Modélisation à objets d'une base de données morphologique pour la langue arabe

Youssef Tahir (1), Noureddine Chenfour (2) et Mostafa Harti (3)
Département de Mathématiques et Informatique – Faculté des sciences DM
B. P. 1769 Fès - Maroc
(1) ytahir@fsdmfes.ac.ma
(2) chenfour@yahoo.fr
(3) mharti@fsdmfes.ac.ma

Résumé – Abstract

Nous présentons, dans ce papier, l'une des étapes principales de La réalisation d'un analyseur morphologique pour le texte arabe. Il s'agit de la conception et l'organisation d'une base de données linguistique pour la langue arabe. Cette base de données doit comporter toutes les primitives linguistiques de l'arabe telles que les particules, les schèmes des verbes, les schèmes des noms dérivés, les affixes et les noms particuliers, ainsi que toutes les concaténations possibles entre elles.

Nous remarquons, cependant, que la réalisation de cette base de données nécessite une organisation bien adaptée qui nous permettra d'exploiter convenablement ces données. Ainsi, nous proposons une organisation particulière basée sur les concepts de classes et d'objets. L'idée de base de cette organisation consiste à considérer chaque primitive morphologique comme un objet. Nous obtenons un ensemble d'objets morphologiques que nous organisons sous forme de classes dont chacune ayant ses propres caractéristiques morphologiques. Les classes morphologiques sont elles-mêmes organisées en se basant sur les deux concepts d'héritage et de composition dans plusieurs paquetages. Ce papier présente, donc, une description détaillée de cette organisation qui nous a permis de réaliser une base de données quasiment exhaustive des primitives morphologiques de la langue arabe avec un nombre d'entrées très réduit.

In this work we are interested by the conception of a morphological database for Arabic language, which could be used for the realization of different Arabic language processing systems. Therefore, it must contain all Arabic language morphological primitives like affixes, nouns, verbs and particles and also all possible concatenations of them (which we called morphological rules).

The realization of such a database requires note only linguistic data collection, but also an adapted organization of the database in order to be exploited appropriately. Hence, we suggest a particular organization based on the two concepts of class and object. Indeed, we considered that every morphological primitive is an object. These objects are gathered in several classes according to the set of affixes that each one accepts. We have also arranged all classes in several packages according to the type of the morphological elements that each one

represents. We present so, by this paper a detailed description of the architecture of our database organization based on the UML diagram modeling. This organization allowed us to obtain a nearly-exhaustive database with a small number of components.

Mots Clés – Keywords

Morphologie, base de données morphologique, modèle objet, descripteurs morphologiques.
Morphology, morphological database, object modeling, morphological descriptors.

1 Introduction

Notre travail de réalisation d'un analyseur morphologique pour la langue arabe se situe dans le domaine de la linguistique calculatoire. Il a pour objectif la mise en place d'un système informatique capable de fournir une représentation morphologique pour chaque mot constituant un texte arabe en entrée. Il est clair qu'un tel système repose sur l'utilisation et la manipulation des données linguistiques de la langue arabe qui représentent sa base de connaissances. Pour cela, nous sommes intéressés dans ce travail par la conception d'une base de données linguistique comportant toutes les primitives morphologiques de la langue arabe.

La réalisation d'une telle base de données nécessite non seulement la collection de données linguistiques mais aussi leur organisation suivant des techniques bien adaptées afin de faciliter leur exploitation. Nous proposons alors une organisation particulière fondée sur les deux concepts essentiels de la conception à objets : le concept de classes et celui d'objets. L'idée principale de cette organisation est de représenter toutes les primitives morphologiques de la langue arabe sous forme d'objets. Ces objets sont organisés dans un ensemble de classes en se basant sur la règle suivante :

« *Les éléments d'une même classe acceptent les mêmes préfixes et les mêmes suffixes ou bien ils sont suffixes ou préfixes des mêmes classes.* »

Nous avons aussi organisé toutes ces classes dans un ensemble de paquetages selon le type des primitives morphologiques que chaque classe représente et en se basant sur deux concepts essentiels : le concept d'héritage et celui de composition ou agrégation entre classes. Ainsi, nous présentons dans ce papier l'architecture de notre base de données linguistique fondée sur les diagrammes de modélisation UML.

Dans ce sens, nous proposons l'utilisation d'un nouveau langage de représentation des connaissances morphologiques qui s'adapte mieux à nos besoins. Il s'agit du langage de spécification des règles morphologiques JMODEL : Java based MOrphology DEfinition Language (Chenfour N., 2003).

