Voyellation automatique de l'arabe

Fathi DEBILI (1) - Hadhémi ACHOUR (2)
(1) CNRS - CELLMA / IRMC - (2) ISG / IRMC
20, rue Mohamed Ali Tahar - Mutuelleville - Tunis - Tunisie
Tél. (216.1) 584 677 - Fax : (216.1) 797 376
Courrier électronique : debili@ehess.fr

Abstract

We tackle the problem of automatic, or at least assisted, vocalization, a problem that arises from the almost universal absence of vowels in Arabic texts.
We show that the problem of vocalization resides in the fact that the majority of Arabic words accept several potential vocalizations and are therefore ambiguous.
In essence, the problem reduces to choosing, in context, the correct vocalization from among several. We focus here on the results obtained by starting with morphological analysis and proceeding to a grammatical (part-of-speech) tagging.
In the proposed system, the vocalic ambiguity is detected by means of a double dictionary of voweled and non-voweled forms. The process of resolution is set in motion starting with morphological analysis and continuing through subsequent steps. The experiments described here concern the treatment as far as grammatical (part-of-speech) tagging.

Résumé

Nous abordons le problème de la voyellation que nous voulons automatique ou du moins assistée, problème issu de l'absence quasi systématique des voyelles dans les textes arabes.
Nous montrons que le problème de la voyellation réside dans le fait que les mots arabes acceptent dans leur majorité plusieurs voyellations potentielles, qu'ils sont donc ambigus. De façon essentielle, le problème revient à choisir en contexte la bonne voyellation parmi plusieurs.

Nous focalisons ici sur les résultats obtenus au sortir de l'analyse morphologique d'abord et de l'étiquetage grammatical ensuite.

Dans le système proposé, l'ambiguïté vocalique est détectée au moyen d'un double dictionnaire non voyellé/voyellé. Le processus de résolution est enclenché dès l'analyse morphologique et se continue dans les étapes ultérieures. Les expérimentations décrites ici concernent les traitements qui vont jusqu'à l'étiquetage grammatical.

I. Introduction

Un texte arabe non voyellé est fortement ambigu. 74% des mots qui le composent acceptent plus d'une voyellation lexicale, et 89,9% des noms qui le constituent acceptent plus d'une voyelle casuelle. La proportion des mots ambigus passe à 90,5% si les comptages portent sur leurs voyellations globales (lexicales et casuelles).

Pour mieux comprendre ces chiffres prenons l'exemple du mot كتب / **ktb** et comptabilisons ses diverses voyellations lexicales et casuelles. Le dictionnaire nous renvoie les sept voyellations lexicales suivantes:

« كَتَبَ / **kataba** » *(il a écrit)*
« كُتِبَ / **kutiba** » *(il a été écrit)*
« كُتُب / **kutub** » *(des livres)*
« كَتْب / **katob** » *(un écrit)*
« كَتَّبَ / **kattaba** » *(il a fait écrire)*
« كُتِّبَ / **kuttiba** » *(faire écrire - forme factitive)*
« كَتِّبْ / **kattibo** » *(fais écrire)*

auxquelles en toute rigueur il convient d'ajouter les deux voyellations correspondant à la segmentation ك+تب / **k+tb**:

« كَتَّبَ / **katabba** » *(comme trancher)*
« كَتَبَ / **katabb** » *(comme 'tranchement')*
ce qui donne neuf voyellations au total.

Pour les noms, le dictionnaire nous renvoie d'autre part l'ensemble des cinq voyelles casuelles suivantes [ً ٌ ٍ]/[a u i an un in][1], ensemble que nous appelons schéma casuel. Comme on le voit, la voyelle casuelle ً ne figure pas dans ce schéma. La raison est qu'elle ne s'applique pas aux graphies: كُتُب / **kutub**, كَتْب / **katob** et تَـب / **tabb** mais aux graphies كُتُبا / **kutuban**, كَتْبا / **katoban** et تَبا / **tabban** qui, dans un dictionnaire de formes, constituent des entrées distinctes.

La combinatoire des voyellations lexicales et casuelles associées à كتب / **ktb** donne donc au total et minimalement 21 voyellations globales.

Comment compter ?

Derrière ces chiffres se cache une double question : Comment mesurer le nombre de voyellations lexicales et casuelles différentes d'une forme simple? Comment effectuer ces mêmes mesures sur les formes agglutinées?

