Alkhalil morpho sys1: A morphosyntactic analysis system for Arabic texts

Article · January 2010

CITATIONS
70

READS
1,254

2 authors, including:

Abdelhak Lakhouaja
Université Mohammed Premier
52 PUBLICATIONS **618** CITATIONS

SEE PROFILE

Alkhalil Morpho Sys[1]: A Morphosyntactic analysis system for Arabic texts

A. BOUDLAL*, A. LAKHOUAJA**, A. MAZROUI**, A. MEZIANE**,
M. OULD ABDALLAHI OULD BEBAH** and M. SHOUL*
*Faculty of Letters and Human Sciences, University Mohamed I, Oujda, Morocco
rahimboudlal@hotmail.com, mshoul@hotmail.com
**Faculty of Sciences, Department of Mathematics and Computer Sciences, University Mohammed I, Oujda, Morocco
abdel.lakh@gmail.com, azze.mazroui@gmail.com,
abdelouafi_meziane@yahoo.fr, medbebaha@yahoo.fr

ABSTRACT

Alkhalil Morpho Sys[1] is a morphosyntactic parser of Standard Arabic words. The system can process non vocalized texts as well as partially or totally vocalized ones. Our approach is based on modelling a very large set of Arabic morphological rules, and also on integrating linguistic resources that are useful to the analysis, such as the root database, vocalized patterns associated with roots, and proclitic and enclitic tables. As an output of the analysis, we have a highly informative table mainly containing vocalization of the stem, its grammatical category, its possible roots associated with corresponding patterns, proclitics and enclitics.

Keywords*: Arabic language processing, analyser, morphosyntactic parser, Standard Arabic.*

1. INTRODUCTION

Arabic language processing (ALP) is a field of research in which many linguistic engineering specialists are interested. Such interest has increased with the written Arabic documents' proliferation which is partly due to the Web popularization and the increase of the means of communication in Arabic. Researches conducted these recent years in such field have led to an important variety of applications such as automatic indexing, information retrieval, machine translation, automatic summarization, automatic word generation, syntactic analysis, morphological analysis, automatic vocalization, spell checking and text analysis systems.

In order to have their performances enhanced, most of these applications require a correct morphological analysis of Arabic words. Consequently, developing a morphological parser capable of correctly processing all Arabic words and offering any possible morphological information remains a necessity for ALP. It is worth noting that such system is still a challenge for linguistic engineering specialists, which is due in particular to the richness and complexity of the Arabic language.

It is in this framework that our contribution takes place and which is added to the efforts already made by the scientific community in such field. Our system takes inspiration from approaches that consider the word a concatenation of morphemes and whose analysis consists first in its segmentation into prefix-stem-suffix before looking for morphological information like the pattern of the stem, its root and its grammatical category.

2. ARABIC LANGUAGE FEATURES

Arabic is a Semitic language written from right to left. Its alphabet is very limited, comprising 28 consonants and six vowels, three of which are short (ُ ِ َ) and the remaining three long (ا ي و). The Arabic language shows two main features that make of it a complex and ambiguous language: agglutination and dispensability of vowel diacritics. Indeed, each Arabic word may be made up of prefixes, one root or stem and suffixes. Determining these lexical units is hampered by word multiple segmentations and the problem lies in distinguishing the appropriate segmentations among the solutions proposed. As for the lack of text vocalization, a very familiar situation, it generates several cases of lexical and morphological ambiguities. For instance, the non vocalized word علم may be read عِلْمٌ « science », عَلَمٌ « flag », عَلِمَ « he knew ». Thus, a non vocalized word out of context has several interpretations, and a right reading and sound comprehension resort then to vocalizing at least one of the letters of the word in question.

2.1. CLASSIFICATION OF ARABIC WORDS

The Arabic lexicon is estimated at $6x10^{10}$ distinct words (or surface forms) [10]. Various classifications of this lexicon can be considered:

[1] The system was developed in collaboration with the Arab League Educational, Cultural and Scientific Organisation (ALECSO) and King Abdul Aziz City for Science and Technology (KACST). We thank Mr. Almoataz Bellah Al-Said from Cairo University for his help during the creation of system databases.