Nous remarquons également que pour la collection des données linguistiques nous nous sommes basés sur un ensemble de travaux effectués dans le domaine de la morphologie arabe (Abdallah Ben Akil B., 1974 ; Ben Hachem Elansari J., 1979; Eljazim A. et Amin M., 1988; Hassan Ezzayat A. et al., 1989; Qubbiche A., 1979).

2 Conception de la base de données

Comme nous l'avons mentionné dans le paragraphe précédent, notre base de données se compose d'un ensemble de classes; chacune représentant une famille de données linguistiques de même nature et de mêmes caractéristiques morphologiques. Ainsi, nous introduisons la notion de *classes morphologiques* (CLM) qui doivent contenir toutes les *composantes morphologiques* (CMM) de la langue arabe telles que les schèmes des verbes, les schèmes des noms dérivés, les particules, les affixes et les noms particuliers. Chacune de ces CMM est définie par un ensemble de caractéristiques morphologiques indiquant le genre, le nombre, la valeur grammaticale, etc. Nous appelons ces caractéristiques des *descripteurs morphologiques* (DM). Afin de bien illustrer la correspondance entre chaque composante morphologique et son ensemble de descripteurs morphologiques nous définissons un deuxième type de classes appelées *classes de propriétés morphologiques* (CLP). Chacune de ces classes contient des descripteurs morphologiques de même nature.

Exemple :

La classe propriété "*Number*" est une classe qui contient tous les descripteurs morphologiques indiquant la propriété nombre modélisée à l'aide du langage JMODEL.

```
property class Number {
        NSg ;   // singulier
        NDl ;   // dual
        NPl ;    // pluriel
}
```

Nous définissons également les *classes de règles* (CLR) qui représentent toutes les concaténations possibles entre les différentes composantes morphologiques définies dans les classes morphologiques. Ainsi chaque classe de règles représente une famille de mots arabes complets. Le dernier type de classes que nous définissons constitue les *classes d'union* (CLU), qui visent à attribuer le même alias à plusieurs classes ou bien parties de classes. L'objectif de l'utilisation de ce type de classes est, d'une part, d'éviter la redondance, et d'autre part, de réduire le nombre des règles de concaténation (Chenfour N., 2003).

Nous avons aussi organisé toutes ces classes en plusieurs paquetages selon le type de chacune. Ainsi, nous avons obtenu quatre paquetages qui contiennent toutes les classes morphologiques : le paquetage des affixes, le paquetage des particules, le paquetage des verbes et le paquetage des noms. Ces quatres paquetages dépendent du paquetage des propriétés qui regroupe toutes les classes de propriétés. La dernière partie de notre base de données est représentée par le paquetage des règles qui contient toutes les classes des règles de concaténations morphologiques (voir la figure 1). Nous remarquons que ce dernier paquetage dépend de tous les autres paquetages.

