Segmentation de textes arabes en unités discursives minimales

Iskandar Keskes[1,2] Farah Beanamara[2] Lamia Hadrich Belguith[1]
(1) ANLP Research Group, MIRACL, Route de Tunis km 10, 3021, Sfax, Tunisie
(2) IRIT, 118, route de Narbonne F-31062 Toulouse Cedex 9
keskes@irit.fr, benamara@irit.fr, l.belguith@fsegs.rnu.tn

RÉSUMÉ

La segmentation d'un texte en Unités Discursives Minimales (UDM) a pour but de découper le texte en segments qui ne se chevauchent pas. Ces segments sont ensuite reliés entre eux afin de construire la structure discursive d'un texte. La plupart des approches existantes utilisent une analyse syntaxique extensive. Malheureusement, certaines langues ne disposent pas d'analyseur syntaxique robuste. Dans cet article, nous étudions la faisabilité de la segmentation discursive de textes arabes en nous basant sur une approche d'apprentissage supervisée qui prédit les UDM et les UDM imbriqués. La performance de notre segmentation a été évaluée sur deux genres de corpus : des textes de livres de l'enseignement secondaire et des textes du corpus Arabic Treebank. Nous montrons que la combinaison de traits typographiques, morphologiques et lexicaux permet une bonne reconnaissance des bornes de segments. De plus, nous montrons que l'ajout de traits syntaxiques n'améliore pas les performances de notre segmentation.

ABSTRACT

Segmenting Arabic Texts into Elementary Discourse Units

Discourse segmentation aims at splitting texts into Elementary Discourse Units (EDUs) which are non-overlapping units that serve to build a discourse structure of a document. Current state of the art approaches in discourse segmentation make an extensive use of syntactic information. Unfortunately, some languages do not have any robust parser. In this paper, we investigate the feasibility of Arabic discourse segmentation using a supervised learning approach that predicts nested EDUs. The performance of our segmenter was assessed on two genres of corpora: elementary school textbooks that we build ourselves and documents extracted from the Arabic Treebank. We show that a combination of typographical, morphological and lexical features is sufficient to achieve good results in segment boundaries detection. In addition, we show that adding low-level syntactic features that are manually encoded in ATB does not enhance the performance of our segmenter.

MOTS-CLÉS : Segmentation discursive, unité discursive minimale, langue arabe.
KEYWORDS : Discourse segmentation, Elementary discourse units, Arabic language.

1 Introduction

La segmentation discursive d'un texte vise à segmenter le texte en unités discursives minimales (UDM) qui ne se chevauchent pas. Ces unités sont ensuite reliées entre elles par des relations rhétoriques afin de construire la structure discursive du texte. La segmentation est donc une étape primordiale dans l'analyse du discours car des erreurs de segmentation peuvent dégrader les performances de l'analyseur. (Soricut et Marcu, 2003) ont montré qu'une bonne segmentation permet de réduire de 29% les erreurs de leur analyseur

© ATALA

discursif.

Une UDM peut être une phrase ou une proposition. Dans une phrase complexe, elle correspond généralement à des clauses verbales, comme *[Un film d'horreur] [qui m'a fait peur]* où la proposition relative introduite par le pronom relatif indique un point de coupure. Ces deux segments sont ainsi reliés par la relation d'élaboration. Une UDM peut aussi correspondre à d'autres unités syntaxiques décrivant des éventualités, comme des syntagmes nominaux ou des syntagmes prépositionnels, comme dans *[Après quelques minutes,] [nous avons trouvé les clés sur la table]* où nous avons une relation d'encadrement temporelle entre ces deux UDM. Une UDM peut être aussi structurellement emboitée dans une autre pour rendre compte des cas d'appositions, de constructions clivées ou encore des cadres adverbiaux, comme dans *[M. Dupont, [un homme d'affaire riche,] a été sauvagement tué]* où « un homme d'affaire riche, » est une apposition.

Plusieurs travaux ont été menés sur la segmentation automatique de discours dans différentes langues. Chaque segmenteur a sa propre définition d'UDM car le repérage des limites des segments dépend principalement de la théorie utilisée. En effet, chaque théorie du discours définit ses propres guides de segmentation. Globalement, la segmentation automatique de discours peut être effectuée selon des techniques à base de règles ou en utilisant des techniques d'apprentissage. Dans la première approche, des règles empiriques identifient les bornes *début* et *fin* de segments en s'appuyant sur une combinaison d'indices de surface (les ponctuations et les marqueurs lexicaux), des informations morphologiques et des informations syntaxiques. Pour l'anglais, citons les travaux de (Le Thanh et al., 2004) qui ont obtenu une F-mesure de 86,9% sur le corpus RST Discouse Treebank (Carlson et al., 2003). (Tofiloski et al., 2009) ont proposé le système de segmentation SLSeg basé sur une analyse syntaxique. Ce système a obtenu une F-mesure de 80-85%. Les approches symboliques ont également été utilisées pour réaliser la segmentation en UDM pour d'autres langues comme l'allemand (Lüngen et al., 2006), l'espagnol (Da Cunha et al. 2010) et le japonais (Sumita et al., 1992). La plupart de ces travaux définissent les UDM dans le cadre de la RST (Mann et Thompson, 1988).

Les méthodes d'apprentissage exploitent le plus souvent des traits lexicaux et syntaxiques pour classer chaque mot de la phrase comme étant une frontière d'une UDM ou non. Toujours dans le cadre de la RST, (Soricut et Marcu, 2003) ont décrit comment segmenter les phrases en UDM sur la base de l'analyseur SPADE, tout en exploitant une analyse syntaxique extensive. (Sporleder et Lapata, 2005) ont montré que l'emploi d'une analyse lexicale couplée à une analyse syntaxique surfacique (catégorie grammaticale et chunk) sont suffisantes pour obtenir de bons résultats. (Fisher et Roark, 2007) ont proposé diverses améliorations de SPADE en utilisant une analyse à états finis. (Subba et Di Eugenio, 2007) ont utilisé un réseau de neurone. Pour les autres langues, citons (Jirawan et al., 2005) pour le thaï qui utilise un système d'apprentissage par arbre décisionnel associé à des règles.