La réponse n'est pas aussi simple qu'il n'y paraît au prime abord. La fusion des voyelles casuelles, du tanwin, des suffixes du pluriel ou du duel, d'une part, et l'agglutination, d'autre part, rendent difficile peu ou prou les comptages. En effet, la distinction informatique de ces différents composants linguistiques n'est pas toujours chose aisée.

S'il est trivial de compter les nombres de voyellations lexicales et casuelles d'un mot comme مدرسة / **mdrst**, qu'en est-il de mots comme مقهى / **mqhy**, مسلمون / **mslmwn**, ou كتابا/ **ktlbl**. Et que deviennent ces nombres lorsque à ces mêmes mots sont agglutinés quelque proclitiques ou enclitiques.

La difficulté vient ici de ce que la voyelle casuelle ne se trouve pas toujours présente en position finale, qu'elle ne prend pas toujours la simple forme d'un signe diacritique codé au moyen d'un seul caractère, et que dès lors elle n'est pas toujours facilement repérable.

Il y a de surcroît que les comptages sont liés à la représentation informatique des données lexicales et aux règles qui leur sont associées : selon que l'on utilise un dictionnaire de lemmes ou un dictionnaire de formes, les comptages autant que les objets (schémas vocaliques ou casuels) sur lesquels portent ces comptages seront différents.

Par exemple بـاب / **blb** aura selon qu'il est issu d'un dictionnaire de lemmes ou d'un dictionnaire de formes les schémas vocaliques et casuels suivants[2] :

Lemme	schéma voc.	schéma casuel
باب	.َ	[ً ا ٍ]

Forme	schéma voc.	schéma casuel
باب	.َ	[ً * ٍ]

La voyelle ا ً / **anl** n'est pas présente dans ce dernier schéma casuel car dans le dictionnaire de formes il y a aussi l'entrée بابا / **blbl** à laquelle

[1] Nous donnons dans la liste suivante la codification des voyelles que nous avons préconisée. Attention cette codification n'est utilisée que pour représenter les schèmes vocaliques. La translitération utilisée par ailleurs pour représenter les mots arabes est ici circonstantielle.
o : ُ, a : َ, u : ُ, i : ِ, A : ً, U : ٌ, I : ٍ, an : ً, un : ٌ, in : ٍ, An : ً, Un : ٌ, In : ٍ.

[2] **Notations** : dans la représentation des schémas vocaliques et casuels le « . » indique l'absence de signe vocalique pour les semi-voyelles (ى ي و ا) / (l w Y y) occupant la position correspondante dans la graphie du mot. De même, dans un schéma casuel, l'« * » dans une position déterminée, indique l'interdiction faite à la graphie d'avoir ce cas, étant entendu que les six positions d'un schéma casuel sont respectivement associées aux six cas : ً ٌ ٍ / a u i an un in.

sont associés les schémas vocaliques et casuels suivants :

Forme	schéma voc.	schéma casuel
بابا	´ .	[* * ا´ * * *]

Enfin il y a que la description informatique introduit parfois des simplifications qui se font au prix de confusions qui l'éloignent du modèle linguistique qu'elle est censée représenter.

Nous donnons ici au travers de différents exemples les conventions et les choix de représentation que nous avons préconisés. Rappelons que nous travaillons au moyen d'un dictionnaire de formes, celui-ci codant les voyellations de la façon suivante :

Formes	schéma vocalique	schéma casuel
مدرسة	´´´´	[´ ´ ´ ´]
	´´´´	[´ ´ ´ ´]
	´´´´	[´ ´ ´ ´]
كتاب	. ´	[´ * ´ ´]
	. ´ ´	[´ * ´ ´]
كتابا	´ . ´	[* * . * * *]
	´ . ´ ´	[* * . * * *]
مسلمون	. ´ ´ ´	[* ´ * * ´ *]
	. ´ ´ ´ ´	[* ´ * * ´ *]
	. ´ ´ ´ ´	[* ´ * * ´ *]
مقهى	´´´	[* * * . . .]
	´´´	[. . . * * *]
صحارى	´ . ´´	[. . . . . .]

