Splitting Arabic Texts into Elementary Discourse Units

ISKANDAR KESKES, Sfax University, Tunisia and IRIT-Toulouse University, France
FARAH BENAMARA ZITOUNE, IRIT-Toulouse University, France
LAMIA HADRICH BELGUITH, Sfax University, Tunisia

In this article, we propose the first work that investigates the feasibility of Arabic discourse segmentation into elementary discourse units within the segmented discourse representation theory framework. We first describe our annotation scheme that defines a set of principles to guide the segmentation process. Two corpora have been annotated according to this scheme: elementary school textbooks and newspaper documents extracted from the syntactically annotated Arabic Treebank. Then, we propose a multiclass supervised learning approach that predicts nested units. Our approach uses a combination of punctuation, morphological, lexical, and shallow syntactic features. We investigate how each feature contributes to the learning process. We show that an extensive morphological analysis is crucial to achieve good results in both corpora. In addition, we show that adding chunks does not boost the performance of our system.

Categories and Subject Descriptors: I.2.7 [**Natural Languages Processing**]: *Discourse; Text Analysis*

General Terms: Languages, Experimentation

Additional Key Words and Phrases: Discourse segmentation, elementary discourse units, Arabic language

ACM Reference Format:
Keskes, I., Zitoune, F. B., and Belguith, L. H. 2014. Splitting Arabic texts into elementary discourse units. *ACM Trans. Asian Lang. Inform. Process.* 13, 2, Article 9 (June 2014), 23 pages.
DOI: http://dx.doi.org/10.1145/2601401

1. INTRODUCTION

Discourse segmentation aims at splitting texts into Elementary Discourse Units (EDUs) which are nonoverlapping units that serve to build the discourse structure of a document. Indeed, EDUs are the entities that have to be linked by coherent relations or the entities that have to be grouped together if a set of EDUs is, as a whole, an argument of a coherent relation. Identifying EDU boundaries is thus an important first step in discourse parsing, since a wrong segmentation degrades the performances of discourse parsers. For instance, Soricut and Marcu [2003] have pointed out that perfect segmentation reduces the number of parser errors by 29%. Several works on automatic discourse segmentation have been undertaken by using rule-based [Le Thanh et al. 2004; Tofiloski et al. 2009] or learning techniques [Fisher and Roark 2007; Sporleder and Lapata 2005]. Most studies have focused on English. We note, however, some efforts for other languages such as French [Afantenos et al. 2010], Thai [Charoensuk et al. 2005], German [Lüngen et al. 2006], Spanish [Da Cunha et al. 2010], and Brazilian Portuguese [Pardo et al. 2004]. As far as we know, no work has investigated EDU segmentation in Modern Standard Arabic (MSA). This article is an attempt to do so using the Segmented Discourse Representation Theory (SDRT) as our formal framework [Asher and Lascarides 2003].

Authors' addresses: I. Keskes (corresponding author), Research Group, MIRACL-Sfax University, Sfax, Tunisia and IRIT-Toulouse University, France; email: keskes@irit.fr; F. B. Zitoune, IRIT-Toulouse University, France; L. H. Belguith, ANLP Research Group, MIRACL-Sfax University, Sfax, Tunisia.

DOI: http://dx.doi.org/10.1145/2601401

Due to the morphological and syntactic properties of MSA, discourse segmentation poses a different set of challenges. In particular, what are the segmentation principles that guide the segmentation process of Arabic texts? How can discourse segmentation deal with Arabic complex morphology where words, especially discourse connectives, are highly ambiguous? What kind of morphological analysis is suitable, that is, shallow versus extensive? Are morphological features sufficient to achieve good results? What is the added value of shallow syntactic features? To answer these questions, we propose a two-step procedure. (1) The first step is the elaboration of an annotation scheme that defines a set of principles to guide the segmentation process. Two corpora that have different genre, audience, and style of writing have been annotated according to this scheme: elementary school textbooks and newspaper documents extracted from the syntactically annotated Arabic Treebank (ATB part3 v3.2) [Maamouri et al. 2010b]. (2) The second step is the elaboration of a feature set to automatically identify EDUs using a multiclass supervised learning approach that predicts nested EDUs. We use state-of-the-art features whose efficiency has been empirically determined such as punctuation, morphological, lexical, and syntactic features [Afantenos et al. 2012; Fisher and Roark 2007; Soricut and Marcu 2003; Sporleder and Lapata 2005]. Their use in Arabic discourse segmentation is, nonetheless, novel. We investigate how each feature contributes to the learning process. In particular, we analyse the effect of shallow and extensive morphological features as well as the effect of chunks. We report our experiments on boundary detection, that is, the ability of the system to classify each token into the correct class, as well as on EDU recognition, namely, the ability of the system to identify EDU boundaries. We show that an extensive morphological analysis is crucial to achieve good results for both corpora. In addition, we show that adding chunks does not boost the performance of our system.

This article is organized as follows. The next section provides a definition of EDUs and highlights the challenges we need to overcome given the specificities of the Arabic language. Section 3 details the characteristics of our data, describes the segmentation guidelines, and presents the inter-annotators agreement study conducted on the two corpora. Section 4 presents our features. Our experiments and results are reported in Section 5. We compare our results to related work in Section 6. We finally conclude by summarizing the main contributions of this work.

2. DISCOURSE SEGMENTATION

2.1. What are EDUs?

Defining segment boundaries is generally theory dependent since each theory defines its own specificities in terms of segmentation guidelines and unit size. Main discourse theories are: the Rhetorical Structure Theory (RST) [Mann and Thompson 1988] in which the discourse structure of a document is a tree where leaves (called nucleus and satellite) are contiguous EDUs and edges are rhetorical relations, the Discourse Lexicalized Tree-Adjoining Grammar (DLTAG) [Webber 2004] where the discourse structure is created by a composition of EDUs anchored by discourse connectives, and the Segmented Discourse Representation Theory (SDRT) [Asher and Lascarides 2003] that attempts to make explicit the interactions between the semantic content of the segments and the global, pragmatic structure of the discourse. The discourse structure is a graph and not a tree as in RST. Other important discourse theories include the Discourse Representation Theory (DRT) [Kamp 1981] where the discourse structure is represented by the Discourse Representation Structure (DRS) formed by discourse referents and DRS conditions, the Intentional Discourse Model [Grosz and Sidner 1986], the Linguistic Discourse Model [Polanyi and Scha 1984; Polanyi 1985; Scha and Polanyi 1988], and the Discourse Graph Bank model [Wolf and Gibson 2006].

In this article, we follow the segmentation principles as defined within SDRT. An EDU is mainly a sentence or clauses in a complex sentence that typically correspond to verbal clauses, as in [*I loved this movie*]$_a$ [*because the actors were great*]$_b$ where the relative clause introduced by the marker *because*, indicates a cutting point. We have here the relation *Explanation(a,b)*[1]. An EDU can also correspond to other syntactic units describing eventualities, such as prepositional and noun phrases, as in [*After several minutes,*]$_a$ [*we found the keys on the table*]$_b$ where we have two EDUs related by *Frame(a,b)*[2]. In addition, an EDU may be structurally embedded in another so as to encode adjuncts like appositions or cleft constructions with discursive long-range effects such as frame adverbials, nonrestrictive relatives, and appositions, as in [*Mr. Dupont,* [*a rich business man,*]$_a$ *was savagely killed*]$_b$where we have the relation *Entity-Elaboration(a,b)*[3].

Our segmentation is more fine-grained than in Wolf and Gibson's framework [2006] or the annotation scheme of RST [Carlson et al. 2003]. Indeed, in RST, EDUs can be simple sentences or clauses in a complex sentence that typically correspond to verbal clauses or to prepositional and noun phrases. In addition, embedding in RST is done artificially since it is handled by the relation labeler (with an ad hoc "same-unit" relation) and not during the segmentation stage.

2.2. EDU Segmentation in Arabic: Main Challenges

In this section we give a brief overview of MSA specificities. For a detailed description of MSA and Arabic Natural Language Processing (ANLP), see Habash [2010]. In the remainder of this article, all examples are extracted from our corpora. They are given in Arabic along with their English translation and their transliteration using Buckwalter 1.1.

Arabic does not have capital letters nor punctuation marks are widely used in current Arabic texts (at least not regularly). Moreover, Arabic discourse tends to use long and complex sentences. We can easily find an entire paragraph without any punctuation. As a Semitic language, Arabic has a complex morphology. Indeed, in addition to a *concatenative morphology*, where words are formed via a sequential concatenation process, Arabic is characterized by the presence of a *templatic morphology* where a templatic morpheme is composed of a root (a sequence of (mostly) three, (less so) four, or very rarely five consonants), patterns (an abstract template in which roots and vocalisms are inserted), and vocalisms that specify the short vowels to use with a pattern. For example the word stem كَتَبَ/katab/to-write is constructed from the root ك - ت - ب/*k-t-b*, the pattern *1V2V3* and the vocalism *aa* [Habash 2010]. Concatenative morphemes can be stems, affixes, or clitics. A clitic has the syntactic characteristics of a word but depends phonologically on another word or phrase. Clitics include prepositions, conjunctions, and pronouns. For instance, prepositions (like ل/li/for), conjunctions (like و/w/and), articles (like ال/Al/the) and pronouns (like ه/h/he), can be affixed to nouns, adjectives, particles, and verbs, causing several lexical ambiguities. Here are some examples.