- Classification according to grammatical category: verbs, nouns and particles.
- Classification on the basis of the derivative and the primitive (non-derivative): the Arabic lexicon contains on the one hand a large part of words that are derived from a root and according to a pattern, and on the other a class of non-derivatives (e.g. foreign words, proper names, demonstratives …); the latter class is limited compared to the former.
- Classification according to whether the word is augmented or unaugmented.

In order to develop our parser, we resorted to the primitive/derivative classification.

2.1.1. DERIVATIVE CLASS

It was split into two subclasses, the verb subclass and the noun subclass.

a) Verb subclass: a verb is an entity that expresses a time-related meaning. There are two verb types: unaugmented verbs (mujarrad: مجرد) and augmented verbs (mazeed : مزيد). The former are called unaugmented because they are made up only by the root letters. They are constituted of three or four letters, but most of these verbs are triliteral. The augmented verbs are obtained by affixing one or several additional letters to the first stem of the (unaugmented) verb. Thus if we add a 'أ' (hamza, i.e. glottal stop) to the verb خرج which means « go out » we will get أخرج, an augmented verb that assumes another meaning, "cause to go out".
   As far as Arabic is concerned, verb conjugation depends on
   - time: perfect, imperfect;
   - number: singular, dual, plural;
   - gender: masculine, feminine;
   - person: first, second, third;
   - voice: active, passive.

b) Noun subclass: it is constituted of nouns and pronouns. Arab grammarians take into consideration several types of noun classification: non derivative/derivative nouns, proper/common nouns, determinate/indeterminate nouns, compound/non-compound nouns.

2.1.2. NON-DERIVATIVE CLASS

It comprises on the one hand proper names and foreign nouns, and on the other particles such as determiners (articles), prepositions, adverbs and conjunctions. Particles play an important role in sentence interpretation. They are useful for situating facts or objects in time or space and they are essential to the text coherence.

2.2. WRITTEN (ORTHOGRAPHIC) WORD MODEL

Morphological analysis is interested in the written word. The latter is a string of characters delimited by two separators which may be two spaces or one space and one punctuation mark. The written word in Arabic bears more information than that in Latin. Some written words stand for a sentence, such as "سنخبركم" which means: 'we will inform you'. The Arabic written word has a structure termed maximal word decomposable into proclitic(s), prefix, stem, suffix(es) and enclitic(s) [9]. Thus we can have the most complex form of an Arabic word if all these constituents co-occur. The inflected form, made of stem, prefix and suffix, is said to be the minimal word, the latter constituting the lexical nucleus of the written word.

Clitics (i.e. proclitics and enclitics) are morphemes that convey grammatical information. For example, in the written word سنخبركم the enclitic كم is the direct object. Clitics constitute a finite set, but combinations can take place between proclitics on the one hand and enclitics on the other to yield an additional list of compound clitics.

As far as verbs are concerned, one-letter prefixes necessarily indicate the person of verbs conjugated in the present, and suffixes can be the endings of verb conjugation. As for nouns, suffixes can be plural and feminine markers.

The stem can be a derivative that will be identified with a root and pattern, or can be a non-derivative.

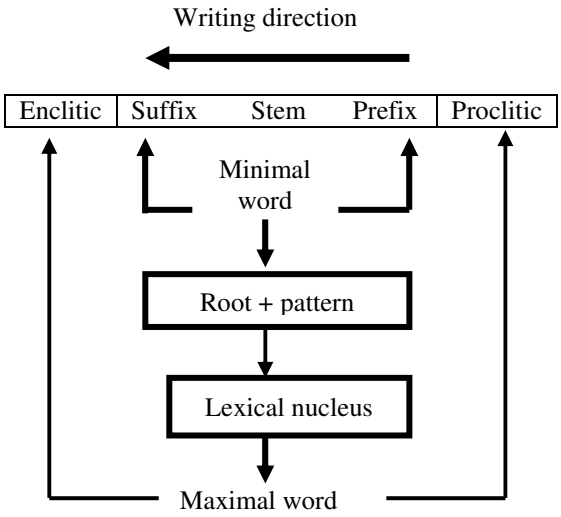

Fig. 1: Decomposition of a written word

3. THE STATE OF THE ART IN MORPHOLOGICAL ANALYSIS

For several decades, numerous researchers have been interested in developing morphological parsers for Arabic. Their interest has given birth to various systems. A large part of such systems, it should be noted, was influenced by their application fields. Thus systems were realized in accordance

