Clause-based Discourse Segmentation of Arabic Texts

Iskandar Keskes[1,2], Farah Benamara[1], Lamia Hadrich Belguith[2]

[1]IRIT-Toulouse University, 118 Route de Narbonne, F-31062 Toulouse - France
[2] ANLP-Research Group, Miracl-Sfax University, FSEGS, BP 1088, 3018, Sfax - Tunisia
E-mail: keskes@irit.fr, benamara@irit.fr, l.belguith@fsegs.rnu.tn

Abstract

This paper describes a rule-based approach to segment Arabic texts into clauses. Our method relies on an extensive analysis of a large set of lexical cues as well as punctuation marks. Our analysis was carried out on two different corpus genres: news articles and elementary school textbooks. We propose a three steps segmentation algorithm: first by using only punctuation marks, then by relying only on lexical cues and finally by using both typology and lexical cues. The results were compared with manual segmentations elaborated by experts.

Keywords: discourse segmentation, Arabic natural language processing, clauses.

1. Introduction

Discourse structure is essential in determining the content conveyed by a text. It affects for example, the temporal structure of a text, the interpretation of anaphoric expressions and presuppositions. Discourse structure has shown to be useful in many NLP applications, such as automatic text summarization (Marcu, 2000) and question answering (Chai and Jin, 2004). Discourse parsing consists in two steps: (1) discourse segmentation which aims at identifying Elementary Discourse Units (EDU), and (2) building the discourse structure by linking EDUs using a set of rhetorical or discursive relations. We deal in this paper with the first step, focusing on Arabic text segmentation.

Discourse segmentation aims at splitting texts into non-overlapping units. This task is theory dependent since each discourse theory defines its own specificities in terms of segmentation guidelines and size of units. A simple way is to consider a sentence as a basic unit (Halliday and Hasan, 1976). However, sentences can be long and can contain several smaller units that can be related with discourse relations. In RST (Mann and Thompson, 1988), EDUs can be simple sentences or clauses in a complex sentence that typically correspond to verbal clauses, as in the sentence below where we have two EDUs:
[This is the best book] [that I have read in along time.]

EDUs can also correspond to other syntactic units describing eventualities, such as prepositional and noun phrases, as in the following examples where we have respectively two and three EDUS:
[After several minutes,][we found the keys on the table.], and,
[Mary Smith,] [who is now in that corner,] [wants to meet you.]

RST does not allow for nested EDUs. On the contrary, other theories like SDRT (Asher and Lascarides, 2003), allow for embedded segments in order to encode adjuncts such as appositions or cleft constructions with discursive long-range effects such as frame adverbials, non-restrictive relatives and appositions, as in:
[Mr. Dupont, [a rich business man,][living on the Paris region], was savagely killed].

Several research works have been undertaken on automatic discourse segmentation for different languages using both rule-based and learning techniques. Segmentation principles rely mainly on: discourse cues, punctuation marks and syntactic information going from parts of speech, chunks to full syntactic parsing, including dependencies. Within the RST framework, recent works include (Seeger and Brian, 2007) (Da Cunha et al., 2010) (Harald et al, 2006) and (Jirawan et al, 2005) for respectively English, Spanish, German and Thai languages. We finally cite (Afantenos et al., 2010) who developed, within the SDRT framework, a discourse segmenter for French texts that handles nested structures.

In this paper, we propose a rule-based approach to Arabic texts segmentation, where segments are sentences, clauses as well as other constructions including prepositional and noun phrases.

2. Related works

In Arabic, discourse segmentation has not been fully addressed mainly because EDU segmentation in Arabic is more complex than other languages. First, Arabic is an agglutinative language in which the clitics are agglutinated to words. Indeed, prepositions (like ف (then)), conjunctions (like و (and)), articles (like ال (the)) and pronouns can be affixed to nouns, adjectives, particles and verbs which causes several lexical ambiguities. For example, فهم / “fahm“ can be a noun (that means understanding) or a conjunction (ف/”fa”/ then) followed by the pronoun (هم/ “hom “/they).

