*İnönü Üniversitesi Uluslararsı Sosyal Bilimler Dergisi*
*Cilt 5, Sayı 1, 2016, s. 161-172.*
*www.inijoss.net*

# ARABIC MORPHEMES AND MACHINE TRANSLATION

Joseph-Gabriel BAUDOUIN
Idleb University, Faculty of Arts, Idlib/SYRIA
lemoquedad@gmail.com

## ABSTRACT

This article aims to familiarize lecturers and researchers with the ambiguities of the Machine translation and to make understand the importance of the morphemes functions to identify words and later sentences' syntax, grammar and meanings. This manner to do is also useful to learn or teach Arabic, because it gives another way to approach linguistics or simply languages. Therefore, for the Arabic language, letters and short vowels are important for lexical and syntactic understanding, for that we should not neglect any of them. For the Turkish language, machine translation encounters same difficulties and more, because of its unusual syntax for Arabs and European people, so we need to think differently to resolve teaching and translation problems. We have the lexical ambiguity, which is introducing word into the syntax, which should be able to link between syntax's words to give the appropriate meaning. This work is trying to open a window to look through it to the language as machine can look and see it.

The Machine Translation is one application among many ones what concern languages' engineering. All of them use languages' process. …….

I begin by introducing an example to illustrate this big science. The problem of the Machine Translation or any processing system is to be able to identify words individually in sentence and combine between them to get the possible meanings. Therefore, there are many problems, to reach this goal. We call these problems ambiguities.

I had defined in my researches, in Lucien Tesnière researches centre, a typology of ambiguities that we encounter in Machine Translation, with goals to improve translation's quality.

One of these ambiguities is the segmental ambiguity; it means how we should segment a word that we want to identify or we cannot identify, as a unit.

We take this lexis to illustrate this ambiguity. This word: "أقال" How it should be identified?

The system (the computer) will identify it as one word that means "Dismiss", but depending of its syntax, this identification can be an error. So lexical identifications can be limited by their syntaxes and meanings.

If we say "أقال المدير الموظف", if this sentence is limited to theses vocabulary, our first identification (dismiss =أقال) is right; and in this case, it means, "The director dismissed the employee." Moreover, this is a right translation.

But if the sentence is "أقال الطالب الحقيقة"; and if it is translated as "The student dismissed the truth"; this translation is wrong or can be wrong, because, in the literal sense, the student could not dismiss the truth; otherwise, in the figurative sense, it can be discussed.

So what is the other translation?

In this other way, we need to propose identifications; for that this other way is the segmentation.

Maybe I need to introduce this term of "segmentation": It is the operation necessary to divide a syntagmatic word, which is composed of many lexical or grammatical units, in smaller units recognisable in dictionary.

If we consider that, this word "أقال" is including two units (identities) and that because the first letter "أ" is an affixal identity, and more precisely, it is a prefixal one[1]. So our word becomes:

"أ+قال" what means, "Did he say / has he said"; so this identification opened a new possibility. So our sentence "أقال الطالب الحقيقة" becomes:

"The student did he say / has he said the truth."

So this last identification, if it is applied in our first sentence, it becomes:

"The director did he say / has he said the employee?"

What is also right, if we consider possibilities.

The essential here is that, the system should give these two possibilities, if it wants to be effective.

So we need to inventory affixal morphemes, with idea to be able to identify them when that is necessary.

As we know the Arabic language is morphological, that means we can identify any word by the knowledge of its morphemes, as these morphemes are prefixal, infixal or suffixal.

By this report, we need to identify all morphemes, as it is nominal, verbal or prepositional. Moreover, define its position in the word, prefixal, infixal or suffixal. As we will see, this method defines the adoptive segmentation to operate on the word. For that, we begin by defining "the morphology."

