Journal of King Saud University – Computer and Information Sciences (2014) **26**, 398–416

King Saud University

Journal of King Saud University –
Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Journal of
King Saud University –
Computer and
Information Sciences

Learning explicit and implicit Arabic discourse relations

Iskandar Keskes [a,*], **Farah Benamara Zitoune** [b], **Lamia Hadrich Belguith** [c]

[a] *ANLP Research Group, MIRACL Lab-Sfax University, Tunisia & IRIT-Toulouse University, France*
[b] *IRIT-Toulouse University, France*
[c] *ANLP Research Group, MIRACL Lab-Sfax University, Tunisia*

Available online 7 October 2014

KEYWORDS

Discourse relations;
Segmented Discourse Representation Theory;
Arabic language

Abstract We propose in this paper a supervised learning approach to identify discourse relations in Arabic texts. To our knowledge, this work represents the first attempt to focus on both explicit and implicit relations that link adjacent as well as non adjacent Elementary Discourse Units (EDUs) within the Segmented Discourse Representation Theory (SDRT). We use the Discourse Arabic Treebank corpus (D-ATB) which is composed of newspaper documents extracted from the syntactically annotated Arabic Treebank v3.2 part3 where each document is associated with complete discourse graph according to the cognitive principles of SDRT. Our list of discourse relations is composed of a three-level hierarchy of 24 relations grouped into 4 top-level classes. To automatically learn them, we use state of the art features whose efficiency has been empirically proved. We investigate how each feature contributes to the learning process. We report our experiments on identifying fine-grained discourse relations, mid-level classes and also top-level classes. We compare our approach with three baselines that are based on the most frequent relation, discourse connectives and the features used by Al-Saif and Markert (2011). Our results are very encouraging and outperform all the baselines with an *F*-score of 78.1% and an accuracy of 80.6%.

1. Introduction

Identifying discourse relations is a crucial step in discourse parsing. Given two adjacent or non adjacent discourse units (clauses, sentences, or larger units) that are deemed to be related, this step labels the attachment between the two discourse units with discourse, rhetorical or coherence relations such as Elaboration, Explanation, Cause, Concession, Consequence, Condition, etc. Relations capture the hierarchical structure of a document and ensure its coherence. Their triggering conditions rely on elements of the propositional contents of the clauses – a proposition, a fact, an event, a situation (the so-called abstract objects (Asher, 1993)) – or on the speech acts expressed in one unit and on the semantic content of another unit that performs it. Some instances of these relations are explicitly marked; i.e. they have cues that help identifying them such as *but, although, as a consequence*. Others are implicit; i.e. they do not have clear indicators, as in *I didn't go to the beach. It was raining*. In this last example to infer the intuitive Explana-

[*] Corresponding author.

Peer review under responsibility of King Saud University.

ELSEVIER **Production and hosting by Elsevier**

http://dx.doi.org/10.1016/j.jksuci.2014.06.001
1319-1578 © 2014 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

tion relation between the clauses, we need detailed lexical knowledge and probably domain knowledge as well.

Automatic identification of coherent relations has received a great attention in the literature within different theoretical frameworks (the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), the GraphBank model (Wolf and Gibson, 2005), the Penn Discourse Treebank model (PDTB) (Prasad et al., 2008), and the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). Each work tackles some aspects of the problem:

- Detection of relations within a sentence (Soricut and Marcu, 2003),
- Identification of explicit relations (Hutchinson, 2004; Miltsakaki et al., 2005; Pitler et al., 2008),
- Identification of implicit relations (Marcu and Echihabi, 2002; Blair-Goldensohn et al., 2007; Lin et al., 2009; Pitler et al., 2009; Louis et al., 2010; Zhou et al., 2010; Park and Cardie, 2012; Wang et al., 2011),
- Identification of both explicit and implicit relations (Versley, 2013),
- Building the discourse structure of a document and relation labeling, without making any distinction between implicit and explicit relations. See for example (DuVerle and Prendinger, 2009; Baldridge and Lascarides, 2005; Wellner et al., 2006; Lin et al., 2010) who proposed discourse parsers within respectively the RST, SDRT, Graph Bank and PDTB frameworks.

Several approaches have been proposed to address these tasks, going from supervised, semi-supervised to unsupervised learning techniques. A large set of features was explored, including lexical, syntactic, structural, contextual and linguistically informed features (such as polarity, verb classes, production rules and word pairs). Although most of the research studies have been done for the English language, some efforts focused on relation identification in other languages including French (Muller et al., 2012), Chinese (Huang and Chen, 2011), German (Versley, 2013), and Modern Standard Arabic (MSA) (Al-Saif and Markert, 2011).

Al-Saif and Markert (2011) proposed the first algorithm that identifies explicitly marked relations holding between adjacent Elementary Discourse Units (EDU) within the PDTB model. In this paper, we extend Al-Saif and her colleague's work by focusing on both explicit and implicit relations that link adjacent as well as non-adjacent units within the SDRT, a different theoretical framework. We use the Discourse Arabic Treebank corpus (D-ATB) which is composed of newspaper documents extracted from the syntactically annotated Arabic Treebank v3.2 part3 (Maamouri et al., 2010b). Each document is associated with complete discourse coverage according to the cognitive principles of SDRT. Our list of relations was elaborated after a deep analysis of both previous studies in Arabic rhetoric and earlier work on discourse relations. It is composed of a three-level hierarchy of 24 relations grouped into 4 top-level classes. The gold standard version of our corpus actually contains a total of 4963 EDUs, linked by 3184 relations. 25% of these relations are implicit while 15% link non adjacent EDUs.

In order to automatically learn explicit and implicit Arabic relations, we use state of the art features. Among these features, some have been successfully employed for explicit Arabic relations recognition such as al-masdar, connectives, time and negation (cf. Al-Saif and Markert, 2011). Others however are novel for the Arabic language and include contextual, lexical as well as lexico-semantic features, such as argument position, semantic relations, word polarity, named entities, anaphora and modality. We investigate how each feature contributes to the learning process. We report on our experiments in fine-grained discourse relations' identification as well as in mid-level relations' and top-level class identification. We compare our approach to three baselines that are based on the most frequent relation, discourse connectives and the features used by Al-Saif and Markert (2011). Our results are encouraging and outperform all the baselines.

The next section gives an overview of SDRT, our theoretical framework. Section 3 presents the data. Section 4 describes our list of Arabic discourse relations. Section 5 details the annotation scheme of the D-ATB corpus, the inter-annotator agreements study as well as the characteristics of the gold standard. In Section 6 we give our features. Section 7 describes the experiments and results. Finally in Section 8, we compare our approach to related work.

2. The Segmented Discourse Representation Theory (SDRT)

SDRT is a theory of discourse interpretation that extends Kamp's Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) to represent the rhetorical relations holding between Elementary Discourse Units (EDUs), which are mainly clauses, and also between larger units recursively built up from EDUs and the relations connecting them.

For annotation purposes, we consider a discourse representation for a text T in SDRT to be a discourse structure in which every EDU of T is linked to some (other) discourse unit, where discourse units include EDUs of T and complex discourse units (CDUs) that are built up from EDUs of T connected by discourse relations in recursive fashion. Proper SDRSs form a rooted acyclic graph with two sorts of edges: edges labeled by discourse relations that serve to indicate rhetorical functions of discourse units, and unlabeled edges that show which constituents are elements of larger CDUs. The description of discourse relations in SDRT is based on how they can be recognized and their effect on meaning, i.e. what is their contribution to truth conditions. They are constrained by: semantic content, pragmatic heuristics, world knowledge and intentional knowledge. They are grouped into *coordinating relations* that link arguments of equal importance and *subordinating relations* linking an important argument to a less important one. SDRT allows attachment between non adjacent discourse units and for multiple attachments to a given discourse unit, which means that the discourse structures created are not always trees but rather directed acyclic graphs. This enables SDRT's representations to capture complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups,[1] as well as crossed dependencies[2] (Wolf and Gibson, 2006; Danlos, 2007).

[1] In a document, an author introduces and elaborates on a topic, 'switches' to other topics or reverts back to an older topic. This is known as discourse popping where a change of topic is signaled by the fact that the new information does not attach to the prior EDU, but rather to an earlier one that dominates it (Asher and Lascarides, 2003).

[2] Suppose a sentence is composed of four consecutive units u1, u2, u3, u4. A cross-dependency structure corresponds to the attachments R(u1, u3) and R'(u2, u4).

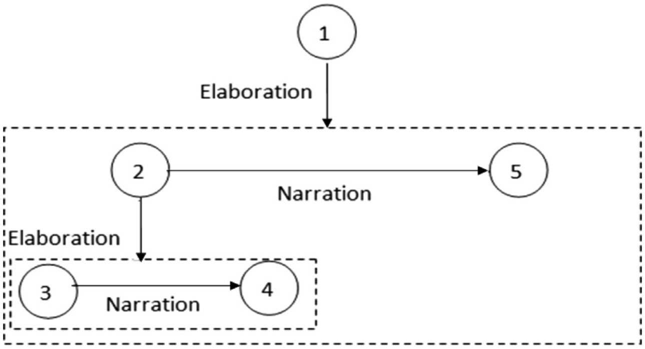

Figure 1 Example of an SDRT-graph.

The SDRT discourse graph is constrained by the right frontier principle that postulates that each new EDU should be attached either to the last discourse unit or to one that is super-ordinate to it via a series of subordinate relations and complex units. Fig. 1 gives an example of the discourse structure of the example (1), familiar from Asher and Lascarides (2003). In this figure, circles are EDUs, rectangles are complex segments, and horizontal links are coordinating relations while vertical links represent subordinating relations.

(1) [John had a great evening last night.]$_1$ [He had a great meal.]$_2$ [He ate salmon.]$_3$ [He devoured lots of cheese.]$_4$ [He then won a dancing competition.]$_5$ To illustrate the importance of SDRT's representation, let us consider the following examples in (2) and (3) taken respectively from the RST Treebank corpus (an English corpus annotated following RST (Carlson et al., 2003) and the Annodis corpus (a French corpus annotated following SDRT (Afantenos et al., 2012), discussed in (Venant et al., 2013):

(2) [In 1988, Kidder eked out a $ 46 million profit,]$_{31}$ [mainly because of severe cost cutting.]$_{32}$ [Its 1,400-member brokerage operation reported an estimated $ 5 million loss last year,]$_{33}$ [although Kidder expects to turn a profit this year]$_{34}$ (RST Treebank, wsj_0604).

(3) [Suzanne Sequin passed away Saturday at the communal hospital of Bar-le-Duc,]$_3$ [where she had been admitted a month ago.]$_4$ [She would be 79 years old today.]$_5$ [...] [Her funeral will be held today at 10h30 at the church of Saint-Etienne of Bar-le-Duc.]$_6$ (Annodis corpus, ER045).

These examples involve what are called long distance attachments. Example (2) involves a relation of Contrast, or Comparison between 31 and 33, but which does not involve the contribution of 32 (the costs cutting of 1988). (3) Displays something comparable. A causal relation like Result or at least a temporal Narration holds between 3 and 6, but it should not scope over 4 and 5 if one does not wish to make Sequin's admission to the hospital a month ago and her turning 79 a consequence of her death last Saturday. It is impossible however, to account for such long distance attachments using the immediate interpretation of RST trees[3] (2). For instance, an Explanation relation between 31 and 32 should not include 33 or 34 in its scope. To handle such difficulties, SDRT adjusts the conception of the discourse structure so that the immediate interpretation is retained.

Two main corpora have been developed following SDRT principles: Discor for English (Reese et al., 2007) and Annodis[4] for French (Afantenos et al., 2012). The Discor corpus analyzed the interaction between document discourse structure and co-reference resolution. This project annotated 60 texts from the MUC 6 and MUC 7 data sets and only experts in the theory did the annotation. The Annodis corpus combined two perspectives on discourse: a bottom-up view that incrementally builds a document discourse structure from EDUs, and a top-down view that focuses on the selective annotation of multi-level discourse structures. The bottom-up approach resulted in the annotation of 86 documents (short Wikipedia articles as well as news articles) with a total of 3199 EDUs and 3355 relations. As far as we know, the Discourse Arabic Treebank corpus (D-ATB) corpus is the first effort toward building recursive and complete discourse structures of Arabic texts (cf. Sections 4 and 5).

3. The data

Arabic Treebank (ATB) v3.2 part3 (Maamouri et al., 2010b) consists of 599 newswire stories from Annahar News Agency. There are a total of 339,710 words/tokens before clitics are split and 402,291 words/tokens after clitics are separated for the Treebank annotation. Each document in this corpus is associated to two annotation levels. First a morphological and parts of speech level and then the syntactic Treebank annotation that characterizes the constituent structures of word sequences and provides categories for each non-terminal node.

