

Presentation of a Model for the Segmentation of Arabic

GHENIMI, A., CARDEY, S.
Centre Lucien Tesnière, France

Résumé:

Whatever the purpose concerns the understanding, the summarizing or the translation of a text, the machine starts by analyzing it. The analysis starts, in its turn, by the segmentation of this text, before starting possible labelling of the segmented units. The segmentation consists, in the case of Arabic, of dividing up the text being processed into paragraphs and sentences, the sentences into words, then the words into affixes and radicals. In this paper, we present the design of a model of segmentation for voweled or unvoweled 'Standard Modern Arabic (SMA)'. We present the criteria, explain the algorithm used to achieve the task and illustrate with examples segmented by an implementation of the model.

1 . Introduction

Arabic is one of the most difficult languages to process by computer. The difficulties are mainly due to two features. The first one is lexical. Both its syntax and its semantics are expressed by a set of vowels, called in Arabic 'حركات' / haraka:t', which are not always

Presentation of a Model for the Segmentation of Arabic

represented in the words of written texts. Arabic is a consonantal language.

The second feature is syntactic. Contrary to English or French, which are analytical languages; they depend on the word order, a system of auxiliaries and prepositional sentences¹ to communicate the relations: 'who does what to whom and in which circumstances', Arabic is an inflected and agglutinative language at the same time. In inflected languages, the position of the elements in the structure does not provide precise syntactic information. Arabic is at the same time an SVO, VSO, VOS, and OVS structure.

The agglutination is the concatenation of affixes to the radical² to express syntactic relationships. For instance, these affixes could be combined: up to three prefixes in the case of a noun, such in 'فبالقلم' and up to four suffixes in the case of a verb, such in 'سألتمونيها'. (Ghenimi, A., 2003)

Before going any further in the discussion, let us start with the presentation of the issue and the segmentation criteria.

¹ We give a definition of the sentence as conceived by most grammarians: "... an assembly of words making a complete sense", "delimited, at its beginning, by a capital letter of the initial letter of the first word and ends by a full stop".

² Radical: In the parts of speech containing an inflection, the radical constitutes the lexeme (the lexical morpheme). In 'يكتبونها', 'يـ' is a prefix, 'ونـ' and 'ها' are flexional morphemes. The radical is 'كتب'.

2 . The segmentation criteria

We define the sentence as the minimal linguistic unit expressing a complete idea. However, in language processing, the identifying criteria of the sentence in French or English texts do not apply to Arabic. Such criteria include: the sentence starts with a capital letter of the initial letter of the first word and ends with a strong mark, like the full stop, the question mark, and the exclamation mark. In Arabic, neither proper nouns, nor the first words of the sentences take capital letters. The capital letter simply does not exist in Arabic. Moreover, in Arabic literature, we can find sentences which are independent from each other without even being separated by any kind of mark. They are coordinated by a conjunction of coordination 'و', 'أو' or 'ثم' / ' and', ' or ' or 'then'.

The second problem is that of the identification of the words of the sentence (Cardey 2006, 1997). This can be done, as in the treatment of the majority of languages, by locating the strings of characters delimited by spaces. Spaces in a sentence are, generally, used as markers of separation of a string of characters and thus, of words. However, in Arabic, these strings correspond rarely to words.

In Arabic, most grammatical categories which are considered as words in other languages appear as prefixes and suffixes. These include articles, personal pronouns, certain prepositions, possessive adjectives, direct and/or indirect objects, etc. A string of characters located between two spaces contains often more than one word. Furthermore, as we find several lexemes in only

Presentation of a Model for the Segmentation of Arabic

one word, we can also find several words, separated by spaces, in only one lexeme (Cardey 2004), as in:

Example:

{ببدا أن} → ‘Nevertheless’,
{رأس المال} → ‘The capital’,
{مركز المسافات المتناسبة} → ‘The barycentre’,

Or, in several units, as in:

Example :

{بأقلامكم} → ‘With your pens’,
{نلزمكموها} → ‘We impose it on you’, etc.

In the absence of alternative criteria, we adopted that of spaces in our approach³ and in our procedures of segmentation of the words of the sentence. Nevertheless, the full stop, the question and the exclamation marks remain, by default, symbols of the end of the sentence. The other punctuation marks are considered as separators.

Among the Arabic morphosyntactic topics, we present the concept of lexical items in what follows. This is necessary for the continuation of this study in particular to find the agglutinative and morphosyntactic constructions of the language.