Second, unlike Indo-European languages, Arabic does not have capital letters which makes the task of text segmentation into sentences harder than the one for other languages such as English, where the capital letters are used as cues for text splitting.

Moreover, Arabic texts can be diacritized, partially diacritized, or totally non diacritized. Most current Arabic documents are not diacritized. Indeed, the diacritics (i.e. orthographic symbols, which represent among other things short vowels) are only used in educational books for beginners. It should be noted that non diacritized texts are highly ambiguous: the proportion of ambiguous words exceeds 90%. For example, the word كتب [ktb] could be diacritized in 21 different ways (Debili et al., 2002). Among these forms, we can cite "كَتَبَ / he wrote" and "كُتُبٌ / books". The same confusion holds between the verb (ذَهَبَ/go) and the noun (ذَهَبٌ/gold). Thus, a non diacritized 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. Hence, the absence of diactritics in Arabic texts is another difficulty which confirms that EDU segmentation in Arabic is more complex than the one for other languages such as English or French.

Most researches on Arabic discourse segmentation aim at splitting texts into paragraphs, sentences or clauses. (Belguith et al., 2005) proposed a rule-based approach to segment non-vowelled Arabic texts into sentences. The approach consists of a contextual analysis of the punctuation marks, the coordination conjunctions and a list of particles that are 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. Star is used in many Arabic NLP systems such as MORPH, a morphological analyser for Arabic texts (Belguith et al. 2005), MASPAR, a Multi-Agent System for Parsing Arabic (Belguith et al., 2008) and Al-Lakas El'eli, an Arabic automatic summarization system (Maâloul et al., 2008).

(Touir et al., 2008) proposed a rule-based approach to segment Arabic texts using connectors and without relying on punctuation marks. Segmentation principles do not follow any discourse theory. They perform an empirical study of sentences and clauses connectors in order to segment Arabic texts while preserving the semantic of its constituents. They introduce the notion of active connectors, which indicates the beginning or the end of a segment and the notion of passive connectors that does 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.

Finally, (Khalifa et al., 2011) proposed a learning approach to segment Arabic texts by exploiting the rhetorical functions of the connector "و / and ". Among the six rhetorical types of this connector, two classes have been defined: "Fasl" which is a good indicator to begin a segment, and "Wasl" which does not have any effect on segmentation. A set of 22 syntactic and semantic features were then used in order to automatically classify each instance of the connector "و" into these two classes. The authors reported that their results outperform the results of (Touir et al., 2008) when considering the connector "و".

Our approach is novel in three ways. First, it relies on an extensive analysis of a large set of lexical cues as well as punctuation marks. It goes thus beyond the method proposed by (Touir et al., 2008) since we handle both a greater number of lexical cues and punctuation marks. Our approach goes also beyond the work of (Khalifa et al., 2011) since their method relies only on one discourse cue. In addition, our analysis was carried out on two different corpus genres: news articles and elementary school textbooks. Corpus analysis allows us to group connectors into different categories depending whether they are (or not) a good indicator to begin or end a segment.

Second, unlike (Belguith et al., 2005), our approach relies on morphological and syntactic information using several dictionaries and orthographic rectification grammar. To this end, we use NooJ linguistic resources (Mesfar, 2008) in order to perform surface morphological and syntactic analysis.

Finally, we propose a three steps segmentation algorithm: first by using only punctuation marks, then by relying only on lexical cues and finally by using both typology and lexical cues. The results were compared to manual segmentations elaborated by experts.

3. Data

We conducted a corpus study on two different corpora: 150 news articles (737 paragraphs, 40532 words) and 250 elementary school textbooks (EST) (1095 paragraphs, 29473 words). The corpus was manually segmented by three linguists. In order to better understand the segmentation principles and due to the complexity of the task, the annotation relies on a consensus.

The distribution of the number of texts and segments per genre is shown in Table 1. We get a total of 4725 segments for the news article and 2625 for the textbooks. 80% of the news articles and 60% of the textbooks were used for building our segmentation patterns. The rest of the corpus was left for test.

