AraSeg : un segmenteur semi-automatique des textes arabes

Zoubeir Mouelhi

AraSeg[*] : un segmenteur semi-automatique des textes arabes

Zoubeir Mouelhi

ICAR - Université Lumière-Lyon2

Abstract

Written in Visual Basic and running on Windows, *AraSeg* is a text segmentor, and grammar manual and dictionary at the same time. The analysis of each sequence of characters is based on the model of analysis of the graphic Arabic word in which a graphic word is regarded as a succession of immediate constituents. The main role of this text segmentor is tokenization, lexical segmentation.

Résumé

Écrit en Visual Basic et tournant sous Windows, *AraSeg* est un segmenteur à la fois à grammaire et à dictionnaire. L'analyse de chaque séquence de caractères est basée sur le modèle d'analyse du mot graphique en arabe dans lequel un mot graphique est considéré comme une suite de constituants immédiats. Le rôle principal de ce segmenteur est l'itémisation, la segmentation lexicale.

Mots-clés : segmenteur, segmentation automatique, analyse de mot graphique, itémisation, analyse lexicale, grammaire segmentale, TAL arabe.

1. Introduction

La segmentation est une étape fondamentale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités d'un certain type qu'on aura définies et repérées préalablement. La segmentation d'un texte informatisé est l'opération de délimitation des segments de ses éléments de base qui sont les caractères, en éléments constituants de différents niveaux structurels : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème, etc.

Les spécificités du système d'écriture et de la morphologie arabes engendrent un grand nombre d'ambiguïtés effectives et virtuelles, morpho-lexicales ou syntaxiques, provoquant une importante explosion combinatoire, surtout au niveau de l'analyse morphologique, ce qui ajoute d'énormes difficultés à l'analyse automatique de l'arabe aussi bien au niveau lexical, morphologique ou syntaxique. La disjonction par exemple, au niveau du système d'écriture, entre l'ensemble des consonnes et celui des voyelles est une source de ces difficultés. Au niveau morpho-syntaxique, le phénomène d'agglutination au sein du mot graphique arabe qui, dans certains cas, peut constituer une phrase complète, en est une autre.

Le segmenteur que nous décrivons dans cet article est un outil de TAL arabe dont la fonction principale est d'opérer une itémisation des textes arabes basée sur une analyse lexicale des

[*] *AraSeg* est une version que nous avons améliorée, d'un segmenteur développé en collaboration avec R. Zaafrani, voir (Zaafrani, 2002).

mots graphiques. Cette segmentation lexicale s'appuie sur le modèle d'analyse, morphologique, du mot graphique en arabe.

2. Écritures segmentées *VS* écritures non segmentées

En TAL, on présente généralement les langues, quant à leur système d'écriture, comme appartenant à deux familles différentes : les langues « avec séparateurs » et les langues « sans séparateurs ». Les langues dites « avec séparateurs » sont celles qui ont des systèmes d'écritures segmentées c'est-à-dire des écritures délimitées par des espaces (*spacedelimited writings*) et où les mots sont nettement séparés par des délimiteurs (espace, signes de ponctuation, caractères spéciaux, ...). A ce type de langues on oppose les langues dites « sans séparateurs ». Elles présentent des systèmes d'écritures non segmentées (*unsegmented writings*) où les mots ne sont pas séparés par des espaces et où les frontières des mots ne sont pas nettes. Le japonais, le chinois et le thaï sont les représentants parfaits de cette deuxième famille de langues.

La langue arabe présente un système d'écriture à l'intersection des deux familles. C'est un système d'écriture qui combine une écriture segmentée et une écriture non segmentée. En effet, une partie des mots graphiques arabes correspondent à des mots minimaux séparés par des délimiteurs. En revanche, une bonne partie des mots graphiques arabes sont composés d'une suite d'unités lexicales agglutinées analysable en termes de mots minimaux et de clitiques et qu'il faut donc segmenter si l'on veut arriver aux unités de base les composant.

