Morphological Analyzer of Arabic Words Using the Surface Pattern

Said Iazzi[1], Abdellah Yousfi[2] , Mostafa Bellafkih [1] , Driss Aboutajdine[1]

[1] Laboratory GSCM-LRIT, FS, Agdal-Rabat, Maroc

[2] Team ERADIASS, FSJES, Souissi-Rabat, Maroc.

Abstract

The present article introduces a system for the morphological analysis of the Arabic language. The system is based on the surface patterns of Arabic words.

Our work in this article purports to deal with Arabic derived nouns. It is based mainly on the building of a database for the surface patterns of the latter. In order to deal with Arabic derived nouns, the article is also based on a previous study by [19] for the analysis of Arabic verbs.

Our approach was tested against a corpus of 2400 Arabic words (400 verbs and 2000 derived nouns), the obtained results are very interesting and show the utility and importance of this approach.

Keywords: *ATNL, Arabic Derived Nouns, Surface Pattern of Words, Morphological Analysis, Degree of Similarity*.

1. Introduction

The Arabic morphological analysis is one of the tools that permit to solve the majority of the problems of Arabic. It has been extensively used in several domains of the automatic treatment of natural languages (ATNL), like documentary research, electronic dictionaries, systems of marking, etc.

Several works have been realized in order to elaborate morphological analyzers of Arabic. They can be aggregated in three approaches [6] [19]:

- **The symbolic approach:** This approach is based on the segmentation of the word into prefixes, infixes and suffixes in order to extract the root of the Arabic word. Several morphological analyzers have been elaborated which rely on this approach [6]; [5] ; Hegazi et ElSharkawi, 1986 ; Koskenniemi, 1983 ; Beesly, 1998 ; El-Sadany et Hashish, 1989 ; Khoja et Garside, 1999 ; Soudi, 2002). Among the well-known analyzers for this approach is that of Buckwalter The latter consists of determining all possible segmentations of the word, looking for the results in the lists of the radicals, the suffixes and prefixes, and verifying, then, if the morphologies of each of the elements are compatible with each other while examining three tables of correspondences: prefix-radical, prefix-suffix, radical-suffix.

- **The statistical approach:** This approach calculates the possibilities and the probabilities that a prefix and a suffix or a radical can appear together in a data base of words (Goldsmith et John, 2001).

- **The hybrid approach:** this approach combines the two previous approaches (Darwish, 2002).

Among the disadvantages of these approaches, we can mention for example:

i. The dictionary of the words is very big, and it is very difficult to construct a dictionary containing all Arabic words. These dictionaries (of words) contain a sort of repetition of the nouns having the same morphological rules ("مجيئ – جاء"،"مشيئ –شاء").

ii. These approaches use several rules at the time of morphological analysis.

To remedy these problems, we developed an independent morphological analyzer of the words dictionary, without using the rules at the time of morphological analysis. Our system uses only the surface patterns of the word to analyze.

2. Construction of the surface patterns base of the derived nouns

2.1 Derived Arabic nouns

The derived nouns are the nouns that can be derived from a verbal root. The number and the nature of these forms vary according to the status of the verb to which they are connected. Among the derived nouns (see Table 1), we can mention (Mesfar, 2008):

- **The Active Participle ' اسم الفاعل ':** is a noun associated to any verb of action (transitive or intransitive) and which designates the agent of the verb, i.e. the one that has done the action. For example, the verb 'ضرب ' has the active noun'ضارب '.

- ***The Passive Participle 'اسم المفعول':*** is a noun associated to any transitive action verb. It designates the patient that undergoes the action or the result of this action. For example the verb 'ضرب ' ' (to hit) has for passive participle 'مضروب' (hit).

- ***Verbal Noun'المصدر':*** is an abstract noun formed on the same root as the verb to which it is associated and expresses the same semantic content as the verb. A verb can have more than one verbal noun.

  For example, the verb ود (to like) admits four different verbal nouns مودة – ودادة- ودا – ودا'.

- ***The similar quality 'الصفة المشبهة'***: the nouns of the similar quality indicate the absolute presence of the quality of the one who did the action, like رؤوف 'gracious.'

- ***The comparative اسم التفضيل ':*** it indicates the common quality of two nouns of which one expresses a superior degree, like أخوف 'more fearful .'

- ***The nouns of places and times ' اسم الزمان والمكان ':*** they indicate the place or the time of the action, like مأخذ , ملعب 'playground.'

- ***The noun of instrument 'اسم الآلة ': it indicates the means by which the action has been achieved, like ملعقة 'a spoon.'***

  In this article, in addition to these nouns, we have also treated the following derived nouns: "اسم الهيئة", " المصدر الميمي", ", masdar sinaai "المصدر الصناعي ", hyperbole "صيغة المبالغة" and "اسم المرة".