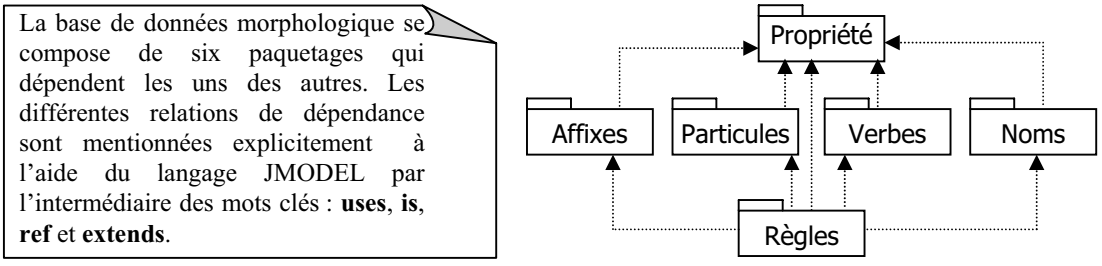

Figure 1: Présentation générale de notre base de données linguistique

2.1 Paquetage des verbes

Le système morphologique des verbes arabes est très particulier. En effet, il est d'une part robuste et totalement régulier dans le cas des verbes sains "الأفعال الصّحيحة" (*/ealeafcAlu eaSSaHIHatu/*)[1] en se basant sur la représentation «*racines – schèmes*», et d'autre part, irrégulier en général dans le cas des verbes non-sains "الأفعال المعتلّة " (*/ealeafcAlu ealmuctallatu/*) et des verbes incomplets "الأفعال النّاقصة" (*/ealeafcAl eannAqiSatu/*) (une famille de verbes se distinguant par la syntaxe et par l'impossibilité de prendre certaines formes de conjugaison, d'où la raison de la nomination "incomplet"). Une étude fondamentale de ce système nous a permis d'établir la représentation « *radical – affixes de conjugaison* » des verbes arabes qui va rendre régulières toutes les règles de conjugaison même dans le cas des verbes non-sains. Cette représentation est basée sur la décomposition de la forme conjuguée d'un verbe arabe en deux parties complémentaires. La première, dite radical, est invariante par rapport au genre, nombre et personne. La deuxième partie, que nous avons appelée affixes de conjugaison (Tahir Y. et al., 2003), dépend du genre, du nombre et de la personne ainsi que du temps de conjugaison (elle est indépendante du schème du verbe à l'infinitif). Elle est représentée par un suffixe dans le cas du passé "الماضي" (*/ealmADI/*) et de l'impératif "الأمر" (*/ealeamr/*), et par un couple « *préfixe – suffixe* » dans le cas du présent "المضارع" (*/ealmuDAric/*). Cette représentation nous permet de rendre la morphologie des verbes arabes une morphologie totalement concaténative.

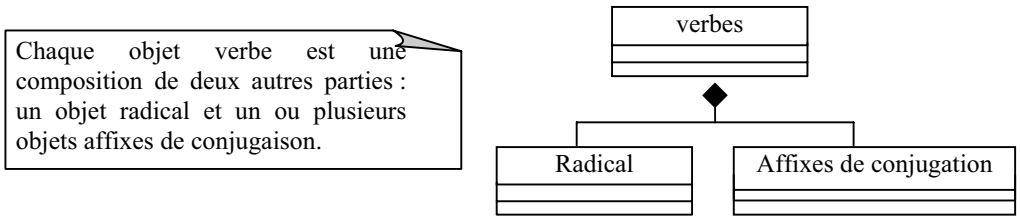

Figure 2 : La représentation (radical - affixes de conjugaison) des verbes arabes

La modélisation de cette structure a nécessité l'introduction d'un paquetage des verbes, qui contient les radicaux de tous les verbes avec une classification en cinq sous paquetages (voir la figure 3) :

Trois principaux paquetages:

- Le paquetage des verbes sains ("الأفعال الصّحيحة") qui contient tous les radicaux des verbes sains (105 radicaux). Ces radicaux sont distribués sur cinq classes morphologiques selon le temps de conjugaison de chacun.
- Le paquetage des verbes non–sains ("الأفعال المعتلّة") qui contient tous les radicaux des verbes non-sains (137 radicaux). Ce paquetage contient également plusieurs classes morphologiques constituées selon le temps de conjugaison et la nature de défectuosité (*/mi~Al/*, */nAqiS/*, */eajwaf/*, etc.) de chaque radical.

[1]Pour une simplicité d'écriture et de manipulation directe par clavier des codes phonétiques nous avons adopté un codage approprié que nous présentons dans le tableau suivant :

ي	و	هـ	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	أ
y	w	h	n	m	l	k	q	f	g	c	V	T	D	S	^	s	z	r	v	d	x	H	j	~	t	b	e

- Le paquetage des verbes incomplets ("الأفعال النّاقصة") qui contient tous les radicaux des verbes incomplets. Ces radicaux sont regroupés en plusieurs classes selon le temps de conjugaison et la nature de chaque radical (verbes de la famille *kAna* (كان), verbes de la famille *kAda* (كاد) et verbes de la famille *Zanna* (ظنّ)).

Avec deux autres paquetages complémentaires :

- Le paquetage des schèmes d'origine : il contient les classes d'origine de chaque radical utilisé dans les trois premiers paquetages. Chacune de ces classes nous permet d'identifier le schème d'origine pour un radical donné en employant un code intermédiaire.
- Le paquetage des unions des radicaux : Afin de réduire le nombre de règles morphologiques de conception des verbes, nous proposons le paquetage des unions des radicaux qui contient un ensemble de classes d'union. Chacune de ces classes regroupe tous les radicaux qui acceptent les mêmes affixes.

