The Implementation of Arabic Subject Markers in the LKB System

Adel Jebali
Université du Québec à Montréal

Abstract

Arabic Subject Markers are interface phenomena (specifically between morphology and syntax). In this paper, I describe them briefly, I give my linguistic analysis within the framework of the Head-Driven Phrase Structure Grammar and I show how I implement them in the LKB system. I show that this system, despite its strength, does not allow for a proper implementation of these units.

Standard Arabic (henceforth Arabic) Subject Markers (henceforth SMs) are morphemes that convey information on the gender, number and the person of subject or topic[1]. They are attached to perfective and imperfective verbs and respect well-defined morphological patterns. They are suffixes when attached to perfective verbs and prefixes and / or suffixes when attached to imperfective verbs. Two examples of these units are provided in (1) and (2). In (1) the SM *-ta*, which encodes the features 2MS, is attached to the perfective stem *katab* "wrote". In (2), the SM *ti :* , which encodes the features 2FS, is attached to the imperfective indicative stem *ktub* "write".

(1) *katab* ***-ta*** *maqa :l* *-a* *-n*
wrote -2MS paper -ACC -INDE
“You (masculine) wrote a paper”

(2) ***t-*** *aktub* ***-i :*** *-na* *maqa :l* *-a* *-n*
2- write -FS -IND paper -ACC -INDE
“You (feminine) write a paper”

Introduction

The study of these linguistic markers go back to the first Arabic grammarians, such as Sibawayhi in the thirteenth century. They have also a privileged place in contemporary linguistic research, although they do not constitute an object of research on their own right. Thus, we find that they are studied by addressing issues affecting the agreement, such as the well-known agreement asymmetries (see, inter alia, (Aoun, Benmamoun, and Sportiche 1994), (Harbert and Bahloul 2002) and (Benmamoun and Lorimor 2006)). (Jebali 2008) is the first work entirely dedicated to the modeling of these markers, a work that includes the description, the linguistic analysis and the implementation.

The problems raised by SMs are mainly morphosyntactic. The question that arises is : are they better treated as arguments or as agreement markers ? The terminology used to designate them in the literature hesitates between disparate labels such as pronouns (even pronominal affixes), agreement markers and clitics. These labels are most often not based on independently motivated criteria and authors who use them do not wonder about the impact of their classification on the analysis of other phenomena encountered in Arabic.

I propose a linguistic analysis of SMs based on independent criteria , I assess the impact of this analysis on other phenomena, I formalize this analysis in the framework of the Head-Driven Phrase Structure Grammar (HPSG) (Ginzburg and Sag 2000), (Pollard and Sag 1994), (Pollard and Sag 1987) and (Sag, Wasow, and Bender 2003). I implement the analysis in the Linguistic Knowledge Building (LKB) system (Copestake 2002) to test its validity and I show some limitations of this system.

Describing SMs

Perfective SMs

When attached to perfective verbs, SMs are suffixes. They precede any other morpheme that could be as well attached to the verb. In the example (3) we have such a case where the SM *-ta* (2MS) is attached to the verbal stem *katab* and the object marker *-ha :* (it) is attached to the left of the SM, any other order being impossible.

(3) *katab* ***-ta*** *-ha :*
wrote -2MS -it
“You wrote it”

Imperfective SMs

SMs attached to imperfective verbs may be prefixes or circumfixes[2]. One example of a prefix SM is given in (4a),

---

[1] See (Li and Thompson 1976).

[2] See (Anderson 1992) for a definition.

	SM	Independent pronoun
2MS	-ta	?anta
2FS	-ti	?anti
2D	-tuma :	?antuma :
2MP	-tum	?antum
2FP	-tunna	?antunna

TAB. 1 – SMs and independent pronouns

where the SM *t-* (2S) is attached to the verbal stem *aktub*. In (4b), the SM is a discontinuous morpheme, i.e. a circumfix, whose two parts are *t-* and *-i :*.

