Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe

Ahmed Hamdi
Aix Marseille Université, LIF-CNRS, Marseille
ahmed.hamdi@lif.univ-mrs.fr

RESUME

Ce travail s'inscrit dans le cadre de l'analyse morphologique et syntaxique automatique de la langue arabe. Nous nous intéressons au traitement de la diacritisation et à son apport pour l'analyse morphologique. En effet, la plupart des analyseurs morphologiques et des étiqueteurs morphosyntaxiques existants ignorent les diacritiques présents dans le texte à analyser et commettent des erreurs qui pourraient être évitées. Dans cet article, nous proposons une méthode qui prend en considération les diacritiques lors de l'analyse, et nous montrons que cette prise en compte permet de diminuer considérablement le taux d'erreur de l'analyse morphologique selon le taux de diacritiques du texte traité.

ABSTRACT

Apport of Diacritization in Arabic Morpho-Syntactic Analysis

This work is concerned with the automatic morphological and syntactical analysis of the Arabic language. It focuses on diacritization and on its contribution to morphological analysis. Most of existing morphological analyzers and syntactical taggers do not take diacritics into account; as a consequence, they make mistakes that could have been avoided.
In this paper, we propose a method which process diacritics. We show that doing so reduces considerably the morphological error rate, depending on the diacritics rate in the input text.

MOTS-CLES : diacritisation, traitement automatique, analyse morphosyntaxique, langue arabe.
KEYWORDS : diacritization, computer processing, morpho-syntaxic analysis, Arabic language.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 247–254,
Grenoble, 4 au 8 juin 2012.

1 Introduction

En plus de sa morphologie fortement flexionnelle, dérivationnelle et agglutinante, la langue arabe se caractérise par l'absence des voyelles courtes (diacritiques) dans la plupart des textes écrits. En effet, contrairement au français, les voyelles courtes arabes ne sont pas des lettres de l'alphabet, ce sont des signes diacritiques qui se rajoutent aux consonnes (lettres) et qui jouent le même rôle que les voyelles dans les autres langues. La diacritisation en arabe est l'opération qui consiste à attribuer des diacritiques aux lettres des mots non diacrités. Cet exercice est à la fois classique et important dans le traitement automatique de l'arabe.

Généralement, les écrits en arabe sont non diacrités et c'est au lecteur de deviner les diacritiques des textes au moment de la lecture. En revanche, les textes religieux et quelques ouvrages scolaires sont entièrement diacrités. D'autres ressources, telles que les textes journalistiques, peuvent être partiellement diacrités. Les diacritiques rajoutés dans ces écrits sont utilisés pour lever des ambiguïtés morphologiques, syntaxiques et parfois sémantiques. Les diacritiques casuels, par exemple, servent à lever l'ambiguïté syntaxique. Ces diacritiques s'associent à la dernière lettre d'un mot à valeur nominale et ils marquent le cas. Ils aident à identifier les fonctions syntaxiques des mots dans une phrase. Les diacritiques affectés aux autres lettres sont appelés lexicaux, ils sont employés pour lever les ambiguïtés morphologiques et sémantiques.

On pourrait faire un parallèle entre la diacritisation de l'arabe et l'accentuation du français. Prenons le mot « presse » comme exemple, ce mot peut être reconnu comme un nom « presse » ou bien un participe passé « pressé ». La différence entre accentuation et diacritisation est que, en arabe, cette opération associe à chaque lettre d'un mot un diacritique.

Dans cet article, nous allons utiliser la convention de translittération définie par (Buckwalter, 2004), nous représentons entre crochets les caractères translittérés. La translittération est l'opération qui consiste à utiliser un autre jeu de caractères pour faciliter la lecture du lecteur francophone. Le diacritique est représenté après la consonne à laquelle il est affecté. Les diacritiques arabes sont classés en trois catégories :

- les diacritiques simples qui sont au nombre de quatre ◌َ [a], ◌ُ [u], ◌ِ [i] et ◌ْ [o], tous ces diacritiques se prononcent de la même façon que leurs translittérations sauf le dernier qui indique l'absence de tout son.
- les diacritiques doubles sont ◌ً [F], ◌ٌ [N] et ◌ٍ [K] : il s'agit de diacritiques casuels, ils produisent, respectivement le même son que les trois premières voyelles simples avec l'ajout du son « n » à la fin. Exemple : ◌ً se prononce « an ».
- le diacritique ◌ّ appelé «chadda», qui a pour effet le doublement de la lettre à laquelle il est associée.

