Reconnaissance de Textes Arabes à Vocabulaire Ouvert

Wady Kammoun, Abdel Ennaji

Reconnaissance de Textes Arabes à Vocabulaire Ouvert

Wady Kammoun – Abdel Ennaji

Laboratoire Perception, Système, Information (PSI)
FRE-CNRS 2645, Université de Rouen
76821 Mont Saint Aignan Cedex, France

Abdel.Ennaji@univ-rouen.fr

Résumé *: Ce papier fait suite aux travaux initiés dans [Kanoun02] portant sur la reconnaissance de textes arabes imprimés. Le prototype proposé ici s'appuie de la même manière sur l'utilisation d'un moteur de vérification morpho-syntaxique de la langue arabe pour filtrer les hypothèses générées à la suite d'une reconnaissance analytique de chaque mot. L'architecture du système proposée dans ce travail consiste à utiliser les outils de vérification linguistique uniquement en phase de post-traitement et permet de traiter l'ensemble du vocabulaire arabe (mots décomposables et non). Ce système est conçu de manière à permettre des interactions avec l'utilisateur en cas d'échec de la reconnaissance pour alimenter les différents lexiques utilisés. Un tel scénario devrait permettre d'envisager un système de reconnaissance à vocabulaire très large et évolutif en ligne. Pour la reconnaissance, un nouvel algorithme de segmentation, des caractéristiques à base de descripteurs de Fourier et un ensemble de graphèmes de base sont introduits. La validation de ce nouveau prototype a été menée sur une nouvelle base de mots.*

Mots-clés *:* Reconnaissance de textes, Analyse lexicale, Analyse morpho-syntaxique.

1 Introduction

Les avancées réalisées dans le domaine de la reconnaissance de l'écriture arabe imprimée se concrétisent aujourd'hui par un certain nombre de systèmes opérationnels. Les performances relevées dans la littérature [AMI 97] [ALB 98] sont souvent sensible à la qualité du document, au vocabulaire utilisé ainsi qu'à la taille et le type de police de caractères. De plus, les retombées commerciales de ce secteur n'étant que très peu évoluées en comparaison avec d'autres scripts tel que le latin, la plupart des travaux restent cantonnés au stade de prototype de laboratoire ou pour des applications très ciblées. Une autre limitation globale à tous les travaux dans ce domaine de recherche est liée au fait que la plupart des systèmes proposent une approche de reconnaissance lettres ou mots au mieux, sans véritablement considérer l'entité texte dans sa globalité. Les notions syntaxique et sémantique véhiculées par un texte se trouvent ainsi difficiles à appréhender, et l'usage de modules dédiés se trouve fortement conditionné par les performances des modules de reconnaissance.

C'est dans cette perspective que notre approche de la reconnaissance de textes arabes se veut une alternative aux approches classiques. L'approche initiée dans [Kanoun02] a permis de valider la démarche globale de recours aux règles linguistiques pour piloter la reconnaissance. Le système proposé initialement permettait de piloter à toutes les phases de traitement l'espace d'hypothèses par filtrage linguistique, mais ne permettait de traiter que les mots décomposables du vocabulaire. Dans ce papier, nous proposons une nouvelle architecture du système, plus simple dans sa conception puisque le filtrage linguistique n'est opéré qu'à la fin du processus de reconnaissance, mais également plus générale puisqu'elle permet également de traiter le vocabulaire non décomposables. Dans cette architecture, l'accent est également mis sur les interactions avec l'utilisateur afin d'alimenter en ligne et de manière évolutive les dictionnaires et lexiques utilisés pour la reconnaissance. Ce dernier aspect est particulièrement intéressant puisque non seulement il devrait fournir un moyen semi-automatique pour la constitution de grands lexiques, mais aussi et surtout de confronter les règles linguistiques a priori aux données traitées en ligne pour les mots décomposables en prendre ainsi compte du contexte de chaque mot dans une perspective fouille de textes.

