Natural Language Engineering **22** (6): 849–879. © Cambridge University Press 2015
doi:10.1017/S1351324915000224

ISO standard modeling of a large Arabic dictionary

AIDA KHEMAKHEM[1], BILEL GARGOURI[1], ABDELMAJID BEN HAMADOU[2] and GIL FRANCOPOULO[3]

[1]*MIRACL Laboratory, FSEGS, University of Sfax, B.P. 1088, 3018 Sfax, Tunisia*
e-mail: khemakhem.aida@gmail.com, bilel.gargouri@fsegs.rnu.tn
[2]*MIRACL Laboratory, ISIMS, University of Sfax, B.P. 242, 3021 Sakiet-Ezzit, Sfax, Tunisia*
e-mail: abdelmajid.benhamadou@isimsf.rnu.tn
[3]*IMMI-CNRS and Tagmatica, Rue John von Neumann, 91405 Orsay, France*
e-mail: gil.francopoulo@wanadoo.fr

(*Received 6 May 14; revised 24 June 15; accepted 24 June 15; first published online 7 September 2015*)

Abstract

In this paper, we address the problem of the large coverage dictionaries of Arabic language usable both for direct human reading and automatic Natural Language Processing. For these purposes, we propose a normalized and implemented modeling, based on Lexical Markup Framework (LMF-ISO 24613) and Data Registry Category (DCR-ISO 12620), which allows a stable and well-defined interoperability of lexical resources through a unification of the linguistic concepts. Starting from the features of the Arabic language, and due to the fact that a large range of details and refinements need to be described specifically for Arabic, we follow a finely structuring strategy. Besides its richness in morphology, syntax and semantics knowledge, our model includes all the Arabic morphological patterns to generate the inflected forms from a given lemma and highlights the syntactic–semantic relations. In addition, an appropriate codification has been designed for the management of all types of relationships among lexical entries and their related knowledge. According to this model, a dictionary named El Madar[1] has been built and is now publicly available on line. The data are managed by a user-friendly Web-based lexicographical workstation. This work has not been done in isolation, but is the result of a collaborative effort by an international team mainly within the ISO network during a period of eight years.

1 Introduction

All natural languages have numerous lexical resources, which are quite useful both for human users and Natural Language Processing (NLP). However, their construction and maintenance are very expensive. Some resources are limited to the specific needs of their applications using various structures without any unified

[1] http://elmadar.miracl-apps.com/.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

linguistic concepts. The reuse of such resources is difficult when the challenge is to merge them in order to obtain a new richer resource. More generally, the exchange remains a difficult (and expensive) issue when nothing has been scheduled for this purpose. To meet this challenge, several projects were conducted such as ACQUILEX (Bogurev *et al.* 1988), GENELEX (Antoni-Lay, Francopoulo and Zaysser 1994), TEI (Véronis and Ide 1996), Eagles (Calzolari, Mc Naught and Zampolli 1996), Multext (Calzolari and Monachini 1996), SIMPLE (Lenci *et al.* 2000) and ISLE (Bertagna *et al.* 2000).[2] These projects led to the emergence of the LMF (Lexical Markup Framework) ISO standard for the lexical structure modeling (ISO 24613) (Francopoulo 2003), (Francopoulo and George 2008) in association with the ISO Data Categories Registry (DCR)[3] following ISO 12620 (Ide and Romary 2004). These standards were designed by a group of sixty ISO experts coming from different cultures, languages and continents. Numerous developments followed in different parts of the world.[4]

Unfortunately, the Arabic language did not immediately benefit from the emergence of these standards, although it is spoken by more than 300 million people around the world, and is the official language of more than twenty countries. The language still uses references to different printed dictionaries based on incompatible lexicographical schools. Only few works tried the application of LMF on the Arabic language out, according to previous revisions of this standard. Some developments were made in morphology (Khemakhem, Gargouri and Abdelwahed 2006), (Romary, Salmon-Alt and Francopoulo 2004), (Salmon-Alt, Akrout and Romary 2005) and some studies were conducted in syntax (Loukil, Haddar and Ben Hamadou 2008). However, these works were developed during the drafting of the LMF standard and were not updated according to the ISO validation.

Obviously, the situation of the Arabic lexical resources is below expectation. In case of human use, for instance, most of the existing electronic dictionaries are digitized from a printed version and are not well structured to facilitate access and browsing (Taleb 2005). This situation is largely due to a lack of modeling efforts to face the linguistic richness and the evolution of the Arabic language. In fact, apart from the common data categories (i.e. verb, noun, particle), the Arabic language uses specific subcategories (e.g. deficient verb, common noun, masdar) involving various and numerous kinds of morphological, syntactic and semantic markers. Besides the lexical meanings and connotations, verbs, nouns and particles have syntactic behaviors. And since Arabic is a derivational and inflectional language, Arabic has specific issues related to the links between the original forms and the generated ones, what inevitably influences the organization of the dictionary. For the derivational aspect, the question is to know which original form (i.e. root, lemma) should be considered in order to derive a new lemma and, associated to this rationale, whether the root has (or not) to be considered as an entry. Some irregular forms of nominal

[2] A more comprehensive history of NLP lexicons may be found in (Calzolari, Monachini and Soria 2013).
[3] www.isocat.org.
[4] An extensive list of LMF lexicons is presented in (Francopoulo 2013).

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

inflection (i.e. broken plural) suggest that the inflected forms be entries. In addition, the evolution of the Arabic language forces the lexicographer to introduce new entries that do not respect the classical derivational process based on a root form. It should be noted that the rather popular Arabic edition of Wiktionary[5] shares our concerns. However, it deals with a superficial structure of resources. Its syntactic model is very light and does not link synonyms to related meanings. Moreover, the relations between meanings and syntactic knowledge are not covered.

In the NLP context, several models have been developed without consensus and are, in general, limited to specific needs while using no common data categories. Therefore, the reuse of most Arabic electronic dictionaries is difficult and their models are incomplete. For example, DIINAR (Abbès, Dichy and Hassoun 2004) covers only the morphological level; Arabic WordNet (Elkateb *et al.* 2006) focuses on the semantic level and some lingware systems use their appropriate lexical resources such as MORPH2 (Chaâben, Hadrich and Ben Hamadou 2010).

From a lexicographic standpoint, the construction of Arabic electronic dictionaries is confronted with some difficulties observed in the two conventional steps (Ben Mrad 1987): (1) the theoretical study of the content and the organization of the lexical entries (LEs) and (2) the practical step of the population. Given the fact that the population step depends on the quality of the theoretical study, we focused on the difficulties related to this latter. We suggested a solution to the problems mentioned above and related to the modeling of a large Arabic lexical resource. Such a solution was performed during the framework of the validating process of the LMF (ISO 24613) (Khemakhem *et al.* 2006), (Khemakhem, Elleuch, Gargouri and Ben Hamadou 2007) and through the population of the DCR (ISO 12620), specifically for Semitic languages. The proposed model now covers all linguistic levels (i.e. morphological, syntactic and semantic) with respect to the specificities of Arabic language (Khemakhem, Gargouri and Ben Hamadou 2012) with the convention that, in the current text, we use the Habash–Soudi–Buckwalter transliteration scheme (Habash, Soudi and Buckwalter 2007). Our model contains the required knowledge for human users and for the NLP in an explicit way while remaining open to extension. The data is associated with a running software tool in order to search, conjugate and collaboratively modify.

The remaining of this paper is organized as follows. We start with the presentation of the Arabic lexicography and give an overview on the related electronic dictionaries. Then, we describe the main ISO standards useful for our work and we specify the standardization of some data categories which are appropriate for the Arabic language. After that, we detail the modeling of the dictionary covering the core and then, the morphological, syntactic, semantic sections. Afterwards, we comment on the implementation of the dictionary and present the systems ensuring its query and practical population. Thereafter, we provide a discussion based on a comparison between the proposed model and those related to some well-known dictionaries.

[5] en.wiktionary.org/wiki/Arabic.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

2 Arabic lexicography

According to Darwish (1956) and Haywood (1960), the lexicography of Arabic dictionaries focuses on two main issues that need to be solved. The first issue is the determination of the nature of LEs comprising roots, derived forms and foreign words. This is the macrostructure. The second issue is the internal description of each LE, which includes speech, phonological, morphological, syntactic and semantic information. This is the microstructure and of course, this distinction is not specific to Arabic lexicons but concerns also English lexicons (Landau 2001) and French ones (Rey-Debove 1971), for instance.

2.1 Macrostructure

Arabic is a derivational language based on roots that may generate derived forms, considered as lemmas of this language. Each root is a sequence of generally three or four letters, but not considered as a word. The description of entries depends mainly on the nature (root or derived form) of the LE. The words that are not of Arab origin can be either integrated into the language as an Arabic word (Almuʿar∼ab, المُعَرَّبْ) that easily supports the application of the derivational aspect and some inflection like دينار diynaAr, or added using their transliteration (aldaxil, الدَّخِيلْ), like computer كَمْبْيُوتَرْ kambyuwtar and Internet أَنْتَرْنَاتْ ‘antarnaAt. The old dictionaries (i.e. until the end of the nineteenth century) used the root as a LE that groups all its derived forms (lemmas). Such arrangement is supported by the first lexicographic school called ‘the radical school’ (المدرسة الجذرية Almadrasaħu Alji∂riyaħu). The retrieval process in such dictionaries is difficult: this process is based on human intuition and education, in particular his ability to know in advance the root of the words. In order to overcome this difficulty, many other kinds of arrangements related to other lexicographic schools were proposed, such as the phonetic school (المدرسة الصوتية Almadrasaħu AlSawtiyaħu), the terminal school (المدرسة القافية Almadrasaħu AlqaAfiyaħu) and the alphabetic school (المدرسة الهجائية النطقية Almadrasaħu AlhijaA’iyaħu AlnuTqiy∼aħu). The most widespread lexicographic school is ‘the alphabetic school’. It considers each lemma as a LE. All these entries are presented in the alphabetical order of letters (أ، ب، ت، ث،...، ي) (Alouani 2008).

2.2 Microstructure

The description of entries depends mainly on the intended use of the dictionary. In old dictionaries, entries have non-regular structures and their contents are varied and based on contexts and examples. However, there are studies that specify the main components of LEs such as the study of Ben Mrad (1987) that defines three parts to describe an LE:

- Part of speech: noun (اسم ‘ism), verb (فعل fiʿl), adjective (صفة Sifħ), adverb (ظرف Ďarf), particle (أداة ‘adaAħ),
- Lexical morphology: gender of noun with their inflected forms,
- Lexical semantic: definition and context.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Nowadays, electronic versions favor the construction of more detailed and more explicit microstructures than those of printed versions. The organizational problem does not arise within the electronic representation, as it was the case for the printed versions. The explicit representation allows seeking knowledge using simple query languages without knowing the organization of the dictionary. It is rather the microstructure which is the most important issue concerning the development of dictionaries notably with the continuously evolving content. Thus, new requirements in dictionaries' contents appeared with the electronic version such as conjugation, semantic relations (i.e. synonym, antonym), syntactic behavior and multimedia knowledge (i.e. sounds, videos, pictures).