We have randomly selected 90 documents from ATB. Our aim was to manually annotate each document with complete discourse coverage according to the cognitive principles of SDRT (cf. Section 2). The annotation of our corpus required three steps: (a) the elaboration of a new hierarchy of discourse relations, (b) the definition of the annotation manual and (c) the manual annotation of our corpus following the annotation guidelines as defined in the manual. The first two steps were performed by three experts in Arabic linguistic while the last step involved two experts in discourse analysis.[5] To achieve these three steps, our corpus was split into three subsets: a development set composed of 13 documents used for defining a novel hierarchy of Arabic discourse relations (cf. Section 4) and annotation training, a set of 7 documents for measuring inter-annotator agreements (cf. Section 5) and finally training and test sets composed of 70 documents for learning Arabic discourse relations (cf. Sections 6 and 7).

In order to avoid errors in determining the basic units (which would make the inter-annotator agreement study tedious), we have discarded discourse segmentation from the annotation campaign. Instead, EDUs are automatically identified and then manually corrected if necessary. The segmentation of our corpus was performed by a multi-class supervised learning

[3] The immediate interpretation of an RST tree R(a,b) is that a and b are respectively the left and the right arguments of R. Given the work on nuclearity, the inferred interpretation of an RST tree is not always the correct interpretation of discourse.

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

[5] Experts involved in manual annotation are not the same experts that have been involved for building the new hierarchy of discourse relations.

Table 1 Characteristics of our data.

	# Documents	# EDUs
Overall corpus	90	6336
Building discourse relations hierarchy + Annotation training	13	911
Inter-annotator agreements study	7	462
Discourse relation learning (training/testing)	70 (gold corpus)	4963

approach using the Stanford classifier that is based on the Maximum Entropy model (Ratnaparkhi, 1997). Each token can belong to one of the three following classes: *Begin*, if the token begins an EDU, *End* if it ends an EDU or *Inside*, if a token is none of the above. Our learning method used a rich lexicon (with more than 174 connectives) and a combination of punctuation, morphological and lexical features. It achieved an accuracy of 0.631 on token boundary recognition. However, this classification does not guarantee that the retrieved EDUs are well-formed (that is, for each begin bracket, there is a corresponding end bracket). To ensure correct bracketing, we performed a post-processing step that consists in adding an end bracket for each opening bracket that has no corresponding end. This step boosted the performance of our system up to 0.130 with an accuracy of 0.769 on EDU recognition. See (Keskes et al., 2014) for a detailed description of our segmentation principles of Arabic texts and for a presentation of our learning method.

Table 1 summarizes the characteristics of our data.

4. Building a new hierarchy of Arabic discourse relations

To our knowledge, the only available resource in Arabic annotated with discourse information is the Leeds Arabic Discourse Treebank[6] (LADTB) that extended PDTB to MSA (Al-Saif and Markert, 2010). This corpus provides a partial discourse structure of a text by focusing on explicit discourse connectives and the annotation of their arguments as well as the discourse relations that link adjacent arguments. PDTB relations are informational and focus on how they are inferred from observable markers in discourse. In addition, no explicit semantic or interpretive effect is given to these relations. In the LADTB, the set of relations is mostly the same as the one used in the English PDTB (Prasad et al., 2008) except that the number of relations was reduced from 33 to 17 (for example, the Contrast subtypes (Opposition vs. Juxtaposition) and Condition (Hypothetical, etc.) were removed) and that two novel relations have been added, namely Background and Similarity.

We propose a semantically driven approach following SDRT, as done in earlier studies on Arabic rhetoric that provided a semantic and a pragmatic analysis of Arabic rhetorical senses (Abdul-Raof, 2012). SDRT focuses on the relation semantic characterization, which allows determining whether two relations are the same, one entails the other, are independent or are incompatible. We follow this approach in the annotation manual to describe a relation independently from its possible discourse markers (too often ambiguous, especially in Arabic), and to focus on what distinguishes relations that are often confused. Compared to (Al-Saif and Markert, 2010), our approach goes beyond the annotation of explicit relations that link adjacent units, by completely specifying the semantic scope of each discourse relation, making transparent an interpretation of the text that takes into account the semantic effects of discourse relations.

Given our semantic-driven approach on discourse, we chose not to reuse the LADTB relations set. Instead, we started from the set of relations already defined within past SDRT-like annotation campaigns and we refined them via a specialization/generalization process using both Arabic rhetoric literature and corpus analysis. This is motivated by general considerations for capturing additional relations and by language-specific considerations for adapting previous relations to take into account Arabic specificities.

We relied on the previous set of 19 relations defined within the Annodis project (Afantenos et al., 2012). In this project, relations were grouped into 7 top-level categories: Causation, Structural, Logic, Reported speech, Exposition/Narration, Elaboration and Commentary. Among these relations, we focused our study on semantic relations that involve entities from the propositional content of the clauses (meta-talk (or pragmatic) relations were discarded). Annodis classification has several top-level classes and some of them contain only one relation (such as Reported Speech and Commentary). To manually annotate our corpus, we wanted to reduce the number of top-level classes and, at the same time, to adapt Annodis relations to the Arabic specificities. Therefore, we decided to build a new classification of Arabic discourse relations by flattening the Annodis hierarchy so as not to influence our experts by the already existing Annodis's top-level classes.

Three experts in Arabic linguistics were involved in this task. We provided them with a precise description of SDRT principles, as well as a definition of the meaning of discourse relations as defined within the Annodis project (henceforth Annodis_set). We have also provided a description of Arabic rhetorical senses as previously defined in earlier studies in Arabic rhetoric (Abubakre, 1989; Al-Jarim and Amine, 1999; Sloane, 2001; Musawi and Muhsin, 2001; Owens, 2006; Abdul-Raof, 2012). We name this set Arabic_set. Then, we asked the experts to collapse these two sets by analyzing how explicit and implicit rhetorical relations are instantiated in our corpus.[7] For each relation R in the Annodis_set, experts look for its corresponding rhetorical senses in the Arabic_set. Five situations may occur:

(1) There is an exact correspondence between the semantic of R and its equivalent in the Arabic_set. Then, the relation R is selected and the experts analyzed how R is signalled in the corpus in order to give a preliminary list of its discourse markers.
(2) There is only a partial correspondence between the semantic of R and its equivalent in the Arabic_set. Then the relation R is selected and the experts further specified its semantic according to the particularities of the Arabic language.

[6] www.arabicdiscourse.net/.

[7] The data used in this step were composed of news paper documents extracted from ATB as well as 25 documents (924 EDUs) extracted from Tunisian Elementary School Textbooks (EST) built by our own.

(3) The semantic of R covers different senses in the Arabic_set and each sense has its own realization in the corpus. Then, R needs to be specialized. New relations are added and the experts were asked to define their semantics along with their corresponding discourse markers.
(4) A group of relations from the Annodis_set corresponds to one sense in the Arabic_set and in addition these relations are often not differentiated in the corpus. In this case, the experts are asked to generalize these relations and to create a new top-level relation.
(5) If there is no correspondence of R in the Arabic_set and no instance of R in the corpus. R is discarded.

This procedure resulted in a new hierarchy of 4 classes: إنشائي/ < n$A}y/Thematic, زمني/zmny/Temporal, بنيوي/bnywy/Structural, and سببي/sbby/Causal with a total of 24 relations, as shown in Fig. 2. In the remainder of this paper, relation names (and examples) are given in Arabic along with their direct English translation (if possible) and their transliteration using Buckwalter 1.1.

5. Annotation campaign

5.1. Annotation guidelines

Two experts in discourse analysis were asked to annotate the D-ATB corpus. We provided them with a precise definition of the meaning of discourse relations (cf. Section 4) and asked them to insert relations between constituents. When appropriate, EDUs can be grouped to form complex discourse units.

The goal of the annotation manual was the development of an intuition for each relation, suitable for the level of the annotators. Occasional examples were provided, and we gave a list of possible markers for each relation but we cautioned that the list was not exhaustive. Indeed, we believe that if the manual mentions all cues for each discourse relations, this will certainly lead to some wrong annotations, especially for ambiguous markers, which are frequent in Arabic. For example, the relation مقابلة/mqAblp/Contrast is often introduced by specific markers in Arabic such as: على العكس/ElY AlEks/however, في المقابل/fy AlmqAbl/however, وعلى عكس ذلك/wElY Eks*lk/unlike, على النقيض/ElY AlnqyD/unlike ..., as in (4). Similarly, main markers of the relation شرط/$rT/Conditional include: س/s/so, لو/lw/if, إذا/ < *A/if, لولا/lwlA/except, متى/mtY/when, مهما/mhmA/whatever, كلّما/kl~mA/whenever, فإنّ/f < n~/so, فقد/fqd/so, ف/f/then ..., as in (5).

(4) [يضحك أخي] 1 [وفي المقابل تبكي أختي.] 2

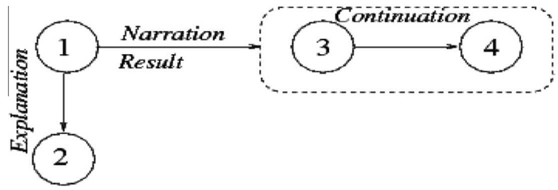

Figure 3 Right frontier principle. In this example, open attachment sites are the unit 4 and the CDU [3,4].

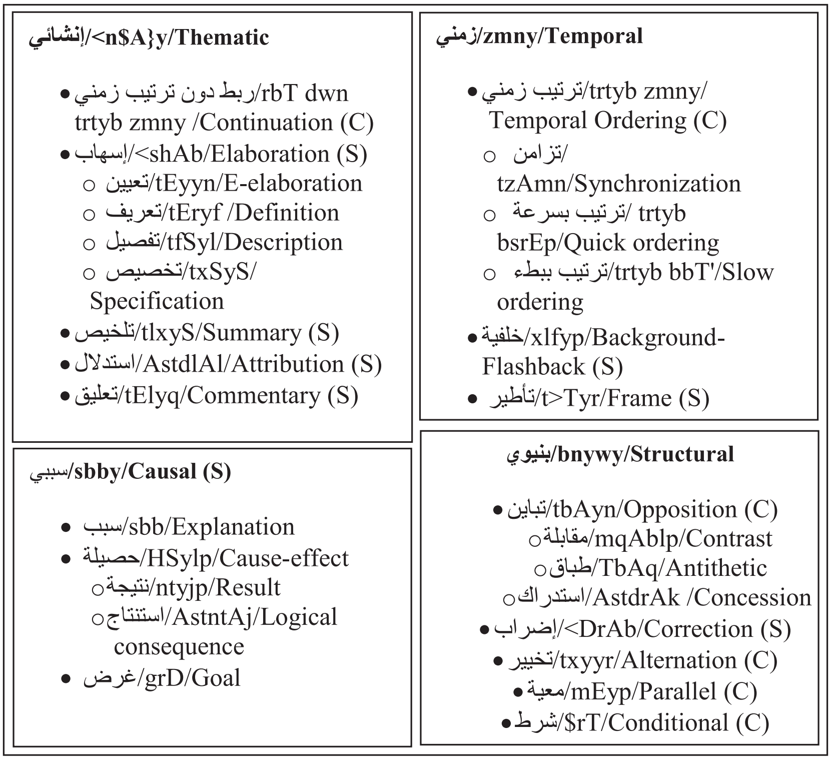
إنشائي/<n$A}y/Thematic
• ربط دون ترتيب زمني/rbT dwn trtyb zmny /Continuation (C)
• إسهاب/<shAb/Elaboration (S)
○ تعيين/tEyyn/E-elaboration
○ تعريف/tEryf /Definition
○ تفصيل/tfSyl/Description
○ تخصيص/txSyS/ Specification
• تلخيص/tlxyS/Summary (S)
• استدلال/AstdlAl/Attribution (S)
• تعليق/tElyq/Commentary (S)
زمني/zmny/Temporal
• ترتيب زمني/trtyb zmny/ Temporal Ordering (C)
○ تزامن/ tzAmn/Synchronization
○ ترتيب بسرعة/ trtyb bsrEp/Quick ordering
○ ترتيب ببطء/trtyb bbT'/Slow ordering
• خلفية/xlfyp/Background-Flashback (S)
• تأطير/t>Tyr/Frame (S)
سببي/sbby/Causal (S)
• سبب/sbb/Explanation
• حصيلة/HSylp/Cause-effect
○ نتيجة/ntyjp/Result
○ استنتاج/AstntAj/Logical consequence
• غرض/grD/Goal
بنيوي/bnywy/Structural
• تباين/tbAyn/Opposition (C)
○ مقابلة/mqAblp/Contrast
○ طباق/TbAq/Antithetic
○ استدراك/AstdrAk /Concession
• إضراب/<DrAb/Correction (S)
• تخيير/txyyr/Alternation (C)
• معية/mEyp/Parallel (C)
• شرط/$rT/Conditional (C)

Figure 2 Hierarchy of Arabic discourse relations used in the D-ATB corpus. (S) and (C) correspond respectively to subordinating and coordination relations.