L'approche affixale sur laquelle se base notre démarche de reconnaissance consiste à isoler pour chaque mot traité ses 4 composantes de base ou morphèmes que sont le préfixe, le suffixe, l'infixe et la racine. Cette décomposition est valable pour le vocabulaire décomposable de la langue arabe qui est principalement constitué des verbes, adverbes participes actifs et passifs, adjectifs, etc. Le vocabulaire non décomposable quant à lui est surtout constitué des noms propres, pronoms, nombres, les noms de pays, les particules, etc. Cette dernière catégorie de mots sera traitée par le système de manière analytique uniquement en cas d'échec de la reconnaissance affixale. Il en découle une architecture séquentielle dans laquelle les outils linguistiques sont utilisés dans un premier temps pour reconnaître et caractériser les mots arabes décomposables à partir d'un lexique dédié. Celui-ci contient uniquement les racines du vocabulaire utilisé ce qui permet de couvrir un vocabulaire de mots 80 fois plus important en moyenne [BEN 93]. En cas d'échec de la reconnaissance affixale,

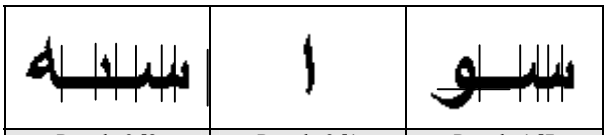
un module purement analytique est activé et utilise pour cela un dictionnaire de mots non décomposables dédié.
Dans la section 2, nous présentons le nouvel algorithme de segmentation proposé pour ce système. L'ensemble des graphèmes de base qui seront à la base de l'approche analytique utilisé sera précisé. La section 3 présente un rappel des principes de l'analyse affixale et présente les caractéristiques des morphèmes de base des mots décomposables. La section 3 aborde le principe retenu pour l'architecture du système de reconnaissance et présente les différentes étapes du processus de reconnaissance. Une partie expérimentations et résultats préliminaires est donnée à la suite de cette même section.

2 Segmentation X-Y pour l'arabe imprimé

L'entité de base considérée au niveau de la segmentation est une ligne de texte dans une image préalablement segmentée en lignes et acquise à une résolution de 300 ppp. Cette ligne subit 4 étapes de segmentation pour enfin arriver à un ensemble de graphèmes élémentaires :

- Segmentation des mots par projection verticale, en supposant que l'espace séparant les mots est plus important que celui séparant les pseudo-mots. Les seuils utilisés ont été choisis empiriquement en se basant sur la base de mots construite.
- Elimination des signes diacritiques et mémorisation leurs coordonnées. Après reconnaissance des graphèmes, il y aura analyse de l'emplacement de ces signes pour la reconnaissance des lettres. On ne fera pas de distinction sur le nombre de ces signes (processus peu fiable en général), et lors de la reconnaissance, toutes les combinaisons possibles sont générées par association des lettres et signes diacritiques. Ce principe a été utilisé dans [KAN 02].
- Segmentation verticale en graphèmes et élimination de la ligne de base. Cette opération se fait sur la ligne de base après estimation de l'épaisseur de celle-ci. Cette estimation est faite sur toute la ligne de texte. Voir fig 1.
- Segmentation horizontale. En cas de présence de jambages, il y aura segmentation en dessous de la ligne de base comme illustré en fig 2.

FIG. 1 : calcul de l'épaisseur de la ligne de base



Pseudo 3 [8 graphèmes]	Pseudo 2 [1 graphème]	Pseudo 1 [7 graphèmes]

FIG. 2 : passage d'un mot à un ensemble de graphèmes

La méthode de segmentation ainsi définie génère un ensemble de 17 graphèmes élémentaires qu'on a pu identifier sur nos bases de mots. Cette liste, présentée au tableau 1, est plus exhaustive que celle proposée dans [Kanoun02] principalement du fait de la prise en compte de mots de taille plus petite (12 et 14).

