

# The experimentation of a HPSG grammar for the Arabic language on the LKB system

Sirine BOUKEDI\*, Nouredine LOUKIL\*\* and Kais Haddar\*\*\*

\* National Engineering School, Sfax, Tunisia. Email: serine\_fss@yahoo.fr

\*\* National Engineering School, Gabes, Tunisia. Email: noureddine.loukil@isimsf.rnu.tn

\*\*\* Sciences Faculty, Sfax, Tunisia. Email: kais.haddar@fss.rnu.tn

**Abstract**— NLP (Natural Language Processing) is a very delicate domain. It covers four levels of treatments: lexical, syntactic, semantic and pragmatic treatments. The Arabic language didn't benefit from the recent research especially in the syntactic level. In this context, our work lies within the conception of a HPSG (Head-driven Phrase Structure) grammar that covers the different syntactic phenomena of this language. The established grammar is based on a type hierarchy permitting to take into account the Arabic language's specificities. It is specified on TDL (Type Description Language) and tested on a parser generated by the LKB (Linguistic Knowledge Building) system.

**Keywords**—Type hierarchy; HPSG grammar; LKB; TDL

## I. INTRODUCTION

The syntactic analysis is a fundamental treatment in NLP (Natural Language Processing) domain. In fact, it represents the primordial phase for other analysis as the semantic one. It's also necessary for several particular applications such as the systems of dialogue Human-Machine, automatic translation and grammatical correction.

Despite this importance, the syntactic analysis has not been spilled in the research domain, especially for the Arabic language, except a few works that studied some particular phenomena as [1], [3] and [12]. Indeed, the syntactic treatment is very delicate. It requires a checking of some constraints and a great complexity in term of time and efforts. Moreover, it is very difficult to decide for the adequate grammar that treats effectively the various Arabic syntactic phenomena. Nowadays, some researchers are looking for some tools (i.e., the generation tools and the heuristic ones) simplifying the parser's construction.

The main objective of this work is to establish a grammar HPSG (Head-driven Phrase Structure Grammar) for the Arabic language that will be specified in TDL (Type Description Language) and validated by the LKB (Linguistic Knowledge Building) system. This grammar is very interesting. In the one hand, it models the different grammatical principles, in the other hand, it represents lexical entries. This representation differs from an entry to another according to the entry's type. Therefore, the originality of this work is located in the identification of a type hierarchy classifying the different Arabic unities. Based on this hierarchy, we bring some modifications on the HPSG to the lexical and syntactic level.

In the following paper, we start by some previous works focused on the syntactic analysis. Then, we propose a type hierarchy for the Arabic language. Based on this hierarchy, we present the established grammar HPSG as well as the different modifications brought to make it compatible with Arabic. Finally, we validate this grammar with the LKB system that generates automatically a parser. This parser is tested on a corpus of sentences. According to the obtained results, we evaluate our grammar and we enclose our work by a conclusion and some prospects.

## II. PREVIOUS WORKS

The study on various previous works showed that there exist two principal approaches to construct a parser. The first approach consists in designing and implementing an own parser based on a very determined analysis algorithm. The second one consists in using a system that generates automatically a parser.

The first approach is important since it is expandable and it allows the designer to master the details of his parser. However, there is always a problem in the choice of the adequate algorithm that can guarantee the parser efficiency (i.e., Chart parsing algorithm). It's also very complicated and requires an elevated cost in term of times and efforts.

For the second approach, it is sufficient to establish an adequate grammar. The generator uses specific algorithms guaranteeing effective results. Thus, the cost is less expensive and the interface ergonomics is generally tested by experienced users. The only problems of this approach are in term of extensibility and maintenance. Indeed, it is necessary to have the right of access to generator's code.

Among the systems generating parsers based on HPSG, we can mention LKB [7] and TRALE, an extension of ALE (Attribute Language Engine) systems [13]. These two generators are similar in the reliability point of view. But in term of accessibility, the LKB interface is more ergonomic and easier to use. This explains the choice of LKB, in most works for parsers generation.

