Proposals for a normalized representation of Standard Arabic full form lexica

Susanne Salmon-Alt[1], Amine Akrout[1], Laurent Romary[2]

[1] ATILF-CNRS, Analyse et Traitement Informatique de la Langue Française , Nancy, France
Susanne.Salmon-Alt@atilf.fr, amine_akrout@yahoo.fr

[2] LORIA, Laboratoire Lorrain de Recherche en Informatique, Nancy, France
Laurent.Romary@atilf.fr

***Abstract* — Standardized lexical resources are an important prerequisite for the development of robust and wide coverage natural language processing application. Therefore, we applied the Lexical Markup Framework, a recent ISO initiative towards standards for designing, implementing and representing lexical resources, on a test bed of data for an Arabic full form lexicon. Besides minor structural accommodation that would be needed in order to take into account the traditional root-based organization of Arabic dictionaries, the LMF proposal appeared to be suitable to our purpose, especially because of the separate management of the hierarchical data structure (LMF core model) and elementary linguistic descriptors (data categories).**

I. INTRODUCTION

Any type of linguistic processing on texts requires a minimal set of lexical resources that will be matched against the actual words encountered in the textual data. This is all the more crucial for character and text recognition where the overall recognition rate will highly rely on the extensiveness of the lexical data available in association with the processing software. Still, if we consider the huge cost of creating and maintaining lexical resources for natural language processing (NLP) – for Arabic, see for example [3], [6], [8], [22] or [28] – we can see that such resources should not be designed in isolation, but should potentially be put in contact with one another for mutual enrichment. In the worst case, the specific context of maintenance of a lexical resource is indeed situated in an intractable network of proprietary lexical databases. As a consequence, we will consider here that there is a strong need for more widely spread methods of specifying lexical structures, so that the conditions under which the corresponding databases may be able to exchange data are precisely defined. Even more, it seems that enough knowledge has been accumulated across the years to consider the idea of an actual international standard for the representation of NLP lexicons that would preserve the possibility of both describing various types of formats and ensuring interoperability between those. Such an enterprise is indeed currently under discussion in the context of ISO committee TC 37/SC 4 in the *Lexical Markup Framework* (*LMF*) project [13], to become the future ISO 24613 standard.

The LMF modeling framework for lexical structures relies on strong previous experience in the specification of lexical databases, such as developed in the context of various European projects like MULTEXT [7], EAGLES [12], ISLE/MILE [4] or Parole [23]. However, as opposed to those, LMF is not just yet another data model for NLP lexicons: it might be best understood as a synthesis and an abstraction over the previous proposals. The underlying idea is to provide a specification platform which allows to use a set of generic building blocks (components) which, combined with elementary descriptors (data categories), is intended to cover not only a wide variety of possible lexical structures, but also a wide range of languages, particularly those behind the traditionally Indo-European set of languages rather well accounted for in the NLP community. The LMF specification principles can be used either as a new descriptive tool for existing lexical resources or as a basis for the design of new lexical databases dedicated to NLP. We wish to illustrate this latter aspect by a case study on using the LMF specification platform for the design of an Arabic morphological full form lexicon. The objectives of this enterprise are twofold: first, we want to evaluate to which extent the current state of LMF makes it applicable to other languages than those so far considered in previous European projects; second, we are about to conceive the skeleton of a large coverage ISO-conformant lexical database for Arabic, to be developed in collaboration with the University of Sfax[1].

II. THE LEXICAL MARKUP FRAMEWORK AT A GLANCE

A. A semasiological view on lexical data

Lexical structures can classically be considered according to the way they organize the relation between words and senses: either senses are considered as subdivisions of the lexical entry (the semasiological view) or on the contrary, it is assumed that words, or better terms, are described as ways of expressing concepts, which are described prior to them (the onomasiological view). The semasiological view is obviously the one that allows an exhaustive survey of lexical content for a

[1] supported by the program "INRIA /Universités Tunisiennes"

given language. In particular, it corresponds to the basis for any classical editorial (or print) dictionary, and also underlies, at least implicitly, most of existing NLP lexicons. From a more theoretical perspective, it has been shown that lexical structures can be modeled as feature structures [16], leading to inheritance properties within entries [15], as partially implemented in the TEI Print Dictionary chapter [19]. It has also been shown that the internal structure of a lexical entry might be configured through different layers : in a two-layered approach, the *form* and *sense* layers are anchored on the Saussurian definition of a linguistic sign and related to the basic notions of *signifier* (sound pattern) and *signified* (concept) [10]. The syntactic behavior of the lexical unit is then systematically subordinated to its semantic description. This is actually the choice made in LMF.

B. The LMF core model

Accordingly, the LMF core model is organized as a purely hierarchical structure built upon the following components (Figure 1):

- the *Lexical database* component, which gathers up all information related to a given lexicon;
- a *Global information* component collecting metadata (e.g. version, contributors, update, etc.);
- a *Lexical entry* component, which corresponds to the elementary lexical unit in a lexical database;
- a *Form* component providing access to surface properties (phonological and graphical realization) and grammatical meaning (inflectional features) of individual word forms;
- one or more *Sense* components, which actually organize the lexical entry, since they can be both repeated and further subdivided into sub- senses.

