A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic

Everhard Ditters
TCNMO, Nijmegen University
The Netherlands
e.ditters@let.kun.nl

0 Abstract

One way to describe the structure of the sentence in Modern Standard Arabic (MSA) is to draft a hypothesis and to test it on authentic data in an automatic processing environment. The result will be a verified theory about sentence structure in MSA as well as a collection of analysed data ready for teaching and research purposes. An operational version of the below discussed Toulouse MSA Sentence Grammar with a rather limited lexicon will be available at Workshop Time.

1 Linguistic description of the sentence

In the framework of generative linguistics (Universal Grammar - UG) Fassi Fehri (1993) has been elaborating an interpretative[1] description of sentence structure in Modern Standard Arabic (MSA). He adapted by means of specific raising rules the basic unmarked VSO (verb – subject – object) sequence in order to accommodate for sentence structures in MSA following a clear SVO (subject – verb - object) sequence. Positive in this approach is the attempt to extent the scope of universal grammar applicability with language facts from other natural languages. Negative is the identification of the SVO order typology with the nominal sentence structure in MSA in general. Instead of Fassi Fehri's (1993) IP-structures we rather prefer to present a single IP-structure description in which the slash represents alternatives:

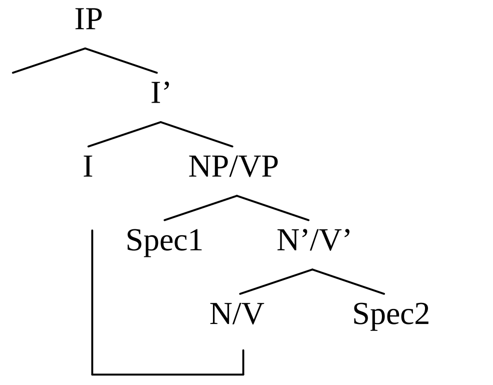

Figure 1

Quoting only parts of our UG-source for MSA data (Fassi Fehri, 1993, pp. 16 e.s.) this figure should be read as:

- the subject is base-generated in Spec1 of VP and the unmarked VSO order is derived at S-structure by raising V to I;
- a VP comment is base-generated in Spec1 of the topic NP and the unmarked SVO order is derived at S-structure by raising N to I;
- other realisations of the comment function than a VP are rewritings of Spec2.

In the framework of corpus linguistics we are elaborating a declarative[2] description of sentence structure in MSA with a specific function and in terms of immediate constituents realising the obligatory and/or optional functions. Both the word order and the mutual relationships and dependencies will be accounted for by means of a two-level attribute

[1] Interpretative here hints at the main objective of the UG-approach and well to explain and describe language facts in order to draft the underlying language independent grammar system.

[2] Declarative here hints at the non-interpretative description of language facts resulting in an account of syntactic and semantic structures the correctness of which is to be tested on authentic data.

grammar formalism, the so-called Affix Grammar over Finite Lattices (AGFL).[3]

We distinguish in MSA two different sentence types with contextual defective forms:

- a nominal sentence with an enonciative function and the following characteristics:
  - o the main or central category involved in topic function, is a noun phrase (NP);
  - o the constituents able to occur in comment function are: another NP, an adjective phrase (ADJP), an adverb phrase (ADVP), a prepositional phrase (PP), a verb phrase (VP) or an clause (CL);
  - o the sequencing of the constituents involved is juxtaposition;
  - o relationships and dependencies between the main category and other components able to occur are expressed by full (rich), partial (poor or weak) and zero agreement of co-occurring morpho-syntactic and/or semantic features;
  - o an optional modifier function accounts for an interrogative, negative or conditional sentence type;
- a verbal sentence with a predicative function and the following characteristics:
  - o the main or central category involved is a verb phrase (VP);
  - o the constituents able to occur in the VP are in form and function lexically defined in the complement structure of the verbal entry;
  - o the sequencing of the constituents involved is government by flexion;
  - o relationships and dependencies between the main category and other components able to occur are expressed by partial (poor or weak) and zero agreement of co-occurring morpho-syntactic and/or semantic features;
  - o an optional modifier function accounts for an interrogative, negative or conditional sentence type.
- contextual defective forms: elliptic occurrences of both types:
  - o one (or more) of the obligatory function realisations but possibly occurring at different and/or lower levels of description is (are) deleted but its (their) semantic value is known from context.