1	2	3	4	5	6	7	8	9
ا	ك	ف	ط	ع	د	ص	ه	٩
10	11	12	13	14	15	16	17	
ر	د	ذ	لا	لأ	ء	ـ	ب	

TAB. 1 : ensemble des graphèmes élémentaires

3 Principe de fonctionnement de l'analyse affixale

L'analyse affixale a été proposée dans [BEN 93] et [GAR 95] pour le traitement automatique de la langue arabe dans un but de vérification et de correction lexicale de textes électroniques arabes non-voyellés. L'analyse affixale se base sur les aspects morpho-phonologiques du vocabulaire pour décomposer un mot en morphèmes de base (préfixe, infixe, suffixe et racine) et de contrôler la cohérence de ces morphèmes avec la racine du mot. Cette approche ne permet de traiter que la catégorie des mots décomposables et ses principes de bases sont présentés ci-après.

3.1 Vocabulaire décomposable

Un mot arabe peut être décomposable ou non. Les mots décomposables sont ceux dérivés d'une racine comme les participes actifs et passifs, les adjectifs, les verbes, etc.(voir tableau 2). Les mots non décomposables sont les nombres, les noms de pays, les pronoms, les particules, etc. [BEN 93]. En l'absence de statistiques, l'observation de quelques textes courants nous laisse supposer que le vocabulaire décomposable représente environ 50% des mots d'un texte. Le reste est constitué d'environ 20% de mots purement non décomposables, et le reste d'articles, de particules, etc. Ces derniers sont pour la plupart des mots courts d'un lexique fini et bien identifié, et on peut supposer que leur reconnaissance peut être menée de manière assez fiable de manière analytique moyennant le développement de moteurs spécifiques.

Un mot décomposable peut être composé d'un couple (préfixe - suffixe) et un radical. Le radical peut à son tour être décomposable en infixe et racine Tableau 2).

	Combinaison affixale				
Mot	Préfixe	Suffixe	Infixe	Racine	Radical
تتكاثرون	تتـ	ون	ا	كثر	كاثر
تنجحوا	تنـ	وا	ϕ	نجح	نجح
يكاتب	يـ	ϕ	ا	كتب	كاتب

TAB 2: exemples de mots décomposables

3.2 Propriétés des morphèmes de base (préfixes, suffixes et infixes)

Les préfixes sont toujours situés en début du mot et sont composé d'au maximum trois lettres. Seuls six lettres de l'alphabet interviennent pour constituer un ensemble de 24 préfixes où l'absence de préfixe est comptabilisée comme hypothèse.
Les suffixes sont situés en fin de mot et sont composés d'au maximum trois lettres. Un ensemble de sept lettres interviennent pour constituer l'ensemble des 29 suffixes possibles.
Cinq lettres de l'alphabet arabe dans leur forme de milieu de mot constituent les infixes du vocabulaire arabe. Le nombre maximum que peut avoir un infixe est de deux lettres. Notons qu'un mot peut avoir comme infixe le redoublement de l'une des lettres de la racine.

3.3 Restrictions affixale et sémantique

Un mot décomposable appartient à la langue arabe si sa combinaison affixale (préfixe, infixe, suffixe) est cohérente [BEN 93]. De plus, le sens d'un mot n'est pas toujours compatible avec celui que peut apporter une combinaison affixale même si elle est correcte. Ainsi, toutes les combinaisons affixales ne sont pas cohérentes, et une racine donnée ne s'associe pas systématiquement à toutes les combinaisons affixales [BEN 93] [Kanoun02].

4 Reconnaissance de mots arabes imprimés

Le système présenté à la figure 3 est fondé sur une reconnaissance analytique pure pour initier la reconnaissance. L'ensemble des hypothèses de mots fait ensuite l'objet d'une analyse affixale selon les principes déjà exposés. L'analyse affixale [GAR 95] est réalisée sur la base d'un dictionnaire de racines et d'un ensemble de règles. En cas d'échec, le mot est considéré comme indécomposable et la reconnaissance est opérée par référence à un dictionnaire de mots indécomposables. En cas d'échec, le mot est proposé à l'utilisateur qui aura le choix d'alimenter un des trois lexiques : celui des racines, des mots ou celui des graphèmes. Ainsi le système est doté de capacités d'étiquetage semi-assisté et de constitution en ligne de dictionnaires servant à la reconnaissance d'un vocabulaire ouvert.

