Linguistics Tools to Develop an Arabic Syntax Analyzer

Prof. Dr. Mohamed El Hannach
Arabic Linguistics Engineering Society in Morocco
P.: 2535 (c) of Fez - Morocco
elhannach@yahoo.com

***Abstract*— This study presents a formal description of a linguistic system for developing an Arabic syntax analyzer based combinatorial grammar. In this context we will present several formal concepts to redefine the structure of the Arabic language in order to adapt it to the automatic processing of at syntactic level, such as support verb, nominalization, and other operators.These concepts provide an effective framework for the implementation of language engineering techniques aimed at integrating the Arabic language within the NLP community**

***Keywords: support verb; nominalization; labeling; distribution; processing; parser; database*.**

I. INTRODUCTION

This study presents a formal description of the linguistic system for developing an Arabic syntactic analyzer based on the theoretical foundations of combinatorial grammar. Combinatorial grammars are known for their power to support the production of language analyzers, many of which are already operational for a good number of natural languages. In this context we will present several formal concepts to redefine the structure of the Arabic language in order to adapt it to the automatic processing at the syntactic level. These concepts include support verb, nominalization, inflected passive restructuring, and various operators. These concepts provide the appropriate environment for the implementation of language engineering techniques aimed at integrating the Arabic language within the NLP industry [1]. Our system syntactic analysis takes place in Arabic NLP works already achieved at different research centers in the world, especially in France, where parsers were implemented for different natural languages including Arabic. We cite as an example [2]: Nooj, Gaspar Monge Laboratory, Institute INRIA and LADL where the data bases have been developed for parsers that are already operational in different applications on the Internet. To give an operational aspect of the present study, we present a database of syntactic structures of Arabic language, using the same techniques already used in research on NLP applied to many natural languages. Our study will be focused on the syntactic distributional matrix of this language that we schematize as follows:

No V W where: W = 0, N1, N2

This structure must generate several transformational structures, represented as follows [17]:

a. Passif: *V No N1* ≡ V-u N1 Prep No
b. Adjectivation: *V No N1* ≡ No V-a Prep N1
c. Nominalisation: *V No N1* ≡ Vsup No V-n Prep N1
d. Restructuration: V No N1 ≡ V No GN (Na Nb)1

We will focus the present study on 'Nominalization' which constitutes a fragment from our large Arabic syntatic database. This syntactic property is the most productive class of verbs expressing a feeling. This work is part of our regular work that we carry out the engineering of the Arabic language in the context of our close collaboration with various research centers in the world [16], [19].

The Arabic syntactic level is a major challenge for Arabic NLP, because of its intersection with the morphological fusionist level, as opposed to the morphological system of European languages. The simple Arabic word consists of a root, either triliteral (three consonants =: ف ع ل) or quadrilateral (four consonants =: ف ع ل ل), and a pattern (= : *al-wazn* and *al-mizan*) [14].The root supports the discretization; that is to say the distribution of vowels on the consonant of the root; it also supports various additions, such as (=: Prefix Suffix and Infix) [14].The morphological system, thus conceived, allows writing the language without vowels, but people read it easily via their morphological competence. Vowels and affixes extend the meaning of the word, before assigning its appropriate lexical and grammatical category, such as: Name: N (noun), V-n (deverbal noun) and V-a (Adjective)[1] or Normal Verb, Support verb, and idiomatic verb. The distribution of vowels and affixes on the components of the root system is via a specific algorithm. Assigning a functional vowel to the first consonant of the root is sufficient for the scheme to determine the lexical or syntactic class of the word[2].

[1]. We adopt the following notation: V: verb, V-n: deverbal noun, V-a: adjective
[2]. We emphasize that the syntax constitut made the main frame of the generation of words, the word in Arabic can not have autonomy outside the combinatorial syntaxiquem and this is, in part, to the absence of the form verbal infinitive in English.