Cette représentation n'est pas comme on le voit sans conséquence sur les comptages. De façon fort simple, elle ne retient au compte des voyelles casuelles d'une graphie que l'ensemble de ses seules voyelles finales. Simplicité donc, mais au double détriment :

* d'abord de schémas casuels dont le nom devient quelque peu usurpé puisque incluant parfois des voyelles qui ne sont que finales et non casuelles (par ex. مُسْـــــلِمُونَ / **musoLimuwna);**
* ensuite de schémas vocaliques incluant parfois les marques du tanwin comme pour مقهى / **mqhy**, ce qui conduit à légèrement amplifier l'ambiguïté lexicale puisque pour de tels mots l'on se retrouve avec une ou plusieurs voyellations lexicales supplémentaires, en l'occurrence ici avec : ´´´ / a o a et la voyellation lexicale supplémentaire ´´´ / a o an. Une ambiguïté lexicale « artificielle » est ainsi créée par l'apparition d'un schéma vocalique incluant la marque du tanwin.

Ces distorsions restent en fait assez marginales. Pour le traitement informatique de la voyellation elles sont sans conséquence. Il n'y a que le taux d'ambiguïté lexicale moyen qui est très légèrement amplifié au détriment de celui de l'ambiguïté casuelle. Le tableau suivant donne précisément les comptages relatifs aux entrées du dictionnaire qui donnent lieu à des schémas vocaliques ou casuels « impropres ».

Mots du type	Voyellation lexicale	Voyellation lexicale	Nb mots du dict.
عصا	´´	´´ (en plus)	35
مقهى	´´´	´´´ (en plus)	744
مسلمون	´ . ´ ´´		14071
مسلمين	´ . ´´		44502
مسلمان	´ . ´´		43747
بابا	´ .	´ . (en remplacement)	29750
صحارى	´ . ´´		680

Les cas qui conduisent à des comptages légèrement erronés au regard de la définition linguistique restent donc en proportion relativement peu nombreux : au total 30 529 sur

les 503 000 entrées que compte le dictionnaire, soit 6% du nombre total des mots non voyellés.

Et si l'on ne devait s'intéresser qu'aux seuls mots ayant reçu une voyellation lexicale supplémentaire ayant pour conséquence d'en augmenter l'ambiguïté, nous constatons que leur nombre est négligeable : 35 + 744 soit au total 779, ce qui donne en proportion 0,15%.
Le codage retenu n'introduit donc au regard de ce qui aurait été souhaitable de comptabiliser qu'une très légère distorsion dont les conséquences dans la caractérisation quantitative du problème de la voyellation ne sont pratiquement pas visibles.
D'autre part, il convient de remarquer que les comptages qui en découlent livrent au fond la véritable mesure des difficultés que nous aurons à résoudre tant il est vrai qu'il nous faut bien lever l'ambiguïté مَقْهَى / مَقْهًى (**maqohay / maqohany**).

II. Ambiguïté vocalique

II.1. Mesure en définition

Nous donnons dans le tableau suivant les comptages liés aux voyellations lexicales et casuelles des 503000 entrées du dictionnaire de formes utilisé. Les deux premières lignes livrent respectivement les proportions d'entrées non ambiguës/ambiguës au regard de la voyellation lexicale, casuelle et globale. La dernière ligne donne le nombre moyen de voyellations lexicales, casuelles et globales par entrée.

Dictionnaire	voy. lexicale	voy. casuelle	voy. globale
non ambigus	56%	57%	44%
ambigus	44%	43%	56%
nb moy. de voy. par mot	**2,08**	**2,5**	**2,9**

II.2. Mesure en usage

Le tableau suivant livre les comptages similaires effectués sur un texte d'environ 23000 unités morphologiques[3] avec répétition. Le nombre des voyellations associées à une forme agglutinée étant obtenu par la combinatoire des voyellations associées aux différentes unités lexicales qui la constituent.

Textes	voy. lexicale	voy. casuelle	voy. globale
non ambigus	25,6%	10,1%	9,5%
ambigus	74,4%	89,9%	90,5%
nb moy. de voy. par mot	**6,2**	**5,07**	**11,5**

Parce que sous l'angle de la solution informatique, le problème de la voyellation est analogue à celui de la réaccentuation automatique, nous donnons dans les deux tableaux comparatifs suivants les comptages similaires relatifs à l'accentuation du français.

