Conception d'un jeu de ressources libres pour le TAL arabe sous Unitex

CONFERENCE PAPER · MAY 2013

DOI: 10.13140/RG.2.1.4088.0485

4 AUTHORS, INCLUDING:

Noureddine Doumi

Taher Moulay University of Saida

8 PUBLICATIONS **1** CITATION

SEE PROFILE

Ahmed Lehireche

University of Sidi-Bel-Abbes

43 PUBLICATIONS **67** CITATIONS

SEE PROFILE

Denis Maurel

University of Tours

131 PUBLICATIONS **200** CITATIONS

SEE PROFILE

Conception d'un jeu de ressources libres pour le TAL arabe sous Unitex

Noureddine DOUMI*, Ahmed LEHIRECHE**, Denis MAUREL***, Moussa ALI CHERIF**

**Université de Saïda, Laboratoire EEDIS - UDL SBA Algérie,*

***Université UDL-SBA, laboratoire EEDIS-UDL SBA, Algérie*

****Université François Rabelais Tours, Laboratoire d'informatique, France*

noureddine.doumi@univ-saida.dz, elhir@univ-sba.dz, denis.maurel@univ-tours.fr et malicherif@gmail.com

RÉSUMÉ. *L'objectif de cet article est la description d'un module arabe libre sur la plateforme Unitex. La compilation d'un jeu de ressources libres au format Unitex a nécessité en premier lieu de proposer un corpus de test, de choisir un jeu d'étiquettes bien adapté et de construire des dictionnaires respectant le formalisme DELA de LADL.*

Nous décrivons chacun des points ci-dessus, notamment la construction des dictionnaires, pour lesquels nous avons conçu un algorithme de génération automatique de graphes pour la flexion des verbes et un autre pour celle des noms. Nous utilisons les principes de la morphologie lexèmatique et définissons pour chaque lexème un ensemble de thèmes servant de base aux différentes flexions. Pour les verbes, nous utilisons cinq thèmes donnés par l'utilisateur; les graphes obtenus engendrent ensuite jusqu'à 264 formes verbales fléchies. Pour les noms communs et adjectifs, nous n'utilisons qu'un seul thème; les graphes obtenus génèrent 63 formes fléchies.

ABSTRACT. *This paper aims to describe the process of building a free Arabic package for the Unitex framework: we proposed a test corpus, we chose a tag set suited to this task and we build dictionaries respecting the LADL DELA format.*

We describe each of the above particularly the building of dictionaries, for which we designed algorithms for automatic generation of verb and noun inflection graphs. We use the word-based inflection foundations and we define for each lexeme a set of themes. For the verbs, we use five themes given by the user and the graphs generate up to 264 inflected verbal forms; for the nouns and adjectives we use one or at most two themes and the produced graphs generate 63 inflected forms.

MOTS-CLÉS: *TAL arabe, morphologie, jeu d'étiquettes, dictionnaire DELA, Unitex, graphe de flexion.*

KEYWORDS: *Arabic NLP, morphology, tag set, DELA dictionary, Unitex, inflection graph.*

1. Introduction

La langue arabe appartient à la famille des langues sémitiques, constitués principalement par l'arabe, l'amharique et l'hébreu. Les langues sémitiques sont caractérisées par 1- un lexique construit à partir de racines trilitères et quadrilitères, 2- d'un système d'écriture de gauche vers la droite et 3- d'un alphabet de type Abjed.

L'arabe est la première langue sémitique en nombre de locuteurs plus de 360 millions et la 7ème langue mondiale en nombre d'utilisateurs dans l'Internet[1].

Utilisée par plus de 22 pays dans le monde, la langue arabe est parlée en plusieurs dialectes mais officiellement il y a qu'une seule langue arabe utilisée par les organisations étatiques et dans les médias officiels du monde arabe.

[1] Internet World Statistique à l'url http://www.internetworldstats.com (consulté en Fev 2013)

2. Arabe et ses dialectes

La langue arabe officielle est divisé en Arabe Classique (AC) et Arabe Standard Moderne (ASM), la première est la langue des textes saints de l'islam : le Coran et le Hadith et du patrimoine culturel, littéraire et scientifique de la civilisation arabo-musulmane. Cependant l'ASM est la langue officielle du monde arabe actuellement, elle est utilisée dans l'enseignement et dans les médias. La différence entre l'AC et l'ASM consiste dans le lexique, l'ASM utilise un lexique plus grand et plus moderne que celui de l'AC (Khoja 2001) et du coté de la grammaire l'ASM a abandonné l'utilisation de quelques formes compliquées de l'ancienne grammaire.

L'AC et l'ASM ont en commun d'être des langues écrites alors que les dialectes de l'arabe sont seulement parlés ; ils sont classés en sept groupes (Habash, 2010) : arabe égyptien (EGY), arabe levantin (LEV), arabe du golfe (GLF), arabe magrébin (MAG), arabe iraquien (IRQ), arabe yéménite (YEM) et arabe maltais (MLT).

3. La description du jeu de ressources

Unitex est une plateforme open source de traitement automatique des langues, développée à l'université Paris-Est Marne-La-Valée (Paumier, 2009). Cet outil est fondé sur la technologie des automates à nombre fini d'états (FST : Finite State Transducers) ; ainsi on peut exprimer les règles morphologiques, syntaxiques et même sémantiques d'une langue par des transducteurs, des réseaux de transitions récursifs, des grammaires algébriques et des réseaux de transitions augmentés.

Les ressources lexicales sous Unitex doivent avoir le format et la syntaxe des dictionnaires DELA de LADL[2]. Unitex a prévu l'ajout des modules de langues sémitiques aux vingt modules de langues qui existent déjà[3]. Les options d'écriture et de lecture des textes et des graphes ainsi que l'encodage utilisé par Unitex (UTF16) permettent d'ajouter facilement la langue arabe à cette plateforme. Un module d'une langue est représenté par un répertoire qui contient toutes les ressources et les fichiers de configurations : dans les sections suivantes nous exposons les différents composants du module arabe que nous avons ajoutés à Unitex.

3.1. L'alphabet

La conception du module arabe commence par le choix de l'alphabet et de l'ordre de cet alphabet pour effectuer un tri des unités lexicales trouvées lors d'un traitement. On a opté pour l'alphabet arabe d'Unicode[4] qui commence de \u0621 jusqu'au \u0652 dans le même ordre. Ce qui constitue les deux fichiers *alphabet* et *alphabet_sort*.