II. SYNTACTIC THEORY

This brief introduction shows that the parsing of Arabic must take into account the morphological level in the automatic processing of the syntax of the language in order to reduce the number of errors that may be encountered during the development of the parser. Thus, we started our parser with the construction of an electronic dictionary of simple and inflected words; this dictionary[3] greatly facilitated our task; it contains over 2.5 million lexical entries coded according to the tables system. Each one of our tables represents a finite state automaton.

In order to develop a reliable parser for Arabic, we combined these approaches with basic descriptive theoretical and methodological principles as follows [5]:

1. The verb is the minimum syntactic unit (V =: simple sentence); this principle is based on the absence of the infinitive form in Arabic, because the verb is always combined with its subject. This means that strictly the form V isolated in the syntactic context does never occur in Arabic. Thus, the following sentence is accepted =: oقلمس while P =: ø قلم* is unacceptable. It is due to the absence of the subject which means that V=: قلم with no subject is beyond the syntactic classification of Arabic. This means that the strict system of the Arabic language is intrinsically syntactic. Moreover, the following three types of names:

   N=: اسم ذات,
   V-n-: مصدر,
   V-a=: صفة

   never find their autonomy outside the syntactic structure. This fact acquires a basic role in the linguistic system of the language.

2. The verbs are divided into three major grammatical categories; ordinary verb, support verb and idiomatic verb. The ordinary verb commands a simple structure as in:

   V No N1=: أدهش هذا الأمر عليا

   As for the support verb (*Vsup*), it is generated by a syntactic operation called nominalization, as applied to the basic sentence above:

   V No N1 ≡ Vsup No V-n Prep N1
   أثار هذا الأمر الدهشة في علي ≡ أدهش هذا الأمر عليا

A third type of verbs appears in idiomatic structures (مسكوك)[4]. For exemple: Ahmad was killed

لفي أحمد حتفه

The phrase thus constructed cannot undergo any morphological or syntactic operation, because of its structural rigidification from its linguistic form called opaque[5].

3. We adopted the syntactic structure matrix that begins with a verb, represented by a verb phrase, excluding the nominal sequence that begins with a name. This process allows us to build a database of syntactic structures limited in number and in combination. We identify these structures as follows[6]:

   a. V No
   b. V No N1
   c. V No Prep N1
   d. V No N1 Prep N2
   e. V No Prep N1 Prep N2

These five basic syntactic forms represent all syntactic classes of Arabic, where the verb as predicate plays the role of a function whose arguments are the N and Prep N. The cooresponding formula may be represented as follows:

$$P = V (x, y).$$

The distributional rules on these five basic shapes allow us to generate thirty-five syntactic classes, each of which admits a set of properties that make it a transformational syntactic class wich the requirement that the verbs must belong to a common semantic field. The base class (b) above result in the following classes[7]:

b1) V No +hum N1 +hum =: ضرب أحمد غليا
b2) V No +hum N1-hum = N1: كتب محمد رسالة
b3) V No nr N1 + hum =: عليا (هذا الأمر, محمد) أدهش
b4) V No-hum N1-hum = حطمت الرياح الأشجار

This distributional system of the basic structures enables us to build a database containing all Arabic syntactic forms, thus forming a lexicon-grammar that covers a large size.

The example we present below clarifies our approach already applied to several natural languages. Consider the structure (b3) above; it is characterized by a set of transformational and distributive properties that clearly distinguish it from other classes of verbs in Arabic.