II.3. Mesure en définition

Dictionnaire	Français	Arabe	
		voy. lexicale	voy. globale
non ambigus	96%	56%	44%
ambigus	4%	44%	56%
nb moy. de voy. par mot	**1,04**	**2,08**	**2,9**

[3] Nous distinguons les unités lexicales, entrées du dictionnaires, des unités morphologiques, chaînes de caractères comprises entre deux séparateurs forts dans un texte, lesquelles sont constituées d'unités lexicales agglutinées. Les proclitiques, les formes simples et les enclitiques sont des unités lexicales. Les formes simples lorsque isolées dans le texte et les formes agglutinées sont des unités morphologiques.

II.4. Mesure en usage

Textes	Français	Arabe	
		voy. lexicale	voy. globale
non ambigus	72%	25,6%	9,5%
ambigus	28%	74,4%	90,5%
nb moy. de voy. par mot	**1,3**	**6,2**	**11,5**

Ces tableaux suggèrent a priori que voyeller est sans doute plus difficile que réaccentuer. Ne part-on pas en effet d'une situation bien plus ambiguë dans le cas de l'arabe : dans le lexique 44% des mots ou même 56% si l'on prend en compte la voyelle casuelle sont ambigus, contre seulement 4% pour le français. En outre, alors que l'ambiguïté fait porter au mot arabe 2 ou 3 voyellations en moyenne, elle n'attribue pour le français que 1,04 accentuations en moyenne.

Les mêmes mesures effectuées en usage accentuent encore plus la différence : 75% (sans voyellations casuelles) / 90% (avec voyellations casuelles) des unités morphologiques constituant un texte arabe sont en effet ambiguës contre seulement 28% pour le français. Là aussi avec de surcroît pour l'arabe un niveau d'ambiguïté nettement plus élevé : de 6,2 (sans) / à 11,5 (avec) voyellations par mot en moyenne contre 1,3 accentuations en moyenne pour le français.

III. Voyellation automatique

Le processus de la voyellation automatique est dans son principe assez simple. Tout comme pour la réaccentuation, il opère en deux temps. Dans un premier temps il associe à chaque mot retenu l'ensemble de ses voyellations potentielles. Dans un second temps, et au travers de toutes les étapes ultérieures du traitement, le processus tente de réduire le nombre de ces voyellations potentielles, la résolution étant obtenue s'il n'en subsiste qu'une seule.

III.1. Quelles connaissances pour la voyellation

La voyellation, comme la réaccentuation, nécessite la mise en oeuvre de connaissances de tous les niveaux : morphologique, syntaxique, sémantique, pragmatique. Le tableau suivant donne des exemples caractéristiques de situations où l'ambiguïté vocalique (lexicale et casuelle) nécessite pour sa résolution de telles connaissances.

Nature des connaissances requises	**Exemples**
Morphologique	رأى, الى, مستفيدون Ces mots ne reçoivent à l'issue du traitement morphologique qu'une seule voyellation. Ils sont donc d'emblée voyellés.
Syntaxique	فقد ساعده كثيرا Nous nous intéressons ici à la voyellation de l'unité morphologique ســاعده. Celle-ci est constituée de l'unité lexicale ســاعد à laquelle est agglutiné l'enclitique ه. Le dictionnaire donne pour ساعد les voyellations potentielles {سَاعَدَ، سَاعِدْ، سَاعِدِّ، سَاعِدٌ، سَاعِدٍ، سَاعِدَ، سَاعِدِ}, et pour ه {هُ، هِ}. La syntaxe opère dans le cas présent à deux niveaux. D'abord au niveau de la forme agglutinée en éliminant les combinaisons illicites : سَاعَدَ+هِ، سَاعِدْ+هِ، سَاعِدَ+هِ، سَاعِدُ+هِ، سَاعِدِ+هُ، سَاعِدَّ+هُ، سَاعِدِ+هِ، سَاعِدٍ+هُ، سَاعِدٍ+هِ Ce qui conduit déjà à une réduction de l'ambiguïté. Ensuite au niveau des relations syntaxiques pour ne retenir au compte de ســاعده que la voyellation associée à son état de verbe + pronom post-verbal, soit la forme voyellée سَاعَدَ+هُ.