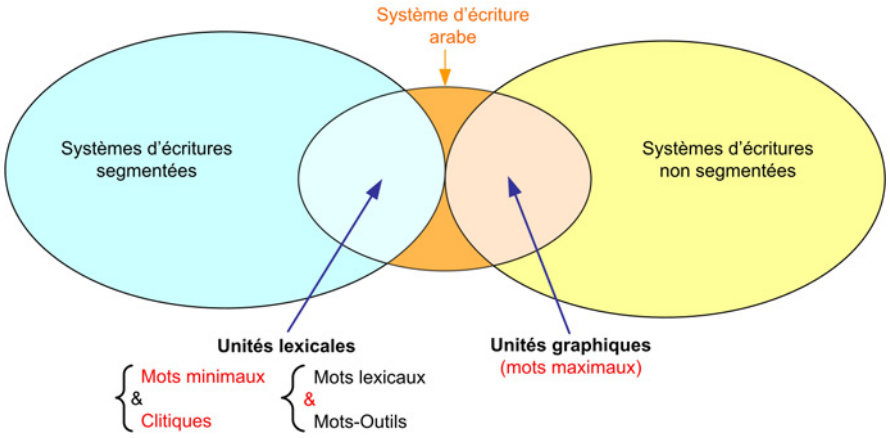

Figure 1. Systèmes segmentés et systèmes non segmentés

3. Les types de segmentation

Il existe plusieurs niveaux d'analyse auxquels on peut s'arrêter pour repérer les différents éléments constituant le texte et en définir les frontières. On peut s'arrêter au niveau de la phrase, au niveau de la proposition ou à celui du syntagme. Mais on peut arriver aussi au niveau du mot graphique, au niveau des unités lexicales ou aller au delà de celles-ci pour arriver aux unités de base les composant : les morphèmes. Selon la visée de l'analyse à entreprendre : lexicale, morphologique ou syntaxique, on peut généralement parler de trois grands types d'application de la segmentation :

- ***L'itémisation*** (*tokenization* ou *word segmentation*) qui est la segmentation d'un texte en mots ou items lexicaux (*tokens*). Ce type de sgmentation est aussi appelé ***segmentation lexicale***.

- ***La segmentation morphologique*** qui va plus loin que la segmentation lexicale en cherchant à isoler les différents constituants des items lexicaux en unités distinctes, plus petites, qui sont les morphèmes.
- ***Le chunking*** qui consiste à isoler les différents constituants du texte en unités indépendantes, supérieures aux mots, comme les propositions, les syntagmes etc. Ce type de segmentation est aussi appelé ***segmentation syntaxique***.

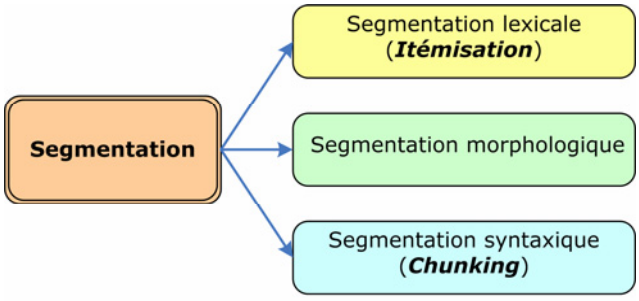

Figure 2. Les types de segmentation

Par segmentation, nous entendons ici la segmentation lexicale ou itémisation qui consiste à segmenter un texte en mots-formes ou items lexicaux. C'est une opération consistant à structurer le texte en passant d'un ensemble continu de caractères à une suite discrète d'items lexicaux.

La segmentation en arabe est basée sur la notion du *mot graphique* et de son analyse. Cette notion est féconde en traitement automatique de la langue arabe depuis les premiers travaux de David Cohen (Cohen, 1961 et 1970) jusqu'aux travaux autour de la base de connaissances DIINAR[1], en passant par les travaux de l'équipe SAMIA[2]. Elle est particulièrement productive en textométrie arabe. Selon que l'on s'attache à analyser le mot graphique, ou mot maximal, en termes de proclitiques, préfixes, base, suffixes et enclitiques, ou seulement en termes de mot minimal et de clitiques, l'on procède à une analyse morphologique pouvant aller jusqu'à la disjonction de la racine et du schème entrant en composition de la base, dans le premier cas, ou simplement à une analyse lexicale mettant en exergue les unités lexicales c'est-à-dire mots lexicaux et mots-outils, dans le second. En textométrie, c'est bien entendu la deuxième analyse qui est pressentie. De ce fait, segmenter un texte arabe, revient donc à analyser ses mots maximaux en mots minimaux et clitiques ; il est cependant évident que les mots graphiques qui sont dépourvus de clitiques sont des mots minimaux et leurs frontières demeurent donc inchangées.