4. Segmentation principles

During the corpus analysis, three different segmentation principles were identified: (p1) using punctuation marks only, (p2) using discourse cues only, and (p3) using both the principles (p1) and (p2).

	Training corpus		Test corpus	
	texts	**segments**	**texts**	**segments**
4th EST	30	604	17	340
5th EST	28	550	15	260
6th EST	30	400	20	301
7th EST	31	541	22	315
8th EST	32	630	25	345
News	100	4725	50	2450
Total	**251**	**7350**	**149**	**4011**

Table1: The training and test corpus

4.1 P1: Punctuation marks principles

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/ wa”, which represents the conjunction (“and”) for English. Borrowed by the Italian typographers, the comma becomes mute in the Latin alphabet. The point 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 parentheses, the exclamation point, the question mark, the three points, etc. have the same values as those of European languages (Belguith, 2009). It should be noted that some punctuation marks in Arabic look different from the European ones. Indeed, the Arabic comma points to the opposite way (،) and it is written on top of the line. Also, the Arabic question mark looks to the opposite side (؟).

The punctuation marks are not widely used in current Arabic texts (i.e., at least not regularly) and when they are used, they do not respect the typography rules[1]. Therefore, their presence cannot guide the segmentation process as for other languages such as English or French which make segmenting Arabic text harder.

During the segmentation process, annotators classify punctuation marks into two categories: **strong indicators** that always identify the end of a segment and **weak indicators** that do not always indicate the beginning or the end of a segment. In our corpus, annotators identify 4 strong indicators: the exclamation mark (!), the question mark (?), the colon (:) and the semi-colon (;), as well as 6 weak indicators: the full stop (.), the comma (,), quotes, parenthesis, brackets ([]), braces ({}) and underscores. The dot and the comma are most frequent in our corpus.

We give below some examples of strong indicators.

(1) [ألقيت كلمة مازالت أحفظها إلى هذا اليوم :][«وطني. أحبّك يا وطني. »]
[I said a word that I still remember still today:] [«My country. I love you dear country. »].

(2) [طردت خليل من المدرسة ؛] [لأنه غش في الامتحان.]
[Khalil was expelled from school;] [because he cheats in the exam.]

In order to handle weak indicators, we design a set of decision rules, such as:

- If the full stop is part of a named entity, it doesn’t represent the end of a segment.

(3) [د. طارق سويدانسالج أمرض مختلفة.]
[Dr. Tarak Swiden has treated various diseases.]

(4) [يعتبر فيتامين ب.2 و ب.12 من اكثر الفيتامينات التي تسلسدسلى مقاومة الزهايمر.]
[The vitamins B.2 and B.12 are considered as the most effective to fight against Alzheimer illness.]

- If the dot is preceded by one word and if this word is not a verb, then dot doesn’t represent the end of a segment.

(5) [وطني. أحبّك يا وطني.]
[My country. I love you dear country.]

- If the comma is followed by a verb or (اشارة اسم/a demonstrative pronoun), then it represents the end of segment:

(6) [ترك بيروت ،] [لذلك كانت زوجته ليست دائماً إلى جانبه]
[He left Beirut,][this is why his wife was not always with him.]

- If an apposition contains only a named entity, then it does not represent the end of a segment, as in :

(7) [كتب الشسلر الكبير،نزار القباني، أشعار كثيرةسن المرأة.]
[The great poet, Nizar Qabani, wrote many poems about woman.]

- For the other weak indicators, i.e quotes, parenthesis, brackets, braces and underscores, they usually indicate the beginning of a segment only if they contain a verbal clause.

(8) [طرق المدير باب القس][(حيّانا ببشاشة)] [وتقدّم إلى معلّمنا.]
[The director knocks the door of the class room][(he smiles)][and then he comes to talk to our teacher.]

(9) [قال المدير “ تحيّة العلم”][فانقطعت كل حركة.]
[The director said “salute the flag”][and then movements have stopped.]