Un mot arabe peut être non diacrité, partiellement ou entièrement diacrité. L'absence des diacritiques dans un mot provoque des difficultés dans le traitement automatique. C'est-à-dire, qu'un mot non diacrité est plus ambigu qu'un mot partiellement diacrité. D'après (Debili, 1998), 74% des mots en moyenne acceptent plus d'une diacritisation lexicale, et 89.9% des noms acceptent plus d'un diacritique casuel. La proportion des mots ambigus est de 90.5% si les comptages portent sur leurs diacritisations globales

(lexicales et casuelles).

Pour illustrer ces ambiguïtés de façon plus claire, prenons l'exemple du mot non diacrité ورد [wrd]. Ce mot peut être reconnu comme étant:

- le verbe وَرَدَ [warada] (*apparaître*), troisième personne du singulier, passé, voix active : «*est apparu*»,
- le verbe وُرِدَ [wurida] (*être apparu*), troisième personne du singulier, passé, voix passive : «*a été apparu*»,
- le verbe وَرَّدَ [war∼ada] (*fleurir*), troisième personne du singulier, passé, voix active : « *a fleuri* »,
- le verbe وُرِّدَ [wur∼ida] (*faire fleurir*), troisième personne du singulier, passé, voix passive : « *a fait fleurir* »,
- le nom وَرْد [warod] (*roses*), cette forme peut prendre cinq voyelles casuelles différentes suivant le contexte.

En comptant, aussi, les deux formes agglutinées وَ+رَدَّ [wa+rad∼a] (*et a rendu*) et وَ+رُدَّ [wa+rud∼a] (*et rends/et a été rendu*), la forme ورد [wrd] présente au total 11 diacritisations potentielles, pour 4 lemmes et 2 catégories grammaticales. Cet exemple montre bien que l'ambiguïté vocalique d'un mot produit des ambiguïtés lemmatiques et grammaticales.

Bien que les diacritiques soient destinés à lever les ambiguïtés lors d'un traitement automatique, la majorité des analyseurs morphosyntaxiques de l'arabe comme celui de Buckwalter (Buckwalter, 2004), Xerox (Beesley, 2005) ou l'analyseur MADA (Habash, 2005) n'analysent que des textes non diacrités à cause du manque de ressources arabes diacritées. Par conséquent, si l'entrée est partiellement diacritée, ces analyseurs commencent par éliminer tous les diacritiques, puis ils font l'analyse comme si l'entrée était non diacritée. Les analyseurs morphosyntaxiques de l'arabe ne profitent donc pas des diacritiques présents dans les textes pour désambiguïser les mots. Dans le cadre du travail présenté ici, nous proposons une méthode qui permet de prendre en compte ces diacritiques. Leur prise en compte améliorera naturellement la diacritisation automatique. Nous nous intéressons ici à étudier l'apport de ces diacritiques sur les autres niveaux d'analyse : l'étiquetage grammatical et l'analyse morphologique.

Dans la section 2 de cet article, nous présentons la méthode proposée, ainsi que l'analyseur auquel nous avons intégré notre proposition. Dans la section 3, nous décrivons les expérimentations réalisées pour évaluer notre travail, et nous donnons enfin les résultats de l'analyse morphologique avant et après l'ajout de notre solution.

2 Description de la méthode

Afin d'étudier d'une façon concrète l'influence des diacritiques sur l'analyse morphologique, nous avons introduit un ensemble de modules dans l'analyseur MADA. Nous commençons par présenter cet analyseur, en nous focalisant sur les erreurs provoquées par la non prise en compte des diacritiques lors de l'analyse. Enfin, nous décrivons en détail la méthode avec laquelle nous visons à améliorer les performances de MADA.

2.1 L'analyseur morphosyntaxique MADA

MADA (*Morphological Analyzer and Disambiguator of Arabic*) (Habash, 2005) est un analyseur morphologique de l'arabe. Cet analyseur réalise la segmentation, la diacritisation, la lemmatisation, l'étiquetage grammatical et l'analyse morphologique.

Les données d'apprentissage de MADA proviennent du corpus *the Penn Arabic Treebank* PATB (Maâmouri, 2004), le corpus d'apprentissage contient 120 000 mots alors que 12 000 mots ont été utilisés pour l'évaluation. Nous présentons les résultats de l'évaluation dans la section suivante.