with specific areas, such as information retrieval, machine translation, automatic word generation. Along with these, other systems can be mentioned, such as those that are commercially oriented, as well as those that are academically oriented whose research teams do not say much about them and do not give any opportunity for third party evaluation. Such concealment partly explains the unsatisfactory research advancement in Arabic in comparison with European languages. In order to situate our system in relation to the available ones, some of the most cited in the literature will be referred to with special focus on their approaches as well as their limitations.

MORPHO3 from RDI [3]: this system represents a hybrid model for Arabic morphological analysis. It combines morphological rules and a statistical knowledge base to solve problems in morphology. This system's main shortcoming is that the rules are built manually, which takes time and requires great knowledge of the Arabic language. The MORPHO3 system output is a morphological analysis of the word including the latter's root and pattern, and the significance of its prefixes and suffixes, etc.

Xerox finite-state morphological analyser [6]: the Xerox morphological analyser designed by Kenneth Beesley uses Xerox finite-state language modelling tools. It generates all the possible lists of morphological features for each word. Among this analyser's shortcomings, we can mention the following:
- overgeneration in word derivation and many erroneous outputs,
- lack of compatibility of the different morphemes of a written word.
A trial version is on the Xerox site www.xrce.xerox.com/research/mltt/arabic.

AraParse morpho-syntactic analyser: it is based on the lexical resources of the European project DINAR (**D**ictionnaire **IN**formatique de l'**AR**abe). This analyser is modular and uses large coverage linguistic resources. It distinguishes three lexicon classes:
- lexicon of minimal words,
- lexicon of minimal inflected forms (prefixes, stem, suffix),
- lexicon of all maximal inflected forms.

This analyser is very costly in terms of maintenance and synchronization, despite the richness of lexical data integrated in the system.

Darwish's Sebawai system [10]: it is an Arabic morphological analyser developed by Darwish in 2002 in only one day!! It is similar to MORPHO3, except that it uses automatic rules instead of manual ones. This shallow analyzer is used in information retrieval. Its coverage is not perfect; it specializes only in extracting the possible roots of a particular word. Its recognition rate of the right root of a word is 84%.

Aramorph Morphological analyzer [7] : designed by Tim Buckwalter, it is downloadable from LDC site at http://www.nongnu.org/aramorph/french/. The text to be analyzed in AraMorph should be transliterated into ASCII before any processing, and the result should be reconverted into Arabic to be intelligible. This well-known analyzer is designed to be integrated into a machine translation application. It is referred to in the literature and is available for evaluation. However, it suffers a number of shortcomings, among which are (see [5,11]):
- absence of general rules of generation: all lemmas are listed manually;
- insufficiency in processing proclitics attached to verbs and nouns for asking questions:
  - أأكتب = do I write?
  - أهو = is it him?
- lack of specifying certain imperative forms: out of 9198 only 22 verbs get their imperative forms in addition to the imperfect, which will make the analysis of verbs expressing order (as in the examples below) very limited.
  - حاول = (you) try !
  - انتظر = (you) wait !

Morphological analyzer at Sakhr : Sakhr Company developed a commercial morphological analyzer [8]. The latter provides the stem analysis for any Arabic word. It covers all the Arabic language, be it modern or classical. This analyzer identifies all the possible stems of a word, after extracting both suffixes and prefixes. Unfortunately, no trial version is available.

4. SPECIFICATIONS OF ALKHALIL MORPHO SYS ANALYZER

In this section, Alkhalil Morpho Sys that we developed will be presented. First, the lexical resources used will be described; then the orthographic word segmentation method will be introduced, as well as the adopted analysis techniques, in addition to the description of the list of the different morphosyntactic tags that the system is able to identify.