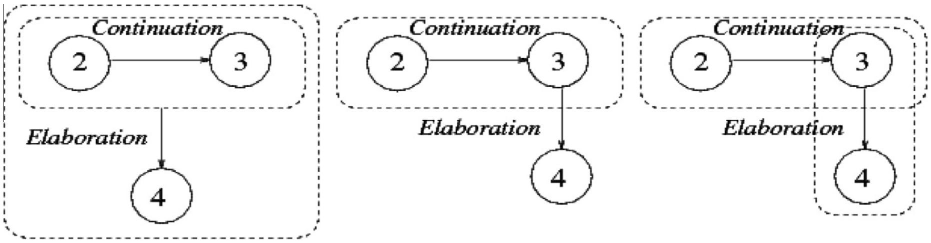

Figure 4 An example of a CDU constraint. Figures in the right and in the middle are correct configurations whereas the one in the left is not allowed because CDUs cannot overlap.

[yDHk >xy]$_1$ [w fy AlmqAbl tbky >xty.]$_2$
[My brother laughs]$_1$ [however my sister cries.]$_2$
مقابلة/mqAblp/Contrast (1,2)

(5) [إذا أصلحت السيارة]$_1$ [و قمت بدهنها،]$_2$
[سأستطيع بيعها]$_3$

[<*A>SlHt AlsyArp]$_1$ [w qmt bdhnhA,]$_2$ [s>stTyE byEhA]$_3$
[If you repair the car]$_1$ [and you paint it,]$_2$ [I can sell it]$_3$
شرط/$rT/Conditional ([1,2],3)[8]
معية/mEyp/Parallel (1,2).

Our annotation manual clearly details the constraints that annotators should respect according to the structural principles of SDRT. This is a first step before moving to non-expert annotation in order to build a discourse bank that studies how well SDRT predicts the intuition of subjects, regardless of their knowledge of discourse theories. Main SDRT constraints concern: unit attachment (no isolated unit in the graph, attachment mainly follows the reading order of the document), right frontier principle (Asher and Lascarides, 2003) (cf. Fig. 3), structural constraints including accessibility, complex units, no cycles, etc (cf. Fig. 4).

5.2. Inter-annotator agreements study

The annotation campaign was as follows. First, we trained our annotators using 13 documents (911 EDUs). During the training, annotators were encouraged to discuss their annotations and to give their feedbacks on the annotation manual. More precisely, we noticed that the document length was a handicap since the document annotation can take two days making the task of connecting all the EDUs in the same whole discourse structure very tedious (each document has around 26 sentences and 8 paragraphs). To overcome this problem, we decided to separately annotate the discourse structure of each paragraph in a document, and then to link these sub-structures with the mid-level relation إسهاب/<shAb/Elaboration in order to guarantee the connectivity of the resulting graph.[9] After the training, annotators were asked to doubly annotate the same 7 documents (462 EDUs) in order to compute the inter-annotator agreements. Finally, we asked the annotators to build the gold standard corpus by consensus (70 documents), by discussing the main cases of disagreement.

Discourse annotation depends on two decisions: a decision about where to attach a given EDU, and a decision on how to label the attachment link via discourse relations. Two inter-annotator agreements have thus to be computed and the second one depends on the first because agreements on relations can be performed only on common links. We relied on the algorithm developed within the Annodis project (Afantenos et al., 2012) to compute both attachment and labeling agreements. The algorithm used for agreements attachment assumes that attaching is a yes/no decision on every EDUs pair, and that all decisions are independent, which of course underestimates the results (see in (Afantenos et al., 2012) for an interesting discussion on the difficulty on how to match/compare rhetorical structures, especially when CDUs have to be taken into account). For attachment, we obtained an *F*-score of 0.890. When commonly attached pairs are considered, we got a Cohen's kappa of 0.750 for the full set of 24 relations. Overall, our results are higher compared to those obtained by Annodis (0.660 F-measure for attachment and a Cohen's Kappa of 0.400 for relation labeling) mainly for two reasons. First, our annotation manual was more constrained since we provided annotators a detailed description of how to build the document discourse structure. Second, we did not focus on the document overall discourse structure but on the paragraph discourse structure which implies shorter distance attachments (an average of 20 EDUs per paragraph in our case vs. an average of 55 EDUs in Annodis). The main disagreement came from non-adjacent EDUs. Indeed, one annotator tended to form CDUs more frequently while the other often produced "flat" structures. For example in (6), the annotation Frame(1,2) and Continuation(2,3) is equivalent to Frame(1,[2,3]) and Continuation(2,3).

(6) [في نظام التعليم الجامعي أ.م.د.،]$_1$
[يدرس الطالب ثلاث سنوات إجازة،]$_2$
[ويدرس سنتين ماجستير.]$_3$

[fy nZAm AltElym AljAmEy >.m.d.,]$_1$ [ydrs AlTAlb vlAv snwAt <jAzp,]$_2$ [wydrs sntyn mAjstyr.]$_3$
[In the university education system L.M.D.,]$_1$ [the student studies three years Bachelor's degree,]$_2$ [then two years Master's degree.]$_3$

[8] The notation [a,b] indicates that a and b are a complex discourse unit.

[9] In news document, paragraphs are about the same main topic. We have then considered that a document is composed of a top node (the main topic of the document) and that each paragraph is a complex discourse unit that elaborates on that topic. Elaboration here refers to a group of discourse relations that connect utterances describing the same state of affairs: reformulation (restatement), specification (particularization), generalization, etc.

We present below an example of an annotated paragraph taken from the document ANN20020115.0003.

(7) [قصفت طائرات أميركية مجمعات كهوف في شرق
أفغانستان،]$_1$ [ضمن الحملة]$_2$ [التي تشنها على مقاتلي تنظيم
"القاعدة" وحركة "طالبان" الإسلامية،]$_3$ [في الوقت الذي تركز
الحكومة الأفغانية المؤقتة على قضايا سياسية مثل تعزيز
الأمن وإمدادات الإغاثة]$_4$ [لإعمار البلاد]5 [التي مزقتها
الحرب.]$_6$ [وأفادت "وكالة الأنباء الإسلامية" الأفغانية]$_7$
[التي تتخذ إسلامآباد مقرا لها]$_8$ [انه تم قصف دون توقف لأحد
غارت الطائرات الأميركية على منطقة جوار على مسافة 30
كيلومترا جنوب غرب خوست.]$_9$ [وقالت:]$_{10}$
["لم يهدأ القصف طوال الساعات الـ 48 الاخيرة".]$_{11}$

[qSft TA}rAt >myrkyp mjmEAt khwf fy $rq >fgAnstAn]$_1$ [Dmn AlHmlp]$_2$ [Alty t$nhA ElY mqAtly tnZym “AlqAEdp” wHrkp “TAlbAn” Al<slAmyp,]$_3$ [fy Alwqt Al*y trkz AlHkwmp Al>fgAnyp Alm&qtp ElY qDAyA syAsyp mvl tEzyz Al>mn w<mdAdAt Al<gAvp]$_4$ [l<EmAr AlblAd]$_5$ [Alty mzqthA AlHrb.]$_6$ [w>fAdt “wkAlp Al>nbA’ Al<slAmyp” Al>fgAnyp]$_7$ [Alty ttx* <slAm |bAd mqrA lhA]$_8$ [Anh tm qSf dwn twqf l>Hd gArt AlTA}rAt Al>myrkyp ElY mnTqp jwAr ElY msAfp 30 kylwmtrA jnwb grb xwst.]$_9$ [wqAlt:]$_{10}$ [“lm yhd> AlqSf TwAl AlsAEAt Al 48 AlAxyrp”.]$_{11}$

[American planes bombed some caves in Eastern Afghanistan,]$_1$ [within the campaign]$_2$ [that aimed at killing “Al Qaida” and “Taliban” fighters,]$_3$ [meanwhile the Afghan Interim Government focused on political issues such as strengthening security and relief supplies]$_4$ [in order to rebuild the country]$_5$ [that was destroyed by the war.]$_6$ [The “Afghan Islamic News Agency” [which is located in Islamabad]$_7$ reported]$_8$ [that American planes have made a non-stop bombing on an area situated 30 km Southwest of Khost.]$_9$ [And it said:]$_{10}$ [”the bombing lasted 48 h.”]$_{11}$ (see Fig. 5).

5.3. The gold standard

Given the good inter-annotator agreements results, annotators asked to build the gold standard by consensus by discussing main cases of disagreements. Table 2 summarizes the characteristics of our gold standard.

The total number of annotated discourse relations is 3 184. The distribution of these relations is presented in Table 4. In these statistics, the relation إسهاب/<shAb/Elaboration used to link paragraphs is not counted. Our gold corpus contains more than 58% of إنشائي/<n$A}y/Thematic relations. The most frequent relation is ربط دون ترتيب زمني/rbT dwn trtyb zmny/Continuation(21.14%). On the other hand, infrequent relations (less than 1%) are: تخيير/txyyr/Alternation, استنتاج/AstntAj/Logical consequence, تلخيص/tlxyS/Summary, مقابلة/mqAblp/Contrast and طباق/TbAq/Antithetic.

Table 3 shows additional statistics. Our gold corpus contains 9% of CDUs. We observe that CDUs are more present as a second argument of a relation. Also, among the relations that link EDUs, 15% concern non-adjacent units. The زمني/zmny/Temporal class and the سببي/sbby/Causal class tend to be more local (more than 90%) whereas the بنيوي/bnywy/Structural class and the إنشائي/<n$A}y/Thematic class are more structural. Among the 3184 relations, more than 25% (802) are implicit, i.e. signaled by any connectors. For example, the relations طباق/TbAq/Antithetic, خلفية/xlfyp/Background-Flashback and تفصيل/tfSyl/Description are often implicit, as in (8).

(8)[كان الفلم مسلي جدا.]$_1$ [ضحك أخي]$_2$ [ورفه عن نفسه .]$_3$

[kAn Alflm msly jdA.]$_1$[DHk >xy]$_2$ [wrfh En nfsh.]$_3$
[It was a very exciting movie.]$_1$ [My brother laughed]$_2$ [and had a good time.]$_3$

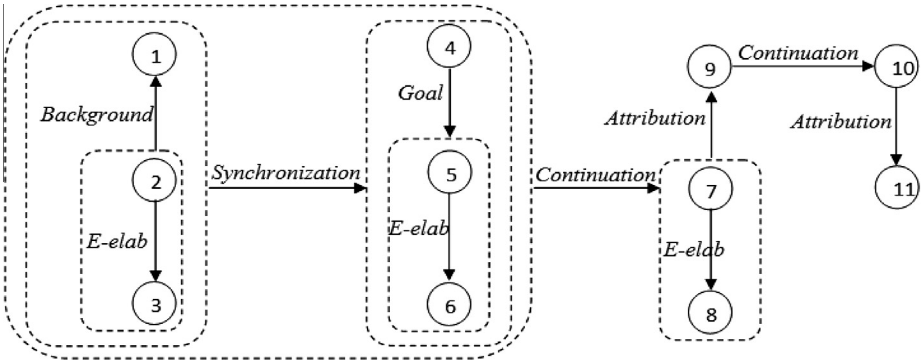

Figure 5 Discourse annotation of example (7).

Table 2 Gold corpus characteristics.

	Texts	Size	Sentences	EDUs	Embedded EDUs	Words + punctuations
D-ATB	70	381ko	1832	4963	542 (9.16%)	39,746

Table 3 Discourse relations distribution of the gold standard.