3.2. Le corpus

Unitex est un logiciel sous licence LGPL[5], donc le corpus distribué avec le module arabe doit être libre de droits. Notre choix s'est porté sur un texte qui représente une légende arabe écrite par Ibn Toufeyl (1110-1185), un philosophe et savant arabe de l'Andalousie : le roman a pour taille 96 Ko (32 pages A4) et est composé de 18 261 formes simples. Nous allons par la suite proposer un texte libre de droits et qui illustre l'ASM, c'est-à-dire un texte qui date au plus du début 19ème siècle

Notre corpus est partiellement diacritisé du fait qu'il contient des mots comprenant des signes diacritiques tels que les signes de nounation (tanwin) et de gémination (shadda). Le reste du corpus qui représente la grande partie des textes n'est pas diacritisé.

3.3. Les jeux d'étiquettes

Pour le choix du jeu d'étiquettes, nous avons opté pour un ensemble de 17 catégories syntaxiques (cf. le tableau 1). Ces catégories sont le résultat d'une étude comparative approfondie des jeux d'étiquettes proposées dans les travaux. Notons que parmi ces catégories il y a des catégories fermées et des catégories ouvertes. Les

[2] Laboratoire d'Automatique Documentaire et Linguistique

[3] Site officiel de l'Unitex : http ://www-igm.univ-mlv.fr/~unitex (consulté en Fev 2013)

[4] www.unicode.org/charts/PDF/U0600.pdf (consulté en Fev 2013)

[5] Site officiel de la fondation GNU http://www.gnu.org/licenses/lgpl.html (consulté en Fev 2013)

traits morphosyntaxiques sont une partie importante du jeu d'étiquettes, dans notre travail nous avons opté pour un ensemble de 21 traits.

Dans Unitex, le jeu d'étiquettes est introduit dans des fichiers de configuration pour être utilisé ultérieurement dans des taches de validation telle que la vérification de conformité des dictionnaires. La figure 1 représente le fichier *morphology* de notre package arabe d'Unitex.

3.4. Le prétraitement

Le traitement de corpus est précédé par un prétraitement dont l'objectif est de normaliser quelques formes qui peuvent exister dans le texte. Notre prétraitement consiste à :

Appartenance	Application	Catégories	Noms	Verbes	Particules	Résiduel	Ponctuation
QAC[6]	Étiquetage morpho syntaxique du Coran	5	34	3	21	15	12
Wajdi et al.	Général	5	29	3	13	15	11
Khoja	Étiquetage morpho syntaxique	5	103	57	9	7	1
El-Kareh et al.	Étiquetage morpho syntaxique	3	46	3	23	-	-
Alqrainy et al.	Vocalisation automatique	3	11	3	7	-	-

Tableau 1. Les jeux d'étiquettes des projets sur le TAL arabe

- Supprimer le caractère kashida (tatouil), c'est-à-dire le caractère - (\u0640), ce caractère est utilisé pour distendre un mot pour qu'il convienne à une mise-en-forme bien donnée par exemple le mot كتب /kataba/ peut être écrit sous différentes formes, comme suit : كـــــــتب ,كتـــــــــب de longueur 20 et 15 respectivement au lieu de 3. Les trois formes sont équivalentes donc la suppression ou le maintien du caractère kashida sont équivalents.

```
Arabic
<CATEGORIES>
Nb : s, p, d
Gen : m, f
Temps : A, I, F, P
Pers : 1, 2, 3
Cas : u, a, i, U, A, I
Mode : u, a, o
Def : r, n
Emph : <E>, n
Constr : <E>, c
<CLASSES>
Nc : (Nb,<var>),(Gen,<var>),(Def,<var>),(Cas,<var>),(Constr,<var>)
Np : (Nb,<var>),(Gen,<var>),(Cas,<var>)
Adj:(Nb,<var>),(Gen,<var>),(Def,<var>),(Cas,<var>)
Adv:
v : (Temps,<var>), (Pers,<var>), (Nb,<var>),(Gen,<var>),(Mode,<var>)
ve : (Temps,<var>), (Pers,<var>), (Nb,<var>),(Gen,<var>),(Mode,<var>)
Prsl : (Pers,<var>), (Nb,<var>),(Gen,<var>)
Dmst : (Nb,<var>),(Gen,<var>),(Cas,<var>)
Rltf : (Nb,<var>),(Gen,<var>),(Cas,<var>)
Prps :
CnjCrd :
Apcp :
Sbjc :
...
```

Figure 1 ***: Extrait du jeu d'étiquettes***

[6] QAC : projet Quranic Arabic Corpus cf. l'url http://corpus.quran.com/

- Supprimer les signes diacritiques à l'exception du signe gémination (shadda). Du fait que les diacritiques sont ajoutés à un mot pour résoudre une éventuelle ambiguïté la suppression de ces caractères apporte un degré d'ambiguïté plus au texte. Mais pour une simplicité de traitement l'utilisateur peut choisir ou non un prétraitement qui supprime les diacritiques (prétraitement facultatif). On note que le maintient du signe de gémination (la shadda) est dû à sa fonction qui est le redoublement du caractère sur lequel il apparut. Par exemple le verbe مدّ /mad~a/ est de longueur 2 (sans diacritiques) mais il est considéré comme un verbe trilitère à cause de redoublement du caractère د : le dernier caractère avec gémination.

- Convertir les nombres hindis en chiffres arabes

- Convertir les signes de ponctuation latins en leurs équivalents arabes

- Corriger les caractères non arabes (par exemple perses) en leurs équivalents arabes.

4. La construction des dictionnaires DELA

Comme il a été indiqué auparavant, les ressources lexicales sous Unitex doivent respecter le formalisme DELA de LADL, qui spécifie la structure des dictionnaires électroniques et de leurs contenus. On trouve dans DELA trois types de dictionnaires (Courtois, 1995) :

- dictionnaires de mots simples, DELAS et DELAF,

- dictionnaires de mots composées, DELAC et DELACF,

- dictionnaires phonémiques, DELAP et DELAPF.

Dans cet article, nous nous intéresserons seulement au premier type : les dictionnaires DELAS et DELAF pour construire le lexique arabe nécessaire au traitement automatique de la langue arabe sous Unitex. Nous rappelons que les mots composés dans la langue arabe représentent un pourcentage important du lexique mais, le traitement de cette partie du lexique est régi par des règles morphologiques différentes. Une autre étude aura pour objet spécifique la morphologie polylexicale.