2 Formal description of the sentence

As we have in mind the automatic analysis of raw corpora our formal grammar accounts for untagged collections of MSA text data. So we will begin our description at text level in order to go down to the lexical level. The formalism chosen is capable to describe natural language structures by means of context-free rewrite rules while a generator automatically translates the formal grammar into a parser.

2.1 The formalism

The first level of description accounts for word order (defined by hyper-rules) making use of a database of terminal items organised by categories (lexical rules). The second level, equally organised in the form of context-free rewrite rules, accounts for relationships and dependencies by means of affixes attached to the non-terminals of the first level but with a limited domain of values (listed in meta-rules), controlling the information flow implicitly (by means of value declarations) or explicitly (by means of predicate rules). The order of the rules is not pertinent since the grammar is materialised in a non-deterministic static formalism.

2.2 The grammar

The general conventions for grammar writing and the use of the AGFL-formalism are as follows. Optional elements in a right hand side are enclosed within square brackets (“[]”). A hash (#) initiates a not-processable comment line. A colon (:) represents the rewrite symbol in hyper-rules, lexical rules and predicate rules. A double colon (::) represents the rewrite symbol in meta-rules. A comma separates members of a rule. A semi-colon separates alternatives. Literal values are enclosed between double quotes. A period ends each rule.

[3] For relevant information about the AGFL formalism and the NLP processing environment we refer to Koster 1971 and 1992, Meijer 1986, van Halteren 1997. The AGFL-documentation and software are available at: www.cs.kun.nl/agfl.

2.2.1 The First level

We first present the hyper-rules in order to account first for the linear sequence of elements constituting the two distinct sentence structures in MSA with obligatory and optional function slots. In § 2.2.3 we discuss the extension of the non-terminals of the first level with affix variables.

2.2.1.1 Hyper-rules

```
# Toulouse Grammar
# Sentence Level
# Part 1: global sentence definition
Start Text.
Text : Utterances.
Utterances : Utterance; Utterance, Utterances.

Utterance :  [Connector], Base, [End Marker].
Connector :  Coord.
Base : S type; S type, Coord, S type.
End Marker :    Mark.
# Part 2: sentence type differentiation
S type :   Sentence; Double S;
           Elliptic S.
Double S :   Condition, Answer;
             Answer, Condition.
Condition :  Cond Part, Sentence.
Answer :     Sentence.
Elliptic S : Elliptic VS; Elliptic NS.
# Part 3: sentence description
Sentence :   Expression.
Expression : Predication; Enonciation.
Predication :   Verbal S.
# Part 4: the verbal sentence
Verbal S :      [S Modifier], VS Head,
                [S Modifier].
Elliptic VS :   [S Modifier], Elliptic VP,
                [S Modifier].
Elliptic VP :   VP Part.
# Part 5: the nominal sentence
Enonciation :   Nominal S.
Nominal S :     [S Modifier], Topic,
                [S Modifier], Comment,
                [S Modifier];
                [S Modifier], Comment,
                [S Modifier], Topic,
                [S Modifier].
Elliptic NS :   [S Modifier], Topic,
                [S Modifier];
                [S Modifier], Comment,
                [S Modifier].
# Part 6: description of sentence components
S Modifier :    S Affirmation;
                S Interrogation;
                S Interrogation, S Negation;
                S Negation; S Separation.
VS Head :       VP.
Topic :         NP; CCL.
Comment :       NP; ADJP; ADVP; PP; VP;
                CL.
S Affirmation :    ADVP.
S Interrogation :  ADVP.
S Negation :       ADVP.
S Separation :     Pers Pron.
```

2.2.1.2 Discussion hyper-rules

The start symbol of our grammar is *Text*. As much as possible the description in terms of alternating function and category layers is maintained. We made an exception for the category *Text*, rewritten into the category *Utterances*, and for the Sentence function *Expression*, rewritten into the functions *Predication* or *Enunciation*.

