Analyses Tools for Non-head Structures

Sirine Boukédi
Faculty of Sciences Economy and Management of Sfax
Sirine.boukedi@gmail.com

Kais Haddar
Sciences Faculty of Sfax
Kais.haddar@fss.rnu.tn

Abstract

Syntactic analysis is a fundamental phase in NLP (Natural Language Processing) domain. This phase occurs in several applications and at different levels. Moreover, it wasn't spilled in domain research, especially for Arabic language. In fact, most of researchers working on Arabic language treated simple structures and neglected complicated ones such as relatives, coordination, ellipse and juxtaposition. In this context, the present work lies within the construction of a HPSG (Head-driven Phrase Structure Grammar) grammar treating Arabic coordination. The established grammar is specified on TDL (Type Description Language) and experimented with a parser generated by LKB (Linguistic Knowledge Building) system.

1 Introduction

The study on Arabic language showed that coordination is one of particular structures. It is frequent in different corpus and occurs with many other phenomena. The interaction with the other structures makes the study very delicate. For this reason, it wasn't spilled in research domain.

Based on a large literature, most of existing researchers treated coordination structure for Roman languages except some works such as (Haddar, 2000) and (Maaloul *et al.*, 2004). In fact, Arabic coordination is very complicate. It covers many forms and different structures. Therefore, there is a big problem in the categorization of Arabic coordination.

Moreover the last researchers found a problem in the choice of the adequate formalism representing the different forms of coordination structures. But most of related works used HPSG. The choice of this formalism is justified. In fact, HPSG offers a complete representation for linguistic entries.

Therefore, our work aims to find an adequate typology classifying Arabic coordination structures and to construct a HPSG grammar representing the different forms of coordination. This grammar will be validated on LKB system.

In the present paper, we start by describing some related works treating coordination structure. Then, we adapted HPSG grammar to represent the different forms of our phenomenon, based on a proposed typology. It should be noted that the established grammar treated simple sentences and complex ones representing the different forms of coordination except cases of interaction with ellipse phenomenon. Finally, we validated our grammar on LKB system after specification in TDL. According to the obtained results, we evaluate our grammar and we enclose our work by a conclusion and some perspectives.

2 Previous works

The study on previous works showed that researchers on coordination structure started since 1970, such as (Hudsan, 1976), (Postal, 1974) and (Rau, 1985), for many languages. The different researches used various grammars. Some works used the GCCA (Applicative Categorical Combinatory Grammar), other works used GI (Interactive Grammar) and other ones were based on HPSG Grammar. But most of them, used this last one (HPSG formalism).

For French language, we can mention (Biskri and Desclés, 2006) who studied coordination structure of similar constituents, based on GCCA grammar. Moreover, (Le Roux and Perrier, 2006) studied constituent and non constituent structures based on XMG tools, a compilation tool, and used the GI formalism.

For Portuguese language, (Villavicencio *et al.*, 2005) studied the coordination of nominal phras

Proceedings of Recent Advances in Natural Language Processing, pages 686–691,
Hissar, Bulgaria, 12-14 September 2011.

es. They identified different strategies of analyses based on the HPSG formalism.

For Bulgarian language, we can mention the work of (Osenova and Simov, 2005) who studied the coordination phenomenon and its interaction with ellipse forms based on HPSG grammar. It should be noted that the formalization was encoded in XML.

According to our research, the study showed that most of the related works treated the coordination of Roman language. But, there is some works treating Arabic coordination such as (Haddar, 2000) and (Maaloul *et al.*, 2004). The proposed typology is similar in most of the related works. The difference between them appears in the choice of the grammar and the analysis tools.

For Arabic works, for example, Haddar (2000) studied syntactic analyses of Arabic coordination based on ATN (Augmented Transition Network) and (Maaloul *et al.*, 2004) studied the coordination of Arabic constituent based on HPSG grammar. This grammar was tested and validated on a constructed system, AICOO.

Based on the proposed typology, these related researches working on Arabic coordination didn't treat all forms of this structure. Therefore, the originality of our work is to construct a HPSG grammar covering all the possible forms of coordination and its interaction with the other phenomena such as the ellipse one. In the following paragraph, we present the proposed typology of Arabic coordination that we adopted from the related works.

3 Proposed typology for Arabic coordination

According to some linguists such as (Abdelwahed, 2004) and (Dahdeh, 1992), the coordination phenomenon joins two or several elements with a particle of coordination (conjunction). In Arabic, there exist nine conjunctions (و، ف، ثمّ، حتّى، لكن، أم، أو، لا، بل).

