pr> or augmented <Au> or quadrilateral <Qr> primitive or augmented, this treatment is done By analyzing these structural characteristics that define it. By the determination of the number <N> of the letters constituting it, their types (Radicales <Rd> or Augmented <Ag>) and their nature (Sain <Sn> or Defectous <Df>) for the primitive trilateral root.

4.2 Generation Phase

The generation process is the direct result of synchronization between linguistic databases, a generation algorithm. It is important to specify that the result of the generation is the consequence of the set of morphological (for words) or syntactic (for sentences) traits that the user introduces into the variable inspector.

4.2.1 Linguistic Databases

The principle of this organization is to represent all the morphological components of the Arabic language in the form of objects. These objects are organized in a set of classes.

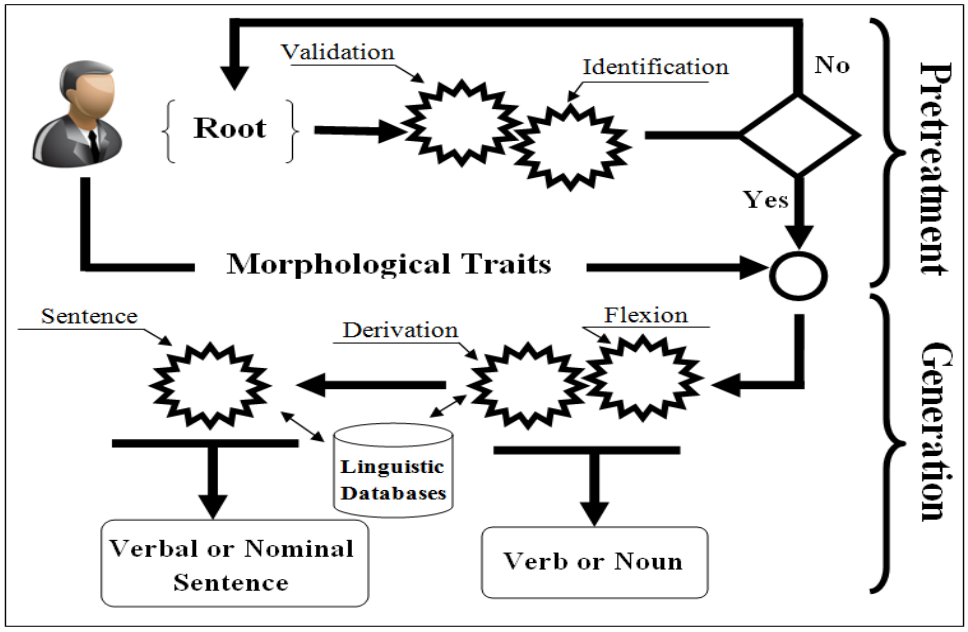

Fig 1: General Architecture of MoGAR

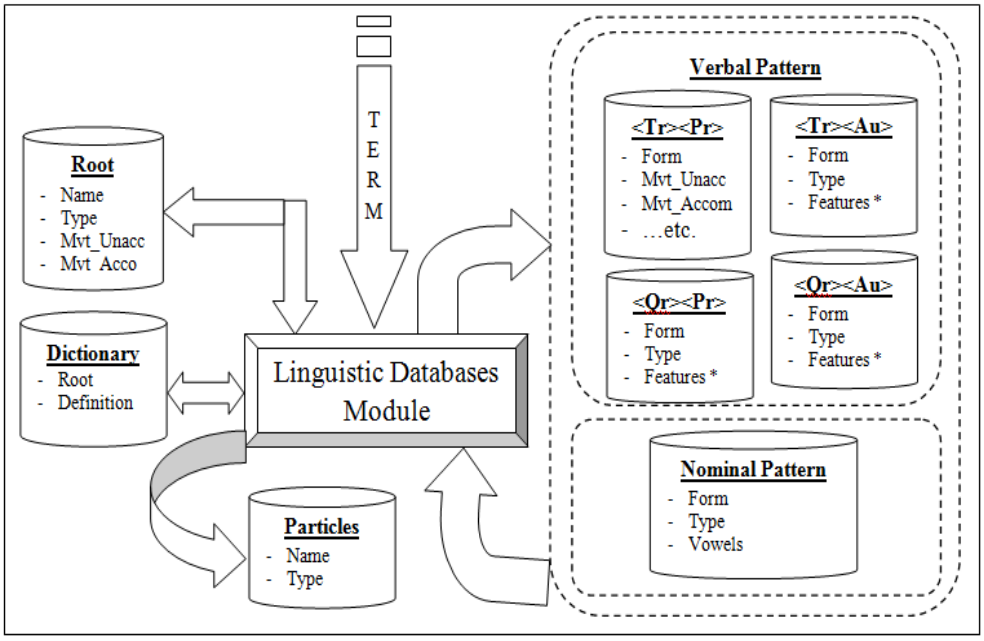

Fig 2: General architecture of linguistic databases.

