The MORPH2 new version: A robust morphological analyzer for Arabic texts

Nouha Chaâben Kammoun [1], Lamia Hadrich Belguith [1],
Abdelmajid Ben Hamadou [2]

MIRACL Laboratory
[1] FSEGS – B.P. – 1088 –3018 Sfax – Tunisia
[2] ISIMS – B.P. – 242 –3021 Sakiet-Ezzit Sfax – Tunisia

Abstract

In this paper we describe a new version of the Arabic Morphological analyzer MORPH2 (a morphological analyzer for Arabic texts). The aim of this work is to deal with the observed problems when evaluating the old MORPH2's version. Then, we present the morphological analysis method used to build this analyzer. We focus on the new step (vocalization and validation) added to our method. This step allows our analyzer to provide fully vocalized outputs with a validation process of noisy solutions. Besides, we present the structure of the lexicon used by our method. It is an XML lexicon that allows more morphological analysis efficiency. It is organized so that it can be used not only for morphological analysis, but also for other linguistic analysis levels. The new version of MORPH2 has been evaluated on an Arabic corpus. The obtained results in terms of recall and precision are respectively 89, 77% and 82, 51%. We noticed that the major causes of failure are the non detection of relation nouns and primitive nouns.

Keywords: Arabic language, morphological analysis, vocalization

1. Introduction

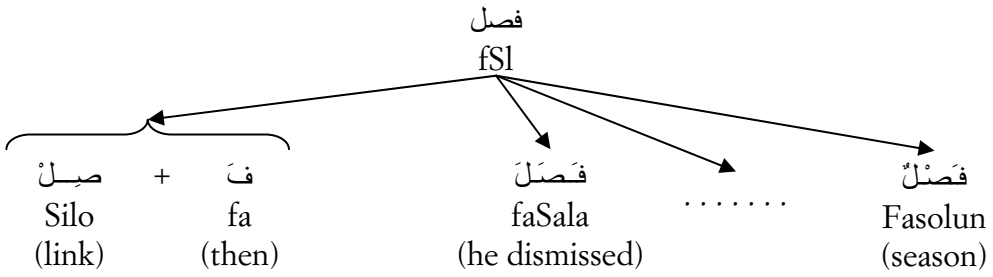
Morphological analysis is one of the crucial parts of Natural Language Processing (NLP). The aim of morphological analysis is to recognize word composition and to provide specific morphological information about words. Such information are useful in the most generic application of NLP such as text analysis, error correction, parsing, machine translation, automatic summarization, etc. Then, developing a robust morphological analyzer is needed.

In this paper, we describe a new version of the Arabic Morphological analyzer MORPH2 (Belguith and Chaâben, 2006). The aim of this work is to deal with the observed problems when evaluating the old MORPH2's version. Then, we present here a brief description of Arabic morphological problems. An overview of morphological analysis' state of the art is then introduced. The next section describes our morphological analysis. We focus especially on the new step (vocalization and validation) added to our method. This step allows our analyzer to provide fully vocalized outputs with a validation process of noisy solutions. Then, a presentation of some lexicon structure is provided. We propose an XML lexicon that allows more morphological analysis efficiency. This lexicon is organized so that it can be used not only for morphological analysis, but also for other linguistic analysis levels. An example of analysis is then presented with a brief description of MORPH2 interface. Moreover, in this paper, we present some generic applications in which MORPH2 was embedded. Finally, we provide the evaluation of our morphological analyzer on an Arabic corpus.

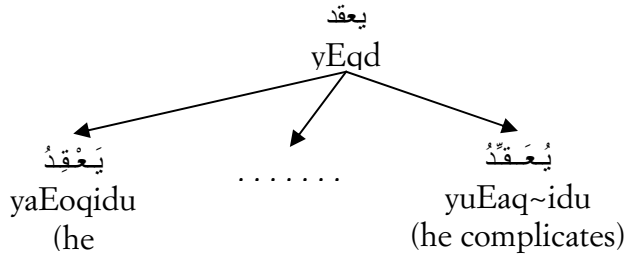
2. Arabic morphological problems

Like all Semitic languages, Arabic is characterized by a complex morphology and a rich vocabulary. Arabic is a derivational, flexional and agglutinative language. In fact, an Arabic word is the result of a combination of a trilateral or quadrilateral root with a specific schema. Then, there are many verbal and nominal lemmas that can be derived from an Arabic root. Besides, from a verbal or nominal lemma many flexions are possibly indicating variations in tense (for verbs), in case (for nouns), in gender (for both), etc. Agglutination in Arabic is another specific phenomenon. Indeed, in Arabic, articles, prepositions, pronouns, etc. can be affixed to adjectives, nouns, verbs and particles to which they are related. Derivational, flexional and agglutinative aspects of Arabic yield prominent challenges in NLP. Thus, many morphological ambiguities have to be solved when dealing with Arabic language. In fact, many Arabic words are homographic: they have the same orthographic form, though the pronunciation is different (Attia, 2006). In most cases, these homographs are due to the non vocalization of words. It means that a fully vocalization of words can solve these ambiguities. We present, in what follows, some of these homographs:

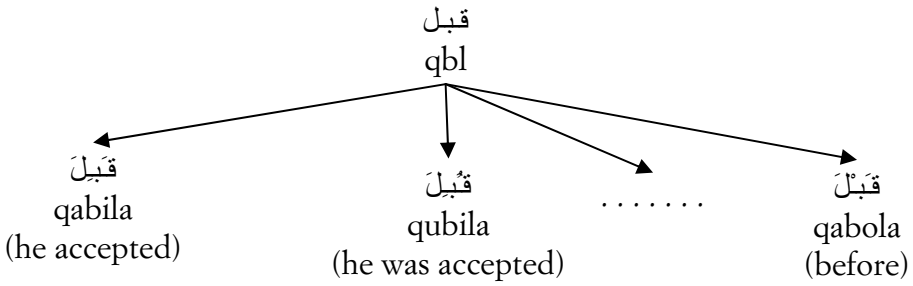
- Homographs due to agglutination: the affixation of clitics in words can produce a form that is homographic with another full form word (Example: the form "فصل").



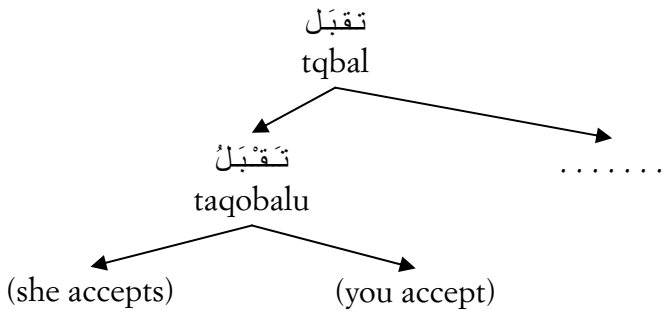
- Homographs due to absence of the character chadda " ˊ ": the presence of chadda inside the word gives a different sense from the one without chadda.



- Homographs due to absence of short vowels: in Arabic, some conjugated verbs or inflected nouns can have the same orthographic form. Adding short vowels to those words makes Differences between them (in pronunciation).



- Homographs due to some homographic prefixes:



3. State of the art

As presented in the previous section, the morphology of Arabic, like all other Semitic languages, is rich and highly complex. The famous root-and-pattern model used in word formation and the writing system of Arabic poses many challenges to computational NLP systems. Arabic computational morphology began in the mid-1980's with Beesley. He used two-level morphology method. In this period, the lack of resources was one of the most challenging problems in computational morphology. Today, with the appearance of few resources and tools for Arabic (even if most of them are not free) many researches in computational morphology have been done. An overview on the state of the art of Arabic computational morphology is presented in the book edited by (Saoudi et al., 2007). They consider two main Arabic computational approaches: the knowledge-base approach and the empirical approach. Knowledge-based methods are based on a lot of linguistic information. They involve a good lexicon representation and some computing techniques to perform morphological analysis. According to different lexicon representations, a large variety of methods belongs to the knowledge-base approach. Saoudi, et al. classify those methods as follows:

- Syllable-based morphology: this class of methods considers syllables to be the primary concept in morphological description.
- Root-and-pattern morphology: as proposed by McCarthy (McCarthy, 1981), stems are formed by a derivational combination of a root morpheme and a vowel melody. The projection of a root on patterns allows stems formation.
- Lexeme-based morphology: lexeme-based morphology supports the claim that the stem is the only morphologically relevant form of a lexeme (i.e. inflectional or derivational morphemes – suffixes, prefixes, infixes and reduplication – are not themselves grammatical elements).
- Stem-based Arabic lexicon with Grammar and Lexis Specifications: the stem-based methods allow the reduction of Arabic word structure complexity, elimination of large numbers of lexical gaps and possibility to associate relevant and specific morphological, syntactic and semantic features with each lexicon entry.

Using knowledge-based methods few morphological analyzers for Arabic have been developed such as Buckwalter Arabic morphological analyzer (Buckwalter, 2004); Xerox two-level morphology system (Beesley, 2001); Sebawai system (Darwish, 2002) for shallow parsing; Abouenour et al., Morphological analyzer (Anouenour et al., 2008) and Attia morphological analyzer (Attia, 2006). Empirical-based methods employ machine learning techniques to extract linguistic knowledge from natural language data directly. The aim of these methods is to learn how to weigh between alternative solutions and how to predict useful information for unknown entities through rigorous statistical analysis of the data.