(4) a. ***t-*** *aktub* *-u* *maqa :l* *-a* *-n*
2S- write -IND paper -ACC -INDE
"You write a paper"
b ***t-*** *aktub* ***-i :*** *-na* *maqa :l* *-a* *-n*
2- write -FS -IND paper -ACC -INDE
"You (feminine) write a paper"

Some morphosyntactic properties

Whether they are suffixes, prefixes or circumfixes, SMs appear in three main constructions : in VSO order, in SVO order and in subjectless constructions. In VSO order, as in the example (5), the subject is an NP and the SM bears no specification of number (it is always singular).

(5) *katab* ***-at*** *at-* *tilmi :dha :t* *-u* *maqa :l*
wrote -3F the- students.F -NOM paper
-a *-n*
-ACC -INDE
"The students (feminine) wrote a paper"

In an SVO order, the SM bears the same number specifications as the preverbal component, but the interpretation is slightly different from the one obtained in the VSO order :

(6) *at-* *tilimi :dha :t* *-u* *katab* ***-na*** *maqa :l*
the- students.F wrote -3FP paper -ACC
-a *-n*
-INDE
"The students (feminine), they wrote a paper"

In subjectless constructions, the SMs bear number specifications and stand for the subject, as in the following example :

(7) *katab* ***-u :*** *maqa :l* *-a* *-n*
wrote -3MP paper -ACC -INDE
"They wrote a paper"

Some morphophonological properties

SMs are small parts of nominative independent pronouns, as shown in table (1).

Morphosyntactic	Morphological
Arguments, agreement markers	Affixes

TAB. 2 – Some results

Analyzing SMs

My linguistic analysis of SMs is based on the answers to 2 questions :

1. what is the morphological status of these units ? Are they best treated as affixes or as clitics ?
2. What is their morphosyntactic status ? Are they best treated as arguments or as non-arguments ?

I answer the first question by appealing to the Zwicky criteria as presented in (Zwicky 1977), (Zwicky 1985b), (Zwicky 1985a) and (Zwicky and Pullum 1983). I use as well the criteria proposed by (Miller 1992).

The second question is treated by distinguishing between grammatical agreement and anaphoric agreement as defended by (Bresnan and Mchombo 1987). I use as well the (Li and Thompson 1976) criteria to distinguish between topics and subjects.

The results of this analysis are presented in the table (2).

Modeling and Implementing SMs

This analysis is modeled within the HPSG framework, as presented in (Pollard and Sag 1994), (Pollard and Sag 1987), (Ginzburg and Sag 2000) and (Sag, Wasow, and Bender 2003). The formal apparatus I propose contains the following elements :

- A multiple inheritance type hierarchy (Carpenter 1992).
- Lexical rules : inflectional rules to derive words from lexemes by adding affixes.
- Lexical entries : the stems from which the verbs and the nouns are derived.
- Principles, such as Head-Feature Principle and Valence Principle.
- Phrases are organized in a type hierarchy included in the larger one.

I implement this analysis in the LKB system of (Copestake 2002). This system provides the following tools :

- A compiler specifically designed for the constraints-based formalisms and HPSG in particular.
- A syntactic parser whose output is a tree.
- A generator.
- A GUI.
- A debugger.

My implementation consists in a modular grammar that LKB uses to parse and generate Arabic sentences containing SMs. The modules that make up this grammar are as follows :

1. A type hierarchy : this sub-system acts as a defining framework for the entire grammar, i.e. it defines the retained types, the features appropriated to each type and the degree of compatibility between those types.
2. Lexical entries : lexemes (verb and noun stems).
3. Phrase structure rules designed as constraints on types (and included in the hierarchy as well).
4. Inflection rules : these rules create, on the fly and only in parsing or at the request of the user, words formed from the lexemes provided in the lexicon.

Type Hierarchy

The type hierarchy (and all the grammar) is written in a simple language called TDL. (8), is an example that shows the description of the type *phrase* in the type hierarchy :

(8)
```
phrase := sign & [ COMPS <> ].
```

This description specifies the mother type (*sign*) and the constraints that the type *phrase* must satisfy (to have a null list of complements COMPS <>).

This definition along with the other type definitions are stored in a text file called `types.tdl`. It is the first component the LKB system loads and verifies in the grammar.