— The word فهم/fhm can be a noun (that means understanding), a verb (that means to understand), or a conjunction (ف/f/then) followed by the pronoun (هم/hm/they).

[1]Explanation(a, b) holds when the main eventuality of b is understood as the cause of the eventuality in a.

[2]*Frame (a,b)* holds when the segment b is on the scope of a frame a. The segment a is generally at the beginning of a sentence and can be a temporal or a spatial adverbial or an adverbial that has a large scope as in *[After 6 years ago,]$_a$ [she got married with John]$_b$ [and she has got two children]$_c$* where the segments b and c are within the scope of the segment a.

[3] *Entity-Elaboration(a, b)* holds when *b* gives more details about an entity introduced in *a*.

— The word وليد /wlyd can be a person name (Waleed), a noun (that can mean a newborn), or the composition و/w/and + لـ/li/for + the noun يد /yd/hand.
— The word فضل /fDl can be a person name (Fadhl) as in (1), a verb (that means to prefer), or the conjunction ف/f/then followed by the verb ضل/Dl/lost.

(1) استقبلت عائلة مصطفى فضل البارحة.

Astqblt EA}lp mSTfY fDl AlbArHp.
Yesterday, I received Mustapha Fadhl's family.

Note, however, that not all complex word structures are ambiguous. For instance, the word استتذكرونها/Astt*krwnhA ([اس/As/will], [تتذكرون/tt*krwn/you remember], and [ها /hA/her]) represents "will you remember her?" in the English language.

Another specificity of Arabic is that word order is fairly flexible. Indeed, the change of certain position of words does not alter the meaning of the sentence. For example, the sentence "the child goes to the school" can be written in Arabic in three forms: ذهب الولد إلى المدرسة / *hb Alwld <lY Almdrsp, الولد ذهب إلى المدرسة / Alwld *hb <lY Almdrsp and إلى المدرسة ذهب الولد / <lY Almdrsp *hb Alwld.

Finally, the most important challenge in ANLP is diacritics. Arabic has 28 consonants that may be interleaved with different long and short vowels. Short vowels are not often explicitly marked in writing. Indeed, they are neither written in the Arabic handwriting of everyday use nor in general publications. Diacritics represent, among other things, short vowels. Arabic texts can be fully diacritized, partially diacritized, or nondiacritized. It should be noted that nondiacritized texts are highly ambiguous. For example, the word ثمن /vmn can be diacritized in 22 different forms. The same confusion holds between the verb ذَهَبَ /*ahaba/go and the noun ذَهَبٌ /*NhabN/gold. Thus, a nondiacritized word could have different morphological features, and in some cases, different POS, especially when it is taken out of its context. In addition, even if the context is considered, the POS and the morphological features could remain ambiguous, as shown in (2).

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

wSf AlTbyb llmryD mjmwEp mn Al>dwyp lmEAljp >lmh wjrHh
The doctor recommended to his patient a set of drugs to treat his pain and injury.

In (2), if the word جرحه/jrHh is recognized as a verb (to injure), we will have a segmentation error since this word is a noun in the context of (2). The cutting point here should be the word لمعالجة /lmEAljp, because the discourse marker لـ/l/for is a good indicator of the relation *Goal*.

3. BUILDING AN ARABIC CORPUS FOR DISCOURSE SEGMENTATION

3.1. Data

Our data come from two different corpora: elementary school textbooks and newspaper documents extracted from the syntactically annotated Arabic Treebank (ATB v3.2 part3).

EST documents are usually well structured. Sentences are short (around 5.6 words per sentence) with quite a simple syntactic structure. They are characterized by the presence of punctuation marks. Document length is also short (around 10 sentences per document). Our colleagues of Sfax University have collected 34 EST documents. They have first randomly selected a set of texts from Tunisian elementary school

INPUT STRING: ان/ An	INPUT STRING: سوريا/swryA
IS_TRANS: An	IS_TRANS: swryA
INDEX: P7W3	INDEX: P7W4
OFFSETS: 3,6	OFFSETS: 6,12
UNVOCALIZED: <n	UNVOCALIZED: swryA
VOCALIZED: <in~a	VOCALIZED: suwriyA
POS: PSEUDO_VERB	POS: NOUN_PROP
GLOSS: that	GLOSS: Syria

Fig. 1. Morphological analysis of the first two words in example (4) (ان and سوريا) as given by ATB manual annotations. The annotation includes: the Arabic word (INPUT STRING), its transliteration (IS_TRANS), its position in the sentence (INDEX), its offsets, its corresponding unvocalized and vocalized words, its part-of-speech (POS), and its English translation (Gloss)..

textbooks (level 4th, 5th, 6th, 7th, and 8th, then they have manually input them into a text file format. This corpus actually contains a total of 622 sentences which corresponds to 8 704 tokens (words+punctuations). Contrary to ATB documents, it is important to note that EST documents are not associated with any kind of manual annotation. (3) is an example of a sentence extracted from this corpus.

(3)على شاطئ الحمّامات، انتصب حصن قديم، يدخله الزّائر من بوّابة مقوّسة، تفضي به إلى أروقة مسقّفة، كأسواق المدينة العتيقة.

ElY $AT}AlHm˜AmAt, AntSb HSn qdym, ydxlh Alz˜A}r mn bw˜Abp mqw˜sp, tfDy bh <lY >rwqp msq˜fp, k>swAq Almdynp AlEtyqp.
In the Hammamet beach, an old fort is erected, in which a visitor can enter from an arched gate, leading him to wrapped corridors that resemble ancient city markets.

ATB documents consist of 599 newswire stories from the An Nahar News Agency. Each document in this corpus is associated to two annotation levels: a morphological and parts-of-speech level and a syntactic Treebank annotation. The second level characterizes the constituent structures of word sequences, provides categories for each nonterminal node, and identifies null elements, co-reference, traces, etc. Contrary to EST, ATB documents are longer (around 25 sentences per document) and sentences are syntactically more complex. We have randomly selected 56 documents from ATB for a total of 1 427 sentences and 31 682 tokens (words+punctuations). The example in (4) presents a short sentence extracted from an ATB document along with the morphological analysis of its first two words (Figure 1) and its associated syntactic tree (Figure 2).

(4) ان سوريا أصبحت ابتداء من مطلع السنة الجارية عضوا غير دائم في مجلس الأمن لمدة سنتين.

An swryA >SbHt AbtdA' mn mTlE Alsnp AljAryp EDwA gyr dA}m fy mjls Al>mn lmdp sntyn.
Since the beginning of the year, Syria became a non-permanent member of the Security Council for a period of two years.

3.2. Annotation Scheme

The annotation scheme defines a set of segmentation principles to guide the segmentation process. Our scheme is inspired from an already existing manual elaborated

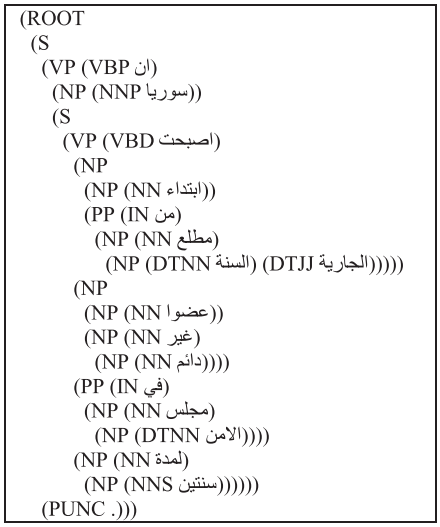

Fig. 2. Syntactic analysis of the example (4) as given by ATB manual annotations.

within the Annodis[4] project that focused on the selective annotation of multilevel discourse structures of French documents following SDRT [Afantenos et al. 2012]. Annodis manual provided annotators with an intuitive introduction to discourse segments, including the fact that discourse segments can be embedded in one another. Detailed instructions were provided describing how to handle segmentation for most of the cases that could naturally arise.

We have adapted this manual to take into account Arabic specificities. First, we have identified similar cases of segmentation, such as simple phrases, conditionals, correlative clauses, and subordinate phrases. Then, we have added Arabic-specific principles to handle cases such as al-masdar (also called the infinitive or the (de)verbal noun) constructions, مبتدأ /mbtd>and خبر /xbr clauses (also referred to as a copular construction or equational sentence), coordinations, and adverbial clauses. In our manual, each segmentation principle is presented along with examples that illustrate main cases of segmentation as well as cases that do not need segmentation. In this section we give basic segmentation cases as well as main segmentation principles.