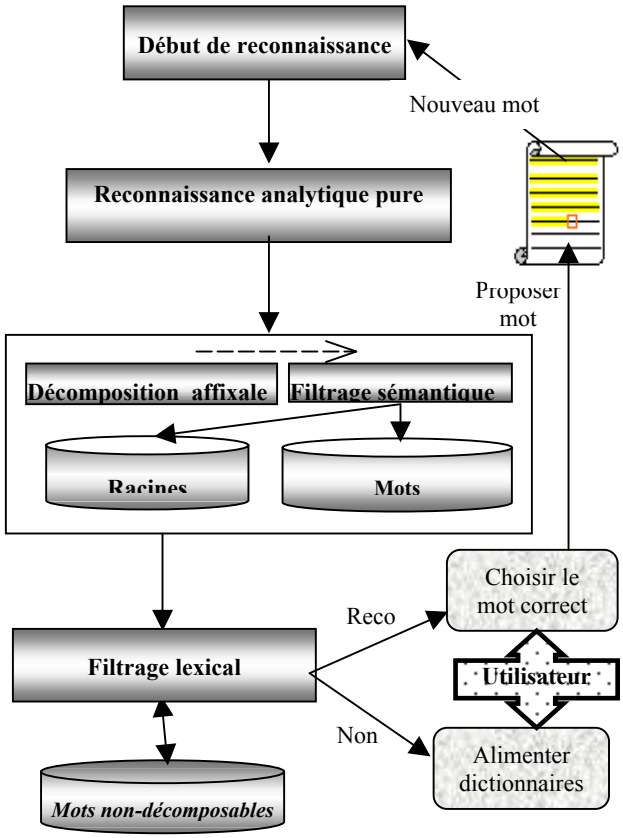


FIG. 3 : scénario de reconnaissance

4.1 Reconnaissance de lettres

Afin d'orienter la reconnaissance des lettres à partir d'une succession de graphèmes, nous avons réalisé un passage par familles. Une famille est un ensemble d'une ou plusieurs lettres morphologiquement semblables et se composant d'un même ensemble de graphèmes. Chacune de ces familles sera analysée à part pour une reconnaissance analytique en utilisant des indices visuels uniquement.
Des règles traduisant des contraintes sur la morphologie de l'écriture arabe imprimée, principalement des indicateurs de surface, sont générées pour alléger l'opération de reconnaissance. Les indices visuels pris en compte sont essentiellement les signes diacritiques et leurs emplacements et la position de la lettre dans un mot (isolée, fin d'un pseudo-mot). L'opération de reconnaissance des graphème est fondée sur l'extraction d'un ensemble de descripteurs de Fourier.
Les règles dégagées pour le passage des graphèmes aux lettres ont été testées pour l'ensemble de notre base de mots avec un taux de reconnaissance très satisfaisant. Quelques erreurs ont été détectées dans des cas très particuliers par rapport aux règles dégagées, mais le principe de reconnaissance adopté permet de compenser la plupart de ces erreurs.

ثثقايلون	تتقايلون	ننقايلون	ثتفايلون	تنفايلون	نثقابلون	ثنقابلون	تثفابلون	نتفابلون
تثقايلون	نتقايلون	ثثفايلون	تتفايلون	ننفايلون	ثتقابلون	تنقابلون	نثفابلون	ثنفابلون
نثقايلون	ثثقايلون	تثفايلون	نتفايلون	ثثقابلون	تتقابلون	ننقابلون	ثتفابلون	تنفابلون
ثتقايلون	تنقايلون	نثفايلون	ثنفايلون	تثقابلون	نتقابلون	ثثفابلون	تتفابلون	ننفابلون

TAB 3 : Reconnaissance analytique du mot "**تتقابلون** "

5 Résultats expérimentaux

Le processus de reconnaissance analytique est illustré pour le mot tatakabalouna "تتقابلون ". Le prototype arrive à détecter tous les graphèmes correctement et propose ainsi un seul ensemble. Lors de la génération des mots possibles, nous retenons toutes les combinaisons possibles par injection des points diacritiques, ce qui nous génère un ensemble de 36 mots possibles illustrés au tableau 3.