## 1 – THE MORPHOLOGY

In introducing this idea, we take the definition showed by al-Labdī Muhammad S. N. in his "Dictionary of the grammatical and morphological terminologies", so he says:

[1] With idea to distinguish between the three positions of affixes, prefix, infix and suffix.

" البُنية : بِنية الكلمة و بناؤها و مبناها ألفاظ مترادفة، تعني كلها ذات اللفظ و تركيبه و مادته و أصوله... فبنية الفعل "نزل" تعني حروفه التي يتكون منها، و الهيئة التي تنتظم هذه الأحرف من حركة أو سكون."[2]

« The morphology: The morphology of the word, its grammatical vocalization and interne vowels, are synonyms. All these terms mean the composition, the material and the radical of the word… so the morphology of the verb « نَزَلَ » (*nazala*) is composed of its letters and short vowels. »

Therefore, we call all these elements morphemes. This definition is easy for a simple word, but for a syntagmatic word, it is not. If we take this affixed word « فَنَزَلَ » (*fanazala*), we need to be capable to define it with its affixes, so we must identify its morphemes and segment the word as a function of its morphemes. What are morphemes?

## 2 - MORPHEMES

Morphemes can have many appellations, as lexeme or grammeme, as a function of its role lexical or grammatical in the word. Therefore, some morphemes can affect the word's grammaticality or lexicality.

### a – Lexical and grammatical morphemes:

As I said, morphemes can have one or two of these functions. The morpheme and its functions have this classification, illustrated by this table:

| Morpheme | Alphabetic | Vocalic |
|---|---|---|
| **Nominal** | X | X |
| **Verbal** | X | X |
| **Prepositional** | X | |

Table (1): Morphemes and its appellations.

This table shows morphemes that can be affixed to nouns, verbs or prepositions. As they can be lexical or grammatical; then lexical, mean lexicology (lexical morphology) and they concern the syntax; grammatical, they concern the sentence's grammar.

The next table shows among letters that are morphemes. Therefore, we mention these morphemes and its functions:

[2] AL-LABDĪ Mohammad, The dictionary of the grammatical and morphological terminologies *(Muᶜğam al-muṣṭalaḥāt an-naḥawiyya wa aṣ-ṣarfiyya*), Beyrouth, Mu'asasat ar-risala, 1986, p. 27.

| Morpheme | Paradigmatic | Grammatical |
|---|---|---|
| ء | X[3] | ءَ ( أَ ) |
| ا | X | ا |
| ب | | بِ |
| ت | X | تَ[4] |
| س | X | سَ |
| ف | | فَ |
| ك | | كَ / كُ / كِ |
| ل | | لَ / لِ / لْ |
| م | X | مْ |
| ن | X | نَ / نْ |
| ه | | هُ |
| و | X | وَ |
| ي | X | ي/يْ |

Table (2): Paradigmatic and grammatical morpheme's functions.

This table shows only letters which have lexical or grammatical functions. These functions are shown by any lexis. The grammatical morphemes modify grammatically the morphology, but the paradigmatic morphemes haven't any morphological (a morphological modification is done by the paradigmatic creation) impact on lexis.

Morpheme can have lexical or grammatical functions, so some morpheme can introduce lexical and syntactic ambiguities.

Let us take these examples, which contain many affixes:

"تمرين" and "تمرين" these words have, without their morphological vowels, the same morphology which we can consider as lexical and grammatical. Therefore, the morphemes "ت" and "ين" affix nouns and verbs. This written form (*graph*) is composed of lexis and its affixes:

[3] This morpheme is borne by the morpheme lexical «ا» (A) for its grammatical and paradigmatic functions, it is also written on the line.

[4] Fastened to the name « الله », it modifies its determinative vowel, so we have « تَالله».

- ت + مر + ين
- ت + مرن (مر + ي + ن)

The verbal roots are respectively: "مَرّ" and "مَرَّنَ"

The difference between of these graphs is categorical. The noun is identified by dictionary's entries, as for the neutral (infinitive verbal form) verb.