Lors de l'analyse d'un texte, MADA produit pour chaque mot toutes ses analyses possibles, ensuite, le modèle SVM (*Support Vector Machines*) est utilisé pour générer une prédiction de quelques traits morphologiques. Enfin, MADA fait la hiérarchisation des analyses retournées, la meilleure analyse étant celle qui s'accorde le plus avec la prédiction.

Comme nous l'avons évoqué ci-dessus, l'un des inconvénients de cet analyseur est qu'il ne prend pas en considération les diacritiques de l'entrée, et peut donc produire des analyses incompatibles avec l'entrée. Une entrée E est non compatible avec une analyse A de MADA, si les diacritiques de E ne sont pas tous présents dans la diacritisation de A. Prenons comme exemple le mot partiellement diacrité كتبت [ktb**a**t], ce mot possède un diacritique [a] associé à la troisième lettre du mot [ktbt]. Les trois premières diacritisastions renvoyées par MADA sont respectivement :

1. كَتَبْتُ [katab**o**tu] (*j'ai écrit*)
2. كَتَبْتَ [katab**o**ta] (*tu as écrit*)
3. كَتَبَتْ [katab**a**t] (*elle a écrit*)

Le diacritique affecté à la troisième lettre de l'entrée dans les deux premières analyses retournées n'est pas identique avec l'entrée. Ainsi, ces deux analyses sont considérées comme incompatibles avec l'entrée. Elles ont entrainé des erreurs qui auraient pu être évitées au niveau de la diacritisation et aussi dans d'autres traits morphologiques tels que le genre et la personne.

2.2 Méthode proposée

Pour remédier à ce problème d'incompatibilité d'analyses, la méthode que nous proposons consiste à restreindre l'ensemble des analyses des mots retournées par MADA à celles qui contiennent des diacritisations compatibles avec les mots diacrités passés en entrée. Pour ce faire, nous avons eu recours aux automates à états finis. Les mots comportant éventuellement des diacritiques, qui sont fournis à MADA ainsi que les sorties de MADA sont représentés sous la forme d'automates. Ce mode de représentation va permettre de réaliser les tests de compatibilité entre l'entrée et la sortie grâce à des opérations standard sur les automates.

Dans un premier temps, nous représentons l'entrée par un automate à états finis A1. Chaque lettre et chaque diacritique correspond à une transition, la transition du diacritique vient juste après celle de la lettre qui lui a été associée. La figure suivante donne un exemple de représentation d'un mot avec un automate à travers le mot كتبت

[ktb**a**t].

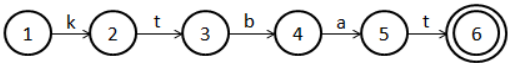

FIGURE 1 – Représentation du mot partiellement diacrité [ktbat] par A1

De la même manière, les diacritisations produites par MADA sont représentées par des automates, sauf qu'on a ajouté une transition vide (ε-transition) à chaque transition qui représente un diacritique. Deux cas sont envisageables, selon que la diacritisation est compatible ou non avec le mot en entrée. Reprenons l'exemple de la section 2.1, la figure 2 présente l'automate de la première diacritisation retournée par l'analyseur كَتَبْتُ [katab**o**tu], alors que l'automate qui correspond à la troisième diacritisation كَتَبَتْ [katab**a**t] est présenté dans la figure 3.

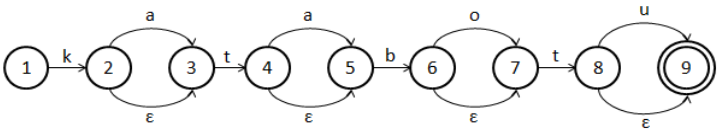

FIGURE 2 – Représentation de la diacritisation [katabotu] par A2

Le choix de l'analyse à retenir passe par la vérification de la compatibilité de l'automate A1 avec les automates A2 et A3. Deux automates sont compatibles si leur intersection est non nulle, c'est-à-dire, s'il existe un chemin commun entre eux de l'état initial à l'état final. Par conséquent, nous pouvons constater que A1 n'est pas compatible avec A2, donc, l'analyse qui contient cette diacritisation devrait être rejetée. En revanche, nous gardons la troisième analyse puisqu'elle contient une diacritisation qui s'accorde avec l'entrée.