Based on some related works (Haddar, 2000) and (Maaloul *et al.*, 2004), coordination structure in Arabic can be subdivided, like Roman languages, in two categories: coordination of constituent and coordination of non constituent. The first category covers cases when the conjunction joins two or several well formed constituents. These constituents can have similar or different categories. The figure 1 represents the coordination of similar constituents. The figure 2 represents the coordination of different constituents.

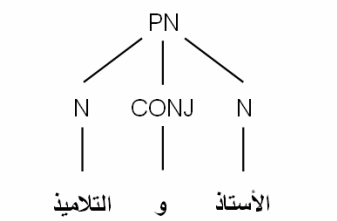

Figure 1. Coordination of constituents having similar categories.

As shown in Figure 1, the conjunction" و, and" joins two compounds having the same category, two defined nouns "الأستاذ, the professor" and "التلاميذ, the pupils".

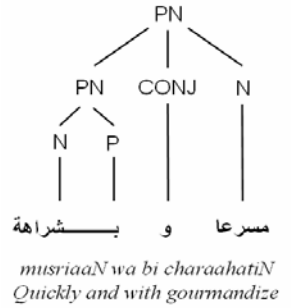

Figure 2. Coordination of constituents having different categories.

However in Figure 2, the same conjunction joins two different compounds. The first one "مسرعا, quickly" is an adverb, the second one ", بشراهة, with gourmandize" is a reduction phrase " مركب بالجر, reduction phrase".

For the second category, coordination of non constituent, the conjunction joins two or several constituents where one of them is incomplete. In fact, it represents the case where there is interaction with the ellipse phenomenon.

According to some references, there exist four forms of ellipse: Right Node Raising, Left Node Raising, Gapping and VP-ellipse. The first form: Right Node Raising, designed the case when the first element that should be at the right of the second compound, is missed.

The second form: the Left Node Raising designed the case when the second element is missed in the left of the second compound of coordination phrase.

For the third form: Gapping, it represented when there exist discontinuities in the second compound of the coordination phrase.

Finally, for the last form, VP-ellipse, it represented the case when the verbal phrase is missed and replaced by a proverb like "كذلك, also".

Based on the proposed typology for the Arabic coordination, we adapted the HPSG grammar. In fact, based on some references such as (Godard, 2006), the coordination phenomenon is a non head structure. Its representation differs from other phenomena. So it necessitates a particular structure. In the following paragraph, we present the HPSG grammar and the different modifications brought to this formalism to represent Arabic language. Then, we present the HPSG structure of Arabic coordination.

4 HPSG for the Arabic language

HPSG is a unification grammar (Pollard and Sag, 1994). It is characterized by a reliable modeling of the universal grammatical principles and a complete representation of linguistic knowledge.

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. The DI schemata, describe a syntactic phenomenon. It should be noted that to compose the various phrases, a set of principles should be verified (i.e., HFP Head Feature Principle).

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

4.1 Arabic features

Referring to previous projects (Elleuch, 2004), (Bahou *et al*, 2005) and (Abdelkader *et al.*, 2006), we have kept some features and have added some others according to the proposed type's hierarchy. As example, we present, in table 1 below, the features characterizing the Arabic particle.

Features	Possible values
PFORM	- Non operative مهمل - Operative عامل
NATP	- elision particle حرف جر - Subjunctive حرف نصب

Table 1. Arabic particle features

Indeed, an Arabic particle can be operative particles or non operative. The coordination particles are classified as non operative particles. In fact, it didn't specify any constraint to the conjoint compounds.

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 present as example the conceived schema for Arabic coordination.

4.2 Arabic schemata

HPSG grammar is based on six schemata. In this work, we adapted each schema to represent an Arabic syntactic phenomenon (the simple one). In the context of our work; we present the conceived schema for Arabic coordination.

To represent coordination structure, a complicate phenomenon, we have represented, at first, the simple one. In fact, coordination interacts with different others phenomenon. All other representations were headed structure. However, the coordination has a particular structure. In fact, according to some references, the coordination is a non headed structure. Godard (2003) showed that the conjunction can't be the head of the phrase. In fact, a coordination particle is non operative. Thus, it can't specify conjunct elements.

