



Journal of Applied Sciences

ISSN 1812-5654

science
alert

ANSI*net*
an open access publisher
<http://ansinet.com>

A Morphological Analyzer for Vocalized or Not Vocalized Arabic Language

Med El Amine Abderrahim and Fethi Breksi Reguig
Laboratoire de Traitement Automatique de la Langue Arabe (LTALA),
Faculté des Sciences de L'ingénieur, Université Abou Bekr Belkaid Tlemcen,
BP 230 Chetouane Tlemcen, Algérie

Abstract: This research has been to show the realization of a morphological analyzer of the Arabic language (vocalized or not vocalized). This analyzer is based upon our object model for the Arabic Natural Language Processing (NLP) and can be exploited by NLP applications such as translation machine, orthographical correction and the search for information.

Key words: Arabic NLP, arabic graphical word

INTRODUCTION

There are tremendous Arabic morphological analyzers. Some of them have a commercial purpose and others are available for research and evaluation. Among most recent and known analyzers are:

- **The morphological analyzer of Xerox:** It adopts the root and pattern as bases for the analysis and uses a set of rules build with the Xerox finite-state technology (Beesley, 2001). However, this latter has a disadvantage, i.e., it presents some defects concerning the rules of derivation and appropriate formation of the forms as well (Attia, 2006).
- **The morphological analyzer of Buckwalter:** It is certainly the more referenced analyzer in the literature, it is available for evaluation, yet, it has a disadvantage that is neglects the use of rules that process linguistic phenomena. For example, each lexicon entry is followed by all forms obtained by inflection, so this will increase the cost of its maintenance (Attia, 2006).
- **The morphological analyzer of Attia:** This analyzer is based upon Xerox finite-state technology. It helps to recognize multiword expressions. It does not handle vocalized texts and targets a particular application (which is syntactic parser) (Attia, 2005, 2006). Because of limitation, it cannot be used in application like automatic voyellation texts.

The development of a morphological analyzer for Arabic hinges on two components: Grammar and

dictionary. Grammar that is rules which validate the arrangement of forms based upon the elements of dictionary. Dictionary contains the vocabulary of the processed natural language, we point out to that the dictionary gives a set of grammatical properties to each word in the language.

There are two ways to make a dictionary:

- The first way is the fact to make a very large; i.e. it contains all the vocabulary, hence the analysis is summarized via a simple access to the dictionary i.e. reducing the effort made for the development of grammar analysis. We note that this method has mostly been used by the following analyzers: Zouari (1998), Achour (1998) and Ben Othmane (1998).
- The second way, is to make a very concise dictionary. It does not need an endeavour to be build, but causes a complexity of grammar. Here are some analyzers used to this method: Hassoun (1987), Al-Shalabi *et al.* (1998), Ben Hamadou (1993), Attia (2000), Beesley (2001), Gaubert (2001), Ouersighni (2002), Zaafrani (2002) and Attia (2005).

Still a third possibility is that there is a mixture: say a grammar and a dictionary and try to make a balance between them. Now let's examine different alternatives to make a reasonable and justified organization of the choice.

CHOICE OF THE ORGANIZATION

There are several organizations for the Arabic language; they depend on the structure of the Arabic

Corresponding Author: Med El Amine Abderrahim, Laboratoire de Traitement Automatique de la Langue Arabe (LTALA),
Faculté des Sciences de L'ingénieur, Université Abou Bekr Belkaid Tlemcen, BP 230 Chetouane Tlemcen,
Algérie Tel: 0021343285685 Fax: 0021343285685

graphical word. We will justify the choice of our organization for the Arabic NLP. But before all, we find it a must to give a short description of the Arabic graphical word.