Dans notre dictionnaire chaque forme fléchie a deux entrées, une entrée entièrement voyellée et l'autre non voyellée et avec signe de gémination. Mais cela à notre avis ne résout pas le problème de la voyellation car les unités lexicales dans un texte de l'ASM se présentent sous trois formes ou plus (a) formes non voyellée ni signe de gémination (b) formes non voyellée et avec signe de gémination s'il se présente dans l'unité (c) formes partiellement voyellée. Le problème de la voyellation partielle peut être résolu en intégrant un algorithme qui calcule toutes les formes de voyellation possibles d'un mot, ce qui est possible grâce à l'entrée entièrement vocalisée de notre dictionnaire. Il faudra voir si nous devons stocker ces formes ou les calculer pendant le traitement.

Dans notre travail, le lexique est divisé en deux grands groupes : catégories syntaxiques fermées et catégories syntaxiques ouvertes.

4.1. Les catégories syntaxiques fermées

En arabe, à l'instar des autres langues il y a des catégories syntaxiques ouvertes, capables d'être enrichies dans le temps et en fonction de l'évolution de la langue, et il y a les catégories syntaxiques fermées qui sont limitées en nombre et dont les fonctions syntaxiques sont bien connues, les mots outils (cf. tableau 3). Les catégories fermées par leur nature limitée et non extensible sont communes entre l'AC et l'ASM.

Dans cet article, nous considérons comme catégories fermées tous les pronoms personnels, relatifs et démonstratifs, toutes les particules ainsi que les verbes d'état.

Le tableau 3 dresse l'ensemble des catégories considérées dans notre travail comme catégories fermées. Ces catégories sont divisées en trois groupes.

4.1.1. Les verbes d'état

Les verbes d'états sont des verbes particuliers, cependant leur flexion est analogue aux verbes normaux. Donc un verbe d'état possède une forme canonique et un nombre bien déterminé de formes fléchies (cf. Figure 2). Les lemmes des verbes d'état sont cités dans tous les livres de la grammaire arabe sous le titre ÂaxawaAtu kaAna et ils sont au nombre de 13 (Elafgani).

Dans notre cas, les verbes d'état sont intégrés dans le dictionnaire des formes simples (DELAS) ; par exemple, une entrée de ce dictionnaire a la forme suivante : صار,$Ve2

Où SaAra représente le lemme non vocalisé, le caractère $ indique le mode morphologique d'Unitex et Ve2 représente le nom du paradigme flexionnel dans lequel le verbe d'état SaAra (devenir) se fléchit, c'est un graphe de flexion qui représente un transducteur.

Code	Catégorie	Catégorie arabe Correspondante
V	Verbe	Verbe
Ve	Verbe d'état	Verbe
Nc	Nom commun	Nom
Npr	Nom propre	Nom
Adj	Adjectif	Nom
Adv	Adverbe	Nom
Prsl	Pronom personnel	Nom
Dmst	Pronom démonstratif	Nom
Rltf	Pronom relatif	Nom
Intrg	Article d'interrogation	Particule
CnjCrd	Conjonction de coordination	Particule
Prps	Préposition	Particule
Sbjc	Particule de subjonctif	Particule
Evdt	Marqueur d'évidentialités	Particule
Apcp	Particule de l’apocopé	Particule
Rstr	Particule de restriction	Particule
Crbr	Marqueur de corroboration	Particule

Tableau 2. Les catégories syntaxiques du module arabe d'Unitex

Article	Catégorie	Code	Nombre
Verbe d'état	Verbe d'état	Ve	13
Pronoms personnels	Nom	Prsl	32
Pronoms démonstratifs	Nom	Dmst	34
Pronoms relatifs	Nom	Rltf	20
Particule de l’apocopé	Particule	Apcp	21
Conjonction de coordination	Particule	CnjCrd	8
Marqueur de corroboration	Particule	Crbr	7
Marqueur d'évidentialités	Particule	Evdt	7
Article d'interrogation	Particule	Intrg	15
Préposition	Particule	Prps	22
Particule de restriction	Particule	Rstr	9
Particule de subjonctif	Particule	Sbjc	6

Tableau 3. La liste des catégories fermées

صِرْتُ,صار.Ve:A1sm:A1sf
صِرْنَا,صار.Ve:A1dm:A1df:A1pm:A1pf
صِرْتَ,صار.Ve:A2sm
صِرْتِ,صار.Ve:A2sf
صِرْتُمَا,صار.Ve:A2dm:A2df
صِرْتُمْ,صار.Ve:A2pm
صِرْتُنَّ,صار.Ve:A2pf
صَارَ,صار.Ve:A3sm
صَارَتْ,صار.Ve:A3sf
صَارَا,صار.Ve:A3dm
صَارَتَا,صار.Ve:A3df
صَارُوا,صار.Ve:A3pm
…

Figure 2 : ***Extrait du DELAF : les entrées d'un verbe d'état***

Par exemple, le graphe Ve2 de la figure 3 génère 96 formes fléchies, on note ici que les verbes d'états se fléchissent en fonction du : sujet (personne, genre et nombre), de l'aspect et du mode ; on note aussi qu'il y a des verbes d'état qui ne se fléchissent pas avec tous les aspects (Elafghani); le tableau 9 montre les types de verbes d'états et le nombre de leurs formes fléchies.