3 Overview of electronic Arabic dictionaries

3.1 Dictionaries for human use

Electronic Arabic dictionaries dedicated to human users are extremely rich in linguistic, lexicographical and cultural information. However, they are generally poorly structured: this is a difficulty for the computer operations which require cleanly coded information. The search tools are unable to extract the best information from the current organization of these dictionaries. Both human experience and intuition are required in such cases. The situation of these dictionaries makes their reuse a costly and laborious process. Moreover, the internal content of LEs is borrowed without any adaptation to the technological innovation. Examples of these dictionaries are the Sakhr[6] computerization of the old dictionaries 'lisan Alarab' لسان العرب, 'Alqamous AlmuwhiT' القاموس المحيط 'AlwasiyT' الوسيط, 'Elganiy∼i' الغنيّ. Moreover, there are some bilingual dictionaries (e.g. Ajeeb, Larousse and Alexandria) that result from a simple translation of words to foreign languages. Other works as Dutch-Arabic bilingual dictionary (Maks, Tiberius and Van Veenendaal 2008) are based on LMF. These dictionaries are incomplete and very poor in linguistic knowledge.

3.2 Dictionaries for NLP use

Most of the existing lexical resources deal with a unique linguistic level and are generally adapted to the specific needs of the intended applications. Examples of lexicons covering the morphosyntactic level are the Buckwalter lexicon (Buckwalter 2004), Morph2 system lexicon (Chaâben *et al.* 2010), ElixirFM lexicon (Smrž 2007), DIINAR lexical database (Abbès *et al.* 2004), Nooj Arabic Dictionary (Mesfar and Silberztein, 2008), AraComLex lexicon (Attia *et al.* 2011), AraLex (Boudelaa and Marslen-Wilson 2010), Unitex lexicon (Doumi *et al.* 2013) and SALMA lexicon (Sawalha, Atwell and Abushariah 2013). For the semantic level, the Arabic WordNet (Elkateb *et al.* 2006) seems to be the most used one. The models and structures of such resources are not extendable and they are difficult to reuse in a different

[6] lexicons.sakhr.com.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

lingware environment. For example, the DIINAR does not represent the lemmas, but it generates them from roots, using a set of rules. Doing so, the meanings cannot be associated to the lemma (absent in this case). Some resources, like Arabic WordNet, focus exclusively on the semantic relations and are of limited linguistic coverage. It is worth noting that some resources are dedicated to Arabic dialects, used as an informal means of communication (Diab *et al.* 2010), (Maamouri *et al.* 2006).

4 ISO standards related to lexical resources

There are two main standards for the modeling of lexical resources, that work together. The first is LMF (ISO 24613), which proposes a consensus for the structure (or skeleton) of the data. The second is the DCR (ISO 12620), which describes the linguistic concepts by means of data categories. The latter is used to enrich the structures defined by means of the former (i.e. LMF).

4.1 Lexical markup framework: ISO 24613

LMF is a meta-model that provides a common standardized framework for the construction of NLP lexicons and Machine Readable Dictionaries (MRD). The specification is an ISO standard defined in Unified Modeling Language (UML), whose aim is to manage data exchange, and enable the merging of a large number of individual electronic sources from extensive global electronic resources. The meta-model is organized as different extensions and an obligatory core package. It should be adorned by data categories that are selected from the DCR.

4.2 Data category registry: ISO 12620

Data Categories represent basic linguistic concepts (e.g. /part of speech/, /grammatical number/) and describe the specific linguistic properties of a language, independently of a specific model. In order to share data categories within the community, the ISO TC 37/SC 4 developed the online DCR, especially used simultaneously with the other standardization activities. The DCR specifies the names and definitions for data categories relevant to the language resource domain as well as the management procedures for a DCR for Language Resources.

4.3 Modeling process

The LMF standardization process for lexical resources construction starts with the selection of the necessary classes to represent the lexical resource from the model that includes a core and a set of extensions such as morphology, morphological pattern, syntactic, semantic and MRD. The presence of the core is mandatory according to LMF. The selected classes are then enriched with data categories that are selected from the DCR. In addition, we can define a set of new data categories that are needed but unavailable in the DCR. This supplemental set of data categories should be registered with the DCR Registration Authority and managed according to ISO 12620.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

5 Specific lexical concepts of Arabic

Most of the Arabic words are characterized by a root and a scheme, which are respectively normalized under the permanent identifiers number 2231 and 2174 in the DCR. These words refer to three categories: noun, verb and particle. Each category has many specific subcategories in order to improve the morphosyntactic analysis. For example, masdar is a subcategory of the noun that can replace the verb in a nominal sentence. This specification is supported by the standard that uses subcategories as well as main categories as part-of-speech (POS). It links the main category and their subcategories by the 'is a' relation. However, there are some values that are necessary for the Arabic language, but that are not registered in the DCR, so we have added several POS values. The specific nominal subcategories are presented in the following Table 1.

Table 1. *The specific subcategories of Arabic nouns*

Perm. identifier	POS	Translation	Transliteration	ISA relation
2221	Affixed Personal Pronoun	ضَمِيرْ مُتَّصِلْ	Damiyr mut∼aSil	pronoun
2225	Diminutive Noun	اِسْمْ التَّصْغِيرْ	Ăsm Alt∼aSγiyr	noun
2220	elative	اِسْمْ تَفْضِيلْ	Ăsm tafDiyl	noun
	intensive	اِسْمْ مُبَالَغَةْ	Ăsm mubaAlaγaħ	noun
	Instrument Noun	اِسْمْ آلَةْ	Ăsm Ālaħ	noun
	Manner Noun	اِسْمْ هَيْأَةْ	Ăsm hayÂaħ	noun
2224	masdar	مَصْدَرْ	maSdar	noun
	Number Noun	اِسْمْ عَدَدْ	Ăsm ςdad	noun
	Once Noun	اِسْمْ مَرَّةْ	Ăsm mar∼aħ	noun
	Place Noun	اِسْمْ مَكَانْ	Ăsm makaAn	noun
2226	Relation Noun	اِسْمْ نِّسْبَةْ	Ăsm nisbaħ	noun
	Time Noun	اِسْمْ زَمَانْ	Ăsm zamaAn	noun
2253	Voice Noun	اِسْمْ الصَوْت	Ăsm AlSawt	noun

Among the characteristics of nouns, there is elInclusion (2219 in the DCR) indicating if the noun may be associated with the article Al 'ال'. The inflection of some nouns stands by a morphological feature: broken plural (2218 in the DCR) that does not have any inflection. The specific subcategory of the Arabic verbs is deficient verb (3007 in the DCR), knowing that these verbs have two subcategories: plain and deficient. Among the characteristics of verbs, we indicate the aspect and mode that are requested in the conjugation. The aspect may take these values: accomplished (cessative), unaccomplished and imperative; the first two values are respectively normalized under numbers 2001 and 2217 in the DCR. The unaccomplished may have the indicative, subjunctive and apocope[7] moods. The

[7] Deletion of a final element in a word.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

latter is normalized under number 2254 in the DCR. The specific subcategories of the Arabic particle are presented in the following Table 2.

Table 2. *The specific subcategories of Arabic particles*

Perm. identifier	POS	Translation	Transliteration	Isa
2227	coordination particle	حَرْفُ الجَرُّ	Harfu Aljar~	particle
2228	distinctive particle	حَرْفُ التَمَيِّز	Harfu Altamy~iz	particle
2229	relative particle	حَرْفُ المَوْصُولْ	Harfu AlmawSuwl	particle
	conditional particle	حَرْفُ الشَّرْطْ	Harfu Alš~arT	particle

6 Core modeling

The LMF core model is the mandatory part of a normalized resource. This core contains two types of information: general and common. The general information describes the entire lexicon of the language, the version, the creation date and the authors. The common information is needed at several linguistic levels such as *LE*[8], *Form* and *Sense*. The LE has a very important role in organizing the various portions of the lexical resource and can serve as a connector that links all related knowledge.

6.1 Main directives

We believe that the core model of the Arabic language should be structured by its derivational aspect, which is based on roots and schemas. A large part of Arabic words are derived forms obtained by using a generation process. This generation process consists in replacing the three consonants of the schema (ل ع ف [f ç l]) [f ç l]) by the consonants of the concerned root, while preserving the same order of letters. The lemma is obtained by replacing the consonants of the schema ف[f]ع [ç] ل[l] respectively with the consonants of the root ك [k], ت [t], ب [b], while maintaining the other components of the schema (م, ة) for the second example. We can classify the schemas into two types: the verbal schemas and the nominal schemas. Thus, from a root, we can generate nouns and verbs according to the used schema. In Figure 1, the verb كَتَبَ [kataba] is obtained using the verbal schema فَعَلَ [façala], whereas the noun مَكْتَبَةٌ [maktabaħũ] is obtained by applying the nominal schema مَفْعَلَةٌ [mafçalaħũ]. Thus, the schema is considered as a mold on which the root flows, as shown in the following figure.

The derivational aspect has influenced Arabic lexicographers in the choice of the type (root or derived form) of the LE for their dictionary. The resources for human use are organized according to the roots while including all their derived forms or according to the derived forms, which represent their roots as complementary

[8] Following UML conventions, the class names are mentioned in uppercase letters. To avoid any confusion with other names, the UML attribute and class names are written in italic within our text.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

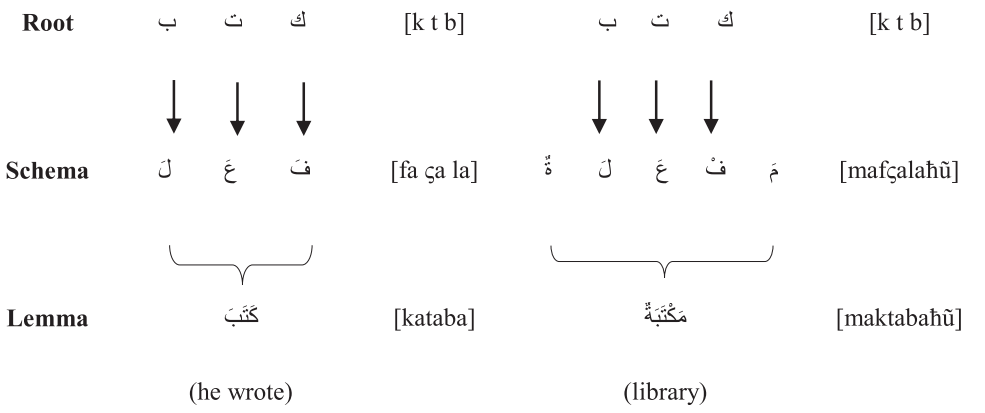


Fig. 1. Example of derivation from the root.