3.2.1. Basic Principles. EDUs are delimited by square brackets. Discourse cues are always at the beginning of a segment, whereas punctuation marks that delimit segment frontiers always appear before the end of a segment. EDUs cannot overlap but they can be embedded in one another (double square brackets are not allowed), as in (5).

(5) [ناقش الأستاذ الامتحان،[الذي أجراه التلاميذ الأسبوع الماضي،] و الدرس الحالي .]

[nAq$ Al>stA AlAmtHAn, [Al*y >jrAh AltlAmy* Al>sbwE AlmADy,] w Aldrs AlHAly .]*
[The teacher explained the exam, [that was passed by the students last week,] and the current lesson.]

[4]w3.erss.univ-tlse2.fr/annodis

An EDU is basically a verbal (refer to (5)) or a nominal clause (مبتدأ /mbtd> and خبر/xbr) (refer to (6)). A cutting point can neither separate a verb from its complement nor a subject from its verb. In addition, segment frontiers can never occur within a chunk or a named entity.

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

[qSft TA}rAt >myrkyp mjmEAt mn Alkhwf.]
[American aircrafts bombed a set of caves.]

(7) [كانت الطفلة جميلة.]

[kAnt AlTflp jmylp.]
[The girl was beautiful.]

3.2.2. Main Segmentation Principles

— *Al-masdar (المصدر /AlmSdr)*. They are segmented only in the indefinite accusative case (منصوب /mnSwb) because this construction generally signals discourse relations. For example, in (8), al-masdar بحثا /bHvA/looking for, explains why Ahmed went to the library.

(8) [اتجه أحمد إلى المكتبة][بحثا عن كتاب الرياضيات.]

[Atjh >Hmd <lY Almktbp][bHvA En ktAb AlryADyAt.]
[Ahmed went to the library][looking for the mathematic book]

We do not segment in other cases (like البحث / *AlbHv* / *search)*, as in (9).

(9) [استمر في البحث عنه في كل المكتبات.]

[Astmr fy AlbHv Enh fy kl AlmktbAt.]
[He is still searching for him in all libraries.]

— *Conditionals (شرط / $rT)*. They are always segmented, as in (10).

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

[>A >SbH AlTqs jmyl,][s>xrj >tnzh.]*
[If the weather is nice,] [I'll go for a stroll,]

— *Correlatives (تلازم/ tlAzm)*. They are always segmented, as in (11).

(11) [كلما أطالع الكتب،][كلما تتحسن ثقافتي العامة.]

[klmA >TAlE Alktb,][klmA ttHsn vqAfty AlEAmp.]
[The more I read books,] [the more I learn.]

— *Coordinations (ربط/ rbT)*. In Arabic, a coordination is introduced by markers such as و/w/and, ف/f/and, ثم/vm/then, أو/>w/or. . . which are highly ambiguous. For instance, the conjunction و/w/and can have six different senses [Khalifa et al. 2011]: (a) والقسم/w Alqsm that means testimony, (b) ورب/w rb that means few or someone, (c) والاستئناف/wAlAst}nAf that simply joins two unrelated sentences, (d) والحال /w AlHAl that introduces a state (refer to (12)), (e) والمعية/w AlmEyp that means the

accompaniment, and (f) والعطف/w AlETf meaning the conjunction of related words or sentences (refer to (13)).

(12) دخل الولد الفصل وهو يبتسم.

dxl Alwld AlfSl whw ybtsm.
The child enters the classroom smiling

(13) انتهت العطلة وبدأت الدراسة.

Antht AlETlp wbd>t AldrAsp.
The holidays ended and the study began.

Our treatment of coordination goes beyond discourse segmentation proposed in Khalifa et al. [2011] since we do not only deal with the marker و /w/and but also with other markers. We segment coordination in four cases: (i) coordination of independent clauses, (ii) coordination of subordinating clauses, (iii) when two verbal phrases share the same object or the same subject, as in (14), and finally (iv) coordination of prepositional phrases that introduce events, as in (15). We do not segment in all the other cases, such as the conjunction between two objects of the same verb.

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

[AstEAd Alr} ys Altwnsy EAfyth][wqAm bAstqbAl AlmwATnyn].
[The Tunisian President regained its health] [and has begun to receive the citizens.]

(15) [أعلنت الحكومة عدم موافقتها على محضر الجلسة] [لعدم توفر الشروط الأزمة]

[>Elnt AlHkwmp Edm mwAfqthA ElY mHDr Aljlsp] [lEdm twfr Al$rwT Al>zmp]
[The government announced his refusal to open the session] [because of a lack of good conditions]

— *Subordinations (* صلة */Slp)*. They are always segmented. Relative clauses are introduced by the relative pronouns الذي/Al*y/ and التي /Alty/ that correspond in English to the following pronouns: which, who, whom, and that (cf. (16)). Some conjunction of subordinations (like أن />n/that, أنّ />n˜/that, إن /<in/if, إذا /<i*A/if-whether, حتى /HtY/so that and طالما /TalamA/as long as) are generally used after a verb of communication or a reported speech verb (cf. (17)). Other markers introduce temporal and causal subordinations such as قبل أن /qbl >n/before-that, لأنّ /l>n˜/because, حين /Hyn/when and غير أن /gyr >n/ nevertheless.

(16) [و في كتاب التكليف [الذي وجهه الى الحكومة الجديدة ،] تمت اتخاذ كل الترتيبات والاستعداد الكامل.]

*[w fy ktAb Altklyf [Al*y wjhh AlY AlHkwmp Aljdydp ,] tmt AtxA* kl AltrtybAt wAlAstEdAd AlkAml .]*
[In the book of reference [which has been sent to the new government,] all the arrangements have been taken.]

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

[wqAl wzyr AldfAE] [An nHw stp ms&wlyn Amyrkyyn wSlwA AlY AlblAd.]
[The Minister of Defense said] [that six U.S. officials had arrived in the country.]

— *Appositions (بدل /bdl)*. They are segmented in most cases. Appositions can be:
  — adjectival phrases;
  — adverbial phrases, introduced either by relative adverbs (such as متى/mtY/when, كيف/kyf/how, لماذا /lmA*A/why, حيث/Hyv/where/) or by regular adverbs (such as حينذاك/Hyn*Ak/at that time, وقتذاك/wqt*Ak/by then and ربما /rbmA/perhaps) as in (18); or
  — nominal or verbal phrases introduced by pseudo-verbs like إن/<n/that, ليت/lyt/hope-that, لعل/lEl/may-be, or by noninflectional verbs like حيا/HyA/come-to, سرعان/srEAn/soon.

Prepositional phrases (introduced by الى/<lY/until, عن/En/about, فى/fY/in, من/mn/from and على/ElY/on) that appear at the end of a clause are not segmented.

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

[An Aljnwd, [Hyv sykwnwn mslHyn,] ystTyEwn AldfAE En Anfshm.]
[The soldiers, [once they are armed,] they will be able to defend themselves.]

— *Adverbials (ظرفية /Zrfyp)*. In some cases, an adverbial can be an EDU. This concerns adverbials that introduce an event or a state, as in (19) where we have a *Goal* relation, and adverbials that are at the beginning of the sentence, as in (20) where we have a *Frame* relation. (21) gives an example of a temporal adverbial introduced by البارحة على الساعة الرابعة والنصف مساء/AlbArHp ElY AlsAEp AlrAbEp wAlnSf msA'/yesterday at four-thirty in the afternoon, that does not indicate a cutting point.

(19) [رجعت مسرعا إلى البيت][حيث كان المطر يتهاطل.]

[rjEt msrEA <lY Albyt][Hyv kAn AlmTr ythATl.]
[I returned quickly at home][where was raining.]

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

[EndmA twfy jdy,][knt SgyrA jdA.]
[When my grand-father died,][I was very young]

(21) [اجتمع المجلس البارحة على الساعة الرابعة والنصف مساء][لمناقشة هذا القانون]

*[AjtmE Almjls AlbArHp ElY AlsAEp AlrAbEp wAlnSf msA'][lmnAq$p h*A AlqAnwn]*
[The council met yesterday at four thirty in the afternoon][in order to discuss this law]

— *Other cases.* We segment reported speech sentences between quotes (this case indicates the *Attribution* relation). We also segment modifiers that begin with possessive pronouns that detail a previously introduced entity (cf. (22)) since this case indicates the *Entity-Elaboration* relation. We do not segment in case of transliteration, Latin characters, and abbreviations, as well as in case of demonstrative pronouns (هذا/h*A/this, هذه /h*h/this and هذان /h*An/these).

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

[wqd~mt lnA SHnA SgyrA][fyh mqrwDAt $hy~p.]
[She gave us a small dish] [containing tasty Makrouts.]