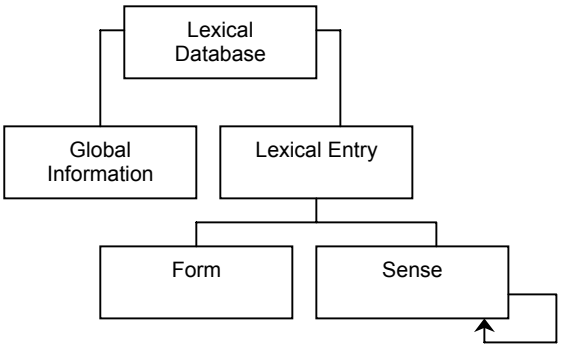

Figure 1. Core components of LMF

C. Data categories

Furthermore, following general principles of linguistic annotation scheme design ([17] et [18]), LMF provides a mechanism for specifying the content of the core meta model components by using elementary descriptors (so-called *data categories*). Data categories reflect basic linguistic concepts, such as /partOfSpeech/, /grammaticalNumber/, or /grammaticalCase/, for example. They are stored and managed independently from the hierarchical structure of the data model: this is indeed the easiest way to allow for recording language specific properties independently of structural properties of the linguistic layers to be described. For instance, the data category /grammaticalGender/ holds two values for French (/masculine/ and /feminine/) and three values for German (/masculine/, /feminine/ and /neuter/). In order to share data categories within the community, the ISO/TC 37 deploys an on-line registry[2] of them, especially for use in conjunction with the other standardization activities. The future LMF standard as such should not provide a specific list of data categories to be used for lexical descriptions. This would by far be too complex, given the potential variety of applications. It is thus expected that implementers will systematically refer to the ISO/TC 37 data category registry to find the adequate descriptive background for their own purpose.

Entry Identifier :	/grammaticalGender/
Profile :	Morpho-syntax
Definition :	Grammatical genders are classes of nouns reflected in the behavior of associated words; every noun must belong to one of the classes and there should be very few which belong to several classes at once.
Explanation:	Grammatical gender is distinguished from natural gender by the fact that grammatical gender requires *agreement* between nouns and the forms of modifiers (demonstratives, articles, adjectives, etc.), whereas natural gender does not.
Source :	Charles F. Hockett, *A Course in Modern Linguistics*, Macmillan, 1958 : 231.
Conceptual Range :	/masculine/, /feminine/, /neuter/, /common/

Object Language :	fr
Name :	genre
Conceptual Range :	{/masculine/, /feminine/}

Object Language :	en
Name :	gender, grammatical gender
Conceptual Range :	{}

Object Language :	de
Name :	Genus, Geschlecht
Conceptual Range :	{/masculine/, /feminine/, /neuter/}

Figure 2. Formal description of data categories

More precisely, the DCR is being established as an organized repository of data elements, which are described according to the principles of ISO DIS 12620 (Terminology and other language resources — Specification of data categories and management of a data category registry for language resources). The principles provide a two level organization of the documentation associated to a data category, as exemplified in Figure 2. At a first level, the general characteristics of the data category are being described through the provision of:

- an identifier that will allow any application to reference it uniquely within the registry (for instance by means of a URI such as *http://www.tc 37.org#grammaticalGender*);

[2] http://syntax.inist.fr

- one or several profiles that will indicate the possible, yet not exclusively, application domains of he data category (such as *morpho-syntax*);
- a definition, which, together with possible explanatory notes, will give the semantics of the data category;
- when applicable, a conceptual domain, providing an open or closed list of values (such as *masculine*, *feminine* etc.), considering that the values in turn are data categories to be registered.

At a second level, a data category can be described in the context of its application to specific languages. Figure 2 thus states that for German (de), *Genus* and *Geschlecht* are two possible names for the data category /grammaticalGender/ and that the conceptual domain only contains three of the values described at the first level (namely /masculine/, /feminine/ and /neuter/).

D. The fully specified model

Given what we have described so far, the actual modeling activity required from the lexical implementer – as reflected currently in the Lexus[3] lexical databank – is limited to the

- selection of relevant components from the meta-model ;
- provision of a selection of relevant data categories from the DCR
- specification of the anchoring component for each data category.

Figure 3 shows for example the anchoring of the DCR data category /grammaticalCategory/ to the meta-model component /lexicalEntry/. This is an implementation choice reflecting the editorial choice to consider that different parts of speech (noun, verb etc.) lead to separate lexical entries in a given dictionary. Another choice would have been to anchor the part of speech on the /sense/ component and to factorize form information for a same entry string.