Table 3. *Examples of morphological information generated*

Lemma	Available Information		Generated Information	
	Root	Scheme	Written in arabic	Translation
كَتَبَ kataba	ك ت ب k t b	فَعَلَ façala	مُجَرَّد صَحِيح mujar~ad SaHiyH	unaugmented intact
أَكَلَ ‘akala	أ، ك ل k l	فَعَلَ façala	مُجَرَّد صَحِيح مَهْمُوز الفَاء mujar~ad SaHiyH mahmuwz AlfaA’	unaugmented intact with initially-Hamzated
وَزَنَ wazana	و ز ن w z n	فَعَلَ façala	مُجَرَّد مُعْتَلّ الفَاء mujar~ad muçtal AlfaA’	unaugmented with intially - defective consonant (by waw)
اِسْتَخْرَجَ ‘istaxraja	خ ر ج x r j	اِسْتَفْعَلَ ‘istafçala	مَزِيد صَحِيح maziyd SaHiyH	augmented intact
اِسْتَقَامَ ‘istaqaAma	ق و م q w m	اِسْتَفْعَلَ istafçala	مَزِيد مُعْتَلّ العَيْن maziyd muçtal Alçayn	augmented with medially- defective consonant (by waw)

information. For NLP use, many works such as DIINAR (Abbès *et al.* 2004), Buckwalter lexicon (Buckwalter 2004), ElixirFM (Smrž 2007) and SALMA (Sawalha and Atwell, 2013) are based on derivational rules for generating morphosyntactic knowledge including the derived forms. In this case, the population of such resources by other levels, particularly the semantic level, hit serious problems because the words are not persistent. The rules used can generate forms, but cannot generate their meanings with their syntactic particularities. For this reason, we have chosen the extensional representation for the derivational aspect. Accordingly, we presented the derived forms and the roots as LEs and we used the morphological relations to ensure the link between these two types of entries. In this way, we can satisfy the different lexicographical schools. In addition, for each derived form, we identify its schema and the relation with its root (see Section 7.1). The schema and the root used for generating a derived form can indicate several other morphological, syntactic and even semantic informations (see Table 3) and avoid their presence in the dictionary as unaugmented or augmented, intact or initially-defective or medially-defective knowledge.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

It shall be noted that a second level of derivation process can be applied on a word (and not on a root) in order to obtain another derived word (see Section 7.1).

6.2 Core model for Arabic

The following Figure 2 specifies the classes and the main data categories of the core model.

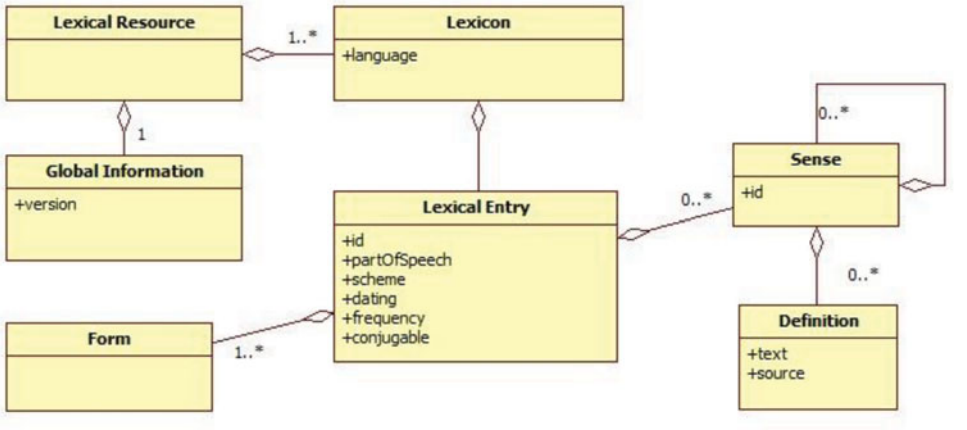


Fig. 2. (Colour online) Core model diagram for Arabic.

The core model is composed of the following classes:

- *Lexical Resource*: It is made up of one or more *Lexicons* and *Global Information* instances.
- *Global Information*: It includes all the general pieces of information of a lexical resource such as title, version and creation date.
- *Lexicon*: It is made up of one or more *LE* instances and characterized by a language. It has a very important role, because it can contain components that are shared by several LEs as *Morphological Pattern* and *Subcategorization Frame* instances.
- *Lexical Entry*: It is the most important class that represents the words of a language and consists of one or more *Form* instances. It can represent roots (i.e. كتب /k t b/, لعب /l ‘ b/), derived forms (i.e. كَاتِبْ /kâtib/ ‘writer’), letters (i.e. ء , ب, ت), invariable words (i.e. كَمْ /kam/ ‘how’, وَ /wa/ ‘and’) or words of the non-Arabic origin (i.e. كُمْبِيُوتَرْ /kambyûtar/ ‘computer’). An LE can bring together several meanings or it may not indicate any meaning as in the case of roots. For each LE, we assigned an identifier.

If an LE is a root, then the Data Category (DC) type should be set as ‘root’. Else, the LE should hold multi-valued POS information. These values are the grammatical categories (for the Arabic language: noun, verb and particle) or the subcategories (e.g. for nouns: common noun, masdar). In addition, each category of words has a set of specific DCs as follows:

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

- Verbs: schema, etymology, etc.
- Nouns: schema, etymology, etc.
- Particles: etymology, etc.
- Letters: nature (defective/saint), characteristic (aloud/whisper), phonetic output, etc.
- *Sense*: This class can have several definitions and several sub-meanings. Each meaning must have an identifier that allows us to ensure the semantic links between them. Besides, it can be characterized by the DC: semanticClass and nature (metaphorical/literal).
- *Definition*: It contains a text and its source. We note that, according to LMF, each meaning may have more than one definition.
- *Form*: It is an abstract class representing a lexeme or a morphological variant. It has subclasses, such as *Lemma* and *WordForm,* which belong to the morphological extension.

7 Morphological modeling

The morphology of an Arabic word is characterized by the derivational and the inflectional aspects. The description of these aspects is given first, in order to argue the morphological model.

7.1 The derivational aspect: complementary description

It should be recalled that the derivational aspect is based on a root and a schema to give a derived form. This can be the basis of the derivation, that is to say, a schema can also be applied on a word (already derived form) to generate a new derived form. So we have two levels of derivation: the first is based on a root, and the second is based on a derived form. For example, as illustrated in Figure 3 مَخْرَجٌ - خَرَجَ - خ ر ج [x r j – xaraja – maxrajũ] belongs to the same derivational chain, which is based on the root 'ج خ ر' [x r j]. This root is combined with the schema 'فَعَلَ' [façala] to give us a derived form «خَرَجَ»[xaraja]. This can be the basis of a derivation, that is to say, it can be combined with a schema «مَفْعَلٌ»[mafçalũ] to give us a new derived form «مَخْرَجٌ»[maxrajũ]. The LE «خَرَجَ»[xaraja] must have a link to ensure the connection with its root «خ ر ج»[x r j], and another link (i.e. a stem link) to ensure the connection with its derived form «مَخْرَجٌ»[maxrajũ].

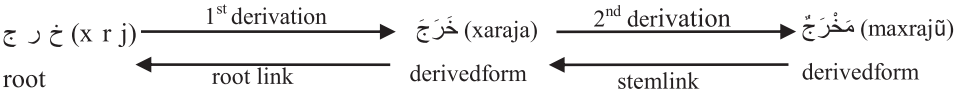


Fig. 3. Derivation.

7.2 The inflectional aspect

For the inflectional aspect, we adopt a representation that includes the minimum lexical information ensuring maximal coverage. The representation is mainly

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

descripted in intension[9] by means of symbolic features and not in extension by listing all inflectional forms. The intensional representation is applied to regular inflections of verbs and nouns (i.e. the dual, the regular feminine plural, see Section 8). The extensional representation is used for irregular inflections, especially the nouns that are broken plural. For the nouns which have an irregular inflection, we save all their inflectional forms with their morphological features. The presence of this information in a dictionary improves the results of the analyzers and provides human users with precise information about the conjugation of Arabic words. The morphological features that characterize the Arabic verbs and nouns are presented in the following table.

7.3 The Arabic morphological model

The LMF standard offers a morphological extension for the representation of the morphological knowledge of a LE. In the following Figure 4, we present the classes and the main DCs of the Arabic morphological extension.

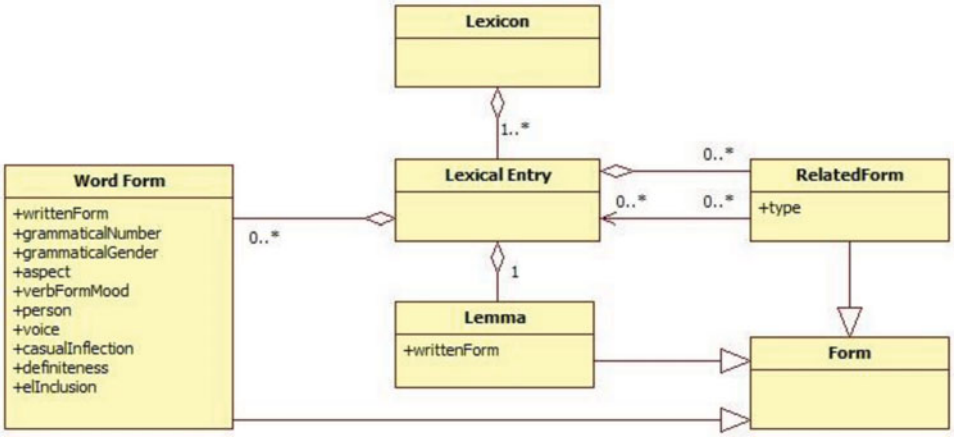


Fig. 4. (Colour online) The Arabic morphological model diagram.

The model of the morphology extension is composed of three classes:

- *Lemma*: This class is required for each LE. It represents the canonical form of a word into the written form DC. If the word is a noun, its canonical form must be singular. If the word is a verb, its canonical form must be accomplished with the third person singular. To maximize the coverage of linguistic information, we opted for the representation of vowelled words. We note that it is unnecessary to submit the unvowelled form of a word since it can be generated. However, we removed the last vowel of the names that

[9] The term 'intension' should not be confused with the term 'intention' which is a more frequent one. According to the Concise Oxford English Dictionary 2011, 'intension' is defined as 'internal content of a concept. Often contrasted with extension'. In contrast, 'intention' is defined as 'conceptions formed by directing the mind towards an object'.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Table 4. *The morphological features for Arabic verbs and nouns*

Verb		Noun	
Feature	Values	Feature	Values
Grammatical gender	Masculine/feminine	Grammatical gender	Masculine/feminine
Number	Singular/dual/plural	Number	Singular/dual/plurial/broken plural
Person	First/second/third	Casual inflection: اللإعراب	Nominative (مَرفوع)/genitive (مجرور)/accusative (مَنصوب)
Aspect: الصيغة	Accomplished (المَاضِي)/unaccomplished (المُضَارَع)/imperative(الأمْر)	Definiteness	Definite/indefinite
Mood	Indicative (المَرفوع)/subjonctive (المَنصوب)/apocope (المَجزوم)	elInclusion	Yes/no
Voice	Active (المعلوم)/passive (المجهول)		

have no importance to the morphological level because its presence creates an explosion of syntactic base. This vowel varies according to the position of the name within the sentence.