Supervised, unsupervised and semi-supervised methods were used for Arabic morphology (Clark, 2007). Appearance of empirical-based methods in Arabic morphology is due to appearance of Arabic corpora (e.g. GigaWord, Arabic TreeBank) which can be used by learning techniques. Few Arabic morphological analyzers based on those methods have been developed ((Clark, 2003) (El Jihed and Yousfi, 2005); (Boudlal et al., 2008)).

4. Proposed method

MORPH2 is based on a knowledge-based computational method (Belguith and Chaâben, 2006). Our morphological analysis method involves five steps (see Figure 1.). In this section, we provide a brief description of the principle of this method. We focus especially on the new step (Vocalization & validation) added to our method.

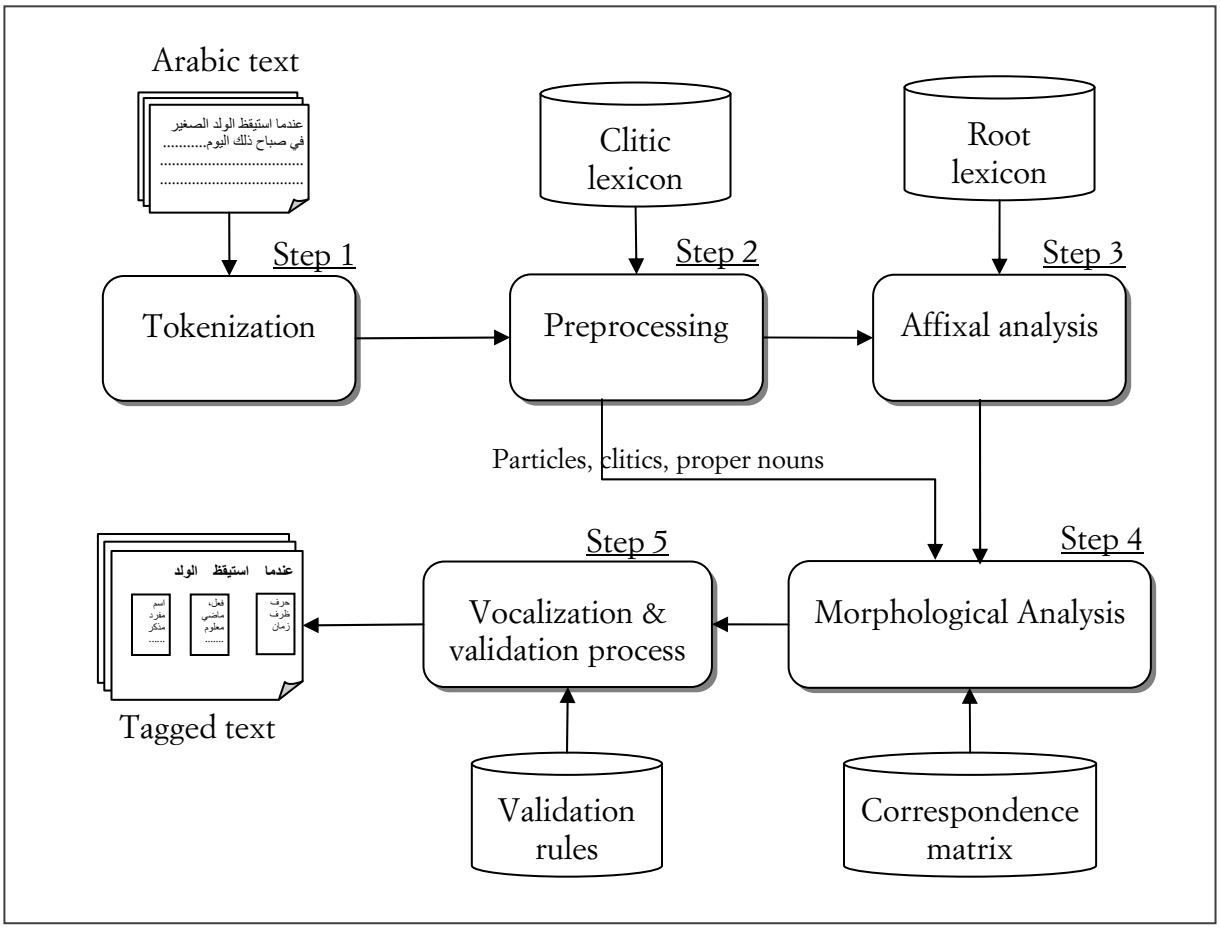


Figure 1. Proposed morphological analysis method

4.1. Principle of the method

As input, our method accepts an Arabic text, a sentence or a word (each word can be non vocalized, partially vocalized or fully vocalized). A tokenization process is applied in a first step. Then, the method determines, in a second step, for each word, its possible clitics. An affixal analysis is then applied to determine all possible affixes and roots. The fourth step consists of extracting the morpho-syntactic features according to the valid affixes and the root. Finally, a vocalization step has been added to our morphological analysis method. The tokenization process consists of two sub-steps. First, using the system "Star" (an Arabic text