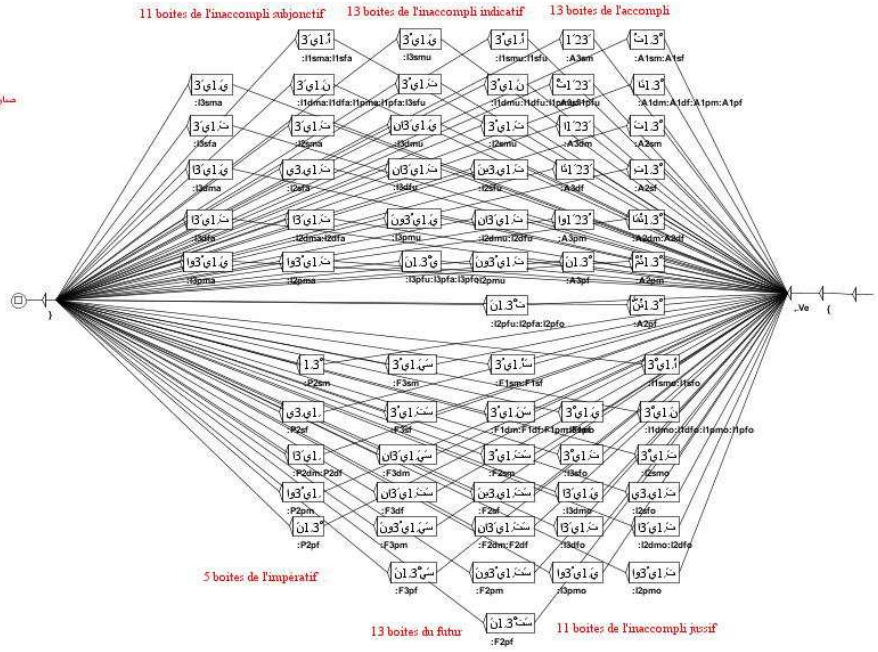

Figure 3 : ***Le graphe de flexion Ve2 du verbe d'état صار /SaAra/***

4.1.2. Les pronoms et particules

Les formes fléchies des autres catégories fermées du tableau 8 (pronoms et particules) sont calculées manuellement et ajoutées directement dans le dictionnaire DELAF. Ceci est du au nombre limité de formes fléchies pour les pronoms et l'absence des formes fléchies pour les particules.

A titre d'exemple les pronoms démonstratifs sont fléchis suivant le genre, le nombre et le cas, c'est-à-dire 18 formes fléchies (2 genres x 3 nombres x 3 cas), les pronoms personnels sont fléchies en genre, en nombre et en personne (2 genres x 3 nombres x 3 personnes). Le nombre peut augmenter si on prend en considération les formes tolérantes aux erreurs d'écriture, par exemple le pronom démonstratif هؤلاء /haŵulA'i/ peut être écrit au moins en trois formes : هؤلاء /haŵulA'i/ (73 200 000 occurrences Google), هاؤلاء /haAŵulA'i/ (324 000 occurrences Google) et هأولاء /haÂuwlA'i/ (78 300 occurrences Google)[7].

[7] Résultat d'une recherche sur www.google.com effectuée en février 2013

4.2. Catégories ouvertes

On désigne par catégories ouvertes les catégories qui évoluent suivant l'évolution de la langue, toutes les catégories non citées dans la section des catégories fermées sont considérées comme ouvertes, en l'occurrence les noms et les verbes. Avant de parler de chacune de ces catégories, présentons le format du dictionnaire DELAS.

4.2.1. Le dictionnaire DELAS

Les entrées de ce dictionnaire représentent les formes canoniques des verbes et des nominaux et leurs paradigmes flexionnels correspondants. Ce dictionnaire servira plus tard à la génération du dictionnaire final. Une entrée doit contenir le lemme du verbe ou du nom/adjectif et le nom du graphe qui représente le paradigme flexionnel, la figure 4 représente un échantillon de notre DELAS.

Verbe d'état	Aspects de flexion	Nombre de formes fléchies
كان /kaAna/	Tous	96
صار /SaAra/	Tous	96
أصبح /ÂaS.baHa/	Tous	96
بات /baAta/	Tous	96
ظل /Ďal~a/	Tous	96
أمسى /Âam.saý/	Tous	96
أضحى /ÂaD.Haý/	Tous	96
ليس /lay.sa/	Accompli	18
مابرح /maAbariHa/	Accompli, inaccompli	36
مازال /maAzaAla/	Accompli, inaccompli	36
مافتئ /maAfatiŷa/	Accompli, inaccompli	36
مادام /maAdaAma/	Accompli	18
ماانفك /maAAn.fak~a/	Accompli, inaccompli	36

Tableau 4. Les verbes d'état arabes

```
درس,$V1
كتب,$V1
أصر,$V10
أقر,$V10
تسامح,$V11
```

Figure 4 : ***Exemple des entrées de DELAS***

4.2.2. Les verbes

La flexion des verbes arabes comme il a été dit au paragraphe 2.2.1 est régulière, mais le nombre de cas dépend des patrons et de la structure saine ou défectueuse du verbe. Pour notre dictionnaire les verbes viennent d'un corpus de test (figure 5) puis un expert linguiste à travers une interface graphique lemmatise et introduit le lemme du verbe à notre programme, un patron du lemme est calculé puis recherché dans la liste des verbes déjà établis; si le verbe existe on l'ajoute directement au DELAS, sinon on calcule son graphe d'une façon automatique.

4.2.3. Le patron flexionnel

Comme il est montré dans l'algorithme de la figure 5, nous avons proposé un patron flexionnel qui définit la façon selon laquelle un verbe se fléchit : ainsi deux verbes ayant le même patron flexionnel sont obligatoirement fléchis de la même façon, même si une différence reste au niveau des consonnes qui composent la base. Ce patron est calculé en partant de l'idée que les caractères d'un verbe susceptibles d'être affectés par la flexion sont :

En début de verbe : ا, أ, إ, آ, و

En milieu de verbe : ا

En fin de verbe : ا, ى, أ, و, ي, ن, ت.

Tous les signes diacritiques : ً, ٌ, ٍ , ْ , ّ , ُ , ِ , َ

L'équivalent patron de caractères cités ci-dessus sont indiqués comme ci-dessous :

