# HPSG Grammar for Arabic Coordination Experimented with LKB System

**Sirine Boukedi[1] and Kais Haddar[2]**

University of Sfax[1/2], Faculty of Science Economy and management[1] and Faculty of Sciences[2]
MIR@CL laboratory[1/2]
sirine.boukedi@gmail.com, kais.haddar@yahoo.fr

### Abstract

The existing applications in Natural Language Process (NLP) weren't reliable. Indeed, many complex phenomena are not treated completely such as relatives, juxtaposition, ellipsis and the coordination, essentially for Arabic grammar. In fact, the problems encountered relate to the choice of the formalism and the parser validating the constructed grammar. Therefore, our work aims to appreciate the different forms of Arabic coordination. Moreover, we plan to represent them with an adequate formalism, the Head-driven Phrase Structure (HPSG). The constructed grammar was validated with Linguistic Knowledge Builder (LKB), a parser generator system.

## Introduction

The coordination is an important linguistic phenomenon. It joins two or several compounds using conjunctions. However, there exist some cases where the elements composing a coordination structure are joined implicitly. This phenomenon interacts with many other syntactic phenomena, such as ellipsis and relatives. Therefore, there exist a large number of coordinated forms.

To treat the different cases of the coordination phenomenon, we should use a reliable formalism and choose a robust parser. The literature showed that there exist two different approaches: conceiving the parser or using a generator system. However, the second approach achieves satisfactory results since a generator system is based on algorithms approved by experts.

Therefore, in our work, we use HPSG formalism (Pollard and Sag 1994). The choice of this formalism is justified. It is a unification grammar characterized by a reliable modeling and a complete representation of linguistic knowledge. HPSG proposes a modularized organization of linguistic knowledge. It minimizes the syntactic rules and attributes importance to the lexicon. The elaborated grammar is validated with LKB system proposed by (Copestake 2002). It is ergonomic and used standard parser algorithm, "chart parsing". The originality of this work is to develop an HPSG grammar based on an adequate hierarchy classifying the different forms of Arabic coordination.

In the present paper, we present some related works treating coordination structure. Then, we give the proposed classification of Arabic coordination. After that, we introduce the HPSG representation of the different cases. Then, we present the TDL specification of the constructed grammar. This language is designed to support essentially the lexicalized grammatical theories and it is easy to extend. Finally, we give the experimentation with Linguistic Knowledge Building (LKB) system and we evaluate the obtained results. In our conclusion, we provide some perspectives and future works.

## Previous works

Researchers on coordination phenomenon started since 1970 for various languages. Our study showed that each work treated some particular forms of coordination using different grammars. Most of the related works considered that the coordination can be subdivided in two categories: constituent and non constituent coordination.

Biskri treated French coordination (Biskri and Desclés 2006), essentially constructions based on the conjunction "et, *and*". In their work, they used the ACCG. The obtained results showed that ACCG grammar is unable to take into account the coordination of elements having different function and nature.

Other researchers like (Djamé and Benoît 2006) used Lexicalized Tree Adjoining Grammar (LTAG). They presented a general approach for elliptical constructions of coordination. The used grammar has a delicate process of treatment. Moreover, it is expensive in terms of efforts and response time. Based on the obtained results, they concluded that the complexity of this process is exponential and depends of the number of derivations. Unlike all these grammars, HPSG is reliable for Natural Language Processing (NLP).

For (Abeillé 2006), she proposed two different solutions: Categorical Grammar (CG) and HPSG grammar. The first solution inserts some predicates using operators like the logical ones. This solution has several disadvantages: the appearance of many ambiguities, the difficulty of using the operators and can't represent elliptical constructions. By the way, HPSG has a clear description of linguistic objects using SAV, based on a detailed type hierarchy.

For Arabic language, some works treated Arabic coordination like (Haddar and Ben Hamadou 2009). The authors present a clause grammar to distinguish between well formed clauses and the uncompleted ones. To prove the feasibility of the proposed approaches, they developed a prototype called ERASE (Ellipsis Resolution of Arabic Sentences). The obtained results are satisfactory but the study on coordination phenomenon was done superficially. In conclusion, There study were incomplete and treats some forms of Arabic coordination. Therefore, we start our work by a large study and we propose a classification for Arabic coordination.