tokenizer based on contextual exploration of punctuation marks, conjunctions of coordination, etc.), the text is divided into sentences (Belguith et al., 2005). The second sub-step detects different words within each sentence. Morphological preprocessing step aims to extract clitics agglutinated to the word. A filtering process is then applied to check if the remaining word is a particle, a number, a date, or a proper noun. Arabic language has specific structural properties. A vocabulary word consists of a trilateral/quadrilateral root to which an affixal combination (consisting of a prefix, one or two infixes and a suffix) is added. Affixal analysis aims to identify basic elements that enter into the constitution of a word to say, the root and affixes (i.e. prefix (P), infix (I) and suffix (S)). This process is done by the five following stages (Ben Hamadou, 1993): (P, S) couple identification, candidate affixal triad identification, lexical filtering, ((R), (P, I, S)) association control and transformation recognition. From one stage to another a filtering mechanism is done allowing the removal of noisy recognized decompositions. The morphological analysis step consists of determining from the (R, P, I, S) form obtained for each word, all possible morpho-syntactic features (i.e, POS, gender, number, time, person, etc.). Morpho-syntactic features detection is made on three stages (Belguith, 1999). The first stage identifies the part-of-speech (POS) of the word (i.e. verb, noun and particle). The second stage extracts for each POS a list of morpho-syntactic features. A filtering of feature lists is made in the third stage.

4.2. Vocalization and Validation step

The aim of the new version of MORPH2 is to deal with problems encountered in the old version. It has also been extended to handle fully vocalized texts in addition to non vocalized Arabic texts. Thus, we added a new step to our morphological analysis method called "vocalization and validation". Using morpho-syntactic features and (R, P, I, S) decomposition, a vocalization process is applied. Each handled word is then fully vocalized according to morpho-syntactic features determined in the previous step. To make vocalization, we interdigitate the root extracted in the affixal analysis level with the corresponding fully vocalized schema (see examples below).

Example: Let the word "تقابلتم") *tqAbltm*) be analyzed. Figure 2 describes the vocalization process. Through the analysis of this word, the detected (R, P, I, S) association, in affixal analysis step, is (قبل, ت, ا, تم). In fact, the affixal triad (ت, ا, تم) has only the corresponding vocalization schema "تَفَاعَلْتُمْ" (*tafaAEalotumo*). So, the interdigitation of this schema with the detected root "قبل" gives as a result the fully vocalized word "تَقَابَلْتُمْ" (*taqaAbalotumo/ you have been contrasted*).

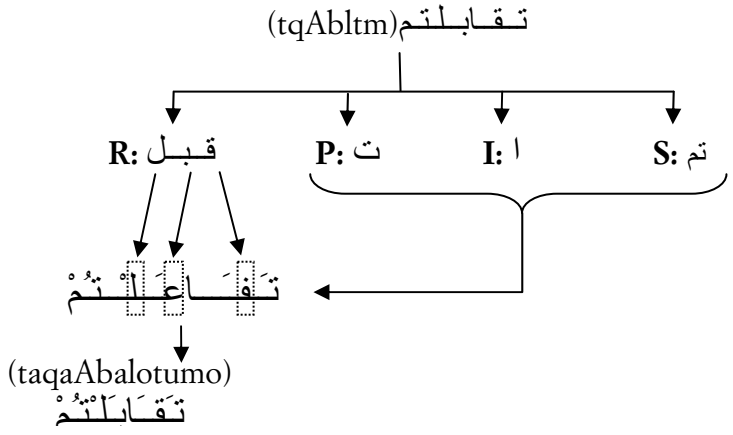


Figure 2. *Vocalization process of the word "* **تقابلتم** *"*

Moreover, in Arabic different orthographic alternation operations can be applied on the combination of a root and a schema, especially when the hamza and the long vowel belong to the root. These orthographic alternation operations include transformation, omission and assimilation. That's why a validation process is needed. The validation process consists of dealing with transformation, omission and assimilation operations and filtering resulting outputs.

- The transformation phenomenon is applied when a root letter has to change while the root is combined with a schema or while a word is conjugated. Example: Let the word "ازدهرت" (*Azdhrt/ she prospered*) be analyzed. Through the analysis of this word, the (R, P, I, S) association detected is (زهر, ا, ت, ت).

