Arabic Language Automatic Processing for generating Automatic Textual translation to French

LASKRI Mohamed Tayeb *

Abstract

An automatic processing system of the arabic natural language is proposed in order to complement existing works on french language. We carried out, in this study, a relatively detailed univestigation of this language own characteristics, which may serve for morpho-lexical as well as for syntaxic and semantic analysis. The proposed system uses techniques relevant to the richness of the arabic language for each particular mode of analysis. Hence, we put forward a technique based on the use of root and scheme concepts for morphological anlysis. Whereas, casual marks and desinence concepts are used for syntaxic analysis and semantic based technique is also proposed for the last case.The system ,thus realised, allow us to process any text written in arabic via the generation, for each sentence, of an internal representation that permits either the answers to questions in the subject or the translation into a target language whose linguistic knowledges are pre-registred.
Satifactory results are obtained with this investigation compared to other well known techniques in the field such as conceptual dependence, Fillmore theory and oriented object approach.

***Key Words** : Artific Intelligence - Natural Language - Textual data processing - Semantic cases Arabic Language Objects - Conceptual dependence - Casual works*

INTRODUCTION

Natural language does not require from the user who is not expert in computer science any learning or additional knowledge.It rather, simplifies for him computer systems [Boubakeur86]. In fact , it is more likely that one would feel less unhibited to use french to ask a question from the machine than actually refer to its language. If the computer fails to understand his question, he generally tends to reformulate it until its final recognition by the system.This way, the user will gradually overcome all the difficulties and constraints imposed by the system.

Where as he might have been compelled to give up if he was to acquire, at first, a specialised language in order to formulated his question [winograd84]. In the course of this study we will be developping a system for arabic language automatic processing which would take into account the specific characteristics pertaining to this language while , at the same time, acknowleding the works of other scholars in linguistics or computer science .From the purely pratical point of view our aim is basically , to provide the system with a phrase in the arabic language as an entry and to obtain in exchange the corresponding phrase in a target language, french, which would be semantically equivalent to the initial phrase inserted in the system. The latter is then expected to identify the given phrase , and provide an adequate translation after having operated a language processing which would be described below.

Before the analysis of any phrase the proposed system, would have to follow a certain number of well ordered and necessary steps varing from the mere processing of a word to the whole phrase and its semantic interpretation. En the case of the majority of this type of systems the steps are well and range from the morpho-lexical analysis, syntactic analysis to the semantic one.

Morpho-lexical analysis

The main objective of this type analysis is to test the affiliation of a given lexical item to the selected linguistic domain, and to prepare all the important information pertaining to it and that may serve for a syntactic analysis . This domain is related to a lexicon which ideally should not include all the possible form of a word ; but only one (agreed upon to be the most representative) should be listed,with the other forms to be identified when applying flexibility rules [Coulon 86].

Conventional Representation of words

Most of the works devoted to the study of the French or English Language so far have adopted the representation of words in the Lexicon for singular infinitive concerning verbs. Generally speaking, it would suffice to add or delete a suffix morpheme to a root in order to obtain a new word in French.

• **Example** : the Words in french language :
étudiera, **étudi**ée, **étudi**ons,.may all be reduced to a commun root **étudi** .
After affixation of morphemes (in bold) to roots :

Blanc (m.s), Blanche(f.s) Blancs(m.p),
Blanche(f.p Cité (m.s), Citée (f.s)

However in Arabic, the case may be different in most situations as the deletion of a prefix or suffixdoes not constitute a valid proof to assert that the then obtained root is the final form to insert on to the lexicon.

Difficulties in representing roots in Arabic : For example, the derivation of the gender feminine or plural in the Arabic Langage is generally accompanied by a whole transformation of its structure and not by the addition of a morphene.

• Examples of words with their corresponding plural or feminine form :

How can we determine which form is the representative one in the lexicon when dealing with these Words. It becomes evident that we cannot depend on such a method of subdivision of words into root and their corresponding morphemes (prefixes or suffixes).

Morphological processing of roots.

The solution we will be providing here is very practical for *this kind of processing and is related to the* Arabic language as it retives upon the notion root which is widely used in the science of language.

a) the notion -root-

The root of a given word is a consonantal succession of three letters and from which we may derive a whole varietes of other words with a similar semantic nuance. Thas every word, in the Arabic, language, can be reduced to a three-letter- root fixed pattern.