Each class represents a family of linguistic data of the same nature and morphological characteristics. Such as the patterns of verbs, patterns of derived nouns, particles, affixes and particular nouns. Each of these classes defined by a set of morphological characteristics; indicating gender, number, sex, person, etc. We call these characteristics the morphological features (see Fig 2).

4.2.2 Operation of the base

The management of the database is dynamic in our system; this flexibility allows the power supply and operation of the database at all times and permanently. This manipulation is done in two ways: the first can be done by the user, it concerns only the roots, this operation is automatically cleared at the beginning of the generation process in case the root does not belong to the root, Set of generator roots. An update module and planned for this task. The second is the responsibility of the administrator because of the sensitivity and value of the information in the system and then it concerns the manipulation of the patterns using an interface that facilitates the feeding of the database by Nominal forms and paradigms of verbal inflection.

Although the database has not been filled, a root dictionary accompanies the user throughout its generation. This dictionary is a digitization of the root dictionary named " تاج العروس من جواهر النفوس".

To simplify the management function of our database, we created 03 modules, which are:

- Root management module: This manages the input of the words according to their grammatical categories, one accesses directly to the base of the generator for its update. It should be noted that this module is interactive in reading mode accompanying the user throughout its generation for all grammatical categories.
- Pattern management module, which manages the entry of patterns according to the type of the roots, accessing directly the base of the forms of the generator for its update. There are five update sub modules. Four for the different types of verbal roots and the fifth concerns the nominal patterns.
- Particle management module: as the particles in the Arabic language are reduced in number, they are predefined in the system. Because of the multitude of use of these particles, this module allows the addition of examples of their uses.

4.2.3 Generation of verbs (Flexion)

The process of bending is an essentially grammatical function. It is marked by the verbal morphological features that are: time or aspect, way, gender, person and number. It does not produce new words but rather grammatical variants of a word appropriate to a syntactic context. In other words, the conjugation of a verb is the form that can take this verb to the active and passive voice. This form varies with mode and appearance (accomplished, incomplete and imperative). It also varies with the person or persons represented by the subject. The bending process is divided into five (05) processing steps which are as follows:

- Choice of the basis of the forms (P01): the choice of the database is based on the nature of the identified root (primitive trilateral, augmented trilateral, quadriliter primitive or quadriliter augmented).
- Selection of the corresponding bent form (P02): depending on the set of variables (aspect, path, person, number and gender), the associated pattern is selected from the database of verbal bending patterns, such as the variable:
  - Aspect = {accomplished, unaccomplished, imperative} = {A, U, I};
  - Channel = {active, passive} = {A, P};
  - Person = {speaker, interlocutor, absent} = {S, I, A};
  - Number = {singular, dual, plural} = {S, D, P};
  - Genre = {male, female} = {M, F}.
- Application of the principle of conjugation (P03): the transformation of the letters of the pattern by those of the root.
- Application of the bending rules (P04): if the verb contains HaM'Zat or a semi consonant, the corresponding rules will be applied.
- Formatting of the final bent form (P05): since the vowels sound of the characters to be concatenated with the consonants. This operation often gives bad forms spelling, the reason why this step is added at the end of the process.

  Example: "ل" + fatha + "ا" gives "لا", after formatting will be "لا". It should be noted that the imperative is only conjugated to the 2nd person.

4.2.4 Generating nouns

The system of generation is a semantic function allowing the formation of new words from the already existing words; the latter is due to the modification of the semantic identity of the base word by affixation. Our system consists of two nested processes, derivation and declination.

- Declination of nouns: The process of declination is the transformation from a singular to a feminine noun into a dual or plural, or the permutation of the latter between determination or indetermination its loss of the general meaning of the word. This process is accomplished by: adding, deleting or changing letters called augments. To illustrate this mechanism, let's analyze the following example (See Table 2).