Les mots, ainsi générés, seront par la suite analysés avec le noyau de vérification affixale. Cette analyse permet d'opérer toutes les opérations de décomposition en préfixe, suffixe, infixe et racine, ainsi que les filtrages lexicaux et sémantiques nécessaires.

La base d'apprentissage de graphèmes utilisée dans ce prototype a été extraite à partir d'une base de 1423 mots, dont 320 sont non-décomposables, et correspondant à trois polices différentes pour les tailles (12, 14, 16, 18, 20 et 22). Une base de test de 300 mots décomposables a également été construite afin de tester les performances du système. Cette base de test a été réalisée avec les mêmes polices pour les tailles 12 et 14. Aucune des deux bases ne contient des mots possédant des anomalies morphologiques. Les racines utilisées pour l'ensemble des deux bases, celle d'apprentissage et l'autre de test, sont 3 et 4-consonantiques contrairement au travaux de [Kan02]. Le choix des tailles de polices utilisées s'appuie sur des constatations faites sur la taille utilisée couramment pour l'écriture arabe imprimée. De la base des 1423 mots conçus pour apprentissage nous avons construit une base d'apprentissage de 1769 graphèmes.

La phase de classification utilisée dans cette phase préliminaire utilise un k-plus-proche-voisin (kppv). Les valeurs des nombre de descripteurs de Fourier ainsi que des voisins pour le kppv sont le résultat d'expérimentations menées pour différentes valeurs. Le processus de reconnaissance permet de générer dans un premier temps toutes les hypothèses de lettres possibles par association des graphèmes proposés par le classifieur. Toutes les hypothèses de mots, de manière exhaustive, sont ensuite proposées à partir des hypothèses de lettres

Le tableau 4 donne une indication sur le nombre de listes générées à la suite de la phase de reconnaissance sur les 2 bases d'apprentissage et de test.

Nombre de listes générées	1	2	4	6	8	12	>12
Base d' apprentissage	59.7	16.1	16.8	5.1	1.8	0.5	0
Base de test	42.3	14.3	21.4	2.6	13	3.7	2.7

TAB. 4 : nombres de listes de graphèmes reconnus pour un même mot (en %)

Nous pouvons constater que le moteur de reconnaissance génère moins de 5 propositions pour près de 80% de la base de test. Ce résultat est intéressant puisqu'il traduit la complexité combinatoire de l'espace de solutions dans lequel les vérifications affixale et lexicale doivent opérer. Dans le même sens, moins de 5% de mots ont générés plus de 10 propositions. Notons enfin la différence assez importante entre les bases d'apprentissage et de test, ce qui dénote une représentativité insuffisante de la base d'apprentissage. Toutefois, cette situation peut paradoxalement illustrer l'apport des vérifications affixales.

Puisque chaque graphème est caractérisé par une distance lors de la de reconnaissance, nous calculons le poids de toute la liste de graphèmes proposée en sommant les distances des différents graphèmes. La figure 4 représente les distances correspondant aux 4 premiers mots proposés pour chacune des deux bases. Dans ce cas, les polices utilisées sont de tailles 12 et 14. Les distances affichées, sont les moyennes des distances des listes de graphèmes générés pour chaque mot de notre base.

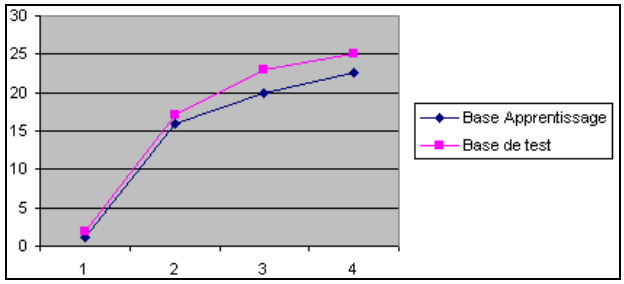


FIG. 4 : Moyenne des distances des 4 premiers mots proposés

La figure 4 montre la variation rapide entre les distances du premier et deuxième mot. Nous remarquons facilement la bonne reconnaissance de la première proposition, ce qui donne une idée sur la pertinence des données de la base de graphèmes.

