Aïda Khemakhem[1], Bilel Gargouri[1], Abdelhamid Abdelwahed[2]
[1]Laboratoire MIRACL, ISIM-SFAX
aida_khemakhem@yahoo.fr, Bilel.Gargouri@fsegs.rnu.tn
[2]LSCA, FLSH- Sfax
abdaws@yahoo.fr

LMF est-il convenable pour la langue arabe ?

I. INTRODUCTION

Les ressources lexicales sont très utilisées dans le domaine du Traitement Automatique des Langues (TAL). Outre la variation et la richesse des connaissances à modéliser, la construction des ressources lexicales (lexique sous forme de fichiers classiques, bases lexicales, …) est confrontée à d'autres difficultés liées, entre autres, à l'absence de format standard pour les structures décrites et à la variation du choix des catégories de données d'une ressource à l'autre. De ce fait, le développement de ressources lexicales réduites et appropriées à des projets particuliers a marqué le domaine du TAL, limitant ainsi les possibilités de réutilisation et de fusion. Ce constat concerne aussi bien les langues latines que l'arabe dont les produits lexicaux restent peu nombreux. La standardisation de la démarche d'élaboration et de représentation des ressources lexicales s'avère donc un besoin pressant en TAL.

Dans le courant des efforts déployés pour cette standardisation, un projet baptisé LMF (Lexical Markup Framework) est en cours de validation par l'un des comités de l'ISO (Organisation Internationale de Standardisation) sous le numéro 24613. LMF décrit une démarche qui part d'un modèle abstrait prenant en considération les caractéristiques de plusieurs langues (européennes, asiatiques et américaines) et couvrant tous les niveaux des langues (i.e., morphologique, syntaxique et sémantique) pour aboutir à une représentation concrète en XML. Bien que certaines illustrations pratiques de LMF ont déjà vu le jour (i.e., *Morphalou* pour le Français), la langue arabe n'est pas encore considérée.

Notre travail s'inscrit dans le cadre de la validation de LMF en tant que norme internationale. L'objectif étant double, tout d'abord, étudier les possibilités d'appliquer les principes et les modèles de LMF pour la langue arabe, ensuite, développer une base lexicale pour l'arabe conforme à la future norme internationale LMF.

Dans la première partie, nous présenterons la proposition LMF et ses différents éléments. Une discussion de la possibilité d'appliquer LMF pour l'arabe sera étalée dans la section d'après. Enfin, nous donnerons une idée sur la base lexicale de l'arabe qui est en cours de construction selon LMF.

II. LMF : LEXICAL MARKUP FRAMEWORK

Le projet LMF [3] est une initiative pour standardiser la représentation des ressources lexicales pour avoir une future norme ISO 24613 élaborée par le comité TC 37/ SC 4. Il se base sur un ensemble de normes : ISO 12620 pour la description des catégories de données, UML pour les représentations conceptuelles et XML pour l'implémentation.

LMF profite de l'expérience des autres travaux qui ne sont pas flexibles puisqu'ils imposent le modèle et son format de représentation. En effet, LMF propose la standardisation du niveau conceptuel sous forme d'un méta modèle qui couvre tous les niveaux de description linguistique. Ce méta modèle est composé d'un noyau et d'un ensemble d'extensions qui traitent, respectivement les niveaux morphologique, syntaxique et sémantique. Il est décoré par des catégories de données sélectionnées à partir du RCD (Registre de Catégories de Données) ou ajoutées par les utilisateurs.

En vue d'atteindre le maximum de flexibilité, LMF ne fixe pas le format de représentation (i.e., DTD), mais propose un format de représentation pivot en XML (Generic Mapping Tool : GMT) qui a un rôle important pour l'échange entre les bases lexicales.

Le noyau de LMF

Le modèle noyau de LMF est organisé selon une structure hiérarchique des classes UML suivantes :

- *Database* : la totalité de la ressource.
- *Lexicon* : un lexique d'une langue donnée.

- *LexicalEntry : le mot en langage courant. C'est l'unité élémentaire dans une base lexicale et qui* porte l'information de la partie du discours.
- *Form : les valeurs orthographiques et phonétiques des unités lexicales avec des spécifications* grammaticales.
- *Sense : les attributs qui décrivent le sens du mot.*

La Figure 1, donnée ci-après, présente le méta modèle noyau de LMF.

