Un système de génération et étiquetage automatique de dictionnaires linguistiques de l'arabe

Mohamed BOUALLEGUE (2), Mohsen MARAOUI (2), Mourad MARS (1, 2)
Mounir ZRIGUI (1)

(1)Labo (UTIC : Equipe de Monastir) – Université de Monastir
Faculté des Sciences de Monastir
Mounir.Zrigui@fsm.rnu.tn
(2)Labo LIDILEM, Université STENDHAL, Grenoble, France
{mohamed.bouallègue,Mohsen.maroui,Mourad.mars}@u-grenoble3.fr

Résumé, L'objectif de cet article est la présentation d'un système de génération automatique de dictionnaires électroniques de la langue arabe classique, développé au sein de laboratoire UTIC (unité de Monastir). Dans cet article, nous présenterons, les différentes étapes de réalisation, et notamment la génération automatique de ces dictionnaires se basant sur une théorie originale : les Conditions de Structures Morphématiques (CSM), et les matrices lexicales. Ce système rentre dans le cadre des deux projets MIRTO et OREILLODULE réalisés dans les deux laboratoires LIDILEM de Grenoble et UTIC Monastir de Tunisie

Abstract. The objective of this article is the presentation of an automatic system of generation of dictionaries electronics of Arabic basing itself on the conditions of morphemic structure and the matrices lexical. This work returns within the framework of two projects MIRTO and OREILLODULE under development to laboratories LIDILEM of Grenoble and UTIC of Tunisia.

Mots clés : génération automatique, dictionnaire, conditions de structure morphématique.

Keywords: automatic system of generation, conditions of morphemic structure, matrices lexical.

1 Introduction

L'arabe classique est une langue qui se caractérise par sa « flexion interne ». Il emploie

pour sa morphologie nominale et verbale, des marques d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas, qui peuvent être des suffixes ou des préfixes.

Contrairement au français, le système de l'arabe utilise pour ses dérivations une racine et non un radical. Cette racine est composée uniquement de consonnes, au groupement desquelles est attachée une notion générale plus ou moins précise. Les différents mots qu'on tire d'une racine à l'aide d'un jeu d'alternance vocalique sont appelés les dérivés verbaux et les dérivés nominaux.

Les grammairiens arabes usent d'une notation particulière qui substitue à la représentation en C1C2C3, une racine réelle de la langue « faAla » exprimant l'idée d'agir, faire.

La création de mots se fait par l'insertion de voyelles à l'intérieur d'une racine

Plusieurs linguistes ont poussé l'idée de génération automatique de lexique à partir des caractéristiques morphologiques et dérivationnelles de la langue arabe et les conditions de structures morphématiques (CSM), comme Cantineau (Cantineau,1960) et Greenberg (Greenberg(1950) .Cette idée est à la base de ce travail où nous présenterons un système de génération et d'étiquetage automatique de dictionnaires arabes, en se basant sur les CSM, les matrices lexicales (ML), ainsi que leurs structures et leurs modes d'accès (ZAAFRANI, 2004) (SILBERZTEIN, 1993). Nous nous focalisons sur le dictionnaire des racines admissibles et attestées. Nous appliquerons des procédures autant que possible automatisées, pour engendrer un maximum d'entrées et d'informations et éliminer les bruits : l'intervention manuelle des spécialistes de la langue s'avère nécessaire dans certains cas bien déterminés et limités pour éliminer ces derniers.

Dans ce qui suit nous allons tout d'abord détailler notre méthode de génération et étiquetage automatique qui se base sur les CSM et les ML : cette génération automatique s'aidant des CSM et ML, constitue l'originalité de cette méthode ainsi que leurs structures et leurs modes d'accès ; puis dans une deuxième étape nous présenterons le système réalisé

2 Génération et Etiquetage automatique du dictionnaire

Le dictionnaire est bien sur un élément capital pour une bonne performance de tout système de traitement automatique de langage naturel, que ce soit en termes de couverture ou de précision. De même le jeu de catégories grammaticales utilisé dans l'étiquetage a une forte influence sur la qualité du système (Chanod Tapaneinen, 1995). Il est clair qu'un système réduit de catégories conduira souvent à de meilleurs taux de réussite qu'un système plus détaillé (Merialdo, 1995).