Caractère	ا	أ	إ	آ	ى	و	ي	ن	ت	َ	ُ	ِ	ّ	ْ
Schème	A	H	V	T	Y	U	I	n	t	a	u	i	s	O

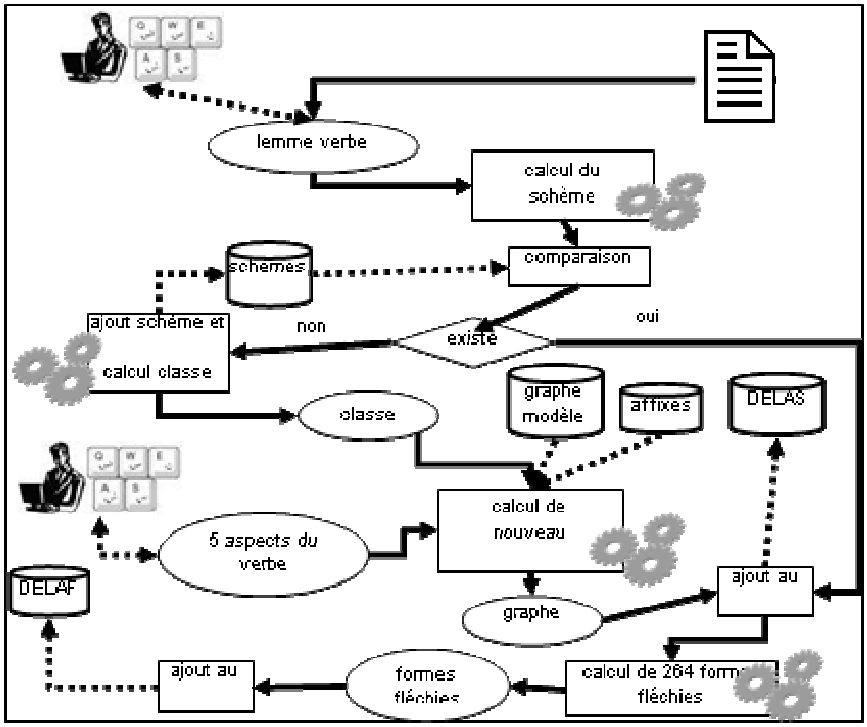

Figure 5 : ***Algorithme général de flexion des verbes***

4.2.4. La classe flexionnelle

Dans le but de générer le paradigme flexionnel sous Unitex (le graphe) d'une façon automatique, on trie les verbes arabes en de multiples niveaux ; en premier lieu, on doit déterminer la classe de verbe selon l'algorithme de la figure 6, dans cet algorithme on a :

salim(verbe) est vrai si le verbe ne contient aucun des caractères ا, ى, ّ

moudhaaf1(verbe) est vrai si le verbe contient ّ et sa longueur ne dépasse pas 2

moudhaaf2(verbe) est vrai si le verbe contient ّ et sa longueur dépasse 2

nakiss1(verbe) est vrai si la fin du verbe est un ا

nakiss2(verbe) est vrai si la fin du verbe est un ى

ajouaf(verbe) est vrai si l'intérieur du verbe contient un ا

```
Algorithme classe flexionnelle
Début
si salim(verbe) alors classe=c1
       sinon si (moudhaaf1(verbe) et nakiss1(verbe)) alors classe=c2
       sinon si moudhaaf1(verbe) et nakiss2(verbe)) alors classe=c3
       sinon si (moudhaaf1(verbe) alors classe=c4
       sinon si (moudhaaf2(verbe) et nakiss1(verbe)) alors classe=c5
```

Figure 6 : ***Algorithme de la classe flexionnelle***

4.2.5. Génération de graphe

Sous la plateforme Unitex on peut élaborer manuellement un graphe à travers un éditeur graphique, par exemple l'édition d'un graphe comme celui de la figure 3 peut dépasser une heure. Dans notre travail le nombre de graphes de flexion est en fonction des types et sous types de verbes arabes. Ainsi le nombre de graphes correspondant est important d'où l'idée de générer automatiquement les graphes. Un algorithme a été proposé pour réaliser cette tache, il a besoin de trois entrées : la classe de flexion, la liste des affixes et le graphe modèle, correspondant au thème. La figure 7 montre un exemple de liste des affixes et la figure 8 un exemple de graphe modèle.

```
**********************************
*Fichier des affixes de verbe de type :كَتَبَ
**********************************
Nombre d'entrées : 184
Accompli actif : 0-12
Inaccompli actif : 13-86
Impératif : 87-96
Accompli passif : 97-109
Inaccompli passif : 110-183

0,-,-,1,تُ
1,-,-,1,نَا
2,-,-,1,تَ
3,-,-,1,تِ
4,-,-,1,تُمَا
5,-,-,1,تُمْ
6,-,-,1,تُنَّ
7,-,-,-,-
```

Figure 7 : ***Exemple de liste d'affixes***

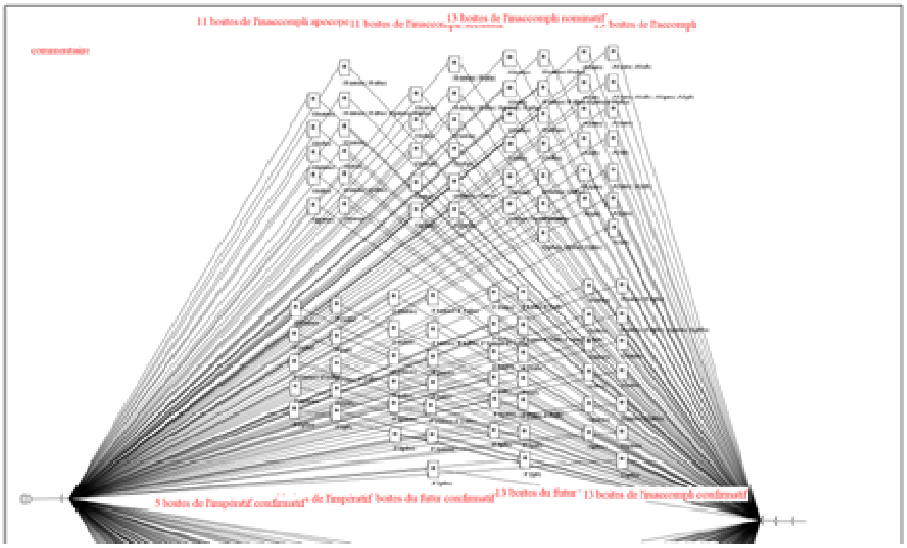

Figure 8 : ***Graphe modèle***

Le graphe modèle contient 184 boites contenant le caractère astérisque (*) et la sortie est le code flexionnel selon le cas, en utilisant des classes java qu'on a développé on génère le graphe de flexion du verbe kataba comme il est montré dans la figure 9.

Par exemple la première boite (la plus en haut à droite) contient 1تُ qui est calculée par la classe et qui a remplacé * dans le graphe modèle et la sortie de la boite reste telle qu'elle /:A1smc:A1sfc pour dire (l'accompli de l'actif de la 1ère personne du singulier masculin et féminin)

Le graphe calculé ne représente pas le paradigme d'un seul verbe mais un ensemble de verbe, pour détecter tous les verbes ayant le même paradigme flexionnel nous avons conçu un petit algorithme qui se base sur la manière dont le verbe est formé. Notre système de génération n'est pas actuellement disponible sous Unitex, mais le sera dans le futur.

