République Tunisienne
Ministère de l'Enseignement Supérieur, de la Recherche Scientifique
et des Technologies de l'Information et de la Communication

Université de Sfax
Faculté des Sciences Économiques et de Gestion de Sfax

THESE EN COTUTELLE

Préparée en vue de l'obtention
du titre de docteur en Informatique

Discourse Analysis of Arabic Documents and Application to Automatic Summarization

Présentée et soutenue le 11/05/2015 par

Iskandar KESKES

Membres du Jury :

Prof. Lamia HADRICH BELGUITH	Professeur - FSEG – Sfax - Tunisie	Directeur de thèse
Prof. Nicholas ASHER	Directeur de recherche CNRS - IRIT - Toulouse - France	Directeur de thèse
Dr. Farah BENAMARA ZITOUNE	Maître de conférences - IRIT- Toulouse - France	Co-directeur de thèse
Dr. Nizar HABASH	Maître de conférences - Université de New York - Abu Dhabi	Rapporteur
Prof. Rim FAIZ	Professeur - IHEC - Tunis - Tunisie	Rapporteur
Prof. Claudette CAYROL	Professeur - UPS - Toulouse - France	Examinateur

Année Universitaire 2014-2015

Acknowledgment

First and foremost, thanks to the beneficent and merciful God who gives me everything I have.

This research thesis would not have been possible without the support of many people.

Therefore, I am deeply grateful to my supervisors Prof. Lamia HADRICH BELGUITH, Prof. Nicholas ASHER and Dr. Farah BENAMARA ZITOUNE for their patience, support, encouragement, supervision, kind advice and interesting follow up which gave me confidence and made the completion of this work possible. I benefited from their advice and valuable discussion, particularly so when exploring new ideas during this research project. I would like to thank them for keeping me focused and stimulating me to take a critical standpoint regarding my research. In our meetings, they were able to give me direct feedback and quickly identify problems, leading to a better quality of the output. They taught me to look at my thesis in a structured way, accompanied with a large dose of common sense.

Not forgetting my best friends Fatma SALLEM, Ahmed DERBEL, Jihen MAAZOUN and Salma DAMMAK thanks for their encouragement, assistance and love.

I wish to express my love and gratitude to my beloved family specially my father Jamil, my mother Monia, my brothers Bilel and Mohamed Amine, my brother in law Oussema and my sister Rania for their understanding & endless love, through the duration of my study. Without the support of all members of my family, I would never finish this thesis and I would never find the courage to overcome all these difficulties during this work.

I would especially like to express my gratitude to my fiance, Ines BOUJELBEN, who has always supported me and helped me overcoming the difficulties without complaining.

Iskander KESKES

Abstract

Within a discourse, texts and conversations are not just a juxtaposition of words and sentences. They are rather organized in a structure in which discourse units are related to each other so as to ensure both discourse coherence and cohesion. Discourse structure has shown to be useful in many NLP applications including machine translation, natural language generation and language technology in general. The usefulness of discourse in NLP applications mainly depends on the availability of powerful discourse parsers. To build such parsers and improve their performances, several resources have been manually annotated with discourse information within different theoretical frameworks. Most available resources are in English. Recently, several efforts have been undertaken to develop manually annotated discourse information for other languages such as Chinese, German, Turkish, Spanish and Hindi. Surprisingly, discourse processing in Modern Standard Arabic (MSA) has received less attention despite the fact that MSA is a language with more than 422 million speakers in 22 countries.

Computational processing of Arabic language has received a great attention in the literature for over twenty years. Several resources and tools have been built to deal with Arabic non concatenative morphology and Arabic syntax going from shallow to deep parsing. However, the field is still very vacant at the layer of discourse. As far as we know, the sole effort towards Arabic discourse processing was done in the Leeds Arabic Discourse Treebank that extends the Penn Discourse TreeBank model to MSA. In this thesis, we propose to go 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. In particular, we propose the first effort towards a semantically driven approach of Arabic texts following the Segmented Discourse Representation Theory (SDRT). Our main contributions are:

- A study of the feasibility of building a recursive and complete discourse structures of Arabic texts. In particular, we propose:
    - an annotation scheme for the full discourse coverage of Arabic texts, in which each constituent is linked to other constituents. A document is then represented by an oriented acyclic graph, which captures explicit and implicit relations as well as complex discourse phenomena, such as long-distance attachments, long-distance discourse pop-ups and crossed dependencies.
    - a novel discourse relation hierarchy. We study the rhetorical relations from a semantic point of view by focusing on their effect on meaning and not on how they are lexically triggered by discourse connectives that are often ambiguous, especially in Arabic.
    - a thorough quantitative analysis (in terms of discourse connectives, relation frequencies, proportion of implicit relations, etc.) and qualitative analysis (inter-annotator agreements and error analysis) of the annotation campaign.

- An automatic discourse parser where we investigate both automatic segmentation of Arabic texts into elementary discourse units and automatic identification of explicit and implicit Arabic discourse relations.
- An application of our discourse parser to Arabic text summarization. We compare tree-based vs. graph-based discourse representations for producing indicative summaries and show that the full discourse coverage of a document is definitively a plus.

Résumé

Dans un discours, les textes et les conversations ne sont pas seulement une juxtaposition de mots et de phrases. Ils sont plutôt organisés en une structure dans laquelle des unités de discours sont liées les unes aux autres de manière à assurer à la fois la cohérence et la cohésion du discours. La structure du discours a montré son utilité dans de nombreuses applications TALN, y compris la traduction automatique, la génération de texte et le résumé automatique. L'utilité du discours dans les applications TALN dépend principalement de la disponibilité d'un analyseur de discours performant. Pour aider à construire ces analyseurs et à améliorer leurs performances, plusieurs ressources ont été annotées manuellement par des informations de discours dans des différents cadres théoriques. La plupart des ressources disponibles sont en anglais. Récemment, plusieurs efforts ont été entrepris pour développer des ressources discursives pour d'autres langues telles que le chinois, l'allemand, le turc, l'espagnol et le hindi. Néanmoins, l'analyse de discours en arabe standard moderne (MSA) a reçu moins d'attention malgré le fait que MSA est une langue de plus de 422 millions de locuteurs dans 22 pays.

Le sujet de thèse s'intègre dans le cadre du traitement automatique de la langue arabe, plus particulièrement, l'analyse de discours de textes arabes. Cette thèse a pour but d'étudier l'apport de l'analyse sémantique et discursive pour la génération de résumé automatique de documents en langue arabe. Pour atteindre cet objectif, nous proposons d'étudier la théorie de la représentation discursive segmentée (SDRT) qui propose un cadre logique pour la représentation sémantique de phrases ainsi qu'une représentation graphique de la structure du texte où les relations de discours sont de nature sémantique plutôt qu'intentionnelle. Cette théorie a été étudiée pour l'anglais, le français et l'allemand mais jamais pour la langue arabe. Notre objectif est alors d'adapter la SDRT à la spécificité de la langue arabe afin d'analyser sémantiquement un texte pour générer un résumé automatique.

Nos principales contributions sont les suivantes :

- Une étude de la faisabilité de la construction d'une structure de discours récursive et complète de textes arabes. En particulier, nous proposons :
    - un schéma d'annotation qui couvre la totalité d'un texte arabe, dans lequel chaque constituant est lié à d'autres constituants. Un document est alors représenté par un graphe acyclique orienté qui capture les relations explicites et les relations implicites ainsi que des phénomènes de discours complexes, tels que l'attachement, la longue distance du discours pop-ups et les dépendances croisées.
    - une nouvelle hiérarchie des relations de discours. Nous étudions les relations rhétoriques d'un point de vue sémantique en se concentrant sur leurs effets

sémantiques et non pas sur la façon dont elles sont déclenchées par des connecteurs de discours, qui sont souvent ambigües en arabe.

- une analyse quantitative (en termes de connecteurs de discours, de fréquences de relations, de proportion de relations implicites, etc.) et une analyse qualitative (accord inter-annotateurs et analyse des erreurs) de la campagne d'annotation.

- Un outil d'analyse de discours où nous étudions à la fois la segmentation automatique de textes arabes en unités de discours minimales et l'identification automatique des relations explicites et implicites du discours.
- L'utilisation de notre outil pour résumer des textes arabes. Nous comparons la représentation de discours en graphes et en arbres pour la production de résumés.

Table of contents

List of Tables 12
List of Figures 13
General Introduction 14
Chapter 1: Background 20
Introduction 21
1. Discourse analysis 21
1.1. Basic notions 21
1.1.1. Discourse connectives 21
1.1.2. Discourse units 23
1.1.3. Discourse relations 24
1.1.4. Discourse structures 24
1.1.5. Discourse cohesion and discourse coherence 24
1.2. Main discourse theories 25
1.2.1. Discourse Representation Theory 25
1.2.2. Rhetorical Structure Theory 26
1.2.3. Segmented Discourse Representation Theory 29
1.2.4. GraphBank model 32
1.2.5. Penn Discourse TreeBank model 34
2. Arabic discourse analysis 36
2.1. Arabic specificities 36
2.2. Arabic particularities at the discourse level 37
2.2.1. General specificities 37
2.2.2. Arabic discourse connectives 38
2.3. Main studies on Arabic discourse processing 42
2.3.1. Hassan et al.'s work 43
2.3.2. Khalifa et al.'s work 44
2.3.3. Al-Saif and Markert's work 45

3. Our approach 47
Conclusion 49

Chapter 2: Manual Annotation For Arabic Discourse Analysis 50
Introduction 51
1. The data 51
2. Discourse segmentation manual 53
2.1. Annotation scheme 53
2.1.1. Basic principles 54
2.1.2. Main segmentation principles into clauses 55
2.1.3. Main segmentation principles into EDUs 59
2.2. Inter-annotators agreement study 63
3. Manual annotation of discourse relations 64
3.1. Arabic rhetoric 64
3.2. Building a new hierarchy of Arabic discourse relations 68
3.2.1. General methodology 68
3.2.2. A detailed description of our hierarchy 71
3.3. Annotation campaign 82
3.3.1. The corpus 83
3.3.2. Annotation procedure 83
3.4. Results 85
3.4.1. Qualitative analysis 85
3.4.2. Quantitative analysis 87
Conclusion 91

Chapter 3: Automatic Discourse Segmentation 92
Introduction 93
1. Related work 94
1.1. EDU segmentation: main approaches 94
1.2. Arabic EDU segmentation 95

2. Rule-based approach 97
2.1. The data 97
2.2. Proposed approach 97
2.3. Experiments and results 100
3. Learning approach 102
3.1. The data 102
3.2. Proposed approach 103
3.2.1. Punctuation features 103
3.2.2. Lexical features 104
3.2.3. Morphological features 104
3.2.4. Syntactic features 105
3.3. Experiments and results 106
3.3.1. Token boundary detection 106
3.3.2. EDU recognition 110
3.3.3. The learning curve 111
Conclusion 112