Therefore, we developed a ternary non headed rule for the coordination phenomenon to obtain the representation below:

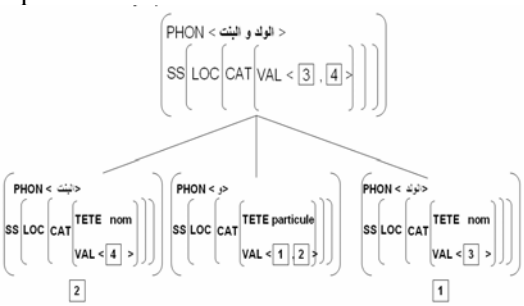

Figure 3. Coordination schema

As represented in figure 3, this structure doesn't contains three non headed daughters: two Fils-conj representing the two compounds of the coordination phrase الولد, (*'aalwaladu, the boy*) and البنت, (*'aalbintu, the girl*) and the Fils-conjonction representing the coordination particle و, (*wa, and*).

To validate our constructed grammar with LKB system, we specified it in TDL (Type Description Language). The choice of LKB platform is justified. In fact, it generates automatically a reliable parser. Some related works such as (Garcia, 2005) used this system and they obtained good results.

In the following paragraph, we give an idea about the specification of the constructed HPSG.

5 TDL specification

According to (Krieger and Schäfer, 1994), the TDL syntax presents an important similitude with the HPSG representation. Therefore, the TDL specification was simple. At the present time, our grammar covers the first category of coordination mentioned in section 3: coordination of constituents.

To specify this grammar, we specified the lexical entries AVMs, the type hierarchy and the syntactic rules representing the different forms of coordination and all possible simple sentences (verbal and nominal).

In the following paragraph, we present an example of TDL specification of an AVM and some schemata.

5.1 TDL specification of an AVM

From a HPSG representation, the TDL specification of an AVM is very simple. We present in the following figure the specification TDL of "هذا, that" (hadha).

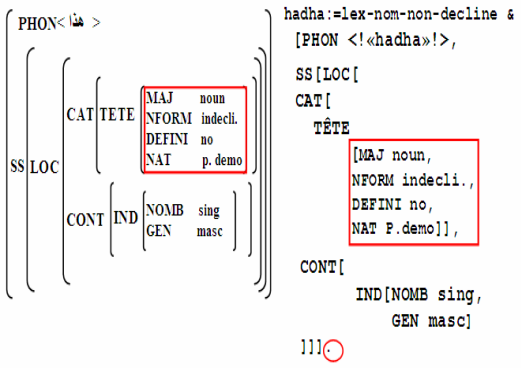

Figure 4. TDL Specification of "that" (hadha, هذا)

In fact, the symbol ":=" designates that "هذا", (*hadha, this*) is an instance of indeclinable nouns. The different constraints are added by the symbol &. The feature structures are delimited by brackets [ ]. Besides, the various attributes values are separated by commas "," and the full stop ". " designates the end of the AVM.

5.2 TDL specification of a schema

To specify the syntactic rule of Arabic coordination, we started by rules representing simple phenomenon. For the coordination structure, we specified two different rules. The first one represents verbal phrases and sentences. The second one represents nominal phrases and sentences. In the following figure we present the TDL specification of the coordination of nominal phrases.

```
regle_coordination_nom := regle-tern-sans-t &
     [SS.LOC[CAT[TETE[DEFINI oui, DEC #dec],
                 VAL [TOPIC < >, SPR <#nontete1>,
                      COMPS <#comps1, #comps2>]],
             CONT.IND [NOMB duel]],
      BRS.BRS-NTETE <[SS #nontete1 &
                      [LOC[CAT[TETE nom &[DEC #dec],
                               VAL [COMPS <#comps1>]]]]],
                     [SS #nontete2 &
                     [LOC[CAT[TETE particule_non_operative,
                              VAL [SPR <#nontete1>]]]]],
                     [SS[LOC[CAT[TETE nom & [DEC #dec],
                                 VAL [SPR <#nontete2>,
                                      COMPS <#comps2>]]]]]>].
```

Figure 5. TDL Specification of the coordination rule

As represented in this figure, this rule joins nominal phrases. It extends from the type regle-tern-sans-t. This type of rules represents non headed structures. In fact, before implementing this syntactic rule, we specified this type of rules. (Figure 6):

```
regle-tern-sans-t :=
  regle-ternaire &
  [BRS struc-sans-tete &
       [BRS-NTETE < #1, #2, #3 >],
   ARGS < #1, #2, #3 >].
```

Figure 6. TDL Specification of the type rule regle-tern-sans-t

In fact, regle-tern-sans-t is a ternary rule having two non-headed daughters joined with a particle of coordination.