Toutes les approches d'apprentissage mentionnées plus haut réduisent la tâche de segmentation à une classification binaire en ignorant les UDM emboitées[1]. Afin de prédire ce type d'UDM, (Afantenos et al., 2010) ont utilisé un classifieur (Maximum Entropy Model) à quatre classes où chaque mot peut être classé au début de l'UDM, à la fin de l'UDM, au milieu ou bien au début et à la fin de l'UDM. Ce classifieur utilise une combinaison de traits lexicaux

[1] La RST traite le cas des segments emboités lors de la détermination des relations (Mann et Thompson, 1988).

© ATALA

(principalement des n-grams et un lexique de marqueurs discrusifs) et syntaxiques (chunk, catégorie grammaticale et chemin de dépendance) et a été évalué sur le corpus ANNODIS (Afantenos et al., 2012), un corpus pour la langue française annoté discursivement selon les principes de la théorie de la représentation discursive segmentée (SDRT) (Asher et Lascarides, 2003). Dans ANNODIS, la proportion d'UDM emboîtées dépasse les 10%. Le classifieur obtient une F-mesure de 58%. Une étape de correction qui consiste en l'ajout des limites manquantes d'UDM améliore sensiblement les résultats de 15%.

Qu'elles soient à bases de règles ou à base de techniques d'apprentissage, la plupart des approches actuelles utilisent une analyse syntaxique extensive. Cependant, plusieurs langues ne disposent pas encore d'un analyseur syntaxique robuste. La question qui se pose alors est comment concevoir un découpage automatique en UDM robuste pour ces langues sans utiliser d'informations syntaxiques ? Dans cet article, nous allons montrer la faisabilité de la segmentation discursive en UDM pour la langue arabe dans le cadre de la théorie SDRT (Asher et Lascarides, 2003), en proposant une méthode d'apprentissage supervisée multi-classes qui prédit les UDM imbriquées. À notre connaissance, ceci est le premier travail qui traite la segmentation discursive pour la langue arabe. Pour ce faire, nous utilisons deux genres de corpus qui ont un style d'écriture différent : des textes de livres de l'enseignement secondaire tunisien (TES) et des textes de journaux annotés syntaxiquement, issus du corpus Arabic TreeBank (ATB part3 v3.2) (Maamouri et al., 2010b). Nous montrons que l'utilisation de traits typographiques, lexicaux et morphologiques est suffisante pour obtenir de bons résultats. De plus, nous montrons que l'utilisation de traits syntaxiques de surface (chunks) n'améliore pas les résultats. Nos résultats montrent que la segmentation du discours en langue arabe est réalisable sans faire recours à la syntaxe

Cet article est organisé comme suit. Nous commençons par exposer les principales difficultés de la segmentation discursive en langue arabe. Nous présentons ensuite les principaux travaux existants dans ce domaine. La section 4 présente notre corpus, le manuel de segmentation ainsi que les résultats de l'annotation manuelle. La section 5 détaille notre méthode de segmentation ainsi que les traits utilisés. Les résultats obtenus sont discutés dans la section 6.

2 Segmentation discursive de textes arabes

Vue la richesse des propriétés morphologiques et syntaxiques de la langue arabe standard moderne (ASM)[2], la segmentation en UDM est une tâche difficile. En effet, contrairement aux langues indo-européennes, la langue arabe n'admet pas de lettres majuscules ce qui rend la tâche de segmentation plus difficile que pour les autres langues, comme le français. De plus, la ponctuation n'est pas utilisée d'une façon systématique ce qui complique la détermination des frontières des segments. Ainsi, le discours arabe tend à utiliser des phrases longues et complexes au point qu'on peut souvent trouver une page sans aucun signe de ponctuation.

Comme les autres langues sémitiques, la langue arabe a une morphologie riche et complexe. Les mots sont formés par un processus de concaténation séquentiel de trois composants (préfixe + racine + suffixe). Ces composants ont des caractéristiques morphologiques et syntaxiques qui varient selon le contexte du mot. Les suffixes et les

[2] Pour plus de détail sur l'ASM et le traitement automatique de la langue arabes voir (Habash, 2010).

© ATALA

affixes peuvent être des prépositions, des conjonctions ou des pronoms. Par exemple, la préposition (comme ف/"fa"/puis), la conjonction (comme و/"wa"/et), l'article (comme ال/"Al"/le) et le pronom (comme ه/"ho"/il) peuvent être affixés à un nom, un adjectif, une particule ou un verbe ce qui induit une très grande ambigüité à la fois lexicale et morphologique. Par exemple, le mot فهم / "fahm", peut être un verbe (comprendre), un nom (compréhension) ou une conjonction (ف/"fa"/puis) suivie d'un pronom (هم/"hom"/ils). Enfin, un mot peut avoir plusieurs affixes et suffixes, comme par exemple, "استتذكرونها" ([ا/"A"/Est-ce-que], [س/"Sa"/allez], [تتذكر/"tata*k~ar"/rappeler], [ن/"na"/vous] et [ها/"hA"/elle]) qui représente en français « Est-ce que vous allez vous rappeler d'elle ? ». Cette richesse morphologique rend la tâche de segmentation beaucoup plus difficile, surtout lors du repérage de marqueurs lexicaux qui sont, en général, de bons indicateurs pour la détermination automatique des frontières de segment.