Chapter 4: Automatic Discourse Relation Recognition 113
Introduction 114
1. Related work 115
2. The features 117
2.1. Al-Saif et al.'s features 119
2.2. New features 120
3. Experiments and results 124
3.1. Overall results 125
3.2. Fine-grained classification 127
3.3. Mid-level classification 132
3.4. Coarse-grained classification 132
3.5. The learning curves 133
Conclusion 134

Chapter 5: Automatic text summarization using SDRT framework 135
Introduction 136
1. Related studies 138
1.1. Numerical and symbolical approaches 138
1.1.1. Numerical approaches 138
1.1.2. Symbolical approaches 141
1.2. Main studies for Arabic 144
2. The data 148
2.1. ADTB corpus 149
2.2. AD-RST corpus 150
3. Content selection algorithms 151
3.1. Tree-based content selection algorithm 151
3.2. Graph-based content selection algoritms and 152
3.2.1. Strict pruning 153
3.2.2. Easy pruning 155
4. Examples 156
4.1. Example from AD-RST corpus 156
4.2. Example from ADTB corpus 157
5. Experiments and results 158
Conclusion 162

General conclusion 163
References 166

List of Tables

Table 1.1. Taxonomy according to Khalifa et al. (2012) 45
Table 2.1. EST details 52
Table 2.2. Characteristics of our data in the gold standard 64
Table 2.3. SDRT relations in Annodis project. 69
Table 2.4. Characteristics of our gold corpus. 83
Table 2.5. Discourse relation distribution in the gold corpus 88
Table 2.6. Discourse relation and argument type in the gold corpus 89
Table 2.7. Some weak discourse connectives and the possible relations that can signal 90
Table 3.1. Training and test distrubition. 97
Table 3.2. Evaluation results of the rule-based approach. 100
Table 3.3. The gold standard corpus characteristics. 102
Table 3.4. Results of the baselines, (B1) and (B2); and the classifiers, (C1) and (C2) 108
Table 3.5. Results of the (C2) classifier with SAMA features on each class. 109
Table 3.6. Confusion matrix of the (C2) classifier on ADTB. 109
Table 3.7. Results of (C2) with SAMA features and (C3) with syntactic features. 110
Table 3.8. Accuracy of EDUs recognition before and after post-processing. 111
Table 4.1. Examples of concepts related by AWN relations. 123
Table 4.2. Overall results for the fine-grained classification. 126
Table 4.3. Overall results for the mid-class and coarse-grained classification. 127
Table 4.4. Detailed results for the mid-level classification (Level2). 132
Table 4.5. Detailed results for the top-level classification (Level1). 133
Table 4.6. Confusion matrix for the coarse-grained classification. 133
Table 5.1. ADTB characteristics. 149
Table 5.2. SDRT discourse relations selected for the summarization task. 149
Table 5.3. Rhetorical frame of the relation **تخصيص**/txSyS/*Specification*. 150
Table 5.4. AD-RST characteristics. 151
Table 5.5. RST discourse relations selected for the summarization task. 151
Table 5.6. Pre-treatment cases. 153
Table 5.7. Algorithm outputs. 157
Table 5.8. Algorithms outputs. 158
Table 5.9. The baseline results. 159
Table 5.10. The results of the proposed algorithms. 160

List of Figures

Figure 1.1. Example of DRS. 25
Figure 1.2. Hierarchy of RST relations. 27
Figure 1.3. Example of RST schema types 28
Figure 1.4. The RST analysis of the “car repair” example. 28
Figure 1.5. Example of an SDRT-graph. 31
Figure 1.6. Hierarchy of coherence relations used in GraphBank (Wolf *et al.*, 2003). 32
Figure 1.7. Graph representation for Example 9. 33
Figure 1.8. The PDTB relations. 35
Figure 1.9. The LADTB relations. 46
Figure 2.1. Morphological analysis of the two first words of Example 2. 53
Figure 2.2. Syntactic analysis of Example 2 as given by ATB manual annotations. 53
Figure 2.3. Figures of speech (**علم البيان**/Elm AlbyAn) in Arabic rhetoric (Abdul-Raof, 2012). 65
Figure 2.4. Word order (**علم المعاني**/Elm AlmEAny) in Arabic rhetoric (Abdul-Raof, 2012). 66
Figure 2.5. Semantic embellishment (**علم البديع** /Elm AlbdyE) in Arabic rhetoric. 67
Figure 2.6. Hierarchy of Arabic discourse relations used in the ADTB corpus 70
Figure 2.7. Right frontier principle.. 82
Figure 2.8. An example of a CDU constraint. 82
Figure 2.9. The discourse annotation for Example 85. 84
Figure 2.10.Two discourse annotations for Example 86 85
Figure 2.11. The distribution of our top-level classes according to their argument types. 89
Figure 3.1. A rule-based approach for discourse segmentation. 98
Figure 3.2. NooJ local sub-grammar for the dot marker. 98
Figure 3.4. NooJ local sub-grammar for DCs and punctuation marks patterns. 99
Figure 3.5. The XML output of our segmentation process. 99
Figure 3.6. The learning curve of (C2) for ADTB. 111
Figure 4.1. Discourse annotations for Example 1. 118
Figure 4.2. Feature impact on the **انشائي**/<n$A}y/*Thematic* relations in terms of F-score. 128
Figure 4.3. Feature impact on the **زمني**/zmny/*Temporal* relations in terms of F-score. 129
Figure 4.4. Feature impact on the **بنيوي**/bnywy/*Structural* relations in terms of F-score. 130
Figure 4.5. Feature impact on the **سببي**/sbby/*Causal* relations in terms of F-score. 131
Figure 4.6. The learning curve of our three level models. 133
Figure 5.1. RST tree for Example 1. 152
Figure 5.2. Example of a discourse structure. 154
Figure 5.3. Example of a discourse structure. 156
Figure 5.4. The discourse annotation for Example 2. 157
Figure 5.5. The discourse annotation for Example 4. 158

General Introduction

I. Context

I.1. Discourse processing

Within a discourse, texts and conversations are not just a juxtaposition of words and sentences. They are rather organized in a structure where discourse units are related to each other to ensure both discourse *coherence* and *cohesion*. Cohesion is defined as linguistic properties of a text that contribute to coherence (Halliday and Hasan, 1976). These properties include anaphoric expressions, the links between references, and lexical items occurring in sentences. Coherence on the other hand refers to the logical structure of discourse where every part of a text has a function and a role to play, with respect to other parts in the text (Webber *et al.*, 2012). Coherence has to do with semantic or pragmatic relations among units to produce the overall meaning of a discourse (Hobbs, 1979; Mann and Thompson, 1988; Grosz *et al.*, 1995). Identifying rhetorical relations is a crucial step in discourse processing. Given two discourse units that are deemed to be related, this step labels the attachment between the two units with discourse relations such as Elaboration, Explanation, Conditional, etc. as in *This is the best book that I have read in along time*, where the second clause introduced by "that" expands or elaborates on the first without giving additional information. Their triggering conditions rely on 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 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 Explanation relation between the two sentences, we need detailed lexical knowledge and probably domain knowledge as well.

Discourse structure is essential in determining the content conveyed by a text. It has shown to be useful for many NLP applications, such as automatic text summarization (Marcu, 2000a), information extraction (Vincent, 2010), automatic translation (Hardmeier, 2012), sentiment analysis (Chardon *et al.*, 2013) and question answering (Chai and Jin, 2004). The usefulness of discourse in NLP applications mainly depends on the availability of powerful discourse parsers. To build such parsers and improve their performances, several resources have been manually annotated with discourse information. These resources can be characterized according to four criteria: the underlying discourse theory (i.e. 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)), the data structure of the discourse (i.e. tree, graph or dependencies), the nature and the hierarchy of relations (i.e. semantic, intentional or lexically grounded) and finally the language. Most available resources are done in English. Recently, several efforts have been undertaken to develop manually annotated discourse information for other languages such as Chinese (Xue, 2005; Zhou and Xue, 2012), Danish (Buch-Kromann *et al.*, 2009; Buch-Kromann and Korzen, 2010), Dutch (Van der Vlieth *et al.*, 2011), Hindi (Oza *et al.*, 2009), Czech (Mladova *et al.*, 2008), Turkish (Zeyrek and Webber, 2008; Zeyrek *et al.*, 2009; Zeyrek *et al.*, 2010), and French (Danlos *et al.*, 2012). Surprisingly, discourse processing in Modern Standard Arabic (MSA) has received less attention despite the fact that MSA is a language with more than 422 million speakers in 22 countries[1].

I.2 Arabic natural language processing

Modern Standard Arabic (MSA) is the universal language of the Arab world. It is a modernized and standardized version of Classical Arabic used in writing and more formal settings, such as education and media. MSA has a complex linguistic structure with a rich morphology and a complex syntax (Al-Sughaiyer and Al-Kharashi, 2004; Ryding, 2005; Habash, 2010). It is mainly characterized by the lack of diacritics (dedicated letters to represent short vowels), complex agglutination, pro-drop structure, and free order word structure. These characteristics make Arabic processing more challenging. For instance, Farghaly and Senellart (2003) estimated that the average number of ambiguities for a token in MSA can reach 19.2 (compared to 2.3 in most other languages). These ambiguities are mainly due to the presence of particular morphological phenomena. Indeed, particles such as prepositions (e.g. ب/b/*by*/*with*[2]), conjunctions (e.g. و/w/*and*), and pronouns (e.g. هم/hm/*them*) can be affixed to words. For instance, the word وبسيارته/wbsyArth/*and by her car* is composed of the conjunction و/w/*and*, the preposition ب/b/*by*, the noun سيارة/sayArt/*car*, and the personal pronoun ه/h/*her*. Furthermore, the lack of vowels in current texts and the multiplicity of the vowel forms could make the analysis and the comprehension of Arabic texts more difficult. For example, the word فضل/fDl can be an Arabic person named entity or a conjunction ف/f/*then* followed by the verb ضل/Dl/*lost*.

Most researches on Arabic NLP resource generation have focused on morphology (Boudlal *et al.*, 2011), lexical semantics (Diab *et al.*, 2008) and syntactic analysis (Maamouri *et al.*, 2010b). There is also a huge literature on Arabic NLP including shallow and deep syntactic parsing (Belguith, 1999; Aloulou, 2005; Diab *et al.*, 2007; Diab *et al.*, 2009; Green and Manning, 2010; Ali Mohammed and Omar, 2011; Bahou, 2012; Marton *et al.*, 2013), morphology analysis (Eskander *et al.*, 2013; Sawalha *et al.*, 2013; Gridach and Chenfour, 2011), question answering (Benajiba *et al.*, 2012; Trigui *et al.*, 2014), automatic translation (Sadat and Mohamed, 2013;

[1] http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day/

[2] All Arabic examples in this thesis are extracted from our corpora. They are given in Arabic along with their English translation and their transliteration using Buckwalter 1.1: http://search.cpan.org/~graff/Encode-Buckwalter-1.1/

Carpuat *et al*., 2012), opinion mining and sentiment analysis (Abu-Jbara *et al*., 2013; Mourad and Darwish, 2013; Abdul-Mageed and Diab, 2012) and named entity recognition (Darwish, 2013; Aboaoga and Ab-Aziz, 2013; Boujelben *et al*., 2013). However, the field of Arabic NLP is still very vacant at the layer of discourse.