3.3. Inter-Annotators Agreement Study

Two Arabic native speakers (undergraduate students in Arabic linguistics) were asked to doubly annotate a set of documents from our corpora following the guidelines given in the annotation scheme. First, annotators were trained on four EST documents

Table I. Characteristics of Our Data in the Gold Standard

	Texts	Size	Sentences	EDUs	Embedded EDUs	Word+PUNC
EST	25	67ko	442	924	86 (10.74%)	6 437
ATB	50	267ko	1 272	2 788	372 (7.49%)	28 288
Total	**75**	**334ko**	**1 714**	**3 712**	**458 (8.10%)**	**34 725**

(75 sentences) and four ATB documents (110 sentences). The training phase for ATB lasted longer compared to EST since ATB documents contain more complexity. This phase allowed for revising the annotation guidelines. Then, each annotator was asked to separately annotate five EST documents and two ATB documents which corresponded, respectively, to 71 and 63 sentences (documents used for training were discarded).

Agreements were computed by counting how often each annotator classified each token as being an EDU boundary. We got an average Cohen's kappa of 0.83 for ATB and 0.89 for EST. We observed four cases of disagreement: (1) lexical ambiguities, especially for discourse markers that appear as clitics (cf. Section 2); (2) long sentences with more than five words (cf. (4) in Section 3.1); (3) the absence of punctuation marks, especially when clauses within a sentence are not separated with punctuations (cf. (14) and (15) in Section 3.2.2); and (4) al-masdar constructions (cf. (23) given next). Cases (2) and (3) were more frequent in ATB documents.

(23) [تشكر أحمد جارته][وفاء لعملها.]

[t$kr >Hmd jArth][wfA' lEmlhA.]
[Ahmed thanks her neighbor][by loyalty to her work.]

In (23), one annotator considered that the word وفاء/wfA' is a cutting point because this word is al-masdar in an indefinite accusative case of the verb وفى/wfY. Hence, the second EDU explains why Ahmed thanks his neighbor. The second annotator, on the other hand, cut at the word لعملها/lEmlhA' because he considered the words جارته وفاء/dyArh wfA' as a named entity (the name of the neighbor). For him, the second EDU explains why Ahmed thanks his neighbor Wafa. Of course, this is an error because, in our example, the word وفاء/wfA' is al-masdar construction and not a named entity.

Given the good inter-annotator agreement results, annotators were asked to build the gold standard by consensus by discussing main cases of disagreements, as discussed earlier. Table I gives statistics about the data in the gold standard. The column WORD+PUNC indicates the number of tokens.

4. AUTOMATIC DISCOURSE SEGMENTATION

We performed a supervised learning on the gold standard by using the Stanford classifier that is based on the Maximum Entropy model [Berger et al. 1996]. Each token can belong to one of the following three classes: *Begin*, if the token begins an EDU, *End* if it ends an EDU, or *Inside*, if a token is none of the preceding[5].

To identify EDU boundaries, we used four groups of features: *punctuation, lexical, morphological*, and *syntactic* features. A feature vector is associated to each token. The features were designed after analyzing the documents used for training as well as those used to compute inter-annotator agreements (which correspond to six ATB documents (181 sentences) and nine EST documents (138 sentences)). Our set of features is given next.

[5]Theoretically, a segment can be reduced to one token. However, we do not observe such cases in our data.

4.1. Punctuation Features

Punctuation marks used today in Arabic writings are those of the European writing system, but they do not necessarily have the same semantic functions. For example, the origin of the comma is to be found in the Arabic letter و /w that represents the conjunction "and" in the English language. The full stop is often used in Arabic to mark the end of a paragraph, whereas the comma, in addition to its coordination function, can also be used to announce the end of a sentence [Belguith et al. 2005]. In Arabic, the other punctuation marks like the parentheses, the exclamation point, the question mark, and the three points have the same values as those of European languages [Belguith 2009].

During the annotation campaign, we have identified two punctuation mark categories (henceforth PMC): *strong* that always identify the end or the beginning of a segment and *weak* that do not always indicate a boundary. We have three punctuation features: (1) TOKEN_PUNC, the PMC of the token to be classified; (2) BEFORE_PUNC, the PMC of the token that precedes the current token; and (3) AFTER_PUNC, the PMC of the token that follows the current token. PMC can take three values: 0 if the token is not a punctuation mark, 1 if it is a strong indicator, and 2 if it is a weak indicator.

4.2. Lexical Features

We consider here both discourse cues such as حيث/Hyv/where, بينما/bynmA/while, عندئذ/End}*/at that time, and a set of specific words, called indicators, that are important for the segmentation process. Indicators can be reporting verbs and propositional attitude verbs (e.g., قال/qAl/say, أعلن/>Eln/announce, إعتقد/<Etqd/believe), noninflectional verbs (e.g., حيّا/Hy~A/come-to, حذار/H*Ar/beware and امين/Amyn/amen), adverbs (e.g., بعد/bEd/after, قبل/qbl/before, من المفروض/mn AlmfrwD/normally, فقط/fqT/only), conjunctions (e.g., حالما/HAlmA/the-moment-that and طالما/ TAlmA /so-often), and particles (e.g., لم /lm/not and لن /ln/never). Like punctuation marks, we have Two Lexical Cue Categories (LCC): *strong* and *weak*. Strong connectors are usually followed by a verb indicating the beginning of a segment. Some of these connectors are: كي/to, ل/l/for, لكن/lkn/but, لكنّ/lkn~a/but, غير أن/ gyr >n/nevertheless, بيد أن/byd >n/however, من أجل أن /mn >jl>n/in-order-to. On the other hand, ambiguous connectors do not always mark the beginning of a segment, such as the connector و /w/and and the particles ثم/vm/then, ف/f/so-then, etc. For example, the particle و can express a new clause, a conjunction between NPs, or it can be part of a word, as in ورشة/wr$p/atelier.

We have four lexical features: (1) TOKEN_LEX, the current token LCC; (2) BEFORE_LEX, the LCC of the token that precedes the current token; (3) AFTER_LEX, the LCC of the token that follows the current token; and (4) TOKEN_BeginLex, a Boolean feature that indicates whether the current token begins with an indicator or with a discourse cue. This last feature deals with agglutinations. LCC can take five values: 0 if the token is not a lexical cue, 1 if the token is a strong discourse cue, 2 if the token is a weak discourse cue, 3 if the token is a strong indicator, and 4 if the token is a weak indicator.

To handle both punctuation and lexical features, we have built a lexicon of segmentation indices where each entry is characterized by its type (a punctuation mark, a discourse cue, or an indicator), its nature (strong or weak), and a list of its possible Parts-Of-Speech (POS). We have also indicated if the lexical entry is composed of other words, such as خلاصة القول/Alqwl xlASp/in-summary and باختصار/bAxtSAr/briefly. If so, we have detailed each element of the composition. We have finally associated to each entry its English translation and an example of its usage in context. Our lexicon contains 174 entries: eleven punctuation marks (four strong: the exclamation mark,

the question mark, the colon, and the semi-colon, as well as six weak: the full stop, the comma, quotes, parenthesis, brackets, braces, and underscores) and 163 lexical cues (83 discourse cues and 80 indicators) among which 76.4% are strong and 23.6% are weak.

4.3. Morphological Features

Our aim is to identify what kind of morphological analysis is suitable for Arabic discourse segmentation, that is, shallow versus extensive. To this end, we propose to use two contextless parsers that provide different morphological information: Alkhalil [Boudlal et al. 2011], a shallow parser, and the Standard Arabic Morphological Analyzer SAMA version 3.1 [Maamouri et al. 2010a], an extensive analyzer. We have thus designed two sets of morphological features, one for each parser output.

Alkhalil gives each token a nonordered list of all its possible forms (by default, the first form of this list is chosen) [Boudlal et al. 2011]. More precisely, it generates the stem, its grammatical category, and its possible roots, where each root is associated to its corresponding patterns, proclitics, and enclitics. Alkhalil does not take into account the context nor punctuation marks. In addition, it does not provide affixes information and its database does not contain information about the closed nouns except their fully diacritized form and their Arabic class name, along with the allowed proclitics and enclitics. For each token, we have six Alkhalil features: (1) STEM, the token stem; (2) POS, the token parts-of-speech; (3) CATEGORY, the token grammatical category; (4) HAS_PREFIX and (5) HAS_SUFFIX that, respectively, indicate if the token has a prefix or a suffix; and (6) PATTERN, the token pattern. All the features are encoded in strings (in Arabic script).