1. Distributional level: this class is characterized by the presence of the noun subject of the non-restricted[8] form, which allows its complementary distribution with a regular completive. For example:

[3] DEMAS (Dictionnaire électronique des mots arabes simples), et DEMAF (Dictionnaire électroniques des mots arabes fléchis).
[4] . Cf. El Hannach 1990
[5] . Cf. El Hannach, Syntaxe des verbes qualitatifs, 2001
[6] . We use the following conventions: No stands for subject; *N1* for first object; *N2* for second object; *Prep* for preposition; *N+hum* for human noun; *N-hum* for non-human noun; N nr for indeterminate noun No =: Subject, N1 =: First object, N2 =: Second object, and Prep =: Preposition
[7] . We derived four sub-classes, but our study will be focused on b3.
[8] . The undeterminate noun is an open distributional position, where we can put any noun: N+hum, N-hum, and 'that P =: Completive.

b3) V No nr N1+hum =: عليا (هذا الأمر, **محمد**) أدهش

is equivalent to:

b3 ') V N1+ hum ('anP)1 =: أدهش عليا أن يرسب في الاختبار

The completive performs the function of the subject, thus that of a full noun. This noun triggers the feeling of astonishment experienced by N1+hum =: عليا. Note that other distributional properties are brought into play for this class; they clearly distinguish it from the rest of the Arabic syntactic classes[9].

2. Transformational level: we note that this class also possesses its own properties which induce a well defined database of syntactic structures. We were able to identify three transformational properties as follows:

1. Adjectivization:

b3-a) V No nr N1+hum ≡ N1+hum V-a Prep N0 nr

علي مندهش من (**محمد**, هذا) ≡ أدهش (**محمد**, هذا) عليا

3. Passif:

b3-b) V Nonr N1+hum ≡ RefV N1+ hum Prep N nr

اندهش علي من (**محمد**, هذا) ≡ أدهش (**محمد**, هذا) عليا

4. Nominalization[10]:

b3-c) V No nr N1 ≡ Vsup No Vn Prep N1

أثار (**محمد**, هذا) الدهشة في علي ≡ أدهش (**محمد**, هذا) عليا

Note that the operation represented by the active nominalization b3-c creates three types of complex nominal groups, each of which is composed of at least two elements: GN = Vn Prep N1, GN = (V-n)1 (Prep N1)2, which requires a special syntactic processing to insert in our Arabic parser.

III. EXPERIMENTAL RESULTS

In this part of our study we present a fragment of our large database on which was focused our analysis[11]. As a component in the development of a parser based on the class of verbs b3-c above, we need to build a table, grouping sets of syntactic properties (P), and to prepare them for automatic processing by the Dyalog program, specializing in the analysis of natural language syntax tables. Given the high number of the syntactic properties (about 100 operations divided into properties and sub-properties), we split the table into three sub-tables: Noma, Nomaj and Nomap[12]. Each sub-table contains the derivatives of the syntactic form of base labeled b3-c This subset from the database NOMA presents tabular matrix that is composed of three columns: the first group contains a single column containing the lexical entries (=: Verbs) is considered as the constant element of the table; a second group is allocated to the distributional properties of the subject, and a third group contains the active nominalization properties. The properties expressed in this case are represented by a metalanguage describing the structures generated by the application of the nominalization operation. The entries containing the operators (+) and (-) express the result of applying the operation nominalization of verbs (verb =: simple sentence). Thus, each entry may represent a finite state automaton. The verb is the input, and the derived sentences are the output. Such a scheme facilitates the task of Dyalog [21].

Columns and rows

We refer to Table 1 in this discussion. Each column contains a single syntactic property. It can be interpreted as representing a structure in which a verb can occur or not. Thus, we have two types of columns: those containing the main properties of the operation of nominalization and those containing sub-syntactic properties of the same operation. Each row corresponds to a form of syntactic paradigm for testing the verbal input. An entry containing the (+) sign indicates that V have the property P; the minus sign (-) indicates that V does not have the property P. Other possible symbols (e. g. *, ?, *?) are completely excluded from our sub-tables ([13]). Note that the first group of columns of our sub-tables is constructed by simple morphological units, i.e., by verbs and non-verbal expressions. The verb is always conjugated at the past aspect of the third person singular [20].