The Arabic Graphical Word (AGW): Let's remind here that we do not start from nothing, the excellent description of the Arabic graphical word, already existing in the literature (Cohen, 1970; Dichy, 1990; Hassoun, 1987) will be useful to us as a starting point. We mean by graphic word (GW) any graphical sequence of characters that can be separated either by delimiters such as blank or punctuation marks. Contrary to the simple structure of the GW of a language such as French or English, the AGW has a complex structure and requires a particular modelling for its assumption by NLP system. Simply put, AGW can either be simple or complex. A simple AGW is an attested word of the language; it is formed by the concatenation of a basis with possible affixes (prefixes and suffixes). Without affixes, it does not constitute an attestable word of the language.

$$\text{Simple AGW} = \text{Prefixes} + \text{Base} + \text{Suffixes}$$

Yet, a complex AGW is formed by the concatenation of a simple AGW and a set of clitics (proclitics and enclitics) (Dichy, 1990; Hassoun, 1987).

$$\begin{aligned} \text{Complex AGW} &= \text{Proclitics} \# \text{Simple AGW} \# \text{Enclitics} \\ \text{Complex AGW} &= \text{Proclitics} \# \text{Prefixes} + \text{Base} + \\ &\quad \text{Suffixes} \# \text{Enclitics} \\ \text{Or: Complex AGW} &= \text{Prébases} + \text{Base} + \text{Postbases} \\ \text{With Prébases} &= \text{Proclitics} \# \text{Prefixes} \\ \text{Postbases} &= \text{Suffixes} \# \text{Enclitics} \end{aligned}$$

Concatenation noted by the symbol # in the above expression, expresses a weak connection, in other words, an attestable complex AGW of the language can be carried out without proclitics, respectively enclitics.

Example: The graphical form 'فَشَكَرَ وَنَهَمَ'

Postbases			Prébases	
Enclitiques	Suffixes	Base	Préfixe	Proclitiques
هم	ون	شكر	ي	فس

On the basis of the AGW structure we distinguish five possible organizations to the lexicon:

- List of all the words of the language (complex AGW).
- List of inflected word (simple AW), proclitics and enclitics.
- List of bases, prefixes, suffixes, proclitics and enclitics.

- List of prébases, roots, patterns and postbases.
- List of roots and patterns.

Let us now try to analyze these various organizations.

- The first is to use a lexicon of all the forms of the complex AGW. The advantage of this organization is to simplify the analysis strategy (a simple lexicon consultation), but its disadvantage is to poses the problem of the lexicon size and access timer; for the Arabic language we count hundreds of million words (approximately 6.10^{10} (Attia, 2000)). So, this organization is not practically possible in spite of its technical feasibility which is related to the current presence of sophisticated computer equipment (memory of significant size, higher speed of the processing)!
- The second is to use a lexicon of all the forms of the simple AGW (i.e., the list of all the inflected words obtained from the bases list) with a list of clitics. This lexicon can be generated automatically by a deriver and a conjugator. The size of the lexicon will be less considerable (estimated at approximately as 6 million forms in (Ouersighni, 2002); this lexicon (prefix-base-suffix) is generated from the lexical data base DIINAR.1 of the SAMIA project.), consequently the procedure of analysis must take account of the construction rules of a complex AGW according to the triplet (simple AGW, proclitics, enclitics). Generally these construction rules are of two types, either of the rules or of the compatibilities rules between components. The algorithm of analysis avoids all the phenomena of inflection and derivation. It processes only the phenomena of agglutination of clitic with the inflected forms.
- The third is to use a list of bases accompanied with a list of clitics and affixes. We note that, the size of the lexicon is reasonable (199254 entities in (Ouersighni, 2002)), but the procedure of analysis will be more complex compared to the previous organization because it has to takes account of the construction rules of the 5-uplet (bases, proclitics, enclitics, prefixes, suffixes). Starting from the lexicon of this organization we note also that we can generate the lexicon of the previous organization by the operation of derivation and conjugation.
- The fourth organization is to use a list of roots, a list of patterns, a list of prébases and a list of postbases. In this organization a word is presented like a quadruplet (prebase, root, pattern, postbase). The interdigitation (intrication) of pattern and root create the radical part. So the lexicon size is minimal (7570 entities in (Attia, 2000)), but the analysis

procedure will be more complex than the other previous organizations, because in addition to the construction complexity of the word in question, another complexity can be added; it is the operation of interdigitation (analysis and/or generation of bases from the roots and patterns).