Although the Arabic language has punctuation marks, written Arabic rarely contains these punctuations. Arabic discourse tends to use long and complex

[1] (Basha, 1912) defined the writing rules of the different punctuation marks and their values in the Arabic text.

sentences, so we can easily find an entire paragraph without any punctuation. Therefore, segmenting according to p1 is not enough.

4.2 P2: Lexical cues principles

Using lexical cues could be a solution to further segment sentences into clauses, as in the following example where we have a contrast discourse relation.

(10) [سيعرف الجميع متى نبدأ][لكن لا يعرفوا متى ننتهي]
[They will know when we start][but they don't know when we finish]

Like punctuation marks, lexical cues were grouped into two classes: ***unambiguous*** and ***ambiguous***. In the first class, connectors are usually followed by a verb which is a strong cue to indicate the end of a segment. Annotators have listed 97 unambiguous lexical cues. Here are some of our rules:

- If a verb is followed by {لـ، كي، حتى، من أجل أن، كيما، لكيلا، لئلا}, it indicates the end of a segment as in:

(11) [فبعض الكتاب يستخدمون كلمات سهلة في مقالاتهم] [**من أجل أن** يفهمها القراء.]
[Some authors use simple words in their articles] [**in order** to be understood by readers.]

- If a verb is followed by one of the lexical cues { إلا، لكن، لكنّ، بيد أن،غير أن، أن، بحيث } or if these cues are proceeded by the conjunction "و" (waw) or "ف" (fā), then it indicates the end of a segment as in:

(12) [يمكنك يا أخي أن تستغني عن مالك بأية حال] [**ولكن** ابتعد عن التبذير.]
[You can spend your money] [**but** avoid to fritter away.]

(13) [نحرص على نظافة المطبخ][**بحيث** يتم التخلص من أي بقايا طعام]
[We keen to clean the kitchen] [**so** as to get rid of any remnants of food]

On the other hand, ambiguous connectors do not always mark the beginning of a segment, as the connector "و / and " and the particles (“ثم” (and), “ف” (So), etc.). For example, the particle “و” can express either a new clause (cf. example (14)), a conjunction between NPs (cf. example (15)), or it can be a part of a word (cf. example (16)).

(14) [فنظر إليّ،] [**و**قال:]
[Then he looked at me,] [**and** he said :]

(15) [فلاحظ البائع **و** الحريف يتناقشان على أسعار البضاعة]
[Then he remarked the customer **and** the client discussing about the products’ prices.]

(16) [كانت كل ورشة عمل تشكو من افتقار أجهزة العمل.]
[Each workshop suffers from a lack of equipments.]

During the annotation process, we observed that the lexical cues principles cannot resolve some ambiguities related to weak indicators (49 ambiguous lexical cues were identified). In addition, we have also observed that some connectors can be easily disambiguated using punctuation marks. We need therefore to use both the punctuation mark and the lexical cue that follows it in the sentence in order to better identify the right segment frontiers.

4.3 P3: Mixed principles

We give in this section some rules that illustrate these principles.

- If comma is followed by the conjunction "و" (waw) or "ف" (fā) and then by a preposition of localisation {على, في, عن, من, إلى }, it indicates the end of a segment, as in:

(17) [كان أهله على عادة كثير من العائلات التّونسيّة يتخلّعون ببلدة المرسى،] [**وعلى** شاطئها البديع بدأ اللّقاء حميما بينه وبين الطّبيعة.]
[Like Tunisian families, her family left Marsa city,] [then, they found themselves at the wonderful Marsa’s beach.]