Phrase Structure Rules

The grammar rules are typed feature structures that describe the way phrases and words are combined to shape other phrases. In the LKB system, the daughters are encoded in the feature ARGS. The value taken by this feature is a list in which the order of components corresponds to the linear order of daughters in the phrase. The following example shows the rule corresponding to phrases whose head is an intransitive verb :

(9)
```
head-complement-rule-0 :=
head-initial-unary &
[ SUBJ # subj,
  TOPIC # topic,
  ARGS < word & [ SUBJ # subj,
                  TOPIC # topic,
                  COMPS <> ] > ].
```

Lexical Entries

In order to test our grammar, we need to have real words to parse and generate sentences. With LKB, all we need is some lexical entries for verb stems, some nouns and some pronouns. These are described using the same syntax as the one used in the type hierarchy. Here's for example the lexical entry for the noun *walad* (boy) :

(10)
```
walad := lexeme-noun & [
PHON.LIST.FIRST "walad",
HEAD.AGR.GENDER masc,
SEM.RELS.LIST.FIRST.PRED "walad-rel"
].
```

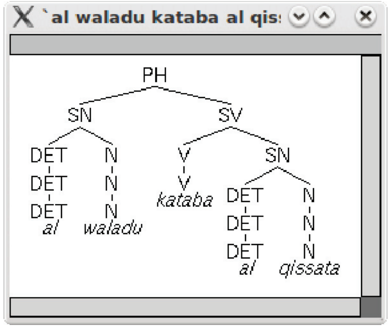


FIG. 1 – A successful parsing of an SVO sentence

This entry specifies the type to which the lexeme *walad* belongs as well as its phonological, morphosyntactic (agreement) and semantic features.

The interpretation of the phrase structure rules and of the lexical entries is somewhat different from that of the type hierarchy. The rightward part after the symbol := in the latter is interpreted as a super-type and as a type in the formers.

Inflection Rules

In my grammar, verbs and nouns are introduced in the lexical entries as lexemes and not as full words. To "derive" real words, we need some inflection rules. Those add affixes (prefixes or suffixes) to the stems as shown in the following example :

(11)
```
lexical-rule-verb-1ms-perfective :=
%suffix (* tu) (na :m nimtu) (xa :f
xiftu)
verb-1ms & [HEAD.AGR agr-1ms].
```

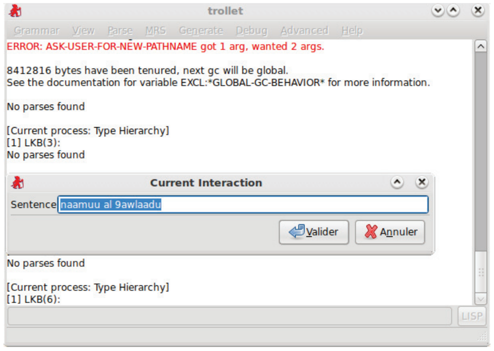
This rule provides the suffix to attach to 1MS perfective verbs (*-tu*). The irregular forms of conjugation are encoded too (such as *xiftu*). The output of this rule is a verb whose agreement features are 1MS.

Results and Problems

My LKB grammar is able to parse a large number of Arabic sentences. 3 examples of the output of the parser are shown hereafter. Figure (1) shows a successful parse. Figure (2) shows the result of an unsuccessful parse (because the sentence is not well-formed) and the figure (3) shows a generation made from a successful parse.

All in all, LKB is a robust system for natural language processing. It is able to parse and to generate well-formed sentences. However, in my experience with this system, I encountered three limitations that seem to be major ones :

1. Large scale grammars are difficult to implement. LKB was designed as an educational tool to show how constraints-based grammars are to be implemented and it

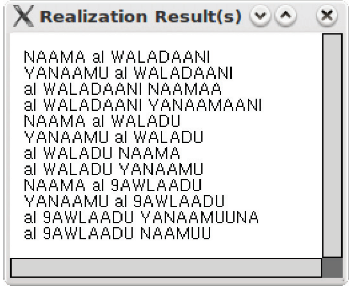


```
ERROR: ASK-USER-FOR-NEW-PATHNAME got 1 arg, wanted 2 args.

8412816 bytes have been tenured, next gc will be global.
See the documentation for variable EXCL:*GLOBAL-GC-BEHAVIOR* for more information.

No parses found

[Current process: Type Hierarchy]
[1] LKB(3):
No parses found
```