4.2.6. Les noms/adjectifs

En arabe les noms communs et les adjectifs sont traités comme une seule classe : les nominaux. La flexion des nominaux est plus complexe que les verbes (Habash 2010) ; cette complexité à notre avis revient au grand nombre de paradigmes flexionnels des nominaux. Les noms et les adjectifs se fléchissent selon leurs classes flexionnelles, par exemple un déverbal est assez régulier par rapport à un nom primitif.

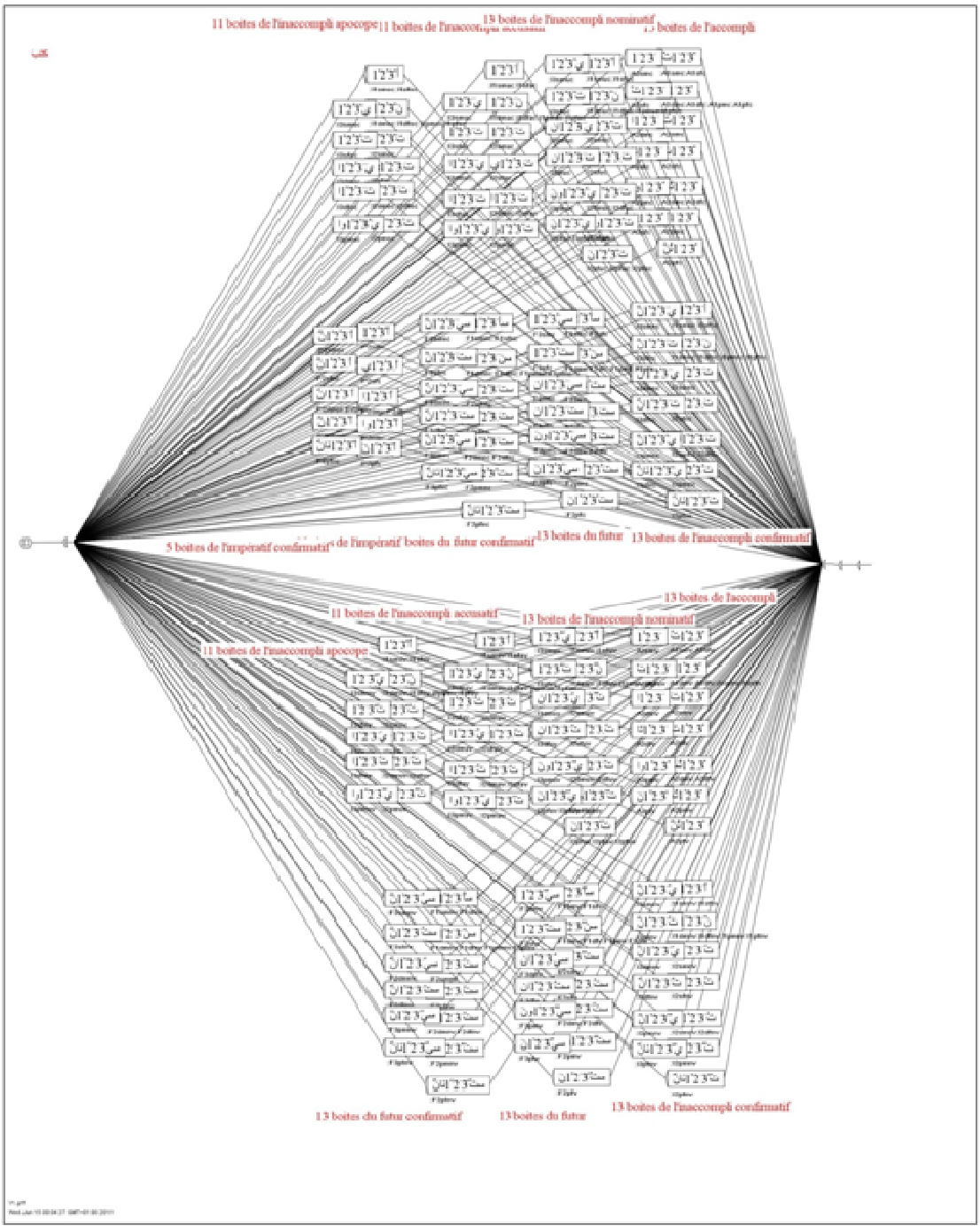

Figure 4 : ***Un graphe engendré automatiquement à partir du graphe modèle de la figure 8***

La flexion des nominaux est en fonction du genre (2 valeurs : masculin et féminin), nombre (3 valeurs : masculin, duel et pluriel), cas (3 valeurs : nominatif, accusatif et génitif), tenwine (3 valeurs : tenwine fetha, tenwine dhamma et tenwine kasra), définition (2 valeurs : défini et indéfini) et construction (2 valeurs : en construction et pas de construction).

Ces traits morphologiques ne se combinent pas toujours entre eux par exemple le tenwine ne se combinent pas avec le duel ni le pluriel régulier masculin ni le défini. Ainsi le nombre de formes fléchies d'un nominal ne dépasse pas 63 formes et sont codés dans le fichier *morphology* de la figure 1. Par exemple la ligne 8 de la figure 10 indique que la forme قَرَارَين /qaraArayn/ est morphologiquement ambiguë, elle peut être smia : un nom singulier, masculin, indéfini et accusatif ou smii : un nom singulier, masculin, indéfini et génitif. La figure 11

représente l'algorithme général de flexion des noms/adjectifs, on remarque qu'il est similaire à celui des verbes avec quelques différences

4.2.7. L'algorithme de flexion des nominaux

A travers une interface, un linguiste parcours un corpus et pour chaque mot s'il est nom/adjectif, il introduit son lemme, on calcule son patron en utilisant le même algorithme que les verbes. Ce patron est recherché dans la liste des patrons des noms qui existent, s'il est trouvé on l'ajoute au DELAS sinon le linguiste détermine sa classe et introduit son pluriel et son féminin en cas d'irrégularité. En utilisant ces entrées et en consultant des graphes modèles on calcule automatiquement le graphe de flexion du nom/adjectif courant, ce graphe sera appliqué sur le lemme pour générer 63 formes fléchies nominales.