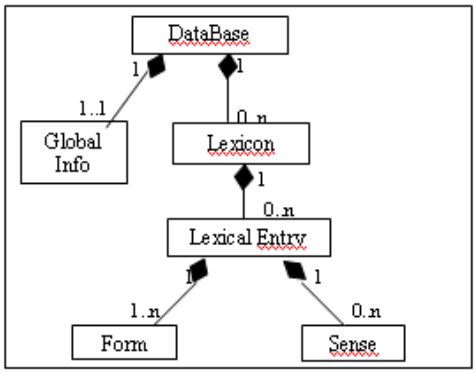

Fig. 1: *Le modèle noyau*

Dans cette figure, une ressource *Database* peut avoir zéro ou plusieurs *Lexicon* qui est composé de plusieurs *LexicalEntry*. Cette entrée lexicale peut avoir une ou plusieurs *Form* et zéro ou plusieurs *Sense*.

II.2 Les catégories de données

Les catégories de données sont des descripteurs linguistiques élémentaires qui permettent de décorer les classes du méta modèle (i.e., */root/*, */grammaticalGender/*). Leur normalisation suit des principes définis par la norme ISO 12620. Elles sont organisées dans un registre de catégories de données accessible en ligne (http://syntax.inist.fr/). Notons que la gestion des catégories de données est indépendante de leur utilisation.

II.3 L'extension morphologique

L'extension morphologique est la partie obligatoire pour la plus part des applications du TAL. Cette extension est traitée de deux manières différentes dans LMF. La première présente les formes fléchies. La deuxième utilise les paradigmes de flexion pour générer les formes fléchies. La Figure 3 présente les classes UML de cette extension. Il s'agit de :

- *LemmatisedForm :* une classe de spécification qui hérite toutes les propriétés de la classe *Form*.
- *InfletedForm* : une forme fléchie correspondant à une forme d'occurrence d'un /*lemmatisedForm*/.
- *InflactionalParadigm* : factorisation d'un ensemble de structures communes à un grand nombre de mots.

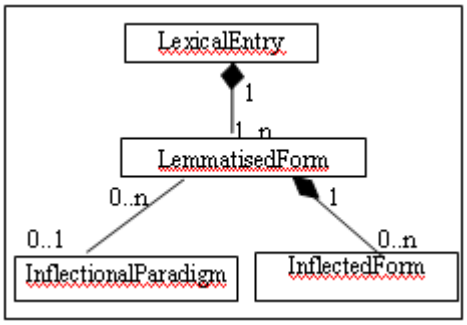

Fig. 3 : *Extension morphologique*

Dans cette figure, une *LexicaleEntry* peut avoir une ou plusieurs *LemmatisedForm* qui peuvent avoir zéro ou plusieurs formes fléchies. La cardinalité zéro est nécessaire pour les entrées invariables (adverbe, pronom personnel,...). La *LemmatisedForm* peut avoir zéro ou une seule *Inflectional Paradigm* qui peut être commune à plusieurs *LemmatisedForm.*

III. APPLICATION DE LMF POUR L'ARABE

Dans cette partie, nous essayons de tester l'applicabilité de LMF pour la langue arabe. LMF offre un cadre générique et simple surtout pour le niveau morphologique que nous allons discuter.

La langue arabe est une langue sémitique qui se caractérise par son aspect dérivationnel et une morphologie complexe. Parmi les caractéristiques propres à l'arabe que nous allons traiter en appliquant les principes de LMF, nous citons la racine, le schème, l'aspect et la vocalisation.

III.1 Le Noyau

En vue de maximiser la couverture des informations linguistiques, nous optons pour la représentation des mots voyellées, les non voyellées peuvent être générées. Cependant, nous allons éliminer la dernière voyelle des noms qui n'a pas de grande importance au niveau morphologique car sa présence engendre une explosion syntaxique de la base.

Pour garder l'aspect dérivationnel de l'arabe, nous ajoutons la racine *(root)* et le schème *(scheme)* dans l'entrée lexicale. La racine est une séquence de trois consonnes ou plus, qui est rattachée à tous les verbes et à la majorité des noms. Elle joue un rôle important pour déterminer l'origine sémantique et la nature morphologique (i.e., sein, hamzé, défectueux).