Table 1. Example of Nouns Declination

Singular (مفرد)		**Dual (مثنى)**		**Plural (جمع)**	
M^{asculine} (مذكر)	F^{eminine} (مؤنث)	M^{asculine}	F^{eminine}	Intact (سالم)	
				M^{asculine}	F^{eminine}
مدرس	**مدرسة**	**مدرسان**	**مدرستان**	**مدرسون**	**مدرسات**

- Noun derivation: The derivation process is the creation of a new noun from a root by affixation (prefix, infixation and suffixation), according to a defined mechanism and forms (Table 3). To do this, it accesses by the identifier of the noun type request in read mode to import the corresponding form of the basis of the nominal patterns. Next, by applying the derivation mechanism generates the base form of the noun. Then, it applies the affixation procedure to generate the final form of the noun.

Table 2. Example of Noun Derivation

Noun	**Root**	**درس**	**علم**	**علّم**
		د ر س	**ع ل م**	**ع ل ل م**
Iinfinitive	**إسم المصدر**	**دِراسَةٌ**	**عُلومٌ**	**تَعْليمًا**
		a study	Sciences	Educated

- When Arabic linguists and grammarians want to indicate a verb, they indicate its inflected form in both aspects (accomplished, uncompleted) and its action noun, such as " درس- يدرس- درس. " The reason why the MoGAr generates automatically without taking into consideration the choice of the user, the action noun and the bending of the verb introduced to the unaccomplished.

4.2.5 Generating the sentence

The sentence is a set of syntactic elements (nouns, verbs, adjective, etc.) organized and structured to convey an idea or a message. In the Arabic language, there are two types of sentences, verbal and nominal.

- Verbal sentence: usually begins with a verb, it can contain the verb, the subject and the object complement (s) that are highlighted, in which case the verb agrees in terms with the subject but not in number. Such as, "حضر الطالب البحث" (the student prepared the lecture), "حضرت الطالبة البحث" (the student prepared the lecture) and " حضر الطلاب البحث" (the students prepared the lecture). When the subject is placed before the verb, the latter agrees in

kind and in number with the subject. Such as, " الطلاب حضروا البحث" (the students prepared the lecture) and when the subject is not expressed, it is included in the verb as: "حضرن البحث" (they (students) prepared the lecture).

- The nominal sentence: it starts with a noun and is represented by two elements, the theme (المبتدأ) and the subject (الخبر).

The order of the words in the simple Arabic sentence is relatively flexible, giving a free choice to emphasize its components. The same sentence can be ordered as follows:

- Verb + subject + complement, such as, " حضر الطالب البحث";
- Subject + Verb + complement, such as, "الطلاب حضروا البحث";
- Complement + Verb + subject, such as, " الـ بحث حضر من الـ طلاب".

4.2.6 Mechanism of generation of a sentence

To construct a sentence, it is to plan its components according to syntactic rules governing its good formation. This operation is divided into two processing processes.

- Generation process: it consists of piloting the three (03) processes of generation (flexion, derivation and declination) on the one hand, and querying the base of the particles and selecting a tool word on demand (Case of a simple sentence with a compound complement).

  Example: "حضر الطالب البحث": simple sentence with simple complement; « حضر الطالب للجامعة": sentence with compound complement.

- Structuring process: according to the rules of formation, the structure of the sentence is metamorphosed according to the choice of the user. For our generators, these choices are limited to a simple sentence, they are counted in three (03) possible choices (VSC, SVC and CVS).

5. PRESENTATION OF MoGAR

Our generator consists of five (05) main screens, three-generation interfaces, an interface for updating the database, and one for statistics.

Fig 4: GUI for verbs generation

Fig 5: GUI for Nouns generation

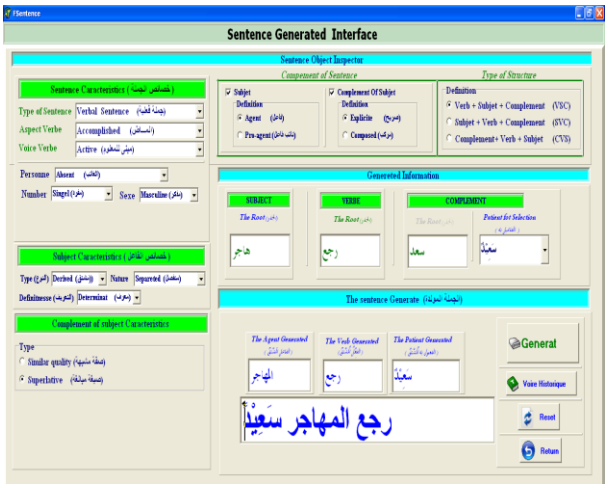

Fig 6: GUI for Sentences Generation

6. TESTS AND RESULTS

6.1 Statistics

Our database is the result of a long research work on the canonical forms of Arabic words as well as patterns.

