Construction of an HPSG Grammar for the Arabic Relative Sentences

Ines Zalila
Faculty of Economics and Management of Sfax
ines.zalila@yahoo.fr

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

Abstract

The paper proposes a treatment of relative sentences within the framework of Head-driven Phrase Structure Grammar (HPSG). Relative sentences are considered as a rather delicate linguistic phenomenon and not explored enough by Arabic researchers. In an attempt to deal with this phenomenon, we propose in this paper a study about different forms of relative clauses and the interaction of relatives with other linguistic phenomena such as ellipsis and coordination. In addition, in this paper we shed light on the recursion in Arabic relative sentences which makes this phenomenon more delicate in its treatment. This study will be used for the construction of an HPSG grammar that can process relative sentences. The HPSG formalism is based on two fundamental components: features and AVM (Attribute-Value-Matrix). In fact, an adaptation of HPSG for the Arabic language is made here in order to integrate features and rules of the Arabic language. The established HPSG grammar is specified in TDL (Type Description Language). This specification is used by the LKB platform (Linguistic Knowledge Building) in order to generate the parser.

1 Introduction

Relative phenomenon has a great importance in all natural languages and in all corpus kinds. That's way researchers in linguistics or in computer sciences pay great attention to this phenomenon (i.e., (Belkacemi, 1998), (Elleuch., 2004) and (Garcia,2006)). Indeed, a phase of parsing of this phenomenon is fundamental for several types of Natural Language Processing (NLP) applications such as grammatical correction and machine translation. Nevertheless, the researches concerning the parsing of relatives, object of this work, have not reached an advanced stage yet. This is due, on the one hand, to the complexity of this phenomenon and, on the other hand, to the interaction with simple and complex linguistics phenomena.

Thus, one of the objectives of this work is to study the various forms of the Arabic relative sentences. This study is based on old grammatical theories (Abdelwahed, 2004), (Belkacemi, 1998) and (Dahdah, 1992), and on discussions with linguists. From the study carried out, we also want to identify all possible syntactic representations of the Arabic relative sentences. The choice of the HPSG is justified by the fact that this formalism has shown great efficiency in several languages such as German.

The elaborated HPSG grammar is specified in TDL (Type Description Language). Based on the elaborated TDL specification, an Arabic parser is generated using the LKB linguistic platform. The generated parser can process complex sentences containing relatives. The originality of this work consists, on the one hand, in the identification of a relative sentences typology, and on the other hand, in the proposition of a HPSG extension detailing under-categorization. This extension is specified in TDL (Type Description Language) (Krieger and Schäfer, 1994), the language supported by LKB platform.

In this paper, we begin with presenting some projects dealing with the relative phenomenon. Then, we give a typology for Arabic relative sentences. After that, we introduce the HPSG formalism and we present modifications made on this formalism to adapt it to the Arabic language. Using this formalism, we elaborate a grammar for the Arabic language which can process relatives and we specify this grammar in TDL. We test this specification by generating a parser in LKB and applying it to a corpus of complex sentences. Finally, we conclude the paper by giving some perspectives of our work.

2 Related works

Researchers on the Arabic Language Processing began in the 1970's. The projects carried out

Proceedings of Recent Advances in Natural Language Processing, pages 574–579,
Hissar, Bulgaria, 12-14 September 2011.

since then and which have proposed parsers based on HPSG are limited.

Most of projects have proposed prototypes of parsers covering some phenomena (i.e., simple sentence, ellipsis). For example, in (Aloulou, 2003) and (Bahou *et al.*, 2003) authors studied the simple Arabic sentences and their representation with HPSG. They proposed some modifications on HPSG to adapt it to the Arabic language. These works are integrated in a multi-agent platform. In (Abdelkader *et al.*, 2006), the elaborated grammar makes it possible to analyze the Arabic nominal sentences. Also, priorities were introduced while applying HPSG schemata.