La segmentation d'un texte arabe est donc cette opération qui consiste à repérer les unités lexicales en délimitant les frontières entre les mots et en fixant précisément les règles qui déterminent les unités segmentables et celles qui ne le sont pas[3].

[1] **DI**ctionnaire **IN**formatisé de l'**AR**abe.

[2] **S**ynthèse et **A**nalyse **M**orphologiques **I**nformatisées de l'**A**rabe.

[3] Pour les règles qui gèrent la segmentation en arabe voir le chapitre 5 intitulé « Norme de dépouillement » de notre thèse de Doctorat (Mouelhi, 2008).

4. L'analyse du mot graphique en arabe

Axés sur la notion du mot graphique, les travaux sur l'analyse automatique de l'arabe ont commencé vers le début des années soixantes. Le texte qui a donné le « coup d'envoi » de ces travaux fut l'article de David Cohen paru en 1961 dans la revue de l'Association pour le Traitement Automatique des LAngues naturelles (ATALA) et intitulé « *Essai d'une analyse automatique de l'arabe* ». Dans ce travail, repris et révisé près de dix ans plus tard[4], D. Cohen propose un schéma général des mots graphiques maximaux entièrement vocalisés analysables en constituants morphologiques ultimes. Cette analyse consiste à décomposer le mot graphique en racine, schème, base, préfixes, suffixes, antéfixes, postfixes[5]. Le *mot maximal* est donc cette unité décomposable en proclitiques, préfixes, base, suffixes et enclitiques. La concaténation des préfixes, de la base et des suffixes formant ce que D. Cohen appelle *mot minimal*, le *mot maximal* peut par conséquent être analysé en proclitiques, mot minimal et enclitiques.

Cependant, bien que précurseurs et productifs, les travaux de D. Cohen restent marqués par deux limites frappantes : la première est que ces travaux ont traité seulement de la langue arabe moderne écrite **entièrement vocalisée**, et la deuxième est qu'ils se situaient exclusivement au niveau de l'analyse et ne s'intéressaient pas à la synthèse dans le sens de la génération de mots minimaux ou maximaux à partir des leurs constituants. Ces deux limites ont d'ailleurs été évitées dans les travaux qui ont suivi et qui ont été, en quelque sorte, le prolongement des travaux de D. Cohen, c'est-à-dire les travaux de SAMIA et les travaux de DIINAR.

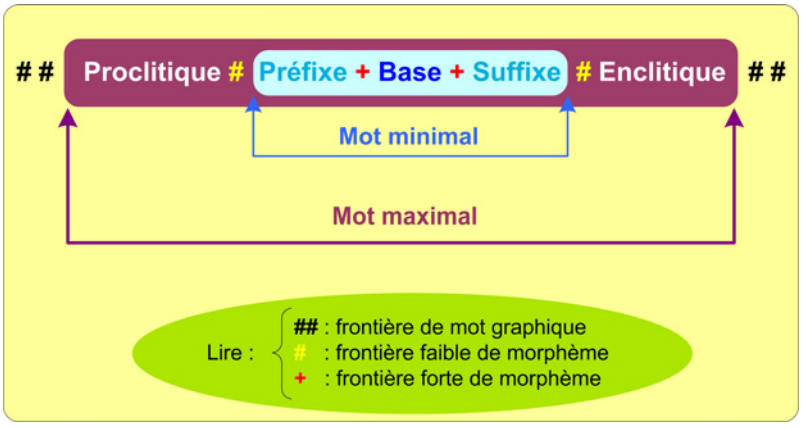

Figure 3. Schéma général du mot graphique en arabe

Le mot minimal est une unité libre minimale au sens de Bloomfield, c'est-à-dire qu'elle peut exister d'une façon autonome sans avoir besoin des clitiques. En revanche, la base, ou noyau lexical, ne peut se passer des affixes et n'est, de ce fait, pas considérée comme une forme libre minimale. C'est pour cette raison que les frontières de préfixe et de suffixe à l'intérieur du mot minimal sont dites des « frontières fortes ». Les frontières de clitiques, quant à elles, sont considérées comme des « frontières faibles ».

[4] D. Cohen, Etudes de linguistique sémitique et arabe, 1970.

[5] Les termes antéfixes et postfixes ont été remplacés par la suite par, respectivement, proclitiques et enclitiques. Voir à ce sujet : (J. Dichy et M. O. Hassoun, 1989).