- The fifth uses a much reduced list of roots and patterns. In this organization we do not speak about lexicon (since we utilise a much reduced one), so the importance is therefore focused entirely on the analysis strategy. The advantage of this approach is evident (no lexicon), but the disadvantages in my opinion will be the analysis strategy complexity and the maintenance complexity (evolutionarity) of analyzers built around this approach; the best way to prove this is, it is easy to compare the effort produced for the lexicon update (organisation 1, 2, 3 or 4) with the analyzer update. Finally the construction of analyzer containing this organization does not privilege the separation between linguistic objects and their processing as well. This organization does not seem to me suitable for the construction of a morphological analyzer especially if we intend to go towards semantics (since we can not get rid of lexicon).

The adopted solution: To choose an organization, three parameters must be taken into account: lexicon size, analysis procedure complexity and the target application. The organization which minimizes the size and complexity will be the most favourable, but unfortunately the size decreases while going from the first organization towards the last one and the complexity increases. If the first organization is not a subject of a reasonable choice, noticing that organization 2 is an alternative of organization 3 (because it is possible to generate the lexicon of organization 2 from organization 3 using a deriver and a conjuguor) then the choice will be made among three possible organizations (i.e., 3, 4 or 5). To go towards semantics level after the morphological analysis means eliminating organization 5 and reducing the choice in two cases organization (3 or 4). To minimize the lexicon size implies choosing organization 4, but the lack of the analysis algorithm complexity value makes the choice difficult. For all these reasons, it appears to us reasonable to adopt the third attitude to represent the lexicon of the system. This allows us, on the one hand, to avoid managing a voluminous and difficult update lexicon and on the other hand, to keep in the base essential elements only allowing us to generate words which we need. On the basis of this organization we developed an object model for Arabic NLP (Abderrahim *et al.*, 2007). And to validate this, we produced a morphological analyzer for vocalized or non vocalized Arabic language.

THE MORPHOLOGICAL ANALYZER (MA)

Our morphological analyzer is based on the framework MALA used as a plat form for the development of the Arabic NLP applications. MALA hinges on two principal components (Fig. 1):

- A linguistic data base (BDL) integrating all the linguistic data suitable for the Arabic language. Conceptually the BDL is represented by a model based on classes (class in the paradigm object). Moreover its implementation is carried out by a set of table in the relational model and among these tables we finds for example the table of nominal bases, verbal bases, words tools, clitics, affixes and compatibilities between clitics...
- A set of primitives or basic methods (PMBDL) for handling BDL.

In addition to the MA we have produced a tool (OL); it is intended for the linguists and it allows to make the update of the BDL in a very simple way.

The development of MALA has many advantages such as for example:

- Separation (between the linguistic data and the programs which handles them)
- Re-use (common set of tools for all Arabic NLP applications)
- Development standardization (allow to build all the applications with the same standards, technologies...)
- Extension and maintenance facility

The input of MA could be a vocalized or non vocalized Arabic text and we note that the first processing step consists in segmenting the input text in forms. Though the space separator being the mark of the forms borders does not cause any problem to the processing. The Fig. 2 will show the principle of the morphological form analysis summarized in the following three steps:

- Devoyellation of the form
- Consultation of the tool word lexicon
- The segmentation operation proceeds with the access to the various tables (clitic and affixes) in order to detect the presence of proclitic, enclitic prefix and suffix in the form. The result would be five segments (proclitic, prefix, radical, suffix, enclitic), furthermore, the operation of analysis will carry out an access to the bases lexicon to check out whether the base exist or not and if exists, the analyzer associates to it the set of its linguistic information (voyellations...).