Researchers having constructed parsers are not numerous, especially for the Arabic language. Most of these works used the first approach. Among these researchers, we mention [6] who presents in his work the first results of SYNTAX, a corpus-based syntactic analyzer. The main objective of this work is to develop a new parser covering French phenomena. This parser was based on the different principles of LEXTER. It identifies

dependence relations extracting phrases. Another parser treats the phenomenon of Arabic relative, proposed by [9]. This parser is based on the HPSG. In the same way, we can mention the work of [1] that proposed a parser for the Arabic nominal sentences. The algorithm of analysis followed by the two works is based on an ascending methodology. Another work, SYNTAXE system, has been proposed by [4] to analyze Arabic texts. This parser is based on a standard algorithm: the Chart-parsing. All these works didn't treat the Arabic language in an effective and robust manner. The assessment of the grammars used is difficult. Besides, the graphic interface requires an improvement for a better presentation, especially for the lexical entries.

Other works used the second approach. Most these works appeared in order to analyze the Latin languages as: [10] and [15]. Indeed, the first work proposed a parser treating the phenomenon of the Spanish relatives. The second work proposed a parser for the French language. The two works were based on HPSG and used LKB as parsers generator. The results obtained are more reliable than the results of the other works having followed the first approach. Therefore, we decide to validate our grammar with the LKB system. In the following paragraph, we present the type hierarchy that we have proposed for the Arabic language.

### III. PROPOSITION OF AN ARABIC TYPE HIERARCHY

The Arabic language is the spoken language by the Arab people. As any language, the Arabic has a grammar which is rather particular and so complicated. This grammar covers four branches (lexicon, morphology, derivation and syntax). In our work, we are interested at the syntactic level.

After a long discussion with linguists and based on some references as [2] and [8], we have proposed the type hierarchy represented in the figure below:

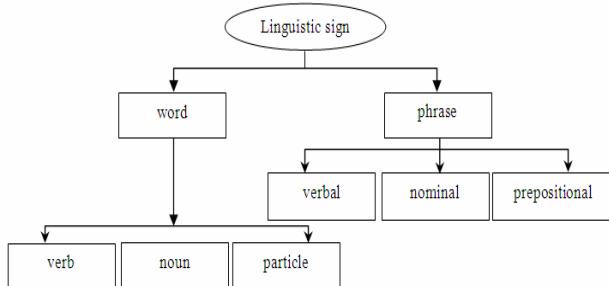


Figure 1. Arabic type hierarchy

Indeed, our study shows that the type root is the linguistic sign «اللفظ». It is subdivided into two sub-categorizations: word «كلمة» and phrase «مركب». A simple word (الكلمة العربية), can be a verb «فعل», a noun «اسم» or a particle «حرف».

According to [2], several criteria are presented to categorize an Arabic verb. It can be subdivided according to the number of letters that compose it or according to whether it is augmented «مزيد» or denuded «مجرد». We choose, in this article to subdivide the Arabic verbs according to the first criterion. Thus, a verb can be trilateral «ثلاثي» or quadrilateral «رباعي», as shown in «Fig. 2.»

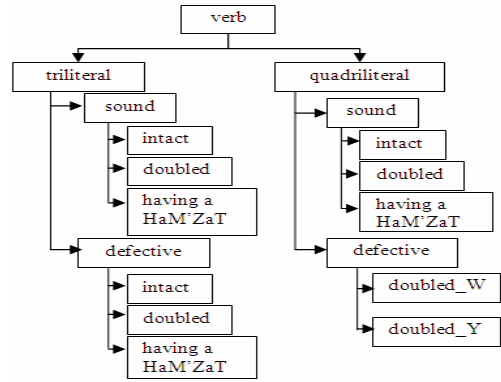


Figure 2. Verb's categories

The above figure shows that a trilateral verb or quadrilateral can be sound «صحيح» or defective «معتل». Each type has different possible values what makes possible to distinguish the various Arabic verbs.

For the nouns «الأسماء», we choose to subdivide them according to there declension «الإعراب». Thus, we find declined nouns «الأسماء المعربة» and indeclinable nouns «الأسماء المبنية», as shown in «Fig. 3.»