5. Le segmenteur semi-automatique *AraSeg*

En plus de la grammaire segmentale dont nous l'avons doté, *AraSeg* est un segmenteur à dictionnaire en ce sens où, dans la mémoire du logiciel est stocké un dictionnaire (il s'agit d'un lexique généré à partir de la base de connaissance DIINAR) et la segmentation est faite par consultation de ce lexique généré. Ce lexique généré est sous forme d'une base de données Access® qui contient, entre autres tables et relations, la table des verbes, la table des noms, la table des déverbaux, la table des mots-outils, la table des proclitiques et la table des enclitiques.

Il est vrai que l'analyse de chaque séquence de caractères est basée sur le modèle d'analyse du mot graphique en arabe dans lequel un mot graphique est considéré comme une suite de constituants immédiats, l'analyseur (morphologique) ayant pour tâche d'identifier les constituants du mot en le décomposant en proclitiques, préfixes, base, suffixes et enclitiques et d'associer à chaque constituant sa ou ses catégories grammaticales ainsi que les traits qui lui sont associés ; mais ce qui nous intéresse dans ce type de segmentation, ce n'est pas ce niveau approfondi d'analyse morphologique : le rôle principal d'*AraSeg* est l'itémisation, la segmentation lexicale. C'est pour cette raison que nous n'utilisons qu'une version allégée de cette analyse du mot graphique en restant à la surface du niveau lexical en ce sens qu'il s'agit uniquement de savoir si le mot ou ses constituants appartiennent ou non à la langue pour pouvoir proposer le cas échéant une (des) segmentation(s) envisageable(s). Cet allègement a pour conséquence de réduire considérablement la taille du lexique généré et de rendre le segmenteur plus rapide et plus performant.

Aussi, devons-nous préciser que l'allègement n'est pas seulement en profondeur ; il est également horizontal parce que nous avons décidé de procéder à ce que l'on appelle l'*analyse unifiée*. En effet, s'agissant, rappelons-le encore une fois, d'une segmentation lexicale, autrement dit d'un découpage en surface du niveau lexical, la distinction à faire entre les différentes acceptions sémantiques, référentielles, etc. d'une unité lexicale n'est pas significative à ce stade-là. Il n'est pas capital de multiplier, par exemple, par huit les possibilités de segmenter un mot graphique commençant par le mot-outil **و** *wa* [et] pour la simple raison qu'il a en arabe huit **و** différents (coordonnant, particule de serment, particule d'accompagnement, *wâw rubba*[6], particule du chiffre 8, etc.) : une analyse unifiée serait ici très appréciée puisque dans les huit cas possibles nous n'obtenons qu'une seule manière de découper ce mot graphique indépendamment des différentes nuances. Dans cette perspective, l'analyseur ne présentera alors qu'une seule façon de segmenter de tels mots graphiques. Un autre exemple de l'*analyse unifiée* : considérons le mot graphique représentant la phrase suivante **أرِنَاهُمَا** *Þarinâhumâ* [montre-les-nous]. Cette phrase est composée du verbe *montrer* conjugué à la deuxième personne du singulier à l'impératif **أر** *Þari* [montre] et de deux compléments d'objet **هُمَا** *humâ* [les] et **نَا** *nâ* [nous]. Le complément d'objet **هُمَا** peut correspondre au pronom personnel complément de troisième personne du duel masculin ou à celui de troisième personne du duel féminin. Et le complément d'objet نَا peut correspondre au pronom personnel complément de première personne du pluriel masculin, de première personne du pluriel féminin, de première personne du duel masculin ou à celui de première personne du duel féminin. Si nous prenons en considération les variations en genre et en

[6] Le mot-outil *rubba*, en arabe *maint*, qui a, sur le plan syntaxique, cette spécificité de régir le cas indirect (génitif) pour le mot postposé, est souvent utilisé accompagné du coordonnant *wâw* و . Cette paire de mots-outils peut être remplacée, en arabe classique, par le seul *wâw* , appelé au demeurant *wâw rubba*, et héritant du pouvoir de *rubba* de régir le cas indirect.

nombre, la combinaison de ces deux compléments nous donne huit analyse possibles que nous résumons ainsi :