4.1. TECHNICAL DESCRIPTION OF THE ANALYZER

The Alkhalil Morpho Sys analyzer is entirely developed in Java; a PERL version however exists. Choosing Java was motivated by various reasons: it is free and highly portable (it runs on all systems: Windows, Unix, Linux, etc.). What is more, it is object-oriented, which facilitates programming, and the Unicode used by Java allows work to be done

directly on Arabic texts without resorting to transliteration. It is also worth noting that the databases that we built are written with the XML extension and that the system has a graphic interface that facilitates its use. Our analyser can process 25 words per second and can analyze files of thousands of words. The output file is a table of 8 columns that correspond to tags and an average number of about 5 lines for a word.

4.2. LINGUISTIC RESOURCES

One of the main steps in the realization of an Arabic morphosyntactic analyzer is the design and organization of the linguistic databases to be used by the system. Alkhalil Morpho Sys linguistic resources are made of a set of classes, each of which representing a category of linguistic data of the same nature and morphological features.

a) Class 'PR' of proclitics: it contains on the one hand the exhaustive list of simple proclitics, on the other hand, the list of compound proclitics attained through the combination of simple proclitics.
Example of simple proclitics and compound proclitics: "و" "ف" "أ" "ب" "ل" "ك" "و" "س" "أو" "أف" "ول" "وس" "فس" "أس" "أوس" "أفس" "فو" "وك" "فك" "أك" "لك" "وب" "فب" "لب" "ول" "فل" "أل" "أول" "أب" "لل" "أفل" "ولل" "فلل" "وبال" "بال" "فبل" "أبل" "لبال"
Such class is afterwards decomposed into three subclasses:
- subclass labelled 'C' of proclitics compatible with all words,
- subclass labelled 'N' of proclitics that are disposed to be concatenated only to nouns,
- subclass labelled 'V' of proclitics that are disposed to be concatenated only to verbs.

b) Class 'EN' of enclitics: it contains a list of simple enclitics and the list of compound enclitics obtained from the combination of simple enclitics.
Examples of simple enclitics and compound enclitics: "ه" "ي" "ك" "كما" " كن" "ها" "هما" "هم" "ناه" "نا "نيهن" "نيهم" "نيها" "نيه "ني" "هن" "ناها" "كهن" "كهم" "كهما" "كها" "كه" "ناهن" "ناهم" "ناهما" "كنها" "كنه" "كموه "كماهم" "كهن" "كهم" "كهما" " "كه" "ونيهما" "ونيها" "ونيه" "وني" "كنهن" "كنهم" "كنهما" "وناهم" "وناهما" "وناها" "وناه" "ونا" "ونيهن" "ونيهم" "وهن" "وهم" "وهما" "وها" "وه" "وناهن"
This class, as is the case with proclitics, is equally subdivided into three subclasses via the compatibility criterion.

c) Class 'PEC' of compatible proclitics and enclitics: it allows identifying proclitics and enclitics that can be simultaneously concatenated to a word.

d) Class 'ND' of non derivable words: these are words that are refractory to a root / pattern analysis. Such class, constituted of proper nouns, functional words like particles, demonstratives, etc., is divided into two subclasses:
   i. subclass 'ND_TW' that is composed of tool words like particles.
   ii. subclass 'ND_PN' that is composed of proper nouns.

e) Class 'PA_V' of patterns of vocalized Arabic words. It is obtained by mainly using the ALECSO (Arab League Educational, Scientific and Cultural Organization) derivational system [12] and NEMLAR corpus [4]. This class contains the possible root derivations. From these derivations and with the assistance of a program we developed, we automatically obtained all vocalized patterns corresponding to each root. This class can in turn be divided into a noun-related subclass: 'PA_N_V', and a verb-related subclass: 'PA_V_V'.

f) Classes 'PA_N_NV' and 'PA_V_NV' of all non vocalized patterns relating to nouns and verbs, respectively. They are obtained from classes 'PA_N_V' and 'PA_V_NV' by eliminating the word vowels in such classes and by keeping then only the unrepeated patterns.

g) Class 'RO' of roots obtained from the ALECSO derivational system [9]. It contains more than 7000 roots.

h) Finally, Class 'R_PA_V', the most important one, containing roots along with their matching vocalized patterns. Each couple (root, vocalized pattern) was assigned the following morphological tags: word nature and syntactic form. This class was built from the ALECSO derivational system and NEMLAR's corpus. Given that the ALECSO derivational system does not cover all the derivatives of a given root (example: جموع التكسير : broken plurals and الأسماء الجامدة: aplastic words), the list was partially completed. This data class is divided into two: a subclass corresponding to nominal patterns and another to verbal patterns.