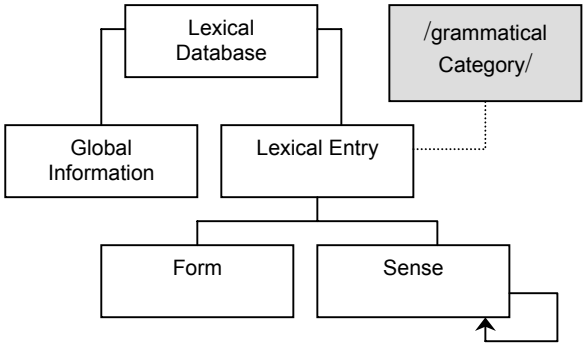

Figure 3. Fully specification of LMF conformant lexica

E. The XML pivot format and user formats

Finally, LMF provides mechanisms to translate the fully specified lexicon model into an isomorphic XML pivot format. The implementers might then chose to express their own model in any user XML dialect – i.e. by deciding to implement a given data category such as /grammaticalCategory/ as an XML attribute rather than an element, or by renaming it as *POS*. The only important fact with respect to the standardization issue is that this proprietary XML dialect must be mappable unambiguously to the LMF-conformant XML pivot structure.

```
<struct type="lexicalDatabase">
        <struct type="globalInformation">...</struct>
        <struct type="lexicalEntry">
                <feat type="grammaticalCategory">...</feat>
                <struct type="form">...</struct>
                <struct type="sense">...</struct>
                <struct type="sense">...</struct>
                ...
        </struct>
        <struct type="lexicalEntry">...</struct>
        ...
</struct>
```

Figure 4. Pivot XML format for the LMF lexicon specification in figure 3

Figure 5. Example of a pivot XML compatible user format

F. Using LMF for a normalized Arabic full form lexicon

In the remainder of this article, we wish to illustrate the use of the LMF specification platform in the context of developing a set of normalized NLP resources for Arabic. Our choice was to start with a morphological full form lexicon, that is a lexicon which lists all inflected forms for a given lexical entry. Those dictionaries are basic resources in the field of NLP. They are needed for any application based on tagged and/or lemmatized input data and in the field of computer-assisted language acquisition. However, most of the existing morphological resources for NLP (*MulText*[4], *LEFFF*[5]) occur as text files, whose lines display the inflected word form, one or more morphological tags (relative to a given tag-set) and the lemma. As opposed to those, we opted for structuring the data into lexical entries rather than along inflected forms. This kind of representation has indeed the advantage of not to be inspired directly by one specific type of usage of such resources (i.e. morphological tagging) and is much easier to extend with respect to syntactic and semantic information. As a consequence, the first step towards this goal has been the design of the skeleton of the lexicon, by determining its underlying macrostructure.

[3] http://www.mpi.nl/lexus

[4] http://www.lpl.univ-aix.fr/projects/multext/MUL5.html

[5] http://atoll.inria.fr/~lclement/lefff/

III. THE MACRO-STRUCTURE OF THE LEXICON

A first important issue is concerned with a fundamental difference in the lexicographical organisation of classical Arabic and Indo-European dictionaries: whereas the basic organisation principles for the latter relies on semasiological and alphabetical criteria, Arabic dictionaries are usually organized into super-entries collecting all derivations from a same tri- or quadri-consonantal root, sometimes considered to stand for a very abstract common meaning[6]. For example, the root *ktb* is a form which does not exist independently, but which has a general meaning related to the notion of writing. It is the entry point for a whole range of verbal or nominal derivates[7]: *kâtaba* (to write), *kattaba* (cause to write), *maktabun* (desk), *maktabatun* (library), *kitâbun* (book) etc.

Especially because of the fastidious usage scenarios for non Arabic users, but also related to arbitrariness in lexeme ordering, there have been several attempts to simplify the direct access to Arabic lexical units ([26], [27]), continuing the precursory work of Ahmed Farid Aššidyăq in the 19th century. However, the access problem (roots vs. lexemes) is not directly crucial to NLP applications, since they are able to carry out non linear researches, as long as the lexicographic entities are properly identified as such. Still, there is no LMF component directly available for modelling the notion of Arabic roots: the lexical entry in the Saussurian sense (an association of form and meaning) is not transferable, since the notion of root is not only a very deep abstraction over forms and meanings without an autonomous existence, but also because it would no longer be distinguishable from the lexemes properly speaking. As a practical consequence, such a choice would conflict with one of the basic LMF principles stating that a lexical entry is a non recursive component, i.e. one that cannot contain other lexical entries.

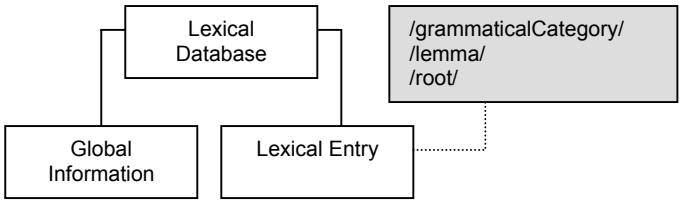

Figure 6. Top level LMF specfication for an Arabic full form lexicon

Our practical solution was therefore to consider as basic lexical entries the lexical derivations (or lexemes) collected below the root level, and to leave the reference to the root to be optionally encoded for each lexeme. With respect to the previously mentioned derivate examples of *ktb*, the lexicon will contain five different entries, each of them pointing to the same root: *kâtaba* (to write), *kattaba* (cause to write), *maktabun* (desk), *maktabatun* (library), *kitâbun* (book). In addition to a /root/ pointer, the minimal set of data categories attached to the lexical entry component informs about the part of speech (/grammaticalCategory/) and gives a conventionally chosen canonical form (/lemma/). So far, we took into account nouns, verbs, adjectives, pronouns and prepositions. Figure 6 shows the top level specification of the lexicon.