## Classification of Arabic coordination

According to (Hamad and Aidi 2012), Arabic coordination can be subdivided on two principal categories: Coordinating attraction (1) and explicative attraction (2).

(1) Taafa [‘alrijaalu fa ‘alnisaa’u] hawla ‘alk`abati
*[Men and women] turned around the Kaaba*

(2) Marartu bi [al faarisi `antara]
*I passed by [the escapee Antara]*

As highlighted in the examples, the first category, coordinating attraction, requires particles. Already the coordinated particles are called particles of attraction. In the next sections, we detail these two categories.

### Coordinating attraction

The coordinating attraction is an explicit relation. This kind of coordination is constructed with conjunctions.

For Arabic language, the elements composing a coordinated structure can be complete or incomplete. Therefore, there exist two different categories: constituent coordination and non constituent coordination. The study on Arabic grammar shows that the two categories require particles. Therefore, we considered them as subtypes of the coordinating attraction.

**Constituent coordination**

The constituent coordination represents the case when the compounds composing a coordination phrase are complete. In fact, there is no lack in the coordination clause. The joined elements can have similar or different categories, as represented respectively in examples (3) and (4).

(3) [‘akala thumma naama] fi ‘aalmanzili
*He [ate then slept] at home*

(4) [‘akala wa bi sor`atiN dhahaba] ‘ila ‘al madrasati
*He [ate and quickly went] to school*

In fact, as represented in sentence (3), the conjunction “thumma, *then*” joins two similar categories (two verbal phrases). However, in the second sentence, it joins a sentence “bi sor`atiN dhahaba, *quickly went*” and a verb “‘akala, *ate*”.

**Non constituent coordination**

The non constituent coordination describes the interaction with ellipsis phenomenon. It represents the case when the coordination clause lacks an element. According to (Haddar and Ben Hamadou 2009), there exist four forms of ellipse: Right Node Raising (RNR), Left Node Raising (LNR), Gapping and VP-ellipse.

RNR represents cases of right factoring (5) in a sentence. In fact, the component factor is at the right of the sentence. Contrariwise, LNR designed the case when the component factor is at the left of the sentence (6). For the third form: Gapping, it represented discontinuities in the second compound of the coordination phrase (7). Finally, for the VP-ellipse, it represented the case when the verbal phrase is missed and replaced by a proverb (8).

(5) [‘akala] Mohamed tufaahataN wa ∅ ‘akhouhu ijaaSataN,
*Mohamed ate an apple and his brother a pear*

(5’) Mohamed [‘akala] tufaahataN wa ‘akhouhu ∅ ijaaSataN
*Mohamed ate an apple and his brother a pear*

(6) ‘akalat- thumma naamat- [hadhihi ‘alkittatu]
*She ate the she slept, this cat*

(7)’istaykadha [‘aalwaladu] fa ghassala ∅ wajhahu
*The boy is waked up so hi washed his face*

(8) ‘akala ‘aalwaladu wa kadhalika [faàla] ‘akhouhu
*The boy ate and so his brother*

The study on Arabic grammar shows that sometimes when we transform a verbal sentence to a nominal one, we can switch from a form to another. (See example (5’)). In fact, after transformation, the example (5’) is no longer an RNR but a gapping form. In this context, we tried to cover the majority of coordination cases and we focused on this class of coordination.

The study on the Arabic grammar showed that there exist some cases when there is no particle in the coordination structure. It represents the explicative attraction. We give in the next section an overview.