SAMA 3.1 is a new version of the Buckwalter Arabic Morphological Analyzer (BAMA) 2.0. SAMA associates to each Arabic word token all its corresponding "prefix-stem-suffix" segmentations. In addition, it lists all known/possible annotation for each solution, with assignment of all diacritic marks, morpheme boundaries (separating clitics and inflectional morphemes from stems), all Parts-Of-Speech (POS) labels, and glosses for each morpheme segment. We have designed 10 SAMA features: (1) LEMMA, the token lemma; (2) POS, the token POS; (3) VOCALIZATION, the token vocalization; (4) PREFIX; (5) SUFFIX; and (6) ROOT that, respectively, give the prefix, the suffix, and the root of the token; (7) PREFIX_INFO; (8) SUFFIX_INFO; and (9) ROOT_INF that, respectively, give the information of the prefix, the suffix, and the root; and finally (10) GLOSS, that indicates the token gloss. All these features are generated by SAMA in transliterated form.

4.4. Syntactic Features

To evaluate the added value of syntactic features to discourse segmentation of Arabic texts (cf. Introduction), we propose to take into account chunks. To obtain them, we chose to rely on manual annotations instead of using a shallow syntactic parser such as AMIRA [Diab 2009]. Indeed, our aim is to test the upper bound for shallow syntax features. If we do not find chunks useful, we do not need to use a parser to predict them. Syntactic features concern only the ATB corpus (we recall that EST documents do not contain any manual annotations (cf. Section 3.1)).

We have only one feature that specifies whether the token to be classified is at the beginning, at the end, or in the middle of a chunk.

5. EXPERIMENTS AND RESULTS

In order to measure the impact of morphological and syntactic features on the performance of our segmenter, we designed three classifiers: (C1) that uses punctuation, lexical; and Alkhalil features; (C2) that relies on punctuation, lexical and SAMA

features; and (C3) that uses punctuation, lexical, SAMA features; and syntactic features. Configurations (C1) and (C2) were run on EST and ATB while (C3) concerns only ATB. Punctuation features are the same for all the three configurations. Lexical features are obtained by checking whether the current token lemma (as given by SAMA) or the current token stem (as given by Alkhalil) is an entry in our lexicon. Our first experiments showed that best results are achieved when using SAMA lemmatization. We have thus decided to use the token lemma as given by SAMA.

For each corpus, we have performed a tenfold cross-validation where 10% of the corpus was left for test. For all the experiments, we have used both character n-grams and word n-grams as features. Best results were achieved with n = 4. Because we have few EDU boundaries, our dataset is skewed (see Table I, Section 3.3 for an overview of our data characteristics). But we did not observe any problem related to the class imbalance in the training set with the parameters we used when building the classifier.

We recall that our aim is to automatically identify a segment. This means that our system has to achieve good performances on:

— token boundary detection, which is the ability of the system to classify each token into the correct class (*Begin*, *End*, and *Inside*);
— EDU recognition, which is the ability of the system to identify an EDU. Here, only the *Begin* and the *End* class matter. In addition, the system has to generate a balanced number of instances of each class in order to ensure a coherent bracketing. In case of an ill-formed EDU, a specific post-processing rule is applied.

We present next our results on each of these two tasks. We end this section by giving the learning curve of our experiments.

5.1. Token Boundary Detection

5.1.1. Analyzing the Impact of Punctuation, Lexical and Morphological Features. Unlike Tofiloski et al. [2009] and Soricut and Marcu [2003] that only measure the score of their segmenter on boundaries inside sentences (to avoid artificially boosting the performance), the evaluation of our system takes into account sentence boundaries since end-of-sentence or end-of-paragraph boundaries are not given automatically but are predicted by our segmenter. Table II gives (C1) and (C2) overall performances in terms of precision, recall, F-score, and accuracy, averaged over the three classes *Begin*, *End*, and *Inside*. Best performances are marked in boldface. We first start with punctuation features to which several features are progressively added; this is marked by the "+" sign in the table. We have also compared the performance of each classifier against two baselines: (B1) that only uses the current token punctuation category (TOKEN_PUNC); and (B2) that uses both the current token punctuation and lexical category (i.e., TOKEN_PUNC and TOKEN_LEX).

Our first baseline (B1), that tests if the current token is a punctuation mark (from the strong or the weak type) or not, performs badly for both corpora. Taking into account both right and left context (by adding BEFORE_PUNC and AFTER_PUNC features) improves the F-score by, respectively, 7.4% for EST and 3.7% for ATB. However, punctuation features alone are not sufficient to achieve good results for both corpora for three main reasons: the absence of regular punctuation marks, especially for ATB, the high frequency of weak punctuation marks (cf. (22)), and the presence of named entities.

(22) [كانت رافعة يدها الطّويلة، فاتحة فمها الواسع،]

[kAnt rAfEp ydhA AlT˜wylp, fAtHp fmhA AlwAsE,]
[She was raising her long arms, opening her wide mouth,]

Table II. Results of the Baselines (B1) and (B2) and the Classifiers (C1) and (C2) in Terms of Precision (P), Recall (R), F-Score (F) and Accuracy (Acc)

		EST				ATB			
		P	R	F	Acc	P	R	F	Acc
Punctuation features	**TOKEN_PUNC** (B1)	0.450	0.416	0.432	0.511	0.237	0.277	0.255	0.422
	+BEFORE_PUNC, AFTER_PUNC	0.575	0.453	0.506	0.684	0.252	0.348	0.292	0.504
PUNC + LEX (B2)		0.581	0.485	0.507	0.686	0.479	0.471	0.487	0.822
Lexical features	+TOKEN_LEX	0.568	0.492	0.513	0.689	0.397	0.415	0.406	0.807
	+BEFORE_LEX, AFTER_LEX, TOKEN_BeginLEX	0.557	0.497	0.515	0.685	0.407	0.455	0.430	0.809
(C1) : Punctuation + Lexical +Alkhalil morphological features	+STEM, POS, CATEGORY	0.581	0.485	0.528	0.694	0.492	0.501	0.496	0.784
	+ PATTERN	0.557	0.497	0.525	0.693	0.511	0.507	0.509	0.798
	+ HAS_PREFIX, HAS_SUFFIX	**0.573**	**0.504**	**0.536**	**0.701**	**0.557**	**0.503**	**0.529**	**0.811**
(C2) : Punctuation + Lexical + SAMA morphological features	+LEMMA, POS, VOCALIZATION	0.897	0.818	0.856	0.911	0.871	0.801	0.835	0.917
	+PREFIX, SUFFIX, ROOT	0.903	0.833	0.866	0.915	0.870	0.811	0.839	0.920
	+PREFIX_INFO, SUFFIX_INFO, ROOT_INFO	**0.919**	**0.853**	**0.885**	**0.919**	**0.888**	**0.810**	**0.847**	**0.923**
	+GLOSS	0.877	0.806	0.840	0.901	0.869	0.807	0.837	0.919

Using the McNema's test, the difference between (C1) and (C2) is significant at $p<0.05$ for both EST and ATB.

Compared to (B1), (B2) obtained better performances. However, the results are similar to those obtained when using (B1) + BEFORE_PUNC + AFTER_PUNC for EST, which shows that segmentation in EST, is less sensitive to the surrounding punctuations of a given token than ATB.

When adding lexical features, EST results remained stable while at the same time ATB results (in terms of accuracy) improved significantly over (B1) + BEFORE_PUNC + AFTER_PUNC by more than 30%. We think that the absence of improvement for EST can be explained by the fact that EST is characterized by regular punctuation marks, which seems to be adequate to reach an accuracy of 0.686. The good results obtained for ATB show that our lexicon is a useful resource for discourse analysis. In addition, we observe that adding contextual lexical features, namely the lexical type (strong or weak) of the left (BEFORE_LEX) and the right token (AFTER_LEX) improves ATB results. Indeed, these features were able to disambiguate cases like in (23) where the adverb بَعْد/baEud/after was identified as a verb بَعُد/baEod/to-move-away by SAMA. However, lexical features cannot deal with other types of ambiguities, like named entities (cf. error analysis at the end of this section).

(23) [أكل الولد تفاحة، بعد غسلها]

[<kl Alwld tfAHp, bEd gslhA]
[The child ate the apple, after he has washed it]

Table III. Results of the (C2) Classifier on Each Class

		EST				ATB			
		P	R	F-score	Acc	P	R	F-score	Acc
(C2)	Inside	0.956	0.961	0.958	0.988	0.938	0.966	0.952	0.922
	Begin	0.971	0.862	0.913	0.920	0.967	0.831	0.894	0.980
	End	0.829	0.738	0.781	0.933	0.735	0.658	0.695	0.944

Table IV. Confusion Matrix of the (C2) Classifier on the ATB Corpus

	Inside	Begin	End
Inside	22 236	325	314
Begin	268	2 588	0
End	**1 022**	4	1 531

Concerning morphological features, the (C2) configuration yields better results compared to (C1) mainly because the SAMA parser gives more morphological information than that given by Alkhalil. Indeed, in addition to Alkhalil's outputs (stem, POS, prefix, and suffix), SAMA provides information about the token root (ROOT_INFO), the token prefix (PRFFIX_INFO), the token suffix (SUFFIX_INFO), as well as the token gloss (GLOSS). Our experiments show that the best score is achieved when adding information of the root, the prefix, and the suffix. However, gloss information does not seem useful for discourse segmentation, since adding it has degraded the average F-score for both corpora. We get similar observations for the pattern feature (PATTERN) in the (C1) configuration since this feature has only a minor impact on the results, especially for EST.