Sémantique	قرأت المدرسة A l'issue de l'étape syntaxique, il subsiste pour قرأت les voyellations potentielles : { ، ، ، } et pour ال+مدرسة : { ْ }+ { ، ، ، }. C'est la compatibilité sémantique qui peut aider ici à retenir la seule combinaison licite قَرَأَتْ الْمُدَرِّسَةُ
Pragmatique	إنما يخشى الله من عباده العلماء A supposer que les étapes syntaxique et sémantique aient joué leurs rôles en éliminant toutes les voyellations potentielles incompatibles ou impropres au contexte, il subsistera pour les mots العلماء et الله deux voyellations donnant lieu à deux lectures différentes : إِنَّمَا يَخْشَى اللَّهَ مِنْ عِبَادِهِ الْعُلَمَاءُ إِنَّمَا يَخْشَى اللَّهُ مِنْ عِبَادِهِ الْعُلَمَاءَ Le choix de l'une ou de l'autre lecture ne peut être effectuée ici qu'au moyen de connaissances extra-linguistiques, en l'occurrence celles qui privilégient ici la première lecture.

Dans le travail présenté ici nous explorons les contributions au traitement de la voyellation de l'analyse morphologique d'une part, et de l'étiquetage grammatical d'autre part. Nous tentons d'en évaluer les apports en termes de résolution ou sinon de réduction de l'ambiguïté vocalique.

III.2.. Analyse morphologique

A proprement parler, l'analyse morphologique ne fait que mettre au jour les diverses vocalisations potentielles des mots d'un texte. Le problème est trivial lorsqu'il s'agit de mots simples : les voyellations lexicales et casuelles sont directement délivrées par le double dictionnaire non voyellé/voyellé. Le mot est ambigu ou non ambigu d'emblée, et s'il est ambigu, l'analyse morphologique ne peut rien faire de plus. Pour un texte donné, c'est statistiquement le cas pour environ 52% des mots qui le composent[4].

Le problème est bien plus complexe lorsqu'il s'agit de formes agglutinées, soit pour 58,27% des unités qui composent un texte. Dans ces cas, l'analyse morphologique se doit de reconnaître toutes les segmentations potentielles licites et associer à toutes les unités lexicales qui en sont issues leurs diverses voyellations potentielles. Or la reconnaissance des segmentations licites n'est pas indépendante de la voyellation des unités ainsi segmentées. L'élimination des segmentations illicites repose en effet sur l'emploi de règles de compatibilité qui font appel aux propriétés linguistiques des unités segmentales précisément voyellées. Le rejet est prononcé lorsque pour une décomposition en *proclitique + forme simple + enclitique* donnée, toutes les combinaisons issues des diverses voyellations respectives de ces éléments sont déclarées incompatibles. Bien sûr, ce processus n'aboutit pas toujours. Et c'est précisément lorsqu'il n'aboutit pas qu'il y a parfois réduction voire résolution de l'ambiguïté vocalique.

C'est ce processus qui conduit par exemple pour une unité comme بكتب / bktb à conserver la décomposition ب + كتب / b + ktb avec respectivement les seules possibilités vocaliques suivantes :

pour ب { } / b {i}, et pour كتب { ، ، ، } / ktb {aoi, aoin, uui, uuin}, les autres possibilités

---

[4] 52% de formes simples contre 48% de formes agglutinées sont les proportions exactes mesurées sur un texte d'environ 23000 unités préalablement analysées à la main. Au regard de l'analyse morphologique ces proportions changent légèrement : la discrimination unités simples / unités agglutinées n'étant plus faites, il y a introduction d'une troisième classe d'unités, celles qui sont potentiellement à la fois simples et agglutinées. Les comptages donnent 41,71% de formes simples, 41,63% de formes agglutinées, et 16,64% de formes ambiguës.

{[illegible]} / {uIa, aIo, aAa, uia, aaa} ayant été éliminées.

La contribution de l'analyse morphologique au processus de voyellation ne se limite donc pas seulement à l'attribution des diverses vocalisations potentielles. Au travers de l'analyse des formes agglutinées, elle entame le processus d'élimination de certaines vocalisations potentielles, donc de réduction de l'ambiguïté vocalique, la résolution étant même atteinte dans certains cas. Le tableau suivant donne précisément une évaluation chiffrée de cette contribution.