- *Word Form*: It presents an inflection form of the LE in its DC written Form. The other DCs vary with the category of the LE. For example, following DCs are used for nouns: grammatical number, grammatical gender, casual inflection, definiteness, elInclusion (see Table 4).
- *Related Form*: Its aim is the representation of the derivational aspect. The DC type allows the specification of the derivation links between the different LEs.

In the following figure, we applied our model on a part of the example presented earlier. It should be noted for the user who is not familiar with UML, that this is an object diagram, in other terms, a diagram for an example, with instances instead of classes (with the class name being underlined, following UML conventions). In contrast, the previous UML diagrams were model diagrams, in other terms, general diagrams.

Figure 5 gives an UML object diagram of two LEs: 'خ ر ج' and 'خَرَجَ' which are connected with two *Related Form* objects. The left object belongs to the entry 'خَرَجَ' in order to link its root, which is another entry. The latter is 'خ ر ج' (a root), which has the second object *Related Form* to connect one of its derived forms. This example is represented in Figure 6 with the identifiers used in the binding being graphically shaded.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

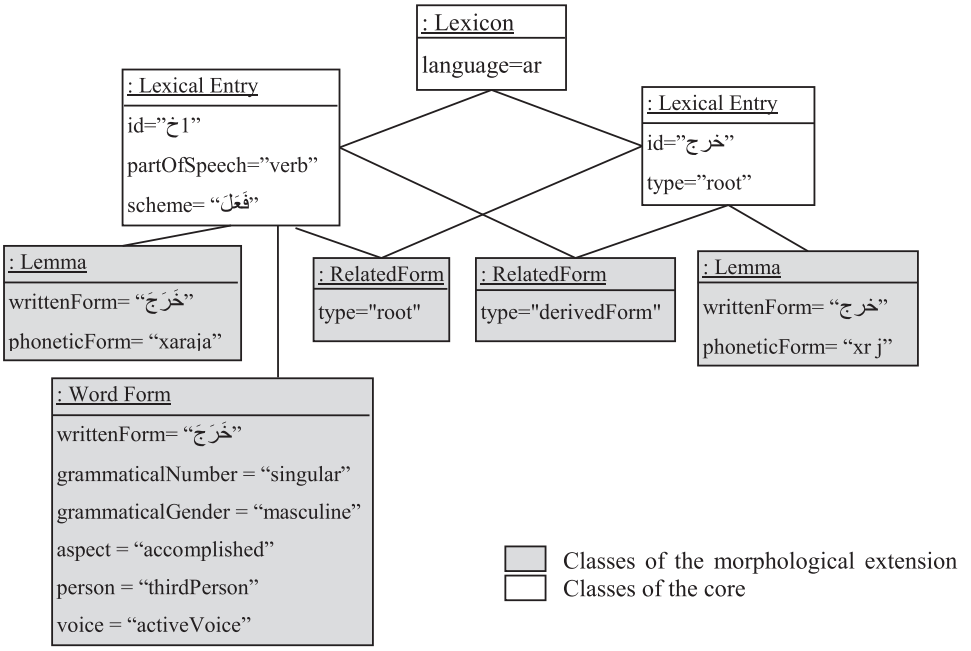


Fig. 5. The object diagram for two entries 'خ ر ج' and 'خَرَجَ'.

8 Morphological pattern modeling

A deep study of the conjugation process facilitates the generation of the inflected forms and promotes the lemmatization process while providing interesting morphological features. In the Arabic language, both nouns and verbs are inflected. Nouns present many irregular forms, whereas the regular ones use phonological rules. However, the verbs inflection is more complex and concerns about 16,000 verbs. In this section, we focus on the Arabic verbs inflection.

8.1 The inflection of Arabic verbs

An Arabic verb is conjugated in two steps:

- The first step is a rule application to get the basis form depending on aspect, mood and voice.
- The second step is the addition of an affix, which depends on gender, number and person.

The following Figure 7 illustrates the conjugation process for the verb 'كَتَبَ'.

The paradigm of conjugation of a verb depends on its root (the nature and the number of consonants) and its schema. We identified forty-two classes of roots and twenty-two classes of schemas. The combination of these two kinds of classes led, after refinement, to 256 paradigms (Khemakhem *et al.* 2007) representing an exhaustive grid of Arabic verbs inflection.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

```
<Lexicon>
<LexicalEntry id="خرج">
…
<Lemma><featatt="writtenForm" val="خ ر ج" />
</Lemma>
<RelatedForm targets="خ1">
<feat att="type" val="derivedForm" />
</RelatedForm>
</LexicalEntry>
<LexicalEntry id="خ2">
…
<Lemma>
<feat att="writtenForm" val="مَخْرَجٌ" />
</Lemma>
<RelatedForm targets="خ1">
<feat att="type" val="stem" />
</RelatedForm>
</LexicalEntry>
```

```
<LexicalEntry id="خ1">
…
<Lemma>
<featatt="writtenForm" val="خَرَجَ" />
</Lemma>
<RelatedForm targets="خرج">
<feat att="type" val="root" />
</RelatedForm>

<RelatedForm targets="خ2">
<feat att="type" val="derivedForm" />
</RelatedForm>
</LexicalEntry>
….
</Lexicon>
```

Fig. 6. XML extract related to the morphological relationship of three entries.

Inflected Form		Suffix	Basis Form	Canonical Form
كَتَبْتُ	←	+تُ	كَتَبْ	كَتَبَ
Katabtu		tu	Katab	Kataba

Fig. 7. An example of calculating the inflected form for the verb 'كَتَبَ'.

8.2 The model of morphological pattern

The LMF standard provides a powerful representation of the morphological patterns, which allows the distinction between the conjugation rule and the addition of an affix to calculate an inflected form. Being designed for all language morphologies, the system is rather complex and due to space limitations, only a short outline is presented here. The following Figure 8 gives the model of the morphological pattern showing the main DCs.

The model of the morphological pattern extension consists of six classes:

- *Morphological Pattern*: It is characterized by an identifier. It has several ordained affixes and zero or more *Transform Set* instances. We will use this class to present an inflectional paradigm bending that generates all the inflected forms of a word. It must be referenced by one or more LEs.
- *Affix*: It presents the affixes used by a pattern. It has the DC written Form and type that can take the values: prefix and suffix.
- *Transform Set*: It includes all the processing needed to compute the inflected forms of a given aspect, mood and voice. It consists of a set of processes

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

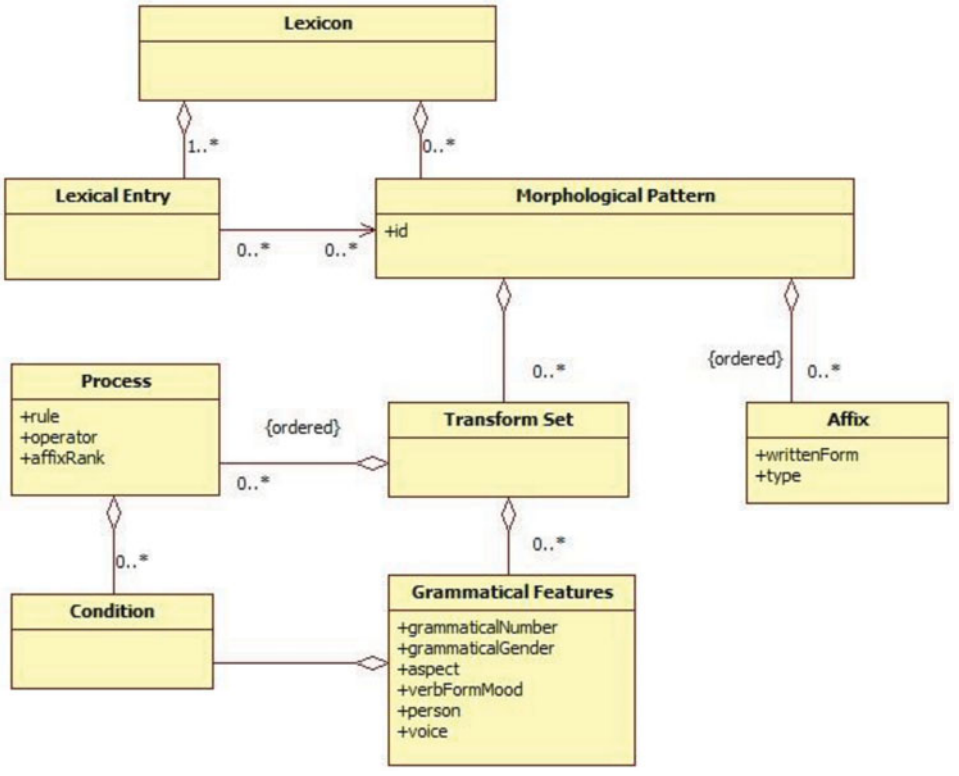


Fig. 8. (Colour online) The Arabic morphological pattern model diagram.

that present the treatments, and one *Grammatical Feature*, which includes the features of a basic form which is common to a set of inflected forms.

- *Process*: It represents a single treatment. This treatment may be either a rule or an operation of adding an affix. If the process presents an operation, then it must have an instance of the *Condition* class.

  - To specify a rule: We use the DC rule, which generates the basic form. This DC is composed of two parts: the first presents the original form of the LE, and the second presents the transformations to apply to the LE: <Feat att = "rule" val = "c{ـ}c{ـ}cـ= c{ـ}c{ـ}cـ"/>. Braces mark these transformations.
  - To specify an operation: we use the DCs operator, affixRank. This operation allows the addition of an affix with respect to a given gender, number and person. Thus, the operator specifies addAffix. The affixRank specifies the serial number of the affix.

- *Condition*: It is made up of a *Grammatical Feature*, which includes the features specific to the addition of an affix.
- *Grammatical Features*: It includes a set of morphological features that characterize an inflected form.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

9 Syntactic modeling

For the Arabic language, there are two types of sentences: nominal and verbal. A nominal sentence is distinguished by the absence of a complete verb (الأفعال التّامة Al'afçaAl Alt∼aAmaħ) and can be influenced by the deficient verbs (الأفعال الناقصة Al'afçaAl Aln∼aAqiSaħ) or the functional particles (الحروف العاملة AlHuruwf AlçaAmilaħ). Thus, besides the complete verbs, the deficient verbs and functional particles have idiosyncratic syntactic information. We specify the subcategorization frames of complete verbs, deficient verbs and functional particles.

9.1 The complete verbs

In the case of complete verbs, the types of subcategorization frame depend on the number of necessary complements to the verb and the need of one or several particles[10] with these complements. In the following Table 5, we present the eight subcategorization frame types of the complete verbs, specified in earlier work (Loukil *et al.* 2008).