Following the phase of specification TDL, we tested the adapted HPSG grammar with the LKB system. In the next paragraph, we give an idea about this system. Then we describe the experimentation and the evaluation of this grammar.

6 Experimentation and evaluation

LKB system is a parser generation tool, proposed by (Copestake, 2002). This system can run on Windows or on UNIX. But the version on Windows can't support Unicode. In fact, LKB is written in LISP using Motif. Therefore, we have added Trollet (TROndheim LingLab Engineering Tool), a tool for multilingual grammar development. It is easy to extend and can be used only on UNIX system. Therefore we installed Ubunto system. Then we install LKB and Trollet. Thus LKB is embedded in Trollet and invisible for the user. This tool replaced the LKB window.

It should be noted that the LKB is based on two types of files: TDL files and LISP files. The first type represents the grammar's files (i.e., types.tdl, rsynt.tdl, lexique.tdl). The second type represents files to parameterize the LKB system. Among these files, we can especially mention the file: "script.lsp'. It is a very important file. It

charges the grammar on LKB. In fact, it indicates the name and the root of each grammar file. In the following paragraph, we describe the stages of syntactic analysis. Then, we present the experimentation of the constructed grammar.

6.1 Stages of syntactic analyses

After charging the constructed grammar on LKB, this system offers a generated parser to analyze a simple sentence or a corpus of sentences. But, it should be noted that this corpus must be segmented in sentences. In fact, LKB system didn't have a module segmenting the tested corpuses.

To analyze a simple phrase, we have to unroll the "parse" menu and choose the "parse input" order. Thus, the LKB system generates a zone of text to type the sentence and as result, a derivation tree appears. The following figure represents the result of a sentence's analysis.

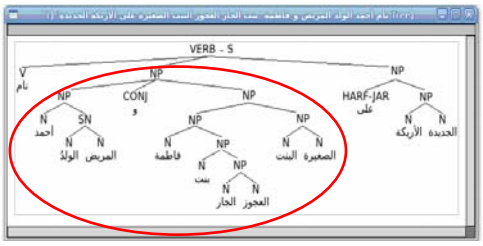

Figure 7. Derivation's tree of a verbal sentence

The derivation's tree in figure 7 is a representation of the following verbal sentence:

نام أحمد الولد المريض و فاطمة بنت الجار العجوز البنت الصغيرة على الأريكة الجديدة

ahmadu the sick boy and fatimatu the girl of the old neighbour, the little girl, are slept on the new couch

As shown in Figure 7, with TROLLET, we resolved the problem of transliteration. Contrary to the LKB on Windows, this system on UNIX can now support Arabic language except that the reading direction of the tree is false. Indeed an Arabic sentence must be read from the right to the left. At the present time, we are looking for a solution for this problem.

The choice of this sentence as example is justified. It represents the interaction of the coordination structure with various syntactic phenomena (i.e. annexing, description). In fact, the subject (circled in red in the figure) is a coordination of constituents joining two nominal phrases.

To analyse a corpus of sentences with LKB, we should specify to it two paths: the path of the file containing the sentences (corpus.txt) composing the corpus, and the path of the file that will contain the results (results.txt). In

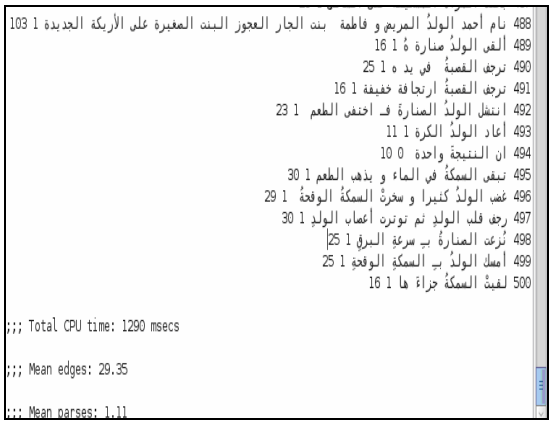

Figure 8. Extract of results.txt

Figure 8 represents our test corpus containing 500 sentences which are extracted from different linguistic resources and containing coordination structures using different particles (و, ثم, ف). It should be noted that in "results.txt", LKB represents the total time of analyses and for each example; it represents the number of tree's derivation and the number of nodes in the creation of the derivation's tree.

According to the obtained results, we evaluated our work. This evaluation is described in the following paragraph.

