A Syntactic Lexicon for Arabic Verbs

Noureddine loukil†, Kais Haddar‡, Abdelmajid Benhamadou†

Multimedia Information Retrieval and Advanced Computing Laboratory
University of Sfax, Tunisia
†{noureddine.loukil,abdelmajid.benhamadou}@isimsf.rnu.tn, ‡kais.haddar@fss.rnu.tn

Abstract

In this paper, we present a modeling of the syntactic lexicon for Arabic verbs based on the Lexical Markup Framework. This ISO standard let us describe the lexical information in a simple way using general guidelines and enable the sharing of resources following the standard. We discuss the syntactic information associated to verbs and the model we propose to structure and represent the entries within the lexicon. To study the usability of the model, we implemented a rule-based system that translates a LMF syntactic resource into Type Description Language. The generated lexicon is used as input for a previously written HPSG grammar for Arabic built within the Language Knowledge Builder platform. Finally, we discuss improvements in parsing results and possible perspectives of this work.

1. Introduction

Computational lexicons that encode syntactic information are known to be difficult to construct. Such a resource provides specifications of syntactic behaviour like surface properties, subcategorization frames, argument realizations and morphology syntax interaction. This kind of information is very useful especially for grammar parsers.

The construction of such resources with the required quality and coverage is time and effort consuming. This is due to the absence of dictionaries or existing lexical databases from which we can extract the syntactic knowledge. This kind of resources is not yet available for less studied languages like Arabic.

On the other side, the NLP community is using a plethora of grammatical theories and representation formats making it difficult to share the valuable resources. In effect, diversity in lexicon representations does not encourage diffusion and exchange between different communities and tend to be an obstacle to build large coverage lexicons. In such situation, the need for a standardization effort is evident and the adoption of a standard will enable NLP application interoperability and ease of maintenance of large lexicons. Evaluation and sharing of results provided by NLP applications is not an easy task because the majority of works generally use proprietary formats and descriptions rather than normalized ones.

In this paper, we identify and discuss possible syntactic information that can be embedded in the syntactic lexicon of verbs in Arabic, based on the syntactic extension of LMF. Then, we present a system of rules enabling the transformation of the LMF compliant lexicon into TDL (Type description Language), a typed feature-based language and inference system, which is specifically designed to support highly, lexicalized grammar theories like HPSG. Finally we discuss the main results and issues that we faced in this work.

2. Arabic LMF syntactic lexicon

LMF (Francopoulo, 2005) provides an extensible architecture that is relevant for modelling both Machine Readable Dictionaries and NLP lexical resources. LMF models consists of a core model and zero, one or more lexical extensions, along with a set of data categories used in specifying the model.

The actual representation of an LMF compatible lexicon is based on XML. As presented in figure 1, the core model is a hierarchical structure consisting of multiple components. The Lexical Entry component represents an actual elementary lexical entry. The Syntactic Behaviour component provides a representation of all the possible syntactic constructions of a lexical entry. Finally, the Sense component provides a semantic description, which can be divided into multiple senses.

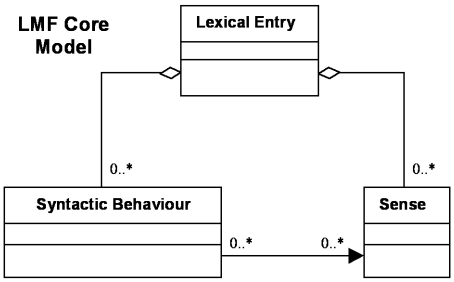

Figure 1: Lexical Markup Framework Core Model.

2.1. Syntax modeling

The Syntax extension specification is given in figure 2 and is built around the Syntactic Behaviour component. A syntactic behaviour is a syntactic formation pattern, which may be adopted by several lexical entries to capture syntactic redundancy in the lexicon. A syntactic behaviour is described by the set of permitted syntactic formations eventually grouped in semantically disjoint subsets.

A Subcategorization frame represents the set of possible syntactic constructions associated to a predicate and actually realized by the combination of several complements or positions. Within it, possible instances of the positions can lead to syntactically correct phrases. In other words, a Subcategorization frame can be seen as valence pattern providing a specification about the order and the nature of permitted positions instances.