Total number of relations	3184
Argument type	
EDU	5798 (91%)
CDU	570 (9%)
Discourse relation and EDU position	
Relations between adjacent EDUs	2706 (85%)
Relations between non adjacent EDUs	478 (15%)
Discourse relation and Argument type	
R (EDU,EDU)	2682 (84.23%)
R (EDU,CDU)	322 (10.11%)
R (CDU,EDU)	112 (3.52%)
R (CDU,CDU)	68 (2.14%)
Discourse relation and Signaling type	
Explicit relations	2382 (74.8%)
Implicit relations	802 (25.2%)

تفصيل/tfSyl/Description (1,[2,3])
ربط دون ترتيب زمني/rbT dwn trtyb zmny/Continuation(2,3)

From Table 3, we also observe that explicit relations are the majority (75%). This concerns relations such as تخيير/txyyr/Alternation, استنتاج/AstntAj/Logical consequence, تلخيص/tlxyS/Summary, and تعيين/tEyyn/E-Elaboration. Explicit relations can be signaled by strong discourse markers that are non ambiguous and generally indicate the same relation. For example, the marker بل/bl/however triggers the relation إضراب/< DrAb/Correction, the marker لكن/lkn/but triggers the relation استدراك/AstdrAk/Concession, and the marker لذلك/l*lk/so triggers the relation غرض/grD/Goal. On the other hand, explicit relations can also be triggered by weak discourse markers that are highly ambiguous and can signal more than one discourse relation or no relation at all. The most frequent weak markers are the clitics و/w/and, ل/l/for-to, and ف/f/so-then. For example, the discourse marker ل/l/for-to can indicate three relations: سبب/sbb/Explanation, نتيجة/ntyjp/Result,

Table 4 Discourse relations frequency in the gold standard.

Discourse relations		Frequency	Percentage (%)
إنشائي/< n$A}y/Thematic	ربط دون ترتيب زمني/rbT dwn trtyb zmny/Continuation	673	21.14
	إسهاب/< shAb/Elaboration	727	22.83
	تعيين/tEyyn/E-Elaboration	482	15.14
	تعريف/tEryf/Definition	50	1.57
	تفصيل/tfSyl/Description	147	4.62
	تخصيص/txSyS/Specification	48	1.51
	تلخيص/tlxyS/Summary	14	0.44
	استدلال/AstdlAl/Attribution	412	12.94
	تعليق/tElyq/Commentary	44	1.38
	Total	1870	58.74
زمني/zmny/Temporal	ترتيب زمني/trtyb zmny/Temporal Ordering	195	6.12
	تزامن/tzAmn/Synchronization	82	5.58
	ترتيب بسرعة/trtyb bsrEp/Quick ordering	52	1.63
	ترتيب ببطء/trtyb bbT’/Slow ordering	61	1.92
	خلفية/xlfyp/Background-Flashback	124	3.90
	تأطير/t > Tyr/Frame	44	1.38
	Total	363	11.40
سببي/sbby/Causal	سبب/sbb/Explanation	111	3.49
	حصيلة/HSylp/Cause-effect	158	4.96
	نتيجة/ntyjp/Result	143	4.50
	استنتاج/AstntAj/Logical consequence	15	0.47
	غرض/grD/Goal	289	9.08
	Total	558	17.53
بنيوي/bnywy/Structural	تباين/tbAyn/Opposition	128	4.02
	مقابلة/mqAblp/Contrast	27	0.85
	طباق/TbAq/Antithetic	12	0.38
	استدراك/AstdrAk/Concession	89	2.80
	إضراب/< DrAb/Correction	44	1.38
	تخيير/txyyr/Alternation	17	0.53
	معية/mEyp/Parallel	93	2.92
	شرط/$rT/Conditional	111	3.49
	Total	393	12.35

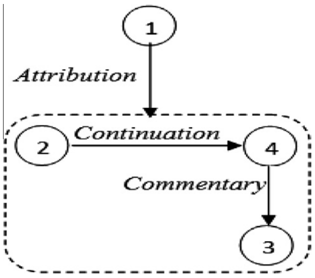

Figure 6 Discourse annotations of (9).

and غرض/grD/Goal. Similarly, the marker ف/f/so-then can indicate the relations نتيجة/ntyjp/Result, ترتيب بسرعة/trtyb bsrEp/Quick ordering, ربط دون ترتيب زمني/rbT dwn trtyb zm-ny/Continuation, and شرط/$rT/Conditional.

6. Features

Building a document discourse structure requires three subtasks: (1) identifying discourse units, (2) "attaching" units to one another, and (3) labeling their link with a coherence relation. In this paper, we focus on the third task. Our instances are thus composed of linked EDUs only.

To perform a supervised learning on the gold standard, we construct a feature vector for each linked couple R(a,b) where R is a discourse relation that links the units a and b (a and b are also called the arguments of R). If a and/or b are complex units, we replace a (resp. b) by its head. Example (9) and its corresponding discourse structure shown in Fig. 6 illustrate this. In this case, we create three vectors that correspond to the relations استدلال/AstdlAl/Attribution(1,2), ربط دون ترتيب زمني/rbT dwn trtyb zmny/Continuation(2,4), and تعليق/tElyq/Commentary(4,3). Finally, in case of multiple relations (i.e. a couple (a,b) linked by different relations), we built as many instances as the number of relations.

(9) [قال وزير الدفاع]$_1$ [إن نحو ستة جنود أميركيين وصلوا
إلى البلاد]$_2$[وان الجنود، [حيث سيكونون مسلحين،]$_3$
يستطيعون الدفاع عن أنفسهم.]$_4$

[wqAl wzyr AldfAE]$_1$ [An nHw stp jnwd Amyrkyyn wSlwA AlY AlblAd]$_2$ [wAn Aljnwd, [Hyv sykwnwn mslHyn,]$_3$ ystTyEwn AldfAE En Anfshm.]$_4$
[The Minister of Defence said]$_1$ [that six U.S. soldiers arrived in the country]$_2$ [and once the soldiers are armed,]$_3$ [they will be able to defend themselves.]$_4$

We designed thirteen groups of features. The first five contain 5 groups (connectives, arguments, al-masdar, tense and negation, length and distance) following (Al-Saif and Markert, 2011).[10] However, compared to (Al-Saif and Markert, 2011), our features are obtained automatically and are not based on the manual annotations of ATB. The 8 remaining features are composed of punctuation, contextual, lexical and lexico-semantic features that have been used in prior work and whose efficiency for detecting both explicit and implicit relations has been empirically determined. They are however new for the Arabic language. Punctuation features were inspired by (Huang and Chen, 2011) and (DuVerle and Prendinger, 2009). Contextual features include textual organization (DuVerle and Prendinger, 2009) (Muller et al., 2012). Lexico-semantic features group polarity and modality (Pitler et al., 2009), named entity (Huang and Chen, 2011), anaphora (Louis et al., 2010) and semantic relations (Subba et al., 2009). Finally, lexical features concern lexical cues with a rich discourse connectives lexicon (Marcu, 2000). Again, all these features do not rely on manual annotations. We use the Standard Arabic Morphological Analyzer SAMA version 3.1 (Maamouri et al., 2010a) for morphological analysis, the Stanford parser (Green and Manning, 2010) for syntactic analysis and various linguistic resources for lexico-semantic features.

We first give all the features already used by Al Saif et al. (namely (F1) to (F5)). Then, we detail our new set of features (namely (F6) to (F13)).

6.1. Al-Saif et al.'s features

(F1) Connectives. We have 6 string features that encode the connective string, the connective lemma, POS of the connective, the position of the connective (begin, middle or end of a unit), the connective type (clitic as ل/l/for-to, simple as لكن/lkn/but, or composed of more than one word as من أجل أن/mn > jl > n/in-order-to), and the syntactic path from the sentence parent to the connective. For example, in (10), the syntactic path of the marker أن/>n/that is the string "(S (NP-TPC-2 (NOUN_PROP)) (VP (PV + PVSUFF_SUBJ:3FS) (NP-SBJ-2 (PP (PREP) (NP (NOUN_PROP))) (SBAR (SUB_CONJ)))".

(10) [نيودلهي أكدت لزو]$_1$ [أن العلاقات مع بيجينغ لن تتأثر
بالتعاون بين بيجينغ وإسلام أباد]$_2$

[nywdlhy >kdt lzw]$_1$ [>n AlElAqAt mE byjyng ln tt>vr byjyng bAltEAwn byn w<slAm >bAd]$_2$
[New Delhi confirmed to Zoos]$_1$ [that relationship with Beijing will not be affected by the cooperation between Beijing and Islamabad]$_2$

(F2) Arguments. We have 7 string features. We encode the surface strings and the POS of the first three words for each argument (that is a total of 6 features) as well as the syntactic category of the argument parent. If the argument is represented by a non-complete tree (as given by the Stanford outputs), we extract the category of the parent shared by the first and the last word in the argument.

(F3) Al-masdar. This is a binary feature that indicates whether the first or the second word of each argument contains al-masdar construction. Al-masdar is a verbal noun construction, frequent in Arabic that names the action denoted by its corresponding verbs. It is a noun category that expresses events without tense. This construction generally signals discourse relations. For example, al-masdar بحثا/bHvA/looking in example (11) explains why Ahmed went to the library.

(11) [اتجه أحمد إلى المكتبة]$_1$ [بحثا عن كتاب الرياضيات.]$_2$

[Atjh >Hmd <lY Almktbp]$_1$ [bHvA En ktAb AlryADyAt.]$_2$

[10] We do not use production rule features since they did not improve Arabic explicit relation recognition in the LADTB corpus (cf. (Al-Saif and Markert, 2011)).

[Ahmed went to the library]$_1$ [to look for the mathematics book.]$_2$
سبب/sbb/Explanation (1,2)

Al-masdar is built from the morphological analyzer Al-Khalil (Boudlal et al., 2011) using well-defined morphological patterns composed of 3 or 4 letter-roots. The patterns can attach suffixes to the root and insert consonant/vowel letters or diacritics into the root. More than 60 morphological patterns can be used to generate al-masdar nouns.

(F4) Tense and negation. We use a string feature to encode the tense assigned to each argument (perfect, imperfect, future or none) and a binary feature to test the presence of negation words in each argument. To detect negation, we rely on a manually built lexicon of 10 Arabic negation words, such as لا/lA/ no and لم/lm/not.

Tense features can help identifying relations from the زمني/ zmny/Temporal class, such as the relations تزامن/tzAmn/Synchronization, and ترتيب ببطء/trtyb bbT'/Slow ordering. Indeed, تزامن/tzAmn/Synchronization holds when the events e1 and e2, introduced in the two units, occur at the same time and when both events are triggered by different subjects (cf. example (12)). On the other hand, ترتيب ببطء/trtyb bbT'/Slow ordering holds when there is a temporal gap between the events denoted by the verbs in the arguments (cf. example (13)). Finally, negation feature can help identifying relations from the بنيوي/bnywy/Structural class, such as the relation إضراب/ < DrAb/Correction where the first or the second argument usually contains a negation.

(12) [كنا نرسم على الحائط،] $_1$ [حينها دخل المعلم.] $_2$

[knA nrsm ElY AlHA}T,]$_1$ [HynhA dxl AlmElm.]$_2$
[We were painting on the wall,]$_1$ [when the teacher arrived]$_2$

(13) [أكمل المعلم الدرس] $_1$ [ثم خرج جميع التلاميذ من القسم]$_2$

[> kml AlmElm Aldrs]$_1$ [vm xrj jmyE AltlAmy*mn Alqsm]$_2$
[The teacher had finished the lesson,]$_1$ [then all the students left the classroom]$_2$

(F5) Length and distance. We have four features. Two have integer values that encode the number of words in each argument and the number of EDUs between the two arguments. One binary feature to deal with the tree distance between the connective and the arguments (0 if the connective and the argument are in the same tree and 1 otherwise). Finally one binary feature to check if both arguments are in the same sentence.

6.2. New features

(F6) Textual organization. We use a string feature to indicate the position of each argument within the document (begin, middle or end of a paragraph[11]) which can be helpful for identifying relations like خلفية/xlfyp/Background-Flashback and تأطير/t > Tyr/Frame (cf. (14)) where the first argument often occur at the beginning of paragraphs. This feature can also help detecting relations such as استنتاج/AstntAj/Logical consequence and تلخيص/tlxyS/Summary (cf. (15)) where the second argument usually appears at the end of paragraphs.