Therefore, we have:

تمرين = Exercise.

However, the identification of our verbal form is not done.

These morphemes have these functions:

<table>
<tr><td colspan="4">تمرين</td></tr>
<tr><td rowspan="2">Noun</td><td colspan="3">تمرين</td></tr>
<tr><td>ن</td><td>ي</td><td>مر</td><td>ت</td></tr>
<tr><td></td><td colspan="2">ين</td><td colspan="2">تمر</td></tr>
<tr><td>Verb</td><td>ين</td><td colspan="2">مر</td><td>ت</td></tr>
</table>

Table (3): Lexical or grammatical morphemes.

The recognition of morphemes is not enough to resolve this ambiguity and distinguish between these two words, but for more easiness, we do not neglect any morphological identification.

I can clarify this table by repeating that the root of our noun here is the verb "marrana" (مَرَّنَ) and the root of our verbal form is "marra" (مَرَّ).

Our system must give us the two possibilities. The choice of one of these categories is decided by its syntactic situation.

These configurations can be:

- هذا تمرين سهل

This sentence means: This exercise is easy.

Then this is a nominal sentence. This lexis derives from a verbal radical by adjunction of some lexical morphemes (here "ت" and "ي").

- أنت تمرين بالسوق

This sentence means: You (female) pass by the market.

This sentence is verbal. Then the verbal form contains affixes that indicate gender, number and tense (with its variants).

### b – Grammatical morphemes:

Many morphemes can have a grammatical function. Therefore, these morphemes can be prefixal, infixal or suffixal. They can affix nouns, verbs and prepositions.

| **Morpheme**[5] | **Grammatical Function** | **Affixations** | **Prefixal** | **Infixal** | **Suffixal** |
|---|---|---|---|---|---|
| ء | Interrogative | Nominal | أَرجل | | |
| | | Pronominal | أَهو | | |
| | | Verbal | أَتفعل | | |
| | Dual | Nominal | | ولدان | ولدا المعلمِ |
| | | Verbal | اضربْ | يقولان | قالا |
| | Vocalic morpheme | Nominal | | | ولداً[6] |
| ب | Completive | Nominal | بِركب | | |
| | Neuter[7] | Pronominal | بِه | | |
| | Neuter[8] | Prepositional | بِأن | | |
| ت | Completive Affirmative | Nominal | تَاللهِ | | |
| | Paradigmatic [9] | Verbal | تَلعب | لَعَبْتُنَّ | لَعِبَتْ |
| ف | Succession et | Nominal | فَرجل | | |

[5]This classification is basic and need to be developed.

[6] It bears the vocalic nominal morpheme « ـً », which can have many grammatical functions.

[7] This neutrality is due of the invariability morphologic of pronouns.

[8] Idem.

[9] This morpheme can have this paradigmatic function (nominal and verbal), a lexical aspect or a grammatical aspect; that depends of the adopted consideration.

| | | | | | |
|---|---|---|---|---|---|
| | conjunction | | | | |
| | Neuter | Pronominal | فَهو | | |
| | Conjunction and consecution | Verbal | فَيركب | | |
| | Neuter | Prepositional | فَإن | | |
| ك | Completive | Nominal | كقلم | قلمكُم<br>قلمكُما<br>قلمكُن | قلمكَ / قلمكِ |
| | Neuter | Prepositional | كما | فيكُم | بكَ / بكِ |
| ل | Completive | Nominal | لِرجل | | |
| | Accusative | Verbal | لِيفعل | | |
| | Neuter | Prepositional | لِأنَّ | | |
| ل | Neuter | Nominal | لَرجلٌ | | |
| | Neuter | Pronominal | لَأنت | | |
| | Neuter | Prepositional | لَفي | | |
| ل | It replaces the morpheme verbal « ل »,when this last is preceded by « فَ » or « وَ » | | | | |
| م[10] | Neuter | Nominal | | قلمكمَا | قلمكمْ (هم) |
| | Neuter | Verbal | | ركبتمَا | ركبتمْ |
| | Neuter | Prepositional | | عليهمَا | عليهمْ |
| ن | Affirmative | Verbal | | | افعلن |
| | | | | | لَيفعلن |