For the complex Arabic sentences, we take as an example the work presented in (Elleuch, 2004). It allows processing of simple sentences as well as complex ones. This work is based on the use of a large number of production and dynamic rules because the HPSG used version is old. Also, we take the research project presented in (Maaloul *et al*, 2004) which deals with Arabic sentences containing joint components and makes modifications on HPSG to adapt it to coordination. Note that all these works are based on their own parser. The relative phenomenon is also studied in (Belkacemi, 1998). This work shows that conjunctive nouns are not considered as determinants but as modifiers.

Concerning, the projects using the second approach which consists in the use of a tool for generation, we find essentially researchers studying Latin languages. For example, the project proposed in (Garcia, 2005) aims to analyze the Spanish relative subordinate clauses. This analysis is made on the LKB platform and is specified in TDL. In the same way, the project presented in (Tseng, 2006) deals with the French phrase affixes.

3 Arabic Relatives

The relative linguistic phenomenon is frequent in sentences and exists in all languages. In this section, we give an overview on the Arabic relative sentence, and then we explain the various forms that can take and we give some ambiguities in the treatment of Arabic relative sentences.

3.1 Definition

An Arabic relative sentence is a subordinate clause that carries out the various grammatical functions of a noun. It can play the role of a topic (مبتدأ), a predicate (خبر), a subject (فاعل) or object (مفعول به). Relative sentences are introduced by a special class of nouns called conjunctive nouns like 'الذي, who' and are followed by a special clause called relative clause.

Relative sentence (Srel) =
Conjonctive Noun (CN) + relative Clause (Crel)
مركب موصولي = اسم موصول + صلة الموصول

The following example illustrates previous rule that describe relative sentence structure:

[التلميذ [الذي نجح في الامتحان]] سافر إلى فرنسا.

[The boy [who succeeded in the examination]] has travelled to France.

In this example, the noun "التلميذ" is modified by the relative sentence Srel "الذي نجح في الامتحان". This relative sentence is composed by the conjunctive noun CN "الذي" followed by verbal clause Crel "نجح في الامتحان".

For Arabic relative sentence, we distinguish those which have an antecedent and others not. The relative ones with antecedent generally make it possible to give information on the antecedent (explanatory relative). In contrast, the relative ones without antecedent are themselves which supplement the means of the sentence (completive relative).

3.2 Relatives Typology

The proposed Arabic relative typology is inspired from the old grammatical theory and the former research tasks. Indeed, it is based on nature of clause which follows the conjunctive noun (Crel). The clause (Crel) can be a verbal phrase (VP), a prepositional phrase (PP) or a nominal phrase (NP). The (Crel) clause nature depends also of conjunctive noun's nature. The categorization of conjunctive nouns (NC) as well as Arabic relative forms is defined in the following sections.

- **Conjunctive noun's nature**

Conjunctive noun's nature plays a role in the categorization of Arabic relative sentences. Indeed, a conjunctive noun (الاسم الموصول) is considered as an indeclined insignificant noun. It occupies the functional head of the sentence and it is semantically co-referent with the antecedent. The conjunctive nouns are categorized as two kinds: Nominal conjunctive (الموصولات الاسمية) and prepositional conjunctive (الموصولات الحرفية). Nominal conjunctive are categorized in two subforms: common conjunctive and special conjunctive. As for the prepositional conjunctive, they are subdivided in two sub-forms: conjunctive ones influencing the verbs and others influencing the nouns.

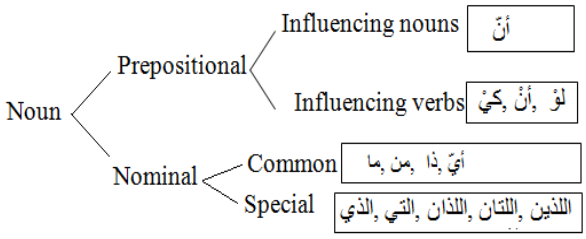

Figure 1: Categorization of conjunctive nouns

Based on elaborated study, explanatory relative is introduced by special conjunctive. All other nature of conjunctive nouns can introduce completive relative. In addition, according to the nature of the relative clause which follows the conjunctive noun, we distinguish two forms.