IV. ARGUMENTS FOR AN EXTENSIONAL PERSPECTIVE

For the encoding of the inflectional information, we adopted deliberately an extensional perspective, i.e. a description of the full set of forms for a given lexical entry: such an extensional representation, comparable to the lexical description level in the *Multext* project [7] can indeed be seen as opposed to an intensional point of view where inflexions would be described by elementary operators or generative rules on inflectional paradigms, factorizing redundant information ([3], [6]). We argue however that the extensional representation should be considered as the primary reference resource, with respect to fine grained lexicographical information, testimony of inflected forms, as well as resource management :

- lexicographical information: previous experience [25] has shown that the extensional perspective is the only one to allow to represent in a flexible way linguistic information specially related to inflectional forms. We can mention here local inflectional variants (fr. *courbattu* vs. *courbaturé*), local gender variation (*amour, orgue, délice*), feminisation (*avocat* vs. *avocate*), spelling variants (*cheik* vs. *cheikh*), defective paradigms (**nous pleuvons*) or the existing of more than one phonological form for a given inflected form (fr. *les* – [le]/[lez]);
- testimony of inflected forms: an extensional full form lexicon is the only one to account easily for a frequent user claim, that is associating inflected forms of a lexical entry with statistics on its occurrence frequency with respect to a reference corpus ;
- maintenance of lexical resources : As long as there is no theory-neuter consensus about the encoding format of linguistic rules, an important advantage of extensional representations is the possibility of acting as a pivot format for merging and comparing lexicons with the perspective of preserving a high editorial and linguistic quality.

Furthermore, we argue that stable and well documented extensional representations of full form lexicons are an important prerequisite for carrying out further research on several complementary tracks, namely normalization proposals for the representation of intensional (or rule-based) lexicons, or integrating additional morphological operations such as derivation .

V. ADAPTING /FORM/ INFORMATION FROM LMF

The next question concerns the anchoring point for the linguistic information related to the list of inflected forms for each lexical entry. /form/ as defined in the LMF core model (cf. Figure 1) represents a spoken word or multi-word phrase, corresponding at a high level of abstraction to the Saussurian *signifier*, as opposed to the *signified*. At this level, there is a one-to-one cardinality between /sense/ and /form/, considering

[6] This does not exclude the possibility of « homonymous » roots, as mentioned in [14].

[7] The question of whether a root is primarily verbal and/or nominal is a theoretical question that does not affect the modelling framework. For discussion, see for example [14].

however that it is possible to factorize form information for polysemic lexical entries in a common /form/ element. /form/ provides so far a framework for specifying the lexical type (word, word form, sentence, phrase, lemma, headword, multi-word expression), the orthography, a particular script or transliteration, phonology (transcription, intonation etc) and grammar (part of speech).

In particular in the context of designing a full form lexicon, one may ask however, to which extent this /form/ information is able to characterize individually each of the word forms (concrete or surface realizations), rather than an arbitrary form (usually called *lemma*) that functions as a shorthand to represent the whole set of word forms in order to create a unique anchoring point for factorizing grammatical and semantic information. Theoretically, any information attached to such an abstract or canonical form should be generic to the whole set of word forms, that is abstracted over individual linguistic behaviour of each of the word forms. Classifying a lexical unit like FACTEUR(1) as a *noun* is indeed an abstraction over grammatical, i.e. combinatorial, properties of each of the inflected forms. However, these combinatorial properties express constraints on the agreement with dependent units which rely *in fine* on semantic valency. Therefore, one may ask to which extent information about part of speech can still be considered as purely "formal" information.

In the same vein, it seems meaningful to factorize at the abstract form level information abstracted over inflectional features of concrete or surface forms. This can be done either by encoding rules for calculating concrete forms from a given input, or via a reference to an extensional list of full forms, such as argued for here. In any case, it appears that inflectional information also is not purely "formal" in nature. Inflectional features bear indeed a (very generic) sense, which has not to be conflated naively with the senses suggested by traditional tags such as *number*, *tense* or *person*, but which is considered by various authors as being subject of a semantic description in the same way as lexical senses ([9], [21]).

Finally, concerning the remaining form information – especially phonology and orthography – it becomes clear that the association of phonological or orthographical information with any canonical form (recall that it is just a unique identifier) is meaningless, as long as those features pertain only to concrete word forms. Fortunately, the distinction between abstract and concrete forms has been introduced in LMF, by creating additionally to the /form/ component an /inflectedForm/ component. The /inflectedForm/ component represents concrete word forms, associated with any type of information (data categories) directly "observable" at the level of concrete forms: pronunciation and orthography within a specified system of transcription, and inflectional features, such as number and tense, as needed in our project.