A chaque fois qu'on ajoute un schème à une racine quelconque nous obtenons un nouveau mot basé sur cette racine et sur son sens original. D'autant plus, dans le cas des verbes, nous serons capables de déterminer son processus de conjugaison et les cas d'utilisation de ce verbe dans la phrase. Le lien entre la racine et le schème est nécessaire pour toutes les extensions du TAL, ainsi, nous avons choisi de les mettre dans l'entrée lexicale du modèle noyau qui est commun à toutes les extensions.

La Figure 4 présente le noyau de LMF avec ses nouveaux attributs en gras.

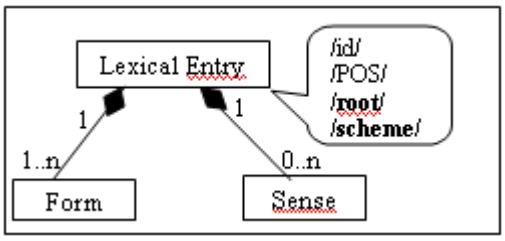

Fig. 2: *Structure noyau pour l'arabe*

La richesse de la langue arabe pose des difficultés pour trouver un consensus sur un classement unique des catégories grammaticales des mots (POS : Part Of Speech), bien que les travaux sur cette langue datent de plusieurs années. Ainsi, nous essayons de limiter ces catégories (i.e., Verbe, Particule, Nom, Nom propre, Nom de nombre) tout en spécifiant les catégories de données pour chaque POS.

III.2 Les catégories de données

En ce qui concerne les catégories de données, nous allons respecter les propositions qui ont été faites dans le cadre de LMF. Nous soulignons que certaines autres catégories peuvent être ajoutées.

La notion de temps est absente dans la conjugaison des verbes, par contre nous avons la notion d'aspect du verbe. Cet aspect comprend trois sortes : *accompli*, *inaccompli* et *impératif*. L'accompli indique que l'action est achevée. L'inaccompli indique que l'action est en train de se réaliser, sans être accomplie. L'impératif indique l'ordre ou la demande.

L'accompli a une seule modalité, alors que l'inaccompli en a trois : l'*indicatif*, le *subjonctif*, et l'*apocopé*. Par contre, l'impératif se rapproche de l'apocopé et ne se diffère que par l'absence des préfixes et des 1er et 3ème personnes [1]. Nous pouvons mentionner que la détermination du temps, dans l'arabe, ne se limite pas à

l'analyse du verbe seulement, mais il faut analyser toute la phrase. Ainsi, nous ajoutons la catégorie de donnée : *(grammaticalAspect)*.

Le genre, selon la norme ISO 12620, peut prendre les valeurs : *masculine, feminine* et *neuter*. Pour l'arabe, on n'utilise pas *neuter* mais nous ajoutons la valeur *commun* qui est nécessaire dans la conjugaison des verbes, car il y a des formes fléchies communes pour le masculin et le féminin tel que "كَتَبْتُ" qui indique la 1er personne singulière, masculine et féminine. Nous remarquons que *neuter* pour l'allemand est différente de *commun* pour l'arabe.

Pour le nombre, nous gardons les valeurs : *singular, plurial,* et *dual*, et nous ajoutons le pluriel brisé (*Brokenplural)* qui est utilisé avec les noms. Le pluriel brisé est un pluriel qui n'a pas de désinence. Il utilise la même racine que le nom au singulier, mais il est construit sur un autre schème emprunté au vaste fond nominal [1].

Reste à remarquer que nous gardons les catégories de données : *personne* et *voix* comme les autres langues. La personne prend les valeurs : *firstPerson, secondPerson* ou *thirdPerson*. La voix prend les valeurs : *actif* et *passif*.

Le tableau suivant récapitule les catégories de données pour la morphologie de l'arabe :

Cat. de données	Les valeurs
grammaticalAspect	accomplished, unaccomplished, imperative
grammaticalMood	indicatif, subjonctif, apocopate
grammaticalGender	masculine, feminine, common
grammaticalNumber	singular, plural, dual, Brokenplural
grammaticalPerson	firstPerson, secondPerson, thirdPerson
grammaticalVoice	actif, passif

Tableau 1: *Catégories de données de la morphologie de l'arabe*

III.3 Cas de l'extension morphologique

Pour la langue arabe, le nombre de formes fléchies est trop important pour qu'une gestion soit facile. A chaque fois que nous ajoutons un mot, il faut ajouter la totalité des formes fléchies associées, par exemple dans le cas d'un verbe quelconque, il faut propager (calculer ou décrire à la main) les 109 formes verbales [2]. Si la génération de la base est manuelle, le risque d'erreur sera trop élevé, le coût sera très important et le temps de réalisation et de vérification sera très long. Ceci nous amène à constater que les paradigmes de flexion jouent un rôle important pour la génération de la base surtout pour les verbes réguliers.