قَرَارُ,قرار.Nc :smiu
قَرَارَ,قرار.Nc :smia
قَرَارِ,قرار.Nc :smii
قَرَارٌ,قرار.Nc :smiU
قَرَاراً,قرار.Nc :smiA
قَرَارٍ,قرار.Nc :smiI
قَرَارَانِ,قرار.Nc :dmiu
قَرَارَيْنِ,قرار.Nc :dmia :dmii

Figure 5 : ***Exemple de flexion des noms communs***

4.3. Le traitement de la cliticisation

Les entrées dans notre dictionnaire de formes fléchies sont stockées sans clitiques alors que les unités lexicales d'un texte traité se présentent sous forme de cliticisation complexe. En arabe un mot renferme plusieurs niveaux de clitiques qui sont en fonction de la catégorie syntaxique de base.

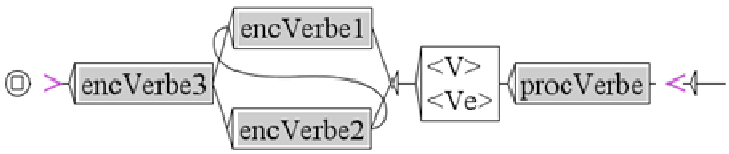

Figure 6 : Graphe de reconnaissance des unités lexicales contenant des clitiques attachées à un verbe.

Dans notre cas la cliticisation est traitée sous forme d'automate de reconnaissance la figure 12 présente un graphe de reconnaissance des unités lexicales contenant des clitiques qui peuvent s'attacher à un verbe.

4.4. Expérimentation et résultats

L'expérimentation de notre système de génération automatique de graphes a donné les résultats cités dans le tableau 14. Dans ce tableau les 1211 premiers verbes du corpus de test sont regroupés par groupes de 100 verbes, on remarque dans la colonne 4 (nouveaux graphes) que le nombre de nouveaux graphes décroit avec l'accroissement des verbes traités et on remarque dans la colonne 5 (taux de graphes/verbe) que le taux décroit presque de la moitié. En augmentant davantage le nombre de verbes traités on arrivera à un point où le nombre de nouveaux graphes tend vers zéro, c'est-à-dire qu'on couvrira tous les graphes possibles. Si on atteint ce stade le traitement d'un nouveau verbe deviendra plus facile et consistera à seulement de lui attribuer son graphe déjà établi.

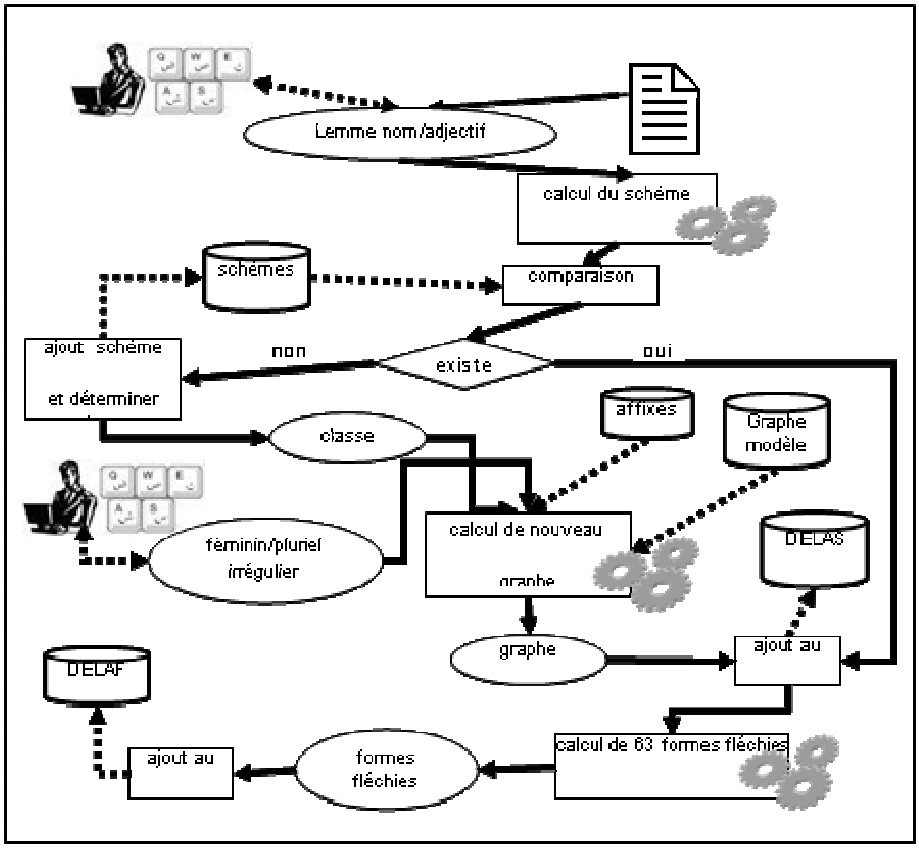

Figure 7 : ***Algorithme de flexion des nominaux***

Groupes de 100 verbes	Total verbes	Total graphes	Nouveaux graphes	Taux verbes/graphes	Taux d'accroissement des graphes
1	100	33	33	3.03	100%
2	200	53	20	3.77	37.73%
3	300	70	17	4.29	24.29%
4	400	78	8	5.13	10.26%
5	500	94	16	5.32	17.02
6	600	104	10	5.77	9.62%
7	700	104	0	6.73	0.00%
8	800	104	0	7.69	0.00%
9	900	120	16	7.50	13.33%
10	1000	120	0	8.33	0.00%
11	1100	120	0	9.17	0.00%
12	1200	146	26	8.22	17.81%
13	1211	155	9	7.81	5.81%

Tableau 5. Résultats de génération automatique des graphes de flexion

Ce qui a été dit pour les verbes est applicable sur les noms et adjectifs et avec un nombre de formes fléchies nettement moins (au plus 63 au lieu de 264 dans le cas des verbes).

5. Conclusion et perspectives

D'après l'expérimentation qui a touché 1211 lemmes de verbes, 1427 noms/adjectifs, 86 pronoms et 95 particules en forme vocalisée et non vocalisée; on conclut que l'arabe se prête facilement à exprimer sa

morphologie complexe en technologie à nombre fini d'état à travers la plateforme Unitex. Un premier package arabe est disponible avec la version 3 d'Unitex. Le tableau 15 donne le contenu des dictionnaires actuels.