In order to model properly linguistic data structures relevant to full form lexica while still being LMF-conformant, we decided to replace the whole /form/ component – standing as a shorthand for a whole inflectional paradigm – with a /wordFormSet/ component, acting as a container for the explicit list of word forms. A single word form fits then the classical Saussurian definition of a linguistic sign. Its description can be split over a /form/ component (dedicated to purely formal aspects, e.g. pronunciation and spelling) and a /inflection/ component that gathers up the grammatical sense due to the inflectional features, as opposed to the lexical sense, which is generally factorized for all inflected forms at the /sense/ side of a lexical entry.

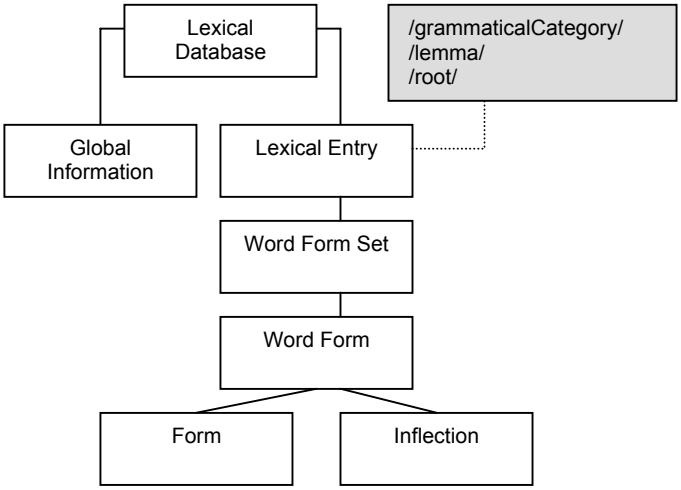

Figure 7. Adapting the LMF /form/ specification to full form lexica

VI. /FORM/-RELATED DATA CATEGORIES

The /form/ component has been reserved to purely formal, e.g. phonological and orthographical realizations of lexical entries. Following [2], we define *orthography* as a human technology consisting of choosing a set of characters and establishing conventions for using them. Languages might have zero, one, or more orthographies. *Transliterations* are orthographies for writing a language in its customary orthographical conventions, but using a symbol set which has a fully reversible one-to-one mapping with the symbol set of the original orthography. An example for Arabic is the Buckwalter transliteration [6]. *Transcriptions* are clearly distinguished from transliterations: they are orthographies devised and used by linguists to characterize the phonology and morphophonology (rather than the original orthography) of a language, especially in order to convey the pronunciation for foreigners which are not comfortable with original orthography. Examples for Arabic include the use of the International Phonetic Alphabet, or transcription adopted officially by the International Convention of Orientalist Scholars in 1936, and used for example in [29].

For our proposal, we chose to unify these three notions under the general concept of (actual or virtual) orthography. Orthographies are represented as one ore more /realization/ features of a given /form/, and qualified with respect to a particular encoding system, the /code/ (Figure 8, 9).

Realization	Code	Status
قتل	Arabic	original orthography
qatala	Akrout [1]	transcription

Figure 8. Qualified realizations for original orthography and transcription

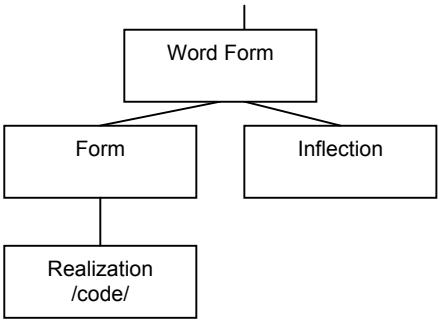

Figure 9. Characterizing the /form/ component with data categories

VII. DATA CATEGORIES FOR ARABIC INFLECTIONS

So far, we have structured the lexicon into lexical entries (and not into roots) and associated each lexical entry with a container (/wordFormSet/) for a set of inflected forms, where each inflected form (/wordForm/) should be described in terms of formal (/form/) and inflectional (/inflection/) properties. At the /inflection/ level, morphological dictionaries typically associate inflected word forms – for example plural noun forms or past tense verb forms – with values for relevant morphological features, such as the number for nouns, or the grammatical tense for verbs.

Therefore, one needs to decide which inflectional properties are considered to be relevant for each of the morpho-syntactic categories included in the lexicon. This work is linguistic in nature and needs a careful analysis of data and traditional linguistic description. Hence, we will not present here the whole list of data categories to be used in our lexicon [1], based mainly based on linguistic work on Arabic grammars ([5], [14], [24]) and the analysis of existing proposals for NLP tagsets ([6], [11], [20]). We rather wish to discuss some general methodological questions in defining data categories for a specific language, taking as examples the morpho-syntactic categories of nouns and verbs in Arabic.