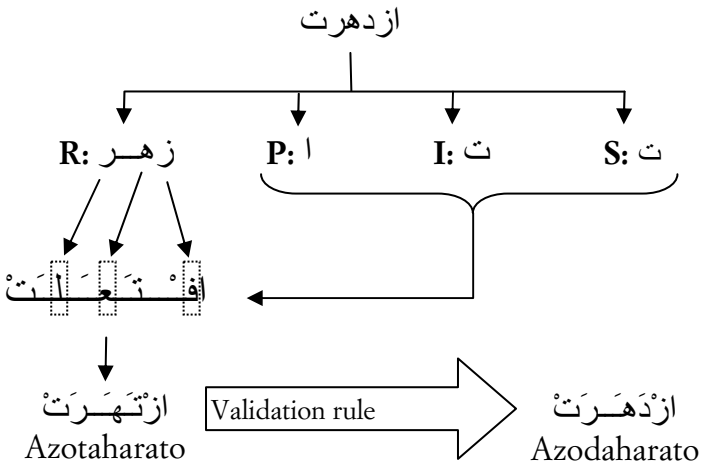


Figure 3. Vocalization and validation process of the word "ازدهرت"

- An omission phenomenon consists of deleting one or two letters while the root is combined with a schema.

  Example: Let the word "يقف" (*yqf/* he stands up) be analyzed. Through the analysis of this word the (R, P, I, S) association detected is (وقف, ي, -, -).

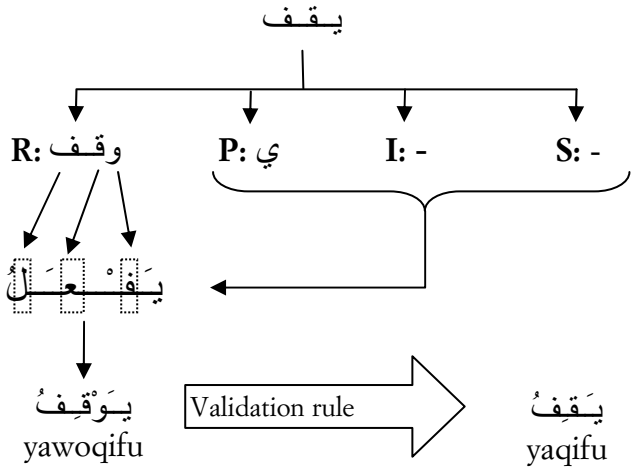


Figure 4. Vocalization and validation process of the word "يـقـف"

- Assimilation phenomenon consists of replacing in some cases two successive consonants, if they are the same, by the chadda " ّ ".

  Example: Let the word "تمدّان" (*tmd~An/ they expand*) be analyzed. Through the analysis of this word, the (R, P, I, S) association detected is (مدد, ت, -, ان).

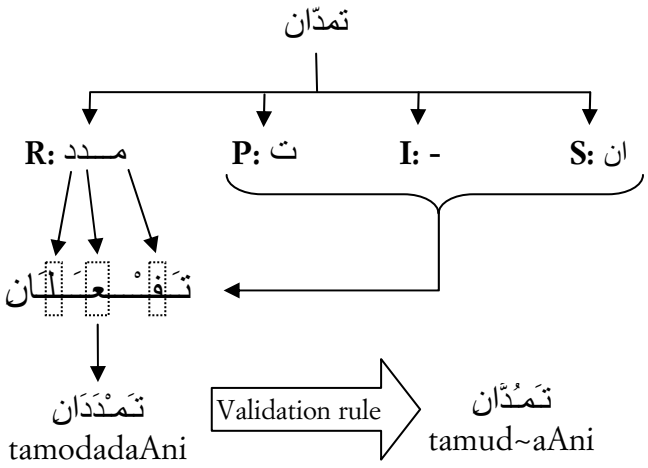


Figure 5. Vocalization and validation process of the word "تمدّان"

Since, the new version of our morphological analyzer must deal with, partially, fully or non vocalized words; the filtering process is used to provide more precision in output results. The process is made by comparing diacritics existing in the initial form input. If, in a position, a difference of diacritic exists between the solution provided by analysis and the input form, then it is deleted from the list of solutions.

5. MORPH2 System

In this section, we present the new version of the MORPH2 Arabic morphological analyzer. MORPH2 is based on the morphological analysis method presented in section 4. The implementation of this method has been done using an oriented object framework. It is made using Java programming language and based on a reduced lexicon and a set of linguistic rules.