Le paradigme de flexion décrit les opérations, les arguments et les positions nécessaires pour chaque flexion. Dans la figure 5, nous présentons comment trouver la forme fléchie "كَتَبْتُ". Dans *InflectionalParadigm,* chaque forme fléchie a un *Morphological FeatureCombiner* qui fait le lien entre les catégories de données *MorphologicalFeature* et les opérations nécessaires pour trouver cette forme.

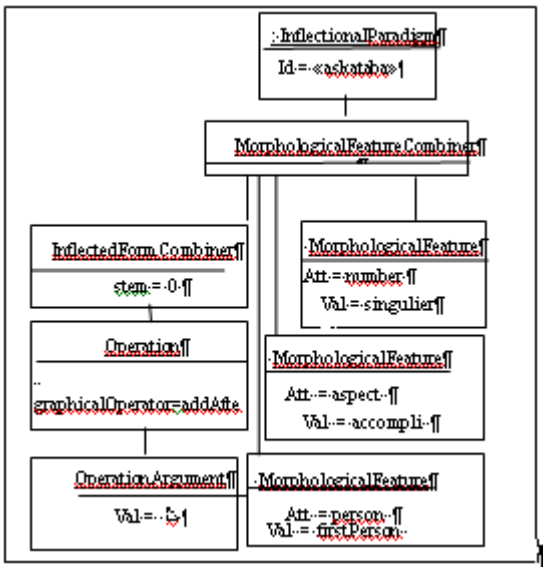

Fig. 3 : *Paradigme de flexion de "كَتَبْتُ"*

III.4 Synthèse

L'étude que nous avons réalisée a montré la possibilité d'appliquer LMF pour l'arabe moyennant certaines adaptations. Ainsi, nous avons procédé par :

- Garder le squelette de LMF
- Ajouter les attributs *root* et *scheme* au niveau de l'entrée lexicale
- Ajouter les catégories de données et les valeurs manquantes : *Aspect* (impératif, accompli et inaccompli), *Nombre* (pluriel brisé), *Genre* (commun) et *Mode* (apocopé).
- Utiliser les paradigmes de flexion pour automatiser la génération de la base.

L'application de LMF pour l'arabe aura le mérite de diminuer la complexité morphologique. En effet, LMF offre un méta modèle simple et donne la possibilité d'ajouter des catégories de données spécifiques. Ces propositions sont faites au niveau conceptuel plutôt qu'au niveau représentationnel pour plus de généricité. Ainsi, nous avons la possibilité de traiter chaque forme dérivée comme une entrée lexicale à part et d'ajouter la racine et le schème comme des attributs. En plus, LMF fourni des modèles permettant la restructuration de ressources lexicales existantes et la possibilité d'interopérabilité et d'intégration de ces ressources.

IV. CONCLUSION

L'étude de l'adaptation de LMF pour la langue arabe a soulevé l'absence de certaines catégories de données spécifiques à cette langue. Cette déficience n'est pas prise en considération ni par LMF ni par la norme ISO12620 (i.e., aspect, pluriel brisé).

En vue de réaliser une base lexicale pour la langue arabe conforme à LMF, nous avons déjà accompli la phase de spécification de la DTD et d'implémentation des interfaces qui facilitent la gestion (i.e., acquisition) et l'exploitation de la base.

Actuellement, nous entamons la phase d'acquisition des entrées lexicales d'ordre morphologique.

REFERENCES[1]

[1] Blachère R., Gaudefroy M., "Grammaire de l'arabe classique", Edition Maisonneuvre-Larose, Paris, 1975.

[1] La liste des références est réduite pour minimiser le nombre de pages

[2] Dichy J., Ammar S., ″Les verbes arabes″, Collection Bescherelle, Edition HATIER, Paris, 1999.

[3] ISO Working Draft for LMF, ISO/TC 37/SC 4 N130 Rev.9. Language resource management − LMF, 2006.