A. Nouns

Arabic nouns bear grammatical gender information. A noun is either feminine or masculine. It has to be noticed that non semantically motivated masculine nouns become feminine in their plural forms. This might be a criterion for not considering the gender as a lexicalized feature (attached to the lexical entry level, such as in Figure 3), but to encode it at the inflection description level. Concerning grammatical number, the peculiarity of Arabic is a three-valued system: one has to distinguish dual, additionally to singular and plural forms. Arabic nouns are also subject to a threefold grammatical case variation: we used the values *nominative*, *accusative* and *prepositional*. The prepositional case covers the form in which a noun appears as an indirect object, in possessive structures and in prepositional phrases. Furthermore, Arabic nouns occur in two different forms, depending of their grammatical definiteness: indefinite (*kitâbun*) and definite. Definiteness is expressed either by the definite article (*alkitâbu*) or in case of determination by personal pronouns in possessive structures (*kitâbî*) or by genitival noun phrases, both without any linking preposition.

Whether the two cases of definiteness has better to be considered as two different data categories or as a variation on definiteness is still under discussion. Another related discussion points is the treatment of pronominal affixes and prepositional affixes (*bikitâb*): we tend to consider those forms rather in terms of composition than inflection. Also an open issue is the specification of the semantic class of the noun. This issue is related to the plural gender change already mentioned: this change occurs only for non semantically motivated masculine nouns. An additional information about the noun's semantic class could therefore be used to maintain the gender as a lexicalized feature, since the plural gender would then be predictable from the noun's semantic. However, we avoided so far to mix up semantic information such as noun class with purely inflectional features.

Data Category Identifier	Conceptual Range
/grammaticalGender/	{/masculine/, feminine/}
/grammaticalNumber/	{/singular/, /dual/, /plural/}
/grammaticalCase/	{/nominative/, /accusative/, /prepositional/}
/grammaticalDefiniteness/	{/indefinite/, /definite/}

Figure 10. /inflection/ data categories for Arabic nouns

B. Verbs

Arabic verbs are subject to a system of inflectional variation, related to the expression of aspect (perfect, imperfect), voice (passive, active), mood (indicative, subjunctive, jussive, and possibly imperative), person (first, second, third), gender (masculine, feminine) and number (singular, plural, dual). Furthermore, the combination of these feature is conditioned by particular co-occurrence constraints: mood distinctions, for example, apply only for imperfect verb forms, and passive voice is incompatible with the imperative mood. In addition to these data categories, Arabic verbs vary also with respect to grammatical number, person and gender.

Data Category Identifier	Conceptual Range
/grammaticalAspect/	{/perfect/, imperfect/}
/grammaticalVoice/	{/active/, /passive/}
/grammaticalMood/	{/indicative/, /subjunctive/, /jussive/}
/grammaticalPerson/	{/firstPerson/, /secondPerson/, /thirdPerson/}
/grammaticalNumber/	{/singular/, /dual/, /plural/}
/grammaticalGender/	{/masculine/, feminine/}

Figure 11. /inflection/ data categories for Arabic verbs

An interesting decision to be discussed is concerned with the conflation of masculine and feminine gender, for example for the first person perfect forms (*katab-tu*). In those cases, one has theoretically different choices:

(1) encode the same form twice, once as being masculine, and once as being feminine;
(2) create a synthetic data category for factorizing the two previous forms under /commonGender/;
(3) don't use the /grammaticalGender/ data category;
(4) consider a genuine data category, such as /neuterGender/ [20].

The last solution (4) should be eliminated because of the lack of special inflection features (as opposed to the German /neuter/ gender) that would justify the introduction of such as class. Solution (3) is not a recommendable choice: in particular, it could leave to a misinterpretation, such as 'Arabic first person perfect forms don't have a grammatical gender'. The two other solutions are slightly equivalent: both state the existence of two different genders, expressed by an underspecified form. The first one (1) would be the solution chosen in a resolutely extensional perspective (double encoding, considering that there exist two different word forms for *katab-tu*), whereas the second one (2) may be understood as a shorthand for the first one: the underlying assumptions are the same (two different word forms), but some of the information has been factorized and can be fully recovered only in case the decoder (human reader or NLP application) has access to a rule that allows him to expand the value /commonGender/ to two different values (/masculine/ or /feminine/).

Another issue is the treatment of the imperative. Whereas the decision to consider it at the same level as the aspectual opposition /perfect/ vs. /imperfect/ ([20]) should be excluded, the questions remains whether it is a separate mood or not. Some authors ([14]) consider indeed the imperative as a variant of the jussive mood, especially because of the defective inflection paradigm (2nd person only) and strong formal similarity with the jussive (excepted the prefix).

VIII. THE FULLY SPECIFIED DATA MODEL

The final data model is shown in Figure 12. Our lexical database contains, besides appropriate metadata, lexical entries or *lemmas* which are characterized by a grammatical category (and optional gloss and root information). Each lemma stands for a set of inflected forms, i.e. word forms. A word form is characterized by a form component – gathering information about various oral or written realisations with respect to a given encoding system –, and an inflection component, gathering inflectional features.

The data categories especially introduced to the DCR for Arabic full form lexica are:

- /root/: the portion of a word that is common to a set of derived or inflected forms, if any, when all affixes are removed is not further analyzable into meaningful elements, being morphologically simple, and carries the principle portion of meaning of the words in which it functions[8];
- /dual/: a grammatical number which refers to two members of the class identified by the noun[9];
- /jussive/: a grammatical mood that indicates commands, permission or agreement with a request. In Arabic, the jussive mood is used in negative past structures, in 1st or 3rd person commands, and in conditional structures.