5.1. MORPH2 lexicon

MORPH2 uses in each stage of analysis a set of data required for processing such as the lexicon of proclitics, enclitic, and particles in the preprocessing step; the lexicon of affixal triads and roots in the affixal analysis stage; the lexicon of derived and primitive nouns and some tables of correspondences. Since, lexicon organization is a basic step in any morphological analysis application; we tried to organize our lexicon so that it allows more morphological analysis performance. MORPH2 is based on an XML lexicon. It contains 5754 trilateral and quadrilateral roots with which corresponding verbal and nominal schemas are associated (see Figure 6.). The combination of roots and verbal schemas provides 15212 verbal stems. The combination of roots and nominal schemas provides 28024 nominal stems. A list of computing rules is also stored in the lexicon (see Figure 6.). An interdigitation process between verbal schemas and corresponding rules inside this list allowed us the detection of 108415 calculated nouns.

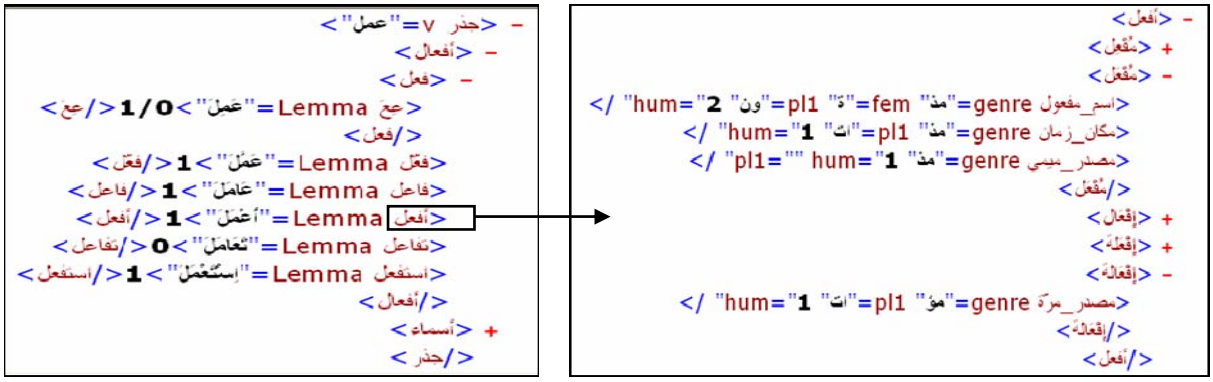


Figure 6. Roots' lexicon and associations rules

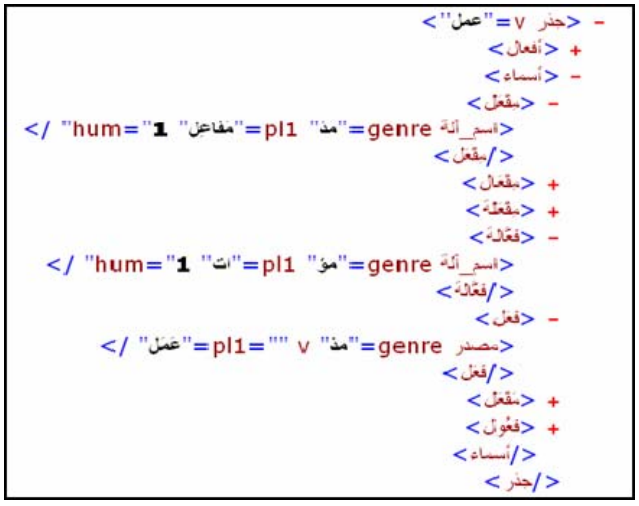


Figure 7. Nominal schema inside a root node

In the Figure 6, we present the structure of the trilateral root "عمل" (*Eml*). This root can be combined with the verbal schemas "فعل", "فعّل", "فاعل", "أفعل", "تفاعل" and "استفعل" to produce respectively the following verbal stems "عمل", "عمّل", "عامل", "أعمل", "تعامل" and "استعمل". The combination of nominal schemas (see Figure 7) and the root is made in the same manner. Calculated nouns result from the combination of a verbal schema with a corresponding computing rule. Let's consider the verbal schema "أفعل". In Arabic, for each verbal stem which schema is "أفعل" a list of possible nouns can be derived. So for each root which can been combined with the verbal schema "أفعل" a list of 8 nominal stems can be computed automatically. While 1704 roots can be combined with the schema "أفعل" then 13632 nominal stems were automatically calculated.

5.2. MORPH2 Interface

As input, the system accepts an Arabic word, sentence or text. Each word in this input can be non vocalized, partially vocalized or fully vocalized. As output, an XML file is provided. This file contains all information extracted by MORPH2. In order to be easy to understand by the user, the system displays the result using an Xsl stylesheet. However, the XML file provided is also useful as format of interchange when the analyzer is embedded inside more generic NLP applications. We present here an example of analysis. As a result of analyzing the word "أفسلّمتكها" (Ofsl~mtkhA/ did I delivred her to you then) by MORPH2, we get the screen presented in Figure 8.