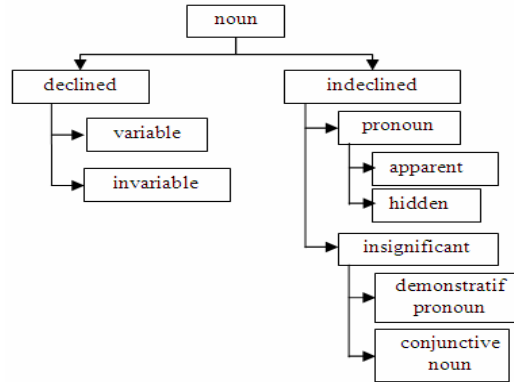


Figure 3. Noun's categories

In fact, a declined noun can be variable «متصرف» or invariable «غير متصرف». It is variable when it varies in the sentence, in gender and in number and invariable when it remains always invariant. Moreover relative pronouns «الموصولة الأسماء» and demonstrative pronouns «الإشارة أسماء» are considered in Arabic as insignificant nouns. They have a meaning only when they are connected with another declined noun.

For particles «الحروف», according to [2], we can classify them in two different categories. The first category represents operative particles «حروف عاملة», which influence either on the nouns or on the verbs. The second represents neglected particles «حروف مهملة» that don't have any influence on the verbs nor on the nouns. «Fig. 4.» illustrates the two distinguished categories.

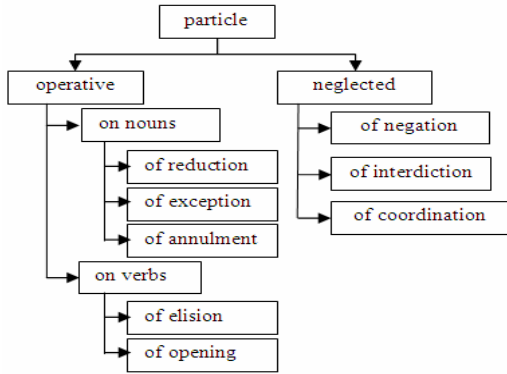


Figure 4. Particle's categories

Based on the proposed type hierarchy for the Arabic language, we have to adapt the HPSG grammar. In fact, the criteria characterizing every type vary from a language to another. Thus, it is necessary to add new criteria specifying an Arabic word. Besides, the word order in the Arabic sentence defers from the Latin sentence. Therefore, the syntactic rules composing the phrases will be different.

In the following paragraph, we present the different modifications made to HPSG, to make it compatible with Arabic language.

#### IV. HPSG FOR THE ARABIC LANGUAGE

HPSG is a unification grammar [14]. It is characterized by a reliable modelling of the universal grammatical principles and a complete representation of the linguistic knowledge. Indeed, each lexical entry contains different types of information (phonological, morphological, syntactic and semantic). Thus, HPSG takes into account a great number of linguistic phenomena and describes constructions with a limited number of operators.

HPSG grammar is based on two essential components: AVMs (Attribute Value Matrix) and a set of immediate domination schemata (DI schemata). An AVM is based on a set of features characterizing a lexical entry. According to the type of the word represented, these features vary from an AVM to another. The following "Fig. 5," represents the general structure of an AVM:

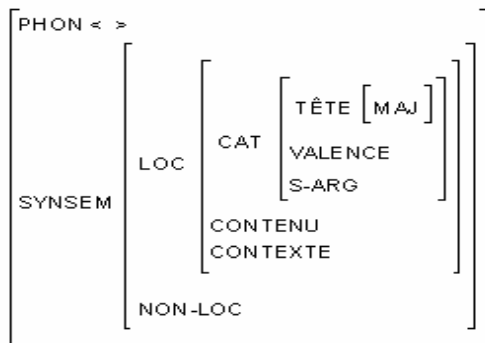


Figure 5. General structure of an AVM

As represented in this figure, each feature describes a type of information, like the features TETE and VAL. The first one covers intrinsic information. The second represents the various components that will be categorized.

For the DI schemata, HPSG is based on six different schemata (i.e., rule of specification 1). Each schema

describes a syntactic phenomenon. To compose the various phrases with those schemata, a set of principles (constraints to be checked) should be verified (i.e., HFP Head Feature Principle).