2.1 Génération automatique du dictionnaire

1. Les conditions de structures morphématiques (CSM)

Les phonèmes de l'arabe sont liés à des restrictions combinatoires et des restrictions séquentielles très strictes qui sont énoncées sous la forme de CSM. Ces conditions sont des règles qui régissent la génération des mots dans la langue arabe : un mot qui enfreint une condition ne peut pas appartenir à l'arabe (HABAILI, 1976).

Cadre théorique:

Soit x l'ensemble des traits possibles définis par la théorie linguistique. Soit C l'ensemble des

28 consonnes de la langue arabe. Soit $C_1C_2C_3$ une racine trilitère, avec C_1, C_2 et $C_3 \in C$. Soit MP[j] [k] la matrice phonologique (avec $1 \leq j \leq 14$ et $1 \leq k \leq 28$) cette matrice représente l'ensemble des traits des consonnes de l'arabe. Soit V l'ensemble des 6 voyelles de la langue arabe. Soit $C_1V_1C_2V_2C_3V_3$ une racine trilitère voyellée, avec V_1, V_2 et $V_3 \in V$. Soit MPv [j][k] la matrice phonologique des voyelles (avec $1 \leq j \leq 14$: l'ensemble des traits des voyelles de l'arabe et $1 \leq k \leq 6$)

Les linguistes dénombrent cinq CSM qui régissent la formation des mots arabes. Ces conditions sont classées en deux types: les restrictions combinatoires et les restrictions séquentielles.

1.1. Restrictions combinatoires

Ces restrictions régissent les spécifications des traits correspondant aux phonèmes de la langue arabes. Dans ce cas trois règles sont à énoncer :

1) CSM1 : tous les phonèmes sont [-aspirés]

Tout phonème de l'arabe est une colonne de x spécifications correspondant à ces x traits, les (x-quatorze) spécifications qui ne sont pas représentées découlent automatiquement des quatorze présentes en vertu de conditions propres à l'arabe classique. La condition CSM1 distingue l'arabe classique de nombreuses langues naturelles qui opposent phonèmes aspirés et non aspirés. C'est l'existence de telles restrictions valables pour tous les phonèmes de l'arabe classique, qui a permis de ne faire figurer que quatorze traits (HABAILI, 1976), parmi x traits possibles définis par la théorie linguistique.

Si ci $\in$ C et ci $C_1C_2C_3$ (avec $1 \leq i \leq 28$) alors MP [aspiré][i] = [0] .	(**1**)

2) CSM2 : tous les phonèmes vocaliques sont [-nasal]

La condition CSM2 exclut les voyelles nasales de l'inventaire des phonèmes de l'arabe classique.

Si vi $\in$ V et vi $C_1V_1C_2V_2C_3V_3$ (avec $1 \leq i \leq 6$) alors MPv[nasale][i] = [0] .	(**2**)

3) CSM3 : tous les phonèmes qui sont [+consonantiques] sont aussi [-syllabiques]

La condition CSM3 exclut les consonnes [+syllabiques]. Cette règle est formulée de la manière suivante:

Si MP[consonante][i] = [-] alors MP[syllabique][k] = [0] .	(**3**)

Outre les restrictions combinatoires entre les valeurs des traits appartenant à un même segment, il existe aussi des restrictions séquentielles.

1.2. Restrictions séquentielles

Ce sont des restrictions qui lient les spécifications de traits appartenant à des segments successifs de la matrice de l'arabe classique, ces restrictions reflètent le fait que n'importe quelle séquence de phonèmes de l'arabe n'est pas un morphème-racine ou un allomorphe possible (variante combinatoire d'un phonème). Par exemple مـــدّ et كـــجب sont des séquences de

consonnes permises par la structure de la langue, mais pas" خخذ " (SAIDANE, 2004).