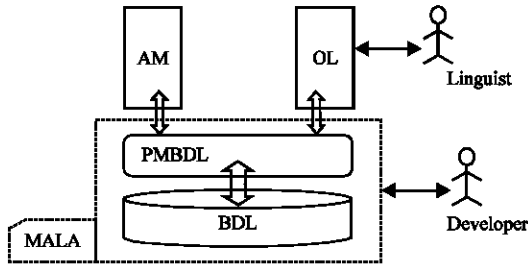


Fig. 1: General architecture of the MALA framework. AM: Morphological Analyser; OL: Tools for Linguist; BDL: Linguistic Data Base; PMBDL: Basic primitives for Handling the BDL; MALA: framework for Arabic NLP

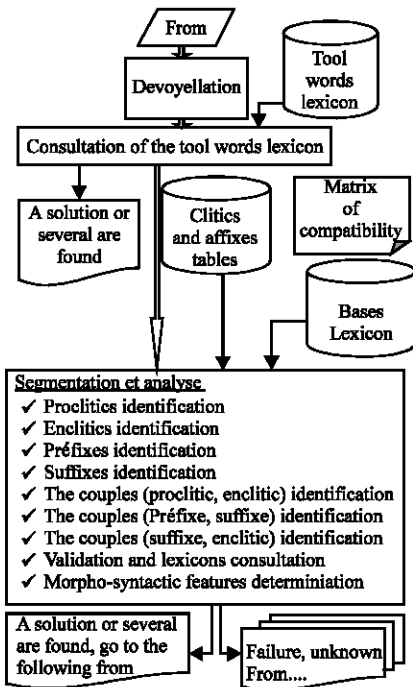


Fig. 2: General architecture of our morphological analyzer

With the output of the morphological form analysis, the analyzer produces a set of information (lexeme, grammatical category, a set of syntactic features...) representing the out-context morphological solution calculated in the used linguistic model.

Text analysis: Contrary to some MA, the analysis of the text at the input is directly applied without pre-processing. The document text is read form by form then the morphological solution for each form is determined.

Before discussing the result of the text analysis, we will describe the output data of the analyzer. Each

1	2	3	4	5	6	7	8	9
Gender	No.	Person	Case	Aspect	Mode	Voice	Transitivity	Human

Fig. 3: The vector representing the morpho-syntactic variable of the morphological solution

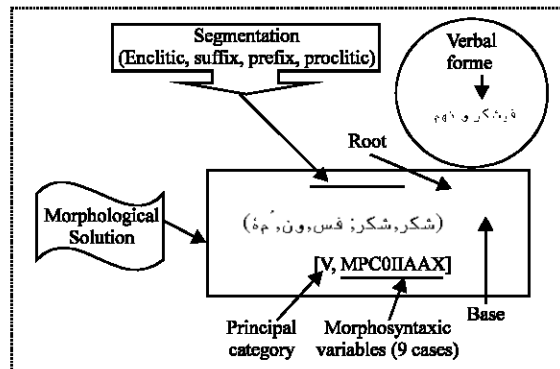


Fig. 4: Example of a morphological solution (the verbal form: 'فيشكر و نهيم')

analyzed form is followed by its morphological solution which includes:

- The base
- The root
- The segmentation into proclitic, prefix, suffix and enclitic
- The principal category (In the Arabic language, there are three categories of forms: nominal, verbal and words tools. The words tool are invariable words, they represent the language constants dictionary): N if Name, V if verb, P if word tool, NBN if a number and PONC if delimiter
- The secondary category (common noun, Name of time and place, assimilated Adjective...) if the principal category is Name)
- The set of morphosyntactic variables values represented by a vector of nine cases (Fig. 3). For example the first case Gender will be able to take one value among {Male (M), Female (F), Not marked (NMG)}. The value NMG (not marked by the gender) concern names which can have the two genders indifferently

The Fig. 4 shows the morphological solution obtained after the analysis of the verbal form 'فيشكر و نهيم'.