- **Relative clauses Typology**

The clause (Prel) which follows the conjunctive noun can be a verbal phrase (VP), prepositional phrase (PP) or nominal phrase (NP). According to these criteria, we identify two forms:

First form: Conjunctive noun followed by a verbal phrase (VP) or prepositional phrase (PP)

This form regroups conjunctive nouns which require the existence of a verbal phrase or prepositional one. For this form, we identify three types of relative's nouns: special nominal conjunctives, common nominal conjunctives, except for the conjunctive "أيّ", and prepositional conjunctives influencing the verbs. We define these various natures of conjunctive nouns as follows.

صافح المدير الذي تكلم كثيرا , البنت التي تحصلت على الجائزة

The director, who spoke a lot, greeted the girl who took the award

In the previous example, the conjunctive noun 'الذي' agrees with its antecedent 'المدير' in gender and number. If the number is conserved and the antecedent's gender was modified the conjunctive noun will be replaced by their correspondent one.

For neutral common conjunctives, they are independent from gender or number 'من، ما،أيٌّ،ذا'. Except for 'أيّ', all neutral common conjunctives require a VP or a PP.

قرأ الولد (البنت) [ما كتب الأب في الرسالة]

The boy (the girl) has read what the father wrote in the letter

The example above illustrates the independence of the common conjunctive 'ما' in gender and number. Conjunctive nouns 'ما' and 'من' do not require an agreement with the verb of VP.

Second form: Conjunctive noun followed by a nominal phrase (NP)

The second form covers conjunctive nouns which require the existence of a nominal clause. These conjunctives are represented by the common nominal pronoun 'أيّ' and the prepositional conjunctives influencing nouns. These natures of conjunctive nouns are detailed as follows.

The conjunctive noun 'أيّ' is a declined common conjunctive noun which refers to all what is human. The conjunctive noun 'أيّ' have a various forms according to the function which plays. Following example illustrates this correspondence:

سيكافئ الأستاذ[أيَّ مجتهد]

The teacher will reward any diligent

سيفوز[أيٌّ مجتهد] بالجائزة

Any diligent will win a prize

Examples above show that the connective noun "أي" can have in a sentence different grammatical functions. In the first example, the connective noun "أي" is a part of the object. So, it is open ending. For the second example, connective noun "أي" is a subject. Then, it is regular.

The prepositional conjunctive noun "أنَّ" requires the existence of nominal phrases after this type of conjunctive. The NP must be open ending.

قال الأب [أنّ الولدَ مريض]

The father says [that the child is sick]

As we already mentioned, prepositional conjunctive noun "أنَّ" is followed by a nominal phrase "الولدَ مريض". This conjunctive does not require an agreement.

As we already mentioned, the relative phenomenon is complex. This complexity is due to the diversity of possible forms and to the ambiguities founded during the analysis like interaction with other linguistic phenomena as ellipsis and coordination. This interaction increases the complexity degree of this phenomenon. The following example illustrates this interaction.

وجد الولد الكتاب [الذي يريد و يرغب]

The child who took the book [which he wants and desires]

In sentence above, we can note that the phenomenon of ellipsis intervenes on the level of the verbs. Indeed, the objects of these two verbs were elided.

In addition, in Arabic language, relatives can be recursive. Indeed, relative sentence can contain another relative sentence. Recursion can contain different types of relative (completive or explanatory). The example above show that the explanatory relative, whose antecedent is "البنت", containing another " البنت التي تحصلت على الجائزة التي تتكون من عدة كتب".

In order to analyze suitably the relatives, some modifications were made to the HPSG formalism.

In the following paragraph, we develop an adequate HPSG grammar.

4 HPSG for Arabic relatives

HPSG (Head-driven Phrases Grammar Structure) is a grammar of unification which was proposed by (Pollard & Sag, 1994). It is considered among best grammars for the modeling of the universal grammatical principles and a complete representation of the linguistic knowledge. Indeed, it make possible to represent in the lexical entries phonological, morphological, syntactic and semantic information.

In order to implement HPSG for the Arabic language, we adopt the already made modifications in order to integrate the particularities of this language (Boukedi et al., 2007). Figure 2 presents the SAV of a conjunctive noun using the majority of features added to the noun's type.