(14) [في نظام التعليم الجامعي أ.م.د.،] $_1$ [يدرس الطالب
ثلاث سنوات إجازة،]$_2$[ثم يدرس سنتين ماجستير،]$_3$
[ثم يدرس ثلاث سنوات دكتوراه.] $_4$

[fy nZAm AltElym AljAmEy > .m.d.,]$_1$ [ydrs AlTAlb vlAv snwAt < jAzp,]$_2$ [vm ydrs sntyn mAjstyr,]$_3$ [vm ydrs vlAv snwAt dktwrAh.]$_4$
[In the L.M.D. courses,]$_1$ [the student studies a three years Bachelor's degree,]$_2$ [two years Master's degree,]$_3$ [then three years Doctorate.]$_4$
تأطير/t > Tyr/Frame (1,[2,3,4])
ترتيب ببطء/trtyb bbT'/Slow ordering (2,3)
ترتيب ببطء/trtyb bbT'/Slow ordering (3,4)

(15) [كان يحدثنا عن مغامراته.] $_1$ [...] $_x$
[وخلاصة القول، كانت جميع مغامراته شيقة.] $_{1+x}$

[kAn yHdvnA En mgAmrAth.]$_1$ [. . .]$_x$ [wxlASp Alqwl, kAnt jmyE mgAmrAth mglqp.]$_{x+1}$
[He told us about his adventures.]$_1$ [. . .]$_x$ [In sum, all his adventures were exciting.]$_{x+1}$
تفصيل/tfSyl/Description (1,x)
تلخيص/tlxyS/Summary (x + 1,[1,. . .,x])

(F7) Punctuation. They can be a good indicator for signaling some discourse relations, such as تفصيل/tfSyl/Description and استدلال/AstdlAl/Attribution (cf. (16)). For each unit, we use 12 features that test for the presence of specific punctuations (!, ?, ., comma,:) as well as of typographical markers ("", (), [], {}, _, -). We use integer values that can vary from 1 to 5 if the unit contains specific features, from 6 to 11 if the unit contains typographical markers, and 0 if the unit does not contain any specific punctuations or typographical markers.

(16) [قال أحمد:] $_1$ ["إن المباراة كانت صعبة"] $_2$

[qAl > Hmd:]$_1$[« < n AlmbArAp kAnt SEbp»]$_2$
[Ahmed said:]$_1$ ["the match was difficult"]$_2$
استدلال/AstdlAl/Attribution (1,2)

(F8) Embedded argument. We use a binary feature to test if the left or the right argument of a relation is an embedded unit. This can help identifying some relations such as تعليق/tElyq/ Commentary and تعيين/tEyyn/E-elaboration (cf. (17)).

(17) [قامت قوات الجيش، [التي اقتحمت المنزل،]$_2$
باعتقال جميع الإفراد] $_1$

[qAmt qwAt Aljy$, [Alty AqtHmt Almnzl,]$_2$ bAEtqAl jmyE AlAfrAd]$_1$
[The army troops, [that stormed the house,]$_2$ arrested all its members]$_1$
تعيين/tEyyn/E-elaboration(1,2)

(F9) Named entities and anaphora. We use two binary features to check the presence of named entities and anaphora. Named entities, pronouns and anaphora are important information for discourse relation recognition. For example, the presence of named entities in the right argument and anaphora in the left argument can help identify the relation تفصيل/tfSyl/ Description (cf. (18)). Moreover, the presence of pronouns and

[11] We relied on carriage return line feed to measure if a given unit is at the beginning, the end or the middle of a paragraph.

Table 5 Examples of concepts related by AWN relations and some discourse relations that they can trigger.

AWN semantic relations	Discourse relations
Near_antonym (ضحك/DHk/laugh, بكى/bkY/cries)	[يضحك أخي]$_1$ [وفي المقابل تبكي أختي.]$_2$ [yDHk > xy]$_1$[w fy AlmqAbl tbky > xty.]$_2$ [My brother laughs]$_1$ [however my sister cries.]$_2$ مقابلة/mqAblp/Contrast (1,2)
Has_holo_part(فريق/fryq/team, لاعب/lAEb/player)	[تألق الفريق التونسي في هذه المباراة،]$_1$ [وبالأخص لاعب الهجوم.]$_2$ [t > lq Alfryq Altwnsy fy h*h AlmbArAp,]$_1$ [wbAl > xS lAEb Alhjwm.]$_2$ [The Tunisian team has shined in this match,]$_1$ [especially the attacker.]$_2$ تخصيص/txSyS/Specification (1,2)
Related_to(جنود/ljnwd/soldiers, مسلح/mslH/military)	[وان الجنود، [حيث سيكونون مسلحين،]$_2$ يستطيعون الدفاع عن انفسهم.]$_1$ [wAn Aljnwd, [Hyv sykwnwn mslHyn,]$_1$ ystTyEwn AldfAE En Anfshm.]$_2$ [and once the soldiers are armed,]$_1$ [they will be able to defend themselves.]$_2$ تعليق/tElyq/Commentary(1,2)
Has_derived (كتاب/ktAb/book, مكتبة/mktbp/library)	[اتجه أحمد إلى المكتبة]$_1$ [بحثا عن كتاب الرياضيات.]$_2$ [Atjh > Hmd < lY Almktbp]$_1$ [bHvA En ktAb AlryADyAt.]$_2$ [Ahmed went to the library]$_1$ [to look for the mathematics book.]$_2$ سبب/sbb/Explanation (1,2)

anaphora in the same argument can help identify the relation معية/mEyp/Parallel (cf. (19)).

(18) [أكل أحمد المربى بشراهة]$_1$ [كأنه لم يذقه قط.]$_2$

[> kl > Hmd AlmrbY b$rAhp]$_1$ [k > nh lm y*qh qT.]$_2$
[Ahmed ate jam greedily]$_1$ [as if he had never tasted it before.]$_2$
تفصيل/tfSyl/Description(1,2)

(19) [نحن موافقون على هذا الحل،]$_1$
[وانتم موافقين أيضا على تطبيقه.]$_2$

[nHn mwAfqwn ElY h*A AlHl,]$_1$ [wAntm mwAfqyn > yDA ElY tTbyqh.]$_2$
[We agree with this solution,]$_1$[and you also agree to implement it.]$_2$
معية/mEyp/Parallel (1,2)

To detect if the arguments contain Arabic named entities, we use the ANERGazet Gazetteers (Benajiba et al., 2007) that contains a collection of 3 Gazetteers: locations (2181 entries), people (2 309 entries) and organizations (403 entries). To test for the presence of anaphora, we manually built a lexicon of 60 Arabic most frequent pronouns and anaphora, such as نحن/nHn/we, انتم/Antm/you, and ه/h/he-it.

(F10) Modality. This binary feature checks the presence of modality in each argument using a manually constructed lexicon composed of 50 Arabic modal words, like أكد/Akd/confirm يرى/yrY/see, يعتقد/yEtqd/think, أوضح/ < AwDH/explain, and لاحظ/lAHZ/remark. Modality can help detect relations like استدلال/AstdlAl/Attribution (cf. example (20)).

(20) [أكد السيد احمد]$_1$
[إن الفريق نزل إلى دوري الدرجة الثانية.]$_2$

[Akd Alsyd AHmd]$_1$ [An Alfryq nzl AlY dwry Aldrjp AlvAnyp.]$_2$
[Mr Ahmed confirms]$_1$ [that the team was relegated to the second division.]$_2$

(F11) Semantic relations. We use Arabic WordNet (AWN), which is one of the best known lexical resources for Modern Standard Arabic (Black et al., 2006). Although its development is based on Princeton's WordNet, it suffers from some weaknesses such as the lack of concepts and some semantic relations between synsets. In our case, we use an enriched version of AWN where semantic relations have been added using a linguistic method based on a set of 135 morpho-lexical patterns (Boudabous et al., 2013). AWN contains about 15,000 entries and 17 semantic relations, like Has_hyponym, Has_instance, Related_to, Near_synonym, Near_antonym, and Has_derived. We build 17 Boolean features, one for each AWN semantic relation R. Each feature tests if there is a concept C1 in the first unit and a concept C2 in the second one, such that R(C1,C2) or R(C2,C1). Table 5 gives some examples of concepts related by AWN relations as well as their corresponding discourse relations. In our corpus, the most frequent semantic relation was Has_hyponym (with 891 instances). The semantic relation Usage_term was absent from our corpus.

(F12) Polarity. To deal with polarity information, we use the translated MPQA subjectivity lexicon (Elarnaoty et al., 2012) that contains more than 8000 English words and their corresponding Arabic translations.[12] Each entry is characterized according to its subjectivity and polarity. Subjectivity can be of two types: strong for terms that are intrinsically subjective such as ابتسامة/AbtsAmp/grin and احترام/AHtrAm/respect and weak for terms that can have an objective or a subjective sense depending on the context, like الأحكام/Al > H-kAm/judgments. Polarity can be of 4 types: positive, negative, both, and neutral.

We associate to each argument two string features: one for subjectivity that checks for the presence of strong or weak opinion words and one that encodes the polarity of that word.

(F13) Lexical cues. We use a rich lexicon of discourse connectives, manually built during the annotation campaign train-

[12] This resource is available through the ALTEC Society at the following address: http://altec-center.org/.

ing session (i.e. 20 documents, 1400 EDUs). It contains 174 entries. For each connective, we specify:

- Its type (discourse cures or indicators). Discourse cues are connectives that have a discursive function such as حيث/Hyv/where, بينما/bynmA/while, and عندئذ/End}*/then. Indicators can be non-inflectional 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, and فقط/fqT/only), conjunctions (e.g. حالما/HAlmA/the-moment-that and طالما/TAlmA/so-often) and particles (e.g. إن/ < n/indeed and أن/ > n/that),
- Its signaling force (strong or weak). Strong connectives trigger one discourse relation, such as كي/ky/to, لكن/lkn/but, غير أن/gyr > n/nevertheless, بيد أن/byd > n/however, and من أجل أن/mn > jl > n/in-order-to. On the other hand, weak connectives are ambiguous. They can trigger different discourse relations or do not trigger any discourse relation. Some of these connectives include the connector و/w/and, حتى/HtY/to, and the particles ل/l/for-to, ف/f/then, etc. For example, the particle و/w/and can signal the relation زمني ربط دون ترتيب/rbT dwn trtyb zmny/Continuation or it can be a part of a word, as in ورشة/wr$p/atelier,
- Its possible parts of speech, and
- The set of discourse relations that it can signal.

Each argument is associated to 7 lexical features. Four are binary and specify whether the argument contains a strong discourse cue, a weak discourse cue, a strong indicator and a weak indicator. One feature gives the list of all possible types of the lexical cue (clitic, simple or composed of more than word). The last two features are strings and give the list of all possible connective parts of speech (as encoded in the lexicon) and the list of discourse relations that it can trigger.

7. Experimentations and results

Our classifier aims to predict both explicit and implicit adjacent and non-adjacent discourse relations. To this end, we carried out supervised learning on the D-ATB corpus, based on the Maximum Entropy model (Berger et al., 1996), as implemented in the Stanford MaxEnt package.[13] For all the experiments, regularization parameters are set to their default value. We used both character n-grams and word n-grams as features. Best results were achieved with $n = 4$. All experiments were evaluated using 10-fold cross-validation. We report on our experiments in fine-grained discourse relations recognition (henceforth, Level 3 with 24 relations), in mid-level classes (henceforth, Level 2 with 13 relations) and also in the top-level classes (henceforth, Level 1 with 4 relations). For each level, we have the same number of instances, i.e. 3184 vectors. See Table 4 (cf. Section 5.3) for a more detailed statistics on each level.

We compare our models to three baselines. The first one (B1) attributes to each instance the most frequent relation. This corresponds to the relation ربط دون ترتيب زمني/rbT dwn trtyb zmny/Continuation for Level 3 and Level 2 and to the relation إنشائي/ < n$A}y/Thematic for Level 1. The second baseline (B2) is based on lexical cues features only (i.e. (F13), as described in the last section). Finally, the last baseline (B3) is composed of (Al-Saif and Markert, 2011)'s features where each instance is represented by a vector composed of all the features (F1) to (F5), which correspond respectively to connectives, arguments, al-masdar, tense and negation, and length and distance.

In the remainder of this section, we first give experiments overall results. Then, we detail the results on each level (Level 1, Level 2 and Level 3). We finally conclude by presenting the learning curves.

7.1. Overall results

We have first measured the effectiveness of each group of features ((F6) to (F13)) on fine-grained discourse relation classification. We built 8 individual classifiers where each model was trained by adding a new group of features to the baseline (B3). The classifiers are compared to the majority baseline (B1) (accuracy = 0.211), to (B2) and to (B3). The results are shown in Table 6 in terms of micro-averaged *F*-score and accuracy (the number of correctly predicted instances over the total number of instances). (*) indicates that the corresponding classifier yields significantly better performance over the baseline (B3) with $p < 0.050$ using Mc Nemar's test. Micro-averaged *F*-score is computed globally over all category decisions. Precision and recall are obtained by summing over all individual decisions as follows:

$$\pi = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + FP_i)},$$

$$\rho = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + N_i)}$$

where M is the number of category decisions. Micro-averaged *F*-measure is then computed as:

$$F\,(\text{micro-averaged}) = \frac{2\pi\rho}{\pi + \rho}$$

We observe that the baseline based on lexical cues (B2) outperforms the majority baseline (B1) in terms of accuracy. When adding connectives (F1) and arguments (F2) features to (B2), the micro-averaged *F*-score on Level 3 was improved by 0.151 over (B1) and by 0.790 over (B2). Moreover, when adding al-masdar features (F3) and tense and negation features (F4) to (B2), we obtain an *F*-score of 0.414 and an accuracy of 0.600 (which is relatively close to the results obtained by

Table 6 Overall results for the fine-grained classification.