Montre-les {*duel*, *masculin*}-nous {*duel*, *masculin*}.	{*duel*, *masculin*}هُمَا{*duel*, *masculin*}أرنَا
Montre-les {*duel*, *masculin*}-nous {*duel*, *feminin*}.	{*duel*, *masculin*}هُمَا{*duel*, *feminin*}أرنَا
Montre-les {*duel*, *masculin*}-nous {*pluriel*, *masculin*}.	{*duel*, هُمَا{*pluriel*, *masculin*}أرنَا *masculin*}
Montre-les {*duel*, *masculin*}-nous { *pluriel*, *feminin*}.	{*duel*, *masculin*}هُمَا{*pluriel*, *feminin*}أرنَا
Montre-les {*duel*, *feminin*}-nous {*duel*, *masculin*}.	{*duel*, *feminin*}هُمَا{*duel*, *masculin*}أرنَا
Montre-les {*duel*, *feminin*}-nous {*duel*, *feminin*}.	{*duel*, *feminin*}هُمَا{*duel*, *feminin*}أرنَا
Montre-les {*duel*, *feminin*}-nous {*pluriel*, *masculin*}.	{*duel*, *feminin*}هُمَا{*pluriel*, *masculin*}أرنَا
Montre-les {*duel*, *feminin*}-nous {*pluriel*, *feminin*}.	{*duel*, *feminin*}هُمَا{*pluriel*, *feminin*}أرنَا

Dans ces huit analyses possibles, une seule et unique segmentation (découpage) résultante est opérée, quel est donc l'intérêt, dans la perspective d'itémisation, de faire figurer huit segmentations identiques si ce n'est le fait d'alourdir le segmenteur automatique ou semi-automatique sans raison valable pour ce genre d'entreprise ! C'est dans ces cas de figures que nous avons opté pour l'analyse unifiée : l'analyseur ne présentera alors qu'une seule façon de segmenter de tels mots graphiques.

La première opération par laquelle commence le segmenteur est bien entendu le repérage des mots graphiques. Les frontières de mots graphiques ne posent pratiquement pas de problèmes dans le système d'écriture de l'arabe. Les mots graphiques sont séparés par des espaces ou par des signes de ponctuation. Une spécificité est, tout de même, à noter concernant l'espace qui fonctionne en général comme séparateur de mots sauf dans les unités polylexicales où il est considéré comme composant de mots complexes. Le repérage et le traitement des unités polylexicales pose un vrai problème aux différents segmenteurs automatiques ou semi-automatiques. Certains sont équipés de dictionnaires de mots complexes qui leur permettent de détecter ces unités dans le corpus à segmenter et de les traiter par la suite en tant qu'unités à part entière.

Une fois le texte à segmenter chargé dans la machine et les frontières des mots graphiques repérées, le système commence à examiner, un à un, tous les mots graphiques du texte. Il consulte d'abord les tables des mots-outils, des proclitiques et des enclitiques aux entrées desquelles il compare les éventuels segments du mot à analyser, repérés suite à un découpage primaire. Le système vérifie ensuite les compatibilités entre tous ces éléments et il propose en fin de compte la ou les segmentations possibles après avoir vérifié et appliqué les règles de la grammaire segmentale. Les segmentations sont ainsi opérées l'une après l'autre, en flot linéaire, c'est-à-dire que dans chaque suite de caractères du texte, la fin du mot segmenté sera le début du mot suivant.

D'une façon plus précise, l'analyse d'un mot graphique sur le plan lexical, c'est-à-dire la vérification de son appartenance (lui ou ses constituants) à la langue, passe par les trois phases suivantes :

↳ **Découpage primaire :**

Cette opération fait appel à la table des proclitiques et à celle des enclitiques permettant de reconnaître et d'isoler les constituants immédiats du mot à vérifier. L'algorithme de découpage utilise des automates de reconnaissance construits à partir de ces tables et permettant de reconnaître un élément même s'il est totalement ou partiellement non voyellé. Le découpage fournit comme résultat trois listes, la première contient tous les proclitiques

reconnus à partir du début du mot, la deuxième tous les mots minimaux reconnus à partir de la fin des proclitiques, et la troisième tous les enclitiques reconnus à partir de la fin du mot.

↳ Vérification des compatibilités :

Considérée comme une vérification de surface, cette étape consiste d'abord à croiser deux à deux les listes obtenues à l'étape précédente pour n'en retenir que les couples compatibles, puis, toujours par croisement, ne retenir que les combinaisons compatibles (triplets). Ceci est fait à l'aide d'une matrice de compatibilité préalablement calculée. Cette matrice est, en fait, intégrée à une matrice de compatibilité plus large utilisée dans l'analyse morphologique : la matrice de compatibilité des prébases et des postbases. Les prébases (respectivement postbases) sont des combinaisons de proclitiques et de préfixes (respectivement suffixes et enclitiques) obtenues en respectant des règles de compatibilité et de cooccurrence issues de la grammaire des formants du mot, formants-noyau et formants-extensions (Dichy, 1990 et 1997).