The example in figure 2 shows that "الذي" is not a significantly declined noun. This information is expressed by the features MAJ and NFORM. As for the feature INDEX, it shows that "الذي" is a singular masculine noun.

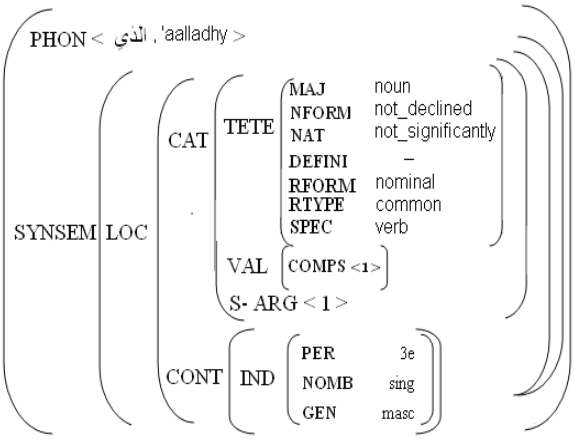

Figure 2: The SAV of the noun "'alladhy **الذِي**"

As it's indicated in previous parts, the immediate dominance (ID) schemata allow the generation of the derivation trees (Pollard and Sag, 1994) and (Blache, 1995). The arabized HPSG formalism must necessarily adapt these schemata in the sense of reading since the Arabic language is written from right to left (Boukedi et al., 2007). As follows, we present the modification of the mark's schemas taking into account the phenomenon of relatives.

Marking schema represents, on the one hand, a son head not having a descent not limited to enclose and on the other hand, a son marker referring HEAD of the marker type. The markers are associated with the feature SYN-SEM | LOC | CAT| MARK. This schema allows generally representing the relative sentences of the Arabic language. Thus, any sentence containing a conjunctive will be represented on this schema.

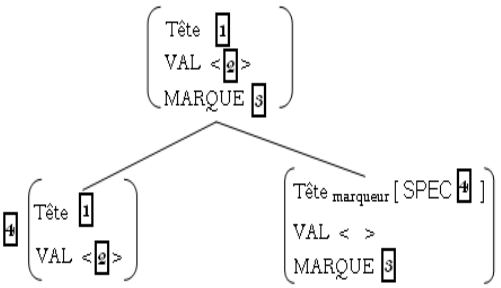

Figure 3: The rule of mark: Modified schema.

For example, the phrase " الذي أكل التفاحة" represents a relative clause whose marker is the conjunctive noun «الذي» followed by a verbal phrase «أكل تفاحة» indexed 4.

Besides the marking rule, we use the modification schema to control the selection of the antecedent by the conjunctive noun. Indeed, the majority of conjunctive noun have an antecedent presented in the form of a noun.

The following phrase represents a relative clause whose modifier is the conjunctive noun « الذي » which modifies its antecedent: the noun « الولد ».

[الولد الذي ...]

[the child who ...]

The figure 4 illustrates the generation of syntax tree for « .... الولد الذي» using the already mentioned rule of modification.

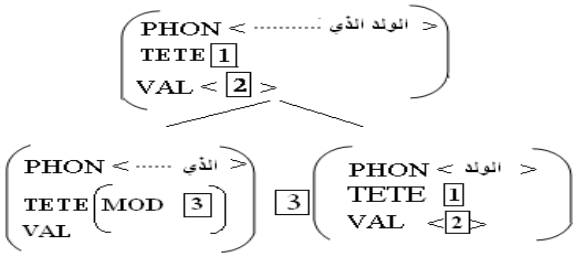

Figure 4: HPSG representation of the phrase **" الولد الذي..."**

In conclusion, the HPSG grammar designed and adapted to the Arabic language makes possible to analyze relative sentences while applying, amongst other things, the rule of marking and modification previously definite.

The elaborated HPSG grammar will be specified on TDL (Type Description Language). Indeed, TDL language is a language syntactically very similar to the attributes-values structures which are the base of HPSG formalism. In the following paragraph, we give an idea on TDL

syntax and specification of the proposed HPSG grammar for Arabic relatives.