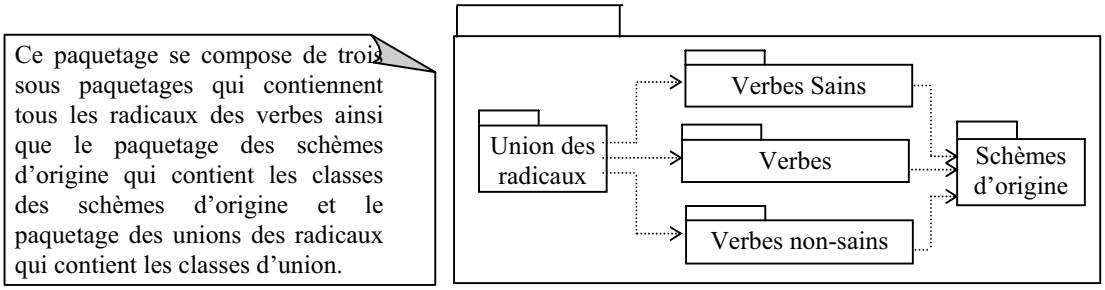

Figure 3: Architecture du paquetage des verbes

2.2 Paquetage des noms

Le système morphologique des noms arabes distingue entre deux catégories des noms. La première catégorie regroupe tous les noms dérivés (noms obtenus en employant les règles de dérivation). Un nom dérivé se caractérise complètement par sa représentation morphologique « racine – schème ». La seconde catégorie regroupe tous les noms particuliers qui ne respectent aucune règle de dérivation. Il est évident que l'étude de cette catégorie exige un lexique qui contient tous les noms particuliers. On distingue également entre les noms ayant la propriété "معرب" /mucrab/ qui possèdent généralement six formes et ceux ayant la propriété "مبني" /mabniyy/ qui possèdent une seule forme. Ces éléments doivent appartenir à une classe finale (qui représente un mot arabe complet). Pour le cas des noms avec la propriété "مبني" nous proposons la représentation de la partie invariante seule. Celle-ci accepte nécessairement l'utilisation d'une famille particulière de suffixes pour construire un nom arabe complet.

Afin de représenter ce système morphologique, nous proposons un paquetage des noms arabes qui se compose de deux sous paquetages : le paquetage des noms dérivés et le paquetage des noms particuliers.

- Paquetage des noms dérivés : Il contient 184 schèmes de noms dérivés que nous avons pus extraire à partir de la littérature arabe et que nous avons organisé en huit classes. Une d'entre elles est une classe abstraite appelée "*nomD*", qui représente la classe mère de toutes les autres classes de ce paquetage. Ces dernières classes sont caractérisées par l'acceptation ou le refus des suffixes de dual, suffixes de pluriels et suffixes de passage au féminin (voir la figure 4).

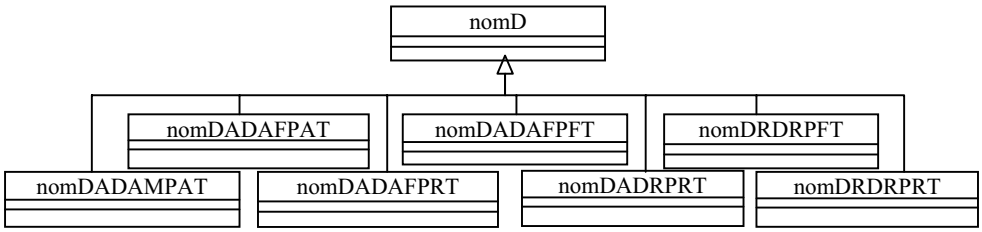

Figure 4: Architecture du paquetage des noms dérivés

- Paquetages des noms particuliers : Ce paquetage contient tous les noms particuliers de l'arabe. Il se Compose de deux sous paquetages; le premier est réservé aux noms particuliers tels que "أسماء الإشارة" /*easmAeu ealeichArati*/, "أسماء الموصول" /*easmAeu ealmawsUli*/, "الضمائر المنفصلة" /*eaDDamAeiru ealmunfaSilati*/ et les nombres. Ce paquetage contient 85 composantes morphologiques ayant toutes la propriété morphologique /*mabniy*/. Il se compose de huit classes dont trois sont abstraites (voir la figure 5). Le deuxième sous paquetage contient tous les autres noms particuliers qui n'appartiennent à aucune des classes du premier sous paquetage. Nous remarquons que ce paquetage est en cours de construction.