By means of the non-terminal *Double S* we account for the occurrence of conditional sentences in MSA. Our research is aiming at the analysis of language data and not language generation. We could have limited ourselves to an alternative rewriting of the sentence modifier (*S Modifier*) into Condition, but we preferred to describe the conditional sentence in MSA here more explicitly.

As we said earlier, elliptic sentences are found where the missing information can be understood from the context in which the sentence occurs. We should be aware that the missing link might have its source at lower levels of the sentence description. In an elliptic verbal sentence, for example, the ellipse will mostly has taken place somewhere in the realisation of the predicative function, the VP.

At this level of description we account by means of sentence modifier affirmation (*S Affirmation*) for a Yes-answer to questions. Interrogation, at this level is restricted to the occurrence of one of the question particles (“hal” or “’a”) followed by a sentence. Wh-question types are treated at phrasal level since the wh-elements are realisations of sentence constituents. The sentence modifier separation

(*S Separation*) accounts for the comment realisation in the form of a definite NP. This modifier is realised in the form of a third person personal pronoun in gender and number agreement with the head of the topic NP.[4]

2.2.2.1 Lexical rules

VP : “VP”.
VP Part : “VP Part”.
NP : “NP”.
ADJP : “ADJP”.
ADVP : “ADVP”.
PP : “PP”.
CL : “CL”.
CCL : “CCL”.
Mark : “.”; “,”; “:”; “;”; “…”; “!”; “?”.
Cond Part : “Cond”.
Coord : “Coord”.
Pers Pron : “Pers Pron”.

2.2.2.2 Discussion lexical rules

In other publications but especially in Ditters (1992)[5] we presented a more or less exhaustive linguistic and formal description of phrasal categories in MSA like the NP, VP,[6] PP, ADJP and ADVP. We, therefore, refer to those publications and include here these categories in the Lexicon. However, the clause[7] (CL), complement clause[8] (CCL) and also the conditional clause (CONCL) surpass the phrasal level and have to be described at sentence level.

However, as shown in the rewriting for ‘Mark’, a lexical rule is a listing of terminal values or literals not further analysable in smaller constituents. The lexical rule for a conditional particle (‘Cond Part’) has in its right hand side a list of alternatives such as ‘law’, ‘’iḏâ’, ‘’in’. The lexical rule for the coordinative particle (‘Coord’) has in its right hand side a list of alternatives such as ‘wa’, ‘fa’, ‘’aw’, ‘’am’, ‘lâkin’, ‘lâkinna’ etc. The lexical rule for the personal pronoun (‘Pers Pron’) contains a listing of lexical items mutually differing in realised values for person, gender and number. The sole possible subcategorisation could be realised by means of affixes.

2.2.3 The second level

In Ditters 1992 we attached 4 affixes (tense, person, gender and number) to the phrasal category VP, 5 affixes to its head function (the same 4 of the VP plus one to account for complementation) and another 6 affixes to the verbal base (radicals, type, derivation, tense, voice and complementation). To the phrasal category NP we attached 6 affixes (np var, definiteness, gender, number, person and case) and 6 to its head function (the same as for the NP but headrealisation instead of np var). To the phrasal category ADJP we attached 4 affixes (definiteness, gender, number and case).

In the meantime we have considerably extended the number of affixes for the VP, the VP Head, the NP and the NP Head in order to account for semantic features linked to the corresponding lexical entries like human or non-human, animated or non-animated, concrete or non-concrete, countable or non-countable etc.