Table 5. *Types of subcategorization frame for complete verbs*

English translation	Arabic
Intransitive verb	فِعْلٌ لَازِمٌ
Transitive verb requiring one direct complement	فِعْلٌ مُتَعَدِي لِمَفْعُولٍ وَاحِدٍ بِنَفْسِهِ
Transitive verb requiring two direct complements	فِعْلٌ مُتَعَدِي لِمَفْعُولَيْنِ
Transitive verb requiring three direct complements	فِعْلٌ مُتَعَدِي لِثَلَاثَةِ مَفَاعِيلٍ
Transitive verb requiring one complement with a preposition	فِعْلٌ مُتَعَدِي لِمَفْعُولٍ بِوَاسِطَةِ أَدَاةٍ
Transitive verb requiring two complements one of them with a preposition	فِعْلٌ مُتَعَدِي لِمَفْعُولَيْنِ أَحَدُهُمَا بِأَدَاةٍ
Transitive verb requiring two complements with prepositions	فِعْلٌ مُتَعَدِي لِمَفْعُولَيْنِ بِوَاسِطَةِ أَدَاتَيْنِ
Transitive verb with three complements: the first is direct and the others are with prepositions	فِعْلٌ مُتَعَدِي لِثَلَاثَةِ مَفَاعِيلٍ بِوَاسِطَةِ أَدَاتَيْنِ

9.2 The deficient verbs

The deficient verbs are used with the nominal sentences that are made up of a topic and an attribute. The types of subcategorization frame depend on the mode of the components of the nominal sentence. We have three types of subcategorization frame for deficient verbs illustrated in the Table 6.

[10] In this case, the particles used are prepositions.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Table 6. *Types of subcategorization frame for deficient verbs*

English translation	Arabic
Regularity topic and opening attribute	يَرْفَعُ المُبْتَدَأ وَ يَنْصِبُ الخَبَرَ
Opening topic and regularity attribute	يَنْصِبُ المُبْتَدَأ وَ يَرْفَعُ الخَبَرَ
Opening topic and attribute	يَنْصِبُ المُبْتَدَأ وَ الخَبَرَ

9.3 The functional particles

Concerning functional particles, these letters are used with nominal sentences. The types of subcategorization frame depend on the mode and the nature (noun, verb) of the components of the nominal sentence. We have retained seven types of subcategorization frames for particles, indicated in Table 7.

Table 7. *Types of subcategorization frame for particles*

English translation	Arabic
The attraction	العَطْفْ
The reduction	الجَرّ
The annulment (Nasskh)	النَّسْخ
The exclusion	الاستثناء
The call	النّداء
The elision	الجَزْم
The openness	النّصْب

9.4 The Arabic syntactic model

The syntactic extension allows the representation of the syntactical context (i.e. in a sentence) of an LE through subcategorization frames. Therefore, we have to specify the possible frames for an LE and their appropriate matching to the various meanings.

The following Figure 9 specifies the model of the syntactic LMF extension.

This model is composed of the following classes:

- *Subcategorization Frame*: It is specific to a lexicon. It is composed of a set of *Syntactic Argument* instances and a *Lexeme Property* instance includes the characteristics of the central node of this frame. It is characterized by an identifier and a type.
- *Syntactic Argument*: It represents an argument of a Frame and has two DCs, syntactic Function and introducer. The introducer is usually a particle that precedes this argument. The syntactic Function can take values that depend on the sentence described by the frame:
  - If the sentence is verbal to active voice, then it can take the values: agent (فَاعِلْ faAcil), complement (مَفْعُولْ بِهِ mafcuwl bihi).

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

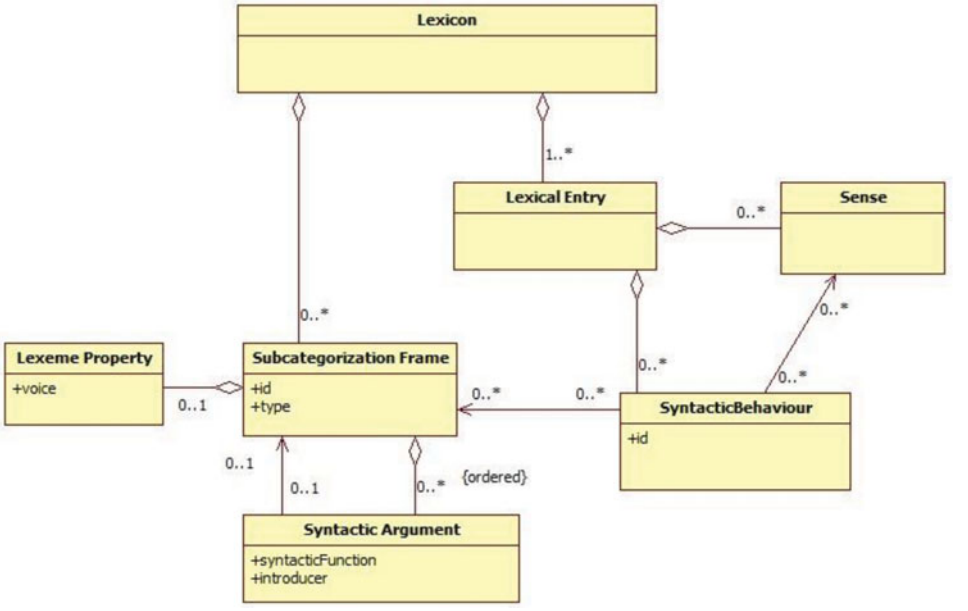


Fig. 9. (Colour online) Syntactic Arabic model diagram.

- If the sentence is verbal to passive voice, then it can take the values: pro-agent (الفَاعِلْ نَائِبْ naA'ib AlfaAçil).
- If the sentence is nominal, then it can take the values: topic (مُبْتَدَأْ mubtada'), attribute (خَبَرْ xabar).

- *Lexeme Property*: It represents the features of the central node of the frame. Generally, the node is a verb that must specify these features as the voice.
- *Syntactic Behavior*: The main role of this class is the specification of meanings which agree with a subcategorization frame as shown in Figure 10. It is characterized by an identifier.

Figure 11 shows the connection between the senses of the LE كَتَبَ *kataba* and the subcategorization frames using a *Syntactic Behavior* class.

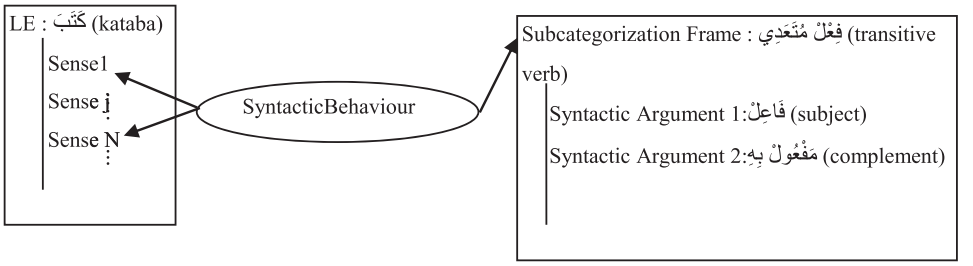


Fig. 10. Example of a syntactic link.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

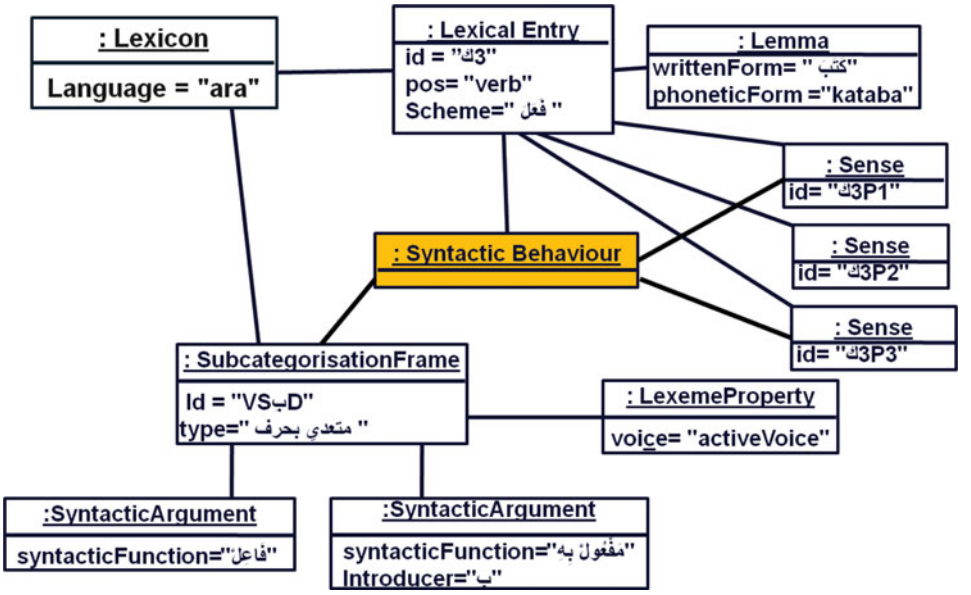


Fig. 11. (Colour online) Example of connecting between an entry and subcategorization frame.

10 Semantic modeling

According to LMF, the semantic knowledge is spread across the core, the semantic and the MRD extensions. In the following Figure 12, we present the selected classes and the DC added in the *Sense* class.

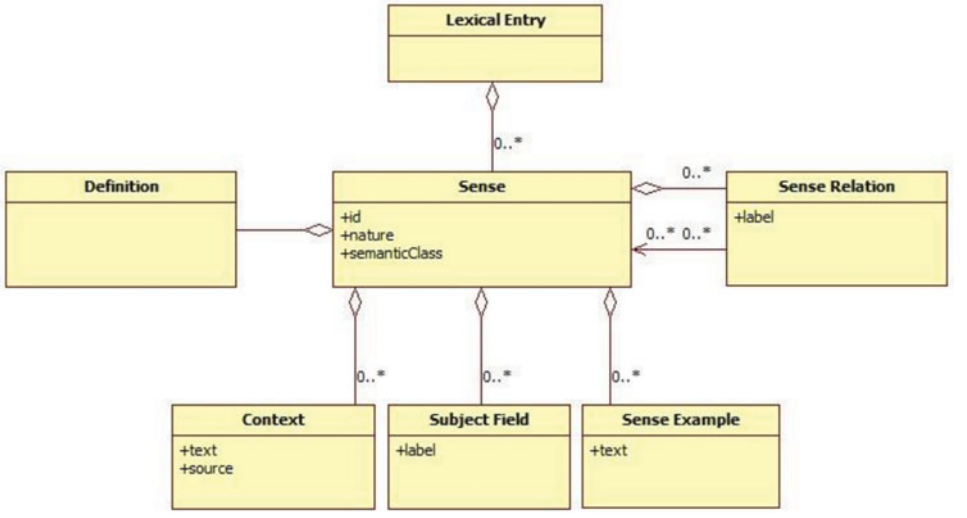


Fig. 12. (Colour online) Arabic semantic model diagram.