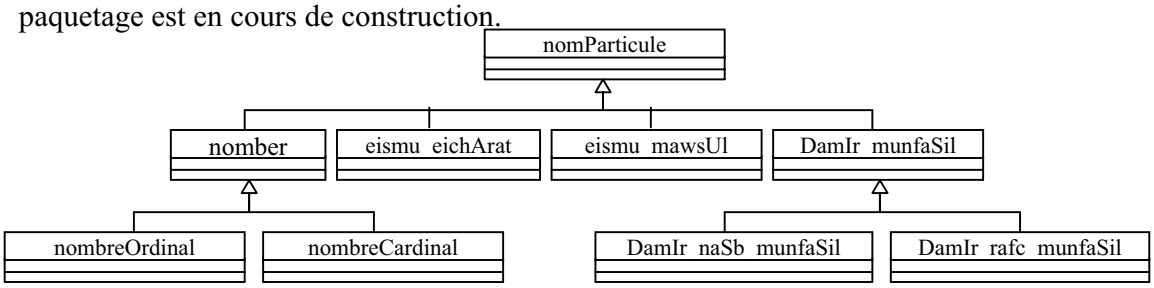

Figure 5: Architecture du paquetage des noms dérivés

2.3 Paquetage des particules

En langue arabe, on distingue entre deux catégories de particules. La première contient toutes les particules dont la forme est constante et que l'on appelle "مبنية حروف" /hurUfun mabniyatun/. La deuxième catégorie, appelée "حروف معربة" /hurUfun mucrabatun/, contient toutes les particules dont la dernière voyelle change pour donner généralement naissance à six formes possibles de la particule selon sa fonction dans la phrase.

La modélisation des classes de particules utilise donc le même principe utilisé pour la représentation des noms particuliers. On obtient un paquetage à 186 particules qui sont organisées en plusieurs classes selon l'ensemble des affixes que chacune accepte. La classe mère de toutes ces classes est la classe "particule" (voir la figure 6).

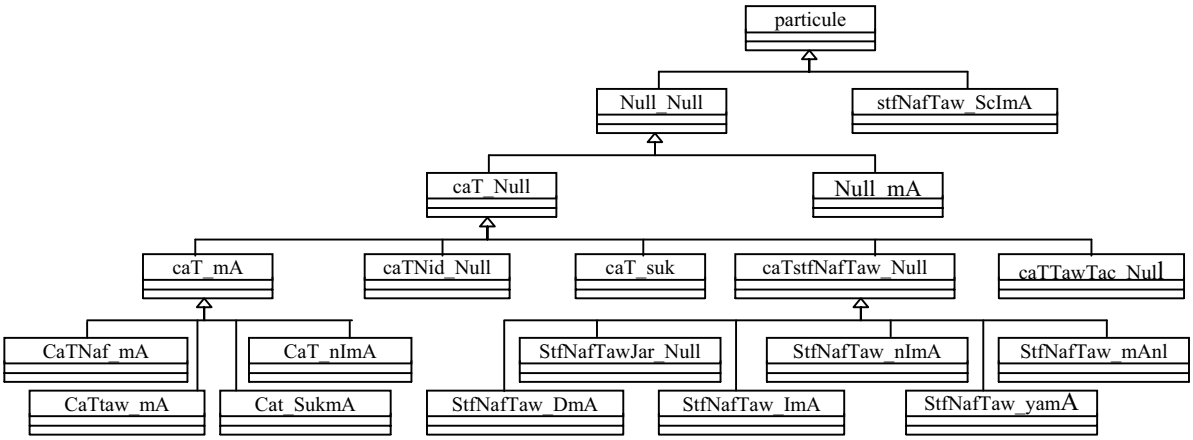

Figure 6: Représentation du paquetage des particules

2.4 Paquetage des affixes

Ce paquetage contient tous les affixes (préfixes et suffixes) connus de la littérature arabe y compris les affixes de conjugaison. Il se compose de trois sous paquetages principaux: le paquetage des préfixes, le paquetage des suffixes, et le paquetage des affixes de conjugaison. Ce dernier contient, également, trois sous paquetages : le paquetage des préfixes du passé "الماضي" (*/ealmADI/)*, le paquetage des affixes du présent ("المضارع") */ealmuDAric/*, et le paquetage des préfixes de l'impératif "الأمر" (*/ealeamr/)* (voir la figure 7).