3. Nominal / verbal lexical items of Arabic

There are two types of lexical item: nominal and verbal. *A lexical item is every isolable and indivisible sequence which admits or not adjuncts by simple concatenation without losing its character of indivisible*

³ . Other procedures and automata will carry out the determination of the lexemes of the text such as compound words.

*sequence from the point of view of its realization*⁴. (Taibi Nacéra, 2003). It thus consists of an أصل / asl or an omnipresent nucleus in the syntactic structure such as: كتاب / kita:b, being able to accommodate in an alternate and/or continuous way incremental units in well defined positions such as: في كتاب الطالب / fi: kita:bi atta:libi. These positions form a structured unit which constitutes what is called the generating scheme.

a) The nominal lexical item is built around a nominal nucleus and adjuncts on its right and left sides. This definition covers both noun and prepositional phrases. Its structure is schematized as follows:

[Preposition] + [Determinant] + NUCLEUS + [Inflection] + [Tanwin] / [Objects] + [Characterizer].

With:

- *Preposition* = { نحو , مع , تحت , فوق , على , إلى , من } / from, of, with, on, under, towards, etc. }, or { لي , لـ , لـ / li, bi, fa, etc}. The prepositions in this last set have the property of concatenating as prefixes immediately before the determinant when it exists. When this last does not exist, they prefix the nucleus directly.

- *Determinant* = { الـ / the },

- *NUCLEUS* = name,

- *Inflection* = { /u/, /a/, /i/ }, for the singular, and { u:n, i:n, a:n }, for the dual and plural, according to the case.

⁴ . The translation from French has been done by us.

Presentation of a Model for the Segmentation of Arabic

- *Tanwin* = { ٲ /un/, ٲ /an/, ٲ /in/}. In the case of unvoweled texts, only /an/ for the masculine gender can be raised, it is represented by 'ٲ'.

- *Object* = Possessive pron. = { كُنْ, كُمْ, كَ, كِ, نَا, يَ } ,
 { هُنْ, هُمْ, هِمْ, هَا, هِيَ, كَمَا } ,
 or a nominal lexical item or a sentence.

- *Characterizer* = is an object or a sentence.

b) The verbal lexical item, according to Hadj-Salah (1979), is the set formed of a verbal nucleus, affix object pronouns, and of certain linguistic units which are called *converters* or *exponents* such as {قد، ما، لن، لم}. The verbal lexical item is defined by three different generating schemes which correspond to the three Arabic verbal moods: the accomplished, the unaccomplished and the imperative mood. We have placed them in this structure, where certain adjuncts cannot be added according to the mood considered.

**[Exponent] + [Converter] + [personal Pron.] +
VERBAL NUCLEUS + [flexional mark] + [affixes
pronouns].**

The verb is followed at the most by two object pronouns, according to its valency. In the case of the accomplished and unaccomplished tenses, a set of word-like-items precedes the verb; they are classified as follows:

- *Exponent* = {إن ، إذن ، كي}, or any interrogation particle = {هل ، أ} ,

- *Converter* = { قد، لا، لما، س، سوف، ما، لن، لم }.

Presentation of a Model for the Segmentation of Arabic

The interrogation particle ‘*ǐ*’ has the property of being placed immediately before the converter when it exists. When this last does not exist, it precedes the personal pronoun subject if it exists (in unaccomplished tenses) or the verbal nucleus directly, if not.

- *Personnel pronoun* = $\{\text{أنا} / \text{أنت} / \text{هو} / \text{Ø}\}$,

The word-like-items, placed after the verb, are the following ones:

- *Flexional mark*: Personal pronouns with the suffix specific to the accomplished tenses,

- $SufA1 = \{ن, وَا, تَن, تَم, تَا, ا, تَمَا, نَا, ت\}$.

Pronouns suffixes specific to the requirement,

- $SufM1 = \{\text{ل, وا, ي, ن, } \emptyset\}$,

Pronouns personal suffixes specific to the unaccomplished tense,

- $Suff1 = \{\emptyset, \text{ان}, \text{ن}, \text{ون}, \text{ين}\}$,

Affix object pronouns valid for both tenses: accomplished and unaccomplished, specific to the transitive verbs.

- Suff2 = { $\emptyset$, هم, هما, ها, ه, كما, كن, كم, ك, نا, ني}.