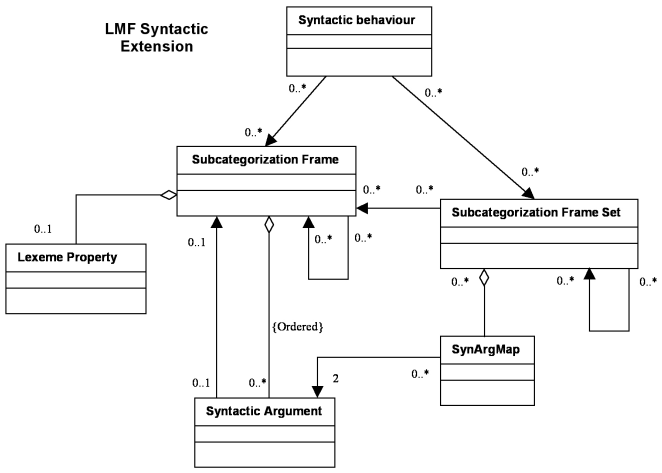

Figure 2: Lexical Markup Framework Syntax Extension.

A lexical entry may have several frames providing each several mandatory or optional positions. Each position proposes possible realizations and their morphological-syntactic descriptions given within the Syntactic Argument component. The Lexeme Property component describe syntactic features special to the lexical entry like tense and mood. In our model, we assume simply that an acceptable syntactic formation for a given verb is embedded in a subcategorization frame (SF).
An SF consists of an ordered list of the arguments required by the verb, and a set of constraints on those arguments like information about complement introducers. Thus, we describe, firstly, the type and the number of arguments, and secondly, the types of constraints associated with arguments.

2.2. Type Hierarchy

Arabic is a Semitic language characterized by a rich morphology and a relatively free word order. The extensional description of the lexicon is a very practical solution to deal with the complexity of morphology. Thus, we can avoid the compilation process common to intensional lexicons in order to obtain the extensional form used by parsers. While using the extensional model, the redundancy inherent to the lexicon can be captured by capturing common descriptions and behaviours.

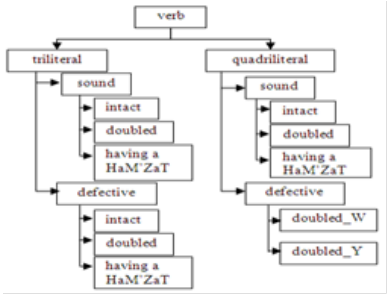

Figure 3: Type Hierarchy for Arabic verb

A verb has two voices: active and passive with some exceptions of verbs with no passive form. Arabic verbs have five moods: indicative, subjunctive, jussive, imperative and energetic. The syntactic constructions accepted by a certain verb depend on its meaning, voice and mood.
According to figure 3, several criteria are presented to categorize an Arabic verb. It can be subdivided according to the number of letters that compose it or according to whether it is augmented or denuded. We choose to subdivide the Arabic verbs according to the first criterion. Thus, a verb can be triliteral or quadriliteral and triliteral or quadriliteral verbs can be sound or defective. Each type has different possible values what makes possible to distinguish the various Arabic verbs.
The verb in Arabic subcategorize for one, two or three required complements, no more. The complement may be direct, i.e. a simple noun, or introduced by a preposition particle from the finite set min, ila, bi,ala, an,fi, lem. In the case that a verb is transitive by particle, it selects which particle can be used. For instance the verb taharraka (to move) accepts only the particles ala, bi, min, lem. This constraint is described by the restriction property that states selected particles for a particular verb.
Arabic shows two different kinds of sentence formations: Verbal and nominal. Nominal sentences adopt the SVO word order and are less frequent. Verbal sentences uses the VSO word order that is more practiced. The fact that each verb can be invoked in SVO or VSO construction varieties is modelled in LMF by using two subcategorization frame sets. The first states the possible SFs in verbal sentences. The second states the possible SFs in nominal sentences.
Our lexicon is built semi automatically using the editor Lexus, a generic tool for the creation of LMF compliant lexicons. Lexus offers a web interface and takes as input the structure of the lexicon and the information for entries. It generates a lexicon compatible with LMF in XML format. The lexicon structure is based on the verbal type hierarchy described in (Loukil et al., b) and (Boukedi et al.,). We considered also proposals on the structure for Arabic LMF lexicon (Loukil et al., a). The followed method has 3 steps: firstly, we specify the subcategorization frames accepted by verbs in Arabic. This is done manually and offline. Then we edit those SFs [1] , as those of figure 4 and figure 5, with the Lexus editor, which perform a compatibility check of the proposed structure with LMF. Finally, we edit Arabic verb lemmas and we affect one or many SFs to each entered verb. The lexicon contains 2500 verb lemmas with an average of 2.7 SF per verb.

2.3. Type Description Language

In a TDL (Krieger and Schafer, 1994) lexicon, entries are encoded as types. Each type is represented by an attribute value matrix (AVM) in which, value can be an atom, a feature structure, an index, a list or even a functional constraint. In case where no value has been specified for a certain attribute, it will have an empty feature structure as a value. Figure 6 shows the TDL specification for the verb jalasa (to sit). Figure 7 introduce a new TDL description for the same verb with syntactic information inherited from the type intransitiveVerb.