Table 3. Number of Roots

	Trilateral		Quadriliter	
	Simple	Augmented	Simple	Augmented
Number of Roots	3659	269	976	30
Total	3928		1006	

Table 4. Number of Verbal Patterns

	Trilateral		Quadriliter	
	Simple	Augmented	Simple	Augmented
Number of Patterns	186	660	62	316
Total	846		378	

Table 5. Number of nominal patterns

Noun	Number of Patterns	Total
Action	61	236
Agent	35	
Patient	24	
Subjects	31	
Intensity	50	
Preference	10	
Time and place	08	
Instrument	17	

6.2 Results

To evaluate the performance of our MoGAR generator, we carried out a series of tests, the results are given in the tables and graphics below.

6.2.1 Case 1: Verbs (intact)

Table 6. Tests and results of verbs generation

Tests	Roots	Correct generation	Percentage
1	540	499	92,41%
2	700	650	92,86%
3	600	575	95,83%
4	450	435	96,67%
5	875	865	98,86%
6	910	900	98,90%

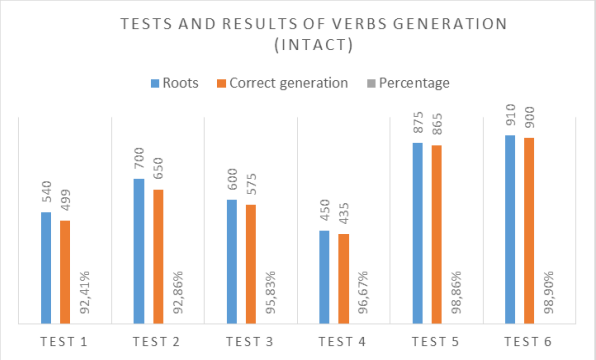

Fig 7: Graphic representation of verbs (Intact) generation.

6.2.2 Case 2: Verbs (Weak)

Table 7: Tests and results of Weak verbs generation

Test	Roots	Correct generation	Percentage
1	112	80	71,43%
2	80	59	73,75%
3	75	60	80%
4	90	77	85,55%
5	105	80	76,19%
6	100	79	79%
Average			77.65%

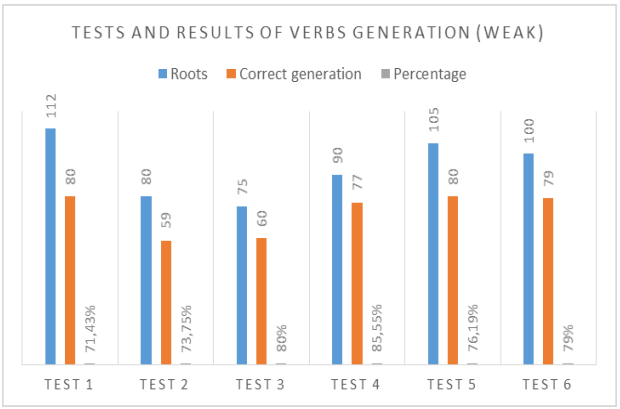

Fig 8: Graphic representation of verbs (Weak) generation.

6.2.3 Case: Nouns

Table 8. Tests and results of Nouns generation

Tests	Roots	Correct generation	Percentage
1	540	400	74,07%
2	700	550	78,57%
3	600	400	66,67%
4	450	400	88,89%
5	875	800	91,43%
6	910	850	93,41%

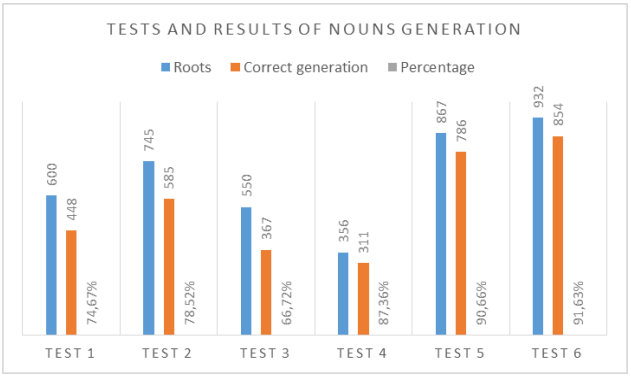

Fig 9: Graphic representation of Nouns generation.

6.3 Discussions

Among the problems encountered in this work are the multitude of derivation patterns of the different types of nouns derived from primitive trilateral roots. This complexity is due to the nature of these verbs (transitivity), the pronunciation in both aspects and the sense of the verb. For example, the nouns of action of primitive trilateral roots have a large number of forms that are all empirical, and the most used ones are enumerated has 15 forms, nouns of similar quality possess 17 empirical forms.