Overall, both corpora achieved good F-scores that are comparable to human results (cf. Section 3.3). An interesting observation comes from punctuation features, in that even if they perform badly when they are used alone, removing them from the features vector has a negative impact on the results for both the two classifiers. For instance, we get an F-score of 0.840 for EST and 0.837 for ATB when running the classifier with SAMA features. Another interesting point is that morphological features alone are not sufficient. Indeed, we get an F-score of 0.713 for ATB and 0.772 for EST when running (C1) and (C2) without punctuation and lexical features. Moreover, when comparing (C1) and (C2), only the *Begin* class is biased (the F-score decreases from 0.899 to 0.540) while the results of the *End* and the *Inside* classes remain stable. Finally, the overall evaluation on EST documents gets similar results compared to those obtained for ATB documents. As expected, we can conclude that discourse segmentation does not rely only on punctuation marks and that text length has no impact on the segmentation. Our results thus demonstrate that our first intuition is wrong when stipulating that segmenting EST documents will be simpler and will achieve better results compared to other corpora. This shows that combining punctuation, lexical, and extensive morphological features is necessary to achieve good segmentation results.

We finally give in Table III the results of our best configuration (C2) per class a. For both corpora, the *End* class gets lower results compared to the *Inside* and the *Begin* class (in terms of F-score).

The error analysis of the outputs of classifier (C2) on the ATB documents shows that our classifier successfully distinguishes between the *Begin* and the *End* classes. In addition, the prediction of embedded EDUs is good in terms of precision (about 0.92, 0.90, and 0.70 for, respectively, the *Inside*, the *Begin*, and the *End* class). As we can see in the confusion matrix in Table IV, main confusions (in bold font) are between the *End* class and the *Inside* class.

Table V. Results of the (C2) Classifier with SAMA Features and the (C3) Classifier with Syntactic Features on Each Class

		ATB			
		P	R	F-score	Acc
(C2) / (C3)	Inside	0.938/0.938	0.966/**0.969**	0.952/**0.953**	0.922/0.923
	Begin	0.967/0.967	0.831/0.831	0.894/0.894	0.980/**0.981**
	End	0.735/**0.744**	0.658/0.650	0.695/0.694	0.944/0.943

The analysis of these confusions shows that most errors come from the presence of named entities and from weak punctuation marks. Examples (24.1) and (24.2) show, respectively, a gold-standard annotation and the output of our classifier. Our system predicts that the word و /w/and is a cutting point because the word أكرم/>krm/Akram has been analysed as the verb أكرم/>krm/to honor, which is, of course, wrong since this word is a named entity.

(24.1) [حصل خالد وأكرم على جائزة.]

[HSl xAld w>krm ElY jA}zp.]
[Khalid and Akram obtained an award.]

(24.2) [حصل خالد][وأكرم على جائزة .]

[HSl xAld][w>krm ElY jA}zp.]
[Khalid][and Akram obtained an award.]

Similarly, examples (25.1) and (25.2) show that our classifier fails to deal with weak punctuation marks. In (25.2) our classifier predicts an EDU boundary after the comma.

(25.1) [لن أعود لشرح الدرس، مرة أخرى.]

[ln >Ewd l$rH Aldrs, mrp >xrY.]
[I won't explain this lesson, again.]

(25.2) [لن أعود لشرح الدرس،][مرة أخرى.]

[ln >Ewd l$rH Aldrs,][mrp >xrY.]
[I won't explain this lesson,][again.]

5.1.2. Analyzing the Impact of Syntactic Features . We have finally assessed the reliability of syntactic features on discourse segmentation of ATB documents (refer to Table V) by adding chunk information to the features vector that achieved best performance in (C2). We observe that adding chunks does not really boost the results. The only improvements (in bold font in Table V) concern the recall of the *Inside* class (+ 0.003) and the precision of the *End* class (+ 0.011). The overall F-score of the (C3) classifier is 0.847, which corresponds to a marginal improvement of 0.010 compared to (C2). Similar observations go for the accuracy measure. We can thus conclude that shallow syntactic features are not useful for Arabic discourse segmentation.

5.2. EDU Recognition

An EDU is correctly recognized if, for each begin bracket, there is a corresponding end bracket. Otherwise, we have to perform a post-processing to ensure correct bracketing. Since the *End* class is the one that performs badly (cf. Table III), we have decided

Table VI. Accuracy (Acc) of EDUs Recognition Before and After Post-Processing

		Acc	
		EST	ATB
(C2) Before pre-processing	EDUs	0.408	0.631
	Embedded EDUs	0.307	0.572
(C2) After pre-processing	EDUs	0.795	0.769
	Embedded EDUs	0.615	0.671

to correct only end bracketing. Post-processing consists in adding an end bracket for each opening bracket that has no corresponding end. Table VI presents our results on both corpora in terms of Accuracy (Acc), before and after post-processing. For this experiment, we have run the classifier (C2) with all the features described in Table II except for the SAMA feature GLOSS (this corresponds to the penultimate line in Table II).

As expected, we observe that post-processing boosts the results for both ATB and EST with more than 0.39 for EST and 0.13 for ATB. The results are more impressive for EST (characterized by regular punctuation marks) because using punctuation features biased the EDUs' recognition results. Concerning embedded EDUs (present in around 11% in the EST corpus and 8% in ATB corpus), we have also observed the same tendencies. The results are, however, lower compared to the ones obtained for nonembedded EDUs. This may be explained by the low frequency of embedded EDUs in each test (around 8 for the EST test and 37 for the ATB test). Finally, we have observed that the performance of our segmenter is sensitive to the length of EDUs in terms of the number of tokens. Indeed, when this length is less than or equal to 3, we get an accuracy of 1.

5.3. The Learning Curve

Finally, in order to analyze how the learning procedure can be influenced by the number of annotated ATB documents, we have computed a learning curve by dividing our corpus into 10 different learning sets. For each set, we performed a tenfold cross-validation, using the features set of the classifier (C2). The learning curve is shown in Figure 3. As we can see, the curve grows regularly between 0 and 5 000 tokens (that is, 10 documents, i.e., around 255 sentences) while it seems to plateau between 5 000 and 25 000 tokens (that is, 50 documents). We can thus conclude that the addition of more than 10 ATB documents will only slightly increase the performance of the segmenter.

6. RELATED WORK

6.1. EDU Segmentation: Main Approaches

Several works have been undertaken on automatic discourse segmentation for different languages by using rule-based or learning techniques. In the first approach, handcrafted rules identify potential cutting points relying on a combination of surface cues (punctuation and lexical markers) and syntactic patterns that encode syntactic categories and parts-of-speech. In the English language, let us cite Le Thanh et al. [2004] that reported an F-measure of 86.9% when evaluating their segmenter against the boundaries in the RST Discourse Treebank (RST-DT) [Carlson et al. 2003]. Tofiloski et al. [2009] built the *SLSeg* system on top of an automatic syntactic parser and showed that their approach outperforms those of other approaches by achieving an F-score of 80–85% in segment boundary. Symbolic approaches have also been used in other languages like German [Lüngen et al. 2006], Spanish [Da Cunha et al. 2010],

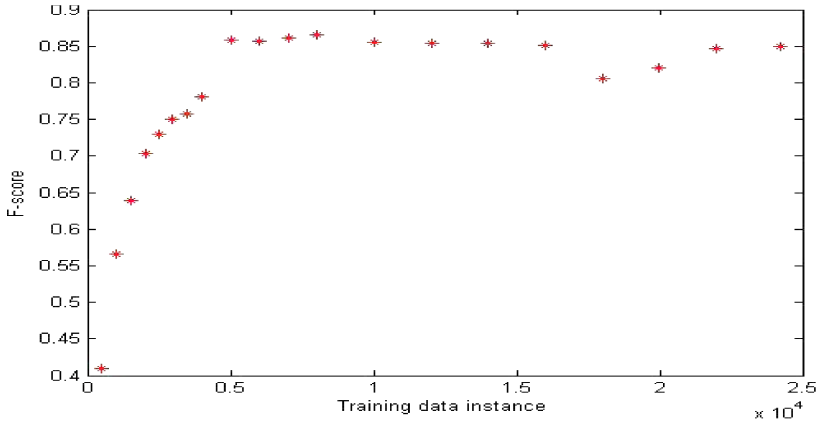

Fig. 3. The learning curve of (C2) for ATB corpus.