[8] http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsARoot.htm

[9] http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsDualNumber.htm

Figure 13 shows a sample of the actual XML implementation of the fully specified model.

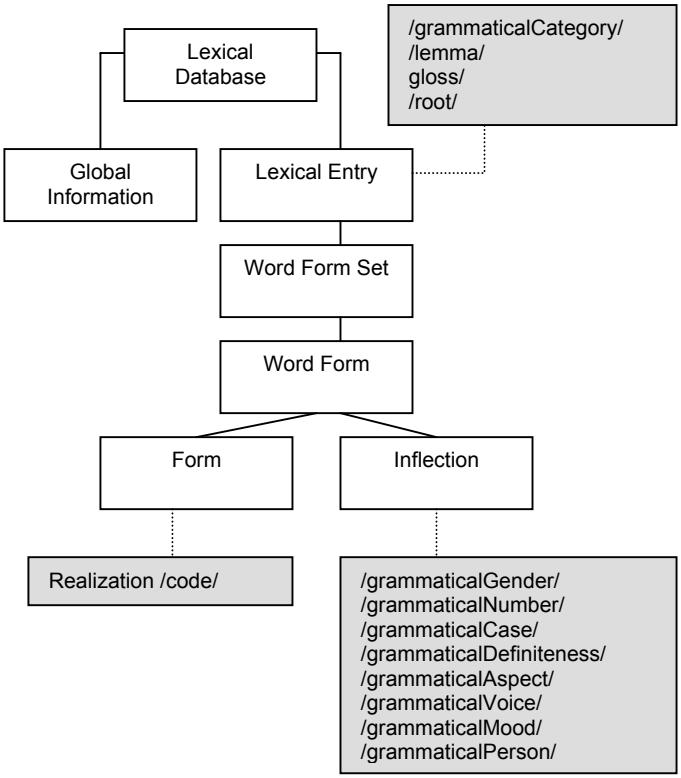

Figure 12. The current LMF specification for Arabic full form lexica

```
<lexicalEntry lemma="kataba" grammaticalCategory="verb" root="ktb" gloss="écrire">
    <wordFormSet>
        <wordForm>
            <form>
                <realization code=""Akrout_2005">katabtu</realization >
            </form>
            <inflection>
                <grammaticalAspect>perfect</grammaticalAspect>
                <grammaticalGender>masculine</grammaticalGender>
                <grammaticalPerson>firstPerson</grammaticalPerson>
                <grammaticalNumber>singular</grammaticalNumber>
                <grammaticalVoice>active</grammaticalVoice>
            </inflection>
        </wordForm>
        ...
        <wordForm>
            <form>
                <realization code="Akrout_2005">taktubâ</realization >
            </form>
            <inflection>
                <grammaticalAspect>imperfect</grammaticalAspect>
                <grammaticalGender>masculine</grammaticalGender>
                <grammaticalPerson>secondPerson</grammaticalPerson>
                <grammaticalNumber>dual</grammaticalNumber>
                <grammaticalMood>subjunctive</grammaticalMood>
            </inflection>
        </wordForm>
        ...
    </wordFormSet>
</lexicalEntry>
```

Figure 13. Example of a newly introduced data category in the DCR

IX. PERSPECTIVES

Obviously, the work presented here opens many perspectives for the development of future activities in the domains of Arabic language processing as well as the related standardization. As stated in the introduction, Arabic character recognition requires the development of wide coverage lexica containing the fully inflected forms to be recognized as acceptable in the OCR process. In General, such lexica will not be developed by one single group but result from the combination of the activities of the various communities that have the competences to compile such lexical data. This is even more important with respect to the complementary lexical resources which cover the dialectal variations in the Arabic world and the lexical specificities of technical communities. Networking those lexical resources can only be done if international standardization principles are applied in a systematic way, to allow unified queries to be remotely applied and results to be homogeneously combined.

This naturally leads us to the standardization agenda that can be derived from the LMF proposal as applied to the Arabic language as in this paper. First, it is necessary to provide an extensive documentation of the descriptors presented here. Such documentation should of course characterize the specific usage of features like number, tense or mood to the Arabic language, but also provide the right terminology to refer to those data categories, so that implementers will quickly find the entry they need in the DCR. For instance, the /jussive/ grammatical mood should be correctly localized to make sure that the names used to refer to it in French (*apocopé*), English (*jussive*) and of course Arabic (*al-majzûm*) are correctly recorded. The next step is then to carry out a similar activity on other lexical description levels, so that not only morphological descriptors are made available and standardized for Arabic, but also specific data categories that are needed for the description of syntactic and semantic constraints in the lexicon .

We should finally draw the attention on the necessity of a strong involvement of the Arabic speaking community in the process of defining and validating the set of data categories that can be considered as sufficient to represent lexical information for this language. As a matter of fact, the membership of ISO committee TC 37/SC 4 does not reflect the importance of Semitic languages in general and Arabic in particular in the linguistic world. This is the only way to make the kind of proposal presented in this paper widely usable by the community.