	F-score	Accuracy
B2 (F13)	0.290	0.422
B3 ((F1) to (F5))	0.432	0.635
B3 + (F6) (*)	0.453	0.654
B3 + (F7)	0.468	0.674
B3 + (F8) (*)	0.442	0.644
B3 + (F9)	0.444	0.646
B3 + (F10) (*)	0.456	0.655
B3 + (F11)	0.453	0.655
B3 + (F12) (*)	0.438	0.649
B3 + (F13) (*)	0.453	0.657
Our model (*)	**0.613**	**0.778**

The bold values in the table represent the total or the average of results.

[13] We experimented with three machine learning algorithms: MaxEnt, NaiveBase and SVM. Best results were achieved by MaxEnt.

(B3)). When evaluating the contribution of individual features on fine-grained relation identification, our results confirm that each individual classifier outperforms all the baselines. Best combinations in terms of accuracy were achieved by adding punctuation features ((B3) + (F7)). On the other hand, the combinations (B3) + (F9) (i.e. named entity and anaphora features) and (B3) + (F8) (i.e. embedding features) resulted in a marginal improvement over the baseline (B3). The combinations (B3) + lexical cues (F13), (B3) + modality (F10), (B3) + textual organization (F6) and (B3) + semantic relations (F11) got almost similar results with an accuracy of 0.650. Among the 8 feature groups, only three get non-significant results over (B3). This can be explained by the fact that punctuation (F7) and named entity (F9) are partially taken into account by Al-Saif et al.'s morphological and syntactic features.

Once we have empirically demonstrated the effectiveness of each group of features individually, we have then assessed the performance of our model when combining all features. We have experimented several combinations. We found that optimal performances were obtained when adding features according to their coverage in the learning corpus. We started by adding to (B3) the features with the lowest frequency (F6) and we ended by adding the features with the highest frequency (F13). The last row in Table 6 shows the scores of our model (B3) + (F6) + (F7) + ... + (F13). The *F*-score and accuracy increase over the baseline (B3) by respectively 0.181 and 0.145. We have also analyzed the performance of our classifier depending on whether the relations link arguments within a sentence or outside the sentence. Our results show that predicting discourse relations within sentences achieved 0.070 better in terms of *F*-score compared to the results obtained when predicting discourse relations outside the sentence. Similarly, the performance of our classifier to predict explicit discourse relations is 0.140 higher than its capacity to predict implicit discourse relations.

Given the good results reached when using all the features for Level 3, we have run the same model for mid-level relation classification (Level 2) and for top-level classification (Level 1). Table 7 presents the results as well as the scores obtained by the three baselines in terms of micro-averaged *F*-score and accuracy. Here again, our models perform significantly better over the baseline B3 with $p < 0.050$ Mc Nemar's test.

Overall, the baseline (B3) gets very good results compared to (B2) with an *F*-score of 0.432, 0.511 and 0.588 respectively, for Level 3, Level 2 and Level 1. However, morphological and syntactic features, as given by Al-Saif and Markert (2011) are *insufficient* for achieving a good performance for our task. Our results are lower to the ones reported in Al-Saif and Markert, 2011 on identifying fine-grained discourse relations (accuracy = 0.70, *F*-score = 0.69) and on class-level relations (accuracy = 0.835, *F*-score = 0.75). This can be explained by three main reasons. Firstly, our classifier is based on features obtained automatically and not on gold standard annotations. Secondly, Al-Saif and Markert's model was trained to classify explicit discourse relations only while ours deals with explicit and implicit relations. Finally, Al-Saif and Markert's model focused on adjacent discourse relations only, while ours treats adjacent and non-adjacent relations.

Table 7 Overall results for the mid-class (Level 2) and coarse-grained (Level 1) classification.

	Level 2		Level 1	
	F-score	Accuracy	*F*-score	Accuracy
(B1)	–	0.211	–	0.587
(B2)	0.381	0.495	0.424	0.558
(B3)	0.511	0.673	0.588	0.697
Our model (*)	**0.653**	**0.778**	**0.758**	**0.828**

Finally, it is interesting to note that our features alone (cf. (F6) to (F13)) lead to lower results compared to (B3) for all the configuration levels. For example, on Level 3, we obtain an *F*-score of 0.370 and an accuracy of 0.500. These results show that using only semantic features (like modality, AWN, MPQA, etc.) cannot outperform the baseline (B3) and that morphological and syntactic features are *primordial* for our task.

7.2. Fine-grained classification

In this section we analyze the impact of each group of features ((F6) to (F13)) in predicting fine-grained relations within the إنشائي/<n$A}y/Thematic, زمني/zmny/Temporal, بنيوي/bnywy/Structural, and سببي/sbby/Causal classes. Figs. 7–10 present respectively how *F*-scores evolve when adding each feature group.

Fig. 7 shows that textual organization (F6) does not have any impact on thematic relations. Both embedding (F8) and named entity and anaphora features (F9) highly influence the results of تعيين/tEyyn/E-Elaboration. This is consistent with the definition of this relation that holds when an entity introduced in the first argument is detailed in the second argument. In Arabic, this relation is often marked by subordinate conjunctions such as الذي/Al*y/that-which-who, التي/Alty/that-which-who, or by possessive pronouns like هو/hw/he-him-it, هي/hy/she-her-it. Similarly, as expected, punctuation features (F7) improve the *F*-score of استدلال/AstdlAl/Attribution by 0.090 over (B3) + (F6). Concerning the other relations, we note that the relation تفصيل/tfSyl/Description reaches its best performance when adding embedding features (F8) while the same features have no impact on the relation تلخيص/tlxyS/Summary. Semantic relations (F11) and polarity features (F12) have a very good impact on تعليق/tElyq/Commentary (+0.070). Indeed, subjectivity is often used to express commentaries, as in (21).

(21) [لعب اليوم المنتخب التونسي.]$_1$
[كان اللعب دون المستوى.]$_2$

[lEb Alywm Almntxb Altwnsy.]$_1$ [kAn AllEb dwn AlmstwY.]$_2$
[The Tunisian team played today.]$_1$ [The game was awful.]$_2$

In Fig. 8, we observe that punctuation features (F7) have a great impact on the performance of the relations ترتيب ببطء/trtyb bbT'/Slow ordering and ترتيب بسرعة/trtyb bsrEp/Quick ordering, since their corresponding *F*-scores increase by respectively 0.150 and 0.180 over (B3). Indeed, these relations usually hold when events within units are separated by commas, as in (22). Embedding features (F8) do not seem to improve the results for all the relations. Named entity and anaphora features (F9) boost the scores of all the relations. This is very salient for تأطير/t > Tyr/Frame with an improvement of more

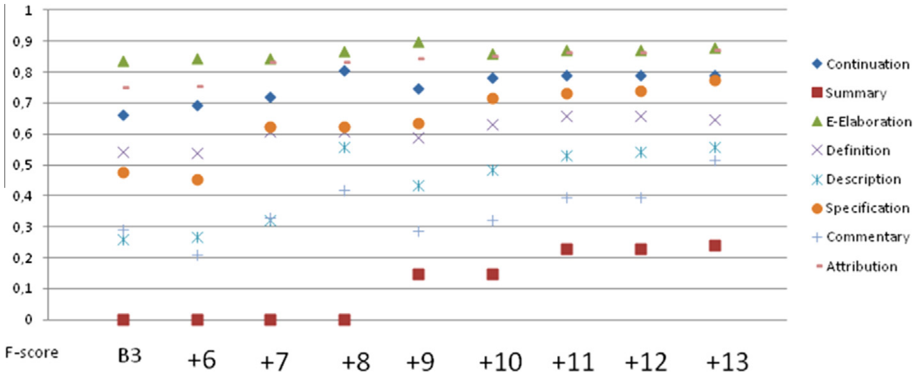

Figure 7 Impact of our features on the إنشائي/< n$A}y/Thematic relations in terms of *F*-score.

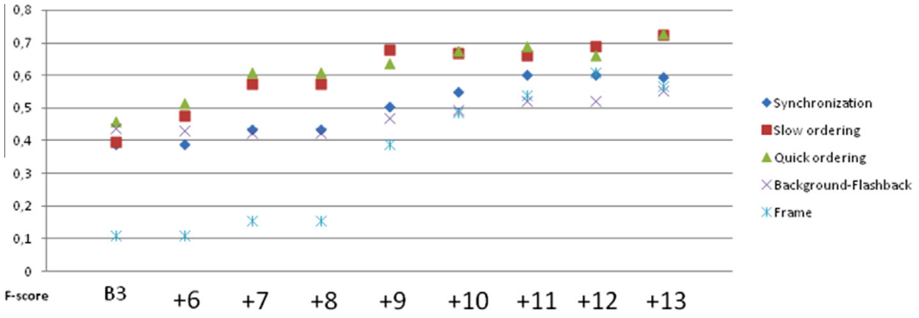

Figure 8 Impact of our features on the زمني/zmny/Temporal relations in terms of *F*-score.

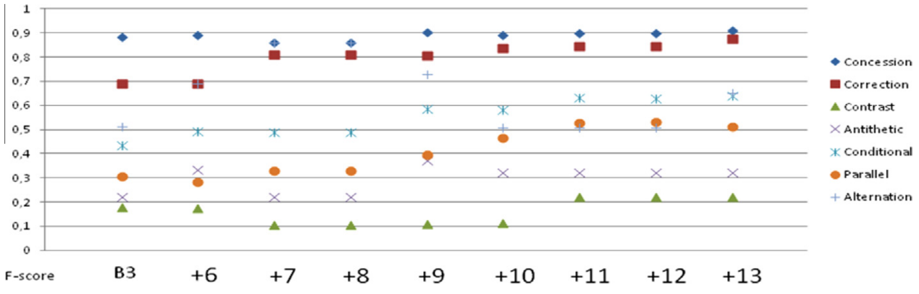

Figure 9 Impact of our features on the بنيوي/bnywy/Structural relations in terms of *F*-score.

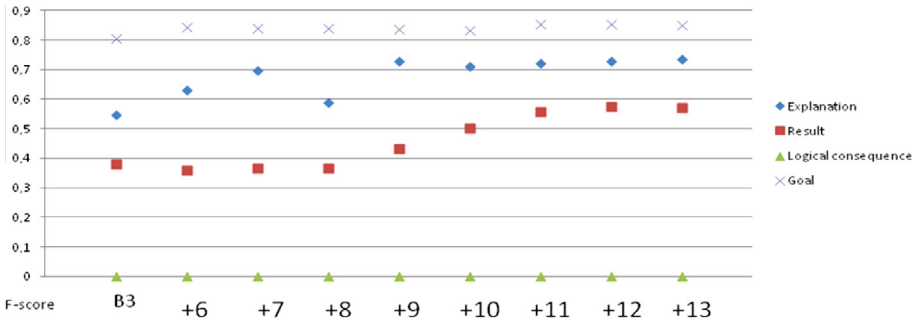

Figure 10 Impact of our features on the سببي/sbby/Causal relations in terms of *F*-score.

than 0.290 over (B3) mainly because the first argument of this relation contains temporal or spatial frames that are often named entities. The other features have a significant impact on all the relations except for lexical cues (F13), polarity (F12) and semantic relation features (F11) that degrade the result of the relation ترتيب ببطء/trtyb bbT'/Slow ordering.

(22) [قاموا بحرق المؤسسات العمومية،]$_1$ [ثم المحلات التجارية،]$_2$ [ثم المنازل.]$_3$

[qAmwA bHrq Alm&ssAt AlEmwmyp,]$_1$ [vm AlmHlAt Altjaryp,]$_2$ [vm AlmnAzl.]$_3$
[They burnt public institutions,]$_1$ [then shops,]$_2$ [then houses]$_3$

Fig. 9 clearly distinguishes between two groups of relations: (a) شرط/$rT/Conditional, تخيير/txyyr/Alternation, إضراب/<DrAb/Correction, and استدراك/AstdrAk/Concession that achieve good results (*F*-score > 0.600), and (b) طباق/TbAq/Antithetic, مقابلة/mqAblp/Contrast, and معية/mEyp/Parallel that perform badly (*F*-score < 0.500).