↳ Interrogation du lexique généré :

Étant donné que chaque mot minimal, chaque proclitique et chaque enclitique correspond à une unité du lexique, toutes les combinaisons issues de l'étape précédente sont comparées, dans cette étape, aux éléments du lexique généré et ce jusqu'à ce qu'une combinaison soit validée : une segmentation est alors proposée en sortie. Si le mot graphique présente plusieurs solutions de segmentation, toutes les possibilités sont retenues et proposées en sortie.

Figure 4. Capture d'écran : écran d'accueil du segmenteur montrant le chemin du fichier à segmenter, le bouton du lancement de la segmentation et le champ affichant le n° du mot graphique eu cours de segmentation

Comme nous venons de le signaler, dans le cas où un mot graphique présente plus d'une segmentation, nous avons choisi d'afficher à l'écran toutes les possibilités de découpage et c'est à l'utilisateur de choisir la bonne segmentation comme le montre la figure suivante (figure 5). Dans la partie supérieure de cet écran, le logiciel présente le mot graphique sujet à une segmentation multiple (à droite), le nombre de segmentations proposées et le numéro de la segmentation encadré par deux flèches, droite et gauche, pour permettre de naviguer entre les différentes analyses possibles (au milieu) et enfin, le bouton qui permet d'enregistrer la segmentation choisie (à gauche). Au milieu de l'écran, est présentée la phrase contenant le mot graphique en question ou, quant celle-ci est trop longue (et c'est souvent le cas en arabe classique), le contexte le plus large possible qui permettrait à l'utilisateur de choisir en toute

connaissance de cause, la bonne segmentation. Dans la partie inférieure de l'écran, le logiciel affiche la segmentation (proclitiques + mot minimal + enclitiques) correspondant au numéro choisi par l'utilisateur, parmi les analyses proposées pour le mot graphique en cours de segmentation.

Figure 5. Capture d'écran : choix de segmentation à faire parmi deux propositions présentées par le segmenteur

En plus de la méthode par règles et celle par dictionnaire sur lesquelles est basée la segmentation, nous avons doté notre segmenteur d'une méthode par apprentissage. Il ne s'agit pas ici d'apprentissage à partir de grands corpus d'entraînement préalablement segmentés comme le font certains segmenteurs pour d'autres langues, ce qui serait un bon procédé, mais il s'agit plutôt d'un module, à la façon d'un OCR, qui reprend les mots graphiques qui n'ont pu être segmentés par le système puis l'utilisateur sélectionne, à partir d'un menu déroulant, un par un ces mots non segmentés et insère manuellement la segmentation voulue en saisissant directement sur clavier les proclitiques, le mot minimal et les enclitiques dans les champs correspondants comme le montre la figure suivante (figure 6). Une fois la segmentation validée, le segmenteur enregistre cette analyse pour l'appliquer à l'avenir, à chaque fois qu'il aura à segmenter le même mot graphique.

Figure 6. Apprentissage du segmenteur : segmentation manuelle des mots graphiques non analysés par le segmenteur et que ce dernier devra enregistrer

Nous avons décidé de séparer, lors de la segmentation, par un caractère spécial différent de l'espace, non seulement les mots graphiques mais aussi les mots minimaux et les clitiques résultant du découpage des mots maximaux. C'est pourquoi le segmenteur a été programmé pour insérer une barre oblique avant et après chacune des unités lexicales obtenues à la sortie de la segmentation : un exemple de fichier de sortie est présenté ci-après, dans la figure 8.