Table 1: An example of the derived nouns in terms of their roots and their pronouns.

Nature and number of pronoun	*Noun Type*	*Root Derivation*	*Root*
مثنى-مذكر	اسم-الفاعل	ضاربان	ضرب
جمع-مذكر	اسم-الفاعل	قائلون	قال
مثنى-مذكر	اسم-المفعول	موليان	ولي
جمع-مؤنث	اسم-المفعول	مضروبات	ضرب
مفرد-مذكر	الصفة المشبهة	أمي	أمو
مفرد-مذكر	التصغير	وويق	وقى
مفرد- مؤنث	المصدر	واقية	وقى
مفرد-مذكر	اسم الزمان و المكان	موجأ	وجأ
مفرد-مؤنث	المرة	وقية	وقى
مفرد-مذكر	صيغة-المبالغة	رعاي	رعى
مفرد-مذكر	اسم-الألة	مرمى	رمى
مفرد-مذكر	المصدر-الميمي	مرعى	رعى
مفرد- مؤنث	المصدر-الصناعي	الوقائية	وقى
مفرد-مذكر	اسم-التفضيل	أخوف	خاف
مفرد-مؤنث	الهيئة	رمية	رمى

2.2 Surface pattern

The word pattern permits to detect the letters which constitute its root. The pattern of "**مكرمون**" is "**مفعلون** ". The letters "**ف،ع،ل** " replace the letters of the root of "**مكرمون** ", and the pattern of "**صارع** " is " **فاعل** " (Youssef, 1999 ; Bahrak, 869-930; Hanafi, 1914 ; Zanjani, 1343).

This type of pattern cannot present the morphological variations of the word (for example the noun**قائل** of the verb **قال**). That is why we proposed an adapted pattern named surface pattern [19].

The method of construction of this new pattern is the following:

If we suppose that the word whose pattern we look for is:

$w = l_1 l_2 ... l_n$ (l_i Character of the word w) and R is its root.

The surface pattern of w is $p = f_1 f_2 ..... f_n$ with :

$$\begin{cases} f_i \text{ is one of the three letters "ف، ع، ل " if } l_i \in R \\ f_i = l_i \text{ if } l_i \text{ is not in R.} \end{cases}$$

And the surface pattern of the root $R = g_1 g_2 ... g_k$ (g_i is a character) is $P = f'_1 f'_2 ... f'_k$ with:

$$\begin{cases} f'_1 = \text{one of the three letters "ف، ع، ل" if } g_i \text{ is a consonant letter at the time of the conjugation of R.} \\ f'_i = g_i \text{ otherwise.} \end{cases}$$

Example:

The conjugation of the word "**رعى**" to the active participle in the 1st person singular is"**راع**"; therefore, the surface pattern of the root "**رعى**" is "**فعى** " and "**فاع**" is the surface pattern of " **راع**".

The surface pattern of "**آجر**" is "**آفع**"and of "**آجرات**" is "**آفعات**".

For the construction of the base of the surface patterns of the Arabic derived nouns, we treated 127 roots that represent almost all possible classes to generate Arabic derived nouns (Youssef, 1999).

Some linguists generated all Arabic derived nouns from these 127 roots, then they conjugated them to the different persons (masculine singular, masculine dual, masculine plural, feminine singular, feminine dual, feminine plural), and from these nouns, they produced the surface pattern of every derived noun.

At the end we obtained more than 6216 surface patterns which represent almost all Arabic derived nouns. (See Table 2).

Table 2: an example of surface patterns in terms of their roots and their pronouns

Root	***Nature and number of pronoun***	***Noun Type***	***Surface Pattern of derived Noun***
فاء	مثنى-مؤنث	اسم-الفاعل	فائيتان
فاء	فمع-مؤنث	اسم-الفاعل	فائيات
فاء	مفرد-مذكر	اسم-المفعول	مفيء
فاء	مثنى-مذكر	اسم-المفعول	مفيئان
فاء	فمع-مذكر	اسم-المفعول	مفيؤون
فاء	مفرد-مؤنث	اسم-المفعول	مفيئة
فاء	مثنى-مؤنث	اسم-المفعول	مفيئتان
فاء	فمع-مؤنث	اسم-المفعول	مفيئات
فاء	مفرد-مؤنث	المصدر	مفيئة
فاء	مفرد-مذكر	المصدر	فيئا

3.The approach used in our morphological analyzer

In the approach already used by [19], we noticed that for the construction of the base of the surface patterns of the verbs, it adds a stage of link of all the possible suffixes and prefixes with the surface patterns of the verbs conjugated. This makes the size of the database of the patterns very big.

In our case, we suppressed this stage and we integrated in the system a phase of segmentation of the word into suffix and prefix before finding the surface pattern of this word.