### Explicative attraction

The explicative attraction is an implicit relation. It is characterized by an inert attracted. His role represents an adjective to explain the attracted. Referring to (Hamad and Aidi 2012), the possible cases that can take the compounds composing the explicative attraction are:

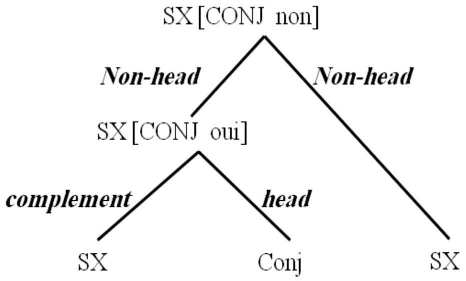
- The last name after the first name
- The last name after the nickname
- The described after an adjective
- The explication after the explicated compound

This type of coordination is very similar to the substitution phenomenon on the syntactic level that makes several cases of ambiguities. In the next section, we give the HPSG grammar for Arabic coordination

## HPSG for Arabic coordination

HPSG is a unification grammar (Pollard and Sag 1994). It is based on Attribute Value Matrix (AVM) for representation and a set of immediate domination schemata (DI schemata). The composition of the different structures is based on a set of principles.

The study showed that all the related researchers working on coordination have considered this phenomenon as a non headed structure. In fact many related works such as (Tseng 2007) argued that the conjunction is a weak head. It inherits an important number of properties from its complement, essentially its head features. For Arabic grammar, this criterion is also true. Indeed, a coordination schema has the following structure.



*General schema of the coordinated structure*

In fact, we have to conceive two different schema. The first one represents a headed structure. It joins the conjunction with the last compound. It represents a complement relation. In fact, the conjunction is the head daughter which chooses the complement compound. The second schema joins this structure with the other elements composing the coordinated structure.

## Experimentation with LKB system

LKB system is a parser generation tool, proposed by (Copestake 2002). This system is specialized for unification grammars such as HPSG grammar. The choice of this platform is justified. In fact, many researchers like (Garcia 2005) and (Laurens 2007) used LKB to validate their work and they obtained reliable results in a short time of response. Moreover, this system is ergonomic and very easy to use. Indeed, LKB used standard parser algorithm, the “Chart parsing”.

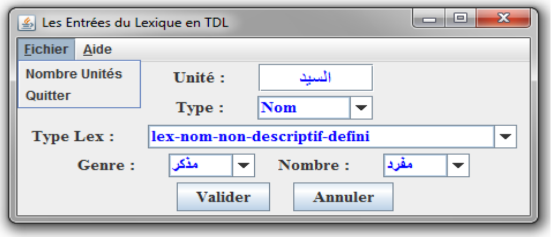
The HPSG modeling starts from a type hierarchy and a set of principles to represent the lexicon, the syntactic rules and the lexical ones. Then, the developed grammar is specified in Type Description Language (TDL). Each type of information is socked in a TDL file. The specified grammar is finally experimented with Linguistic Knowledge Builder (LKB) system, using a test corpus created from the standard corpus Arabic Tree Bank (ATB). In the following section, we present an overview on the TDL specification. Then, we give the experimentation of the constructed grammar.

### TDL specification

HPSG formalism is based on AVMs, to describe the different lexical entries and schemata representation. Each AVM is composed from a set of features. The values attributed to each feature have a type. Therefore, in the grammar development, we star by developing a type hierarchy classifying the lexical unities.

Thus, the TDL specification is based essentially on three TDL files. The first one “types.tdl” represents the specification of the type hierarchy. The second file “lexique.tdl” specifies the lexical entries. The third file “rsynt.tdl” represents the developed syntactic rules.

It should be noted that it is easy to add a lexical entry in the lexicon. However, this task requires many time. Therefore, we developed an application in JAVA “lex-editor” that adds automatically the unities in the lexicon. Moreover, it checks the presence of the unity in the lexicon and accounts the number of entries.



Unité : السيد

Type : Nom

Type Lex : lex-nom-non-descriptif-defini

Genre : مذكر

Nombre : مفرد

*Lex-edito Interface*

Besides, we developed some lexical rules to make the lexicon extensional. In fact, this type of rules generates automatically the derived forms of an entry.