Une autre spécificité qui s'ajoute, est que le système d'écriture de l'arabe est diacritique. En effet, l'alphabet arabe est composé uniquement de consonnes et chaque consonne peut avoir différentes prononciations. Pour surmonter ce problème, les symboles orthographiques, appelés signes diacritiques sont utilisés. Les signes diacritiques représentent, entre autres, les voyelles courtes. Actuellement, la plupart des documents arabes ne sont pas accompagnés par des signes diacritiques. Il faut noter que les textes non diacritiques sont très ambigus et la proportion des mots ambigus dépasse 90% (Debili et al., 2002). Par exemple, le mot كتب/"ktb" peut être écrit sous 21 formes morphologiques différentes (كَتَبَ/"kataba"/il-écrit et كُتُبٌ/"kutubN"/livres) (Debili et al., 2002). L'exemple suivant montre un cas d'ambigüité.

(1) وصف الطبيب للمريض مجموعة من الأدوية لمعالجة ألمه وجرحه.

Le médecin a prescrit au patient une ordonnance pour traiter sa douleur et sa blessure.

Dans cet exemple, si une analyse automatique reconnait le mot جرحه/"jerHihi" comme un verbe (blesser), nous aurons une erreur de segmentation puisque ce mot est un nom (blessure). Le point de coupure ici, devrait être le mot لمعالجة/"limeEalajati"/pour-traiter car le marqueur de discours ل/"li"/pour est un bon indicateur pour la relation *But*.

Enfin, l'ordre des mots en langue arabe est relativement flexible. En effet, le changement de position de certains mots ne change pas forcément le sens de la phrase. Par exemple, la phrase « l'enfant va à l'école» peut être écrit en langue arabe sous trois formes: « ذهب الولد إلى المدرسة », « الولد ذهب إلى المدرسة » et « إلى المدرسة ذهب الولد ».

3 Travaux existants

La plupart des travaux en segmentation discursive de textes arabes traitent la segmentation en paragraphes, phrases ou clauses. (Belguith et al., 2005) ont proposé une approche à base de règles pour segmenter des textes arabes non-voyellés en phrases. L'approche consiste en une analyse contextuelle des signes de ponctuation, des conjonctions de coordination et une liste de particules qui sont considérées comme des critères de segmentation. Les auteurs ont déterminé 183 règles implémentées par le système STAr. (Touir et al., 2008) ont proposé une approche par règles guidée uniquement par des connecteurs lexicaux (la ponctuation n'est pas prise en compte) pour segmenter les textes arabes en clauses. Les auteurs introduisent la notion des connecteurs actifs, qui indiquent le début ou la fin d'un segment

© ATALA

et la notion de connecteurs passifs qui n'impliquent pas un point de coupure. Le même connecteur peut être actif ou passif en changeant d'un contexte à un autre. (Khalifa et al., 2011) ont proposé une méthode d'apprentissage pour la segmentation des textes arabes en clauses en exploitant uniquement les fonctions rhétoriques du connecteur "و/et". Les auteurs ont défini six sens pour ce connecteur : (1)والقسم/"wAw Aloqasam", (2)ورب/"wAw rob~a", (3)والاستئناف/"Alo<isti'onAf", (4)والحال/"wAw AloHAl", (5)والمعية/"wAw AlomaEiy~at" et (6)والعطف/"wAw AloEaTof". Parmi ces six sens, deux classes ont été définies : «Fasl » (1, 2 et 3), qui est un bon indicateur de segmentation, et «Wasl» (4, 5 et 6) qui n'a pas d'effet sur la segmentation. Un ensemble de 22 traits syntaxiques et sémantiques, ont ensuite été utilisées afin de classer automatiquement chaque instance du connecteur "و" dans ces deux classes. Enfin, (Keskes et al., 2012) ont utilisé une approche à base de règles pour la segmentation de textes arabes en clauses. Trois principes de segmentation ont été proposés : (p1) en utilisant uniquement des signes de ponctuation (21% de F-mesure), (p2) en s'appuyant uniquement sur des indices lexicaux (53,5% de F-mesure) et (p3) en combinant les signes de ponctuation et les indices lexicaux afin de faire face à l'ambigüité des indices lexicaux (68% de F-mesure).

À notre connaissance, le travail le plus proche du notre est celui de (Al-Saif et Markert, 2011) qui proposent d'identifier automatiquement le rôle discursif des connecteurs de discours puis de repérer les relations explicitement marquées dans le corpus ATB v 2.0 part 1[3]. Les auteurs utilisent les principes d'annotation du Penn Discours Treebank (PDTB) (Prasad et al., 2008). Nous rappelons que les segments du discours dans PDTB sont généralement des unités plus grandes que les UDM. En effet, ces unités peuvent être une clause ou un ensemble de clauses. La segmentation dans PDTB nécessite trois étapes: (a) l'identification du connecteur de discours (explicite et implicite), (b) l'identification des deux arguments de ce connecteur (à savoir Arg1 et Arg2) et (c) le repérage des frontières de ces arguments. Arg1 peut être situé dans la même phrase que le connecteur discursif ou dans la ou les phrases précédentes. Lorsqu'Arg1 et Arg2 sont dans la même phrase, on peut avoir plusieurs cas: Arg1 apparait devant Arg2, Arg1 venant après Arg2 et Arg2 emboitée dans Arg1 comme dans l'exemple (2).

(2) [ان الأطفال متعبون[و يشعرون بالنعاس]arg2 خلال الدرس .]arg1

[Les enfants sont fatigués [et ont envie de dormir]$_{arg2}$ pendant le cours.]$_{arg1}$

En cas d'emboîtement de segment (connecteurs de subordination, connecteurs de coordination et adverbes de discours), l'arbre syntaxique complet de la phrase sera nécessaire afin d'extraire Arg1 et Arg2 (Lee et al., 2008). (Al-Saif et Markert, 2011) n'ont décrit que l'étape (a) relative à l'identification des connecteurs et n'ont pas traité les UDM emboitées. De plus, ils n'ont donné aucune indication sur la façon dont les étapes (b) et (c) peuvent être réalisées automatiquement.