Le fait qu'il n'existe aucun morphème-racine dont la représentation phonologique soit "كـجب" n'est la séquence d'aucune contrainte structurelle, il s'agit seulement d'une lacune accidentelle : Il s'agit d'une combinaison admissible par la structure de la langue, mais qui est absente du lexique. En revanche, des séquences telles que "خخـــد" ou "ذبـــذ" ne sont pas des morphèmes-racines possibles en arabe classique. La première enfreint la restriction qui est exprimée par la condition CSM4 et la seconde celle qui est exprimée par CSM5 :

CSM4 : La condition CSM4 exclut de l'ensemble des morphèmes-racines possibles en arabe classique toute séquence de phonèmes formée de deux segments identiques, en première et en deuxième consonne radicale.

Si c, d $\in$ C et c, d $\{C_1C_2C_3\}$ (tel que c = C_1 et d = C_2) alors (c $\neq$d).	(**4**)

CSM5 : La condition CSM5 interdit des consonnes identiques qui sont [+continu, +voisé] en première et troisième consonnes radicales.

Si [(MP[continu][i] = [+], MP[voisé][i] = [+]) et (MP[continu][l] = [+], MP[voisé][l] = [+]) alors ci,dl $C_1C_2C_3$.	(**5**)

Puisque les CSM n'opèrent que sur chaque allomorphe pris isolément, elles ne rendent compte que des contraintes à l'intérieur d'un même morphème. D'où, le processus de génération des verbes nécessite l'utilisation d'autres outils, qui sont les matrices lexicales (MOUSSA, 1973).

En vue du traitement automatique, nous avons représenté les conditions de structure morphématique sous forme d'une matrice binaire dans laquelle le chiffre 0 indique l'absence du trait correspondant et le 1 la présence du trait :

MATRICE PHONOLOGIQUE (incomplètement spécifiée)

Graphème	ب	ت	ط	د	ض	ك	ق	ف	ث	ذ	ظ	س	ص	ز	ش	ج	خ	غ	ح	ع	م	ن	ل	ر	و	ي	ه	أ	i	a	u
Phonème	b	t		d		k	q	f	T			s		z	Σ		ξ				m	n	l	r	w	ɸ	h	E	ˊ	ˋ	ˆ
sonnant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			1	1	1	1	1	1	1	1	1
syll																												0	1	1	1
cons																							1	1	0	0	0	0			
cont	0	0	0	0	0	0	0	1	1	1	1					0				1			0	1	1	1	1	0			
nas	0			0																	1	1	0								
ant	1	1	1	1	1			1	1	1	1	1	1	1	0	0							1	1		0	0	0			
cor	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0			
voix	1	0	0	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1			1	1	1	1	0	0			
strident	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0			0	0	0	0	0	0			
rhizo			1		1						1		1																		
haut	0	0		0		1	0	0	0	0		1		1	1	1	0	0					0	0	1	1			1		1

bas	0	0		0			0	0	0	0							0	0	1	1			0	0			1	1	0	1	0
arrière			1		1	1	1				1		1		0	0	1	1	1	1					1	0	0	0	0	0	1

2. Matrices Lexicales

2.1. Matrices Lexicales Trilitères (MLT)

À partir du dictionnaire référence"تـــاج الـــعروس ", « taj el arous », nous avons établis des matrices lexicales bidimensionnelles qui représentent la position des consonnes dans une racine trilitère (HADDAD, 2004). Aux 28 consonnes de la langue arabe correspondent donc 28 matrices lexicales. En effet nous avons transformés les racines trilitères du dictionnaire en des matrices décrivant les racines attestées.

Ce sont des matrices binaires Mi, avec $1 \leq i \leq n$ (n = 28 : nombre des consonnes).

Mi [j][k] exprime les raines $C_iC_jC_k$ (avec i,j et k Є [1..28])(exemple كتب KTB), tel que :

Mi[][] indique la lettre qui est en première position dans la racine $C_iC_jC_k$ (ك) K

Mi[j][] indique la lettre qui est en deuxième position dans la racine $C_iC_jC_k$ (ت) T

Mi[][k] indique la lettre qui est en troisième position dans la racine $C_iC_jC_k$ (ب) B

Cette modélisation matricielle nous a permis de distinguer les cas suivants :

- Si Mi [j][k]= 1 alors la racine $C_iC_jC_k$ est une racine attestée par le dictionnaire تـــاج العروس (exemple كتب)
- Si (Mi [j][k]= 0) alors la racine $C_iC_jC_k$ n'est pas attestée par le dictionnaire " تـــاج العروس ".(exemple طضذ).