Among the few efforts, we cite (Mathkour *et al.*, 2008), (Khalifa *et al.*, 2012) and (Sadek *et al.*, 2012) within the RST framework as well as Al-Saif et al.'s approach within the PDTB model (Al-Saif and Markert, 2010). These studies proposed a two-steps algorithm for discourse analysis of Arabic texts: first discourse connective recognition by identifying the discourse and the non discourse usage of Arabic connectives linking adjacent discourse units, then discourse connective interpretation. Recently, Al-Saif and Markert (2010) proposed the Leeds Arabic Discourse Treebank (henceforth LADTB) the first resource for Arabic annotated with discourse information. LADTB extends the PDTB model to MSA. It provides a partial discourse structure of a text by focusing on explicit discourse connectives, annotation of their arguments as well as discourse relations that link adjacent arguments. This corpus has been used in (Al-Saif and Markert, 2011) to identify explicitly marked relations holding between adjacent arguments.

II. Contributions of the thesis

In this thesis, we propose to go 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. In particular, we propose the first effort towards a semantically driven approach to annotate Arabic texts with discourse information following the Segmented Discourse Representation Theory (SDRT). The annotation starts by segmenting documents into Elementary Discourse Units (EDUs) that have to be linked by discourse relations, to form Complex Discourse Units (CDUs), which in turn may be linked via discourse relations to other discourse units. The main contributions of this work are:

- A study that tackles the feasibility of building recursive and complete discourse structures of Arabic texts. In particular, we propose:
    - an annotation scheme for the full discourse coverage of Arabic texts, in which each constituent is linked to other constituents. A document is then represented by an oriented acyclic graph, which captures explicit and implicit relations as well as complex discourse phenomena, such as long-distance attachments, long- distance discourse pop-ups and crossed dependencies.
    - a novel discourse relation hierarchy. We study rhetorical relations from a semantic point of view by focusing on their effect on meaning and not on how they are lexically triggered by discourse connectives that are often ambiguous, especially in Arabic. Given our semantic-driven approach, we choose not to reuse the set of LADTB discourse relations. Instead, we start from the relations already defined within past

SDRT-like annotation campaigns (cf. Discor (Reese *et al.*, 2007) for English and Annodis (Muller *et al.*, 2012; Afantenos *et al.*, 2012) for French) and propose to refine them via a specialization/generalization process using both Arabic rhetoric literature (Abouenour *et al.*, 2012) and an examination of relations in the corpus. 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.

- a thorough quantitative analysis (in terms of relation frequencies, proportion of implicit relations, etc.) and qualitative analysis (inter-annotator agreements and error analysis) of the annotation campaign.

- An automatic discourse parser where we investigate both automatic segmentation of Arabic texts into elementary discourse units and automatic identification of explicit and implicit discourse relations.

- An application of our discourse parser to Arabic text summarization. We compare tree-based vs. graph-based discourse representations for producing indicative summaries and show that the full discourse coverage of a document is definitively a plus.

III. Outline of the thesis

The thesis is organized around five chapters.

Chapter 1 provides some backgrounds on discourse analysis, including the notions of discourse connectives, Elementary Discourse Unit (EDU), discourse structures and discourse relations. We then survey main theories of discourse including Discourse Representation Theory (DRT), Rhetorical Structure Theory (RST), Segmented Discourse Representation Theory (SDRT), GraphBank model and Penn Discourse TreeBank model (PDTB). The third part of this chapter provides an overview of the linguistic properties of the Arabic language as well as a presentation of the Arabic particularities at the discourse level. This chapter ends by an overview of main research work on Arabic discourse processing highlighting the main contributions of this work.

Chapter 2 discusses the discourse structure annotation scheme. The annotation requires three steps: (1) segmenting the document into EDUs, (2) attaching these units and (3) labelling the attachment by means of discourse relations. This chapter is composed of three main parts. The first one focuses on the first step above and presents a set of principles to guide the segmentation process. Two corpora that have different genre, audience and style of writing have been annotated according to this scheme: Elementary School Textbooks (EST) and newspaper documents extracted from the syntactically annotated Arabic Treebank (ATB) (Maamouri *et al.*, 2010a). We detail the characteristics of our data and present the inter-annotators agreement study conducted on the two corpora. The second part of the chapter is dedicated to the step (2) and the step (3). It presents the new hierarchy of discourse relations and the annotation scheme. We end

this part by giving quantitative and qualitative results of the annotation campaign that we conducted on the ATB corpus. The last part of this chapter presents our analysis of Arabic signalling devices used to trigger Arabic discourse relations. We detail in particular our lexicon that we built during the training stage of the annotation campaign. This work has been published in two papers: in the Transactions on Asian Language Information Processing (TALIP) for EDU annotation scheme (Keskes *et al.*, 2014a) and in a paper under revision at the Language Resources and Evaluation journal (LRE) for discourse structure annotation (Keskes *et al.*, 2015).

In Chapter 3, we propose two approaches to automatically identify EDUs: a rule-based and a learning based method. The first one implements most of the segmentation principles described in the last chapter using a set of dedicated rules to segment Arabic texts into clauses. Although the rules achieved relatively well, we noticed that their construction is very time consuming and that they fail to further segment a clause into EDUs. In the second step, we propose a set of features to automatically identify EDUs using a multi-class supervised learning approach that predicts EDUs as well as nested EDUs. We analyze the effect of shallow and extensive morphological features as well as the effect of chunks. We report on our experiments on boundary detection as well as on EDU recognition. We show that an extensive morphological analysis is crucial to achieve good results for both corpora. In addition, we show that adding chunks does not boost the performance of our classifier. This work has been published in four papers: in the International NooJ 2012 Conference (NooJ) (Keskes *et al.*, 2012a) and in the International Conference on Language Resources and Evaluation (LREC) (Keskes *et al.*, 2012b) for the rule-based approach, and in the Natural Language Processing (TALN) (Keskes *et al.*, 2013) and in the Transactions on Asian Language Information Processing (TALIP) (Keskes *et al.*, 2014a) for the learning approach.

In Chapter 4, we explore a wide range of features to automatically learn both explicit and implicit Arabic relations. Among these features, some have been successfully employed for explicit Arabic relation recognition such as *al-masdar*, connectives, time and negation, etc. (cf. (Al-Saif and Markert, 2011). However, others are novel for Arabic. They include contextual, lexical and lexico-semantic features such as argument position, semantic relations, word polarity, named entity, anaphora, modality, etc. We investigate how each feature contributes to the learning process. Finally, we compare our approach according to three baselines, which 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. This work has been published in the Journal of King Saud University Computer and Information Sciences (JKSU-CIS) (Keskes *et al.*, 2014b).

In Chapter 5, we show how the discourse parser described in Chapter 3 and Chapter 4 can be used in a practical NLP applications. We investigate automatic summarization and in particular a discourse-based approach to produce indicative summaries of Arabic documents. It consists in selecting the most relevant EDUs in the text according to three discursive criteria: the semantics

of discourse relations, their nature (coordinating vs. subordinating) and the document discourse structure (tree vs. graph). To measure the impact of discourse structure on producing indicative summaries, we evaluate our algorithms by comparing their performances against the gold standard summaries manually generated from two different corpora that have two different frameworks: ADTB, annotated according to the Segmented Discourse Representation Theory (SDRT) and the Arabic Discourse RST corpus (AD-RST) (Keskes *et al.*, 2012d), annotated according to the Rhetorical Structure Theory (RST). In each corpus, we perform two evaluation settings. The first one evaluates automatic content selection algorithms when inputs are given by gold standard discourse structures while the second one is an end-to-end evaluation that takes as input the outputs generated by the partial discourse parser described in the previous chapters. This work has been published in International Computing Conference in Arabic (ICCA) (Keskes *et al.*, 2012c).

Eventually, in Conclusion, we provide an overview of this work and emphasise its progresses and limitations. We also expose our perspectives for future work.

Chapter 1: Background

Table of contents

Introduction 21
1. Discourse analysis 21
   1.1. Basic notions 21
   1.2. Main discourse theories 25
2. Arabic discourse analysis 36
   2.1. Arabic specificities 36
   2.2. Arabic particularities at the discourse level 37
   2.3. Main studies on Arabic discourse processing 42
3. Our approach 47
Conclusion 49

Introduction

Discourse analysis is defined as the analysis of language "beyond the sentence". It takes into account the larger discourse context in order to understand how it affects the sentence meaning. In order to narrow down the range of possible meanings of discourse, some linguists have proposed different views and definitions, such as:

-"Discourse is written as well as spoken: each utterance assuming a speaker and a hearer as discourse." (Benvenisle, 1971)

-"An individualizable group of statements and sometimes as a regulated practices that counts for a number of statements." (Foucault, 1972)

-"Text analysis focuses on the structure of written language, as found in such text as essays, notices, road sings and chapters." (Cristal, 1987)

As a modern discipline, discourse analysis is an attempt to discover linguistic regularities in discourse using grammatical, phonological, and semantic criteria, such as cohesion, anaphora, inter sentence connectivity, etc. Moreover, discourse analysis is not just one approach, but also a series of interdisciplinary approaches that can be used to explore discourse coherence. Indeed, discourse analysis principles, assumptions, dimensions of analysis and methodologies (segments, markers, relations, etc.) can be changed when the corpus or the language are changed.

This chapter is organized around three parts. The first one introduces the necessary background about discourse analysis and defines the most important notions used throughout this dissertation. The second part presents an introduction to Arabic language processing focusing on Arabic specificities, Arabic particularities at the discourse level and an overview of main research work on Arabic discourse processing. The last part presents our approach and highlights its major contributions regarding related work.

1. Discourse analysis

In this section, we outline the basic notions related to discourse analysis and discourse processing. In particular, we provide a general definition of discourse connective, discourse unit, discourse structure, discourse relation, discourse cohesion and discourse coherence.

1.1. Basic notions

1.1.1. Discourse connectives

A discourse connective (DC) is a lexical item that relates two different abstract objects in discourse like events, states or propositions (e.g. *although, however, because, therefore, then* and *while*) (Asher, 1993). It can have several grammatical categories such as conjunctions (e.g. *and,*

or, *for* and *so*), subordinations (e.g. *as*, *like*, *than* and *if*), prepositional phrases (e.g. *about*, *after*, *before* and *except*) and adverbs (e.g. *soon*, *never*, *still*, *well* and *quite*). Various labels were used for lexical items with a similar or closed function of DCs: cue phrases (Knott and Dale, 1994), discourse connectives (Blakemore, 1992), discourse operators (Redeker, 1991), discourse particles (Schorup, 1985), discourse signaling devices (Polanyi and Scha, 1983), pragmatic connectives (Stubbs, 1983), discourse pragmatic markers (Fraser, 1988), semantic conjuncts (Quirk *et al.*, 1985) and sentence connectives (Halliday and Hasan, 1976). In the present study, we choose to use the term discourse connective (DC for short), as it is widely used in the discourse processing community.