Textes	voy. lexicale	voy. casuelle	voy. globale
Avant applications règles de compatibilité			
non ambigus	25,6%	10,1%	9,5%
ambigus	74,4%	89,9%	90,5%
nb moy. de voy. par mot	**6,2**	**5,07**	**11,5**
Après applications règles de compatibilité			
résolus	29,1%	12,6%	10,9%
ambigus	70, 9%	87,4%	89,1%
nb moy. de voy. par mot	**4,5**	**3,9**	**7,5**

Relatif à l'analyse d'un texte d'environ 23000 unités complètement non voyellées, la chadda étant en particulier elle aussi absente, ce tableau montre ce qu'apporte en sus l'analyse des unités agglutinées. Pour la voyellation globale par exemple, l'on passe de 9,5% d'unités simples reconnues non ambiguës d'emblée, à 9,5% + 1,4% d'unités simples et agglutinées résolues, soit 10,9%. On observe en même temps une diminution substantielle du degré de l'ambiguïté vocalique : le nombre moyen de voyellations potentielles pour une unité morphologique passe de 11,5 à 7,5. Les colonnes donnant les résultats liés à la voyellation lexicale et à la voyellation casuelle se lisent de la même façon. On observe là aussi les mêmes tendances.

III.3. Etiquetage grammatical

L'étiquetage grammatical n'est pas indépendant de la voyellation. En effet, les cheminements syntaxiques qu'il construit sont liés aux étiquettes grammaticales potentielles qui sont associées non pas aux mots non voyellés mais aux diverses instanciations voyellées potentielles de ces derniers. Les vocalisations sont donc intimement liées aux étiquettes grammaticales, et dès lors, l'ambiguïté vocalique à l'ambiguïté grammaticale. Si donc les vocalisations sont une condition à la détermination des diverses étiquettes grammaticales potentielles d'un mot, inversement, la réduction de cet ensemble d'étiquettes n'est pas sans conséquence sur la définition de l'ensemble des vocalisations potentielles de départ.

La question est de savoir jusqu'à quel point la réduction ou, mieux, la levée des ambiguïtés grammaticales contribue-t-elle à la réduction ou résolution de l'ambiguïté vocalique.

Pour répondre à cette question, nous nous sommes livrés à deux expérimentations.

Dans la première, la situation choisie est idéale. C'est celle où toutes les ambiguïtés grammaticales sont correctement levées. Dans la seconde, les conditions expérimentales sont normales, celles où l'étiquetage est automatique et donc non complètement résolu.

Textes	voy. lexicale	voy. casuelle	voy. globale
Après étiquetage manuel			
résolus	76,5%	98,9%	76,3%
ambigus	23,5%	1,1%	23,7%
nb moy. de voy. par mot	**1,39**	**1,01**	**1,4**
Après étiquetage automatique			
résolus	72,1%	86,6%	68,5%
ambigus	27,7%	13,4%	31,5%
nb moy. de voy. par mot	**1,46**	**1,14**	**1,51**

Commentaires

Nous remarquons tout d'abord que dans la première expérimentation, quand bien même la levée des ambiguïtés grammaticales est entièrement réalisée, la résolution des ambiguïtés lexicales n'est obtenue que pour 76,5% des mots, tandis qu'elle plafonne à 98,9% pour l'ambiguïté casuelle. Ces performances représentent en fait les seuils qui ne pourront jamais être dépassés au sortir de l'étiquetage grammatical.

Les résultats affichés dans la seconde partie du tableau sont donc à évaluer à l'aune de ces seuils.

Bibliographie

Fathi DEBILI, Christian FLUHR
Modularité et construction d'informations linguistiques pour une approche industrielle du traitement automatique du langage naturel, Colloque Informatique et Langue naturelle, Nantes, 12-13 octobre 1988.

Marc EL-BEZE, Bernard MERIALDO, Bénédicte ROZERON, Anne-Marie DEROUAULT
Accentuation automatique de textes par des méthodes probabilistes, Technique et science informatique N°6/1994.

Djamal Eddine KOULOUGHLI
Grammaire de l'arabe d'aujourd'hui, Pocket - Langues pour tous, 1994.

Michel SIMARD
Réaccentuation automatique de textes français

Emna SOUISSI
Etiquetage grammatical de l'arabe voyellé ou non, Thèse de doctorat, Université de Paris VII, Octobre 1997.