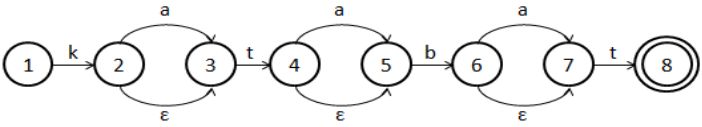

FIGURE 3 – Représentation de la diacritisation [katabat] par A3

3 Expérimentation et évaluation

Pour évaluer les prédictions de MADA avant et après l'ajout de l'option d'analyse des textes diacrités, nous avons eu recours à un corpus de test partiellement diacrité (1.3% de diacritiques) qui contient 25 295 mots préalablement annotés à la main, c'est-à-dire qu'à chaque mot, lui est attribué son analyse morphologique correcte dans le contexte où le mot apparait. Ces analyses contiennent les diacritisations, les étiquettes grammaticales et d'autres valeurs des différents traits morphologiques.

Afin de construire d'autres ressources textuelles partiellement diacritées, nous sommes partis du corpus de test entièrement diacrité, et nous l'avons dépourvu, aléatoirement, d'un taux variable de diacritiques. De cette manière, 10 corpus de test ont été obtenus, ils contiennent un pourcentage de diacritiques qui varie entre 10% et 100%.

Rappelons que MADA ne prend pas en considération les diacritiques présents dans le corpus en entrée, il produit alors les mêmes résultats d'analyse pour chacun des corpus de test. Ses performances au niveau de la diacritisation, de l'étiquetage grammatical et de l'analyse morphologique sont présentées dans le tableau suivant :

Critère	Diacritisation	Etiquetage grammatical	Analyse morphologique
Performance	86.38%	96.09%	84.25%

TABLE 1 – Performances de MADA sur notre corpus de test

Une analyse morphologique est estimée correcte, si toutes les valeurs prédites des traits morphologiques sont conformes avec l'analyse annotée dans le corpus de référence. Ainsi, l'analyseur a produit environ 86% de bonnes diacritisations pour sa meilleure analyse, 96% des catégories grammaticales correctes et 84% de bonnes analyses.

Les tests ont été réalisés, aussi, avec la prise en compte des diacritiques. Tel que nous en faisons l'hypothèse, l'expérience a montré que plus le corpus contient des diacritiques, plus les performances de MADA devraient s'améliorer (cf. table 2).

Taux de diacritisation	Performances MADA		
	Diacritisation	Etiquetage grammatical	Analyse morphologique
1.3%	86.97%	96.41%	84.91%
10%	88.47%	96.79%	86.28%
40%	91.74%	97.12%	89.48%
70%	94.85%	97.33%	92.51%
100%	98.01%	97.49%	95.59%

TABLE 2 – Performances de MADA dans l'analyse des corpus diacrités

Le tableau 2 illustre l'apport de la diacritisation dans l'analyse morphologique de l'arabe. En effet, les performances de l'analyse morphologique passent de 84.25% à 95.59% si MADA prend en considération les diacritiques présents dans les textes entièrement diacrités. L'amélioration est significative, également, dans la diacritisation et l'étiquetage grammatical. Nous remarquons, aussi, que même si le texte est entièrement diacrité, les

performances de la diacritisation de MADA n'atteignent pas 100% et s'arrêtent au niveau de 98%. Cela est dû aux mots non reconnus par MADA.

La courbe ci-dessous décrit l'évolution des performances de MADA en fonction de la proportion des diacritiques présents dans le corpus de test.