Sentence naamuu al 9awlaadu

```
No parses found

[Current process: Type Hierarchy]
[1] LKB(6):
```

FIG. 2 – No parse



```
NAAMA al WALADAANI
YANAAMU al WALADAANI
al WALADAANI NAAMAA
al WALADAANI YANAAMAANI
NAAMA al WALADU
YANAAMU al WALADU
al WALADU NAAMA
al WALADU YANAAMU
NAAMA al 9AWLAADU
YANAAMU al 9AWLAADU
al 9AWLAADU YANAAMUUNA
al 9AWLAADU NAAMUU
```

FIG. 3 – Generation of VS sentences

is not suitable for large scale grammars. This is a serious limitation in my case as the SMs cannot be implemented without implementing most of the Arabic grammar.

2. The morphological module is very limited and does not allow us to add at the same time prefixes and suffixes and it does not allow at all to add infixes to stems. This poses no problem for English, but Arabic has a rich morphology and imperfective verbs, for example, are often formed with both prefixes and suffixes.
3. The generation is a resource consuming process and it fails to produce sentences with more than 4 or five words.

Conclusion

I introduced the Arabic SMs and their linguistic behavior. I analyzed this behavior in the framework of HPSG and I implemented it within the LKB system. It turned out that this implementation, even if it manages to provide successful parsing and generation, may not be complete because of limitations within the system. I think that these limitations can be overcome especially as the LKB system is open source and can therefore be changed to better suit the restrictions imposed by the grammar of Arabic.

References

Anderson, S. R. 1992. *A-Morphous Morphology*. Cambridge : Cambridge University Press.

Aoun, J. ; Benmamoun, E. ; and Sportiche, D. 1994. Agreement, word order, and conjugation in some varieties of arabic. *Linguistic Inquiry* 25(2) :195–220.

Benmamoun, E., and Lorimor, H. 2006. Featureless expressions : When morphophonological markers are absent. *Linguistic Inquiry* 37(1) :1–23.

Bresnan, J., and Mchombo, S. A. 1987. Topic, pronoun and agreement in chichewa. *Language* 63(4) :741–782.

Carpenter, B. 1992. *The Logic of Typed Feature Structures*. Cambridge, England : Cambridge University Press.

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

Ginzburg, J., and Sag, I. 2000. *Interrogative Investigations : The Form, Meaning and Use of English Interrogative Constructions*. Stanford : CSLI Publications.

Harbert, W., and Bahloul, M. 2002. Postverbal subjects in arabic and the theory of agreement. In Ouhalla, J., and Shlonsky, U., eds., *Themes in Arabic and Hebrew Syntax*. Dordrecht : Kluwer Academic Publishers. 45–70.

Jebali, A. 2008. *La modélisation des marqueurs d'arguments de l'arabe standard dans le cadre des grammaires à base de contraintes*. Doctoral dissertation, Université du Québec à Montréal.

Li, C. N., and Thompson, S. A. 1976. Subject and topic : A new typology of language. In Li, C. N., ed., *Subject and Topic*. New York : Academic Press. 457–489.

Miller, P. H. 1992. *Clitics and Constituents in Phrase Structure Grammar*. New York : Garland.

Pollard, C., and Sag, I. A. 1987. *Information-Based Syntax and Semantics, Vol. 1*. Stanford : Stanford University Press, CSLI.

Pollard, C., and Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. Stanford : Stanford University Press, CSLI.

Sag, I. A. ; Wasow, T. ; and Bender, E. 2003. *Syntactic Theory : A Formal Introduction*. CSLI Publications.

Zwicky, A. M., and Pullum, G. 1983. Clicticization vs. inflection : English n't. *Language* 59(3) :502–513.

Zwicky, A. M. 1977. *On Clitics*. Bloomington : Indiana University Linguistic Club.

Zwicky, A. M. 1985a. Cliticization versus inflection : The hidatsa mood markers. *International Journal of American Linguistics* 51(4) :629–630.

Zwicky, A. M. 1985b. Clitics and particles. *Language* 61(2) :283–305.