It is the strength of the AGFL formalism to allow for the attachment of an unlimited number of affixes to a non-terminal on the condition that the domain of their respective individual values is finite.

[4] See for examples Cantarino, 1974, Vol. I, p. 35 and also Ditters 2002, the examples (17) and (18). Another way to account for this kind of nominal sentence structure is to add the category ‘Nominal S’ to the list of alternatives of comment realizations.

[5] A new edition is forthcoming.

[6] In Ditters 1992 (p.228 and p.332) we discussed the occurrence of the ellipse within the NP and the VP. We limited ourselves here to the mention of a terminal value for *VP Part* in the lexical rules.

[7] A clause is an embedded sentence. We refer for its description to the rewriting of *Sentence* in our Toulouse Grammar.

[8] A complement clause can be rewritten as a sequence of a conjunction followed by a conjunctive complement. Depending on the type of conjunction realization (’an, ’anna or combinations of these conjunctions) the conjunctive complement is realized in the form of a Verbal S and Nominal S respectively according to the rules of our Toulouse grammar. A ’an type of conjunction governs the mood value of the verbal entry by imposing the subjunctive. A ’anna type of conjunction imposes an accusative case value at the topic realization.

2.2.3.1 Meta-rules

```
# Part 1: sentence
cat          ::  Alter; Change; Cumul;
                 Nrest ; Restr.
nature       ::  Possible; Real; Unreal.
# Part 2: VP
complementation :: Diprep; Ditr; Intr; Prep;
                    Prepacc; Trans; Triprep;
                    Tritr.
derivation     ::  I; derived.
derived        ::  II; III; IV; V; VI; VII; VIII;
                   IX; X.
mood         ::  Imper; Indic; Juss; Subj.
radicals     ::  Three; Four.
tense        ::  Perfect; mood.
type         ::  Daw; Double23; Lqy; Normal ;
                 Ray; Rmy; Waw1; Waw2 ;
                 Waw3; Ya3.
voice        ::  Active; Passive
# Part 3: NP and ADJP
animatedness     ::  Animated; Nonanimated.
case             ::  Acc; Gen; Nom; Inv.
concreteness     ::  Concrete; Nonconcrete.
countability     ::  Countable; Noncountable.
definiteness  ::  Def; Indef.
gender        ::  Fem; Masc.
headrealisation ::  Com; Count; Elative; Intn;
                    Min; Nad; Neg; Nnum;
                    Nomcom; Num; Pers;
                    Prop; Ques; Vera; Vern;
                    Voc.
humanness     :: Human; Nonhuman.
nonplur       :: Coll; Dual; Sing
np var       ::  Compar; Count; Nomcom;
                 Reform; simple.
number       ::  nonplur; plur.
person       ::  First ; Second ; Third.
plur         ::  Explu; Inplu.
simple       ::  Com; Com Con; Com Con Pom.
```

2.2.3.2 Discussion meta-rules

At sentence level we like to attach an affix ‘cat’ to a sentence initialising connector. The reason is that sentences in MSA may start with a connective identical with a connective that, at phrasal level, is used for co-ordination. There the affix will help to control agreement phenomena connected with alternative, cumulative, restrictive or non-restrictive co-ordination. Concerning conditional sentences we like to distinguish between ‘real’, ‘unreal’ and ‘possible’ conditions.

We also we have to declare the value domain for the affixes linked with the phrasal constituents of the sentence such as the VP, NP and ADJP. Information about the complement structure of the lexical verbal entry plays an important role in the analysis process. For this reason the affix ‘complementation’ is attached to the verbal head with the values: di-prepositional, ditransitive, intransitive, transitive etc. This affix ‘complementation’ is also linked to the verbal derivative since the verbal noun and participles in MSA may govern other constituents like their corresponding verbal entries.