For transitive verbs, two object pronouns of the Suff2 type can concatenate in suffixes, bound by a particle which we call connection affixes⁵ (الوصل).

⁵ . The particles of connection (الوصل), also called affixes of conjugation, still attested in Standard Arabic are the following ones. After the pronoun suffixes of the verb of the categories SufAlou Suff2', there is addition of ' ' و ' / ' waw' for the masculine gender and

Presentation of a Model for the Segmentation of Arabic

- Affixes of connection = {ي, و}, as in (6) and (7) below.

{و} after the personal pronouns of plural masculine:
{كم, تم}

And {ي} after the pronouns of the female singular:
{ت, لك}.

However, when it is the case, the 1st pronoun must belong to the subsets:

Suff2' = {كما, كن, كم, لك, لنا, ني}, or

SufA1 = {ن, نوا, تن, تم, تا, لا, تما, نا, ت}, and 2nd with the subset:

Suff2" = {هم, هما, ها, ه}, Ø }.

Examples:

- (1) كي يسأله / In order to ask him.
(2) قد انتهت / she has finished.
(3) لم تحسب / You have not/she has not calculated
(4) أسوف تحس / Are you going to calculate?
(5) أحسبت ؟ / Have you/has she calculated?
(6) أنلزمكموها ؟ / Do we impose it to you?
(7) كتبتِها. / You have written it.

4. Applying the model

The operations that we carry out in our application are applied to a text in Arabic. This last is segmented in sentences - according to the criteria given previously, (based mainly on strong marks). The sentences are then

of 'ي' / 'ya' for the feminine one before the plural and singular affixes of Suff2" category, as in: 'سمعتوهم' / you have heard them.

Presentation of a Model for the Segmentation of Arabic

Another problem, that the procedure solves correctly, is that of diacritic signs, considered as the vowels of Arabic.

For uniformity reasons, our procedure ignores, quite simply, these signs and sends the words deprived of vowels to the analysis.

Example:

كَتَبْتُ for: كَتَبْتُ, كَتَبْتُ, or
كَتَبْتُمْ for: كُتِبْتُمْ, كُتِبْتُمْ; etc.

Note that the vowels are only ignored for the segmentation of the words into morphemes⁷. They can be noted if the analyzer takes them into account in its calculations.

Arabic words, in their turn, are often composed of several minimal linguistic units which we call morphemes. In our application, we distinguish two classes of morpheme: that of the grammatical morphemes and that of the lexical morphemes. The first class is a closed list and contains a limited number of units: determinants, pronouns, prepositions, marks of gender, of the number, of the person and time: {ي, أ, ت, ن, ل, ب, هـ, هما, ها, هـ, كما, كن, كم, لك, نا, ن, وا, تن, تم, تا, ل, تما, نا, ت, هم}, which is equivalent in English to (of, with, he, it, s, es, them, their, our, your, etc).

Among this class, we have gathered the elements which appear in affixes in a package called package of affixes. This last is subdivided, in its turn, into: a sub-

⁷ . We use the concepts of Andre Martinet, 1965. The morpheme is regarded as the smallest segment which carries a form and a meaning. The interest of the concept of morpheme is especially relevant when it is confronted to that of the word.

package of prefixes and a sub-package of suffixes. The treatment of this class is done at the level of the parser and does not appear in the database. On the other hand, the second class is open; it contains the radicals of the lexicon of the language which are extremely numerous and this second class accommodates regularly new units. It is the vocabulary of the language itself, and nothing else.

Examples of the operation of segmentation are given later in the paper.

We now present the algorithm which we have designed to achieve the operation of segmentation.

5. The algorithm for the segmentation of words into morphemes

First, we built a lexicon database of the language. For conceptual reasons, only the radicals, exclusively consonantal, of Arabic words are represented as entries in the tables which constitute the database. These entries are indexed on their alphabetical order. In other words, these radicals constitute the primary key of the database. Moreover, each radical is provided with the grammatical category to which it belongs.

We organized all the grammatical categories in several packages according to the type of each one: The package of the verbs (transitive verbs, intransitive verbs, defective verbs), the package of the nouns (common nouns, proper names), the package of adjectives (qualifier adjectives, numeral adjectives, etc), and finally the package of the particles (emphatic particles, prepositions, conjunctions, demonstrative pronouns, etc.).

Presentation of a Model for the Segmentation of Arabic

From right to left, until the end of the text, and from the first until the last word of each sentence, the parser seeks the word in the database (Table1).