In fact, HPSG grammar was conceived for the Latin languages. In order to use it for the Arabic language, we present in the following paragraph the different modifications made to HPSG. These modifications were made on the features and schemata level.

#### A. Arabic item feature

Referring to previous projects [1], [4] and [9], we have kept some features and have added some others according to the proposed type's hierarchy. As we have already seen, a linguistic sign (word or phrase) can be characterized by its declension (الإعراب). Therefore a new feature: "DEC" is necessary to specify if it is declined (معرب) or indeclinable (غير معرب).

According to "Fig. 2," a triliteral or quadrilateral verb can be sound (سالم) or defective (معتل). Thus, the features, characterizing the verb type are gathered in the table below:

TABLE I. ARABIC VERB FEATURES

Features	Possible values
RADICAL	- triliteral ثلاثي - quadrilateral رباعي
VFORM	- sound صحيح - defective معتل
TYPE	- intact سالم - doubled مضعف
VOICE	- Passive منفي للمجهول - Active منفي للمعلوم
ASPECT	- accomplished ماضي - unaccomplished مضارع - Imperative أمر
ROOT	- the verb's root (جذر)

The different features characterizing the Arabic verb are mentioned on the HEAD feature (TETE) level. In "Fig. 6," above, we give an illustrative example of an AVM.

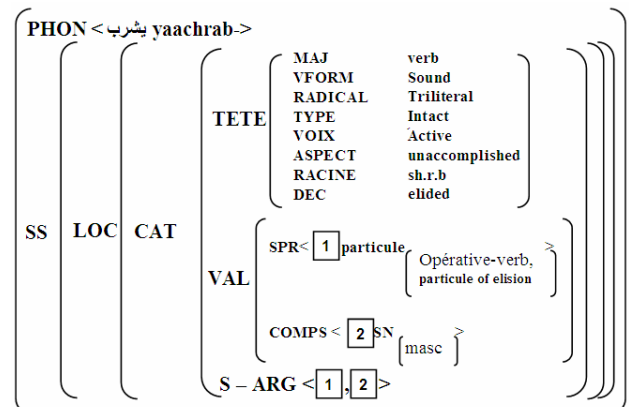


Figure 6. An example of AVM

Based on this AVM, we conclude that the verb « yaachrab- » (يشرب) is in an elided form (مجزوم). It is preceded by an elision particle (حرف جزم), (referred by SPR feature) and followed by a masculine noun (referred

by COMPS feature). The order of these two components is respected by the S-ARG feature.

According to “Fig. 3,” a declined noun can be variable (متصرف) as the common nouns or invariable (غير متصرف) as the proper nouns. For the indeclinable nouns, they regroup personal pronouns (الضمائر), conjunctive nouns (relative pronouns) (الأسماء الموصولة) and demonstrative nouns (أسماء الإشارة). Thus, the features characterizing the noun type are presented in table below:

TABLE II.  
ARABIC NOUN FEATURES

Features	Possible values
NFORM	- Declined <span style="float:right">معرب</span> - Indeclinable <span style="float:right">منبني</span>
DEFINITE	- yes if it is defined <span style="float:right">معرف</span> - no otherwise
NAT	- demonstrative nouns <span style="float:right">اسم إشارة</span> - conjunctive nouns <span style="float:right">اسم موصول</span> - ...
ADJ	- Yes if it can be an adjective - no otherwise

In the same way, the features characterizing the Arabic nouns are mentioned on the HEAD feature level. It should be noted that on the level of the VAL feature, we can differentiate the various types of nouns. In fact, a declined noun can be preceded by another indeclinable. However, an indeclinable noun cannot have a precedent. It only categorizes some complements.

The Arabic particle, presented in “Fig. 4,” can be categorized in operative particles and inoperative ones. Thus, the features characterizing the particle type are presented in the table below:

TABLE III.  
ARABIC PARTICLE FEATURES

Features	Possible values
PFORM	- Non operative <span style="float:right">مهمل</span> - Operative <span style="float:right">عامل</span>
NATP	- elision particle <span style="float:right">حرف جر</span> - Subjunctive particle <span style="float:right">حرف نصب</span>

The modifications brought to this formalism cover not only the features but also the different schemata of the HPSG grammar. In the following paragraph, we are going to present the different modifications brought to the schemata.