Figure 8. MORPH2's interface

6. MORPH2 embedded in NLP applications

Using an oriented-object framework for implementation and XML for lexicon representation, MORPH2 has been easily incorporated inside some larger applications for Arabic NLP. In fact, the system was embedded inside the following generic applications:

- DECORA (i.e. A system of agreement error detection and correction for non vocalized Arabic texts **(**Boujelen et al., 2008)): DECORA is able to detect and correct agreement errors in gender, number, person, determination, human/non-human semantic features and tense. Its process of detection and correction is based on three phases: morpho-syntactic analysis; syntagmatic analysis for error detection and multi-criteria analysis for error correction.
- MASPAR (i.e. A multi-agent system for parsing Arabic (Belguith et al., 2008)): MASPAR is a multi-agent system for parsing Arabic non vocalized texts. This system provides as an output the syntactic structure of sentences. In failure case, it is able to extract partial structures. This system is based on a multi-agent architecture. MORPH2 was incorporated as an agent collaborating with the other agents (i.e. agent "Tokenization", agent "Syntax", agent "Ellipse" and agent "Anaphora"). STAr system (Belguith et al., 2005) is used as agent "Tokenization". It is able to decompose an Arabic text into sentences. As an input, the agent "Syntax" receives from the agent

"Morphology" (i.e. MORPH2) a list of all possible morpho-syntactic features for each word in the sentence. The objective of the agent "Syntax" is to find the right parsing of a sentence using HPSG grammar (Head-driven Phrase Structure Grammar) (Bahou et al., 2006).

- QASAL (i.e. A Question Answering System for Arabic Language (Brini et al., 2009)): QASAL uses techniques from IR and NLP to process a collection of Arabic documents. This system is able to process factoid and definition questions. It accepts as an input an Arabic question written in MSA (Moderne Standard Arabic) and generates as an output the most relevant passages which are likely to contain the candidate answers. It is composed of three main modules: Question analysis module; Passage retrieval module and Answer extraction module.
- The anaphora annotating system of Mezghani et al. (2008): the target of this system is to produce an annotated resource that could be used for automatic anaphora resolution process of Arabic. It uses an XML-based scheme for annotation. In order to accelerate the process of identification of anaphoric entities by the human annotator, a module which could automatically detect the pronouns is integrated. This module is based on the morphological analyzer MORPH2 and a set of syntactic patrons to identify each pronoun in the text.

7. MORPH2 Evaluation

The old version of MORPH2 has been evaluated on a non vocalized Arabic corpus. The obtained results in terms of recall and precision are respectively 69,77% and 68,51 % (Belguith and Chaâben, 2006). We noticed that the transformation and the omission of either long vowels or letter hamza are the major causes of failure. The new version of MORPH2 takes into account all these transformations and omissions. It has also been extended to handle not only non vocalized Arabic texts, but also partially and fully vocalized texts. Thus, it has to improve recall and precision values.

To evaluate the new version of MORPH2, we use the same corpora used in the old version but with maintaining possible existing diacritics. The corpus consists of a collection of various Arabic texts. It contains about 51404 words. The aim of our morphological analyzer is to provide all possible morpho-syntactic features for each word without taking into account the context in which it occurs. That's why we make a preprocessing task on our corpus to maintain only 23121 different words. We assume that different words are those which forms inside the corpus are different (e.g. "علّم", "علّمه" and "علّمه").

Evaluation operation consists of calculating for each word in the corpus its recall and precision measures taking into account all its possible decompositions. The average of those measures is, then, calculated. We obtain 89, 77% for the recall measure and 82, 51% for the precision measure. Failure cases are (in most of the cases) due to the non detection of relation nouns ("اسم النّسبة") and primitive nouns (non- derived). In fact, in this stage, MORPH2 is not able to handle relation nouns (e.g. "ثقافيّ" (*vqAfy~/ cultural*), "استقلاليّة" (AstqlAly~p/ *independent*), etc.).

8. Conclusion and perspectives

In this paper, we have outlined some problems of computational Arabic morphology. Then, we presented our morphological analysis method. It is a knowledge-based method which aims to solve most cases of morphological ambiguities. We, also, presented the new version of our Arabic morphological analyzer MORPH2 based on the above cited method. MORPH2 is implemented in an oriented-object framework using Java programming language. It is based on a reduced XML lexicon. The lexicon is organized so that it can be used not only for morphological analysis, but also for other linguistic analysis levels. MORPH2 has been evaluated of on an Arabic corpus of 51404 words. The obtained results are very encouraging (i.e. recall = 89, 77% ; precision = 82, 51%). Failure cases are (in most of the cases) due to the non detection of relation nouns ("اسم النّسبة") and primitive nouns (non-derived).

As perspectives, we plane to deal with problems encountered during MORPH2's evaluation. An extension of our lexicon to cover more primitive nouns and relation nouns is then intended. Also, we intend to add a POS tagging step in order to take into account the context in which a word occurs. This step allows our system to filter morpho-syntactic lists output.