In appendix A we give the analysis result of the sentence:

فاذا اجاء المطالب لهم لم يقاوموا مدافعته فيحتاج صاحب الدولة حينئذ إلى الاستظهار بسواهم من أهل النجدة

EVALUATION

Because of the unavailability (at of our currently knowledge) of Arabic annotated corpora, an objective evaluation is impossible, but we will try a manual experimentation which consists in taking texts randomly (of limited size) from Internet and try to analyze them manually, where each form of these texts is segmented into clitics, affixes and bases; and consequently a set of morpho-syntactic variables is associated to them (The two evaluation texts (Text 1 and Text 2) as well as the result of our analyzer are available in the address:

http://www1.univ-tlemcen.dz/~itala/index_fichiers/Page748.htm).

The Table 1 shows the various characteristics of the experimentation texts.

The Fig. 5 shows a screen printing of the morphological analyzer.

To evaluate the performances of our analyzer, we have used measurements of noise and silence utilized in the field of information retrieval to measure for example

the performances of the search engines. In our case, a noise is recorded, for an analyzed form, if the analyzer gives one non valid solution. For example for the form: (أهل) (form n°16 example in appendix A), the analyzer proposes five possible solutions.

The fifth solution presents a noise, whose origin is a bad segmentation. We talk about silence when the morphological analyser produces one or more solutions without the expected exact solution. We point out that the causes of silence are of two types: Incomplete dictionaries or constraining rules and to solve the problem of silence we must: update the dictionaries and re-examine the rules. In our case we have updated our dictionary with all the bases and word tools which exist in the texts of evaluation in order to eliminate the first cause of silence (We voluntarily manually listed all the bases and words tools

Table 1: Numerical characteristics of the analyzed texts

Numerical characteristics	Text 1	Text 2
Size of text or the number of AGW (forms)	654	362
No. of different forms	396	239
No. of nominal forms	323	202
No. of verbal forms	96	62
No. of words tool	229	98



Fig. 5: Screen printing of the morphological analyzer

Table 2: Morphological solutions of the form: (أهل)

Solution 1 (correct solution) (nominal form: relations)	(أهل , أهل , أهل , أهل) [N,B,MC000000H]
Solution 2 (verbal form: to marry)	(أهل , أهل , أهل , أهل) [V,0000A00AH]
Solution 3 (verbal form: to qualify)	(أهل , أهل , أهل , أهل) [V,0000A00AH]
Solution 4 (verbal form: to emerge)	(أهل , هائل , هائل , هائل) [V,0000A00AH]
Solution 5 (case of noise) (word tool: is it, is it, double interrogative)	(هل , هل , هل , هل) [P,X]

Table 3: Result of the texts machine analysis

Numerical characteristics	Rate text 1 (%)	Rate text 2 (%)
Correct analysis	100.00	100.00
Noise	2.14	1.90
Silence	0.00	0.00
Ambiguity	51.22	41.16

in the texts of evaluation and consigned them in the various dictionaries used by the analyzer. This leads us eliminating the assumption of incompleteness of our dictionaries as source of silent).

In addition to noise and silence, we calculated the rate of ambiguity corresponding to the rate of forms which have several morphological solutions; we note that it is the case of the previous form (five morphological solutions are proposed by the analyzer). We considered an analysis as being correct if the analyzer proposes the correct morphological solution of the analyzed form such as the first solution of the previous example (Table 2).

After machine morphological analysis we observed the following results (Table 3):

The rate of correct analysis (100%) supposes that for each analyzed form, the analyzer proposes at least one correct morphological solution. In other words, the analyzer does not produce silence (0%) admitting that our dictionaries are theoretically exhaustive (something not true if we target a wide coverage). This report leads us to declare our analysis algorithm is correct. However if we exclude the problem of the incompleteness dictionaries, the noise ratio recorded to 2.14 and 1.9% will suggest that there are invalid segmentations. The origin of this problem is not the algorithm of segmentation itself, but it is the validations rules of the AGW segments, or quite simply it is the linguistic model used which comprises inconsistency. Consequently it cannot capture linguistic reality suitable to the AGW structure, which is the major disadvantage of this model. Finally it is noticed that there are several ambiguous forms 51.22 and 41.16%. The origin is not the analyzer itself, but it is the Arabic language (because of the lack of vowels and the problem of agglutination). The only way of regulating the problem of ambiguity in this case, is to develop tools for the disambiguation.