A DC has three main basic functions:

- explicitly marks discourse relations that link parts of discourse. In Example 1, the DC بينما/bynmA/*while* marks the *Synchronisation* discourse relation,

  (1) [أحمد يهتم بالحديقة] [بينما سلمى ترتب البيت]

  [>Hmd yhtm bAlHdyqp] [bynmA slmY trtb Albyt]

  [Ahmed takes care of a garden] [while Salma arranges the house]

- contributes to discourse coherence,

- guides the discourse interpretation.

A DC can be used at the sentential level or at the level of larger textual units. In each level, discourse connectives can be ambiguous. Indeed, a DC can:

- has a discourse or a non discourse usage, i.e. a DC can trigger a discourse relation or not. In Example 2, the word و/w/*and* is a DC that marks the *Continuation* discourse relation, however in Example 3, it has a non discourse usage.

  (2) انتهت العطلة وبدأت الدراسة

  Antht AlETlp wbd>t AldrAsp

  The holidays ended and the study began

  (3) اجتمع المجلس البارحة على الساعة الرابعة والنصف مساء

  AjtmE Almjls AlbArHp ElY AlsAEp AlrAbEp wAlnSf msA'

  The council met yesterday at fourth hour and a half in the afternoon

- triggers one or several discourse relations. In Example 4, the DC ل/l/*to*/*because* marks the *Goal* discourse relation, however, in Example 5, it marks the *Explanation* discourse relation.

(4) [اضرب الباحثون] [ليُظهروا استياءهم]

[ADrb AlbAHvwn][lyuZhrwA AstyA'hm]

[The researchers are on strike][to show their dissatisfaction]

(5) [رجعت مسرعا إلى البيت] [لتهاطل الأمطار]

[rjEt msrEA <lY Albyt] [lthATl Al>mTAr]

[I returned quickly at home] [because it was raining]

1.1.2. Discourse units

Discourse Units (DUs) are non overlapping text spans that serve to build a discourse representation of a document. They can be clauses, sentences, paragraphs or dialogue turns. Defining DU boundaries is generally theory dependent since each theory defines its own specificities in terms of the segmentation guidelines and the size of units. For example, in the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), DUs are spans which are mainly delimited by discourse connectives and punctuations, as in *[Farmington police had to help control traffic recently] [when hundreds of people lined up to be among the first applying for jobs at the yet -to-open Marriott Hotel.]* where the sentence is segmented using the DC *when.* These spans are generally clauses called *nucleus* or *satellite* (see Section 1.2.2). In the Discourse Lexicalized Tree-Adjoining Grammar (DLTAG) (Webber, 2004; Riley *et al.*, 2006), DUs can be anchored by discourse connectives or can also remain lexically unrealized when DUs are adjacent clauses without DC such as *[Mary walked towards the car.][The door was open]* (see Section 1.2.4). In the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), DUs are Elementary Discourse Units (EDUs) and that are semantically represented in a Segmented Discourse Representation Structure (SDRS) (see Section 1.2.3). Roughly, the difference between spans and EDUs is in term of size and segmentation principles.

In the present study, an EDU is mainly a sentence or clause in a complex sentence that typically correspond to a verbal clause, as in *[I loved this movie]*$_a$ *[because the actors were great]*$_b$ where the relative clause introduced by the discourse connective *because*, indicates a cutting point. An EDU can also correspond to other syntactic units describing eventualities, such as prepositional and noun phrases, as in *[After several minutes,]*$_a$ [*we found the keys on the table]*$_b$. In addition, an EDU may be structurally embedded in another in order to encode adjuncts such as appositions or cleft constructions with discursive long-range effects such as frame

adverbials, non restrictive relatives and appositions, as in *[Mr. Dupont, [a rich business man,]*$_a$ *was savagely killed]*$_b$.

1.1.3. Discourse relations

A discourse relation (or rhetorical relation) is a description of how two DUs are logically connected to one another. In fact, discourse relations considered key for the ability to properly interpret or produce discourse and they referred to the semantic or pragmatic connections that bind one DU to another. These relations capture the hierarchical structure of a document and ensure its coherence such as *Elaboration*, *Explanation*, *Cause*, *Concession*, *Consequence*, *Condition*, etc. Their triggering conditions rely on elements of the propositional contents of the clauses, that is DCs. Discourse relations, based on the presence or absence of DMs, are divided into two groups: explicit (also called signalled) and implicit (also called unsignalled) relations. To infer the implicit relation between the clauses, we need detailed lexical knowledge and probably domain knowledge as well.

In the present study, discourse relations are both explicit and implicit relations that link adjacent or non adjacent discourse units, to form complex discourse unit, which in turn may be linked via discourse relations to other discourse units or complex discourse units. We study discourse relations from a semantic point of view by focusing on their effect on meaning and not on how they are lexically triggered by discourse connectives that are often ambiguous.

1.1.4. Discourse structures

Like DUs, discourse structure is generally theory dependent since each theory defines its own specific structure. Main discourse theories are: the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) in which the discourse structure of a document is a tree where leafs (called nucleus and satellite) are contiguous arguments and edges are rhetorical relations, the Discourse Lexicalized Tree-Adjoining Grammar (DLTAG) (Webber, 2004; Riley *et al.*, 2006) where the discourse structure is created by a composition of arguments anchored by discourse connectives, and the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) where the discourse structure is a graph, since two EDUs can be linked by more than one discourse relation.

In the present study, we focus on building a directed graph where nodes represent discourse segments or groups of discourse segments, and labeled directed arcs represent coherence relations holding between nodes.

1.1.5. Discourse cohesion and discourse coherence

Discourse theories hypothesize that discourse is coherence, that is to say, they assume that the different constituent parts of discourse are dependent from others, and it is possible to establish links between them. These theories seek to explain why certain discourses are seen as consistent

and others as inconsistent. Coherence refers to the logical structure of discourse where every part of a text has a function, a role to play, with respect to other parts in the text (Taboada and Mann, 2006). Coherence has to do with semantic or pragmatic relations among units to produce the overall meaning of a discourse (Hobbs, 1979; Mann and Thompson, 1988; Grosz *et al*., 1995).

Concerning cohesion, it is defined as linguistic properties of text that contribute to coherence (Halliday and Hasan, 1976). It groups the grammatical and lexical relationships that exist between parts of a discourse. These properties include anaphoric expressions, links between references, and lexical items occurring in sentences.

Within a discourse structure, discourse units are related to each other to ensure both discourse cohesion and coherence.

1.2. Main discourse theories

In this section, we present the main existing discourse theories that tend to represent the discourse structure of a text.

1.2.1. Discourse Representation Theory

Starting with the mediation of the discourse anaphora by discourse referents, Kamp and Reyle developed the Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993) which has been designed specifically to deal with the two-way interaction between utterance and context. The connection between information and truth is of paramount importance and they are the crucial ingredients. Based on explicit semantic representations (instead of working with first-order formula syntax), called Discourse Representation Structure (DRS), DRT approach describes the objects mentioned in a discourse and their properties and uses a new discourse processing method to deal with discourse anaphora. For example, Figure 1.1 represents the DRS of "*Peter moves. He speaks.".*

X Y
Peter (X) Moves (X) Y=X Speak (Y)

Figure 1.1. Example of DRS.

In Figure 1.1, a DRS is presented as a box-like structure, with so-called *discourse referents* in the box top part and *conditions* upon these discourse referents in the box lower part. The discourse referents are variables representing all the entities in the DRS. The conditions are the logical statements about these entities. There are two discourse referents in this example (X and Y), denoting "Peter" and "He", respectively. Discourse referents are entities mentioned in the discourse to which pronouns potentially can refer. In our example, an anaphoric link has been established between "He" and "Peter" by virtue of the condition Y=X.

In DRT, interpretation is involved into two main steps: first, the construction of semantic representation, referred to as Discourse Representation Structures (DRSs) (cf. Figure 1.1), from the input discourse and, second, a model-theoretic interpretation of those DRSs. We can represent these two steps as follow:

Discourse -> DRS -> interpretation

The dynamic part of meaning resides in how the representations of new segments of discourse are integrated into the representation of the already processed discourse and what effect this has on the integration of the representations of subsequent, further segments of discourse.

A new version of DRT architecture was proposed by Van Der Sandt and Geurts (Van Der Sandt, 1992; Geurts and Van Der Sandt, 1999; Geurts, 1999; Kamp, 2001a; Kamp, 2001b), based on a general treatment of presupposition (Soames, 1984): a presupposition is a requirement which a sentence imposes on the context in which it is used. In case the context does not satisfy the presuppositions imposed by the sentence, presuppositions are modified or updated to a new context, which does satisfy them. This new version construction proceeds bottom-up: the representations are constructed from syntactic trees by assigning semantic representations to the leaves of the tree and then building representations for complex constituents by combining the representations of their immediate syntactic parts (Kamp *et al.*, 2011).

1.2.2. Rhetorical Structure Theory

The Rhetorical Structure Theory (RST) is a theory of discourse organization by means of discourse relations that hold between text segments. It was created by Mann and Thompson on 1988 for text summarization purposes. This theory has been greatly used in descriptive linguistics, computational linguistics and NLP (from text analysis to text generation). RST focuses on a rhetorical analysis, which aims at structuring the text using semantic relations and intentional relations between the discourse units of the text. These rhetorical relations can be described in terms of the purposes of the writer and its assumptions about the reader. The identification of relations between larger segments of texts yielded a natural hierarchical description of the rhetorical organization of the text. For RST, it is required to segment firstly the text into spans (discourse units), which then become the minimal elements of the analysis. This segmentation is carried out in a simple way, one intended to be as neutral as possible in influencing the analysis process. A span can have nucleus statue – primordial segment for text coherence – or it can have satellite statue – optional segment for text coherence. The most common type of relation is nucleus-satellite relation where the first span is a nucleus and the second span is a satellite. Four components are defined in RST for describing text structures: *Relations*, *Schemas*, *Schema applications* and *Structures*.

- *Relations*: Relations hold between two non overlapping nucleus or/and satellite spans. In case all spans are nuclei, the relation is multinuclear. RST defines a set of twenty-

three rhetorical relations that link two spans. The hierarchy of the rhetorical relations is presented in Figure 1.2.