IV. CONCLUSION

We presented a formal description of a sample of our linguistic database as a basis for the development of an Arabic parser. Our analysis is considered as an application of the theoretical framework named combinatorial-Grammar developed at Pennsylvania University by Z. S. Harris ([14]), and methodological framework named lexicon - grammar developed at LADL [15] and IGM [16] by M. Gross. We focused our analysis on the class of verbs of the form show in b3-c. This class is characterized by a subject unrestricted regularly, wich is in complementary distribution with a completive subject (No nr=: *'anP or ' anna) P*), and a direct object (N =: human). Both distributional properties are supported by the operation of the main verb

[9] . Cf. El Hannach 1988, and M. Gross 1975

[10] . We have three types of Nominalizations in Arabic, but we focused our study on Active Nominaliszation, See El Hannach 1988.

[11] . See the sample of data base below.

[12] . Noma=: Nominalization active, Nomap=: Nominalization passive, and Nomaj=: Nomonalization adjectivale.

[13] . M. Gross 1975

[14] . Cf. Structure mathématique du langage, éd. Dunod 1970.

[15] . Laboratoire d'Automatique Documentaire et Linguistique, CNRS, Université Paris 7, France.

[16] . IGM: Institut Gaspard Monge, Paris Sud University.

nominalization becomes a name مصدر (=*: V-n*) by the insertion of one of the following *Vsup*: (=: *sabbaba 'Athara, xalaqa, 'ahdatha, ba'atha' a'ata 'adhfa, Jalaba*). Each of these *Vsup* triggers a semantically equivalent syntactic structure with 3b-c.

Our syntactic processing allows us to build a database based on the extension approach (as opposed to the intension approach adopted by the generative grammar). Our tables (i.e., matrices) are presented in a matrix composed of columns representing the distributional and transformational proprieties, and rows representing the results of the application of the properties to lexical entries. The treatment of these data, constructed as a finite state automaton, is carried out by the Dyalog software, developed at IGM, Paris-Sud University.

REFERENCES

[1] Farghali, A., Arabic Computational Linguistics. CSLI, 2012.

[2] Tolone, E., Analyse syntaxique à l'aide des tables du lexique-grammaire du français. PhD dissertation, University Paris-Est, 2011.

[3] Sagot, B. and Alii, "Intégrer les tables du lexique-grammaire à un analyseur syntaxique robuste à grande échelle," in TALN, 2009.

[4] El Hannach, M., "Syntaxe des verbes psychologiques de l'arabe." Thèse de Doctorat d'Etat, 1988.

[5] Gross, M., Méthodes en syntaxe. Paris: Hermann, 1975.

[6] de la Cleregeri, E. and Alii, "FRGM: Evolution d'un analyseur syntaxique tag du français." Rapport INRIA, Université Paris 7, 2010.

[7] Silberztein, M. "Dictionnaires électroniques et analyse automatique de textes." Masson, 1993.

[8] Al-Ghamdi, M. and Alii, "Automatic restoration of Arabic diacritics : A simple, purely syntactical approach." 2010.

[9] Blanc, O. and Alii, "Journées lexique-grammaire et lexique syntaxique et sémantique." Rapport IGM, 2009.

[10] Harris, Z., Structure mathematique du langage. Paris: Dunod, 1971.

[11] Muller, M.-P., "Langage-grammaire-automates." Les Mathématiques. net, 2005.

[12] Senellart, J., "Outils de reconnaissance d'expressions linguistiques complexes dans des grands corpus." Thèse de doctorat, LADL, 1999.

[13] El Hannach, M., "*Syntaxe des verbes qualitatifs de l'Arabe*", 2001

[14] Neme, A., "*A lexicon of arabic verbs constituted on the basis of semantic taxinomy and using finite-state transducers, affiliation information*" IGM, 2011

[15] ILF, *"LexSynt (Lexique syntaxique et interface lexique-grammaire*,", 2006

[16] LADL, *"Les bases de données du LADL: Analyse automatique des langues naturelles, Aspects technologiques*." Paris 1989

[17] LADL, Select paper on Lexicon – Grammar" Vol. : 1-3, Paris (1973-1999)

[18] LEON, J., *Traitement automatique des langues*, Paris 2001.