4 Segmentation manuelle

4.1 Corpus

Nous avons utilisé deux genres de corpus qui ont un style d'écriture différent : des textes de livres de l'enseignement secondaire tunisien (TES) et des textes de journaux annotés

[3] Nous utilisons le corpus ATB v3.2, c'est une version révisée de ATB v2.0 utilisé par (Al-Saif et Markert, 2011)

© ATALA

syntaxiquement du corpus Arabic TreeBank (ATB part3 v3.2) (Maamouri et al., 2010b). Les documents du corpus TES sont généralement bien structurés et non-voyellés. Les phrases sont courtes (environ 5,6 mots par phrase) avec une structure syntaxique simple. Ils sont caractérisés par la présence régulière de signes de ponctuation. Les documents sont également courts. Nous avons collectés 34 documents pour le corpus TES.

Le corpus ATB v3.2 part3 est composé de 599 textes du journal Al Nahar. Chaque document dans ce corpus est associé à deux niveaux d'annotation. D'abord, une annotation morphologique fournie pour chaque mot des informations morphologiques, sa translittération et sa traduction en anglais. Le second niveau comporte l'annotation syntaxique de chaque phrase du texte sous forme d'arbre syntaxique. Contrairement aux TES, les textes ATB sont plus longs et les phrases sont syntaxiquement plus complexes. Nous avons choisi au hasard 16 documents de l'ATB.

4.2 Manuel et guide d'annotation

Notre manuel est inspiré du manuel de segmentation élaboré par les partenaires du projet ANNODIS (Afantenos et al., 2012) et qui explique le principe de segmentation de textes pour le français. Nous avons repris ce manuel et l'avons adapté à la spécificité de la langue arabe.

Les UDM sont délimités par des crochets. Par convention, les connecteurs de discours sont toujours au début d'un segment alors que les signes de ponctuation qui délimitent les frontières de segments apparaissent toujours avant la fin d'un segment. Les UDM ne peuvent pas se chevaucher, mais elles peuvent être emboitées les unes aux autres (les doubles crochets ne sont pas autorisés), comme dans l'exemple suivant:

(3) [أصلح الأستاذ الامتحان،[الذي أجراه التلاميذ الأسبوع الماضي،] خلال حصة الدرس .]

[L'enseignant a corrigé l'examen, [qui a été donné aux étudiants la semaine dernière,] pendant le cours.]

Une UDM est essentiellement une clause verbale (comme dans l'exemple (4)) ou une clause nominale (مبتدأ/ « mubotada» et خبر/ « xabar», comme dans l'exemple (5)). Un point de coupure ne peut jamais séparer un verbe de son complément ou un sujet de son verbe. Aussi, un point de coupure ne peut jamais se produire au sein d'un chunk ou d'une entité nommée.

(4) [قصفت طائرات أميركية مجمعات من الكهوف.]

[Des avions américains ont bombardé un ensemble de grottes.]

(5) [كانت الطفلة جميلة.]

[La fille était belle.]

Nous présentons, ci-dessous, quelques principes de notre segmentation.

- Cas des conditionnels (شرط/"$aroT"). On segmente toujours dans ces cas, comme dans l'exemple suivant :

© ATALA

(6) [إذا أصبح الطقس جميل،][سأخرج أتنزه.]

[S'il fait beau,] [je vais faire une promenade.]

- Cas des corrélations (تلازم/"talAzum"). On segmente toujours dans ces cas, comme dans l'exemple suivant :

(7) [كلما أطالع الكتب،][كلما أتعلم المزيد من المصطلحات]

[Plus je lis des livres,] [Plus j'apprends de nouveaux termes.]

- Cas des coordinations (ربط/"rabot"). En langue arabe, la coordination est indiquée par des marqueurs tels و /"wa"/et, بحيث/"bihayv"/donc, ل/"li"/pour ... qui sont très ambigus. Par exemple, la conjonction (و/"wa"/et) peut avoir six sens différents (Khalifa et al., 2011) (voir section 3). En présence de la coordination, nous segmentons dans quatre cas: (i) la coordination entre des clauses indépendantes, (ii) la coordination entre des clauses subordonnantes, (iii) lorsque deux clauses verbales partagent le même objet ou le même sujet, comme dans l'exemple (8), et enfin, (iv) la coordination entre des syntagmes prépositionnels qui introduisent des événements, comme dans l'exemple (9). Nous ne segmentons pas dans tous les autres cas (comme dans l'exemple (10), où nous avons une conjonction entre deux objets du même verbe).

(8) [استعاد الرئيس التونسي عافيته][وقام باستقبال المواطنين.]

[Le Président tunisien est rétabli] [et a commencé à recevoir les citoyens.]

(9) [أعلنت الحكومة عدم موافقتها على التحاور] [لعدم توفر الشروط الأزمة.]

[Le gouvernement a annoncé qu'il refuse la négociation] [à cause de l'insuffisance des conditions requises.]

(10) [اتخذ الملك كل الترتيبات ومعدات السلامة.]

[Le roi a pris toutes les dispositions et les mesures de sécurité.]

- Cas des subordinations (صلة/"silat"). Nous segmentons toujours dans ces cas. Ils sont introduits par : (a) des conjonctions de subordination comme أن/">un"/pour, أن/">~aun"/que, إن/"<ino"/si, سوى/"Siwa"/ sauf et إلا/"<IoA"/moins (qui sont généralement utilisés après un verbe de communication ou lors d'un discours rapporté (comme dans l'exemple (12))), (b) des pronoms relatifs الذي/"ala* y"/qui, التي/"alaty"/qui ... (comme dans l'exemple (11)), ainsi que (c) par des marqueurs de subordination temporelle et/ou causale comme قبل أن/"qabola> un"/avant-que, غير أن/"gayora un~a>"/alors-que, حين/"Hiyna"/quand et لأن/"li>~aun"/parce-que.

(11) [يحتوى كتاب التكليف [الذي وجه الى الحكومة الجديدة ،] على كل الترتيبات المتخذة.]

[Le livre de référence [qui a été envoyé au nouveau du gouvernement,] contient toutes les

© ATALA

dispositions qui ont été prises.]