B. Arabic schemata

As we already mentioned, HPSG is based on six different schemata. In this work, we adapted each schema to represent an Arabic syntactic phenomenon. We give, in this paper some examples of adapted schemata.

The first schema (rule of specification 1) was considered to represent any noun phrase preceded by an indeclinable noun (i.e., pronouns). The following figure illustrates a representation of a nominal phrase with this schema.

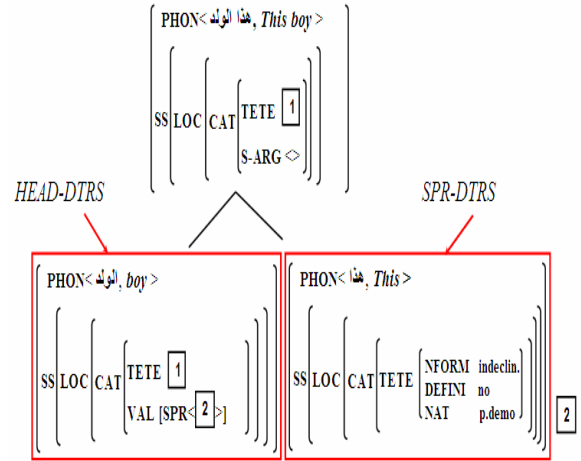


Figure 7. Illustrative example of the schema 1

As represented in the figure above, the name “الولد, boy” represents the head-daughter (HEAD-DTRS) of the noun phrase “هذا الولد, this boy”. It categorizes as specifier (Head-SPR) the demonstrative noun “هذا, this”, indexed 2.

The second schema (rule of specification 2) was adapted to represent the different forms of Arabic nominal sentences. In fact, it covers various forms of topics and attributes. The topic (NP, Nominal Phrase) represents the HEAD-DTRS. It categorizes the attribute (NP or VP, Verbal Phrase) as a COMPS-DTRS. Thus, this adapted schema represents sentences: (NP + NP) or (NP + VP) where:

$$NP \rightarrow N \mid NP$$

$$VP \rightarrow V \mid VP$$

According to [5], the third schema (rule of complementation) represents phrases having a not limited number of complements. Therefore, we have adapted this schema to represent different type of nominal phrases and verbal sentences. Among the nominal phrases treated with this schema, we can mention: the Annexed phrases (المركبات الإضافية) as “the neighbour’s son” (ولد الجار) and the substitution phrases (المركبات البديلية) as “Faatimatu the prophet’s girl” (فاطمة بنت الرسول).

These various phrases and sentences are treated with the same schema. The difference between there representations appears on the level of the constraints specification characterizing the HEAD-DTRS and the COMPS-DTRS.

The modification rule represented by the schema was adapted to represent adjectival phrases (المركبات النعتية). This schema is characterized by a modifier (MOD-DTRS) having the MOD feature. This feature selects the HEAD-DTRS.

To validate this grammar with the LKB, we have specified it in TDL. In the following paragraph, we give an idea about the grammar’s specification.