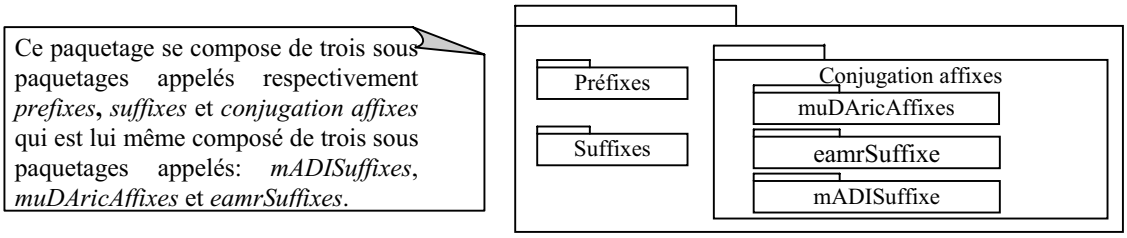

Figure 7: Représentation du paquetage des affixes

- Paquetage des préfixes : Ce paquetage contient 23 préfixes et il se compose de plusieurs classes. Trois d'entre elles sont des classes abstraites : la classe "*Prefixe*", qui est la classe mère de toutes les autres classes, la classe "*PrefixeDef*" qui est la classe mère des deux classes des préfixes de définition, et la classe "*PrefixeHJar*" qui est la classe mère des deux classes des préfixes "حروف الجر" /*hurUf ealjar*/. Ce paquetage contient également 11 classes morphologiques (voir la figure 8).

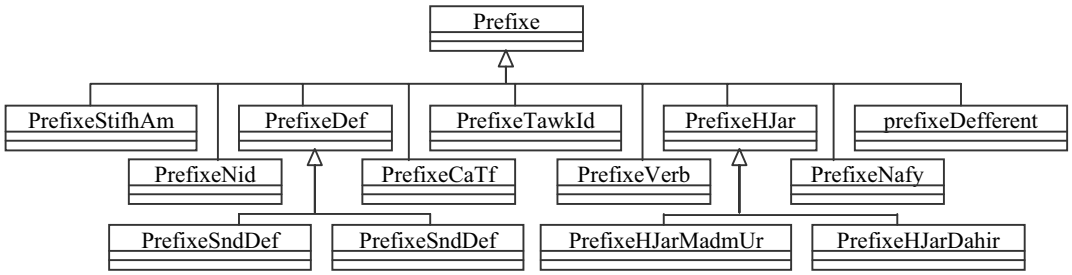

Figure 8: Représentation du paquetage des préfixes

- Paquetage des suffixes : Ce paquetage contient 54 suffixes repartis en quinze classes. Quatre d'entre elles sont des classes abstraites : la classe "*Suffixe*" qui est la classe mère de toutes les classes de ce paquetage, la classe "*CasSuffixe*" qui représente tous les suffixes de cas, la classe "*DamirMuttaSil*" qui est la classe mère des trois classes regroupant les suffixes de "الضمير المتصل" /*eaDDamIr ealmuttaSil*/, et la classe "*NumberSuffixe*" qui représente les suffixes indiquant le nombre (Voir la figure 9).