For the first group (a), textual organization features (F6) did not provide any improvement over the baseline (B3), except for تخيير/txyyr/Alternation. Punctuation features (F7) boost the results of إضراب/<DrAb/Correction whereas the features (F8) to (F13) seem to have a non negligible impact on this relation. Lexical cues (F13) slightly increase the results of تخيير/txyyr/Alternation, شرط/$rT/Conditional and إضراب/<DrAb/Correction, which are often signaled in Arabic by specific markers like إما/<mA/either, أو/>w/or, أم/>m/or, and سواء/swA'/either for تخيير/txyyr/Alternation (cf. (23)), س/s/so, لو/lw/if, إذا/<*A/if, and لولا/lwlA/except for شرط/$rT/Conditional, and بل/bl/however for إضراب/<DrAb/Correction.

(23) [إما أن ارتاح قليلا]$_1$ [أو أشاهد التلفاز]$_2$

[<mA >n ArtAH qlylA]$_1$ [>w>$Ahd AltlfAz]$_2$
[Either I'll sleep]$_1$ [or I'll watch TV]$_2$

For the second group (b), we observe a different behavior where the features (F7) to (F10) degraded the results of مقابلة/mqAblp/Contrast while at the same time, their contributions on the two other relations of this group are mitigated. Semantic relations (F11) have a very good impact on مقابلة/mqAblp/Contrast (+0.10). Indeed, antonyms are often used to express contrasts, as in (24). It is however surprising that we did not observe the same positive effect of these features on the relation طباق/TbAq/Antithetic since this relation holds when there is a verb in the first argument and its negation in the second argument or when the two verbs are antonyms, as in (25). We think that this can be explained by the low frequency of this relation in the dataset (0.38%). Another interesting finding is that semantic relation features (F11) boost the results of معية/mEyp/Parallel by more than 0.060 over (B3)+(F6) to (F10). Indeed, this relation indicates that two units share the same event and have semantically similar constituents, which is captured by some semantic relations of Arabic WordNet such as Near_syonym.

(24) [يضحك أخي]$_1$ [وفي المقابل تبكي أختي.]$_2$

[yDHk>xy]$_1$ [w fy AlmqAbl tbky >xty.]$_2$
[My brother laughs]$_1$ [however my sister cries.]$_2$

(25) [يضحك أخي]$_1$ [ويبكي.]$_2$

[yDHk>xy]$_1$ [wybky.]$_2$
[My brother laughs]$_1$ [and cries.]$_2$

Finally, Fig. 10 shows that our model fails to predict infrequent relations, like استنتاج/AstntAj/logical-consequence. غرض/grD/Goal and سبب/sbb/Explanation led to the best *F*-scores with respectively 0.851 and 0.735. When adding embedding features (F8), the *F*-score of the relation سبب/sbb/Explanation degrades by 0.111. Named entity and anaphora features (F9) boost the scores of the relations سبب/sbb/Explanation and نتيجة/ntyjp/Result whereas these features have no impact on the other relations. Lexical cue features (F13) have no impact on the causal relations.

Overall, we can conclude that each added feature has its own specificities. Some of them are useful for predicting some discourse relations, while they have at the same time a negative impact on predicting other relations. Adding textual organization and punctuation features ((F6) and (F7)) has significantly improved the results of discourse relations that generally hold at the beginning of the paragraph or relations that link arguments containing specific punctuations (like استدلال/AstdlAl/Attribution, ترتيب ببطء/trtyb bbT'/Slow ordering, and ترتيب بسرعة/trtyb bsrEp/Quick ordering). However, these features perform badly on non-adjacent discourse relations (like نتيجة/ntyjp/Result, تفصيل/tfSyl/Description and خلفية/xlfyp/Background-Flashback). Modality (F10), WordNet (F11) and polarity (F12) features contribute to improve the recall, especially for implicit discourse relations. Finally, adding lexical cues features (F13) have a significantly good impact on the discourse relations that are signaled by strong connectors. However, (F13) decreases the results of discourse relations that are signaled by clitics (و/w/and,ف/f/so, and ل/l/for).

Error analysis at Level 3 shows that our model fails to discriminate between the relations غرض/grD/Goal and سبب/sbb/Explanation (cf. example (26)), the relations استدلال/AstdlAl/Attribution and تعيين/tEyyn/E-Elaboration, and the relations تعيين/tEyyn/E-Elaboration and تفصيل/tfSyl/Description.

(26) [وصف الطبيب للمريض مجموعة من الأدوية]$_1$ [لمعالجة ألمه وجرحه.]$_2$

[wSf AlTbyb llmryD mjmwEp mn Al>dwyp]$_1$[lmEAljp >lmh wjrHh]$_2$
[The doctor prescribed his patient a set of drugs]$_1$ [to treat his pain and injury.]$_2$
Gold corpus: غرض/grD/Goal (1,2)
Predicting relation: سبب/sbb/Explanation (1,2)

7.3. Mid-level classification

Table 8 presents the detailed results for the mid-level classification using all features in terms of precision, recall, *F*-score, and accuracy. The last row presents the average precision, the average recall, and the average *F*-score as well as the overall accuracy of the model. Best results are achieved by the relation استدلال/AstdlAl/Attribution (*F*-score = 0.854) while the lowest score has been obtained by the relation تلخيص/tlxyS/Summary (*F*-score = 0.240).

Error analysis at this level shows that the most frequent confusions concern the relations إسهاب/<shAb/Elaboration and the relations of the سببي/sbby/Causal class especially when these relations are implicit (cf. example (27)). Other

errors include the distinction between the relations استدلال/AstdlAl/Attribution and إسهاب/<shAb/Elaboration.

(27) [لقد استغنيت عن هذا الكتاب،]$_1$

[انه لا يحتوي على معلومات قييمة،]$_2$

[lqd Astgnyt En h*A AlktAb,]$_1$ [Anh lA yHtwy ElY mElwmAt qy-ymp,]$_2$
[I do not need this book,]$_1$[it does not contain any important information,]$_2$
Gold corpus: سبب/sbb/Explanation (1,2)
Predicting relation: إسهاب/<shAb/Elaboration (1,2)

7.4. Coarse-grained classification

Table 9 presents our results on the coarse-grained classification using all the features in terms of precision, recall, *F*-score, and accuracy. The last row presents the average precision, the average recall, and the average *F*-score as well as the overall accuracy of the model. The frequency of each class in the D-ATB corpus is indicated between brackets. Our model achieves an *F*-score of 0.758 and an overall accuracy of 0.828, which is relatively close to the results obtained by relation recognition in English (see the related work section).

Table 10 shows major confusions. Main errors (in bold font) are between إنشائي/<n$A}y/Thematic and سببي/sbby/Causal classes.

7.5. Learning curves

In order to analyze how the number of annotated documents influences the learning procedure, we have computed a learning curve, by dividing our corpus into 10 different learning sets. For each set, we performed a 10-fold cross-validation for each classification level. The learning curve is shown in Fig. 11. For Level 1, the curve grows steadily between 0 and 2000 discourse relations (that is 45 documents, i.e. around 1200 sentences) while it seems to plateau between 2000 and 3184 discourse relations (that is 70 documents). We can thus conclude that the addition of more than 45 documents will only slightly increase the performance of the classifier. However, the curve for Level 2 seems to plateau between 2400 and 3184 discourse relations while the curve of Level 3 seems to plateau between 2800 and 3184 discourse relations.

Table 8 Detailed results at the mid-level classification (Level 2).

Level 2	Precision	Recall	*F*-score	Accuracy
Continuation	0.776	0.830	0.802	0.883
Elaboration	0.816	0.846	0.830	0.922
Attribution	0.843	0.868	0.854	0.959
Conditional	0.734	0.566	0.621	0.975
Cause-effect	0.798	0.808	0.802	0.931
Goal	0.825	0.878	0.851	0.973
Background-Flashback	0.634	0.511	0.548	0.971
Opposition	0.804	0.734	0.747	0.982
Parallel	0.651	0.493	0.550	0.979
Temporal ordering	0.694	0.655	0.661	0.959
Correction	0.941	0.775	0.822	0.996
Commentary	0.533	0.370	0.423	0.988
Frame	0.746	0.490	0.581	0.992
Alternation	0.513	0.458	0.456	0.995
Summary	0.330	0.188	0.240	0.997
Total	**0.709**	**0.631**	**0.653**	**0.778**

The bold values in the table represent the total or the average of results.

Table 9 Detailed results at the top-level classification (Level 1).

Level 1	Precision	Recall	*F*-score	Accuracy
إنشائي/<n$A}y/Thematic	0.892	0.919	0.905	0.870
سببي/sbby/Causal	0.764	0.698	0.729	0.886
بنيوي/bnywy/Structural	0.713	0.709	0.711	0.923
زمني/zmny/Temporal	0.688	0.684	0.686	0.932
Total	**0.764**	**0.752**	**0.758**	**0.828**

The bold values in the table represent the total or the average of results.

Table 10 Confusion matrix for the coarse-grained classification.

	Thematic	Causal	Structural	Temporal
Thematic	1727	**112**	52	45
Causal	**82**	422	21	27
Structural	38	34	261	33
Temporal	32	37	34	227

8. Related work

We present in this section the main existing work on discourse relations recognition, by grouping them according to their corresponding theoretical frameworks.

Marcu and Echihabi (2002) proposed the first unsupervised learning approach to detect RST discourse relations, such as Contrast, Explanation-Evidence, Condition and Elaboration that hold between arbitrary spans of texts. They showed that word pair features are important cues for detecting implicit relations. Saito et al. (2006) extended this approach and experimented with a combination of cross-argument word pairs and phrasal patterns to recognize implicit relations between adjacent sentences in a Japanese corpus. (Blair-Goldensohn et al., 2007) further extended this first unsupervised model by using syntactic filtering and topic segmentation. Several authors have also proposed supervised approaches based on

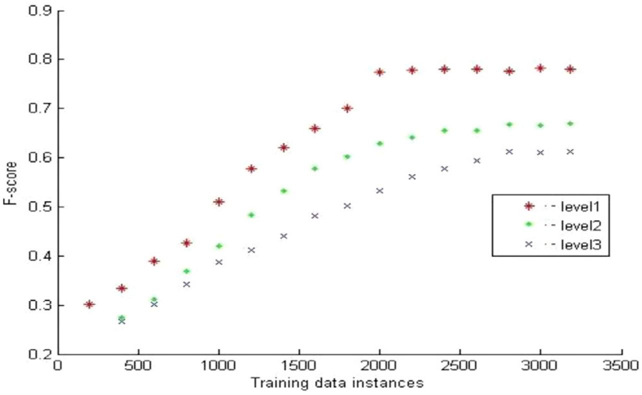

Figure 11 The learning curve of our three level models.

manually annotated data. For English, the RST Discourse Treebank (RST-DT) (Carlson et al., 2003) built on the top of the syntactically annotated Penn Treebank, is one of the well-known RST resources. Relations in RST-DT are grouped into 18 classes, which are further specified into 78 relations, which are organized by nuclearity (nucleus-satellite or multinuclear rhetorical relations). Soricut and Marcu (2003) developed a sentence-level discourse parser using syntactic and lexical features and showed a strong correlation between syntactic and discourse information. Subba et al. (2009) proposed a first-order logic learning approach to relation classification using lexical and linguistic information and compositional semantics.[14] DuVerle and Prendinger (2009) developed a full RST structure parser using a rich features space including lexical, semantic, and structural features. To overcome the problem of infrequent discourse relations in the training set, Hernault et al. (2010a) proposed a semi-supervised discourse relations classification using state of the art features including word pairs, production rules and lexico-syntactic context at the border between two units of texts. Feng and Hirst (2012) extended the HILDA discourse parser (Hernault et al., 2010b) by exploring various rich linguistic features for text-level discourse parsing such as verb classes, semantic similarities, clue phrases, production rules and contextual features that encode the discourse relations assigned by the preceding and the following text span pairs. Finally, Sadek et al. (2012) proposed a rule-based approach to automatically determine RST relations such as Causal, Evidence, Explanation, Purpose, Interpretation, Base, Result, and Antithesis. These relations were then used in a question answering system to answer non-factoid questions (“Why” and “How to”).

To date, two SDRT-like parsers exist. One has been developed for appointment scheduling dialogues (Baldridge and Lascarides, 2005) and the other was developed on top of the Annodis corpus, a French manually built resource with discourse information (Muller et al., 2012). Baldridge and Lascarides (2005) represented discourse structures as headed trees and model them with probabilistic head-driven parsing techniques. They combined lexical features, features inspired from syntactic parsing and dialogue-based features and showed that the last group of features has a great impact on the performance of their model. Muller et al. (2012) proposed a text-level discourse parsing algorithm by performing an A^* global search over the space of possible discourse structures while optimizing a global criterion over the set of potential coherence relations. Best results were achieved with MaxEnt and A^*.