5 TDL Specification

TDL specification of the proposed HPSG grammar requires knowledge about its syntax. The TDL language is a language syntactically very similar to the attributes-values structures which are the base of HPSG formalism. Thus, there are several similarities between HPSG and TDL syntax (Krieger and Schäfer, 1994). These similarities can easily specify HPSG grammars in TDL. Indeed, the addition of the constraints on types is done by the symbol "&". Besides, the co-indexations are preceded by the symbol "#". The comments are preceded by the symbol ";". Moreover, a new type definition is done with the assistance of the symbol ":=". As in HPSG, the feature structures are delimited by brackets [ ].

In order to generate with the LKB a parser dealing with relative sentences, it is necessary to translate into TDL a HPSG lexicon, grammatical rules and a type hierarchy. We propose here an example of TDL specification of marking rule:

```
Regle-marque:=regle-bin-t-fin &
    [SS.LOC.CAT [VAL #val,  MARQUE
   #marque], BRS [BRS-NTETE
   <[SS.LOC.CAT[TETE.SPEC #tete,
     MARQUE #marque]]>,
   BR-TETE [SS #tete &
    [LOC.CAT.VAL #val]]]].
```

Once the syntactic rules are implemented in TDL and gathered in a TDL file named "*rsynt.tdl*", we pass to the experimentation of the grammar implemented in TDL.

The specified linguistic resources (proposed type hierarchy, lexicon and syntactic rules) are used as an input to LKB platform in order to experiment the constructed HPSG grammar. In the next paragraph, we give an idea about LKB platform. Then, we experiment and evaluate the established Arabic grammar.

6 Experimentation and evaluation

Linguistic Knowledge Building (LKB) system is a generation tool, proposed by (Copestake, 2002). It is based on two types of files: TDL files and LISP files. The first type represents the grammar's files. In fact, this grammar is based on seven TDL files: lexicon, type, type-lex, type-rules, rsynt, noeuds and roots.

The second type represents files to parameterize LKB system. It is based on five LISP files. Among these files, we can especially mention the file: "script.lsp' which indicates the name and the repertory of each grammar file.

The evaluation of the constructed grammar is based on a corpus of 500 sentences containing essentially relatives. Besides, the test corpus contains other linguistic phenomena such as the elision, the call, the description. The used lexicon contains approximately 3000 words (~2500 verbs, 450 nouns and 50 particles). It is formed mainly of the corpus words.

For the tested sentences, we note that the generated parser could correctly build their syntactic structures in a reasonable time. In addition, 2% of the sentences do not produce derivation trees, 84% of sentences have only one analysis and 14% have at least two derivation trees.

For the remaining sentences, the failure is due to the existence of more than one derivation tree for the same sentence. In fact, this problem was encountered in previous works using LKB system such as (Garcia, 2005) and (Tseng, 2006). In our work, we introduced other constraints more specific, to resolve the encountered problem according to the proposed type hierarchy. Nevertheless, ambiguous cases persist. This is caused mainly by ambiguities found during relative sentences analysis. Figure 5 represents an example of sentence covering ambiguous cases.

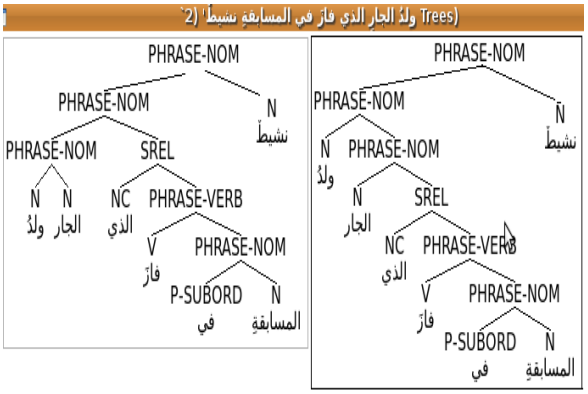

Figure 5: Implementation TDL of "**الذي**"

Indeed, the relative phrase "ولد الجار الذي فاز في المسابقة نشيط" can refer to the noun or to the nominal group "ولد الجار". This nominal group represents an annexed phrase.