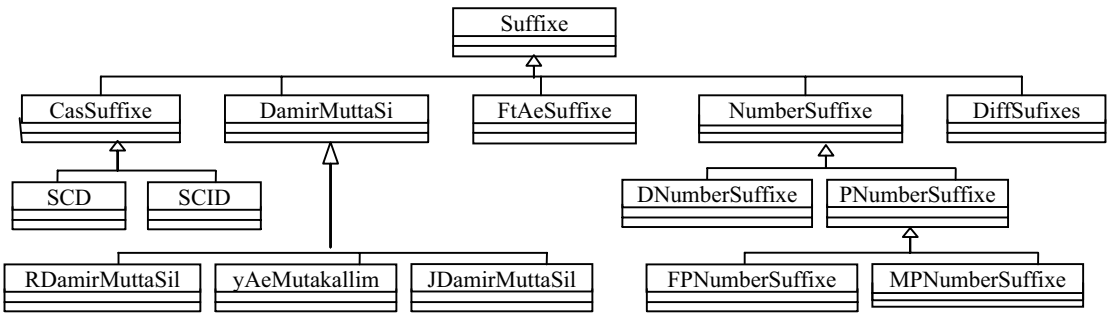

Figure 9: Représentation du paquetage de suffixes

- Paquetage des affixes de conjugaison : Ce paquetage contient 4 préfixes et 69 suffixes de conjugaison. Il se compose de trois sous paquetages contenant chacun les classes des affixes de conjugaison relatives à l'un des trois temps principaux connus dans la littérature arabe. Ainsi, le paquetage "*mADIsuffixe*" contient 25 suffixes de conjugaison du passé (/*ealmADI*/) regroupés tous dans une seule classe morphologique appelée "*MADISuffixe*". Cette classe se compose de quatre partitions appelées respectivement "*regularG1*", "*regularG2*", "*manqUSwAwI*", "*manqUSyAeI*". Chacune de ces partitions contient les suffixes de /*ealmADI*/ adaptés pour une catégorie particulière de radicaux. Le deuxième sous paquetage, appelé "*muDAricAffixes*", contient 4 préfixes et 22 suffixes du présent (/*ealmuDAric*/), qui sont regroupés dans deux classes morphologiques : la première représente les suffixes de /*ealmuDAric*/, appelée "*muDAricSuffixe*", et l'autre, dite "*MuDAricPrefixe*", représente les préfixes de /*ealmuDAric*/. Le dernier sous paquetage, appelé "*eamrSuffixes*", contient tous les 22 suffixes de l'impératif (/*ealeamr*/). Il se compose de quatre classes dont la première, appelée "*RegularEamrSuffixe*", est adaptée aux radicaux des verbes sains. Les autres classes (appelées respectivement "*manqUSYAeIEamrSuffixe*", "*manqUSWAwIEamrSuffixe*" et "*manqUSEamrSuffixe*") sont adaptées pour les radicaux des verbes non-sains particuliers.

2.5 Paquetage des propriétés

Ce paquetage contient toutes les classes de propriétés morphologiques connues dans la langue arabe et il se compose de quatre sous paquetages (voir figure 10). Le premier sous paquetage, appelé "*VpptPackage*", contient toutes les classes des propriétés morphologiques employées pour caractériser les verbes arabes. Le deuxième, dit "*PpptPackage*", regroupe toutes les classes de propriétés morphologiques employées pour caractériser les particules. Le troisième sous paquetage, dit "*NpptPackage*" contient toutes les classes de propriétés morphologiques caractérisant les noms arabes. Le dernier sous paquetages, dit "*GpptPackage*", contient toutes les propriétés utilisées pour la caractérisation des trois types des mots arabes (verbe, nom, et particule). D'autre part, nous pouvons distinguer entre deux types de classes de propriétés morphologiques selon la possibilité pour une même composante morphologique d'accepter plus de deux descripteurs morphologiques de la même classe.

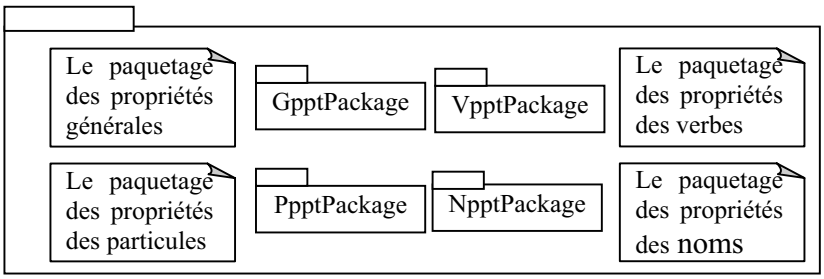

Figure 10: Architecture du paquetage de propriétés

2.6 Paquetage des règles

La dernière partie de notre base de données linguistique est le paquetage des règles. Ce paquetage regroupe toutes les règles de concaténation possibles entre toutes les composantes morphologiques définies dans les paquetages précédents. L'objectif de chaque règle de concaténation est de construire un nouveau mot à partir d'un ensemble des composantes morphologiques. Une autre possibilité des classes de règles est d'affecter (dans certains cas) de nouveaux descripteurs morphologiques (descripteurs qui n'appartient pas aux ensembles

des descripteurs morphologiques des composantes de la concaténation) au résultat de concaténation.

Exemple :

- eal + mufaccil + Ani {GMa} — *Cette règle vise non seulement à construire le mot "ealmufaccilAni" mais aussi à affecter également le descripteur morphologique "GMa " au résultat de concaténation*
- *eal + mufaccil + at + Ani {}* — *La composante morphologique "at" possède le descripteur morphologique de genre féminin "GFe". La règle ne peut donc affecter aucun descripteur morphologique de type genre au résultat de concaténation.*

Ces règles sont organisées dans plusieurs classes selon le type de concaténation de chacune ainsi que la nature de ses composantes.