The grammar rules describing verbal morphology make use of the affix ‘derivation’ in order to account for the different stems in MSA. In the rewriting of ‘derivation’ we find at the right hand side the literal value ‘I’, indicating the first stem, and another affix ‘derived’ which, in separate meta-rule, is rewritten into other literal values referring to the other derived stems. Other relevant information concerning the VP is stored in affixes defining the value for aspect (‘mood’ and ‘tense’), the number of ‘radicals’ of the verbal root, voice and the type of verb (specific labels for strong verbs, hollow verbs etc.).

The affixes used with the NP and ADJP are of different categories. Some distinguish wordclasses (‘headrealisation’) or types of NPs (‘np var’) such as comparatives, numeral phrases, reformulation, construct states etc., other are referring to semantic features (‘animatedness’, ‘concreteness’, ‘countability’ etc.)

2.2.4 Instantiation of affix values

The Arabic equivalents for demonstratives have the value ‘Def’ for the affix ‘definiteness’. The counterparts of the indefinite pronouns are marked for ‘Indef’. The Arabic equivalent for the personal pronoun ‘he’ is marked for third person masculine singular (‘Third’, ‘Masc’, ‘Sing’). These values are inherited features, which they bring along when realised as constituent in a sentence.

Like nominal lexical entries also verbal lexical entries have inherited features by which they govern the arguments able to occur in their environment. However, when realising the comment function in a nominal sentence or in connection with an explicit subject in a verbal sentence we are confronted with syntactic phenomena like agreement and concord. In this cases realised affix values of the topic or the explicit subject impose specific values at the head of the VP. We are then speaking of derived affix values. In a general way we can say that all kind of linguistic modifiers if susceptible to morphological variation have derived affixes and inherit their values from the lexical entry they modify.

The instantiation of affix values may take place in an implicit and an explicit way. Identical affix names in the right hand side of a rewrite rule imply identical affix values. Here we are dealing with implicit instantiation. The imposition of an affix value can take place by imposing a specific affix value for a specific non-terminal in a specific context or by an explicit listing of acceptable combinations by means of predicate rules.

2.2.4.1 Predicate rules

```
# Part 1: concord rules sentence level
topic and comment
    (humanness, gender, number,
                    gender1, number1).
# Value listing
topic and comment
    (Human, gender, number,
                gender, number) : .
topic and comment
    (Nonhuman, gender, nonplur,
                    gender, nonplur) : .
topic and comment
    (Nonhuman, gender, plur,
                    Fem,    Sing) : .
# Part 2: agreement rules at phrasal level
verb explicit subject (gender, number,
                          gender1, number1).
noun adjective
    (humanness, gender, number,
                    gender1, number1).
# Value listings
verb explicit subject (gender, Sing,
                          gender, nonplur): .
verb explicit subject (Fem,    Sing,
                          gender, plur): .
noun adjective
    (Human, gender, number,
                gender, number): .
noun adjective
    (Nonhuman, gender, nonplur,
                    gender, nonplur): .
noun adjective
    (Nonhuman, gender, plur,
                    Fem,    Sing): .
```

2.2.4.2 Discussion of predicate rules

In our formal grammar of sentence structure in MSA we use predicate rules to describe the characteristics of concord between constituents within a sentence as well as the characteristics of agreement between the elements within a constituent. A predicate rule is an alternative member at the right hand site of a rewrite rule listing the affix names, which values are to be mutually conditioned. This predicate rule is then as left hand side rewritten into an empty rule listing the acceptable values of the affixes concerned.

In part 1 we see the topic and comment predicate rule linking the gender and number value realised in the comment with the value for humanness, gender and number of the head of the topic NP. It doesn’t matter whether the comment function is realised by an NP, an ADJP or a VP. As can be seen in the value listing, only for a plural Nonhuman head of the NP topic a special form of concord is assumed.