• Example : from the root كتب we may derive the following stems : كاتب كتابة مكتب كتاب

b) the notion of fixed pattern

In order to generalize and set up the different rules regulating the processing of the Arabiclanguage morphology, grammarians have istablished a morphological patters based on a three-letter-system and which is conventionally know as فعل (FA ALA) (the act of doin and comprising all possible derivations. We will say, f example, that the equivalent of كتب (has written) is t proposed pattern فعل .
The commutation of the two words will give us :

ك : is the "fa" from the verb: ف
ت : is the " ain" from the verb : ع
ب : is the " lam" from the verb : ل

This may be to summaraised the following table :

ب	ت	ك
ك	ع	ا ا

Figure 1.Supperposition of word and its pattern

The number of patterns may be great or limited a may find a patter for every word.
These patterns may be discrived as mould where roots ar print new derivates (sec fig 3).

item	pattern	root
زكتبة	زاعكة	كتب
جأزعة	أعكةا	جزع
زثج¤	زاعك	ثج¤

Figure 2 Examples of stems with their fixed patter

a) *Root extraction*

As we have already mentioned a pattern ma mould commun a ceretain number of words. and a *remains to be cast is the root*

• *Examples : the words* مكتب معبر مسكن *have a comm* pattern مفعل and their successive roots are : كتب عبر سكنز

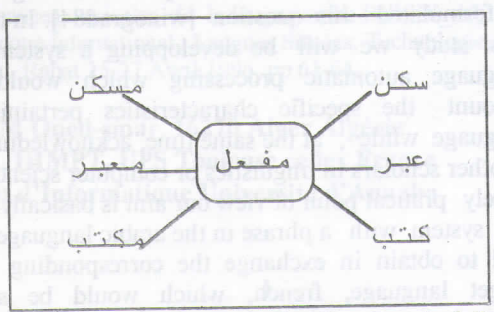


Fig 3. Root extraction examples from words following pattern مفعل

Here all the possible patterns (in a limited number) may we stored in a file (see fig.5)

أفعل	15
افتعل	16
استفعل	17
فاعل	18
...	...

Figure 4. part of a pattern file

Given a word, one may compare its form with those of the different pattern stored in the file in order to mathch its corresponding ones.

A superposition of these two entites will then enable us to identify the principal letters of the word(pertaining to the root).

• **Example** : concerning the word اقترب , we may conclude as a result of its superposition with its pattern which is افتعل (n° 16), that its root is is the three-letter word قرب the "fa" of the verb is : ق ; the "ain" of the verb is : "ر"; the "lam of the verb is ب

ب	ر	ت	ق	ا
ل	ع	ت	ف	ا

Fig 5. Root extraction of the word

Syntactic analysis

A phrase from the Arabic language differs from an English or French phrase, in that the former with its rigid structuring allows very little room for the application usual methods of automatic processing concerning the other languages, and there requises a special processing which presents a great degree of simplicity when compared with other languages. Among these aspects we may may mention its predisposition for word permutation within a phrase thanks to presence of affix markers in word endings.

Importance of casual markers

In the light of this approach, we may, in addition tothe use of the classical method in syntactic analysis widely spread in the study of english french, refer to a complementary method based on the recognition of several syntactic cases by way of casual markers.

The syntactic function of each word within the phrase is then identifiable through its ending and this proves the ulmost importance of these markers.

• **Example** : the presence of the short vowel "Fatha" at the end of the word.enables us to identify its complement as a direct object of the transitive verb whatever his position with in the phrase may be, whereas the short rowel (Dhama) at the end of the word refes to a subject position with in the phrase in question..

Semantic Analysis

Semantic analysis plays a major role in the study of natural language, in the sense that it helps us derive the meaning of surface structures especially when specific tools and methods are used to extract these meanings.

Difficulty of internal representation

It is essential to set up an adequate internal representation of the varions connotative and denotative aspects linked a specific meaning.

The problem to solve here resides in how to establish a pattern of internal and universal representation to describe any idea, independenty of its written or oral surface structure?

The very essence of this question and the difficulties tied to its answer emanates from the fact human beings do not preserve entire phrases inther memory, but just the ideas conveyed behind these phrases. On the other hand it is possible to retransmit the same idea, but with different phrases and even is another language such as Arabic or french provided one has a little knowledge of the rules of the language in question. Thus, an idea may be expressed is french, English, Spanish or chinese, but it always conveys a unique internal representation.