Nous pouvons schématiser comme suit cette représentation:

$\mathbf{M_i}$ **[j][k]** C_i = **ت** 1 **ك** C_k 28

1

ب C_j 0000000000**1**0000000000000

28

Figure 1 : Représentation de la matrice lexicale

Le tableau suivant représente un extrait de la matrice lexicale correspondant à la consonne « tha » **ظ** en deuxième position :

ي	و	ه	ن	م	ل	ك	ق	ف	ع	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	أ	
0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	أ
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	ب
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ت
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ث
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	ج

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ح
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	خ
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	د
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ذ
1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	ر

2.2. Matrices Lexicales Quadrilatères (MLQ)

Les matrices lexicales quadrilatères sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine quadrilatère. En s'inspirant de "تـــاج الـــعروس", et du "الشامل في تصريف الأفعال العربيّة", nous avons pu établir 28 matrices comme suit :

Soit Mi une matrice, avec 1≤ i ≤28. Soit Q une représentation d'une racine quadrilitère quelconque attestée par la langue arabe ; soit C1C2C3 une représentation d'une racine trilitère attestée et qui a donnée la racine quadrilatère Q, avec C_1, C_2 et C_3 ∈ C. Mi [j][k] exprime les racines $C_iC_jC_k$ (avec i, j et k ∈ [1..28])

Ces matrices bidimensionnelles sont formulées de la manière suivante :

- Si Mi [j][k]= 1 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de $C_iC_jC_k$ par dérivation avec le schème "فاعل", comme "كاتب".
- Si Mi [j][k]= 2 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de CiCjCk par dérivation avec le schème "فعّل", comme "بعّد".
- Si Mi [j][k]= 3 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de $C_iC_jC_k$ par dérivation avec le schème "أفعل" , comme "أبعد".
- Si Mi [j][k]= 4 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de CiCjCk par dérivation avec le schème "فعلل" , comme "زلزل".
- Si Mi [j][k]= x, avec x ∈ [أ ب ج د...ه و ي] , alors Q = $C_iC_jC_k$ x , comme "حوقل".
- Sinon (Mi [j][k]= 0) alors la racine $C_iC_jC_k$ n'est pas attestée par le dictionnaire " " تاج العروس.

M_i [j][k] C_i = ت 1 ك C_k 28

1

ب C_j 01023**1**240 س ل

28

Figure 2 : Représentation de la matrice lexicale quadrilitère

2.2 Etiquetage automatique du dictionnaire

L'étiquetage consiste à affecter à chaque lexème toutes les informations ; morphologiques, syntaxique et statistiques (*classe grammaticale, code de conjugaison pour le verbe, code de déclinaison pour le nom, indice d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas, fréquence d'apparition,…*). Nous présentons dans ce qui suit quelques observations concernant la morphologie de l'arabe, en rappelant toutefois qu'il ne s'agit pas de discuter le bien-fondé linguistique des divers concepts que nous allons présenter, mais nous voulons simplement apprécier dans quelle mesure chaque concept nous aide à atteindre notre objectif, à

savoir l'étiquetage automatique du dictionnaire arabe.

1. Caractère dérivationnel

Il semble généralement admis qu'en arabe il existe un mécanisme morphologique qui permet de décrire une large partie des formes canoniques du lexique. Ce mécanisme est basé sur la notion de schème. Un schème est généralement défini comme un modèle qui décrit un groupe de mots partageant certaines propriétés linguistiques (phonologiques, morphologiques, syntaxiques et sémantiques). Sous l'angle morphologique et selon notre centre d'intérêt, le schème est simplement une sorte de fonction dans laquelle viennent se couler les racines pour former des mots. La combinaison d'un petit nombre de schèmes (14 schèmes nominaux et 10 schèmes verbaux) avec l'ensemble des racines attestées suffirait donc pour décrire la majorité des mots arabes. Cette observation nous a permis de générer de façon automatique la presque totalité du lexique arabe en partant de l'ensemble des racines et de l'ensemble des schèmes .