- If it appears there in its entirety, it finds its morphosyntactic properties, and then passes to the following word.

- If not, the program calls upon the procedure which determines the possible prefixes and isolates them from the word under process. This procedure is named *Prefix*. What is left of the word is checked in the lexicon list (Table1) as an entry.

- If it is found, the morphosyntactic properties are retrieved.

- If not, another procedure, that which determines the possible suffixes and isolates them from the word is called. This procedure is named *Suffix*.

- If the rest of the word appears in the table, the same series of operations as that for the other cases is executed.

- If the word does not appear in the lexicon table, the part of the word deprived of the prefixes and the possible suffixes is taken as entry and looked for in the table.

- If it is not found, this means that the word is indeed absent and an informative message is generated before passing to the following word. (Ghenimi, A., 2007)

Thus, when a string of characters like: (يقطعون / he cuts them) is met in a sentence, the program decomposes it in prefix, radical and suffix, displays the outputs which

would be something like the following before it passes to the next word.

- Prefix: (يـ),
- Radical: (قطع),
- Suffix: (هم).

6. Parasitic segmentations

The possible affixes, given in the corresponding sub-packages, are not exclusively used for this purpose. These consonants could really be parts of words as in the examples bellow:

Examples:

In the case of prefixes:

- {بـ} in : {بالغ ، بنر} ,
- {فـ} in : {فاكهة ، فعل} ,
- {سـ} in : {ساعة ، سيارة} ,
- {يـ ، تـ ، أـ} in : {ينبوع ، تمساح ، أملس} ,
- or {لا} in : {لافتة} .

In the case of suffixes, as in :

- {ميزان : ان} : (a balance),
- {سأهم : هم} : (has participated),
- {قانون : ون} : (a law),
- {قرين : ين} : (an associate),
- {بدن : ن} : (a body).

Parasitic decompositions are eliminated by the main program of the application during the analysis, as in the cases illustrated in the examples above. The rule is simple: each time the parser finds one of these particles,

designed as a prefix or a suffix in the corresponding sub-package, it isolates it and consults the database. If the rest of the word does not constitute an entry in the lexicon table and the word before segmentation constitutes one, the decomposition is rejected and the word is considered in its whole (before decomposition).

However, certain decompositions remain ambiguous. The case where both entries (with and without affixes) appear in the lexicon.

Examples:

- (1). {بحر}: is analyzed as only one morpheme: (sea) or as two morphemes: the preposition {بـ} and the substantive {حر}: (a free), and the whole gives: 'with a free'.
- (2). {دوران}: also is analyzed as only one morpheme: (rotation) or as two morphemes: the particle⁸ {ان} and the substantive {دور} (a role), and the whole gives: two roles.

In these cases, it is the entry with the affixes which is retained by default, in order not to take the translation, for example, of: (مكتب / office) instead of: (مكتبة / library). The word: (مكتبة) should appear in its totality as being a radical in the table of the lexicon. In fact, it is a radical.

However, while having chosen, by default, the largest string which matches with a lexicon entry, this does not mean that we have resolved the problem. The correct segmentation of the examples (1) and (2) above,

⁸ . In Arabic, the particle 'ان' can suffix a common noun or an adjective to indicate that the number of individuals concerned is two.

Presentation of a Model for the Segmentation of Arabic

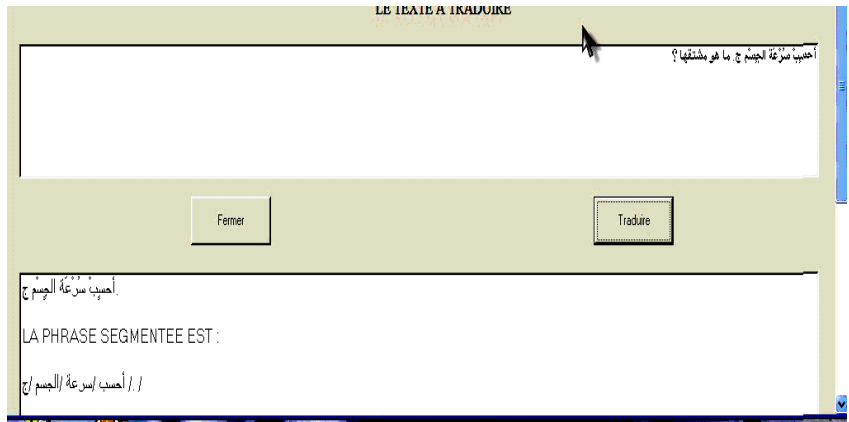
for instance, could well correspond to {بـ} + {بحر} or {دوران} + {ان}. Neither can the vowels resolve the problem. Only a further contextual analysis might reject inadequate decompositions. These possible analyses go beyond this study and are not however dealt with by our model.