4.3. STEPS OF ANALYSIS

morphosyntactic analysis is carried out according to the following five steps:

4.3.1. PREPROCESSING

In order to facilitate subsequent steps, our method starts by preparing the input text. The system proceeds first with segmenting the text into words,

then with normalizing these words by removing both kashida and diacritics. At this stage, any string of characters which is other than Arabic is also eliminated. Our analytical method stores in memory a complete copy of actual vowels of input words, in order to reject the analysis results incompatible with these diacritics.

4.3.2. SEGMENTATION

This step deals with the orthographic word obtained after preprocessing. The system regards it as a series of constituents (proclitics+stem+enclitic) and aims at identifying them. Thus the system proposes all conceivable segmentations by going through the proclitic and enclitic lists defined above. A check of the compatibility of proclitics and enclitics resulting from each segmentation is made. The check in question is made by using the class defined in 3.2.c.

4.3.3. ANALYSIS OF THE STEM

The diacritical vowels being absent, the same stem can lead to various interpretations. It can assume an interpretation that corresponds to a non derivable word, a second interpretation that may refer to a noun and a third one to a verb. Consequently, the system proceeds to a three-phase analysis of the stems of each segmentation validated in the previous step.

a) First phase: the word is analysed as being a non derivable word by checking whether the word stem coming from segmentation belongs to class 'ND' of non derivable words. Validation is then carried out taking into account compatibility criteria identical with those developed in 3.3.4 below. As for valid segmentations, the system will provide the corresponding morphological features. Afterwards, the system moves on to the next phase.
b) Second phase: the system handles the stem as if the latter were a derivable noun. It starts by checking whether the proclitic and/or the enclitic obtained during segmentation are noun-compatible, i.e. whether they belong to class 'N' or to class 'C'. In such case, the system identifies from the stem the possible roots and patterns following the steps below:
   i. using class 'PA_N_NV' to assign the stem the reference patterns having the stem length
   ii. extracting the possible roots by identifying additional letters in the chosen patterns
   iii. making sure that the suggested root belongs to class 'RO' of roots
   iv. using class 'R_PA_V' to check afterwards that the root obtained from a pattern accepts the latter as being that of a possible derivative
   v. assigning, in addition to the valid couple (root, pattern), the associated morphological tags and the possible vowels to the studied stem. Such assignment is possible by using classes 'R_PA_V' and 'PA_N_V'.
c) Third phase: finally, the system deals with the stem regarding it as a verb. Such processing is similar to the previous one, except that verb classes are used here.

4.3.4. RESULT SCREENING

The results obtained from the previous analysis will undergo the following screening processes:

a) concordance between proclitics and enclitics with the output syntactic features:
   - to check the concordance of the stem ultimate character's diacritical short vowel (العلامة الإعرابية) with the proclitic syntactic function,
     - e.g. : the prepositions "ب" and "ك" appear only with nouns in genitive case (الأسماء المجرورة).
   - to check the concordance of the word nature with the enclitic,
     - e.g.: No concordance between the enclitic pronoun "هم" and passive verbs.
b) concordance of the hamza allography (ء, أ, إ, ئـ , or ؤ) in the system's proposed solutions with that of the input word,
   - e.g.: the hamza "ؤ" can not be followed by the short vowel kasra ِ
c) concordance of the vocalizations of the system's proposed solutions with those that may exist in the input word.

4.4. DISPLAY OF THE MORPHOSYNTACTIC ANALYZER'S RESULTS

For a given word, Alkhalil Morpho Sys enables thus the identification of the whole of the possible solutions associated with their morphosyntactic features.