X. REFERENCES

[1] Akrout A. (2005). Modélisation d'un lexique flexionnel. Application à l'Arabe Classique. Mémoire de DEA, Université de Metz.

[2] Beesley K. (1998). Romanization, Transcription and Transliteration. The Document Company - Xerox. http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html

[3] Beesley K. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research : Status and Plans. ACL/EACL 2001, July 6th, Toulouse, France, 2001

[4] Bertagna F., Calzolari N., Lenci A., Zampolli A. (2000). ISLE – Computational Lexicons Working Group. The Multilingual ISLE Lexical Entry (MILE) : a discussion paper. http://www.tagmatica.fr/doc.htm

[5] Blachère, R., Gaudefroy-Démombynes, M. (1975). Grammaire de l'arabe classique. 3ème édition, G.P. Maisonneuve & Larose, Paris, France.

[6] Buckwalter T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium. University of Pennsylvania, 2002. LDC Catalog No.: LDC2002L49. http://www.ldc.upenn.edu/Catalog/docs/LDC2004L02/readme.txt

[7] Calzolari N., Monachini M. (1996). Multext – Common Specifications and Notation for Lexicon Encoding. http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX1.html

[8] Cavalli-Sforza, V., Soudi, A. , Mitamura, T. (2000). Arabic Morphology Generation using a Concatenative Strategy, in The Proceedings of NAACL-2000. http://acl.ldc.upenn.edu/A/A00/A00-2012.pdf

[9] Creissels D. (1995). Eléments de Syntaxe Générale. Presses Universitaires de France, Paris.

[10] de Saussure F. (1916). Cours de linguistique générale. Ré-édition 1997. Payot.

[11] Dichy J. (2000). Morphosyntactic Specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects. Proceedings of the *ACIDA' 2000* conference, Monastir (Tunisia), 22-24 March 2000, *Corpora and Natural Language Processing* vol., 55-60.

[12] EAGLES – Expert Advisory Group on Language Engineering Standards. Reports of the Computational Lexicons Working Group. http://www.ilc.cnr.it/EAGLES96/browse.html#wg2.

[13] Francopoulo G., George M. (2005). ISO/TC 37/SC 4 N130 Rev.7. Language resource management – Lexical markup framework (LMF). http://www.tagmatica.fr/doc.htm.

[14] Larcher P. (2003). Le système verbal de l'arabe clasique. Presses de l'Université de Provence, France.

[15] Ide N., Kilgarriff A., Romary L. (2000). A Formal Model of Dictionary Structure and Content. Proceedings of Euralex 2000. Stuttgart, 113-126.

[16] Ide N., Le Maitre J., Véronis J. (1995). Outline of a Model for Lexical Databases. Current Issues in Computational Linguistics: In Honour of Don Walker, Pisa, 283-320.

[17] Ide N., Romary L. (2004). A Registry of Standard Data Categories for Linguistic Annotation. Proceedings of LREC 2004, Lisbonne, Portugal.

[18] Ide N., Romary L. (2004). International standard for a linguistic annotation framework. *International Journal of Natural Language Engineering*, vol. 10 n° 3-4, p. 211-225.

[19] Ide N., Véronis, J. (1995). Encoding dictionaries. Computers and the Humanities, 29(2), 167-179.

[20] Khoja, S., Garside, R., Knowles, G. (2001). A tagset for the morphosyntactic tagging of Arabic. Corpus Linguistic Conference, Lancaster. http://archimedes.fas.harvard.edu/mdh/arabic/CL2001.pdf

[21] Mel'cuk I. (1993). Cours de Morphologie Générale. (vol I). Presses de l'Université de Montréal et CNRS Editions.

[22] Mohamed Maamouri and Ann Bies (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. Workshop on Computational Approaches to Arabic Script-based Languages, COLING.

[23] PAROLE – Report on the moprhological layer. http://www.ub.es/gilcub/SIMPLE/reports/parole/parole_morph/paromor_2.html#2.6

[24] Neyreneuf M., Al-Hakak G. (1996). Grammaire active de l'Arabe littéral. Librairie générale française. Paris.

[25] Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete : from LMF to Morphalou. Workshop on Electronic Dictionaries, Coling 2004, Geneva, Switzerland. http://www.atilf.fr/perso/salmon-alt/

[26] Sawaie, M. (1999). La Crise de la Terminologie Arabe au XIXe Siècle, Introduction Historique Générale. 1ère édition, Damas, Institut Français de Damas et Maison de l'Occident Musulman, 99-114.

[27] Shivtiel S. (1986). Root-dictionaries or alphabetical dictionaries : a methodological dilemma. In Devenyi K. et al., Colloquium on Arabic Lexicology and Lexicography – CALL. Budapest.

[28] Tahir, Y., Chenfour, N. & Harti, M. (2004). Modélisation à objets d'une base de données morphologiques pour la langue arabe. JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès.

[29] Wehr H. (1952). Arabisches Wörterbuch für die Schriftsprache der Gegenwart. Erstauflage Leipzig 1952.