Each sense is characterized by its nature (literal or metaphorical) and semantic class (i.e. العطف AlçaTfu, الظن AlĎan,). In addition, we selected two classes of the semantic extension:

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

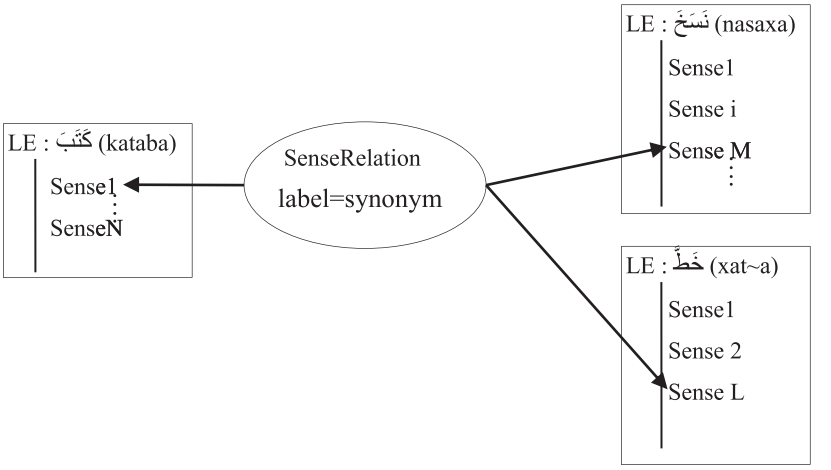


Fig. 13. Illustration of the semantic link principle between 'كَتَبَ', 'خَطَّ' and 'نَسَخَ'.

- *Sense Relation*: It allows connecting the senses belonging to different LE. There are several types of relationships. We quote mainly the synonym and the antonym. We present the type of link in the DC label.
- *Sense Example*: It represents an instance or a sample of meaning.

We selected two classes from the MRD extension:

- *Subject Field*: It is used when the meaning is specific to a particular area.
- *Context*: It represents an example of using an LE in the frame of a given meaning. It has the DC text that takes a sentence and a source representing the origin of this sentence.

Figure 13 illustrates the principle of semantic link between entries supporting the meaning of the verb 'to write' using *Sense Relation*. In this illustration, we link the synonymous meanings of the three LE, but not the entries. The same example is given in the following figure, when applying the normalized model. Figure 14 shows an example of link between the two entries 'كَتَبَ' *kataba* and 'نَسَخَ' *nasaxa* through their synonymous meanings using the *Sense Relation* object.

11 Syntactic/semantic interface

LMF specifies the interface between syntax and semantics within the semantic extension. This interface consists mainly of syntactic arguments and their semantic equivalent, an example and a definition as shown in Figure 15 below.

- *Predicative Representation*: This class represents the interface between syntactic and semantic levels. It is specific to one meaning, it consists of a syntactic behavior, a semantic predicate and the correspondence between the two.
- *Semantic Predicate*: This class represents a general meaning with its associated arguments that are represented in the *Semantic Argument* class. The examples of the predicate can generate a new meaning which is presented by the *Definition* class.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

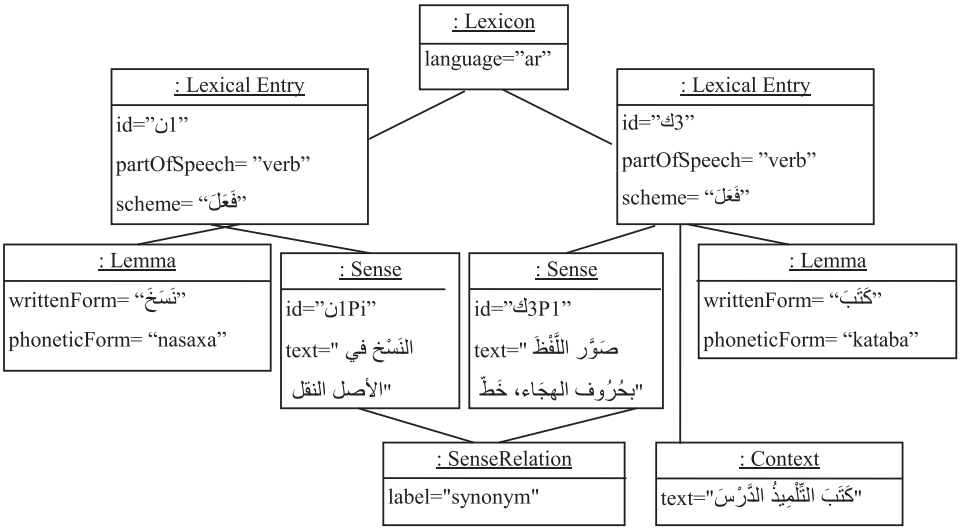


Fig. 14. Part of the object diagram for the entries ‘كَتَبَ’ and ‘نَسَخَ’.

- *Semantic Argument*: It is a class representing each argument of a given semantic predicate.
- *SynSemCorrespondence*: This class represents a set of *SynSemArgMap* instances for a given *Subcategorisation Frame* instance.
- *SynSemArgMap*: It is a class that links each semantic argument to its corresponding syntactic argument.

The originality of our model is in the representation of the semantic predicate (using *Semantic Predicate* and *Semantic Argument* classes) and its link with the syntax through *Predicative Representation* class that sets the interface between syntactic and semantic levels. It is specific to one meaning and consists of a syntactic behavior, a semantic predicate and the correspondence between the two instances.

In the previous example, given in Figure 16, we show how to ensure the link between a meaning of LE, one syntactic behavior, one semantic predicate and their correspondence using *Predicative Representation* class. The semantic predicate has two semantic arguments. The first has the following attributes: thematic role = ‘agent’ and semantic class = ‘human’. The second argument has the attributes: thematic role = ‘instrument’ and semantic class = ‘inanimate concrete’. In addition, we link the semantic arguments to their corresponding syntax. Thus, the specification of the semantic features in each argument is useful in filtering, checking the compatibility and enriching of the semantic representation.

12 Implementation

12.1 XML serialization

As recommended by LMF, we use XML to implement our model. Furthermore, with the richness and complexity of this model, notably the various relationships between

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

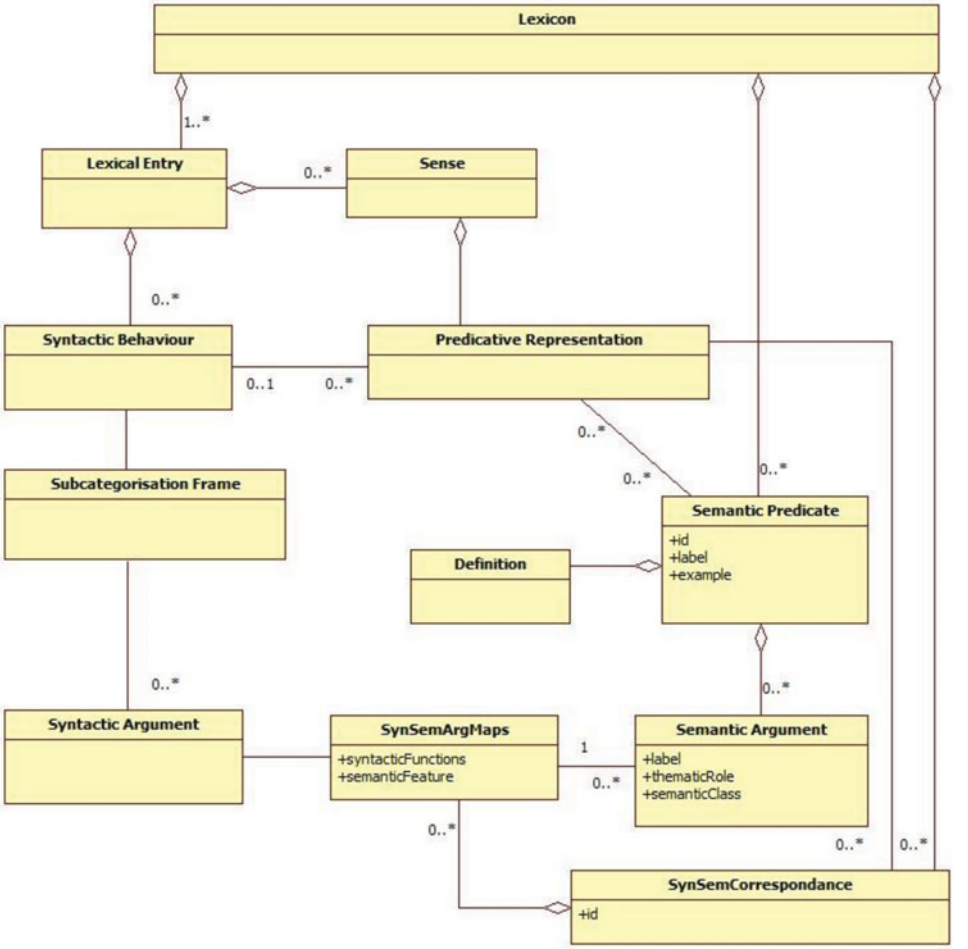


Fig. 15. (Colour online) Arabic syntactic–semantic model diagram.

the dictionary knowledge (e.g. LE, sense, subcategorization frame), we resort to a rigorous knowledge codification. Moreover, in order to manage the huge number of LEs and their related knowledge properly, we propose an appropriate organization in order to facilitate its population and exploitation.

12.2 Collaborative acquisition tool

The richness of the Arabic language and the fineness of the proposed model make the population process of the dictionary difficult and time consuming. To overcome these problems, we are in favor of a collaborative approach. This is why we developed a user-friendly website which allows the lexicographers to participate in the populating of the dictionary. This website is a system formed by a graphical interface that covers the whole description of a *LE* and a set of API (Application Programming Interface), which controls the acquisition of knowledge using completeness and consistency rules. The implemented tool consists of a main Web page (see below) for the

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

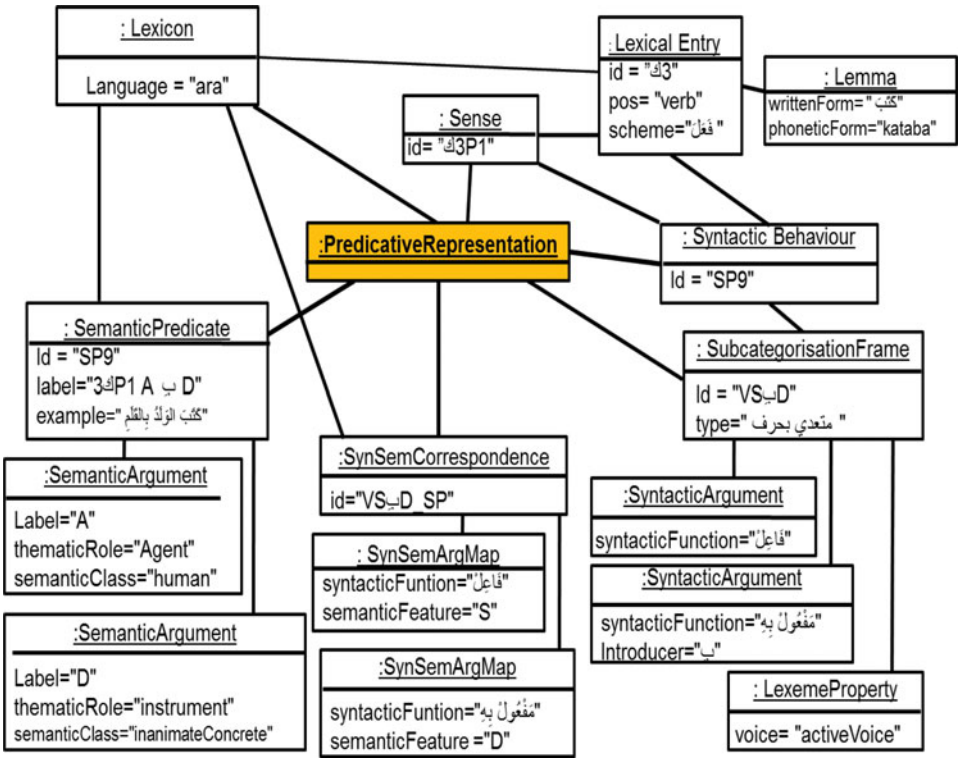


Fig. 16. (Colour online) Example of syntactic–semantic link.

acquisition of general information and appropriate for Web Pages for each linguistic level. All the Web Pages act dynamically according to the type of knowledge like POS and subcategorization frame.