7. Examples of segmentation

We present here a micro-text segmented by our application, according to the above-mentioned algorithm.

أحسبُ سرعةَ الجسمِ ج. ما هو مشتقها؟
أنتم، يجب عليكم أن تكتبوا الحل.

The segmentation of this text gave what follows:

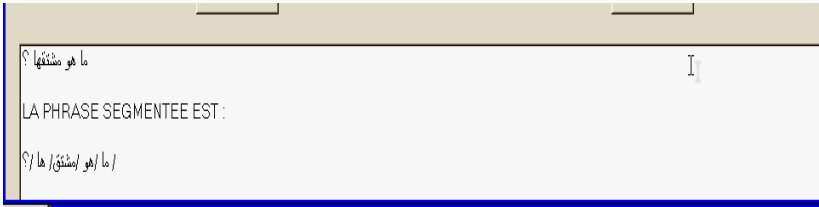


The segmentation of the sentence gives:

أحسبُ سرعةَ الجسمِ ج. → أحسبُ / سرعةَ / الجسمِ / ج. /

Presentation of a Model for the Segmentation of Arabic

We notice the disappearance of the vowels after the segmentation: those diacritics, called in Arabic 'haraka:t'.



The segmentation of the sentence gives

ما هو مشتقها؟

→ ما / هو / مشتق / ها / ؟ /



The segmentation of the sentence gives

→ أنتم، يجب عليكم أن تكتبوا الحل.

/ . / حل / ال / وا / كتب / ت / أن / عليكم / يجب / ، / أنتم /

8. Conclusion

The segmentation, although not always univocal, remains feasible. Two procedures as well as a segment of the main program interact to achieve the task correctly. Our approach to segmentation succeeds in processing Arabic texts provided or deprived of vowels and arrives at results usable by any kind of post processing; in particular labelling, translation or data-mining. All the various ways of separating the words of the sentence are taken into account by the implemented code. Moreover, the separators, reference marks of the segmentation, are allowed to combine with each other severally and be placed between two consecutive words, such as: one or more spaces, a comma and a space, a space, then a comma or a bracket or colon, and then a space, etc.

Références

- ALSHARAF H., CARDEY S., GREENFIELD P., SHEN Y., (2004), *Problems and Solutions in Machine Translation Involving Arabic, Chinese and French*, Proceedings of the International Conference on Information Technology, ITCC 2004, April 5-7, 2004, Las Vegas, Nevada, IEEE Computer Society, Vol. 2, pp. 293-297.
- CARDEY Sylviane, GREENFIELD Peter (2006), *Systemic Linguistics with Applications*, in *Linguistics in the Twenty First Century*, Cambridge Scholars Press, United Kingdom.
- CARDEY S., EL HARROUCHY Z., GREENFIELD P., (1997), *La forme des mots nous renseigne-t-elle sur leur*

nature ?, in Actualité scientifique, AUPELF-UREF, La mémoire des mots, pp. 305-315.

- DEBILI Fethi, HADHEMI Achour, EMNA Suissi, (2002), *La langue arabe et l'ordinateur. De l'étiquetage grammatical à la voyellation automatique*, in Correspondances, (Vol. 71), Tunis : Institut de Recherche sur le Maghreb Contemporain.

GHENIMI Abdelouafi, (2003), *La langue arabe : un système de micro-systèmes*, BULAG N°28, Presses Universitaires Franc-comtoises, Besançon.

- GHENIMI Abdelouafi, (2007), *Conception d'un modèle de traduction arabe – français appliqué au domaine des mathématiques*, Doctorate Thesis, Université de Franche-Comté, Besançon.

- HADJ-SALAH Abderrahmane, (1987), *Linguistique et phonétique arabes*, in Applied Arabic Linguistics and Signal and Information processing, Hemisphere, P.C., New-York.

- TAIBI Nacéra, (2003), *Vers un système d'analyse morphosyntaxique de la langue arabe tolérant les fautes*, in Al-Lissaniyat, revue algérienne des sciences et technologies du langage, N° 8, Alger.