[1] 17 SFs in our lexicon

$\langle SubcategorizationFrameid = "TVDv"$
$label = "Trans.verb, Directcompl.inverbalsent."\rangle$
$\langle LexemePropertyposition = "1"$
$partOfSpeech = "verb"$
$mood = "indicative"$
$voice = "active"/\rangle$
$\langle SyntacticArgument$
$function = "subject"$
$syntacticConstituent = "NP"/\rangle$
$\langle SyntacticArgument$
$function = "object"$
$syntacticConstituent = "NP"/\rangle$
$\langle/SubcategorizationFrame >$

Figure 4: Subcategorization frame 1

$\langle SubcategorizationFrameid = "TVDv"$
$label = "Trans.verb, Directcompl.inverbalsent."\rangle$
$\langle LexemePropertyposition = "1"$
$partOfSpeech = "verb"$
$mood = "indicative"$
$voice = "active"/\rangle$
$\langle SyntacticArgument$
$function = "subject"$
$syntacticConstituent = "NP"/\rangle$
$\langle SyntacticArgument$
$function = "object"$
$syntacticConstituent = "NP"/\rangle$
$\langle/SubcategorizationFrame >$

Figure 5: Subcategorization frame 2

2.4. Lexicon transformation rules

In this section, we discuss a set of rules allowing the transformation from an LMF compliant representation to a TDL equivalent description. We notice that most of the information present in LMF possesses its own correspondent in TDL. The identified rule system is based on the correspondence between LMF and TDL. Every LMF lexical entry is described in TDL using features structures and every LMF data category will be translated into a TDL attribute. In order for the system to function correctly, there is an offline step that has to be fulfilled. In fact, we have to construct TDL types for each syntactic behavior along with the general TDL type of the verb. So, it is generally a one to one correspondence. Each rule possesses a determined structure as illustrated in figure 8.

A transformation rule is composed of two parts: The first one titled IN contains the path to reach the LMF attribute value. The second part titled OUT is a simple feature structure that represent a TDL type and denotes the corresponding type in TDL. This structure has the corresponding path of the feature whose value will be affected by the transformation. To illustrate that, we give the rules of figure 8 and figure 9.

In rule of figure 10, the data category PartOfSpeech corresponds to MAJ in TDL. Indeed, PartOfSpeech gives the grammatical category of the lexical entry, in the same way for the feature MAJ in TDL. PartOfSpeech and MAJ represent the same morphological information. If PartOfSpeech possesses the value verb, so MAJ takes it also. We define also specific rules for the syntactic features transformation. The rule of figure 11 illustrates this type of rules. Rule 2 states that for a given lexical entry, the id (identifier or name) of a syntactic behaviour is translated in TDL by forcing the TDL AVM of the lexical entry to inherit from a TDL type of the same name. This TDL type is just the offline constructed equivalent of LMF SyntacticBehavior Component. The overall transformation system is based upon the specifications of 32 rules[2].

JALASA_verb:= Lex-verbe &
[PHON $\langle "jalasa"\rangle$,
$SS[LOC[CAT[HEAD[VFORM complet,$
$RADICAL\ trilitere,$
$MOJARRAD\ +,$
$ROOT\ jls,$
$PATTERN\ faala,$
$ASPECT\ accomplished,$
$PERSON\ 3]]]]$.

Figure 6: TDL specification of the verb jalasa (to sit)

JALASA_verb:= Lex-verbe & Intransitive_verb &
[PHON $\langle "jalasa"\rangle$,
$SS[LOC[CAT[HEAD[VFORM complet,$
$RADICAL\ trilitere,$
$MOJARRAD\ +,$
$ROOT\ jls,$
$PATTERN\ faala,$
$ASPECT\ accomplished,$
$PERSON\ 3]]]]$.

Figure 7: TDL specification with syntax inherited from Intransitive_verb

3. Evaluation and discussion

In order to verify the correctness and the usefulness of the transformation tool, the semi-automatically created lexicon of Arabic verbs is transformed to a TDL lexicon. The generated lexicon is evaluated qualitatively by the LKB system: A parsing system according to the HPSG theory and TDL as a representation language for lexical input. (Boukedi et al.,) proposed a type hierarchy and a minimal HPSG grammar for Arabic running on the LKB system. We use this grammar to compare a manually sample edited HPSG lexicon with its equivalent automatically generated one.