Brazilian Portuguese [Pardo et al. 2004], and Japanese [Sumita et al. 1992]. Most of these systems used the RST framework.

Learning approaches, on the other hand, usually exploit lexical and syntactic features to classify each token in a sentence as being an EDU boundary or not. Within the RST framework, Soricut and Marcu [2003] described how to split sentences into EDUs on top of SPADE, a sentence-level discourse parser. They made an extensive use of the syntactic tree and each token is modeled by taking into account syntactic dominance features (the token itself, its parent, and its siblings). Sporleder and Lapata [2005] used the RST-DT corpus and labeled each token with four different tags: B-NUC and B-SAT for nucleus, and satellite-initial tokens, and I-NUC and I-SAT for noninitial tokens. For the segmentation task, they performed a binary classification, where each span (and not tokens) can have a *Begin* or an *Inside* label. Span boundaries are given by the gold standard. Using this method, they showed that employing lexical and low-level syntactic information (such as parts-of-speech and syntactic chunks) is sufficient to achieve good performance. Their approach is comparable to Soricut and Marcu [2003]. Fisher and Roark [2007] proposed various improvements of SPADE by using finite-state analysis. Subba and Di Eugenio [2007] used a neural network (multilayer perceptron) while Hernault et al. [2010] used conditional random fields to train a discourse segmenter on the RST-DT corpus. For other languages, we cite Charoensuk et al. [2005] who proposed a hybrid approach for Thai using a decision-tree learning system and some heuristic rules.

All previously cited learning approaches do not deal with embedded EDUs and hence boundary detection is reduced to a binary classification task. However, nested EDUs can be frequent, as observed in the ANNODIS corpus [Afantenos et al. 2012], a discourse-level annotated corpus for French following SDRT principles. In this corpus, the proportion of embedded EDUs was about 10%. To predict nested structures, Afantenos et al. [2010] performed a four-way classification using the Maximum Entropy Model. Each token can be either a "left" or a "right" boundary of an EDU, "both" if an EDU contains only one token, or "none" if the token is in the middle of a segment. The segmenter made an extensive use of lexical and syntactic features and got an F-measure of 58%. A rule-based post-processing step increased the results up to 73%.

Current state-of-the art approaches in discourse segmentation make an extensive use of syntactic information going from chunking to deep syntactic parsing, including dependencies. However, some languages lack reliable deep syntactic parsers. Sporleder

and Lapata [2005] have already shown that good results can be reached only by chunking and that their approach can be portable to languages for which deep parsers are not available. We wanted here to go further by analysing to what extent EDU segmentation is feasible without using shallow syntactic information. We adopt a multiclass classification approach as done by Afantenos et al. [2012]. We use a combination of state-of-the-art features to predict nesting. To the best of our knowledge, the use of these features for Arabic discourse segmentation is novel.

6.2. EDU Segmentation for Arabic

Most research on Arabic NLP resource generation has focused on morphology [Boudlal et al. 2011], lexical semantics [Diab et al. 2008], and syntactic analysis [Maamouri et al. 2010b]. There is also a huge literature on Arabic NLP including shallow and deep syntactic parsing [Ali Mohammed and Omar 2011; Diab et al. 2007, 2009; Green and Manning 2010; Marton et al. 2013; Nivre 2007], morphology analysis [Eskander et al. 2013; Gridach and Chenfour 2011; Sawalha et al. 2013], question answering [Bebajiba et al. 2010; Trigui et al. 2012], automatic translation [Carpuat et al. 2012; Sadat and Mohamed 2013], opinion mining and sentiment analysis [Abdul-Mageed and Diab 2012; Abu-Jbara et al. 2013; Mourad and Darwish 2013], and named entity recognition [Aboaoga and Ab-Aziz 2013; Boujelben et al. 2013; Darwish 2013].

On the discourse level, however, little work has been done. Among them, let us cite Belguith et al. [2005] that proposed a rule-based approach to segment nonvoweled Arabic texts into sentences. The approach consists of a contextual analysis of the punctuation marks, the coordination conjunctions, and a list of particles considered as boundaries between sentences. The authors determined 183 rules to segment texts into paragraphs and sentences. These rules were implemented in the STAr system, a tokenizer based on the proposed approach. Touir et al. [2008] proposed a rule-based approach to segment Arabic texts using connectors without relying on punctuation marks. Segmentation principles did not follow any discourse theory. They performed an empirical study of sentence and clause connectors and introduced the notion of active connectors, which indicate the beginning or the end of a segment and the notion of passive connectors that do not imply any cutting point. Passive connectors are useful only when they co-occur with active connectors since this might imply the beginning or the end of a segment. Khalifa et al. [2011] proposed a learning approach to segment Arabic texts by only exploiting the rhetorical functions of the connector و/w/and. Among the six rhetorical types of this connector (cf. Section 3.2.2), two classes have been defined: "Fasl", which is a good indicator to begin a segment, and "Wasl", which has no effect on segmentation. A set of 22 syntactic and semantic features was then used in order to automatically classify each instance of the connector و/w/and into these two classes. The authors reported that their results outperform those of Touir et al. [2008] when considering the connector و/w/and. Finally, Keskes et al. [2012] used a rule-based approach to segment Arabic texts into clauses. They proposed three segmentation principles: (p1) using only punctuation marks, (p2) relying only on lexical cues, and (p3) using both punctuation marks and lexical cues. Better results were achieved by the third principle. The authors reported that major errors are due to lexical ambiguities of discourse cues.

The closest research to ours is the one done by Al-Saif and Markert [2010, 2011] that, respectively, described how to recognize discourse connectives and how to automatically identify explicitly marked discourse relations within the Penn Discourse Treebank (PDTB) framework [Prasad et al. 2008]. Discourse segmentation in PDTB tends to larger units than EDUs since arguments can be as small as a nominalization or as large as several sentences. Segmentation in PDTB requires three main steps: (1) identifying discourse connectives, (2) identifying the locations of Arg1 and

Arg2, and (3) labeling their extent. Arg1 can be located within the same sentence as the connective or in some previous sentences of the connective. When Arg1 and Arg2 are in the same sentence, we can have several cases: Arg1 coming before Arg2 as in (25), Arg1 coming after Arg2 as in (26), and Arg2 embedded within Arg1 as in (5) (see Section 3.2.2).

(25) [تعرضت اضرار]$_{arg1}$ [نتيجة لاصطدام.]$_{arg2}$

[tErDt ADrAr]$_{arg1}$ *[ntyjp 0lASTdAm]* $_{arg2}$
[Suffered damages]$_{arg1}$ *[as a result of the collision.]* $_{arg2}$

(26) [في حين انها حامل،]$_{arg2}$ [إيناس لا تأخذ الطائرة.] $_{arg1}$

[fy Hyn AnhA HAml,] $_{arg2}$ *[<ynAs lA t>x* AlTA}rp.]* $_{arg1}$
[While she is pregnant,] $_{arg2}$ *[Ines did not take the plane.]* $_{arg1}$

In case of embedding (subordinating connectives, coordinating connectives and discourse adverbials), the full syntactic parse tree of the sentence is needed in order to extract the Arg1 and Arg2 spans. Al-Saif and Markert [2011] have described only the step (1) given before and did not treat embedded EDUs. In addition, they did not give any indication of how the steps (2) and (3) given earlier can be automatically performed for Arabic texts.

7. CONCLUSION

The field of Arabic NLP is still very vacant at the layer of discourse. Our article proposed the first corpus and the first approach that tackle discourse segmentation in terms of elementary discourse units for Arabic texts. A subset of this corpus can be uploaded at the following address: https://sites.google.com/site/iskandarkeskes85/corpus. Our main contributions are the following.

— We given an Arabic corpus that includes a discourse-level annotation. Indeed, the only existing work towards producing an Arabic discourse Treebank is the work Al-Saif and Markert [2011] that extends the Penn Discourse Treebank (PDTB) to MSA. In this corpus, annotated elements are the discourse connectives and their two arguments.
— We provide a multiclass supervised learning approach that predicts EDU boundaries and not only discourse connectives, as in Al-Saif and Markert [2011]. Our approach uses a rich lexicon (with more than 174 connectives) and relies on a combination of punctuation as well as morphological and lexical features. Our results show that EST segmentation is very sensitive to punctuation features, contrary to ATB, where punctuations are not widely used. In addition, contextual lexical features have a strong effect on the results, especially for ATB, which shows that ATB documents tend to use more complex words than for EST. For both corpora, we have shown that extensive morphological features are more suitable than shallow morphological analysis, since best scores were obtained when adding information of the root, the prefix, and the suffix. Finally, we have shown that Arabic discourse segmentation is feasible on both corpora without any use of shallow syntactic information (chunks).
— We discern EDU frontiers even in the case of absence of discourse markers (that is, in the case of implicit relations) that represent 25% of cases in our data. Al-Saif and Markert [2011] have treated only the cases of explicit markers.