Désormais, le problème n'est pas aussi simple. L'application systématique du principe de dérivation décrit ci-dessus nous permet de générer artificiellement certains mots ne répondant pas aux critères d'appartenance à la langue. La raison en est qu'en réalité, certains schèmes ne peuvent pas aller avec certaines racines. Si la majorité des mots de la langue peuvent toujours être ramenés à une racine et à un schème, toutes les racines et tous les schèmes ne peuvent être croisés pour former des mots de la langue arabe.

Toutefois, pour que le lexique généré soit exempt d'erreurs, nous devons accompagner cette procédure automatique d'une autre procédure manuelle. Selon les démarches possibles suivantes :

- Mener à posteriori une opération de correction, qui vise à éliminer les mots générés à partir de croisements inopportuns entre racine et schèmes.

- Effectuer en amont, un travail préparatoire permettant de définir les appariements possibles entre racines et schèmes; et procéder en aval, à l'élimination des mots éventuellement générés de façon abusive. Plus cet appariement initial sera fin, moins il y aura de mots incorrects générés.

Nous avons choisi la seconde démarche qui nous semble plus exhaustive en vue que cette tâche est confiée au lexicographe (c'est le linguiste qui vérifie les schèmes et les dérivées).

Les appariements possibles sont assurés par les matrices de dérivation verbale et nominale ; Ces matrices binaires définissent, pour chaque racine, les dérivées possibles (verbales pour la matrice de dérivation verbale et nominale pour la matrice de dérivation nominale). Elles sont formulées de la manière suivante :

Figure 3 : Matrice de dérivation Verbale

Si MDv[i][j] = 1 alors le schème ch_j coïncide avec v_i et le dérivé $Dv_{i,j}$ existe

Si MDv[i][j] = 0 alors le schème ch_j ne coïncide pas avec v_i et le dérivé $Dv_{i,j}$ n'existe pas

Ces matrices seront générées une seule fois par un linguiste, mais leur utilisation est indispensable dans la génération automatique des dérivés verbaux et nominaux.

2. Caractère flexionnel

"*L'arabe est une langue à flexions. Elle emploie, pour la conjugaison du verbe et pour la déclinaison du nom, des indices d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas, qui sont en général des suffixes*." [CHI 98]

Les principes de suffixation et de préfixation des mots permettant la conjugaison des verbes et la déclinaison des noms sont connus et, ils donnent l'impression d'être facilement automatisables. Nous serions donc tentés d'automatiser l'opération de génération de manière complète les formes fléchies.

3. Fréquence d'apparition

Bien que le dictionnaire ne comporte pas véritablement de probabilités, il indique pour chaque mot sa fréquence d'apparition : cette fréquence est déduite à partir d'un corpus étiqueté, le corpus sur lequel nous avons travaillé est constitué d'un ensemble de textes variés, représentant un volume global de 100 000 mots.

D'où la complexité du dictionnaire est de la forme :

Complexité =

Où V désigne la taille du vocabulaire, $A(m_i)$ dénote l'ambiguïté grammaticale du mot m_i , c'est-à-dire le nombre de classes grammaticales différentes qui peuvent être affectés au mot m_i.

3 Description du système réalisé

3.1. Les dictionnaires

Le système permet de générer automatiquement cinq dictionnaires de racines trilitères et quadrilitères étiquetés :

- Le dictionnaire théorique (21952 racines = (28)^3). Il contient toutes les racines trilitères théoriquement possibles de l'arabe standard.
- Le dictionnaire des racines admissibles (20415 racines) : Il contient les racines qui n'enfreignent aucune des (CSM).
- Le dictionnaire des racines attestées (7836): Il contient les racines utilisées dans la langue arabe et qui sont tirées des tableaux de répartitions construits à partir du grand dictionnaire arabe(الصحاح لإبن الجوهري)).
- Le dictionnaire des racines admissibles (13023 racines): Il contient les racines admissibles par la langue arabe mais non attestées. Ces racines peuvent être utilisées pour enrichir la langue arabe par d'autres mots nouveaux.
- Le dictionnaire des racines quadrilitères (4000 racines) : Il contient les racines quadrilitères attestées ; qui sont tirées des matrices lexicales quadrilitères.