To test our HPSG grammar, we used a corpus of 205 transliterated sentences. This corpus was created from a lexicon of 781 words. It covers various structures of nominal and verbal sentences. Results are presented in table 1. In the original test, 85% of sentences were analyzed correctly. The failure cases (0 analysis) are due to the absence of rules treating some particular syntactic phenomena (i.e.,

[2]The number of rules can be extended to deal with special cases of lexical entries.

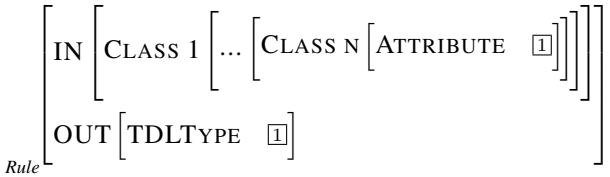

Figure 8: Type transformation rule

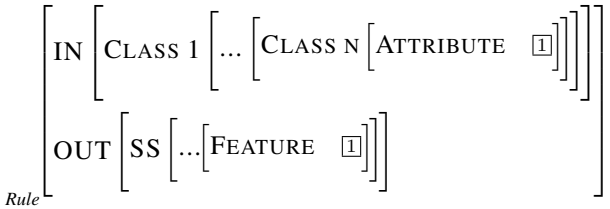

Figure 9: Lexical entry transformation rule

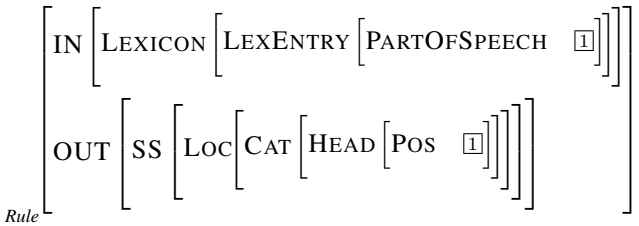

Figure 10: Rule example 1

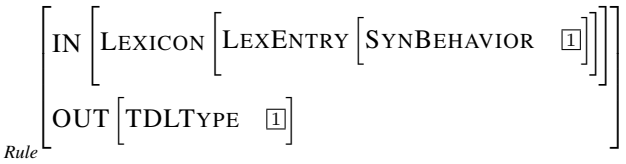

Figure 11: Rule example 2

relative phenomenon, coordination phenomenon). The ambiguous cases (2 analysis) are due to a no precise specification of the constraints specification of some syntactic rules. The same experiment with the new generated lexicon shows a slight progress in parsing results due to the introduction of syntactic information. The quality of the generated lexicon is thus approved to be as useful as manually created lexicons.

Derivation Trees	Analysed sentences (manual lexicon)	Analysed sentences (generated lexicon)
0	25	17
1	175	190
2	5	8
	205	205

Table 1: Parsing results

The assessment in terms of acceptability and parsing results has shown a comparable quality in TDL entries. The results are very encouraging. In fact with a system of 32 rules, we can transform the entire lexical information in the LMF lexicon to its TDL correspondent. Tests with the LKB system do not seem to show parsing improvements. But we think this is because the HPSG grammar does not contain specifications that encompass sufficient syntactic phenomena. Changing the rules without touching the source code can easily modify the transformation system. Such flexibility is obtained thanks to the modular design of the system.

4. Conclusion and perspectives

In this paper, we proposed an approach for modeling a normalized lexicon of Arabic verbs based on LMF and a transformation system enabling the generation of a TDL lexicon from an LMF representation. A system of rules has been designed following a study depended of the two formalisms. The system will be a very useful tool for different applications like the Arabic syntactic analysis which constitutes an essential stage in the linguistic processing. Many improvements are planned like extending the rule system to cope with semantic information and enhancing the lexicon with more syntactic information like diathesis and morphology-syntax interaction.

5. References

S. Boukedi, N. Loukil, K. Haddar, and A. Benhamadou. The experimentation of a hpsg grammar for the arabic language on the lkb system. In *Proceedings of the 3rd International Conference on Arabic Language Processing*.

G. Francopoulo. 2005. Lexical markup framework (lmf = iso 24-613). Technical report, INRIA Loria.

H-U. Krieger and U. Schafer. 1994. Tdl: A type description language for hpsg. part 2: User guide. Technical report, German's research center for artificial intelligence.

N. Loukil, K. Haddar, and A. Benhamadou. Normalisation de la reprsentation des lexiques syntaxiques arabes pour les formalismes d'unification. In *Proceedings of the Lexicon Grammar Conference*.

N. Loukil, K. Haddar, and A. Benhamadou. Towards a syntactic lexicon of arabic verbs. In *Arabic Local Languages workshop- 6th Language Resources and Evaluation Conference*.