Semantic analysis of a phrase in Arabic

As any phrase in any given language, a phrase in Arabic is also a linguistic background which carries a certain meaning to be determined : the whole process of understanding this idea, despite the linguistic analysis it may nvolve, will consist of a set of operations allowing us to move from its external structure to its internal representation. " And given a phrase in a natural language, one may be interested to discover its constituants, the various relations existing between them, and on the basis of this one may them establish the internal strucure reflecting its meaning " [Bonnet4].

Identification of semantic cases

It is important to note that the constituants of a given phrase contribute to the coining of the corresponding idea even if taken in isolation each constituant may convey adifferent meaning. Thus, in order to understand the meaning of a phrase, it is essential to identify the semantic case and

the role played by each and every constituant within the said phrase.

a) *The fillm ·e approach*

Our interest in this particular theory lies in the fact it introduces common features that suits not only description of Arabic but also other languages such as french or English where syntatic cases may be selected to form a paradigm in natural language : " traditional categories such as subject and object are only surface manifestations of the functions of greater fundamental cases " [Bonnet 84]. The central focus in casual grammars consist in attributing the v··b a special role and consider it as the principal constituant of the phrase where by the study is completed with an analysis the relations between the nominal groups and the verb.

• Example : the verb WALK requises the presence of an agent who executes the action of walking, a point of departure from which he started and a destination towards which he is walking. involve consist of aset of operations allowing, will us to move from its external structure to its internal representation. " And given a phrase in a natural language, one may be interested to discover its constituants, the various relations existing between them, and on the basis of this one may them establish the internal strucure reflecting it· meaning " [Bonnet 4].

Identification of semantic cases

It is important to note that the constituants of a given phrase contribute to the coin ng of the corresponding idea even if taken in isolation each constituant may convey a different meaning. Thus, in order to understand the meaning of a phrase, it is essential to identify the semantic case and the role played by each and every constituant within the said phrase.

a) The fillmore approach

Our interest in this particular theory lies in the fact it introduces common features that suits not only description of Arabic but also other languages such as french or English where syntatic cases may be selected to form a paradigm in natural language : " traditional categories such as subject and object are only surface manifestations of the functions of greater fundamental cases " [Bonnet 84]. The central focus in casual grammars consist in attributing the verb a special role and consider it as the principal constituant of the phrase where by the study is completed with an analysis the relations between the nominal groups and the verb.

• Example: the verb WALK requises the presence of an agent who executes the action of walking, a point of depart· from w'.ch he started and a destination towards which he is walking.

b) *Semantic cases in Arabic*

The nature itself of the Arabic language with its casual grammar facilitates the identification of semantic cases.

• **Example** : In the phrase برمج المهندس الحاسوب (the enginner has programmed the computer) we may easily identify المهندس as the agent who executes the action برمج; الحاسوب as the direct object of the executed action and the marker dhamma indicates that the verb برمج is an active one.

Hence the method of identification of semantic cases in Arabic may be explained in terms of the rules which helps detect the semantic case of each word composing a phrase. A few examples may be given here :

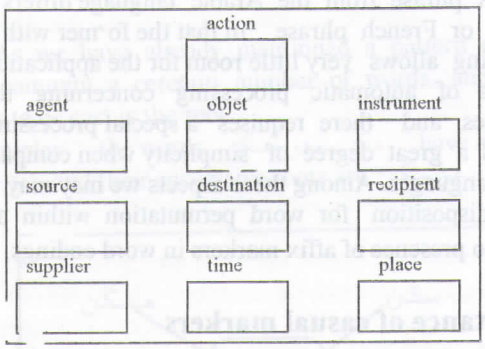
• Example of AGENT : syntactic case = ject
• Example of OBJECT : syntatic case = object complement
• Example of INSTRUMENT :grammatical case : dative and prepartive
• Example of SOURCE : grammatical case : dative and prep = من a place noun serving as direct object of some well known.

verbs such as غادر تارك as in غادر المسافر المدينة:

Internal organisation of the phrase

We have concentrated our choice on frames partly because we agree with Minsky that the method of frames is the closest to the one used by human beings.

Generally speaking a frame comprises a whole set of slots leaving room for the different concepts contained in a phrase to be represented. Figure7, shows to a certain degree of generalisation the various aspects of this representation..



F'gure 6.General frame for representing a phrase in natural language .

A phrase having an active verb such as مشى implies the filling of several useless slots that cannot be used as representing directly the object, the insrument and recipient.