3 Réduction (f,c,l)

Dans la langue arabe, les verbes et les noms dérivés sont complètement caractérisés par la représentation « *racine–schème* » (Fassi Fehri A., 1982, Soudi A. et al., 2001 et Ennaji M. et Sadiqi F., 1992). L'utilisation de cette représentation permet d'appliquer une même règle morphologique à tout un ensemble de verbes ou de noms dérivés ayant le même schème. Afin de réduire le nombre d'entrées de notre base de données, nous proposons de bénéficier de cette représentation pour le cas des verbes sains, verbes non-sains et noms dérivés, ce qu'on ne peut pas faire dans le cas des verbes incomplets, les noms particuliers, les particules et les affixes. En effet, chacune de ces quatres dernières catégories nécessite un lexique qui contient tous les éléments de la classe. Ainsi nous présentons par le tableau suivant le nombre des éléments de chaque paquetage de notre base de données linguistique.

	Nombre de classes morphologiques	Nombre de classes de propriétés	Nombre de composantes morphologiques	Nombre de descripteurs morphologiques
Paquetage des verbes	59	5	496	8
Paquetage des noms	15	11	184	46
Paquetage des particules	6	5	64	186
Paquetage des préfixes	10	3	23	5
Paquetage des suffixes	14	9	54	5
Paquetage des propriétés	0	15	0	25
Paquetage des exceptions	6	0	140	0
Total	104	48	**960**	275

Figure 11 : Un tableau la représentation statistique de notre base de données morphologique

4 Conclusion

Le travail que nous avons présenté constitue une étape préliminaire pour le traitement automatique de la morphologie arabe. Notre base de données morphologique peut être exploitée pour la réalisation des différents systèmes de traitement automatique de la langue arabe tels que les systèmes de traduction automatique, les systèmes de synthèse de la parole à partir du texte, les systèmes de classification automatique de texte, etc.

La réalisation de cette base de données a nécessité une étude approfondie des différentes particularités morphologiques de la langue arabe. Ce travail nous a permis, ainsi, d'élaborer une bibliothèque de toutes les primitives morphologiques de la langue arabe.

Le formalisme orienté objets que nous avons proposé nous a permis d'implémenter notre base de données en utilisant, pour la première fois, un nouveau langage de représentation des connaissances morphologiques JMODEL, mieux adapté à nos besoin. D'autre part la représentation « *racine – schème* » et la représentation « *radical – affixes de conjugaison* » que nous avons proposées ainsi que les concepts d'héritage et de composition nous ont permis de réaliser une base de données morphologique pour la langue arabe avec un nombre d'entrées très réduit.

Enfin, cette modélisation morphologique nous permettra de générer automatiquement un automate fini déterministe qui engendre toute la morphologie arabe.

Références

ABDALAH BEN AKIL B. (1974), شرح ابن عقيل على ألفية ابن مالك، دار الفكر، بيروت، لبنان

BEN HACHEM ELANSARI J. (1979), مغني اللبيب عن كتب الأعاريب، الطبعة الخامسة، دار الفكر، بيروت– لبنان

CHENFOUR N. (2003), Réalisation d'un Langage de définition de la morphologie JMODEL, *JERM2003,* Fès - Maroc.

ELJAZIM A., AMIN M. (1988), النحو الواضح في قواعد اللغة العربية، الطبعة الثامنة عشر، دار المعارف بيروت، لبنان.

ENNAJI M., SADIQI F. (1992), *Introduction to Modern Linguistics,* Afrique orients.

FASSI FEHRI A. (1982), *Linguistique arabe forme et interprétation,* Faculté des Lettres et des Sciences Humain, Rabat.

HASSAN EZZAYAT A., ABDELKADER H., MUSTAPHA I., ANNAJAR M. (1989), المعجم الوسيط، دار الدعوة، استانبول.

QUBBICHE A. (1979), الكامل في النحو والصرف والإعراب، دار الجيل، بيروت، لبنان.

SOUDI A., CAVALI-SFORZA V., JAMARI A. (2001), A Computational Lexeme-Based Treatment of Arabic Morphology, *Arabic NPL workshop at ACL/EACL,* Toulouse - France.

TAHIR Y. CHENFOUR N., HARTI M. (2003), Realization of a morphological analyzer for Arabic language text, *Moroccan – American workshop on the information technology*, march 17-19.