The production of a text or a sentence is a linguistic act and therefore an integral part of the domain of linguistics. The generation of sentences must be both syntactically and semantically correct. Even if the generated sentence structures are simple sentence structures, they are still quite correct.

7. CONCLUSION

The Arabic language has very specific morpho-syntactic phenomena. This particularity is linked to its flexional and derivational morphology, but also to the multiplicity of its forms and the high precision of meaning by its marks.

The creation of this generator was intended to generate the two grammatical categories of the Arabic language (Verbs and Nouns), in addition to the production of simple, syntactically correct sentence. Its realization required an in-depth study of the different morphological characteristics (features: gender, voice, aspect, person, etc.) and syntactic (SVO, VSO and CVS) of the Arabic language.

From the obtained results, it can be said that our generator MoGAr produced a quality generation that is for the verbs and the differentiable nouns, this performance is translated by the rate of generation which is of the order of: 85%. In terms of the sentence, the effort we made enabled us to get on the right path for a more advanced generation.

To conclude, it is believed that a deeper development of our generator, will allow it to generate syntactically and semantically correct verbs, nouns and phrases, by strengthening the properties of the roots, and the links between the classes that compose the sentences. The semantic part had to be provided by an enriched lexicon using a structure of features. Another goal is to move on to text generation, which can be seen as the purpose of the complete generation process. In order to exploit our tool in large systems, like automatic translators.

8. REFERENCES

[1] Shaalan, K., Abdel Monem, A., Rafea, A., & Baraka, H. 2009. Syntactic Generation of Arabic in Interlingua-based Machine Translation Framework. Third workshop on Computational Approaches to Arabic Script-based Languages (CAASL3). Machine Translation Summit XII. Canada.

[2] Azza, A., Shaalan, K., Rafea, A., and Baraka, H. 2008. Generating Arabic text in multilingual speech-to-speechmachine translation framework, Journal Machine Translation Springer Netherlands, Pages 205-258.

[3] El-Barbary, O. 2016. Using Arabic Skeleton Morphology and Maximum Entropy forArabic Document Classification, British Journal of Mathematics & Computer Science, ISSN: 2231-0851.

[4] Baloul, S. 2003. Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé, Doctorat thesis, university of Mans, (France).

[5] Boudelaa, S., Pulvermüller, F., Hauk, O., Shtyrov, Y., Marslen-Wilson, W. 2010. Arabic morphology in the neural language system. Journal of Cognitive Neuroscience, 22(5):998-1010.

[6] Soudi, A., Van den Bosch, A., and Neumann, G., 2007. Arabic computational morphology. ISBN978-1-4020-6046-5 (e-book), Published by Springer.

[7] Zemirli, Z., Sellami, M., and Vigourou, N. 1997. Modélisation des règles phonologiques dans un système de génération automatique de la langue arabe, 1iere Journées scientifiques, Avignon (France).

[8] Blachère, R and Gaudefroy-Demombynes, M. 1975. Grammaire de l'arabe classique (morphologie et syntaxe), G.P. Maisonneuve & Larose, Editeurs à Paris, 508 p.

[9] Shaalan, H., Samih, Y., Attia, M., Pecina, P., and Van Genabith, J. 2012. Arabic Word Generation and Modelling for Spell Checking. In Proceedings of the Eight International Conference on Language Resources and Evaluation, pp. 719-725, Istanbul, Turkey.

[10] Beesley, K. 1996. Arabic finite-state morphological analysis and generation. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), volume 1, pp. 89–94, Copenhagen, Denmark.

[11] Beesley, K. 2001. Arabic Finite-State Morphological Analysis and Generation of Arabic at XeroxResearch. In Actes de ACL/EACL2001. Toulouse.

[12] Cavalli-Sforza, V., Soudi, A., and Mitamura, T. 2000. Arabic morphology generation using a concatenative strategy. In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000), PP. 86–93, Seattle, Washington, USA.

[13] Habash, N. 2004. Large scale lexeme based Arabic morphological generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04). Fez, Morocco.

[14] Shaalan, K., Monein, A., and Rafea, A. 2006. Arabic Morphological Generation From Interlingua (A Rule-Based Approach), International Conference on Intelligent Information Processing (IIP), September 20-23, Adelaide, Australia,.

[15] Habash, N. and Rambow, O. 2006. MAGEAD:A Morphological Analyzer and Generator for the Arabic Dialects, In Proceedings of Coling-ACL.

[16] Habash, N. 2007. Arabic Morphological Representations for Machine Translation, Arabic Computational Morphology, Publisher Springer Netherlands.