## Evaluation and discussion

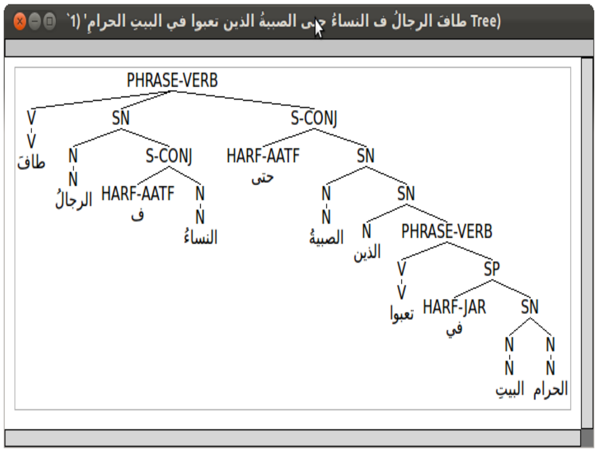
In the present work, we treated different cases of coordination. Indeed, the constructed grammar treats many forms of interaction with ellipsis, some relatives and many embedded forms. The constructed corpus from ATB includes 600 sentences containing 370 coordinations. The table below presents the different treated forms and gives the result of each form.

| Treated forms | Number of sentences | Fail | Success |
|---|---|---|---|
| simple | 40 | 4 | 36 |
| coord+rela | 100 | 30 | 70 |
| embedded forms | 50 | 5 | 45 |
| RNR | 40 | 15 | 25 |
| LNR | 45 | 20 | 25 |
| Gapping | 35 | 15 | 20 |
| VP-ellipse | 60 | 60 | 0 |

As we can see, in this table, the fail percent is too reduced comparing to the success one. In fact, the fail resumes two types of problems: ambiguity and no analyze since we have encountered some syntactic problems. First able, the study showed that there exist some cases very similar to the coordination at the syntactic level. The same constraints are described to many phenomena. This makes many ambiguity cases.

Moreover, many sentences can't be analyzed. In fact, the coordination interacts with the juxtaposition. This phenomenon is very frequent and represents another complex phenomenon that we didn't treat.

To give an overview on the obtained results, we give in the following figure the result of a sentence extracted from the constructed corpus, representing a success case.



*Result of a Successful Example*

This sentence contains coordination, relative and represents an embedded form. It is composed of two coordinated phrase. The principal phrase is joined with the conjunction «hattae, *even* », which contains another phrase based on the conjunction « fa, *and*».

## Conclusion and perspectives

In this paper, we proposed a typology for coordination structure in Arabic language. Based on this hierarchy, we adapted the HPSG grammar. In fact, we defined a particular structure for this phenomenon. Then we validated the constructed grammar with the LKB system. The experimentation was done on a corpus of 600 sentences. According to the obtained results, we evaluated the elaborated grammar.

As perspectives, we are going to treat other particular phenomena and specify more constraints to eliminate the ambiguous cases. Indeed, we are actually, working on juxtaposition. It is very frequent and interacts with coordination. Furthermore, we aim to construct a converter permitting to convert the lexical entries of XML in TDL in order to facilitate the development of the lexicon.

## References

Abeillé, A. 2006. Coordination: two challenges for syntactic theories, LLf, University, Paris 7.

Biskri, I. and Desclés J. 2006. Coordination of different categories in French, Quebec University, Canada.

Copestake, A. 2002. Implementing Typed FeatureStructure Grammars. *CSLI publications.*

Djamé, S. and Benoît. 2006. Modélisation et analyse des coordinations elliptiques par l'exploitation dynamique des forêts de dérivation. *TALN,* leuven.

Garcia, O. 2005. Une introduction à l'implémentation des relatives de l'espagnole en HPSG-LKB. Research memory.

Haddar K. and Ben Hamadou A. 2009. An Ellipsis Resolution System for the Arabic language. *The International Journal of Computer Processing of Languages* 22(4): 359–380

Hamad, Kh., and Aidi, H. 2012. أثر العطف في التماسك النصي في ديوان على صهوة الماء . *Journal de l'Université islamique de recherches en sciences humaines* : 327-356.

Krieger H. and Schäfer U. 1994. TDL: A Type Description Language for HPSG. Part1: Overview. Technical reports, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany.

Laurens, F. 2007. Implémentation des types de phrases et des types de constructions coordonnées du Français avec la plateforme LKB, Technical Report, laboratoire LLF.

Tseng J. 2007. La grenouille : Grammar report, Delph-In summit.