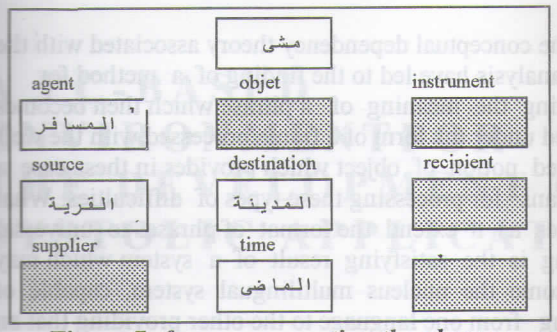


Figure 7. ***frame representing the phrase*** :
مشى المسافر من القرية الى المدينة

Verb classification and conceptuel dependency

One of the most respected theory in the field of natural language understanding is Roger Schank's conceptuel dependency theory in which he suggest a verb classification This method of classifying verbs in categories has led to the reduction of each group of verbs into a primitive which would stand as the representative and which would allow a common processing for all the verbs within the group instead of duplicating it for each verb. The set of these primitives must then be reduced and at the same time, enable us to all the important aspects relevant to the purport of the meaning.

We will then given verbs such as اخذ and سلم which means respectively (take and give) have to select primitive ATRANS (possession transfer) as these two verbs both implies the transfer of something from a supplier to a recipient اخذ : of which the recipient is the agent him self
سلم : of which the supplier is the agent him self.

One solution would be to list before each verb the name of its primitive and in its combination with other thems within the phrase, and then refer to its field (fig.8)

اخذ	ATRANS	.....
سلم	ATRANS	.....
سمع	ATRANS	.....

Figure 8. Portion taken from the dictionary of definitions

Practical use and role relations of objects

Having clearly explained the representation method we have adopted it is now important to select a structure of adequate data permitting its praticad applicability : we have chosen the object formalism which is widespread because of its consistency and clarity. One way of arriving at this is to state each primitive as object of which the facets are the names of all the different possible cases (slot names).This means that to each primitive will correspond an object.

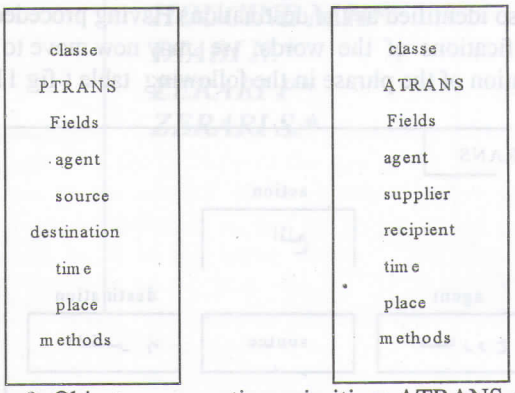


Figure 9. Objects representing primitives ATRANS and PTRANS

Moreover, one may note several fields such as action, time and place, are reproduced with each object, in their capacity of being comman to all phrases. Another adequate solution taking into consideration the advantages of objects would consist of stating a phrase class which would contain all the features comman to every phrase and of which the sub-classes constaining only essential elements would be the objects describing the primitives (fig.10). This way, we have insured that each, instance of one the primitives proceeds from the mother class which is PHRASE.

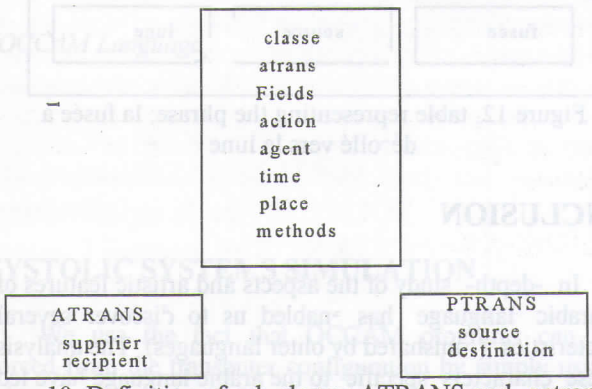


Figure 10. Definition of sub-classes ATRANS and qptrans from the class PHRASE

Phrase translation

As we have already said, of the main objectives of this study is the automatic translation of a phrase from the arabic language to the french language which requires as a prerequisite a good understanding to generate with out ambiguity its internal representation.

- Example : given the following phrase for translation قلع الصاروخ نحو القمر (the space ship has taken off to the moon).

the word اقلع is immediatty identified as the action of the

phrase the word الصاروخ being nominative accompanying as active verb is identified as the agent of the action..the word نحو being a particle introducing the name of a place which is القمر is also identified as the destinations.Having proceded to the identifications of the words, we may now move to the representation of the phrase in the following table (fig 11).