Example: the word "**فواقيانهم**" after the extraction of the prefix "**ف**" and of the suffix "**هم**" we find " **واقيان**". Thus, the surface pattern is "**واعيان**".

We look for the patterns of the word in the set of surface patterns having the same length. In this work, we were able to formulate the function that measures the similarity between the word to analyze and the surface patterns. This function has been formulated as follows:

$$f(m\,;w) = \sum_{i=1}^{N} 1_{[m_i;w_i]} \quad \text{with :}$$

$$1_{[m_i;w_i]} = \begin{cases} 1 & \text{if } m_i = w_i \,(m_i = ف, = ع, = ل) \\ f(m,w) = 0 & \text{Else, and we leave the algorithm} \end{cases}$$

$m_i : i^{th}$ Character of the pattern m

$w_i : i^{th}$ Character of the pattern w

The function f produces a set of solutions of surface patterns that we mark by S:

$$S = \{ m \in P_{L(w)} \quad / \quad f(m, w) > 0 \}$$

$P_{L(w)}$: the set of all surface patterns of L(w) length.

L(w) : the length of the word w .

Example:

$$f(\text{'واقيان' ; 'وافيان'}) = 6$$

$$f(\text{'واقيان' ; 'فاعيان'}) = 6$$

$$f(\text{'واقيان' ; 'متفاعى'}) = 0$$

Subsequently, for every surface pattern m_k of the word w we look for its roots R_{k_r} . To find the roots of the word w, we seek the positions of characters "**ف**"، "**ع**"، "**ل**" in the surface patterns of the word w, and extract the characters associated with these positions. The roots of the

word w are found by replacing these characters in the surface patterns of root.
For example, for the word **قانلون** we find the surface patterns:

- **فاعلون** with the surface pattern **فعل** for its root
- **فانلون** with the surface pattern **فال** for its root.

After the application of our method we find the two following solutions:

قانِلون ⟵ فاعِلونَ– فعل ⟵ قَنَلَ

قانِلُون ⟵ قانِلون– فَالَ ⟵ قال

As the root **قَنَلَ** doesn't exist in Arabic, we keep only the second solution **قَالَ** (see Tab 3)

Table 3: Example of the results of the morphological analyzer of words

Nature and number of pronoun	*Type nom*	*Words pattern (w)*	*Roots pattern*	*Words Roots*	*Words (w)*	*words*
أنت	الأمر	افع	فعى	فأى	افئ	وافئكما
مثنى-مؤنث	اسم-الفاعل	واقيتان	وفى	وقى	واقيتان	كواقيتانهم
مثنى-مؤنث	اسم-الفاعل	فاعيتان	فعى	وقى	واقيتان	كواقيتانهم
فمع-مؤنث	اسم-المفعول	مفوءات	فاء	ساء	مسوءات	بمسوءاتكم
فمع-مؤنث	اسم-المفعول	مفوعات	فاع	ساء	مسوءات	بمسوءاتكم
فمع-مؤنث	اسم-المفعول	منفاعات	انفاع	انصاع	منصاعات	بمنصاعاتهم
مفرد-مذكر	اسم-الفاعل	منفاع	انفاع	انصاع	منصاع	ولمنصاعهم

4. The implementation

To test our approach, we first constructed all surface patterns of the derived Arabic nouns. This stage has been achieved by linguists, and they used a set of Arabic references (Mustapha, 1999 ; Bahrak, ; Hanafi et al., 1914 ; Zanjani, 1343).

For the implementation of our approach, we developed a program in java which consists of the following parts (see diagram 1).

- Part 1 segment the word into suffixes, the prefixes and root.
- Part 2 look for the surface patters of the solutions given by part 1.
- Part 3 look for the roots from all surface patterns returned by part 2.
- Part 4 verify the validity of these roots while verifying if they exist in the base of the roots or not.

This approach has been tested on 2400 words (400 verbs and 2000 derived nouns). These words are different from those used in the phase of surface patterns construction.

The global error rate found is 3.9%. The majority of these errors spring mainly from the insufficiency of the surface patterns data base. There are some derived nouns whose surface patterns don't exist in our patterns base. For the rest of the errors, they come from the phase of generation or the construction of these surface patterns.