In this main web page, there are four zones (Figure 17). The first zone is for the lemma of the LE. The second is for the category specification (verb, noun and particle). The category specification is mandatory and it constrains the other zones. For example, in zone 3, if the LE is a particle, the scheme, stem and root are not considered. In zone 4, which encompasses the three sides, the morphological side does not appear since particles do not have morphological characteristics and are invariable.

12.3 Dictionary

In order to accelerate the feeding of the dictionary, we converted the content of print dictionaries into a normalized dictionary (Khemakhem *et al.* 2009). We carried out the conversion of El-Ghany dictionary [http://www.alfaseeh.com]. Doing so, we obtained 40% of our current entries. A group of lexicographers have participated in the validation of the conversion results and have enriched the obtained entries by specific knowledge according to the proposed model such as subcategorization frames, semantic classes, thematic roles and semantic predicates. Twenty students and four professors worked on a partial time basis over a period of eight years, and it is difficult to give accurate figures concerning the manpower spent in these manual operations.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

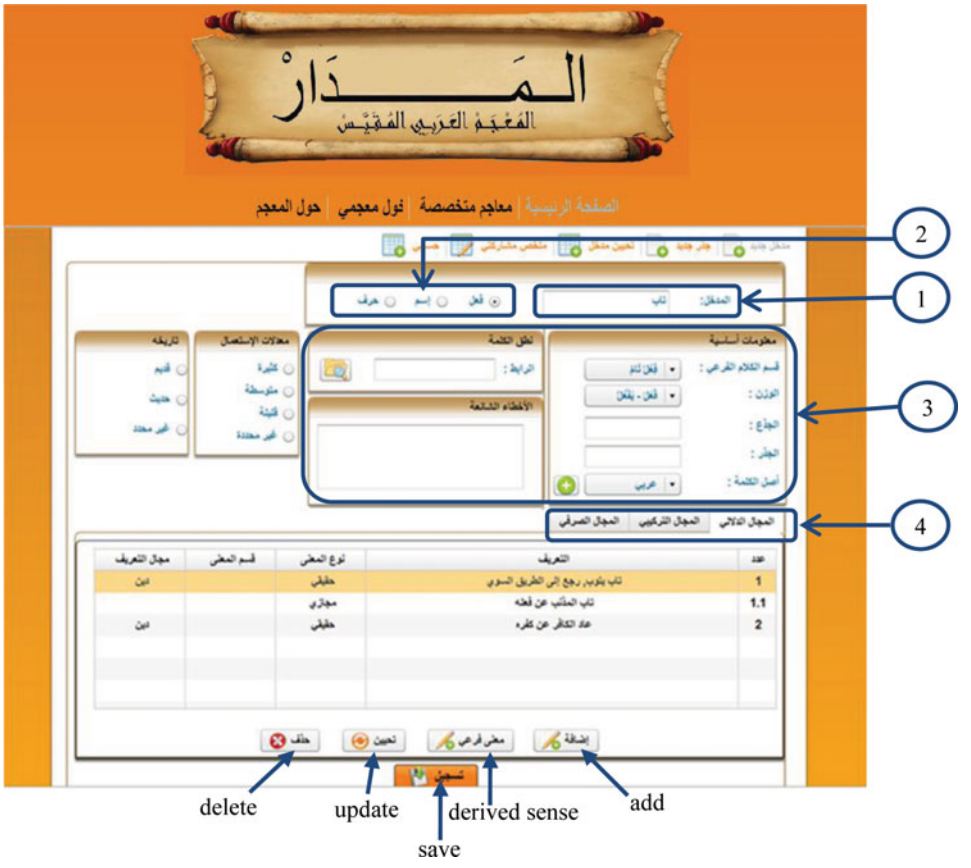


Fig. 17. (Colour online) Example of a population interface for a verb.

Concerning the API associated to the built dictionary, we implemented the tools of the collaborative acquisition/population that will be available online as soon as possible. Moreover, we have implemented and made available the query tools for human use on the Web: http://elmadar.miracl-apps.com. Theses query tools cover, among others, the service of generating inflectional forms based on morphological patterns of the model.

At present, the lexicon contains 37,418 LEs that are distributed according to their categories (see Table 8), knowing that we have used the word 'others' that gather the LE categories which have a low number of occurrences, such as diminutive noun, once noun and particle.

In this following Table 9, we give the number of morphological relationships.

13 Discussion

The aim of the present section is to compare El Madar (i.e. our lexicon) with the models of some well-known dictionaries such as DIINAR, Buckwalter, ElixirFM, Al-Ghany and Wiktionary. To this end, we retain the criteria related to the kind of use, the compliance to standards and the linguistic coverage while highlighting

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Table 8. *Distribution of entries according to their category*

Category of lexical entry	Number of entries
root	3 709
verb	10 796
masdar	12 286
common noun	5 342
present participle adjective	2 067
past participle adjective	1 057
relation noun	366
intensive	402
adjective	148
others	1 245

Table 9. *Distribution of morphological relationships*

Relationship	Number
Root - derived form	17,622
Stem - derived form	15,311
Derived form – root/stem	14,819

the models fineness. Thus, as it is shown in Table 10 given below, we clearly note that El Madar is the only one to deal with both human and NLP uses and to apply normalized data categories. Concerning the morphological coverage, we observe that several dictionaries including El Madar are rich enough. They generally apply an extensional representation of derived words. This can be beneficial to each derived word to fix its knowledge such as meanings. Only the DIINAR dictionary uses the linguistic rules to generate derived words. However, regarding the inflectional aspect, El Madar as well as Wiktionary and DIINAR resort to the intensional representation using inflectional rules, whereas the other dictionaries use an extensional representation. As for syntactic richness, we can state that only El Madar has a high coverage. The others either have an average coverage, like ElixirFM, or a weak and even null one, as Buckwalter. In addition, El Madar is the only one that represents syntactic behaviors. Regarding the semantic representation, only human use dictionaries including El Madar are rich enough. However, El Madar is characterized by the representation of semantic classes of entries. It is also the only dictionary that links synonyms and antonyms to the appropriate meanings and not to the entries, while providing flexibility of choice of the semantic relationship through the link *Sense Relation*. Moreover, the other novelty of this dictionary stems from its reliance on syntactic and semantic knowledge (Khemakhem, Gargouri and Ben Hamadou 2011).

El Madar widely fulfills the requirements of human use while ensuring a good coverage at all linguistic levels. In addition, it is very powerful for the NLP use and contains finely structured knowledge favoring a selective access. In this regard,

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Table 10. *Comparison of Arabic dictionaries models*

		Normalize dictionary	Al Ghany	Arabic Wiktionary	DIINAR	Buckwalter	ElixirFM
Kind of use (NLP/Editorial)		NLP + E	E	E	NLP	NLP	NLP
Use of normalized data categories		Yes (DCR)	No	No	No	No	No
Morphology	Richness of knowledge	High	Weak	High	High	High	High
	Derivation (extensional/intensional)	Ex	Ex	Ex	Int	Ex	Ex
	Inflection (extensional/intensional)	Int	No	Int	Int	Ex	Ex
Syntax	Richness of knowledge	High	Weak	Weak	Weak	No	Medium
	Syntactic description	Yes	No	No	No	No	No
Semantics	Richness of knowledge	High	High	High	No	No	No
	Synonym/antonym link	sense	entry	entry	No	No	No
	Semantic classes	Yes	No	No	No	No	No
Syntax-semantic dependence		Yes	No	No	No	No	No

several Lexical Web services have been developed to query the normalized dictionary for both human and NLP uses (Ben Abderrahmen *et al.* 2006), (Ben Abderrahmen, Gargouri and Jmaiel 2009). Furthermore, the normalized model is so rich and explicit that it enhanced our team to propose an approach to generate domain ontologies (Baccar, Gargouri and Ben Hamadou 2010), (Baccar, Gargouri and Ben Hamadou 2011), (Baccar, Gargouri and Ben Hamadou 2012).

14 Conclusion, current and future work

In this paper, we propose a normalized model of an electronic dictionary for the Arabic language. The originality of our model lies in offering several gathered features that, to our knowledge, no other Arabic dictionary provides. Besides, the advantages provided by the use of ISO standards (i.e. LMF-ISO 24613, DCR-ISO 12620) such as unification of linguistic concepts, flexibility, extensibility and interoperability, the proposed model is useful for human as well as NLP uses. Moreover, it covers all linguistic levels (i.e. morphological, syntactic, semantics) and their intersection aspects. Furthermore, the model is finely structured and rich in relationships in such a way that it allows us to seek all linguistic concepts and relations between them. The morphological patterns provide the generation and lemmatization of the corresponding derived forms by means of a factorized and optimal set of linguistic structures. The dictionary promotes the analyses of Arabic texts by providing explicit knowledge like syntactic behaviors and semantic predicates, useful for the syntactic–semantic analyses. In order to illustrate this model, we developed an appropriate tool for the collaborative population and query for large lexicon. At present, the dictionary contains more than 37,000 entries covering various categories (i.e. root, verb, noun, adjective) connected by different links (i.e. morphological, semantic, syntactic–semantic). An appropriate codification has been used in order to optimize both knowledge representation and access time. To this end, an appropriate API is now implemented, based on the Web services approach (Ben Abderrahmen *et al.* 2009), (Baklouti *et al.* 2013). This piece of software is used to test a new generation approach of domain ontologies, which shows very promising results. We intend to expand the use of the developed collaborative population system by making it available publicly on the Net.

Our current work is to confront the lexicon with the Penn Arabic Treebank to compute a precise percentage of coverage. Our future work will be to confront the lexicon with pages which are collected automatically from the Web in order to be sure that most frequent modern neologisms, buzz words and proper nouns are effectively covered.