[10] This morpheme designates the plural (tow and more.)

| | | | | | |
|---|---|---|---|---|---|
| | Completive | Nominal | | قلمنا | قلمهنَّ |
| | Neuter | Verbal | نَعمل | | يعملونَ |
| | | Prepositional | | | فيهنَ / فينا |
| ه | Completive | Nominal | | قلمهُم | قلمهُ |
| | Accusative (COD) | Verbal | | حملتهُم | حملتهُ |
| | Completive | Prepositional | | عليهم | عليهِ |
| و | Conjunctive | Nominal | عليٌ وَ أحمدُ | | |
| | Completive | | وَ اللهِ | | |
| | Appellative | | وا أماه | | |
| | Conjugation | Verbal | | يعملون | |
| | Conjunctive | | وَ يعملون | | |
| | Conjunctive | Pronominal | أنت وَ هو | | |
| | | Prepositional | وَ عليه | | |
| ي | Completive | Nominal | | | قلمي |
| | Accusative | | | قلمين | قَلَمَيْ |
| | (Conjugation | Verbal | يكتب | تكتبين | (لم) تكتبي |
| | | Prepositional | | | عليْ |
| | Adjectival | | | | دهنيّ |

Table (4): Morphemes and grammatical functions.

All these words in this table are ambiguous and need segmentation, because we cannot identify them by consulting dictionary. Therefore, we know, from this table, which morpheme we should segment.

**c- Paradigmatic morphemes:**

The paradigmatic morphemes have other utilities; they can be used to create mechanism able to generate neologisms and all possible lexical forms.

| Morpheme | Paradigmatic Function | Prefixal | Infixal | Suffixal |
|---|---|---|---|---|
| ء | | | | رجاء / وقاء |
| أ / أُ / إ | Nominal/verbal | أنزل/ إنزال | | |
| ا | Nominal/verbal | اضرب / اضراب | كاتب / كتابة... | أ |
| ت | Nominal/verbal | تفعّل / تفاعل | اكتتب / اكتتاب ... | صالحات |
| س | Nominal/verbal | سيفعل | استفعل / استفعال | |
| ل | Nominal/verbal | | فعلل / تفعلل | |
| م | Nominal/verbal | مفعول / مفعال... | | |
| ن | Nominal/verbal | نلعب | انفعل / انفعال | معلمون/معلمان |
| و | Nominal/verbal | | معلمون / يعلمون | أخو |
| ي | Nominal/verbal | يعلمون | معلمين / معلميْن | أخي/أخيك |

Table (5): Morphemes and paradigmatic functions.

These morpheme's classifications have many utilities, which can used in linguistics, grammar and Machine Translation.

Identifications and segmentations are necessary for any lexical analysis; that means, a first identification it is not enough to close other identifications.

The identification of these morphemes resolves many ambiguities from the lexical, syntactic to semantic one.

**d- Syntactic models:**

The next step is the right management between the sentence's elements. If we look at sentences, only from these two aspects of lexis or syntax, with all segmentations possible, we could not arrive to the right translation. To get the right translation we need the semantics, which can certify and justify connections between words.

For our first examples:

- "أقال المدير الموظف"
- "أقال الطالب الحقيقة"

The verb "أقال" should be understood with its lexical neighbours; otherwise, we do not know how we can connect them semantically. Therefore, the identification of each lexical item should provide with all necessary information. For this reason the system of translation from Google to other translators, if they want to ameliorate their results they should include all these informations[11].