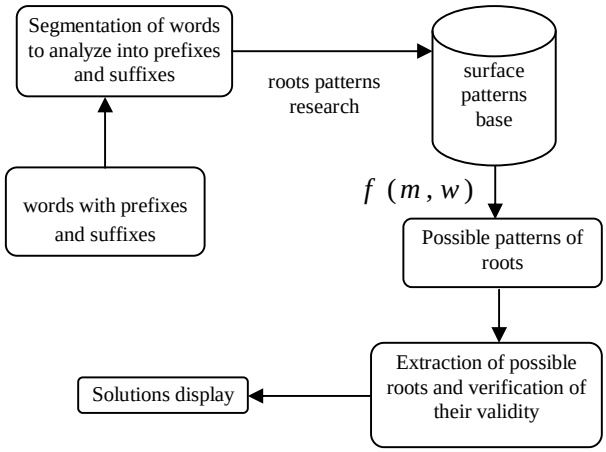

Diagram 1: the stages of our morphological analyzer of the derived Arabic nouns

5. Conclusion:

Our contribution was the treatment of Arabic derived nouns which have not been treated in the case of [19]. Afterwards, we formulated the function that measures the similarity between the words and the surface patterns. Moreover, we could reduce the size of the patterns base while eliminating the phase of prefixes and suffixes addition to the surface patterns. The error rate found is reasonable, and shows the interest of our approach. Subsequently, we will treat the non-derived nouns.

References

[1] Al-Kharashi, I. and Evens.M. (1994). Comparing words, stems, and roots as index terms in an Arabic information retrieval system.
[2] Al Fedaghi.S and Al-Anzi (1989). A new application to generate Arabic Root-Pattern Forms, Procedings of the 11th National Computer Conference and Exhibition, March, Dahran, Saudia Arabia, 391-400.
[3] Bahrak. (بحرق، جمال الدين محمد بن عمر بن مبارك الحميري الحضرمي ، فتح الأقفال وحل الإشكال بشرح لامية الافعال. (869- 930هـ

[4] Beesly.KR (1998). Arabic Morphology Using Only Finate-State Operations, Proceedings of the Workshop on Computational Approaches to Semetic languages. Montreal, Quebec, pp 50-57.
[5] Buckwalter.T (2002). Buckwalter Arabic Morphological Analyzer. Version 1.0. Linguistic Data Consrtium, catalog. Number LDC2002L49 and ISBN 1-58563-257-0.
[6] Darwish.K. (2002). Building a shallow Arabic morphologi-cal analyser in one day. in Proceedings of the ACLWorkshop on Computational Approaches to Semitic Lan-guages, Philadelphia, PA, 2002.
[7] El-Sadany.T.A and Hashish.M.A (1989). An Arabic Morphological System. IBM Systems Journal. Vol.28, No.4, 600-612.
[8] Goldsmith and John.A (2001). Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2), 153-198.
[9] Hanafi. (1914). حنفي بك ناصف-محمد بك دياب-مصطفي طموم-محمود طبعه.قواعد اللغه العربيه لتلاميذ المدارس الثانويه ,افندي عمر-سلطان بك محمد مصر سنه 1914 اديان. علوم الدين
[10] Hegazi.N and ElSharkawi.A (1986). Natural Arabic Language Processing, Proceedings of the 9th National Computer Conference and Exhibition, Riyadh, Saudi Arabia, 1-17.
[11] Khoja.S and Garside.R (1999). Stemming Arabic text. Computer Science Departement, Lancaster University, Lancaster, UK.
[12] Koskenniemi and Kimmo (1983). Two Level Morpology. A General Computational Model for Word-form Recognition and Production. Publication No. 11, Dep. of General Linguistics, University of Helsinki, Helsinki.
[13] Mesfar.S (2008). thèse. analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. université de France-comte, école doctoral.
[14] Mustapha.G (1999). مصطفى الشيخ .الغلاييني. جامع الدروس العربية, العصرية ، المكتبة
[15] Otakar Smrz (2007). "Functional Arabic Morphology Formal System and Implementation". Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague. Doctoral Thesis.
[16] Saliba.B and Al-Dannan.A (1989). Automatic Morphological Analysis of Arabic: A Study of Content Word Analysis. Proceedings of the First Kuwait Computer Conference, Kuwait, March, 3-5.
[17] Sam.A et Youssef D (1999). مجموعة ديشي، يوسف عمار، سام باريس، هاتييه .بيشرال الأفعال العربية
[18] Soudi.A. (2002). A Computational Lexeme-Based, Treatment of Arabic Morphology. Doctorat d'état, Mohamed V University.
[19] Yousfi.A (2010). The morphological analysis of Arabic verbs by using the surface patterns. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 11, May 2010.
[20] Yousfi.A Author Sabri.S and Bouyakhf.E (2006). Système d'analyse morphologique des noms Arabes. JETALA (Journées d'Etudes sur le Traitement Automatique de la Langue Arabe), 2006, Rabat 5-7 juin, 2006.
[21] Yousfi.A. Author Sabri.S. and Bouyakhf.E. (2006). Système d'analyse morphologique des noms Arabes. MCSEAI'06 December 07-09, 2006, Agadir.
[22] Zanjani. (1343 هـ) الزنجاني . متن البناء و متن التصريف العربي .القاهرة 1343 هـ ,مطبعة مصطفى البابي الحلبي