In part 2 we have two different predicate rules: the first controlling the different possibilities of value agreement between a verbal head with its explicit subject in a verbal sentence; the second describing value agreement between the head of an NP and a postmodifier whether the later is realised in the form of an NP, an ADJP or a relative clause (Rel cl).[9]

Finally, a well-considered choice of affix names like the rewriting of number into ‘plur’ or ‘nonplur’ considerably decreases the number of alternatives required for the

[9] A PP in postmodifying function to the head of an NP is not susceptible to the imposition of affix values from the head. This realisation of postmodification is simply expressed by juxtaposition.

empty predicate rules defining the concord and agreement phenomena at sentence and phrasal level.

3 Integration of the levels

We discussed so far the different types of rules available within the AGFL-formalism for the description of the syntactic and semantic structure of sentences in MSA. What is needed is to combine the different levels of description. In other words, affix names have to be attached to non-terminals of the first level. Predicate rules have to be inserted there where appropriate for the filtering out of undesired combinations. A text corpus of data should be prepared and/or pre-processed for the testing of the grammar on the data.

During the compilation of the grammar the AGFL processing environment is checking the grammar for mistakes and faults against the syntax of the rules, formal inconsistencies in the description, missing definitions of right hand side members and occurrences of left recursion. The compiler equally provides statistical information about the number of rules, the number of non-terminals, affix names and affix values.

Once the grammar has correctly been compiled a cyclic process of refining the formal grammar leads to the filtering out of undesired results and the creation of a corpus of analysed data that can be stored in a linguistic database.

4 Summary and conclusion

In this paper we presented a linguistic model for the syntactic and semantic description of the sentence in MSA. We distinguished two different sentence types each with its characteristics.

The linguistic description consists of an analysis in terms of immediate constituents enriched with a component accounting for relationships and dependencies between constituents as well as between the elements within a constituent. The analysis alternates between a function and a category level until the final entries have been reached. The generalisation of the description makes use of heads and modifiers as well as slots and fillers.

The processing of large corpora of text data in an adequate, consistent and coherent way requires the use of automatic tools. A two-level affix formalism for the description of natural languages is available together with a parser generator, which automatically translates the formal grammar into a parser.

The AGFL formalism has been discussed and is applied for the description of the sentence structure in MSA. In the past the same formalism and processing environment have successfully been used for the description of the syntactic structure and the automatic analysis of a number of modern western languages. It has successfully been used for the automatic analysis of phrasal categories in MSA. We just end its application for the description of sentence structure in MSA.

5 References

Cantarino, Vicente. 1974-5. *Syntax of Modern Arabic Prose*. 3 Vols. Bloomington: Indiana University Press.

Ditters, Everhard. 1992. *A Formal Approach to Arabic Syntax: The Noun Phrase and the Verb Phrase*. PhD Nijmegen University. Nijmegen: Luxor.

Ditters, Everhard. forthcoming. Distinct(ive) Sentence Functions in Descriptive Arabic Linguistics, in: Parkinson, Dilworth (ed.): *Perspectives on Arabic Linguistics XV*. Amsterdam: John Benjamins.

Ditters, Everhard. forthcoming. *A Formal Approach to Arabic Syntax: The Sentence*. Amsterdam: Rodopi.

Fassi Fehri, Abdelkader. 1993. *Issues in the Structure of Arabic Clauses and Words*. Dordrecht: Kluwer.

Halteren, Hans van. 1997. *Excursions into Syntactic Databases*. Amsterdam: Rodopi.

Koster, Kees. 1971. Affix Grammars, in: Peck, J (ed.): *Algol 68 Implementation*. Amsterdam: North-Holland. Pp. 95-105.

Koster, Kees. 1992. Affix Grammars for Natural Languages, in Albas, H. and B. Melichar (eds.): *Attribute Grammar Applications and Systems, SLNCS, 545*, Springer, pp. 469-484.

Meijer, Hans. 1986. *Pro Grammar: A Translator Generator*. Nijmegen: Bloembergen Santee.