This verb "أقال" should include in its entry in the dictionary its connections with possible other lexis. If we develop this idea, it means, we need to write the longest sentence possible to account the neighbours. So, for example, we write:

أقال المدير الموظف من عمله في يوم الأحد و ذلك لسبب / بسبب عدم تقيده بأوقات الدوام.

The sentence contains elements, which should be considered when our verb is identified.

We can include this sentence in a table, to define neighbours, so:

| Lexis | أقال | المدير | الموظف | من عمله | | في يوم الأحد | | | لسبب / بسبب إهماله | |
|---|---|---|---|---|---|---|---|---|---|---|
| Category | Verb | Noun | Noun | من | عمله | في | يوم | الأحد | لسبب / بسبب | إهماله |
| | ماض+هو | ال+مدير | ال+موظف | حرف جر | عمل+هو | حرف جر | ظرف زمان | ال+أحد | ل/ب+سبب | إهمال+هو |

Table (5): Sentence's components.

This sentence can be written categorically:

[11] Google is doing that in margin of the window of translation, as they are doing with their service online; they joining these informations and leaving user do that alone and choose the meaning what he wants, and that doesn't mean, he is choosing the right one.

Verb+noun+noun+preposition(place)+noun+preposition+time
adverbial+noun+preposition+noun.

By this manner, this representation is creating syntactic models, which can simplify the syntactic identification and resolve two important ambiguities the lexical and the syntactic.

## 3- CONCLUSION:

The knowledge of morphemes is primordial to identify word manually or automatically; what means, by human, teacher, student or any user, or by machine. This knowledge is useful for students when they begin to study grammar and morphology; it can simplify courses and makes Arabic grammar and morphology more attractive; this way can be to simplify the teaching of other languages, like the Turkish. The problem of machine's identification needs an exhaustive algorithm able to identify morphemes and engage the necessary segmentations; because as we have seen, there are many possibilities, and only an excellent linguist-researcher can envisage them and program the right algorithm.

For example the algorithm of the system's "Translate.Google.com.tr" has a lot of mistakes due to the incomplete identification of lexes and syntaxes. This article can be useful for researchers in Machine translation for the Turkish language. This step needs a lot of researches to scan all problems, to find solutions and elaborate the right algorithms to correct and improve lexical and syntactic Identifications.

## BIBLIOGRAPHY

**Baudouin Joseph-Gabriel**, (2011), Morphemes and grammatical functions in Arabic, SOAS LFG meeting, 05 March 2011, London, UK.

**Baudouin Joseph-Gabriel**, (2010), Les ambiguïtés de la langue arabe pour un traitement automatique, Éditions Universitaires Européennes, Sarrebruck, Allemagne.

**Baudouin Joseph-Gabriel**, (2008), *Désamiguïsation morphologique de l'arabe (Arabic's morphological disambiguation)*, Deuxième colloque international en traductologie et TAL, 7-8 juin 2008, Oran, Algérie.

**Baudouin Joseph-Gabriel**, (2007), *La lexie, sa graphie, sa morphologie, ses affixations, sa*

*syntaxe, sa grammaire et sa sémantique (Lexis, its written form, its morphology, its affixations, its syntax, its grammar and its semantics)*, Colloque international en traductologie et TAL, 9-11.04.2007, Oran, Algérie.

**Al-Labdī Mohammad**, *le dictionnaire de la terminologie grammaticale (Mu$^{c}$ğam al-muṣṭalaḥāt an-naḥawiyya wa aṣ-ṣarfiyya*), Beyrouth, Mu'asasat ar-risala, 1986

**Chenfour Noureddine**, Harti Mostafa, Tahir Youssef, *Modélisation à objets d'une base de données morphologique pour la langue arabe*, JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, 19-22 avril 2004.

**Aloulou Chafik**, Analyse syntaxique de l'Arabe : Le système MASPAR, Récital 2003, Batz-sur-Mer, 11-12 juin 2003.

http://www.sciences.univ-nantes.fr/irin/taln2003/articels/aloulou.pdf.