- If comma is followed by the conjunction "و" (waw) or "ف" (fā) and then by a possessive noun {له, لها, لهما, لهنّ, لنا,لهم, لي , لك, لكم, لكما, لكنّ}, it indicates the end of a segment, as in:

(18) [رأيت أختي في الخارج،] [**لها** دمية تتكلم.]
[I saw my sister outside,] [with a talking doll]

- If a comma is followed by a demonstrative pronoun {تلك, هذه, هذا, ذاك, ذلك, لهذا, بهذا, لهذه, بهذه, لذلك, بذلك} and then by a word that is not a verb, then, we do not have a segment frontier, as in:

(19) [وقف معلّمنا سي حامد، **هذا** اليوم،أمامنا ينظر في وجوهنا مليّا.]
[Mr. Hamed, our teacher, was standing up, looking at us.]

5. Our approach

In order to assess the validity of the previous segmentation principles, we designed three discourse segmenters. The first two ones are based respectively on the principles p1 and p2 while the last one is based on the principle p3. To build the third segmenter, we propose a three steps segmentation algorithm. First, texts are segmented according to p1. This leads to a first segmentation level which is refined according to the principle stated in p2. The final segmentation is obtained by applying the principle p3. Each step has its own patterns coupled with linguistic resources (Mesfar, 2008) like dictionaries of verbs, nouns, adjectives as well as morphological and syntactic surface analysis in order to resolve the agglutination problem. These dictionaries are used to recognize the type of indicators as well as their right and left contexts. The figure below describes the general architecture of our system. The output is an XML file that contains the segmented text.

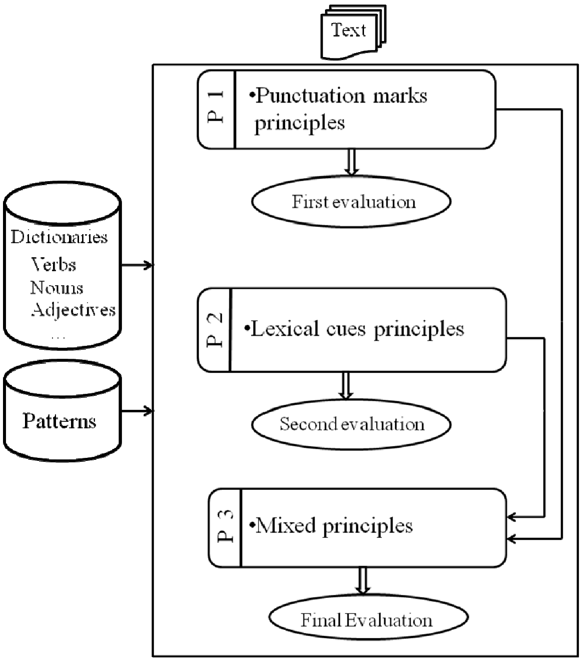

Figure 1: A rule-based approach to discourse segmentation

Our segmentation process is implemented using the linguistic platform NooJ (Silberztein, 1993). NooJ is a linguistic development environment that can parse texts of several million words in real time. It includes tools to construct and maintain large coverage lexical resources, as well as morphologic and syntactic grammars. Based on this platform, we built our patterns using a set of linguistic Arabic resources. The patterns presented previously are described in NooJ local grammars. These local grammars are used in NLP applications as finite-state transducers ranged from morphological analysis to finite-state parsing.

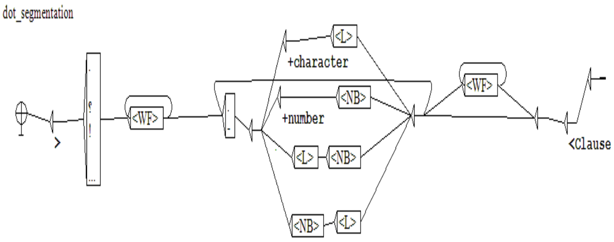

Figure 2: NooJ local sub-grammar for dot marker.

The figure 2 presents an example of a NooJ local grammar for the segmentation using dots: if there is an abbreviation in the beginning or in the middle of a sentence, the dot does not represent the end of a segment.

We give below an example of a text segmented following the principle p3.