6.2 Evaluation

To test our HPSG grammar, we used a corpus of 500 sentences. This corpus was created from a lexicon of 2000 words. It covers different sentences containing coordination structures using different particles. As we have mentioned before, LKB system or TROLLET didn't have a segmentation module to cut sentences composing the corpus. At the present time, it is done manually. Therefore, we missed much time in this phase. But, we are looking for a segmentation tool toad it to our system.

At the present time, we treated only constituent's cases and we are working on ellipse phenomenon and its interaction with the coordination one. According to the file "results.txt", we obtained the following results.

Number of derivation's trees (n)	Number of sentences having n analysis
0	50
1	420
2	30
	500

Table 2. Obtained results

In fact, 84% 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). In fact, at the present time we treated constituent structures. We have not yet treating ellipse phenomenon and its interaction of the coordination phenomenon. The ambiguous cases (2 analyzes) are due to a no precise specification of the constraints specification of some syntactic rules.

7 Conclusion and perspectives

In this article, 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 specified syntactic rules treating simple sentences and coordination structures. The specified grammar was validated with the LKB system. The experimentation was done on a corpus of 500 sentences. According to the obtained results, we evaluated our grammar.

As perspectives, we are going to treat ellipse phenomenon and its interaction with coordination structures. Then, we will treat other particular phenomena and specify more constraints to eliminate the ambiguous cases. Moreover we consider developing lexical rules to make our lexicon extensional. 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

Abdelkader A., Haddar K., and Ben Hamadou A. 2006, *study and analyses of nominal sentences in HPSG*, TALN, Louvain, 379-388.

Abdelwahed A. 2004, *'alkalima fy 'attourath 'allisaany 'alaraby , الكلمة في التراث اللساني العربي*, Aladin library, Sfax – Tunisie, 1-100.

Bahou Y., Hadrich Belguith L., Aloulou C. and Ben Hamedou A. 2005, *Adaptation and implementation of HPSG grammars to parse non-voweled Arabic texts*, Faculty of Economics and Management of Sfax.

Blache P. 2001, *Propriety grammar : constraints for NL*, Hermès Sciences, Paris.

Biskri I., Desclés J., 2006, *coordination of different categories in French*, Quebec university, Canada.

Bourigault D.and Fabre C. 2000, *Linguistic approach for syntactic analyse of corpuses*, Sémantisue et corpus, 131-151.

Copestake A. 2002, *Implementing Typed Feature Structure Grammars*, CSLI Publications, Stanford University.

Dahdah A. 1992, *معجم قواعد اللغة العربية في جداول و لوحات*, Librairie de Nachirun ebanon, 5ème edition.

Elleuch S. 2004, *syntactic analyse of arabic language based on HPSG formalism*, DEA memory on Information system and new technology, 55-88.

Garcia O. 2005, *Une introduction à l'implémentation des relatives de l'espagnol en HPSG–LKB*, Research memory.

Godard D. 2003, *Syntacyic problems of coordination and recent propositions in syntagmatic grammars*, study journey in the university of Paris 7.

Haddar K. 2000, *Formel caracterization of Arabic ellipse and recouvrement processus in Arabic language*, University of Tunis II – Sciences Faculty of Tunis.

Hudson R. 1976, *Conjunction Reduction, gapping and right-node raising*, 535-562.

Krieger H. and Schäfer U. 1994, *TDL: A Type Description Language for HPSG,* Part 1 and Part 2, Research Report, RR, 94-37.

Le Roux J., Perrier G. 2006, *coordination modelisation in Interactive grammars*, TAL, Volume 47, n° 3, 89-113.

Maaloul H., Haddar K., and Ben Hamadou A. 2004, *Arabic coordination: HPS analyse*, MCSEAI 2004, 8th maghrébine conference on GL and IA, Sousse, Tunisie : 487- 498.

Osenova P. and Simov K., *Special Linguistic phenomena in the Bulgarian HPSG_based Treebank*, BulTreeBank Project, Linguistic modelling laboratory, IPP, Bulgarian Academy of Sciences.

Pollard C. and Sag I. 1994, *Head-drive phrase structure grammars*, CSLI series, Chicago University Press.

Postal P. 1974, *On raising*, MIT Press, Cambridge/MA.

Rau F. L. 1985, *the understanding and generation of ellipses in a natural language system*, Berkeley Artificial Intelligence Research project.

Villavicencio A., Sadler L. and Arnold D. 2005, *An HPSG account of closest conjunct Agreement in NP Coordination in Portuguese*, Proceedings of the HPSG05 conference, Department of Informations, University of Lisbon.