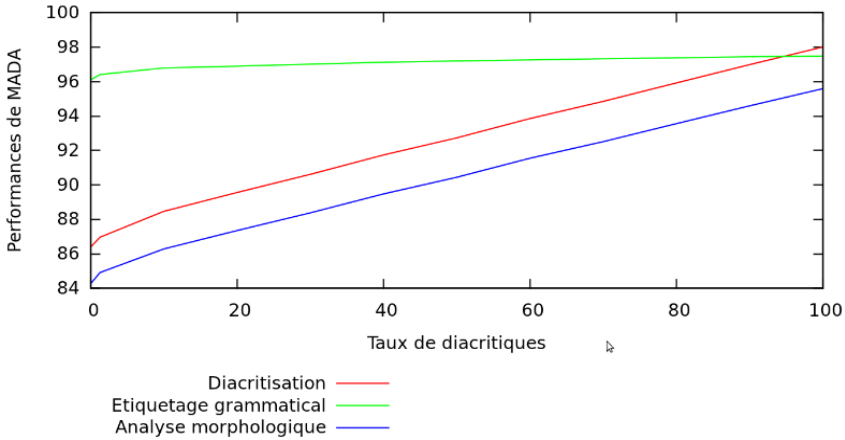

FIGURE 4 – Evolution des prformances de MADA

Ces courbes montrent l'apport des diacritiques dans la désambiguïsation morphologique. Plus il y en a dans l'entrée, plus les performances sont élevées. Nous remarquons, également, que chaque courbe présente deux pentes, avec une pente maximale lorsque le taux de diacritiques de l'entrée est entre 0% et 1.3%. Cela vient du fait que les diacritiques rajoutés à l'entrée sont « naturels », ajoutés par des humains pour désambiguïser des mots considérés comme difficiles à comprendre par le lecteur quand ils sont sous la forme non diacritée. Il est donc normal que l'apport de ces diacritiques soit important. La deuxième pente est constante puisque les diacritiques sont rajoutés d'une façon aléatoire.

4 Conclusion

Dans toute analyse linguistique, la détermination des traits morphologiques d'un mot dans son contexte constitue une étape importante. En arabe, cette détermination est rendue plus difficile par le fait que la majeure partie des mots de la langue sont ambigus. En effet, l'absence des diacritiques en arabe écrit rend les niveaux d'ambiguïté très élevés par rapport aux autres langues. Avec ce travail, nous avons voulu tester l'influence des diacritiques sur les performances pour l'analyse morphologique, l'étiquetage grammatical et la diacritisation automatiques de l'arabe. Les résultats obtenus prouvent que la prise en compte des diacritiques présents dans les textes améliore considérablement toutes ces analyses.

Cette étude pourrait être étendue à d'autres niveaux d'analyse automatique, à savoir syntaxique et sémantique. En effet, seules les diacritiques permettent de distinguer le sujet et l'objet dans ces deux phrases verbales[1] اصطحب الرجلُ الولدَ (*l'homme a accompagné l'enfant*) et اصطحب الرجلَ الولدُ (*l'enfant a accompagné l'homme*). Nous poursuivons nos travaux de thèse dans cette perspective, afin de mesurer l'influence de la diacritisation dans l'analyse syntaxique de l'arabe.

Remerciements

Je tiens à exprimer ma reconnaissance la plus sincère à mes encadrants Mme Nuria GALA et M. Alexis NASR, pour tout le temps qu'ils m'ont consacré, leur directives précieuses et leurs suivis réguliers.

Mes plus vifs remerciements s'adressent aussi à mes collègues de l'équipe TALEP pour leurs sympathies et leur soutien.

Références

BEESLEY, R. (2005). Xerox Arabic Morphological Analysis and Generation Romanization, Transcription and Transliteration.

BUCKWALTER, T. (2004). Buckwalter Arabic Morphological Analyser Version 2.0. *Linguistic Data Consortium (LDC) Catalog Number LDC2004L02, ISBN 1-58563-324-0*.

DEBILI, F. et ACHOUR, H. (1998). Voyellation automatique de l'arabe. *Actes du Workshop on Computational Approaches To Semitic Languages,* Université de Montréal.

DEBILI, F. et SOUISSI, E. (1998). Etiquetage grammatical de l'arabe voyellé ou non. *In Proceedings of the Workshop on Computational Approaches to Semetic Languages, Stroudsburg*.

DEBILI, F., ACHOUR, H. et SOUISSI, E. (2002). La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique. *Correspondances de l'IRMC, N°71, Tunis*.

HABASH, N. et OWEN, R. (2005). Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. *In Proceedings of the Conference of the American Association for Computational Linguistics*, New York.

HAJIC, J., SMRZ, F., BUCKWALTER, T. et JIN, H. (2005). Feature-Based Tagger of Approximations of Functional Arabic Morphology. *Actes de la quatrième conférence sur les Treebanks et les théories linguistiques,* Université de Barcelone.

MAAMOURI, M., BIES, A. et BUCKWALTER, T. (2004). The Pen Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. *In EMAR Conference on Arabic Language Ressources and Tools,* le Caire.

[1] Une phrase verbale dans la langue arabe est une phrase qui contient un verbe.