[كان دبّان قطبيّان نائمين على مقربة من برج المراقبة,][عندما وصلت مع عالم الطبيعة "جون كروغر".][وما إن سمعا صوت الطوّافة][حتّى انطلقا يعدوان مذعورين.][تتجمّع مئات من الدببة القطبيّة كلّ خريف قرب رأس تشرشل على الشاطئ الغربيّ لخليج هدسون في كندا.][وهي تقتات في الشّتاء والرّبيع وأوائل الصيف بحيوانات الفقمة][الّتي تصطادها فوق الجليد.][علما أنّها أضخم الحيوانات اللاّحمة][الّتي تدبّ على وجه الأرض.][ومع تقاصر ساعات النّهار تنتقل الدببة شمالا ببطء][وتتجمّع على مقربة من رأس تشرشل ورؤوس أخرى في انتظار تراكم الجليد من جديد.]

6. Evaluation and results

Our three discourse segmenters, that follow respectively the principles p1, p2 and p3, have been evaluated on the test set for both news articles and textbooks. Table 2 summarizes the obtained results.

	Segmen-tation level	Precision	Recall	F-Measure
EST	P1	46%	44%	45%
	P2	68%	64%	66%
	P3	**86%**	**85%**	**85,5%**
Journal	P1	20%	22%	21%
	P2	55%	52%	53,5 %
	P3	**69%**	**67%**	**68%**

Table 2: Evaluation results

As expected, the first level segmentation (i.e., based on punctuation marks) performs bad. The obtained results for textbooks are better than those for news articles mainly because textbooks are usually well structured and they are characterized by the presence of punctuation marks. Main errors come from weak punctuation marks. For instance, our rules for dots do not perform well in case of the presence of abbreviations at the end of the segment, since this does not imply a cutting point (cf. example (20)).

(20) [حصل على جائزة البنك الإسلامي للتنمية في الاقتصاد الإسلامي لعام 1411 هـ.]
[He obtained the Islamic bank award for the developpement in the Islamic economy for the year 1411 H.]

We also observe that our rules for commas often fail mainly because our system do not correctly handle lexical ambiguities, as in:

(21) [أكل الولد تفاحة، بعد غسلها]
[The child has washed the apple,] [then he ate it],

where the adverb بعد / (after) was identified as a verb بعد/ (to move away).

The second level segmentation obtained better results compared to the first level for the two corpora which shows that lexical cues are good indicators to segment sentences into clauses. Results for textbooks are however better compared to news articles mainly because textbooks writing style are very simple which make the number of ambiguity cases most frequent in journal articles. As for segmentation principle p1, main errors come from lexical ambiguities, as in:

(22) وصف الطبيب للمريض مجموعة من الأدوية لمعالجة ألمه وجرحه
According to the doctor's instruction, the patient has to take a lot of drugs to treat his pain and his injury,

where the system identifies a cutting point before وجرحه, since the morpho-syntactic analysis consider this word as a verb and not as a noun. Errors come also from the syntactic parser, as in:

(23) استقبلت عائلة مصطفى فضل البارحة
Yesterday I received Mustapha Fadhl's family,

where, the named entity فضل is parsed as a conjunction ف and a verb ضل (lost) which implies a beginning of a segment.

Finally, segmenting using both punctuations and lexical cues gives the best results. This shows that using morphological and syntactic information is helpful to disambiguate some lexical connectors as well as weak punctuation marks. Of course, mixed principles have their limits because, in some cases, both punctuation marks and lexical connectors are omitted, as in:

(24) فأخذنا نقرأ بعض الصّفحات معا نناقش ما احتوتها
We have read together some pages and then we have discussed about their content,

where we have two segments related by the rhetorical relation goal.

The main challenge in Arabic discourse segmentation remains the disambiguation of discourse cues. In fact, Arabic being an agglutinative language, we have to go beyond standard morpho-syntactic analysis, in order to deal with lexical ambiguities. We thus need semantics. Interesting efforts in this direction include the work of (Khalifa et al., 2011) on the connector "و /wa" that can be used efficiently in our framework to improve the results of our system when using the principle p3.