V. TDL SPECIFICATION

According to [11], the TDL syntax presents an important similitude with the HPSG representation. Therefore, we choose this language to specify our adapted grammar. In the following paragraph, we present an example of an AVM’s specification. Then we present an example of an adapted schema’s specification.



```

168 bakae hadha 'aalwaladu 'aalmiskynu 1 9
169 bakae 'aalmiskynu 1 3
170 rajlaaa 'aaldamu 'ilae 'aalwaladi 'aalsaghiri waladu 'aaljaarati faatimati 0 2
171 rajlaaa 'aaldamu 'ilae wajhi hadha 'aalwaladi 'aalsaghiri 1 24
172 hade-a tanaffusu hu 1 6
173 khalaasa 'ahmadu waladu 'aaljaari 'aalaaajuzi sutrata hu 1 19
174 jalasa 'ilae maktabi hi jalasa 'aalmufakkiri 0 9
175 jalasa 'aalwaladu jalasa 'aalrajuli 'aalmufakkiri 1 12
176 jalasa 'aalwaladu 'ilae maktabi hi 1 15
177 jaa-a 'aalaylu 'aaldaamu 1 6
178 ghattee 'aalZalaamu 'aalhaaliku 'aalsemaa-a 'aalseafiyata 1 11
179 tu-dhinu 'aalsemaa-u bi 'intihaa-i 'aalohitaa-i 1 15
180 wakafa 'ahmadu kurba 'aalnaafidhati 1 8
181 'ilae 'aalnujumi 'aaljamilati yanZuru 1 8
182 tata-allaku hadhihi 'aalnujumu 'aaljamilatu bi 'aalhaaati ha 1 27
183 'udhunu 'aalmiskynu murhafatuN 'ilae 'aalbaabi 0 6
184 'udhunu 'aalwaladi 'aalmiskynu murhafatu 'aalhissi 2 12
185 yafrahu 'aalwaladu 'aalbaaba 1 5
186 samiaaa hadha 'aalwaladu 'aalsaghiru tarkataN kabyratan 1 14
187 hadhihi 'aalarkatu tahuzzu 'aalkalba 0 7
188 ina 'aalwalada 'aalsaghira 'ahmada ka 'aalhaaridi 1 15
189 ina 'aalwalada fi taryki 'aalhayaati 1 12
190 ina 'ahmada yanchudu 'aalraahata 0 6
191 yanZuru 'aalwaladu 'aalsaghiru 'ilae 'aalduuniyaa 'aaljamilati 1 21
192 ina hadhihi 'aalnaZrata 'aaljamilata naZrata 'aalbaraa-ati 1 20
193 Dehika 'aalwaladu 'aalsaghiru 'ilae 'aalwujudi 1 15
194 ina 'aal-aalaati 'aalZiraasaaiyyata tekaddamat- 1 8
195 tekaddamat- hadhihi 'aalwasa-ilu tekaddumalN aaZymalN 1 10
196 fi 'aalaaahudi 'aalKadymati maHaarythuN hadythaatuN 1 11
197 ina zawjata 'aaljaari faatimata 'imra-atuN jamilatuN 1 14
198 talbaa hadhihi 'aalzawjatu 'aalfaatinatu 'aalchiyaaba 'aalidimechkiyyata 1 14
199 talbaa hadhihi 'aalzawjatu 'aalfaatinatu thiyaabaN mutarrazatan 1 14
200 aalae hadhihi 'aalchiyaabi rusumuN synnyatuN 1 10

;;; Total CPU time: 1688 msec

```

Figure 11. An extract of "results.txt"

As represented in this extract, the LKB system presents as results the number of each sentence, its text and two numbers. We take, as an example the framed sentence. In fact, the first number (1, for this sentence) represents the number of tree's derivation and the second one (3) represents the number of nodes in the creation graph of the derivation's tree. In the following figure, we give the derivation's tree and the creation graph of this sentence.

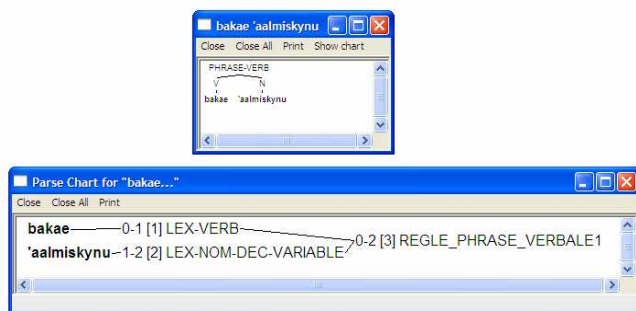


Figure 12. Result of "the poor cried"

Moreover, the LKB system presents in bottom of this file, the total time of analysis. According to this file we have evaluated our grammar.

## B. Evaluation

To test our HPSG grammar, we used a corpus of 200 transliterated sentences. This corpus was created from a lexicon of 781 words. It covers various structures of nominal and verbal sentences. According to the file "results.txt", we obtained the following results.