Wellner et al. (2006) proposed to automatically learn explicit and implicit relations using the Discourse GraphBank corpus (Wolf and Gibson, 2005) as a training set. They used shallow syntactic information, modal parsing (identifying subordinate verb relations and their types), temporal ordering of events and lexical semantic typing including similarity measures between words using a variety of knowledge sources.

The development of several manually annotated resources following the PDTB model has encouraged researches to investigate both explicit and implicit relations recognition in several languages using supervised learning techniques. In the English language, experiments have been done using the PDTB v2.0 (Prasad et al., 2008) corpus that groups relations into a taxonomy of 16 relations at the middle level and 4 coarse top-level classes (Temporal, Contingency, Comparison, Expansion) for a total of 33 relations. Pitler et al. (2008) and Pitler et al. (2009) respectively investigated automatic detection of explicit and implicit relations using lexical, syntactic and linguistically informed features. Lin et al. (2009) implemented an implicit discourse relations model by using the same features as in (Pitler et al., 2009) and by adding constituency parse features such as production rules and dependency parse features. Zhou et al. (2010) detected implicit relations by automatically inserting discourse connectives between arguments using a language model. Louis et al. (2010) focused on implicit relations that link adjacent arguments and experimented with co-reference information, grammatical role, information status and syntactic form of referring expressions. Park and Cardie (2012) provided a systematic study of state of the art features (word and Pairs, the first, the last, and the first three words of each argument, polarity, verbs, inquirer Tags, modality, context and production rules) for learning implicit discourse relations and identified feature combinations that optimize F1-score using the forward selection algorithm. Wang et al. (2011) proposed a typical/atypical perspective to select the most suitable training examples for implicit discourse relations recognition. For Chinese, Huang and Chen (2012) used lexical and shallow syntactic features such as named entity, collocated words, punctuations and argument length. Finally for Arabic, Al-Saif and Markert (2011) proposed a two-step algorithm for Arabic discourse analysis: first discourse connective recognition by identifying the discourse and the non-discourse usage of Arabic connectives linking adjacent arguments, then discourse connective interpretation. They used state of the art features, extracted from the ATB gold standard parsers, and showed that production rule features degraded their performances. They achieved an accuracy of 0.770 on a fine-grained discourse relations and an accuracy of 0.835 on class-level discourse relations.

9. Conclusion

In this paper, we proposed the first model that automatically identifies explicit and implicit Arabic discourse relations that link adjacent as well as non-adjacent discourse units. We used the Discourse Arabic Treebank corpus (D-ATB), the first resource that makes explicit the interactions between the semantic content of discourse units and the global, pragmatic structure of the discourse, following the Segmented Discourse Representation Theory framework. Rhetorical relations were built from a semantic point of view and were defined according to their effect on meaning relying on Arabic rhetoric literatures and corpus analysis. Our hierarchy of discourse relations is organized around 4 top-level classes with a total of 24 relations.

We proposed a supervised learning approach that uses several kinds of features. We analyzed how each feature contributes to the learning process. We first experimented with morphological and syntactic features, as already done by (Al-Saif and Markert, 2011). Our results show that these features are crucial for discourse relation recognition but they are not sufficient for achieving good results. When adding contextual, lexical and lexico-semantic features, our results have

[14] The set of relations used by the authors mixes the classification proposed by Moser et al. (1996) and Marcu (1999).

been boosted for all the configurations (fine-grained discourse relations, mid-level classes and also top-level classes). We compare our approach against three baselines that are based on the most frequent relation, discourse connectives and the features used by (Al-Saif and Markert, 2011). Our results outperform all the baselines.

We plan to extend this work by building an SDRT parser for Arabic. We also plan to use this parser for Arabic text summarization.

References

Abdul-Raof, H., 2012. Arabic Rhetoric. A Pragmatic Analysis. Routledge, ISBN10: 0–415–38609–8..

Abubakre, R.D., 1989. Bayan in Arabic Rhetoric: An Analysis of the Core of Balāgha. Intec Printer Limited, Ibadan.

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., Vieu, L., 2012. An empirical resource for discovering cognitive principles of dis-course organisation: the annodis corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC 2012.

Al-Jarim, A., Amine, M., 1999. البلاغةالواضحة/Al-Balagha al-Wadiha. Editor: Dar Al-maaref. ISBN: 977-02-5784-2.

Al-Saif, A., Markert, K., 2010. The Leeds Arabic discourse Treebank: annotating dis-course connectives for Arabic. Proceedings of the International Conference on Language Resources and Evaluation, (LREC 2010), Valletta, Malta.

Al-Saif, A., Markert, K., 2011. Modelling discourse relations for Arabic. The proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2011), Edinburgh.

Asher, N., 1993. Reference to Abstract Objects in Discourse. Kluwer, Dordrecht.

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

Baldridge, J., Lascarides, A., 2005. Probabilistic head-driven parsing for discourse structure. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNNL), Ann Arbor, 2005.

Benajiba, Y., Rosso, P., Benedi, J.M., 2007. ANERsys: an Arabic named entity recognition system based on maximum entropy. CICLing, Springer-Verlag, Berlin, Heidelberg, pp. 143–153.

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

Black, W., Elkateb, S., Vossen, P., 2006. Introducing the Arabic WordNet Project, International WordNet Conference, 2006.

Blair-Goldensohn, S., McKeown, K., Rambow, O., 2007. Building and refining rhetorical semantic relation models. In: HLT-NAACL, pp. 428–435.

Boudabous, M.M., Chaâben, N., Khedher, N., Hadrich Belguith, L., Sadat, F., 2013. Arabic WordNet semantic relations enrichment through morpho-lexical patterns, The First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), Sharjah, UAE, February 12–14, 2013.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M., 2011. Alkhalil morpho sys: a morpho-syntactic analysis system for Arabic texts.

Carlson, L., Marcu, D., Okurowski, M.E., 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt, J., Smith, R. (Eds.), Current Directions in Discourse and Dialogue. Kluwer, New York, pp. 85–112.

Danlos, L., 2007. Strong generative capacity of RST, SDRT and discourse dependency DAGs. Constraints in Discourse. Benjamins, editor A. Benz, P. Khnlein.

DuVerle, D.A., Prendinger, H., 2009. A novel discourse parser based on support vector machine classification. Proceedings of ACL, 2009.

Elarnaoty, M., AbdelRahman, S., Fahmy, A., 2012. A machine learning approach for opinion holder extraction in Arabic. Int. J. Artif. Intell. Appl. 3 (2).

Feng, V., Hirst, G., 2012. Text-level discourse parsing with rich linguistic features. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2012), Jeju, Korea.

Green, S., Manning, C. 2010. Better Arabic parsing: baselines, evaluations, and analysis. COLING 2010.

Hernault, H., Bollegala, D., Ishizuka, M., 2010a. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, October. Association for Computational Linguistics, pp. 399–409.

Hernault, H., Prendinger, H., duVerle, D., Ishizuka, M., 2010b. HILDA: a discourse parser using support vector machine classification. Dialog. Discourse 1 (3), 1–33.

Huang, H., Chen, H., 2011. Chinese discourse relation recognition. Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, Thailand. November 2011, pp. 1442–1446.

Huang, H., Chen, H., 2012. Contingency and comparison relation labeling and structure prediction in Chinese sentences. Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Seoul, South Korea, 5–6 July 2012, pp. 261–269.

Hutchinson, B., 2004. Acquiring the meaning of discourse markers. In the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pp. 684–691, Barcelona, Spain, pp. 684–691.

Kamp, H., Reyle, U., 1993. From Discourse to Logic. Dordrecht, 1993.

Keskes, I., Benamara, F., Belguith Hadrich, L., 2014. Splitting Arabic texts into elementary discourse units. Forthcoming in Transactions on Asian Language Information Processing (TALIP).

Lin, Z., Kan, M., Tou, H., 2009. Recognizing implicit discourse relations in the Penn discourse Treebank. EMNLP, pp. 343–351.

Lin, Z., Tou H., Kan, M., 2010. A PDTB-styled end-to-end discourse parser. Technical report, School of Computing, National University of Singapore.

Louis, A., Aravind, Joshi, K., Prasad, R. Nenkova, A., 2010. Using entity features to classify implicit discourse relations. SIGDIAL Conference, pp. 59–62.

Maamouri, M., Bies, A., Kulick, S. Krouma, S., Gaddeche, Zaghouani, W., 2010b. Arabic Treebank (ATB): Part 3 Version 3.2. Linguistic Data Consortium, Catalog No.: LDC2010T08.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., Kulick, S., 2010a. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium, Catalog No.: LDC2010L01.

Mann, W.C., Thompson, S., 1988. Rhetorical structure theory: toward a functional theory of text organization. Text 8 (3), 243–281.

Marcu, D., 1999. Instructions for Manually Annotating the Discourse Structures of Texts. Technical Report, University of Southern California, 1999.

Marcu, D., 2000. From discourse structures to text summaries in Workshop Intelligent Scalable Text Summarization ACL, Madrid, Espagne, 2000, pp. 82–88.

Marcu, D., Echihabi, A., 2002. An unsupervised approach to recognizing discourse relations. ACL, pp. 368–375.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., Webber, B., 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. TLT 2005.

Moser, M.G., Moore, J.D., Glendening, E., 1996. Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units. University of Pittsburgh, Department of Computer Science.

Muller, P., Afantenos, S.P., Asher, N., 2012. Constrained decoding for text-level discourse parsing. Proceedings of COLING.

Musawi, A., Muhsin, J., 2001. Arabic rhetoric. In: Sloane, Thomas O. (Ed.), Oxford Encyclopaedia of Rhetoric. Oxford University Press, Oxford, pp. 29–33.

Owens, J., 2006. A linguistic history of Arabic. Published to Oxford Scholarship Online Print ISBN-13: 9780199290826, doi: 10.1093/acprof:oso/9780199290826.001.0001.

Park, J., Cardie, C., 2012. Improving implicit discourse relation recognition through feature set optimization. Proceedings of the 13th Annual SIGdial Meeting on Discourse and Dialogue, SIGDIAL.

Pitler, E., Louis A., Nenkova, A., 2009. Automatic sense prediction for implicit discourse relations in text. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., Joshi, A., 2008. Easily identifiable discourse relations. Proc. COLING.

Prasad, A., Miltsakaki, R., Dinesh, E., Lee, N., Joshi, A., Webber, B., 2008. The Penn discourse Treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Reese, B., Hunter, J., Denis, P., Asher, N., Baldridge, J., 2007. Reference Manual for the Analysis and Annotation of Rhetorical Structure. Technical Report. Department of Linguistics, The University of Texas, Austin.

Sadek, J., Chakkour, F., Meziane F., 2012. Arabic Rhetorical relations extraction for answering "Why" and "How to" questions. In: Bouma, G., Ittoo, A., Métais, E., Wortmann, H. (Eds.), NLDB 2012. LNCS, vol. 7337, Springer, Heidelberg, pp. 385–390.

Saito, M., Yamamoto, K., Sekine, S., 2006. Using phrasal patterns to identify discourse relations. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006), New York, USA, June, pp. 133–136.

Sloane, T.O., 2001. Encyclopedia of Rhetoric. Oxford University Press, New York, p. 837, p. xii.

Soricut, R., Marcu, D., 2003. Sentence level discourse parsing using syntactic and lexical information. HLT-NAACL.

Subba, R., Eugenio, B., 2009. An effective discourse parser that uses rich linguistic information. Uman Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, Boulder, Colorado, June, pp. 566–574.

Venant, A., Asher, N., Muller, P., Denis, P., Afantenos, S., 2013. Expressivity and comparison of models of discourse structure. Proceedings of the SIGDIAL 2013 Conference.

Versley, Y., 2013. Subgraph-based classification of explicit and implicit discourse relations. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013), pp. 264–275.

Wang, L., Lui, M., Kim, S.N., Nivre, J., Baldwin, T., 2011. Predicting thread discourse structure over technical web forums. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 13–25.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., Sauri, R., 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue.

Wolf, F., Gibson, E., 2005. Representing discourse coherence: a corpus-based study. Comput. Linguist., 249–287.

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

Zhou, Z., Xu, Y., Niu, Z., Lan, M., Su, J., Lim, T.C., 2010. Predicting discourse connectives for implicit discourse relation recognition. COLING (Posters), pp. 1507–1514.