[19] Laporte, E., *"Traitement automatique du mot: Etat de l'art*" IGM, 2001

[20] Silberztein, M., "INTEX", ASSTRIL, Paris. 1999-2000

[21] IGM, *Unitex, Programme de recherche unicode*, Paris, 2002

Nominalization =: *Vsup N0 nr V-n Prep N1 +hum*																				No						Verb
																				Completive						
Jalaba No Det V-n Li N1	'adhfa No V-n 'ala N1	'a'ta No Det V-n Li N1	Ba'atha No Det V-n Fi	Harraka No V-n Fi N1	Harraka No Det V-n Fi	'ahdatha No V-n Li N1	'ahdatha No Det V-n Li	'adxala No V-n 'ala N1	'adxala No Det V-n 'ala	Xalaqa No V-n Li N1	Xalaqa No Det V-n Li	'athara N1 (V-n N1)	'athara No Det V-n Fi	'athara No V-n Fi N1	Sabbaba No (V-n	Sabbaba No Det V-n Li	Sabbaba V-n Li N1	No=: V-n W	No=: Kawn P	No=: 'anna P	No=: 'an P	No=: Nnr	No=: +Concret	
+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	+	+	طرب
-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	+	+	+	+	+	+	-	أطفأ
+	-	+	+	+	+	+	-	-	-	+	+	-	-	-	-	-	-	+	+	+	+	+	-	أظلم
-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	+	-	أكل
+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	أكمل
-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	ألجم
-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	ألزم
+	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	ألم
+	-	-	+	+	+	+	-	-	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	ألهب
-	-	-	+	-	+	-	+	-	-	-	-	-	-	+	-	-	-	+	-	+	+	+	-	ألهم
+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	+	-	+	+	-	+	+	+	-	أمات
+	-	+	-	-	-	+	+	+	+	-	-	-	-	-	+	-	+	+	-	+	+	+	-	أنحل
-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	+	+	أنذر
+	-	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	أنعش
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	+	+	+	+	أنقذ
+	-	+	+	+	-	+	-	-	+	+	+	+	-	+	-	+	+	+	+	+	+	+	-	أنشي
-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	+	-	+	+	+	+	+	+	+	أنهك
-	-	-	+	-	+	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+	+	+	+	أنهض
-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	+	+	أصاب
+	-	-	+	-	+	+	+	-	-	-	+	+	-	+	+	-	+	+	+	+	+	+	+	صحى
-	-	+	+	-	+	+	-	+	+	-	-	-	-	-	-	-	-	+	-	+	+	+	-	أصلح
+	+	+	+	-	+	-	-	+	+	+	+	-	-	+	-	-	-	+	+	+	+	+	-	أضاء
+	+	+	+	+	+	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+	+	+	+	أضجر
+	-	-	+	-	+	-	-	-	-	-	-	+	-	+	-	-	-	+	+	+	+	+	+	ضحك
+	+	-	+	-	+	+	-	-	+	-	-	+	+	+	-	-	-	+	+	+	+	+	-	أضرم
+	+	+	+	-	+	+	+	-	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+	أضني
+	+	+	+	-	+	+	-	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	-	ضعف
-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	+	+	+	+	+	+	أعاق
-	+	-	+	-	+	-	-	+	+	+	+	+	+	+	-	-	-	+	+	+	+	+	+	أعجب
+	-	-	+	-	+	-	-	-	-	-	+	-	-	+	-	-	+	+	+	+	+	+	+	أعدي
-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	+	+	أعلي
-	-	+	+	-	-	+	-	-	-	-	+	-	-	-	+	-	+	+	+	+	+	+	-	أعمي

Table 1: Sample of the table of the class of active *Nominalization* (b3-c)