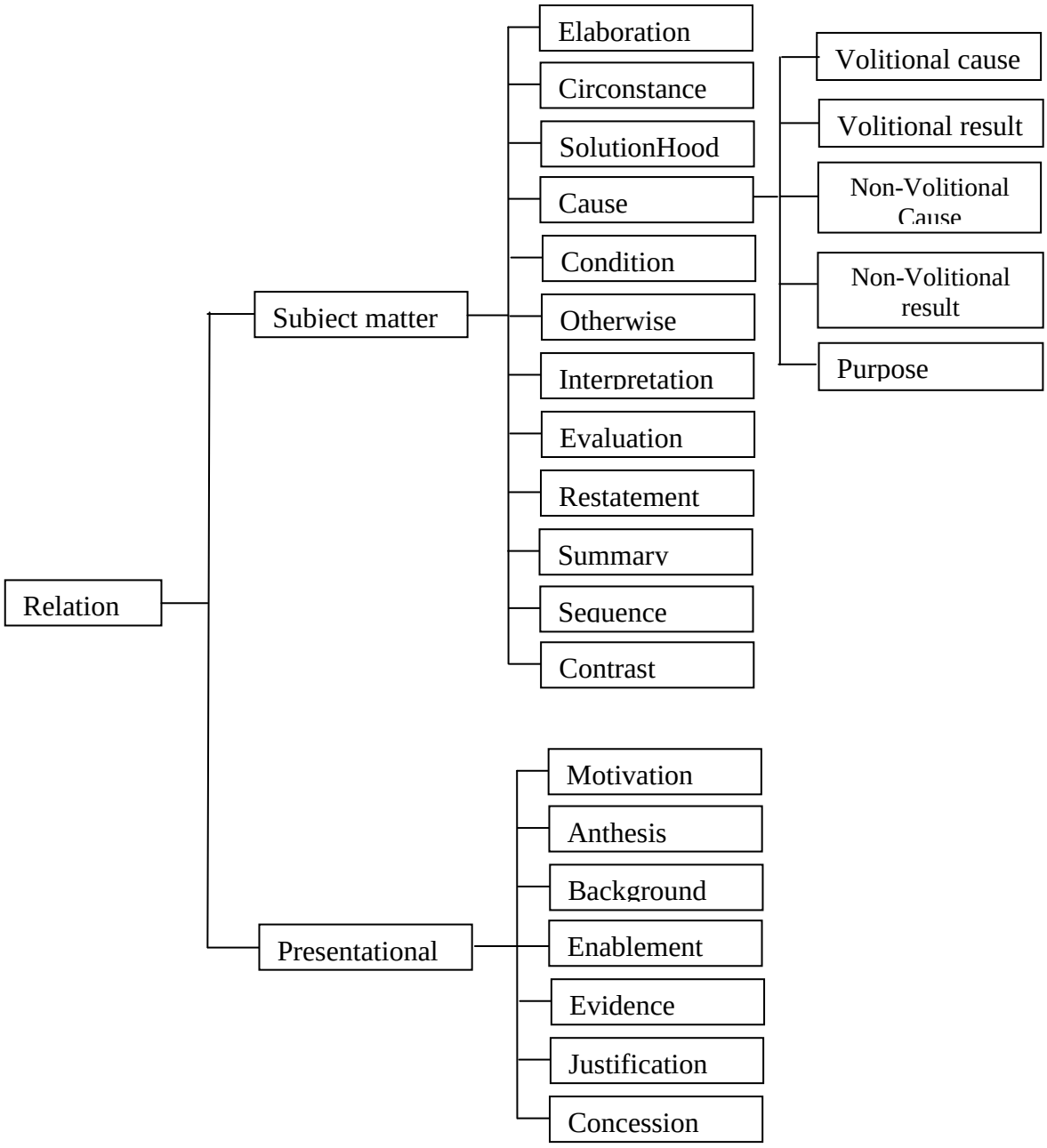

Figure 1.2. Hierarchy of RST relations.

RST relations are applied recursively in a text, until all parts of the text are constituents in an RST relation. The result of such analyses is that RST structure are typically represented as trees, with one top level relation that encompasses other relations at lower levels.

- *Schemas*: RST represents the rhetorical organization of the text using rhetorical structure schemas, which obey constraints of completeness. Each schema indicates how a particular span of text is analyzed in terms of other spans. Conceptually, these Schemas are the basic organizational building blocks of the theory. They are considered to be abstract patterns of text structure comprising a small number of constituent text spans. Given a text, RST determines the possible trees by providing specifications about what relations hold between text spans, and how certain spans are related to the whole collection. There are five types of schema in RST, which are represented by the five diagrams in Figure 1.3.

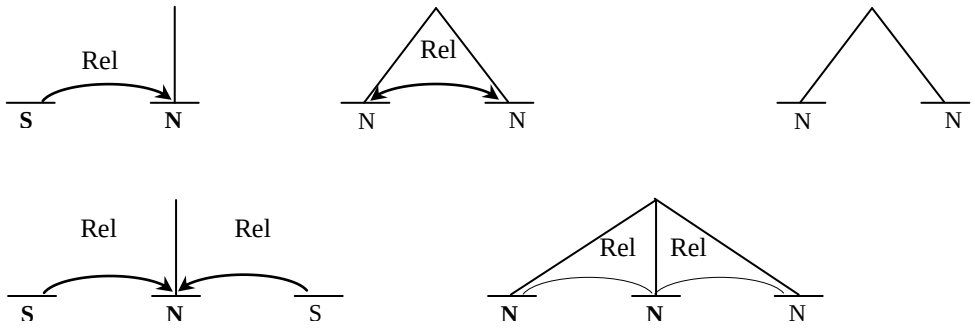

Figure 1.3. Example of RST schema types

To illustrate this component, we can refer to the example "car repair" cited in (Mann and Thompson, 1988), which presents the relation background between (A) and (B):
(A)*I am having my car repaired in Santa Monica (1522 Lincoln Blvd.) this Thursday 19th.*
(B)*Would anyone be able to bring me to ISI from there in the morning or drop me back there by 5 pm please?*
The RST analysis of this example is shown in the figure 1.4:

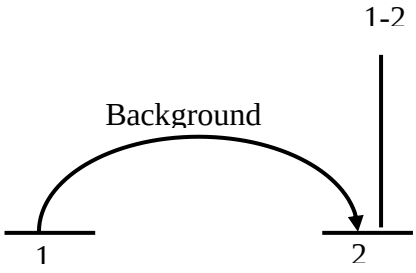

Figure 1.4. The RST analysis of the "car repair" example.

- *Schema Applications*: The schema applications define the ways a schema can be instantiated using several conventions. Three conventions are used to determine the possible application of a schema. Conventions include *unordered spans* where the order of the nucleus and satellite spans is not constrained by the schema, *optional*

relations where all individual relations are considered to be optional, but at least one relation among them must hold (the case of multi-relational schemas), and *repeated relations* where the relation of a schema can be applied many times in the text structure by the application of this schema.

- *Structures*: The composition of the schema applications determines the structure of an entire text. A structural analysis of a text is a set of schema applications which is determined by four constraints: *completedness* where the set includes one schema application representing all text spans, *connectedness* where each span in the analysis is either a discourse unit or a constituent of another schema application of the analysis, *uniqueness* where each schema application is characterized by a different set of spans and *adjacency* where spans of each schema application constitute one larger text span.

Based on RST, many available resources were developed. The RST Discourse Treebank (RST-DT) (Carlson *et al*., 2003) built on the top of the syntactically annotated Penn Treebank (Marcus *et al*., 1993) represents one of the well-known RST resources for English. Relations in RST-DT are grouped into 16 classes, which are further specified into 78 relations, organized by nuclearity (nucleus-satellite or multinuclear rhetorical relations). Similar efforts have been done for building RST-based corpora for German (Stede, 2004), Dutch (Van der Vlieth *et al*., 2011), Portuguese (Pardo *et al*., 2004) and Spanish (Da Cunha *et al*., 2010). We finally note marginal efforts for Arabic (Mohamed and Omer, 1999), Finnish (Sarjala, 1994), and Russian (Sharoff and Sokolova, 1995) with, to our knowledge, no information neither on the availability of these corpora nor on their associated annotation scheme.

1.2.3. Segmented Discourse Representation Theory

The Segmented Discourse Representation Theory (SDRT), developed by (Asher and Lascarides, 2003), 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 as a discourse structure in which every EDU of T is linked to some (other) discourse units, 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 representations to capture complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups[3], as well as crossed dependencies[4] (Wolf and Gibson, 2006) (Danlos, 2007).

SDRT models discourse coherence via defaults and non monotonic reasoning. Annotations in SDRT start from Elementary Discourse Units (EDU), and define hierarchical structures by constructing complex segments (CDUs) from EDUs in recursive fashion. However, SDRT goes beyond adjacent discourse units allowing for the creation of a directed acyclic graph which captures complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups, as well as crossed dependencies, etc. The discourse structure has multiple parented nodes and crossing arcs, which allow to adequately represent discourse structure (Danlos, 2007) (Wolf and Gibson, 2005). To illustrate the importance of such representation, let us consider the following examples in (RST) and (Annodis) taken respectively from the RST TreeBank corpus (Carlson *et al.*, 2003) and the Annodis corpus (Afantenos *et al.*, 2012), discussed in (Venant *et al.*, 2013):

(RST)*[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).

(Annodis)*[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 (RST) 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). 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 attachment using

[3] 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).

[4] 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).

the immediate interpretation of RST trees[5]. (RST), for instance, also involves an *Explanation* relation between 31 and 32, which 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.

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 segments (more details on these constraints are given in Chapter 2). Figure 1.5 gives an example of the discourse structure of Example 6, 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.

(6) [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.]$_5$

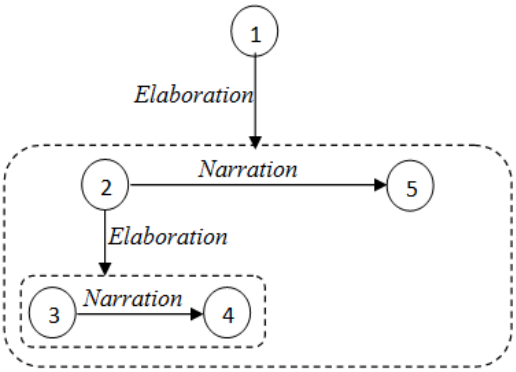

Figure 1.5. Example of an SDRT-graph.

Two main corpora have been developed following SDRT principles: The Discor corpus for English (Reese *et al.*, 2007) and the Annodis[6] corpus for French (Afantenos *et al.*, 2012). The Discor corpus analyzes the interaction between discourse structure and co-reference resolution. This project annotates 60 texts from the MUC 6 and MUC 7 data sets where only experts performed the annotation in the theory. The Annodis corpus combined two perspectives on discourse: a bottom-up view that incrementally builds a 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 short Wikipedia articles as well as news articles with a total of 3,199 EDUs and 3,355 relations. Both naïve and experts were involved in the annotation

[5] 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.

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

campaign. We finally cite efforts for adapting SDRT to Mandarin (Jiun-Shiung, 2005). As far as we know, this work did not provide any available annotated corpora.