TABLE IV.  
OBTAINED RESULTS

Number of derivation's trees (n)	Number of sentences having n analysis
0	25
1	175
2	5
	200

In fact, 85% of sentences were analyzed correctly. The failure cases (0 analyzes) are due to the absence of rules treating some particular syntactic phenomena (i.e., relative phenomenon, coordination phenomenon). The ambiguous cases (2 analyze) are due to a no precise specification of the constraints specification of some syntactic rules.

## VII. CONCLUSION AND PERSPECTIVES

In this article, we proposed a type hierarchy categorizing the Arabic word. Based on this hierarchy, we adapted the HPSG grammar. To validate it with the LKB system, we have specified this grammar on TDL. The experimentation was effected on a corpus of 200 sentences. According to the obtained results, we evaluated our grammar.

As perspectives, we are going to reduce the number of failure cases. We will treat other particular phenomena and specify more constraints to eliminate the ambiguous cases. Moreover we consider developing a converter permitting to convert the lexical entries of XML in TDL in order to facilitate the development of the lexicon.

## REFERENCES

- [1] A. Abdelkader, K. Haddar, and A. Ben Hamadou, « Etude et analyse de la phrase nominale arabe en HPSG », *Traitement Automatique des Langue Naturelles, Louvain*, pp. 379-388, 2006.
- [2] A. Abdelwahed, « 'alkalima fy 'attourath 'allisaany 'alaraby, 'alkalima fi التراث اللساني العربي », *Librairie Aladin 1ère édition, Sfax - Tunisie* : pp. 1-100, 2004.
- [3] CH. Aloulou, « Analyse syntaxique de l'Arabe: Le système MASPARE », *RECITAL, Nantes - France*, 2003.
- [4] Y. Bahou, L. Hadrich Belguith, C. Aloulou and A. Ben Hamedou. *Adaptation and implementation of HPSG grammars to parse non-voiced Arabic texts*, Faculty of Economics and Management of Sfax, 2005.
- [5] P. Blache, « Les Grammaires de Propriétés: des contraintes pour le traitement automatique des langues naturelles ». *Hermès Sciences, Paris*, 2001
- [6] D. Bourigault and C. Fabre, « Approche linguistique pour l'analyse syntaxique de corpus », *Sémantisme et corpus*, pp. 131-151, 2000.

- [7] A. Copestake, *Implementing Typed Feature Structure Grammars*, CSLI Publications, Stanford University, 2002.
- [8] A. Dahdah. « معجم قواعد اللغة العربية في جداول و لوحات », Librairie de Nachirun ebanon, 5ème edition, 1992.
- [9] S. Elleuch, « Analyse syntaxique de la langue arabe basée sur le formalisme d'unification HPSG ». *Mémoire de DEA en Système d'information et Nouvelles Technologies, Sfax, Tunisie* : pp. 55-88, 2004.
- [10] O. Garcia, *Une introduction à l'implémentation des relatives de l'espagnol en HPSG-LKB*, Mémoire de recherche, 2005.
- [11] H. Krieger and U. Schäfer, « TDL: A Type Description Language for HPSG ». *Part 1 and Part 2, Research Report, RR*, pp. 94-37, 1994.
- [12] H. Maaloul, K. Haddar, and A. Ben Hamadou, «La coordination arabe : étude et analyse en HPSG », *MCSEAI 2004, 8ème conférence maghrébine sur le GL et l'IA, Sousse, Tunisie* : pp. 487- 498, 2004.
- [13] Meurers and Detmar W., «A Web-based Instructional Platform for Constraint-Based Grammar Formalisms and Parsing». *In Dragomir Radev and Chris Brew (eds.), Effective Tools and Methodologies for Teaching NLP and CL, New Brunswick, NJ: The Association for Computational Linguistics*: pp.18 – 25, 2002.
- [14] C. Pollard and I. Sag, «Head-drive phrase structure grammars», *CSLI series, Chicago University Press*, 1994.
- [15] J. Tseng, « Implémentation HPSG avec LKB: La Matrix et la Grenouille », *Séminaire HPSG-UFRL, Paris 7, 14(12)*, 2006