References

Abbès R., Dichy J. and Hassoun M. (2004). The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program, in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages - COLING 2004* – University of Geneva, 28th August 2004 : 15-22.

Abouenour L., El Hassani S., Yazidy T., Bouzouba K. and Hamdani A. (2008). Building an Arabic Morphological Analyzer as part of an Open Arabic NLP Platform. In *Workshop on HLT and NLP within the Arabic world: Arabic Language and local languages processing Status Updates and Prospects At the 6th Language Resources and Evaluation Conference (LREC'08*), Marrakech – Morocco, May 2008.

Attia M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *The Challenge of Arabic for NLP/MT Conference, the British Computer Society Conference*, London, October 2006, pp. 48-67.

Bahou Y., Belguith Hadrich L., Aloulou L. and Ben Hamadou A. (2006). Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés. In *Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'06),* Tours France, 25- 27 janvier 2006.

Beesley K. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *Proceedings of the Arabic Language Processing: Status and Prospect -39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Belguith Hadrich L., Aloulou C. and Ben Hamadou A. (2008). MASPAR : De la segmentation à l'analyse syntaxique de textes arabes. In *Revue Information Interaction Intelligence* I3, vol 7, N° 2: 9-36.

Belguith Hadrich L. and Chaâben N. (2006). Analyse et désambiguïsation morphologiques des textes arabes non voyellés. In *Actes de la 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, pp. 493-501, Belgique, 2006.

Belguith Hadrich L., Baccour L. and Mourad G. (2005). Segmentation des textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. In *Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, pp. 451-456.

Belguith Hadrich L. (1999). *Traitement des erreurs d'accord de l'Arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritères pour la correction*. Thèse de doctorat en informatique, Faculté des Sciences de Tunis.

Ben Hamadou A. (1993). *Vérification et correction automatique par analyse affixale des textes écrits en langage naturel : le cas de l'Arabe non voyellé.* Thèse d'Etat en informatique, Faculté des sciences de Tunis.

Boudlal A., Belahbib R., Lakhouaja A., Mazroui A., Meziane A. and Ould Abdallahi Ould Bebah M. (2008). A Markovian Approach for Arabic Root Extraction. In *The International Arab Conference on Information Technology (ACIT'2008)*, Hammamet, December 16-18, 2008.

Boujelben M., Aloulou C. and Belguith Hadrich L. (2008). Toward a detection/correction system for the agreement errors in non-voweled Arabic texts. In *The International Arab Conference on Information Technology (ACIT'2008)*, Hammamet, December 16-18, 2008.

Brini W., Ellouze M., Mesfar S. and Belguith Hadrich L. (2009). An Arabic Question-Answering system for factoid questions. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'09)*, Dalian, China, Sep. 24-27, 2009.

Buckwalter T. (2004). Issues in Arabic Orthography and Morphology Analysis. In *The Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*, Geneva.

Clark A. (2007). Supervised and Unsupervised Learning of Arabic Morphology. In *Arabic Computational Morphology*, A. Soudi, A. van den Bosch and G. Neumann (eds.), Springer, 2007, pp. 181-200.

Clark A. (2003). Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of the tenth Annual Meeting of the European Association for Computational Linguistics EACL 2003*, pp. 59-66.

Darwish K. (2002). Building a Shallow Arabic Morphological Analyzer in One Day. In *Actes du workshop Computational approaches to Semitic languages*, 47-54, 2002.

Dichy J. (1997). *Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot*. Meta 42, printemps 1997, Québec, Presses de l'Université de Montréal: 291-306.

El Jihad A., Yousfi A. (2005). Etiquetage morpho-syntaxique des textes arabes par modèle de Markov caché. *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues RECITAL'05*, Dourdan, 06-10 Juin 2005, pp. 649-654.

Hammami Mezghani S., Belguith Hadrich L. and Ben Hamadou A. (2008). Anaphora in Arabic Language: Developing a corpora annotating tool for anaphoric links. In *The International Arab Conference on Information Technology (ACIT'2008)*, Hammamet, December 16-18.

McCarthy J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12: 373-418.

Ouersighni R. (2002). L'analyse morpho-syntaxique de l'Arabe voyellé ou non voyellé : Le système AraParse. In *Actes de l'assemblée internationale du traitement automatique de la langue arabe.*

Soudi A., Neumann G. and Van den Bosch A. (2007). Arabic Computational Morphology: Knowledge-based and Empirical Methods. In *Arabic Computational Morphology*, A. Soudi, A. van den Bosch and G. Neumann (eds.), Springer, pp. 3-14.

Zaafrani R. (2004). Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. In *Actes de la 11^{éme} Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2004).*