For nouns these features are as follows:

a) for non derivable words, the system gives:
   - vocalization
   - the proclitic and the enclitic associated whenever they exist,
b) for derivable words, the system generally proposes several solutions. For each of these solutions, the system displays:
   - vocalization
   - the proclitic and the enclitic associated whenever they exist,
   - the nature of the noun:
     - different verbal noun types (مصدر أصلي، مصدرميمي)
     - active participle (اسم فاعل)
     - passive participle (اسم مفعول)
     - time and place nouns
     - instrumental noun اسم الآلة
   - the associated vocalized pattern,

- the root,
- the syntactic form of the noun :
  - gender (masculine or feminine)
  - number (singular, dual or plural)
  - syntactic form

For verbs, the system determines:

- vocalization,
- the associated proclitic and enclitic whenever they exist ,
- the verb nature:
  - tense of conjugation: imperfect, perfect, imperative,
  - active verb (مبني للمعلوم) or passive verb (مبني للمجهول)
- the associated vocalized pattern,
- the root,
- the syntactic form of the verb:
  - triliteral or quadriliteral verb,
  - primitive or derived verb,
  - conjugation person,
  - transitive or intransitive verb.

As for particles, the system determines the following features:

- vocalization
- nature of the particle (particle of coordination, preposition etc..)

5. ASSESSMENT

Because of the unavailability of a test corpus of an important enough size and which represents all morphological cases of Arabic words, it is presently rather hard to assess our system. Nevertheless, the system was selected among several morphological systems in the last competition which was organized by ALECSO. Such choice was based on the results of our system's analysis carried out on ALECSO's test corpus. Following this competition, the Alkhalil system has become an open source application, which will enable its assessment by the scientific community and its future users. The source code and all system databases and documentation can be freely downloaded from [1].

6. CONCLUSION AND PROSPECTS

In the present article, we have introduced a new method of morphosyntactic analysis of fully, partially or not vocalized Arabic words. This method is based on morphosyntactic rules programming and on integrating some linguistic resources. We have itemized the different analysis steps and the results attained. Prospectively, we intend to complete the database, particularly that of aplastic nouns and their derivations. As our analyzer also proposes possible vocalizations, we consider the development, in the near future, of a text vocalization system based on our morphosyntactic analyzer. Likewise, using this system's analysis outputs, we will develop a tagger of Arabic words in a sentence by exploiting the context.

REFERENCES

[1] Alkhalil Morpho Sys ,Version 1.0, 2010, http://www.alecso.org.tn/index.php?option=com_content&task=view&id=1302&Itemid=956&lang=ar

[2] Al-Sughaiyer I. and Al-Kharashi I., "Arabic Morphological Analysis Techniques: A Comprehensive Survey," Journal of the American Society for Information Science and Technology, vol. 55, no. 3, pp. 189-213, 2004.

[3] Attia. M., *A large-scale computational processor of the Arabic Morphology and applications*. A Master's thesis, Faculty of Engineering, Cairo University, Egypt, 2000.

[4] Attia. M., Yaseen. M., and Choukri. K., "Specifications of the Arabic Written Corpus produced within the NEMLAR project", www.nemlar.org, 2005.

[5] Attia, M., "An Ambiguity-Controlled Morphological Analyser for Modern Standard Arabic Modelling Finite State Networks", *in Proceedings of the challenge of Arabic for NLP/MT'*, The Britich computer society, London, 2006.

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

[7] Buckwalter. T., "Arabic Morphological Analyser Version 1.0", Linguistic Data Consortium numéro LDC2002L49, 2002.

[8] Chalabi. A., "Sakhr Arabic Lexicon" , *In NEMLAR International Conference on Arabic Language Resources and Tools*, pages 21.24. ELDA, 2004.

[9] Cohen D., *Essai d'une analyse automatique de l'arabe*, in *Etudes de linguistique sémitique et arabe*. Mouton, p. 49-48, 1970.

[10] Darwish K., "Building a Shallow Morphological Analyser in One Day!", *in Proceedings of the Workshop on computational Approaches to Semetic Languages in annual meeting of the Association for Computational Linguistics (ACL-02),* Philadelphie, USA, 2002.

[11] Mesfar S., "Analyse Morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard". A Doctorat thsis, Franche-Compté University, 2008.

[12] Sarf, www.alecso.org.tn