(12) [وقال وزير الدفاع] [ان ستة مسؤولين اميركيين وصلوا الى البلاد.]

[Le ministre de la Défense a dit que] [six fonctionnaires américains sont arrivés au pays.]

- Cas des appositions (بدل/"badal"). Nous segmentons dans la plupart des cas. Les appositions peuvent être des phrases adjectivales, des locutions adverbiales ou des groupes nominaux ou verbaux introduits par des pseudo-verbes comme إن/"<~un"/c'est-le, ليت/"layta"/espérer, لعل/"laEal~a"/peut-être. Les locutions adverbiales sont introduites par des adverbes relatifs tels que متى/"Matay"/quand, كيف/"kayfa"/comment, لماذا/"lima*A"/pourquoi, حيث/ "Hayvu"/où ou des adverbes réguliers tels que حينذاك/"AkaHiyna*"/à-cet-instant, وقتذاك/"waqta*Aka"/à-cet-instant et ربما/"rub~Ama"/peut-être). L'exemple (13) montre un cas d'une locution adverbiale. Les syntagmes prépositionnels (introduits par إلى/"<Ilay"/jusqu'à-ce-que, عن/"Ean"/de, فى/"fiy"/dans, من/"min"/de et على/"EalaY"/sur) qui apparaissent à la fin d'une clause ne sont pas segmentés.

(13) [ان الجنود، [حيث سيكونون مسلحين،] يستطيعون الدفاع عن انفسهم.]

[Les soldats, [quand ils seront armés,] seront en mesure de se défendre.]

- Cas des adverbiaux (ظرفية/"Zarofiy~at"). Dans certains cas, un adverbial peut être une UDM. Cela concerne les adverbiaux qui introduisent un événement ou un état. L'exemple (14) montre un cas de la relation But, alors que, l'exemple (15) présente un cas d'adverbial qui est en début de la phrase et qui indique une relation de Frame.

(14) [رجعت مسرعا إلى البيت][بسبب تهاطل الأمطار.]

[Je suis retourné rapidement à la maison] [à cause de la pluie.]

(15) [عندما توفي جدي،][كنت صغيرا جدا.]

[Quand mon grand-père est décédé,] [j'étais très jeune.]

- Nous segmentons en cas de discours rapporté entre guillemets et pronoms possessifs (comme dans l'exemple (16)) car ils indiquent respectivement la relation attribution et la relation élaboration d'entité. Nous ne segmentons pas en cas de translittération en caractères latins, d'abréviations et en cas des pronoms démonstratifs (هذا/"h`*A"/ce, هذه/"h`* ihi"/ce …).

(16) [وقدّمت لنا صحنا صغيرا][فيه مقروضات شهيّة.]

[et elle nous a donné un petit plat] [contenant des gâteaux délicieux.]

© ATALA

4.3 Calcul de l'accord inter-annotateur

Deux annotateurs natifs arabes ont annoté notre corpus selon les orientations définies dans le manuel d'annotation. Les annotations ont été réalisées en deux étapes. D'abord une phase de formation où les annotateurs ont été invités à annoter 4 documents du corpus TES puis 4 documents du corpus ATB (les deux corpus sont non-voyellés). Cette étape a permis de réviser le manuel d'annotation. Ensuite, chaque annotateur a annoté séparément 5 documents du corpus TES (ayant une moyenne de 20 phrases par document) puis 2 documents du corpus ATB (ayant une moyenne de 35 phrases par document). Les documents utilisés lors de la phase d'entraînement n'ont pas été pris en considération dans les étapes suivantes. La phase d'entraînement pour le corpus ATB a été plus longue que celle pour le corpus TES car ses documents sont plus longs et ses phrases sont plus complexes. Nous obtenons une mesure kappa de l'accord inter-annotateur de 0,83 pour ATB et 0,89 pour TES. Les principaux cas de désaccord proviennent de l'ambiguïté lexicale, en particulier pour les marqueurs discursifs.

Compte tenu des bons résultats d'accord, les annotateurs ont ensuite été invités à construire notre corpus de référence par consensus. Sur un nombre total de 706 UDM pour le corpus ATB nous avons 13,17% UDM imbriquées et sur 924 UDM pour le corpus TES nous trouvons 9,30% UDM imbriquées. Le tableau 1 présente les caractéristiques du corpus de référence.

	Textes	UDM	UDM emboîtées	Mots+ponctuations
TES	25	924	86	6437
ATB	10	706	93	7600
Total	**35**	**1630**	**179**	**14037**

TABLE 1 – Caractéristiques du corpus de référence

5 Traits d'apprentissage

Pour identifier les limites des UDM, nous avons conçu quatre groupes de traits d'apprentissage: typographiques, lexicaux, morphologiques et syntaxiques. Un vecteur de caractéristiques est associé à chaque token (mot ou ponctuation).

Traits typographiques : lors de la campagne d'annotation, nous avons identifié deux Catégories de Signes de Ponctuation (CSP): les ponctuations *fortes* qui identifient toujours la fin ou le début d'un segment (comme « : ») et les ponctuations *faibles* qui ne correspondent pas toujours à la limite de segment (comme « , »). Nous avons trois traits typographiques: (1) **PUNC** : la CSP du mot à classer, (2) **PPUNC** : la CSP du mot qui précède le mot à classer et (3) **FPUNC** : la CSP du mot qui suit le mot actuel. La CSP peut prendre trois valeurs : 0 si le mot n'est pas un signe de ponctuation, 1 s'il s'agit d'une ponctuation forte et 2 s'il s'agit d'une ponctuation faible.