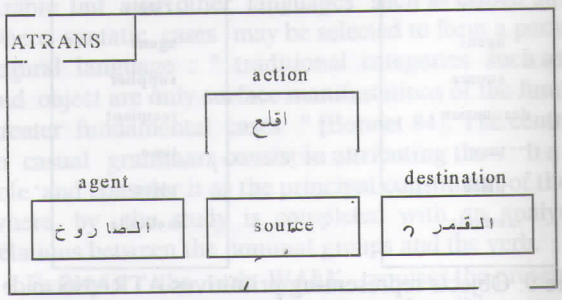


Figure 11 table representing the phrase اقلع الصاروخ نحو القمر

Once we have generated the internal representation, we need to substitute each slot by its equivalent in french. This gives the following automatic transla ions summed up in this figure

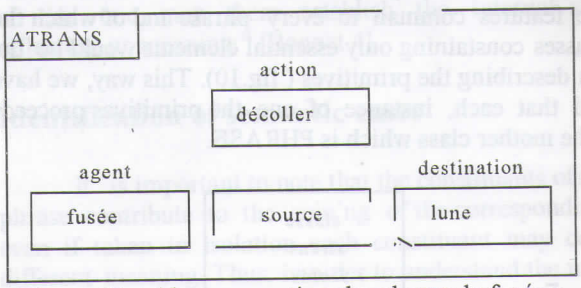


Figure 12. table representing the phrase: la fusée a décollé vers la lune

CONCLUSION

In -depth- study of the aspects and artistic features of the arabic language has enabled us to discover several characteristics (unshared by ohter languages). The analysis of these characters specific to the arabic language have led us to the setting up of a particular method (or, let us say, suveral methods if we consider the different layers of analysis) for an easier and more flexible processing.

Our adequate use the spread notion of roots and fixed paradigms (neglected difficultives in ther s udies) have enabled us to solve a good number of difficultives which were persistant in the construction of words and phrases and their transformations.

The notion of inflexion, casual markers affixed to word ending has contributed to the identification of several syntactic structures whatever their position may be.

And finaly the casual aspect of the arabic language has provided a good asset for the semantic treatment.

The conceptual dependency theory associated with the fillmore analysis have led to the finding of a method for representing the meaning of a phrase which then becomes formalised under the form of a frame processed with the well appreciated notion of object which provides in these case a good means for processing these types of difficulties. What encourages us to extend the format of phrases to universal processing is the satisfying result of a system which may well become the nucleus multilingual system, capable of translating from one language to the other providing that an adequate inter face might be embedded on to it.

REFERENCES

BOUBAKEUR O. (1986) Al-Semss :Un système expert en morphologie, syntaxe et sémantique pour l'étude de la langue arabe. Thèse 3ème cycle, Toulouse France

COLMERAUER A., KANOUI H., PASSERO R. et ROUSSEL Ph. (1973) .Un système de communication Homme-Machine en Français. Rapport de recherche sur le contrat CRI no72-18Groupe en Intelligence ARtificielle. Luminy, France

COULON D.(1977) Description générale d'un système de réponse aux questions.
CRIN Rapport technique. Nancy, France

JAYEZ J.H. (1984) Aspects de la compréhension des langues naturelles. l'informatique professionnelle, N°19 janvier. pages 55-71

KAYSER D. , FOSSE P. , KAROUBI M. LEVRAT B. et NIGAUD L. (1987) A strategy for reasoning in natural language. Vol. 1 N°3. Pages 205-231

KOULOUGHLI D.E.(1987) Vers un traitement automatique de la prédiction verbale en arabe standart moderne. T.A. informations. Revue internationale du traitement automatique du langage. Vol. 28 N°1 Paris, France.

LASKRI M.T. (1991) Préparation automatique d'un domaine d'application dans le cadre d'un système support de thésaurus à langage naturel. Conférence Internationale sur la science des systèmes informatiques. Marseille, France.

LASKRI M.T., BOULAKRADECHE M. et KNIPPEL J.M. (1994) Analyse du langage naturel à base de connaissances. Colloque International "Consensus Ex Machina" Paris, France

LASKRI M.T. (1995) Sémantique du langage naturel à travers un système support de thésaurus. Thèse de Doctorat d'Etat en Informatique. Annaba, Algérie.

MEHDI S.A. (1986) Arabic language parser. Man-Machine studies. Vol 25 N°5 pages 593-611. Londres , England.

*** Institut Informatiqu - University of Annaba Alger**