Dans l'étude actuelle, nous ne proposons aucune alternative pour les anomalies causées par un

accolement, non toléré habituellement, entre les lettres. Cet accolement pourra générer dans certains cas un rejet ou bien une erreur de reconnaissance. Les autres anomalies, comme l'accolement des diacritiques à quelques lettres ne génèrent souvent pas de rejet mais généralement une mauvaise reconnaissance. Le rejet n'est pas encore géré dans la version actuelle du système et constitue une des perspectives de ce travail. Le nombre moyen observé de mots générés par la phase de reconnaissance est de 18 mots pour la base d'apprentissage (utilisée pour la constitution de la base de graphèmes) et de 24 mots pour la base de test.
A la suite de l'analyse affixale et vérification lexicale de la racine dans le lexique des racines, dans le cas de mots décomposables, un ensemble de mots valides est proposé. Le tableau 5 donne le taux d'existence du bon mot parmi les n premiers proposés.

	Top 1	**Top 2**	**Top 3**	**Top 4**
Base d'apprentissage	82.1%	96.4%	97.7%	99.8%
Base de test	81.3%	95.7%	96.4%	99.7%

TAB. 5 : taux de reconnaissance des mots décomposables

Après analyse affixale, nous obtenons un taux de reconnaissance de plus de 99% au top-4 (nombre maximal des solutions proposées) Le taux de mal-reconnaissance est de 0.2% pour la base d'apprentissage et 0.3% pour la base de test. Ces taux correspondent à la non détection du bon graphème.
Afin de tester le taux de reconnaissance global du prototype pour les mots décomposables et non décomposables, nous avons réalisé nos tests sur l'ensemble des mots des deux bases. Nous avons trouvé que dans 99.7% des cas le bon mot figure dans la liste des 4 mots proposés.

6 Conclusion

Dans ce papier nous avons proposé une architecture de système de reconnaissance de textes arabes fondée sur le recours à un moteur de vérification morpho-syntaxique du vocabulaire pour la reconnaissance tel que introduit dans [KAN02]. Cette contribution apporte en plus la possibilité de traiter le vocabulaire non décomposable de la langue arabe et permet d'envisager des interactions utilisateur pour alimenter les différents lexiques du systèmes pour le traitement de très grands vocabulaires. Les résultats préliminaires sont encourageants. Les travaux actuels portent sur la fiabilisation des différents modules du système et l'enrichissement de la base d'apprentissage pour favoriser le terrain à des analyses lexicale et sémantique à des fins de fouille de textes.

Références

[AlBadr ALB 98] B. AL-BADR, R. HARALICK, A segmentation-free approach to text recognition with application to Arabic text, *IJDAR'98*, Vol. 1 Issue 3 (1998) pp. 147-166.

[AMI 97] A. AMIN, W. MANSOOR, Recognition of printed Arabic text using neural networks, *Proc. of ICDAR'97,* pp. 612 – 615.

[BEN 93] A. BEN HAMADOU, *"Vérification et Correction Automatiques par Analyse Affixale des Textes Ecrits en Langage Naturel: le cas de l'Arabe non Voyellé"*, Thèse de Doctorat, Université de Tunis II, Mars 1993.

[GAR 95] B. GARGOURI, Une Expérimentation de l'Approche Orientée Objet dans le Traitement des Langues Naturelles : le cas de la Correction Orthographique, *mémoire DEA ISG-Université de Tunis III. 1995*

[KAN 02] KANOUN S, ENNAJI A, LECOURTIER Y, ALIMI A: Linguistic Integration Information in the AABATAS Arabic Text Analysis System; *IAPR-IWFHR'02*; pp 389-395, 2002