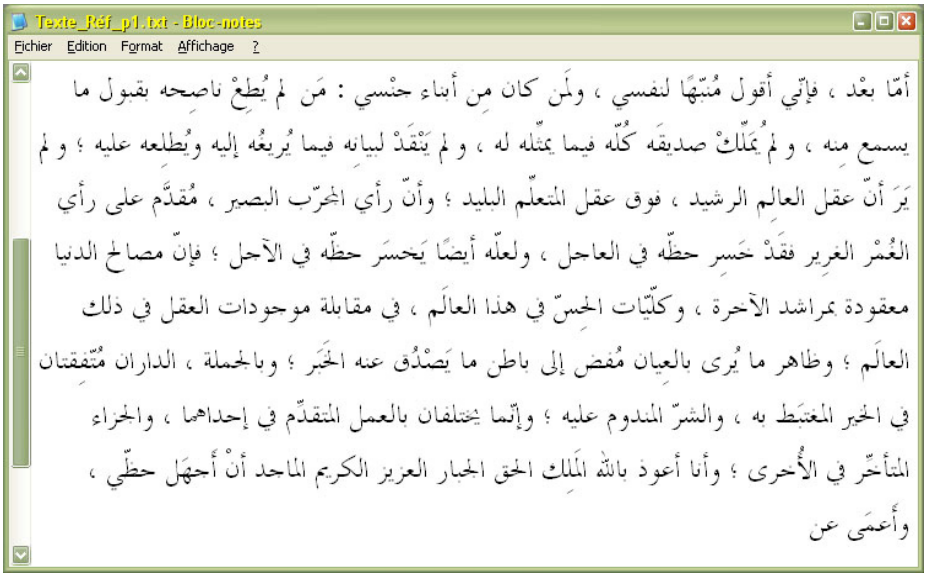

Figure 7. Fichier à segmenter avant de le soumettre au segmenteur

Le texte soumis à la segmentation doit être sous format texte seul « .txt » (figure 7) ; le fichier de sortie rendu par le segmenteur et contenant le texte segmenté est aussi sous format texte mais avec l'extension « .seg » (figure 8), une extension que nous avons créée à cet effet.