1.2.4. GraphBank model

The discourse GraphBank (Wolf *et al.*, 2003) is a model with a less-constrained annotation protocol. Wolf and Gibson motivated from an empirical linguistic perspective. Humans annotate all discourse relations in a text using a protocol that imposes no structural constraints on the representations to estimate empirically the degree to which trees (or graphs) are adequate representations of discourse structures. The authors encourage annotators to make explicit all coherence relations that hold between any two discourse units in a text. When they apply this annotation protocol on a large collection of texts, they observe that the discourse structures that are created in this manner look more like graphs than like trees. Because the links in the resulting graphs cross often, their results strongly suggest that trees are an inadequate representation for discourse structures. On the bases of their corpus analysis, Wolf and Gibson estimate that in order to obtain tree representations from the graph representations in their corpus, one would have to delete approx, which are 12% of the coherence relations identified by the annotators. This process loses important information.

The discourse GraphBank collects a database of texts annotated with coherence relations. The data is composed of 135 news articles from AP Newswire and Wall Street Journal, annotated with hierarchy of coherence relations presented in Figure 1.6.

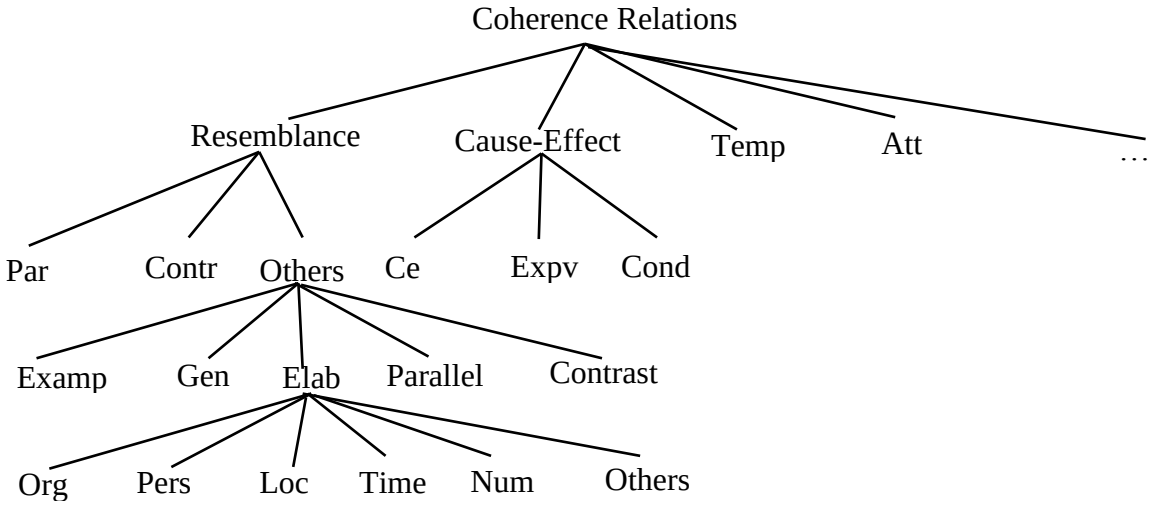

Figure 1.6. Hierarchy of coherence relations used in GraphBank (Wolf *et al.*, 2003).

As illustrated in Figure 1.6, we can mention the resemblance relation that presents the contrast and commonalities between discourse segments. This class includes three sub-relations such as *parallel*, *contrast* and *others*. The parallel relation is symmetrical and infers a set of entities from discourse segments, as in Example 7.

(7) *[John organized rallies for Clinton,][and Fred distributed pamphlets for him.]* (Example extracted from (Wolf and Gibson, 2005))

Also, the contrast relation is symmetrical and infers contrast between members of discourse segments, as in Example 8.

(8) *[John supported Clinton,][but Mary opposed him.]*

The resemblance relation includes also other relations like *elaboration*, example, *generalization*, etc. Borisova and Redeker (2010) have investigated the use of the relation "same" in the Discourse Graphbank (Wolf *et al.*, 2003) that connects the parts of a discontinuous discourse segment.

The main goal of the discourse Graphbank was to define a descriptively adequate data structure for representing discourse coherence structures. The best discourse structure is a graph, rather than a tree. The GraphBank represents a significant advance in corpus-based investigation of discourse coherence structure. Wolf and Gibson investigated the impact of discourse coherence structures on other linguistic processes and natural language applications (e.g. anaphora resolution, automatic summarization and information retrieval), and developed and tested discourse parsing algorithms. Authors showed that tree structures are inadequate to represent discourse coherence structure.

Although GraphBank was adequate to establish different classes of coherent relations (such as causal, elaborative, temporal, intentional relations), this model does not take into account the role of lexical discourse markers, discourse segments, co-reference, entities, and events. Example 9 is extracted from the GraphBank, with the discourse structure shown in Figure 1.7.

(9) 1. *Farm prices in October edged up 0.7% from September*
2. *as raw milk prices continued to rise,*
3. *the Agriculture Department said.*
4. *Milk sold to the nation's dairy plants and dealers averaged $14.50 for each hundred pounds,*
5. *up 50 percent from September and up $1.50 from October 1988,*
6. *the department said.*

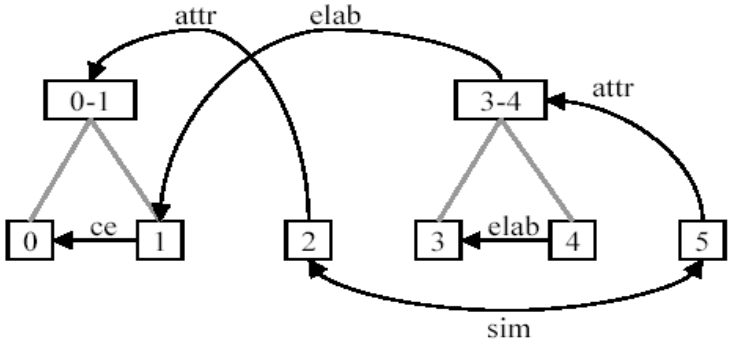

Figure 1.7. Graph representation for Example 9.

Annotations in the Discourse GraphBank differs significantly from other resources since annotators were asked to annotate all discourse relations that could be taken to hold between a discourse segment and any segment to its left. Moreover, GraphBank assumes that the discourse structure of a text is a directed graph where nodes represent discourse segments or groups of discourse segments, and labeled directed arcs represent coherence relations holding between nodes. However, no structural constraints are imposed on the resulting graphs (such as the right frontier principles), which makes the Graph Bank discourse structure one of the most complex.

1.2.5. Penn Discourse TreeBank model

In the Penn Discourse TreeBank (PDTB) (Webber *et al.*, 2006), the identification of discourse structure is approached independently of any linguistic theory by using discourse connectives rather than abstract rhetorical relations. PDTB assumes that connectives are binary discourse level predicates conveying a semantic relationship between two abstract object-denoting arguments. The set of semantic relationships can be established at different levels of granularity, depending on the application. The annotation in PDTB requires three main steps: identifying discourse connectives, identifying the locations of their two arguments Arg1 and Agr2, and labeling their extent. Arg1 can be located within the same sentence as the connective or in some previous sentences of the connective. PDTB follows a lexically-grounded approach to the annotation of discourse relations (Webber *et al.*, 2003). Discourse relations, when realized explicitly in the text, are annotated by discourse connectives - expressing them, thus supporting their automatic identification. For example, the causal relation in (10) is annotated by marking the discourse connective as a result as the expression of the relation.

(10) *U.S. Trust, a 136-year-old institution that is one of the earliest high-net worth banks in the U.S., has faced intensifying competition from other firms that have established, and heavily promoted, private-banking businesses of their own. As a result, **U.S. Trust's earnings have been hurt.***

PDTB adopts a theory-neutral approach, which makes no commitments to what kinds of high-level structures may be created from the low-level annotations of relations and their arguments. Using this approach, the annotated corpora can be used within different frameworks and provided a resource to validate the various existing theories of discourse structure. This theory neutrality represents the interaction between the structure at the sentence level and the structure at the discourse level (Lee *et al.*, 2006). Additionally, PDTB provides sense labels for each relation following a hierarchical classification scheme. Annotation of senses highlights the polysemy of connectives, making PDTB useful for sense disambiguation tasks (Miltsakaki *et al.*, 2005). Figure 1.8 presents the PDTB relations, which group relations into a taxonomy of 16 relations at the middle level and 4 coarse top-level classes (*Temporal*, *Contingency*, *Comparison* and *Expansion*) for a total of 33 relations.

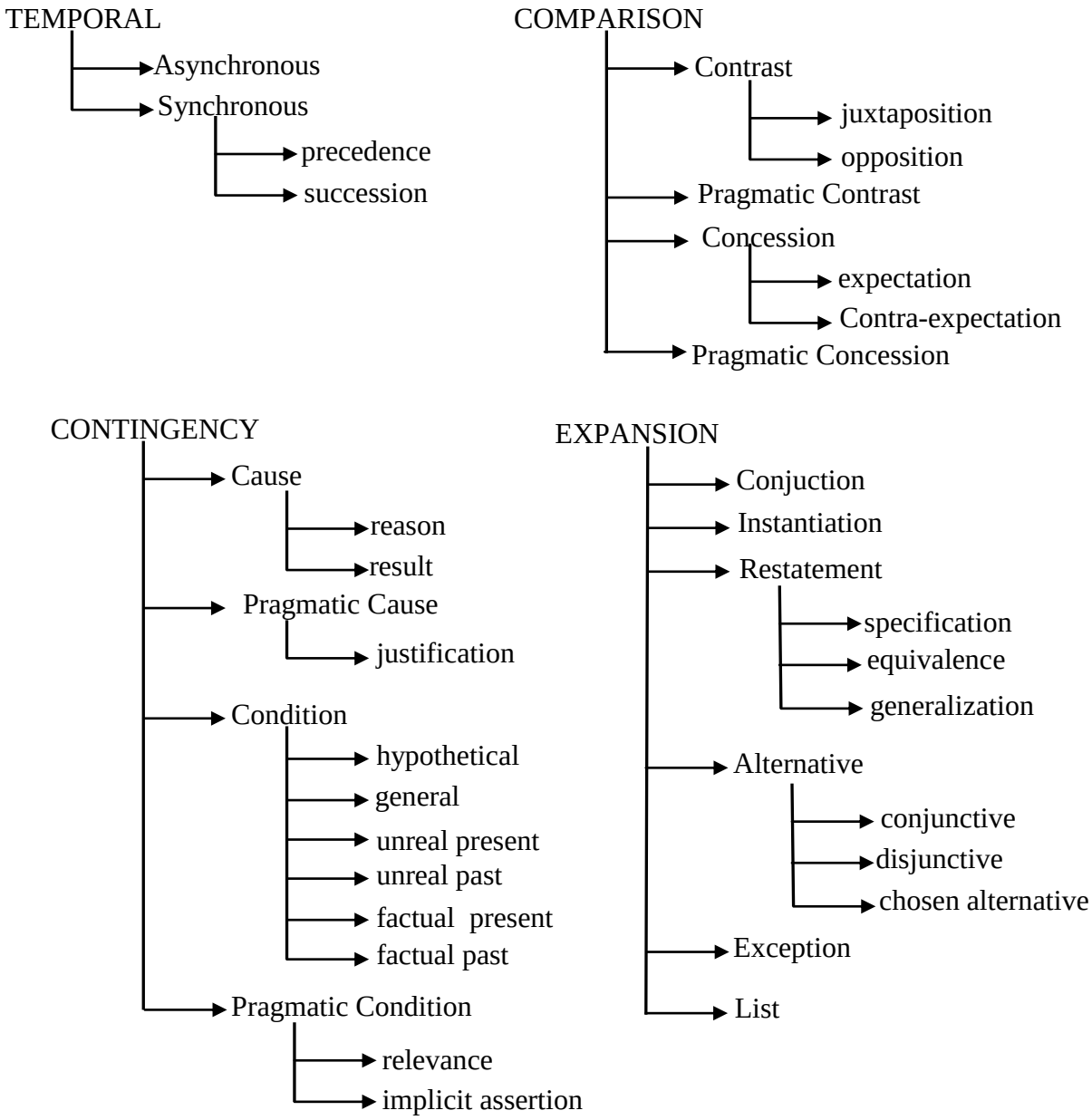

Figure 1.8. The PDTB relations.