We also follow the work done within the W3 C Ontology-Lexica Community Group[11] based on LMF and Lemon[12] concerning the relation between ontologies and lexicons in association with Resource Description Framework (RDF), but the specifications are not stable enough to be concretely applied.

[11] www.w3.org/community/ontolex/.
[12] http://lemon-model.net/.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

References

Abbès, R., Dichy, J., and Hassoun, M. 2004. The architecture of a standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages - COLING 2004*, pp. 15–22. University of Geneva.

Ait Taleb, S. 2005. Dictionnaires électroniques arabes: le modèle des dictionnaires de Sakhr. Revue of the "*Association Marocaine des Etudes Lexicographiques*", pp. 15–31. Number 3–4.

Alouani, M. 2008. «المعاجم العربية», *مجلة البحوث الإسلامية*, Number 11, pp. 227–259. [online] available on www.alukah.net/Articles/Article.aspx?CategoryID=78& ArticleID=417. (Accessed in 2008).

Antoni-Lay, M., Francopoulo, G., and Zaysser, L. 1994. A generic model for reusable lexicons: the Genelex project. *Literary and Linguistic Computing* **9**(1), pp. 47–54. (Report Project Eureka GENELEX).

Attia, M., Pecina, P., Lamia, T., Toral, A., and Van Genabith, J. 2011. A lexical atabase for modern standard Arabic interoperable with a finite state morphological transducer. In C. Mahlow, and M. Piotrowski (eds.), *Systems and Frameworks for Computational Morphology. Second International Workshop, SFCM*, vol. 100, 1st ed., pp. 98–118. Series: Communications in Computer and Information Science. Zurich, Switzerland.

Baccar, F., Gargouri, B., and Ben Hamadou, A. 2010. Towards generation of domain ontology from LMF standardized dictionaries. In *The 22*nd *International Conference on Software Engineering and Knowledge Engineering (SEKE)*, Redwood City, San Francisco Bay, USA, pp. 515–520.

Baccar, F., Gargouri, B., and Ben Hamadou, A. 2011. Domain ontology generation using LMF standardized dictionary structure. *The 6th International Conference on Software Paradigm Trends (ICSOFT)*, Seville, Spain, pp. 396–401.

Baccar, F., Gargouri, B., and Ben Hamadou, A. 2012. LMF dictionary-based approach for domain ontology generation. In *Semi-Automatic Ontology Development: Processes and Resources,* pp. 106–130. IGI Global editions.

Baklouti, N., Fakhfakh, F., Gargouri, B., and Jmaiel, M. 2013. OWL-LingS editor: a tool for semantic description of linguistic Web services. In *3rd International Conference on Cloud Computing and Services Science (CLOSER)*, Aachen, Germany, pp. 224–227.

Ben Abderrahmen, M., Chaari, F., Gargouri, B., and Jmaiel, M. 2006. Des services orientés besoin pour l'exploitation des bases lexicales normalisées.In *10th Maghrebian Conference on Software Engineering and Artificial Intelligence MCSEAI*, Agadir, Morocco, pp. 451–456.

Ben Abderrahmen, M., Gargouri, B., and Jmaiel, M. 2009. LMF-QL: a graphical tool to query LMF databases for NLP and editorial uses. In *Human Language Technology. Challenges of the Information Society*, vol. 5603, pp. 279–290. Lecture Notes in Computer Science.

Ben Mrad, I. 1987. *دراسات في المعجم العربيّ*. Dar Al-Gharb Al-Islami, Beyrouth, Liban.

Bertagna, F., Calzolari, N., Lenci, A., and Zampolli, A. 2000. ISLE — computational lexicons working group. In *The Multilingual ISLE Lexical Entry (MILE): a discussion paper*. http://www.tagmatica.fr/ doc.htm.

Bogurev, B., Briscoe, E. J., Calzolari, N., Cater, A., Meijs, W., and Zampolli, A. 1988. Acquisition of lexical knowledge for natural language processing systems (ACQUILEX), *Proposal for ESPRIT Basic Research Actions N° 3030*. Cambridge, United Kingdom.

Boudelaa, S., and Marslen-Wilson, W. 2010. Aralex: a lexical database for modern standard Arabic. *Behaviour Research Methods* **42**(2): 481–487.

Buckwalter, T. 2004. Buckwalter Arabic morphological analyzer version 2.0. *LDC catalog number LDC2004L02*.

Calzolari, N., Mc Naught, J., and Zampolli, A. 1996. EAGLES Final Report, *EAG-EB-EI*, Pisa, Italy.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Calzolari, N., and Monachini, M. 1996. Multext — common specifications and notation for lexicon encoding. http://www.lpl.univ-aix.fr/ projects/multext/LEX/LEX1.html

Calzolari, N., Monachini, M., and Soria, C. 2013. LMF-historical context and perspectives. In *LMF Lexical Markup Framework,* pp. 1–18. Wiley-ISTE, London.

Chaâben, N., Hadrich, L., and Ben Hamadou, A. 2010. The MORPH2 new version: a robust morphological analyzer for Arabic texts. In *10th International Conference on the Statistical Analysis of Textual Data (JADT)*, Rome, Italy. http://jadt2010.uniroma1.it/

Darwish A. 1956. المعاجم العربية مع اعتناء خاص بمعجم العين للخليل, Egypt.

Diab, M., Habash, N., Rambow, O., Al Tantawy, M., and Benajiba, Y. 2010. COLABA: Arabic dialect annotation and processing. In Proceedings *of the Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages at LREC*, Malta, pp. 66–74.

Doumi, N., Lehireche, A., Maurel, D., and Alicherif, M. 2013. Conception d'un jeu de ressources libres pour le TAL arabe sous Unitex. In *the 6th International Conference on Traductologie and TAL* – [hal-01024409-version 1].

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., and Alkhalifa, M. 2006. Arabic WordNet and the challenges of Arabic. In *Proceedings of Arabic NLP/MTConference*, London, U.K., pp. 15–24.

Francopoulo, G. 2003. Proposition de normalisation de norme des lexiques pour le traitement automatique du langage. *INRIA/LORIA-ACTION SYNTAXE*, Version-1.3.

Francopoulo, G., and George, M. 2008. ISO/TC 37/SC 4 N453 (N330 Rev.16). *Language resource management- Lexical Markup Framework (LMF).*

Francopoulo, G. 2013 (editor). *LMF Lexical Markup Framework*. Wiley-ISTE. London.

Habash, N., Soudi, A., and Buckwalter, T. 2007. On Arabic transliteration. In A. van den Bosch, and A. Soudi (eds.), *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pp. 15–22. Berlin: Springer-Verlag.

Haywood, J. A. 1960. *Arabic Lexicography: Its History, and its Place in the General History of Lexicography*, Leiden, E. J. Brill.

Ide, N., and Romary, L. 2004. A registry of standard data categories for linguistic annotation. In *Proceedings of LREC*, Lisbon, pp. 135–138.

Khemakhem, A., Gargouri, B., and Abdelwahed, A. 2006. LMF est-il convenable pour la langue arabe?, *Journées sur le Traitement Automatique de la Langue Arabe JTALA,* Rabat, Morocco.

Khemakhem, A., Gargouri, B., Abdelwahed, A., and Francopoulo, G. 2007. Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613. In *The Conference Traitement Automatique des Langues Naturelles*, Toulouse, France, pp. 133–142.

Khemakhem, A., Elleuch, I., Gargouri, B., and Ben Hamadou, A. 2009. Towards an automatic conversion approach of editorial Arabic dictionaries into LMF-ISO 24613 standardized mode. In *2nd International Conference on Arabic Language Resources and Tools- MEDAR,* Cairo, Egypt.

Khemakhem, A., Gargouri, B., and Ben Hamadou, A. 2011. Modélisation syntaxico-sémantique normalisée pour la langue arabe. In *30ème Colloque International sur le Lexique et la Grammaire LGC*, Nicosie, Cyprus, pp. 453–464.

Khemakhem, A, Gargouri, B., and Ben Hamadou, A. 2012. LMF standardized dictionary for Arabic language. In *International Conference on Computing and Information Technology,* Al-Madinah Al-Munawarah, Saudi Arabia.

Landau, S. 2001. Dictionaries: *The Art and Craft of Lexicography*, 2nd ed. Cambridge, United Kingdom: Cambridge University Press.

Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., and Zampolli, A. 2000. SIMPLE linguistic specifications*, SIMPLE LE4–8346 EC Project, Deliverable D2.1 & D2.2, WP02, Final version*, ILC et Université de Pisa. http://www.ub.es/gilcub/SIMPLE/simple.html# Specifications.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press

Loukil, N., Haddar, K., and Ben Hamadou, A. 2008. Towards a syntactic lexicon of Arabic verbs. In *HLT& NPL within the Arabic World: Arabic Language &Local Languages Processing – Status Updates &Prospects-LREC*, pp. 93–96. Marrakech, Morocco.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. 2006. Developing and using a pilot dialectal Arabic treebank. In *Proceedings of LREC*, pp. 17–34. Genoa, Italy.

Maks I., Tiberius C., and Van Veenendaal R. 2008. Standardising bilingual lexical resources according to the Lexicon Markup Framework. In: *Proceedings of LREC*, Marrakech, Morocco, pp. 1723–1727.

Mesfar, S., and Silberztein, M. 2008. Transducer minimization and information compression for NooJ dictionaries. In *International Conference Finite-State Methods and Natural Language Processing – FSMNLP*, Joint Research Centre of the EC, Ispra, Italy, pp. 110–121.

Rey-Debove, J. 1971. *Etude Linguistique et Sémiotique des Dictionnaires Français Contemporains*, pp. 317–323. Mouton. The Hague, Netherlands.

Romary, L., Salmon-Al, S., and Francopoulo, G. 2004. Standards going concrete: from LMF to Morphalou. In *Workshop on Electronic Dictionaries, Coling,* Geneva, Switzerland, pp. 22–28.

Salmon-Alt, S., Akrout, A., and Romary, L. 2005. Proposals for a normalized representation of standard Arabic full form lexica. In *Second International Conference on Machine Intelligence (ACIDCA-ICMI 2005)*, Tozeur, Tunisia.

Sawalha, M., Atwell, E., and Abushariah, M. A. M. 2013. SALMA: Standard Arabic Language Morphological Analysis. In *Proceedings of the 1st International Conference on Communications, Signal Processing, and their Application,* Sharjah, UAE, pp. 1–6.

Smrž, O. 2007. ElixirFM - implementation of functional Arabic morphology. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* Prague, pp. 1–8. Czech Republic.

Véronis, J., and Ide, N. 1996. Encodage des dictionnaires électroniques: problèmes et propositions de la TEI. In D. Piotrowsky (ed.), *Lexicographie et informatique - Autour de l'informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy*, pp. 239–261. Paris, Didier Erudition.

https://doi.org/10.1017/S1351324915000224 Published online by Cambridge University Press