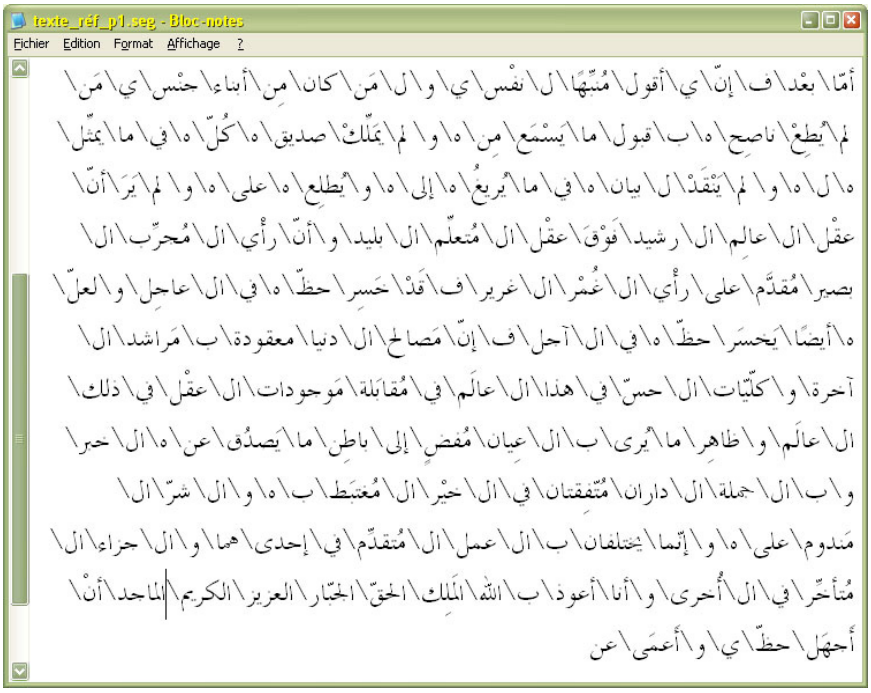

Figure 8. Le fichier de sortie après segmentation

En ce qui concerne les performances de notre segmenteur, elles sont visiblement correctes. Mais avant d'exposer les résultants obtenus en termes de pourcentage de réussite, il faut d'abord rappeler qu'une partie des mots graphiques, ceux qui présentent des analyses multiples, sont segmentés directement par l'utilisateur qui choisit, pour chaque mot graphique, la bonne segmentation parmi les différentes possibilités proposées par le segmenteur. Ce qui veut dire que pour ces mots graphiques, la segmentation validée est considérée comme totalement exacte parce que choisie en toute connaissance de cause par l'opérateur-linguiste (ou supposé comme tel). En revanche, les mots graphiques que le segmenteur a analysés et segmentés automatiquement sans demander l'avis de l'opérateur humain, peuvent comporter de toute évidence des erreurs qui peuvent être dues à plusieurs facteurs : elles peuvent être dues à un manque dans la base de données des mots-outils, à quelques anomalies au niveau de la grammaire segmentale qu'il faudra corriger, ou peuvent simplement provenir de quelques cas d'ambiguïtés effectives (Mouelhi, 2008). Les mots graphiques non analysés du tout, parce que considérés comme mots minimaux ou non reconnus par le segmenteur, peuvent également renfermer des erreurs ; ils sont en effet, réinjectés tels quels à leurs places respectives dans le texte de sortie. Après beaucoup de tests sur de petits textes, nous avons utilisé le segmenteur *AraSeg* pour segmenter un corpus[7] de 61 177 occurrences correspondant à 37 457 mots graphiques.

Sur les 37 457 mots graphiques que comporte le corpus, 34 027 ont été analysés par le segmenteur soit 90,84 % du corpus, et 3 430 mots graphiques n'ont pas pu être analysés soit 9,16 % de l'ensemble du corpus.

Pour juger de la fiabilité du segmenteur, nous avons donc calculé le taux d'erreur pour chacun des petits fichiers (correspondant aux pages du corpus) aussi bien pour les mots graphiques

[7] Il s'agit du premier volume d'un texte arabe classique (de trois volumes) intitulé « *Al-'ImtâÝ wa-l-Mu'ânasa* » de Abû Íayyân at-TawÎîdî (932-1024).

analysés que pour ceux que le segmenteur n'a pas pu segmenter. Le taux d'erreur général a par conséquent été calculé ; les résultats obtenus sont les suivants :

Pour les 90,84 % de mots graphiques qui sont analysés, le taux d'erreur varie entre 2,61 % et 16,86 % avec un taux moyen de 7,06 %. Rapporté à l'ensemble du corpus, ce taux d'erreur est donc de 6,41 %.

Parmi les 9,16 % de mots graphiques qui n'ont pas été analysés par le segmenteur, le taux de ceux qui auraient dû être segmentés est de 69 % (nous les considérons comme des erreurs). Rapporté à l'ensemble du corpus, ce taux d'erreur chute à 4,87 %.

Le taux d'erreur général est donc de 11,28 %. Autrement dit, l'efficacité du segmenteur est de l'ordre de 88,72 % (mots inconnus compris). Elle est tout de même de l'ordre de 93,59 % si nous ne considérons que les mots reconnus. Ce sont des performances largement correctes.

Le segmenteur a été écrit en Visual Basic et tourne sous Microsoft Windows. Pour l'instant, il est en usage interne ; mais il pourrait être mis à la disposition des chercheurs, sous une forme ou une autre, dans un futur proche, une fois que la question des droits aura été résolue.

Références

Berment V. (2004). *Méthodes pour informatiser des langues et des groupes de langues « peu dotés »*, Thèse de Doctorat, Université Joseph Fourier, Grenoble.

Cohen D. (1961). Essai d'une analyse automatique de l'arabe. In *ATALA* (revue de l'Association pour le Traitement Automatique des LAngues naturelles), Paris.

Cohen D. (1970). *Etudes de linguistique sémitique et arabe*. Mouton, La Hague - Paris.

Dichy J. (1990). *L'écriture dans la représentation de la langue : La lettre et le mot en arabe*, Thèse d'Etat, Université Lumière-Lyon 2.

Dichy J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. In *Méta*, vol. 42, N° 2, juin, Montréal, pp. 291-306.

Dichy J., Hassoun M. (1989). *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe, travaux SAMIA I*, Fondation postuniversitaire interculturelle (Conseil International à la Langue Française), Paris.

Kosawat K. (2003). *Méthodes de segmentation et d'analyse automatique de textes thaï*, Thèse de Doctorat, Université de Marne-La-Vallée.

Mouelhi Z. (2008). *Essai de lexicométrie d'une œuvre arabe : le vocabulaire de Tawhîdî dans al-'ImtâÝ wa-l-Mu'ânasa*, Thèse de Doctorat, Université Lumière-Lyon 2.

Palmer D. (2000). *Tokenisation and sentence segmentation*, Handbook of Natural Language Processing, Robert Dale et al. Editors

Zaafrani R. (2002). *Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère*, Thèse de doctorat, Université Lumière-Lyon2.