Discourse relations in PDTB are regrouped into two types depending on how the relations are signalled in text: "explicit" relations that are signaled by discourse connectives, as a result in Example 10 (Arguments of Explicit connectives are unconstrained in terms of their location, and can be found anywhere in the text) and "implicit" relations that link two adjacent sentences in the absence of an explicit connective. In all cases, discourse relations are assumed to hold between two and only two arguments. Because there are no generally accepted abstract semantic categories for classifying the arguments to discourse relations as have been suggested for verbs

(e.g. agent, patient, theme, etc.), the two arguments to a connective are simply labelled Arg1 and Arg2. In the case of explicit connectives, Arg2 (which is in bold in Example 10) is the argument to which the connective is syntactically bound, and Arg1 is the other argument. In the case of relations between adjacent sentences, Arg1 and Arg2 reflect the linear order of the arguments, with Arg1 before Arg2.

PTDB corpora are available for English PDTB (Prasad *et al.*, 2008), Chinese (Xue, 2005; Zhou and Xue, 2012), Danish (Buch-Kromann *et al.*, 2009; Buch-Kromann and Korzen, 2010), Dutch (van der Vliet *et al.*, 2011), Hindi (Oza *et al.*, 2009), Czech (Mladova *et al.*, 2008), Turkish (Zeyrek and Webber, 2008; Zeyrek *et al.*, 2009; Zeyrek *et al.*, 2010), Modern Standard Arabic (Al-Saif and Markert, 2010), and French[7] (Danlos *et al.*, 2012).

2. Arabic discourse analysis

We give in this section a brief overview of MSA specificities. For a more detailed description of MSA and Arabic Natural Language Processing (ANLP), see (Habash, 2010). Then, we introduce Arabic discourse connectives and main studies on Arabic discourse analysis.

2.1. Arabic specificities

Arabic does not have capital letters and punctuation marks are not widely used in current Arabic texts (at least not regularly). Moreover, Arabic discourse tends to use long and complex sentences. We can easily find too long paragraph with only one punctuation at the end (e.g. dot).

As a Semitic language, Arabic has a rich morphology. Indeed, in addition to a *concatenative morphology*, where words are formed via a sequential concatenation process, Arabic is characterized by the presence of a *templatic morphology* where a templatic morpheme is composed of a root (a sequence of (mostly) three, (less so) four, or very rarely five consonants), patterns (an abstract template in which roots and vocalisms are inserted) and vocalisms that specify the short vowels to use with a pattern. For example the word stem كَتَب/katab/*to write* is constructed from the root ك - ت - ب/k-t-b, the pattern *1V2V3* and the vocalism *aa* (Habash, 2010). Concatenative morphemes can be stems, affixes or clitics. A clitic has the syntactic characteristics of a word but depends phonologically on another word or phrase. Clitics include prepositions, conjunctions, and pronouns. For instance, prepositions (such as ف/f/*then*), conjunctions (such as و/w/*and*), articles (such as ال/Al/*the*) and pronouns (such as ه/h/*he*) can be affixed to nouns, adjectives, particles and verbs, which causes several lexical ambiguities. Here are some examples:

[7] The French Discourse Treebank methodology differs in at least two points from the initial PDTB guidelines: it aims at providing a full coverage of a text and uses a new hierarchy of discourse relations, which is based on RST, SDRT and PDTB.

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

—The word وليد/wlyd can be a person name (Waleed), an adverb that means « derived-from » or the composition و/w/*and* + ل /li/*for* + the noun يد/yd/*hand*.

—The word فضل/fDl can be a person name (Fadhl) or the preposition ف/f/*then* followed by the verb ضل/Dl/*lost*.

Moreover, complex word structures are ambiguous. For instance, the word استتذكرونها/Astt*krwnhA ([اس/As/*will*], [تتذكرون/tt*krwn/*you remember*], and [ها/hA/*her*]) represents in English "will you remember her?"

Another specificity of Arabic is that word order is fairly flexible. Indeed, the change of certain position of words does not change the meaning of the sentence. For example the sentence "the child goes to the school" can be written in Arabic in three forms: ذهب الولد إلى المدرسة/ *hb Alwld <lY Almdrsp, الولد ذهب إلى المدرسة/ Alwld *hb <lY Almdrsp and إلى المدرسة ذهب الولد/ <lY Almdrsp *hb Alwld. Note that each form begins with the constituent (i.e. the verb, the subject or the object) to be shared on.

Finally, the most important specificity challenge in ANLP is diacritics. Arabic has 28 consonants, which may be interleaved with different long and short vowels. Short vowels are not often explicitly marked in writing. Indeed, they are typically not written in the Arabic handwriting of everyday use and in general publications. Diacritics represent, among other things, short vowels. Arabic texts can be fully diacritized, partially diacritized, or non diacritized. It should be noted that non diacritized texts are highly ambiguous. For example, the word ثمن/vmn/*price* can be diacritized in 22 different forms. The same confusion holds between the verb ذَهَبَ/*ahaba/*go* and the noun ذَهَبٌ/*NhabN/*gold*. Thus, a non diacritized word could have different morphological features, and in some cases, different POS, especially when it is taken out of its context. In addition, even if the context is considered, the POS and the morphological features could remain ambiguous.

2.2. Arabic particularities at the discourse level

2.2.1. General specificities

According to Koch (1983) and Ostler (1987), Arabic writings are characterized by repetition, balance, and coordination. Compared to other languages, Arabic writers prefer coordination at the expense of subordination with an extensive use of coordination particles (such as و /w/*and* and ف/f/*then*) (Othman, 2004). For instance, Reid (1992) compared 768 essays written in English by Arabic, Chinese, Spanish, and English native speakers in order to determine whether these essays differ in terms of cohesion devices. He found that Arabic writers used significantly more

coordinate conjunctions than the other three languages. The abundance of coordination in written Arabic texts makes short sentences very rare to exist. Arab writers tend to write very long sentences, some of which could be a paragraph long with one full stop at the end.

A second specificity is that Arabic has neither capitalization nor strict rules for punctuation. This can make tasks such as clause boundary detection and named entity recognition more difficult, as shown in Example 11 where the word فضل/fdl indicates a person name ("Fadhl"). The same word can also correspond to the verb "to prefer" or the conjunction ف/f/*then* followed by the verb ضل/dl/*to lost*. This last case can lead a discourse segmenter to consider the word فضل /fdl as the beginning of an elementary discourse unit since the conjunction ف/f/*then* is a good indicator for the discourse relations *Result*, *Narration* and *Continuation*.

(11) استقبلت عائلة مصطفى فضل البارحة.

Astqblt EA}lp mSTfY fDl AlbArHp.

I received Mustapha Fadhl's family yesterday.

2.2.2. Arabic discourse connectives

In Arabic, DCs and their role in discourse interpretation do not receive a great attention in the literature. The most studied Arabic DC is probably the particle و/w/*and* (Cantarino, 1975; Wright, 1975; Fareh and Hamdan, 1999). Historically, the coordination و/w/*and* was addressed by Abd Al-Kader Al-Jarjeni, a well-known Arabic linguist who identified six different rhetorical senses on the basis of rules called "Fasl and Wasl"' which mean "identifying segmentation places in a text" (Hemeida, 1997). "Fasl" signals a discursive function, as in Example 12 where the second DC و/w/*and* triggers the relation *Continuation* while "Wasl" aims at connecting units together without any specific discursive usage such as to express oaths or accompaniment. For example the first و/w/*and* in Example 12 has non discourse usage. "Fasl and Wasl" rules have been used in (Khalifa *et al.*, 2011) to automatically segment Arabic discourse into clauses. Authors classified the six meaning of connective و/w ((a) والقسم /w Alqsm) that means testimony, (b) ورب /w rb that means few or someone, (c) والاستئناف /wAlAst}nAf that simply joins two unrelated sentences, (d) والحال/w AlHAl that introduces a state, (e) والمعية /w AlmEyp that means the accompaniment, and (f) والعطف/w AlETf that means the conjunction of related words or sentences) into two classes : *fasl* which is a good indicator to begin of segment (contains (a), (b) and (c)) and *wasl* which has no effect on segmentation (contains (d), (e) and (f)).

(12)أكل أحمد و أكرم التفاحة وخرجا.

>kl >Hmd w >krm AltfAHp wxrjA.

Ahmed and Akram eat an apple and went out.

In the same context, Taha et al. (2013) studied the discursive functions of the connective (و/w/*and*). The authors oriented their study towards 19 rhetoric functions of this connective. (Salman, 2003) classified this connective into several classes: *concessive discourse marker*, *additive discourse marker*, *intrasententially-connecting concessive discourse marker*, *introductory discourse marker*, *ending marker* (marks the end of the speech), etc.

There are little discussions of other DCs in Arabic. Among the few studies, Alansari (2003) focused on the connective ب/b/*by* and showed that it can have only one discourse usage among 14 different rhetorical functions. Hussein (2008b) studied the connector ف/f/*then* within the relevance theory framework. Ryding (2005) analysed DCs connecting clauses within a sentence such as بل/bl/*rather* while Hussein (2008a) and Alhuqbani (2013) focused on Arabic contrastive DCs such as لكن/lkn/*but* and بينما/bynmA/*when*/*whereas*/*w*hile.