Traits lexicaux : Nous considérons deux types d'indices lexicaux : des connecteurs discursifs comme حيث/"Hayovu"/où, بينما/"bayonamA"/alors-que et ل/"li"/pour et un ensemble de mots spécifiques, appelés indicateurs qui sont importants pour le processus de segmentation. Les connecteurs peuvent être des verbes d'attitude propositionnelle (par exemple قال/ "Qala"/dire, أعلن/">aEolana"/annonce, إعتقد/"<iEotaqada"/croire, …), des adverbes (par exemple بعد/"baEoda"/après, قبل/"qabola"/avant, من المفروض/"mina AalomaforuWD"/normalement, فقط/"faqaT"/uniquement), des conjonctions (par exemple

© ATALA

حالما/"HaAlamA"/dès-que et طالما/"Talama"/tant-que) et des particules (par exemple لم/"lam"/non et لن/"lan"/jamais). Comme les signes de ponctuation, nous avons deux Catégories d'Indices Lexicaux (CIL) : forts et faibles. Dans la première classe, les connecteurs sont généralement suivis d'un verbe qui est un indice fort pour la détermination du début de segment (comme لأن/"li>ana"/parce-que). Dans la deuxième classe, les connecteurs ambigus ne marquent pas toujours le début d'un segment (comme حيث/"Hayovu"/où). Nous avons quatre traits lexicaux : (1) **LEX** : la CIL du mot à classer ; (2) **PLEX** : la CIL du mot qui précède le mot actuel ; (3) **FLEX** : la CIL du mot qui suit le mot actuel et (4) **Blex :** un booléen qui indique si le mot courant commence par un connecteur ou un indicateur. Cette dernière caractéristique traite des cas d'agglutination. La CIL peut prendre cinq valeurs : 0 si le mot n'est pas un indice lexical, 1 si le mot est un indice de discours fort, 2 si le mot est un indice de discours faible, 3 si le mot est un indicateur fort et 4 si le mot est un indicateur faible.

Pour gérer à la fois les caractéristiques typographiques et lexicales, nous avons construit un lexique des indices de segmentation où chaque entrée est caractérisée par son type (signe de ponctuation, indice de discours, et indicateur), sa nature (forte ou faible) et la liste des catégories morphologiques possibles. Nous avons également indiqué si l'entrée lexicale est composée d'autres mots, comme خلاصة القول/"xelASita Aaloqawoli"/en-résumé. Si c'est le cas, nous détaillons chaque mot de cette entrée lexicale. Nous avons associé à chaque entrée sa traduction en anglais et un exemple de son utilisation. Notre lexique contient 174 entrées : 11 signes de ponctuation et 163 indices lexicaux (83 indices discursifs et 80 indicateurs), parmi lesquels 76,4% sont forts et 23,6% sont faibles.

Traits morphologiques : nous avons utilisé SAMA 3.1 qui est une mise à jour de l'analyseur morphologique pour l'arabe (BAMA 2.0) (Maamouri et al., 2010a). SAMA 3.1 considère chaque mot comme préfixe+racine+suffixe et énumère toutes les solutions possibles d'annotation, avec l'affectation de tous les signes diacritiques. Pour chaque mot, nous avons 10 caractéristiques morphologiques : (1) **LEM** le lemme du mot, (2) **POS** la catégorie morphologique du mot, (3) **COV** la vocalisation du mot, (4) **PREF**, (5) **SUFF** et (6) **ROOT** qui indiquent respectivement le préfixe, le suffixe et la racine du mot, (7) **PREF_POS**, (8) **SUFF_POS** et (9) **ROOT_POS** qui indiquent respectivement la catégorie morphologique du préfixe, du suffixe et de la racine, et finalement (10) **GLOSS**, qui indique la traduction en anglais du mot. Toutes ces caractéristiques sont générées par SAMA sous forme translittérée (codé en ASCII).

Traits syntaxiques : ces traits sont extraits à partir des annotations syntaxiques de l'ATB. Les documents du corpus TES ne sont pas concernés. Nous n'avons qu'un seul trait qui spécifie si le mot à classer est au début, à la fin ou au milieu d'un chunk.

6 Évaluation et résultats

Nous avons effectué un apprentissage supervisé en utilisant le classifieur Stanford basé sur le modèle d'entropie maximale (Berger et al., 1996). Chaque mot peut appartenir à l'une des trois classes suivantes: ***début***, si la UDM commence par ce mot, ***fin*** si elle se termine par ce mot ou ***milieu***, si le mot est au milieu de l'UDM. Nous n'avons pas trouvé de problèmes liés au déséquilibrage de la fréquence des classes lors de l'apprentissage. Le tableau 2 présente la fréquence de chaque classe dans le corpus TES et le corpus ATB.

© ATALA

	TES	ATB
Milieu	4589	6188
Début	924	706
Fin	924	706
Total	**6437**	**7600**

TABLE 2 – Fréquence des classes dans le corpus de référence

Afin de mesurer l'impact des traits morphologiques et syntaxiques sur les performances de notre segmentation, nous avons conçu deux classifications : (C1) utilise les traits typographiques, lexicaux et morphologiques et (C2) utilise tous les traits, y compris les traits syntaxiques. Nous avons également testé nos résultats par rapport à deux Baseline : une (B1) qui consiste à utiliser que le trait typographique PUNC, et l'autre (B2) qui combine le trait PUNC et le trait lexicale LEX. Les résultats présentés dans le tableau 3, le tableau 4 et le tableau 5 sont les moyennes de 5 validations croisées en testant à chaque fois sur 20% du corpus de référence (2 textes du corpus ATB et 5 textes du corpus EST). Pour mesurer l'effet de chacun de ces types de traits, nous présentons les résultats en commençant par les traits typographiques, puis nous ajoutons un à un les autres traits. Les valeurs du tableau 3 représentent la moyenne des trois classes : début, milieu et fin.