Besides, there is another problem at the level of lexicon. This problem was encountered also in previous projects working on Arabic language such as (Alnajem and Alzhouri, 2008), (Bahou *et al.*, 2003) and (Elleuch., 2004). In our work, we have added an interface written in JAVA which can enrich the file "lexique.tdl" by new words automatically and without knowing TDL

syntax. Moreover, this lexicon can easy be extended using tools that we have developed in our laboratory like the translator from LMF toward TDL (Fehri et al, 2006).

7 Conclusion and Perspectives

In this paper, we have constructed an HPSG grammar for Arabic language treating particularly relative sentences. For this reason, we have proposed a type hierarchy categorizing Arabic words in different types. According to the proposed type hierarchy, we brought some modifications to HPSG grammar in order to treat Arabic specificities. The constructed grammar was experimented on LKB platform. Therefore, we specified Arabic HPSG with TDL language. This TDL specification is original, in our work since it integrates some operations and verifies certain concepts such as inheritance, adjunction and recursion. The evaluation phase shows that obtained results are satisfactory.

As perspectives of this work, we aim to test our parser on a larger corpus. We plan also to extend the HPSG description to cover other linguistic phenomena. Also, we plan to extend this work to cover semantic analysis. However, more works should be carried out to cover linguistic ambiguities such as recursion.

References

Abdelkader A., Haddar K. and Ben Hamadou A., « Etude et analyse de la phrase nominale arabe en HPSG », Traitement Automatique des Langue Naturelles, Louvain, UCL Presses de Louvain: 379-388, 2006.

Abdelwahed A., « ‘alkalima fy ‘attourath ‘allisaany ‘alaraby , الكلمة في التراث اللساني العربي », Librairie Aladin 1ère édition, Sfax – Tunisie : 1-100, 2004.

Alnajem S. and Alzhouri F., “An HPSG Approach to Arabic Nominal Sentences”, Journal of the American society for information Scienceand Technology: 422 – 434, 2008.

Aloulou C., « Analyse syntaxique de l’Arabe: Le système MASPAR », RECITAL, NantesFrance, 2003.

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

Belkacemi C., «The relative marker: a definite marker substitute? », ArOr Archiv Orientální, 66/2, 142-148, Based on Arabic dialects, 1998.

Blache P., «Les Grammaires de Propriétés: des contraintes pour le traitement automatique des langues naturelles». Hermès Sciences, Paris, 2001.

Boukedi S., Haddar K. and Abdelwahed A., « Vers une analyse des phrases arabes en HPSG et LKB ». GEI 2008, 8ème Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique, Sousse, Tunisie : 487- 498, 2008.

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

Dahdah A.. « معجم قواعد اللغة العربية في جداول و لوحات », Librairie de Nachirun Lebanon, 5ème edition, 1992.

Elleuch S., « Analyse syntaxique de la langue arabe basée sur le formalisme d’unification HPSG ». Mémoire de DEA en Système d’information et Nouvelles Technologies, Tunisie : 55-88, 2004.

Fehri H., Loukil N., Haddar K. and Ben Hamadou A., “Un système de projection du HPSG arabisé vers la plate-forme LMF ». JETALA, Maroc, 1-11, 2006.

Garcia O., « Une introduction à l’implémentation des relatives de l’espagnol en HPSG–LKB », Mémoire de recherche, 2005.

Haddar K. and Ben Hamadou A., « Un système de recouvrement des ellipses de la langue arabe ». Proceedings of VEXTAL, San Servolo V.I.U. 22(11) : 159-167, 1999.

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

Loukam M. and Laskri M., « Vers la modélisation de la grammaire de l’arabe standard basée sur le formalisme HPSG », Actes JED’2007, Journées de l’Ecole Doctorale, 27(5), Annaba/Algérie, 2007.

Maaloul H., Haddar K. and Ben Hamadou A., «La coordination arabe : étude et analyse en HPSG », MCSEAI 2004, 8ème conférence maghrébine sur le GL et l’IA, Sousse, Tunisie : 487- 498, 2004.

Meurers W. D., «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: 18 – 25, 2002.

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

Tseng J., « Implémentation HPSG avec LKB: La Matrix et la Grenouille », Séminaire HPSG-UFRL, Paris 7, 14(12), 2006