Catégorie syntaxique	Entrées DELAS	Entrées DELAF voyellées	Entrées DELAF non voyellées	Pourcentage de couverture
Catégories fermées				
verbe d'état	13	768	902	
Adverbe de lieu/temps		86	95	
Particules		164	205	Plus de 95%
Pronoms		181	198	
Catégories ouvertes				
Verbe	1211	319899	486240	
Nom/adjectif	1427	75309	89388	Indéterminé
Noms propres				
Prénoms		8353		
nom de ville		7977		Indéterminé
nom de pays		802		
Taille du dictionnaire (en entrées)				**990046**

Tableau 6. Contenu des dictionnaires disponibles sous Unitex

Les dictionnaires et les graphes conçus seront étendus par des dictionnaires de mots polylexicaux et des dictionnaires de noms propres afin d'utiliser Unitex pour des applications de haut niveau telles que la reconnaissance des entités nommées dans les textes arabes.

Bibliographie

Al-Najem S. R., *Inheritance-based Approach to Arabic Verbal Root-and-Pattern Morphology*, Arabic Computational Morphology, Springer, 2007, p. 67-88

Alqrainy S., Ayesh A., *Developing a tagset for automated POS tagging in Arabic*, WSEAS transactions on computers, 2006, 5(11), p. 2787-2792.

Atwell E., Al-Sulaiti L., Al-Osaimi S., Abu Shawar B. , 'A Review of Arabic Corpus Analysis Tools', *JEP-TALN 04, Arabic Language Processing*, Fes, 19-22 April 2004.

Beesley KR., 'Arabic Finite-State Morphological Analysis and Generation', Proceedings of the 16th conference on Computational linguistics, Vol 1, 1996 Copenhagen, Denmark: Association for Computational Linguistics, pp 89-94.

Bosch A., Marsi E., Soudi A., *Memory-based Morphological Analysis and Part-of-speech Tagging of Arabic*, Arabic Computational Morphology, Springer, 2007, p. 201-217

Buckwalter T., *Issues in Arabic Morphological Analysis*, Arabic Computational Morphology, Springer, 2007, p. 23-41

Cahill L., *A Syllable-based Account of Arabic Morphology*, Arabic Computational Morphology, Springer, 2007, p. 45-66

Cavali-Sforza V., Soudi A., Arabic Computational Morphology : A trade-off Between Multiple Operations and Multiple Stems, Arabic Computational Morphology, Springer, 2007, p. 89-114

Christopher D. M., Hinrich S., *Foundations of statistical natural language processing*, Massachusetts Institute of Technology, 6ème édition, 2003, p. 140

Clark Alexader, *Supervised and Unsupervised Learning of Arabic Morphology*, Arabic Computational Morphology, Springer, 2007, p. 181-200

Courtois B., Buts et méthodes de l'élaboration des dictionnaires électroniques du LADL, LADL, CNRS, Université Paris 7, cahier du ciel 1994-1995

Darwish K., Oard D. W., *Adapting Morphology for Arabic Information Retrieval*, Arabic Computational Morphology, Springer, 2007, p. 245-262

Debili F., Ben Tahar Z., Souissi E. Analyse automatique vs analyse interactive: un cercle vertueux pour la voyellation, l'étiquetage et la lemmatisation de l'arabe. 14e Conférence TALN & 11e Rencontre RECITAL, Toulouse: IRIT Press, 2007, p. 347-356.

Diab M., Hacioglu K., Jurafsky D., *Automatic Processing of Modern Standard Arabic Text*, Arabic Computational Morphology, Springer, 2007, p. 159-179

Dichy J., Farghaly A., *Grammar-Lexis Relations in the Computational Morphology of Arabic*, Arabic Computational Morphology, Springer, 2007, p. 115-140

El-Mounjid, المُنْجِد في اللُّغَة و الاعلام /Almun.jid fy All~uγaħ w AlAçlAm/, Beyrouth, Liban, Dar El-machreq SARL Publishers 1991

Fradin B., *Recherches actuelles en morphologie*, Ecole doctorale de Poděbrady, Février 2006.

Fradin Bernard. 2003. *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.

Fradin B., F. Kerleroux & M. Plénat (eds). 2009. *Aperçus de morphologie du français*. Saint-Denis: Presses Universitaires de Vincennes.

Mourad Gridach and Noureddine Chenfour (2011), Developing a New Approach for Arabic Morphological Analysis and Generation, CoRR abs/1101.5494.

Guessoum A., Zantout R., Arabic Morphological Generation and its Impact on the Quality of Machine Translation to Arabic, Arabic Computational Morphology, Springer, 2007, p. 287-302

Habash Nizar Y., *Arabic Morphological Representation for Machine Translation*, Arabic Computational Morphology, Springer, 2007, p. 263-285

Habash Nizar Y., Soudi A., Buckwalter T., *On Arabic Transliteration*, Arabic Computational Morphology, Springer, 2007, p. 15-22

Habash Nizar. Y., *Introduction to Arabic Natural Language Processing*, USA, Morgan & Claypool publishers, 2010.

Hamlaoui A., شَذَا العُرْف في فَنِ الصَرْف */šaðaA Alçur.f fy fani AlSar.f/*, Beyrouth, Liban, Resalah Publishers, 1ère édition 2007

Khamkham A., *ArabicLDB : une base lexicale normalisée pour la langue arabe*, mémoire de Master en systèmes d'information et nouvelles technologies, Université de Sfax, Tunisie 2006

Khoja S., Garside R., Knowles G., *A tagset for the morphosyntactic tagging of Arabic*, Actes de International conference CL2001,

Khoja Shereen, *APT : Arabic Part-of-Speech Tagger*, Actes the student workshop at NAACL-2001, p. 20-25

Larkey Leah S., Ballesteros L., Connell M., E., *Light Stemming for Arabic Information Retrieval*, Arabic Computational Morphology, Springer, 2007, p. 221-243

Otakar Smrz. (2007), 'ElixirFM. Implementation of Functional Arabic Morphology', In ACL 2007 Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pages 1–8, Prague, Czech Republic, 2007.

Paumier S., *Unitex 2.1 Users manual*, institut Gaspard-Monge, Université Paris-Est Marne-La-Vallée, 2009

Soualha M., Atwell E., توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية /twĎyf qwAçd AlnHw wAlSrf fy bnA' mHll Srfy llγħ Alçrbyħ/

Soudi A., Neumann G., Bosch A., *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, Arabic Computational Morphology, Springer, 2007, p. 3-14.