		TES			ATB		
		Précision	Rappel	F-score	Précision	Rappel	F-score
Traits typographiques	PUNC (B1)	0.450	0.416	0.432	0.267	0.287	0.277
	+PPUNC, FPUNC	0.575	0.453	0.506	0.281	0.332	0.304
PUNC+LEX (B2)		*0.547*	*0.511*	*0.528*	*0.479*	*0.436*	*0.456*
Traits lexicaux	+LEX	0.875	0.745	**0.804**	0.770	0.647	**0.703**
	+PLEX, FLEX, BLEX	0.870	0.762	0.812	0.761	0.663	0.708
Traits morphologiques	+LEM, POS, VOC	0.897	0.818	**0.856**	0.868	0.805	**0.835**
	+PREF, SUFF, ROOT	0.903	0.833	0.866	0.869	0.806	0.836
	+PREF_POS, SUFF_POS, ROOT_POS	0.919	0.853	**0.885**	0.877	0.816	**0.845**
	+GLOSS	0.877	0.806	0.840	0.866	0.801	0.832

TABLE 3 – Les résultats détaillés de la classification (C1)

		TES			ATB		
		Précision	Rappel	F-score	Précision	Rappel	F-score
C1	Milieu	0.956	0.961	0.958	0.938	0.966	0.952
	Début	0.971	0.862	0.913	0.967	0.831	0.894
	Fin	0.829	0.738	0.781	0.727	0.650	0.686
C2	Milieu	-	-	-	0.938	**0.969**	**0.953**
	Début	-	-	-	0.967	0.831	0.894
	Fin	-	-	-	**0.744**	0.650	**0.694**

TABLE 4 – Les résultats finaux des classifications (C1) et (C2)

Nous remarquons que l'utilisation de traits typographiques uniquement (Baseline B1), ne donne pas de bons résultats, surtout dans le corpus ATB. En effet, les textes de ce corpus se caractérisent par la présence non régulière de signes de ponctuation, contrairement à ceux de TES. La combinaison des traits lexicaux et des traits typographiques (Baseline B2) améliore beaucoup les résultats dans les deux corpus (tableau 3), ce qui prouve l'importance des informations lexicales. Pour ces deux types de traits, nous remarquons que la prise en considération des contextes gauche et droit du mot améliore les résultats surtout dans les cas des indicateurs faibles. Les résultats on été amélioré (plus de 30% pour le corpus TES et

© ATALA

plus de 40% pour le corpus ATB). En effet, si un indicateur faible est accompagné d'un signe de ponctuation, il sera un bon marqueur de segmentation. L'exemple (17) montre que si l'indicateur faible بعد أن/"bada"/après-que est précédé par une virgule, il constitue un point de coupure.

(17) [أكل الولد تفاحة،][بعد أن قام بغسلها]

[Le garçon a mangé la pomme][après l'avoir lavé]

En ajoutant les traits morphologiques générés suite à une analyse extensive par SAMA (10 traits), nous avons pu couvrir la majorité des cas de segmentation : conditionnel, corrélation, coordination, subordination, opposition, etc. Nous notons, aussi, l'effet positif de l'ajout de traits contextuels surtout au niveau morphologique. Donc, l'ajout de ces traits donne les meilleurs résultats. Cependant, le trait sémantique (Gloss) n'a pas d'impact sur la segmentation discursive de textes arabes, il a dégradé la moyenne de F-mesure de 3.5% pour le corpus TES et de 1.3% pour le corpus ATB, car Gloss est la traduction du mot en anglais sans prendre en considération son contexte.

Toutefois, l'ajout de traits syntaxiques n'a pas d'influence sur la détermination des frontières des segments (tableau 4). Les résultats obtenus montrent que l'utilisation d'une analyse lexicale et morphologique extensive (analyse à bas niveau) aboutie à une bonne segmentation discursive sans avoir recours à une analyse syntaxique.

Enfin, Nous avons effectué une correction qui consiste à corriger les frontières des UDM de droite à gauche. Le tableau 5 présente les résultats de la reconnaissance des UDM avant et après la correction.

		Exactitude	
		EST	ATB
Sans correction	UDM	0.408	0.372
	UDM emboîtées	0.307	0.285
Avec Correction	UDM	**0.795**	**0.764**
	UDM emboîtées	**0.615**	**0.571**

TABLE 5 – Les résultats des UDM avec et sans correction

La correction a été réalisée sur les UDM obtenus par le classifieur C1, sans le trait Gloss, sur le corpus EST et le corpus ATB. Cette correction corrige juste les fermetures des UDM qui ont les résultats les moins bonnes. Le tableau 5 montre que la correction améliore la reconnaissance des UDM de 38.7% pour le corpus EST et de 39.2% pour le corpus ATB.

Il reste cependant quelques cas qui ne sont pas bien pris en compte par notre approche. Nous citons essentiellement les erreurs dues à l'analyseur morphologique (SAMA) et à l'ambigüité lexicale, comme le montre l'exemple (18), où le nom propre « أكرم » a été analysée par SAMA comme étant un verbe alors que ce mot est un nom propre. Le couplage de SAMA avec un outil d'extraction d'entités nommées pourrait aider à réduire ces erreurs.

(18) [حصل خالد وأكرم على جائزة.]

[Khalid et Akram ont obtenu un prix.]

© ATALA

7 Conclusion

Dans cet article, nous avons proposé une méthode d'apprentissage pour la segmentation de textes arabes en unités de discours minimales. Cette méthode prédit également les UDM imbriqués. À notre connaissance il s'agit du premier travail qui s'adresse directement à la segmentation du discours en langue arabe. En effet, le seul travail existant tend à produire un discours Treebank arabe (Al-Saif et Markert, 2010) qui étend le discours Penn Treebank (PDTB) pour l'arabe standard moderne (MSA). Dans ce corpus, les éléments annotés sont les connecteurs de discours et leurs relations signalées et non pas la structure discursive complète du texte. Nous avons proposé une approche multi-classe d'apprentissage supervisé qui prédit les frontières des UDM et non seulement les connecteurs de discours. Notre approche utilise un lexique riche (avec 174 connecteurs) et s'appuie sur une combinaison de caractéristiques typologiques, lexicales et morphologiques. Cette approche a les avantages suivants : 1) détecter les frontières des UDM même en cas d'absence de marqueurs du discours (c'est-à-dire, dans le cas des relations implicites, ce qui représentent 15% des cas dans nos corpus). 2) La prise en compte d'UDM emboitée pendant la phase de segmentation.