7. Conclusion

In this paper, we proposed a rule-based approach for Arabic texts segmentation into clauses. Based on a linguistic study of two different corpora (news articles and elementary school textbooks), we identified three segmentation principles: one based on the exclusive use of punctuation marks, the second relies on lexical cues and the last one is based on a combination of the first two principles. Our results show that the third principle is the best segmentation algorithm and that segmenting elementary school books yield better results compared to news articles.

For the moment, our method relies on morphological and syntactic information using several dictionaries and orthographic rectification grammar. Arabic texts segmentation needs in addition a semantic analysis in order to resolve lexical ambiguities.

As future work, we intend to segment clauses into minimal units to take into account appositions, adverbial frames, etc. Then, we plan to study how those EDUs are discursively related using SDRT theory (Asher & Lascarides, 2003) as our formal framework.

8. References

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 2010* (Valletta, Malta, 2010).

Asher, N.; Lascarides, A. (2003). Logics of Conversation. *Cambridge University Press.*

Basha, A. Z. (1912).الترقيم وعلاماته في اللغة العربية. (Punctuation and its marks in Arabic Language).

Belguith Hadrich, L. (2009). Analyse et résumé automatiques de documents : Problèmes, conception et réalisation, Habilitation Universitaire en Informatique, soutenue le 2 mai 2009, FSEGS, Université de Sfax, Tunisie.

Belguith Hadrich, L. ; Aloulou, C. and Ben Hamadou A. (2008). «MASPAR : De la segmentation à l'analyse syntaxique de textes arabes», *Information Interaction Intelligence I3, CÉPADUÈS-Editions,* mai 2008, Vol. 7, n° 2, p. 9-36.

Belguith, L.; Baccour, L. and 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.*

Chai, J. Y. and Jin, R. (2004). Discourse structure for context question answering. *In HLT-NAACL Workshop on Pragmatics of Question Answering*.

Da Cunha, I.; SanJuan, E. and Torres M. (2010). Discourse segmentation for Spanish based on shallow parsing. *Proceedings of the 9th Mexican* *international conference on Advances in artificial intelligence, (MICAI'10),* p. 13-23. Springer-Verlag.

Debili, F. ; Achour, H. and Souissi, E. (2002). La langue arabe et l'ordinateur, de l'étiquetage grammatical à la voyellation automatique. *Correspondances n° 71* juillet-août 2002.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harald, L.; Csilla, P.; Maja, B.; Mirco, H. and Henning L. (2006). Discourse segmentation of German written text. In: Tapio Salakoski, Filip Ginter, Sampo Pyysalo, Tapio Pahikkala (eds.): *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006).* Berlin: Springer, 2006.

Jirawan, C.; Thana, S.; and 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, 2005.

Khalifa, I.; Feki, Z. and Farawila, A. (2011). Arabic Discourse Segmentation Based on Rhetorical Methods. International *Journal of Electric and Computer Sciences IJECS-IJENS,* Vol: 11(1).

Maâloul, M. H. ; Ellouze, M. and Belguith Hadrich, L. (2008). Al Lakas s El'eli / اللّخاص الآلي: Un système de résumé automatique de documents arabes. *9th International Business Information Management Conference, IBIMA'08,* Marrakech, Maroc, 4-6 janvier 2008. pp 1260 – 1268.

Mann, W.C. and Thompson, S. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3). p. 243-281.

Marcu, D. (2000). The Theory and Practice of Discourse Parsing and Summarization. *The MIT Press*.

Mesfar, S. (2008). Analysis morpho-syntaxique automatic recognition of named entites in Standard Arabic. PHD thesis, University of Franche-Comté, France.

Seeger, F. and Brian, R. (2007). The utility of parse-derived features for automatic discourse segmentation*. In ACL*.

Silberztein, M. (1993). Dictionnaires électroniques et analyse automatique des textes : Le système INTEX. Masson–Paris.

Touir, A.; Mathkour, H. and Al-Sanea, W. (2008). Semantic-Based Segmentation of Arabic Texts. *Information Technology Journal. Vol: 7(7).*