For the moment, we have run our experiments by considering Alkhalil features and SAMA features separately. It would be interesting in the future to run our classifiers

by combining features from both sets. Another improvement could be to use a context-aware parser such as MADA [Habash et al. 2009]. Discourse segmentation is the first step towards discourse analysis. An annotation of ATB documents with coherent relations within the SDRT framework is currently underway.

REFERENCES

Abdul-Mageed, M., and Diab, M. 2012. AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*.

Aboaoga, M., and Ab-Aziz, M. J. 2013. Arabic person names recognition by using a rule based approach. *J. Comput. Sci. 9*, 7, 922–927.

Abu-Jbara, A. King, B. Diab, M., and Radev, D. 2013. Identifying opinion subgroups in Arabic online discussions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics - Short Papers (ACLShortPapers'13)*.

Afantenos, S. D., Denis, P., Muller, P., and Danlos, L. 2010. Learning recursive segments for discourse parsing. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'10).*

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., and Vieu, L. 2012. An empirical resource for discovering cognitive principles of discourse organisation: The annodis corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*.

Ali Mohammed M. and Omar N. 2011. Rule based shallow parser for Arabic language. *J. Comput. Sci. 7,* 10, 1505–1514.

Al-Saif, A., and 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'10)*.

Al-Saif, A., and Markert, K. 2011. Modelling discourse relations for Arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*.

Asher, N. and Lascarides, A. 2003. *Logics of Conversation*. Cambridge University Press.

Bebajiba, Y. Rosso, P. Abouenour, L. Trigui, O. Bouzoubaa, K., and Belguith, HL. 2010. Question answering for semitic languages. In *Natural Language Processing Approaches to Semitic Languages*, Pr. Imed Zitouni Ed., Springer, 345–347.

Belguith, H. L. 2009. Document analysis and summarisation: Problems, conception and implementation. In *Habilitation at Faculty of Economics and Management of SFAX*.

Belguith, H. L., Baccour, L., and Mourad, G. 2005. Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. In *Proceedings of the 12th Conference on Natural Language Processing*.

Berger, S., Pietra D., and Della V. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist. 22*, 1, 39–71.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., and Bebah, M. 2011. Alkhalil morpho sys: A morphosyntactic analysis system for Arabic texts. http://www.itpapers.info/acit10/Papers/f653.pdf.

Boujelben, I. Jamoussi, S., and Ben Hamadou, A. 2013. Enhancing machine learning results for se-mantic relation extraction. In *Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems*. 337–342.

Carlson, L., Marcu, D., and Okurowski, M. E. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and New Directions in Discourse and Dialogue*, Springer, 85–112.

Carpuat, M., Marton, Y., and Habash, N. 2012. Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment. *Mach. Translat. 26*, 1–2. 105–120.

Charoensuk, J., Suvakree, T., and Kawtrakul, A. 2005. Thai element discourse unit segmentation for Thai discourse cues and syntactic information. In *Proceedings of the 9th National Computer Science and Engineering Conference*.

Da Cunha, I., San Juan, E., and Torres M. 2010. Discourse segmentation for Spanish based on shallow parsing. In *Proceedings of the 9th Mexican International Conference on Advances in Artificial Intelligence (MICAI'10)*. Springer, 13–23.

Darwish, K. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*.

Debili, F., Achour, H., and Souissi, E. 2002. La langue arabe et l'ordinateur, de l'étiquetage grammatical à la voyellation automatique. *Correspondances de l'IRMC 71*, 10–28.

Diab, M. 2009. Second generation Amira tools for Arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools.*

Diab, M. Hacioglu, K., and Jurafsky, D. 2007. Automated methods for processing Arabic text: From tokenization to base phrase chunking. In *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*. Springer, 159–179.

Diab, M., Moschitti, A., and Pighin, D. 2008. Semantic role labeling systems for Arabic using kernel methods. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'08).*

Eskander, R. Habash, N. Bies, A. Kulick, S., and Maamouri M. 2013. Automatic correction and extension of morphological annotations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'13)*.

Fisher, S. and Roark, B. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 488–495.

Green, S. and Manning, C. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10).* 394–402.

Gridach, M. and Chenfour, N. 2011. Developing a new system for Arabic morphological analysis and generation. In *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP'11).* 52–57.

Grosz, B. and Sidner, C. 1986. Attention, intention and the structure of discourse. *Comput. Linguist. 12*, 175–204.

Habash, N. 2010. *Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies*. G. Hirst Ed., Morgan and Claypool.

Habash, N., Owen R., and Ryan R. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*.

Hernault, H., Prendinger, H., duVerle, D., and Ishizuka, M. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue Discourse 1*, 3, 1–33. http://elanguage.net/journals/index.php/dad/article/view/591.

Kamp, H. 1981. A theory of truth and semantic representation. In *Formal Semantics: The Essential Readings.* Wiley, 189–213.

Keskes, I., Benamara, F., and Belguith, H. L. 2012. Clause-based discourse segmentation of Arabic texts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation.*

Khalifa, I., Feki, Z., and Farawila, A. 2011. Arabic discourse segmentation based on rhetorical methods. *Int. J. Electric Comput. Sci. 11*, 1.

Le Thanh, H., Abeysinghe, G., and Huyck, C. 2004. Generating discourse structures for written text. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04).* 329–335.

Lüngen, H., Lobin, H., Bärenfänger, M., Hilbert, M., and Puskas, C. 2006. Text parsing of a complex genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB'06)*, B. Martens and M. Dobreva Eds.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. 2010a. Standard Arabic morphological analyzer (sama) version 3.1. Linguistic Data Consortium, Catalog No. LDC2010T08.

Maamouri, M., Bies, A., Kulick, S. Krouma, S., Gaddeche, F., and Zaghouani, W. 2010b. Arabic Treebank (ATB): Part 3 Version 3.2. Linguistic Data Consortium, Catalog No.: LDC2010T01.

Mann, W. C. and Thompson, S. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text 8*, 3, 243–281.

Marton, Y. Habash, N., and Rambow O. 2013. Dependency parsing of modern standard Arabic with lexical and inflectional features. *J. Comput. Linguist. Archive 3*, 1, 161–194.

Mourad, A. and Darwish, K. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 55–64.

Nivre, J. 2007. Incremental non-projective dependency parsing. In *Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07)*. 396–403.

Pardo, T. A. S., Nunes, M. G. V., and Rino, L. H. M. 2004. DiZer: An automatic discourse analyzer for Brazilian Portuguese. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence*. Lecture Notes in Computer Science, vol. 3171, Springer, 224–234.

Prasad, A., Miltsakaki, R., Dinesh, E., Lee, N., Joshi, A., and Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.*

Polanyi, L. 1988. A formal model of discourse structure. *J. Pragm. 12*, 601–639.

Polanyi, L. and Scha, R. 1984. A syntactic approach to discourse semantics. In *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting on Association for Computational Linguistics (COLING'84).* 413-419.

Polanyi, L. and Van Den Berg, M. 1996. Discourse structure and discourse interpretation. In *Proceedings of the 10th Amsterdam Colloquium on Formal Semantics*.

Sadat, F. and Mohamed, E. 2013. Improved Arabic-French machine translation through preprocessing schemes and language analysis. In *Proceedings of the 26th Canadian Conference on Artificial Intelligence (AI'13)*. 308–314.

Sawalha, M. Atwell, E. S., and Abushariah M. 2013. SALMA: Standard Arabic language morphological analysis. In *Proceedings of the 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*. 1–6.

Soricut, R. and Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*. 149–156.

Sporleder, C. and Lapata, M. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 257–264.

Subba, R. and Di Eugenio, B. 2007. Automatic discourse segmentation using neural networks. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*. 189–190.

Sumita, K., Ono, K., Chino, T., Ukita, T., and Amano, S. 1992. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on 5th Generation Computer Systems*. 1133–1140.

Tofiloski, M., Brooke, J., and Taboada, M. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP Conference of Short Papers*. Association for Computational Linguistics, 77–80.

Touir, A., Mathkour, H., and Al-Sanea, W. 2008. Semantic-based segmentation of Arabic texts. *Inf. Technol. J. 7*, 7.

Trigui, O., Hadrich-Belguith, L., Rosso, P., Ben Amor, H., and Gafsaoui, B. 2012. IDRAAQ: New Arabic question answering system based on query expansion and passage retrieval. In *Notebook Papers of CLEF LABs and Workshops (CLEF'12)*.

Webber, B. L. 2004. D-LTAG: Extending lexicalized tag to discourse. *Cogn. Sci. 28*, 5, 751–779.

Wolf, F. and Gibson, E. 2006. *Coherence in Natural Language: Data Structures and Applications*. MIT Press.

Received January 2013; revised March 2014; accepted March 2014