La segmentation du discours est la première étape vers l'analyse du discours. Une annotation des documents TES et ATB avec des relations de discours dans le cadre de la SDRT est actuellement en cours.

Références

AFANTENOS, S. D., DENIS, P., MULLER, P. et DANLOS, L. (2010). Learning recursive segments for discourse parsing. *In Proceedings of the International Conference on Language Resources and Evaluation*, (LREC 2010), Valletta, Malta

AFANTENOS, S., ASHER, N., BENAMARA, F., BRAS, M., FABRE, C., HO-DAC, M., DRAOULEC, A. L., MULLER, P., PERY-WOODLEY, M.-P., PREVOT, L., REBEYROLLES, J., TANGUY, L., VERGEZ-COURET, M. et VIEU, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*

AL-SAIF, A. et MARKERT, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic, *In Proceedings of the International Conference on Language Resources and Evaluation,* (LREC 2010), Valletta, Malta

AL-SAIF, A. et MARKERT, K. (2011). Modelling Discourse Relations for Arabic. *The proceedings of Empirical Methods in Natural Language Processing*, (EMNLP 2011), Edinburgh.

ASHER, N. et LASCARIDES, A. 2003. Logics of Conversation. Cambridge University Press.

BELGUITH, H. L., BACCOUR, L. et MOURAD, G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. *12th Conference on Natural Language Processing (TALN'2005)*, Dourdan.

BERGER, S., PIETRA D. et DELLA V. (1996). A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 39–71.

CARLSON, L., MARCU, D., et OKUROWSKI, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *In Jan van Kuppevelt and Ronnie Smith, editors,*

 © ATALA

Current Directions in Discourse and Dialogue. Kluwer, Dordrecht.

DA CUNHA, I., SANJUAN, E. et TORRES M. (2010). Discourse segmentation for Spanish based on shallow parsing. In *Proc. of the 9th Mexican international conference on Advances in artificial intelligence, (MICAI 2010),* 13-23. Springer-Verlag.

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

FISHER, S. et ROARK, B. (2007). The utility of parse-derived features for automatic discourse segmentation. *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, 488–495, Prague, Czech Republic.

HABASH, N. (2010). Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies, Graeme Hirst*, editor. Morgan & Claypool Publishers.

JIRAWAN, C., THANA, S., et ASANEE K. (2005). Element Discourse Unit Segmentation for Thai Discourse Cues and Syntactic Information. *The 9th National Computer Science and Engineering Conference*, 27-28 October.

KESKES, I., BENAMARA, F. et BELGUITH, H. L. (2012). Clause-based Discourse Segmentation of Arabic Texts, *The eighth international conference on Language Resources and Evaluation (LREC 2012),* Istanbul, 21-27 may 2012.

KHALIFA, I., FEKI, Z. et FARAWILA, A. (2011). Arabic Discourse Segmentation Based on Rhetorical Methods. International Journal of Electric and Computer Sciences IJECS-IJENS, Vol: 11(1).

LE THANH, H., ABEYSINGHE, G. et HUYCK, C. (2004). Generating discourse structures for written text. *In Proc. of the 20th International Conference on Computational Linguistics (COLING)*, pages 329–335, Geneva/Switzerland.

LEE, A., PRASAD, R., JOSHI, A., et WEBBER, B. (2008). Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank. *Proc. Constraints in Discourse III Workshop.*

LÜNGEN, H., LOBIN, H., BÄRENFÄNGER, M., HILBERT, M. et PUSKAS, C. (2006). Text parsing of a complex genre. In Bob Martens and Milena Dobreva, editors, Proc*. of the Conference on Electronic Publishing (ELPUB 2006)*, Bansko, Bulgaria.

MAAMOURI, M., BIES, A., KULICK, S. KROUMA, S., GADDECHE et ZAGHOUANI, W. (2010b). Arabic Treebank (ATB): Part 3 Version 3.2. Linguistic Data Consortium, Catalog No.: LDC2010T08.

MAAMOURI, M., GRAFF, D., BOUZIRI, B., KROUNA, S., BIES, A. et KULICK, S. (2010a). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium, Catalog No.: LDC2010L01.

MANN, W.C. et THOMPSON, S. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3). 243-281.

PRASAD, A., MILTSAKAKI, R., DINESH, E., LEE, N., JOSHI, A. et WEBBER, (2008). The Penn discourse treebank 2.0. *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

SORICUT, R. et MARCU, D. (2003). Sentence level discourse parsing using syntactic and lexical

 © ATALA

information. In *HLT/NAACL*, Edmonton, Canada.

SPORLEDER, C. et LAPATA, M. (2005). Discourse chunking and its application to sentence compression. *In Proc. of the HLT/EMNLP Conference*, Vancouver, 257–264.

SUBBA, R. et DI EUGENIO, B. (2007). Automatic discourse segmentation using neural networks. *In Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, Trento, Italy.

SUMITA, K., ONO, K., CHINO, T., UKITA, T. et AMANO, S. (1992). A discourse structure analyzer for Japanese text. *In Proceedings of the international conference on fifth generation computer systems,* Tokyo, Japan ,1133–1140.

TOFILOSKI, M., BROOKE, J. et TABOADA, M. (2009). A syntactic and lexical-based discourse segmenter. *In Proceedings of the ACL-IJCNLP 2009 Conference*, 77–80, Suntec, Singapore.

TOUIR, A., MATHKOUR, H. et AL-SANEA, W. (2008). Semantic-Based Segmentation of Arabic Texts. Information Technology Journal. Vol: 7(7).

WOLF, F. et GIBSON, E. (2006). Coherence in Natural Language: Data Stuctures and Applications. MIT Press.

© ATALA