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

Ahmed HADDAD (1), Mounir ZRIGUI (2), Mohamed Ben AHMED (3)

(1) Laboratoire RIADI (unité de Monastir), Faculté des Sciences de Monastir

ahmed.haddad@ensi.rnu.tn

(2) Laboratoire RIADI (unité de Monastir), Faculté des Sciences de Monastir

mounir.zrigui@fsm.rnu.tn

(3) Laboratoire RIADI, Ecole Nationale des Sciences Informatiques

mohamed.benahmed@riadi.rnu.tn

Mots-clés : dictionnaires électroniques, Conditions de Structures Morphématiques, matrices lexicales, Restrictions combinatoires, Restrictions séquentielles.

Keywords: electronic dictionaries, Conditions of Morphemic Structures , lexical matrix, Combinative circumscriptions, Sequential circumscriptions.

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 du laboratoire RIADI (unité de Monastir). Ce système entre dans le cadre du projet "oreillodule": un système embarqué de synthèse, traduction et reconnaissance de la parole arabe.

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.

Abstract the objective of this article is the presentation of a system of automatic generation of electronic dictionaries of the classic Arabian language, developed within RIADI laboratory (unit of Monastir). This system enters in the setting of project "oreillodule": an embedded system of synthesis, translation and recognition of the Arabian word.

In this article, we will present the different stages of realization, and notably the automatic generation of these dictionaries basing on an original theory: Conditions of Morphemic Structure (CSM), and the lexical matrixes.

1 Introduction

Plusieurs linguistes ont poussé l'idée de génération de lexique à partir des conditions de structures morphématiques(CSM), comme Cantinau et Greenberg (HABAILI, 1976). Cette idée est à la base de ce travail où nous présenterons un système de génération automatique des dictionnaires arabes, en se basant sur les CSM et les matrices lexicales (ML), leurs structures et leurs modes d'accès (ZAAFRANI, 2004) (SILBERZTEIN, 1993). 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.

2 Base théorique :

2.1 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 Є 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 Є 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.

2.1.1 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.

$$\text{Si } ci \in C \text{ et } ci \subset C_1C_2C_3 \text{ (avec } 1 \leq i \leq 28) \text{ alors MP[aspiré][i] = [0] .} \quad (1)$$

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 $\subset$ $C_1V_1C_2V_2C_3V_3$ (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.

2.1.2 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 مدّ et كجب 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 "كجب" 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 "خخد" ou "ذبذ" 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 $\subset$ $\{C_1C_2C_3\}$ (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 $\subset$ $C_1C_2C_3$. (**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).

2.2 Matrices Lexicales

2.2.1 Matrices Lexicales Trilitères (MLT)

Ce sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine trilitère, ces matrices sont extraites de la référence "تاج العروس ", avec quelques

transformations afin de l'utiliser dans ce travail (HADDAD, 2004). Aux 28 consonnes de la langue arabe correspondent 28 MLT. Les 28 matrices sont issues d'une statistique élaborée par Amr Helmi MOUSSA sur le dictionnaire تاج العروس, en transformant les racines trilitères du dictionnaire en des matrices décrivant les racines attestées par ce dictionnaire.

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

Nous distinguons les cas suivants :

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

Nous pouvons schématiser comme suit cette représentation:

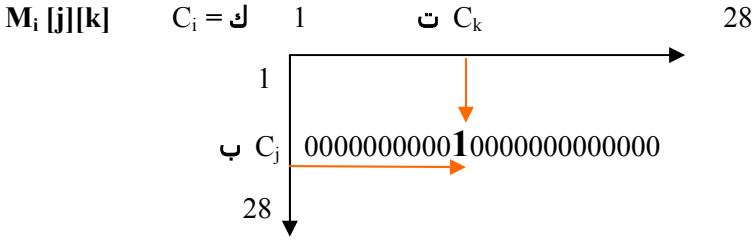

Figure 1 : Représentation de la matrice lexicale

2.2.2 Matrices Lexicales Quadrilitères (MLQ)

Les matrices lexicales quadrilitères sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine quadrilitère. En s'inspirant de "تاج العروس "، et du "الشامل في تصريف الأفعال العربيّة "، nous avons pu établir 28 matrices comme suit :

Soit Mi une matrice, avec $1 \leq i \leq 28$. 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 quadrilitère Q, avec C_1, C_2 et C_3 Є C. Mi [j][k] exprime les racines $C_iC_jC_k$ (avec i, j et k Є [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 Є [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.

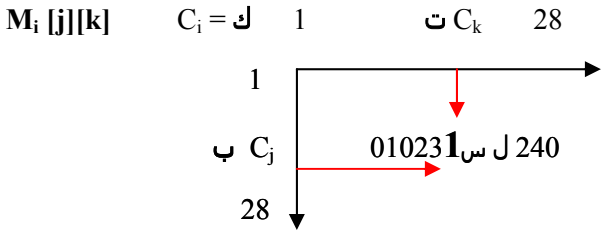

Figure 2 : Représentation de la matrice lexicale quadrilitère

3. Description du système réalisé :

3.1. Les dictionnaires :

Le système développé est composé de deux fonctions qui sont :

1. la génération des dictionnaires,
2. la consultation des dictionnaires.