DISCUSSION

Comparing our analyzer to others is a very difficult task and remains subjective because there is no standard in term of criteria capable of making this confrontation; each analyzer has its own exit and a well target specific application. However we will be able to advance these some remarks.

Our analyzer differs from the existing Arabic language analyzers in the following aspects:

- Do not target a specific application
- Do not carry a pre-treatment of the input text
- Can be used in analysis as in generation
- Can be used for vocalized texts and not vocalized texts
- Hinge on a reliable and coherent model

In addition to these advantages, the realization of this analyzer is simplified (one algorithm of the forms segmentation and another for the segments validation) by the fact that it rests on a framework which implements the model of the AGW. The idea of the AGW modelling and the separation between task of linguist and developer has never been approached in the existing analyzers. It opens a new prospect for the development of a new generation of applications for the Arabic NLP.

Being limited, the size of our current lexicon (compared to other lexicon, such as that of Buckwalter) does not permit to us a claim to a wide cover. An operation to an update lexicon is necessary.

CONCLUSION

Having justified our choice for the organization of the lexicon (in order to build our model for the modelling of the linguistic objects relating to the Arabic language) we proposed a framework for the development of Arabic NLP applications. A remarkable consequence is the separation between the task of linguist and developer. For the validation, a morphological analyzer which can be based on the framework is built. This analyzer can be exploited by NLP applications such as machine translation, orthographical correction and the search for information, etc.

By disregarding the problem of limited size of our current lexicon, a first experimentation showed us that the analyzer gives good performances; a rate of correct analysis of 100% was observed. Furthermore it seems to us that a tool for the disambiguation is necessary to supplement it.

ANNEXE A

The following page represents the result of the sentence analysis:

فإن اجاء المطالب لهم لم يقاوموا مدافعته فيحتاج صاحب الدولة حينئذ إلى الاستظهار بسواهم من أهل النجدة

[P, V] (مَنْ) [P, V] (مَنْ) [P, X] (مِنْ) [P, N] (مَنْ, مَنْ ; ([N, B, MS000000N (مَنْ, مَنْ ; ([V, 0000A00AX] [, PONC] أهل (أهل, أهل ; ([N, B, MC000000H (أهل, أهل ; ([V, 0000A00AH] (أهل, أهل ; ([V, 0000A00AH] (أهل, هلل ; ([V, 0000A00AH] (هلل ; أ, ([P, X] [, PONC] النجدة ; (نَجْدَة, نَجْد ([N, B, FS000000X] [. , PONC]	[V, ZXC0IXAAX] (خْتَاَج, حَوْج (ف, ي, , [V, ZXC0IXPAX] (, , [, PONC] صاحب (صَاِحِب, صَحْب ; (, , [N, F, MS000000H (صَاِحِب, صَحْب ; (, , [V, 0000A00AH] [, PONC] الدولة (دَوْلَة, دَال (ال, , , [N, D, FS000000N [, PONC] حينئذ (حِيْنئِذِ) [P, X] [, PONC] إلى (إِلَى) [P, N] [, PONC] الاستظهار (اِسْتِظْهَار, ظَهْر (ال, , , [N, A, MS000000X [, PONC] بسواهم (سِوَى ; ب, هَمْ) [P, N] [, PONC] من (مَنْ)	[P, X] (لَمْ) [P, V] (لَمْ, لَمْ ; (, [N, A, MC000000X (لَمْ, لَمْ ; (, , [V, 0000A00AX] [, PONC] يقاوموا (قَاوَم, قَوْم (ر, ي, وَا, , [V, MPC0ISAAX] (قَاوَم, قَوْم (ر, ي, وَا, , [V, MPC0IAAAX] (قَاوَم, قَوْم (ر, ي, وَا, , [V, MPC0ISPAX] (قَاوَم, قَوْم (ر, ي, وَا, , [V, MPC0IAPAX] [, PONC] مدافعته (مُدَاَفَعَة, دَفْع (, , , [N, A, FS000000X] (مُدَاَفِيع, دَفْع (, , , [N, F, FS0X0000X (مُدَاَفَعَة, دَفْع (, , , [N, A, FS000000X] [, PONC] فيحتاج (خْتَاَج, حَوْج (ف, ي, , [V, ZXC0IXAAX] (خْتَاَج, حَوْج (ف, ي, , [V, ZXC0IXPAX] (خْتَاَج, حَوْج (ف, ي, ,	فإن (إِذَا ; ف,) [P, X] (إِذَا ; ف,) [P, X] [, PONC] جاء (جَاء, جِيَا ; (, , [V, 0000A00AX] [, PONC] المطالب (مَطَالِب, طَلَب (ال, , , [N, B, MT000000H (مَطَالِب, طَلَب (ال, , , [N, F, MS000000X (مَطَالِب, طَلَب (ال, , , [N, G, MS000000X [, PONC] لهم (لَهُمْ) [P, N] (هَمْ, هَمْ ; ل, (, [N, A, MS000000H (هَمْ, هَمْ ; ل, (, [N, B, MS000000H (هَمْ, هَمْ ; ل, (, [N, A, MS000000H (هَمْ, هَمْ ; ل, (, [N, B, MS000000H (هَمْ, هَمْ ; ل, [P, N] [, PONC] لم (لَمْ)
---	--	---	---