Certaines racines trilitères attestées n'obéissent pas à une ou plusieurs CSM : nous avons créé un sixième dictionnaire (203 racines) qui regroupe ces racines, avec pour chacune, l'affichage de la CSM qui n'est pas vérifiée. Exemple : la racine (ببب) est attestée mais ne vérifie pas la condition CSM4.

3.2 Génération du dictionnaire principal

1. La forme des dérivées des racines trilitères (entrées)

Le processus de dérivation se réalise à partir d'une racine trilitère par préfixation, infixation ou suffixation selon des modèles (ou schèmes) en nombre limité. Il y a deux types de formes :

- Les formes des dérivés verbaux.
- Les formes des dérivés nominaux.

Ces formes sont représentées par les deux figures suivantes :

Figure 4 : Les formes des dérivés verbaux

Figure 5. Les formes des dérivés nominaux

A partir du dictionnaire attesté, des dérivées verbales conjuguées pour tous les pronoms de l'arabe (أنا, هو, هي, نحن,…) et dans les cinq temps de la langue arabe (المرفوع المضارع, المضارع المنصوب, المضارع المجزوم, الماضي, الأمر) et des dérivées nominales canoniques qui ont subis une déclinaison, Nous avons généré le dictionnaire capital qui contient les mots et les verbes comme ils se présentent dans l'écriture et dans les textes arabes

La forme générale associée à chaque entrée de ce dictionnaire est composée des huit zones suivantes :

- Zone 1 : la forme canonique de l'entrée (entrée sous forme normée c'est à dire la racine trilitère).
- Zones 2,3 et 4 : l'ensemble de dérivées verbales : ces dérivées respectent des modèles (schèmes) précis et sont générées selon le besoin par un algorithme approprié, avec le code de dérivation et du code de conjugaison associé à chaque dérivée verbale.
- Zone 5, 6,7 et 8 l'ensemble des dérivées nominaux : ces dérivées respectent des modèles (schèmes) précis et sont générés selon le besoin par un algorithme approprié, en plus du code morphosyntaxique et le code de dérivation, une information flexionnelle est associée à chaque dérivée nominale.

Figure 6 : Association entrée / information

Conclusion

La génération automatique du dictionnaire des racines trilitères et quadrilitères en utilisant les CSM et les ML fait l'originalité de ce travail. Ce dictionnaire sera à la base de toute analyse morphosyntaxique de l'arabe, il regroupe les racines du grand dictionnaire (معجـــم الـــصحاح), auquel on peut ajouter d'autres dictionnaires.

Le dictionnaire principal, résultat de ce système, contient 368762 formes de l'arabe : ce dictionnaire est généré automatiquement à la demande de l'utilisateur donc ne pose pas de

problèmes d'encombrement en mémoire.

Références

A. HADDAD, (2004). Un système de génération automatique de dictionnaires linguistiques et thématiques de la langue arabe. Mastère en informatique, Ecole Nationale des Sciences de l'informatique, TUNISIE.

A.H. MOUSSA, (1973). Statistical study of Arabic roots in moijam arous. Kouweit.

CANTINEAU, Jean (1960) . Etudes de linguistique arabe. Mémorial Jean Cantineau, Paris, Klincksieck

CHANOD, J.-P et TAPANAINEN,P. (1995) : "Creating a tagset, lexicon and guesser for a french tagger " In Proceedings of EACL SIGDAT workshop on From Texts To Tags: Issues In Multilingual Language Analysis.

GREENBERG, Joseph (1950): "The Patterning of root morpheme in Semitic" in Word, vol. 6, pp: 162-181.

H. HABAILI, (1976). Contraintes de structure morphématique en Arabe, DEA en linguistique, Canada, université de Montréal.

MERIALDO B, (1995):"Modèles probabilistes et étiquetage automatique" TAL, volume 36, 1995.

M. SILBERZTEIN, (1993).Dictionnaires électroniques et analyse automatique de textes (Le système INTEX). (Masson, Paris)

R. ZAAFRANI,(2004). Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. JEP-TALN 2004, Fès, Maroc.

T. SAIDANE, A. HADDAD, M. ZRIGUI, M. BEN AHMED, (2004). Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones
JEP-TALN 2004, Fès, Maroc.