3.1.1. La génération automatique de cinq dictionnaires de racines trilitères et quadrilitères arabes :

- Le premier dictionnaire est théorique (21952 racines = (28)^3). Il contient toutes les racines trilitères théoriquement possibles del'arabe standard.
- Le deuxième dictionnaire(20415 racines) : c'est le dictionnaire des racines trilitères admissibles. C'est-à-dire les racines qui n'enfreignent aucune des (CSM).
- Le troisième dictionnaire (7836): c'est le dictionnaire des racines trilitères attestées ; c'est-à-dire utilisées dans la langue arabe et qui sont tirées des tableaux de répartitions construits à partir du grand dictionnaire arabe(الصحاح لإبن الجوهري).
- Le Quatrième dictionnaire (13023 racines): c'est le dictionnaire des 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 cinquième dictionnaire (4000 racines) : c'est le dictionnaire des racines quadrilitères attestées ; qui sont tirées des matrices lexicales quadrilitéres.

Le schéma suivant illustre le diagramme de données relatif à la génération automatique des différents dictionnaires :

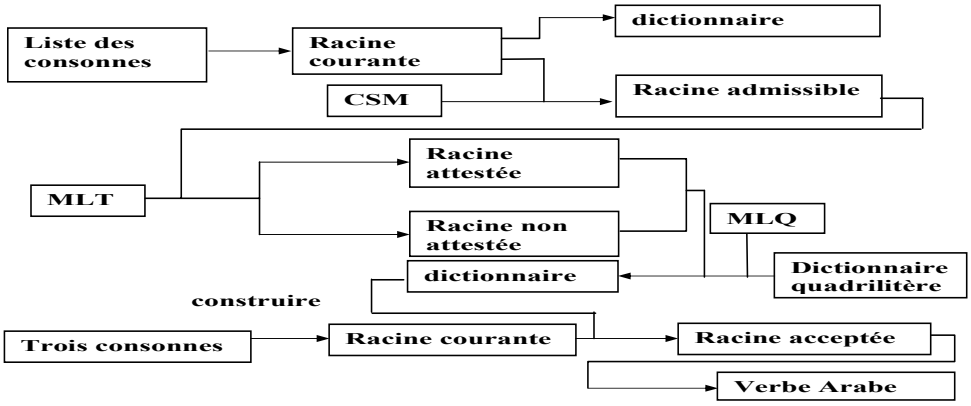

Figure 3 : Diagramme de génération des différents dictionnaires arabes

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.

3.1.2. La consultation de ces dictionnaires dans le but, de la recherche d'une racine, ou l'affichage d'une liste de racines d'un dictionnaire bien déterminé, ou sa mise à jour.

3.2 Structure interne des données du dictionnaire électronique :

Pour réaliser le dictionnaire capital, nous avons adopté la structure de listes chaînées. Cette structure permet un accès simple et une recherche facile des informations. La figure suivante décrit cette représentation interne ainsi que les différents niveaux des données :

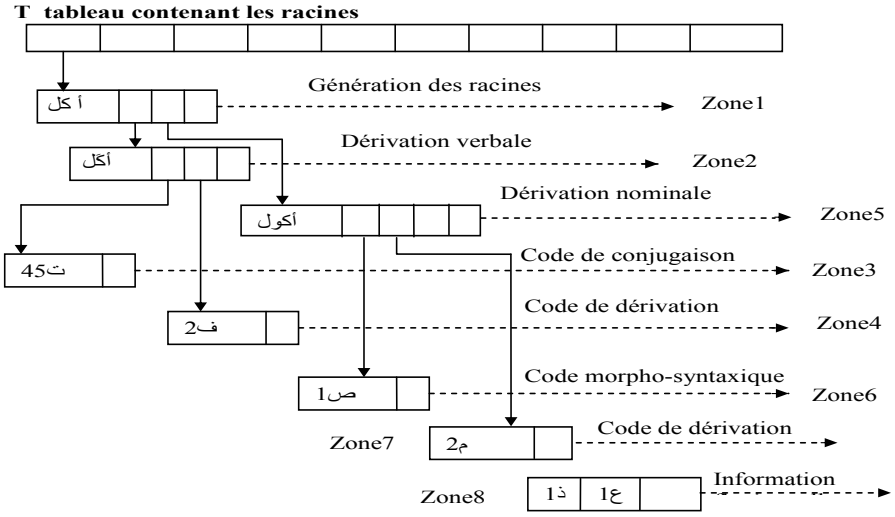

Figure 4 : Structure interne des données

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 morpho-syntaxique de l'arabe, il regroupe les racines du grand dictionnaire (معجم الصحاح), auquel on peut ajouter d'autres dictionnaires.

Le dictionnaire capital, resultat de ce système, contient 36876216 verbes et mots 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

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

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

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. Kouyet .

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, Pr. 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.