REFERENCES

- Abderrahim, M. El A. *et al.*, 2007. Object model for vocalized or not vocalized Arabic. JeTIC'2007. Bechar 21/22 avril, 2007.
- Achour, H., 1998. Contribution to the study of the automatic voyellation problem of Arabic. Ph.D Thesis University, PARIS VII.
- Al-Shalabi, R. *et al.*, 1998. A computational morphology system for arabic. <http://acl.ldc.upenn.edu/W/W98/W98-1009.pdf>.
- Attia, M.A., 2000. A large-scale computational processor of the Arabic morphology and applications. M.Sc. Thesis, Cairo University, Egypt.
- Attia, M.A., 2005. Developing a robust arabic morphological transducer using finite state technology. 8th Annual CLUK Research Colloquium, Manchester.
- Attia, M.A., 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. The Challenge of Arabic for NLP/MT Conference, October. The British Computer Society. London.
- Ben Hamadou, A., 1993. Checking and autocorrection by affixale analysis of the texts written in natural language: The case of not vocalized Arabic. Ph.D Thesis Tunis.
- Ben Othmane, C.Z., 1998. Lexicographical synthesis with the detection and the correction of the Arab faulty. Ph.D Thesis, University of Paris-Sud, December.
- Beesley, K.R., 2001. Finite-state morphological analysis and generation of arabic at xerox research: Status and plans in 2001. Proceedings of the Arabic Language Processing: Status and Prospect-39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France.
- Cohen, D., 1970. A tentative of Arabic machine analysis. Studies of Semitic and Arabic linguistics. Mouton 1970 Paris, pp: 49-78.
- Dichy, J., 1990. The writing in the language representation: The letter and the word in Arabic. Thesis University lumière LYON 2.
- Gaubert, C., 2001. Minimal strategies and rules for Arabic automatic processing. Ph.D Thesis, University AIX-MARSEILLE I.
- Hassoun, M.O., 1987. Design of a dictionary for the Arabic automatic processing in various contexts of application. Ph.D Thesis, University Claude Bernard Lyon I, Juillet.
- Ouersighni, R., 2002. Design and realization of robust morpho-syntactic analysis for Arabic: Used for detection and diagnoses it faults of agreement. Ph.D Thesis, University Lumière-Lyon2, 2002.
- Zaafarani, R., 2002. Development of an interactive environment for training with computer for Arabic foreign language. Ph.D Thesis, University LYON II.
- Zouari, L., 1998. Automatic construction of a dictionary for the morpho-syntactic analysis for vocalized or not vocalized Arabic writing. Ph.D Thesis, University of Paris XI Orsay, avril.