Some other studies have established specific empirical studies for Arabic discourse connectives. For instance, Alhuqbani (2013) and Hussein (2008a) studied the connective "but" and its translated forms in Arabic. Alhuqbani (2013) uses a judgment test which is done on 48 examples of the connective "but" that is made by *Arabic-English speaking informants* and 5 *English native informants*. This connective has four possible translations in Arabic: بل/bl, بينما/bynmA, لكنَ/lkna, and لكن/lkn). Obtained results show that the connective لكن/lkn has the same discursive functions of the connective "but" (contrastive discourse functions: correction, contrast, denial of expectation and cancellation). Nevertheless, only the connective لكن/lkn includes all the discourse functions of the connective "but" (so exact equivalence). None of connectives (بل/bl (correction), بينما/bynmA (contrast), لكنَ/lkna (expectation) include all these discourse functions.

In the same context, another studies of Chaalal (2010) demonstrated the difficulty to translated Arabic discourse connectives to the English ones. Indeed, the connective (ف/f) can have five discourse functions and then five possible translations:

- Sequential (then): ذهبت الى بغداد فالبصرة/*hbt AlY bgdAd fAlbSrp/*I went to Baghdad then to Basra*.
- Result (so): أحب أحمد المسرح فأبدع فيه/>Hb >Hmd AlmsrH f>bdE fyh/*Ahmad loved theatre and so he excelled in it*.
- Causal (because): لا تبكي فإن البكاء ضعف/lA tbky f<n AlbkA' DEf/*Do not cry because crying is weakness*.
- Explanation (For example): هناك أخطاء تاريخية كثيرة في المسلسل. فإغتيال الملك كان طعنا و ليس سما. / hnAk >xTA' tAryxyp kvyrp fy Almslsl. f<gtyAl Almlk kAn TEnA w lys smA. /*There are various historical mistakes in the series that should have been checked. For example, the king was stabbed not poisoned.*

- Contrast (but): دعاني صديقي فلم أجب دعوته/dEAny Sdyqy flm >jb dEwth/*My friend invited me to visit him, but I turned down his invitation*.

Another classification is proposed by Hussein (2008b) who classified the connective (ف/f) into four classes according to the discursive function: 'sequentiality', 'immediacy', 'non intervention', and 'causality'.

As far as we know, the work done within the Leeds Arabic Discourse Treebank (Al-Saif and Markert, 2010) is the sole efforts towards a detailed description of the discursive usage of Arabic DCs. Drawing partly from initial lists of Arabic DCs (Alfarabi, 1990; Alansari, 2003; Ryding, 2005), Al-Saif and Markert (2010) built a list of 107 DCs. They are categorized according to their type, position (at the beginning or the middle of a sentence) and syntactic status. Type can belong to five classes:

- *Simple* for DCs identified through one word, such as أو/>w/*or*.
- *Clitic* which is one or multiple letters attached to a word, such as ف/f/*then*.
- *More than one token* which is syntactical/non syntactical phrase, such as بيد ان /byd An/*but*.
- *Modified* which is a changed form of the principal connective, such as بالرغم /bAlrgm/ and رغم ان/rgm >n/ which present a modified form of the connective رغم /rgm/*although*.
- *Paired* which is two separated parts non adjacent to the connective, such as رغم أن...الا ان/An…AlA A/*although*/*despite* and اذا .. ف/f…A*A/*if…then*.

Syntactic categories can be:

- *Coordination conjunction* such as لكن /lkn/*but*, او /Aw/*or* and و/w/*and*,
- *Subordination conjunction* can be simple (لان/lAn/*because*, بينما /bynmA/*while* and حيث/Hyv/*where*/*since*) or paired (رغم أن...الا ان/rgm >An…AlA An/*although*/*despite*, ف اذا ../ f…A*A */if…then*), *adverbial* such as نتيجة ل /ntyjp l/*as a result* (can be simple or paired such as ف...طالما /TAlmA.. f/*as long as*),
- *Prepositional phrases* such as بالتالي /bAltaly/*consequently*.
- *Nouns* can be simple, such as بغية/bgyp/*desire* and نتيجة/ntyjap/*result* or combined nouns with a preposition, such as فضلا عن/fdlA En/*as well as* and بيد ان/byd An/*but*.
- *Preposition* can be clitic attached to al-masdar, such as ل/l/due to/for and ب/b/*by* and some subordination conjunctions, such as بعد/bEd/*after*, قبل/qbl/*before* and منذ/mn*/*since* which correspond to prepositions attached to al-masdar.

According to (Al-Saif and Markert, 2010), English and Arabic DCs share basic discourse characteristics (function, position and type). Major differences come from clitics and (some) nouns that are considered to be connectives in Arabic but not in English.

During the study and the annotation of LADTB corpus, authors are faced to many ambiguity problems. They identified four main factors of ambiguity:

i) the rich morphology of Arabic and precisely the problem of clitics that are agglutinated to words,
ii) the connective can be occur as a noun like al-masdar (discourse connective) or another noun (in this case, it is a simple connective),
iii) the absence of hamzah (ء) in unvowled texts of LADTB corpus,
iv) and finally the second version of ATB part 1 (the source of TBPC) contains several errors at the morphological annotation and transliteration.

In addition to these ambiguity factors, several connectives do not have correspondent connectives in English language (generated from the PDTB corpus). For example, the connective (اثر/Apr/*after*) is translated as (after) which presents the exact translation of the connective (بعد/bEd/*after*). However, the connective (اثر/Apr/*after*) can have other meanings and can be translated in some cases into (since).

Furthermore, other connectives lose their discourse functions in the English translation, such as (اما/AmA/*or*) and (و/w/*and*) in the beginning of the sentence. In addition, all conditional relations are marked by the DC "if" in the PDTB corpus. In contrast, authors found in the LADTB corpus several DCs (e.g. لو/lw/*if*, لولا/lwlA/*if*, طالما/TAlmA/*if*, اذا/A*A/*if*, ما/mA/*if* and حال/Hal/*if*), that mark conditional relations. Moreover, we can justify the variety and the heterogeneity of connectives by the fact that several simple English connectives have as a correspondence an Arabic connective of a type "more than token" (e.g. الا ان/AlA An/*but*, انما/AnmA/*but*, بل /bl/*but*, بيد ان/byd An/*but*, بيد/byd/*but*, لكن/lkn/*but* and غير ان/gyr An/*however*/*but*).

After the annotation of LADTB corpus, Al-Saif and Markert (2010) have extracted 91 discourse connectives and 16 derived forms. The authors noted that a small set of connectives is common for both languages. Finally, the LADTB corpus contains 6,328 connectives in which 74% are clitics and 4% are more than a token. The most ambiguous connective is (و/w/*and*) and it has 2,400 occurrences in the corpus. Then, the authors used a supervised learning method based on morpho-syntactic features to classify these connectives into two classes: discursive usage and non discursive usage. This method achieved an F-measure of 80%.

In this dissertation, we follow Al-Saif et al's definition of Arabic DCs. In addition, we consider that signalling includes other phenomena than DCs, as suggested by (Taboada and Das, 2013). In their study on signalling of coherence relations, Taboada and Das (2013) proposed a taxonomy of

signals organized in 8 groups with a total of 39 signalling devices: (1) DC (conjunction, adverbial, prepositional phrase, etc.), (2) reference, including personal, demonstrative and comparative references, (3) lexical, such as indicative phrase/word, (4) semantic (synonym, antonym, hyponym, lexical chain), (5) morphological, mainly tense, (6) syntactic, such as non finite/relative clause and parallel structure, (7) graphical signals such as colon, dash, bullet and finally (8) genre that deals with attribution and pyramid scheme. In our study, we restrict other signals to specific words, called *indicators* that are important cues for discourse analysis. Indicators can be reported speech, non inflectional verbs (such as احترس /A.htrs/*beware*), references (including personal and demonstrative reference), some adverbs (such as فقط/fqt/*only*), conjunctions (such as طالم/tAlmA/*as long as*/*so far as*), particles (such as لم/lm/*not* and لن/ln/*never*) and punctuations. In Example 13, the word فيه/fyh (composed of the preposition فى/fy followed by the possessive pronoun ه/h) triggers the relation *Entity Elaboration* since the second segment provides a detailed description of the dish introduced in the first segment. Similarly, punctuations can sometimes indicate a discourse relation. For example, ":" can trigger the relations *Elaboration* or *Attribution*.

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

[wqd~mt lnA SHnA SgyrA][fyh mqrwDAt $hy~p.]

[She gave us a small dish] [containing tasty Makrouts.]

Finally, it is important to note the difficulty of translating Arabic connectives into English (Fareh *et al.*, 1999; Chaalal, 2010; Emara, 2014). For instance, Chaalal (2010) showed found five possible English translations of the DC ف/f depending on its discursive function: temporal succession ("then"), result ("so"), causal ("because"), contrast ("but") and finally exemplification ("for example"). These studies demonstrate that some connectives in Arabic do not have their equivalent in English while some lose their discursive function when translated into English. In addition, different connectives in Arabic can be translated into the same connective in English. Similarly, some Arabic DCs have the same equivalent English connective. Sometimes, it is necessary to add other adverbs such as "rather" or "shortly" to the English connective to get the same usage of its corresponding Arabic connective.

2.3. Main studies on Arabic discourse processing

As far as we know, there are only three main researches on Arabic discourse processing: (Hassan *et al.*, 2008) and (Khalifa *et al.*, 2012) that proposed a taxonomy of Arabic discourse relations within the RST framework and (Al-Saif and Markert, 2010) that created the first corpus in Arabic annotated with discourse information following the PDTB model.

2.3.1. Hassan et al.'s work

Hassan et al. (2008) proposed discourse parser using RST. Authors built a framework of applying RST on Arabic language in order to rhetorically parse, understand, and summarize Arabic texts. They extract the Arabic rhetorical relations based on studying the English relations, analyzing Arabic corpus and understanding and using the Arabic cue phrases. Since the analysis was done on the English corpus, authors part from the hypothesis "*the rhetorical relations that were identified in English text can serve in the processing and analysis of Arabic texts*". Due to the differences between the Arabic and English languages, the English rhetorical relations can not be used in their present forms for the Arabic text. To cope with this problem, authors started by studying the Arabic corpus to extract some Arabic rhetorical relations that reflect the essence of the Arabic texts. In fact, authors pick an English relation, and then they scan the Arabic rhetoric and literature references (Gabawah, 1972; Aubadah, 1983; Abdulmuttalib, 2003; Alansari, 2003) for this relation, to see if this relation is explicitly signaled. If so, the relation is added to the Arabic relations list; otherwise, the relation is ignored. In the second step, authors looked into the Arabic rhetoric and literature references that have been written by Arabic language scholar for the relations that connect the Arabic clauses. In the third step, authors scan the Arabic corpus to obtain the DCs of each relation.

Finally, authors identified 11 relations: *Condition*, *Joint*, *Interpretation*, *Antithesis*, *Justification*, *Confirmation*, *Sequence*, *Result*, *Example*, *Base* and *Explanation*. Each relation is characterized by a *Status* that specifies the rhetorical status of the units (satellite_nucleus or nucleus_satellite), *Position* that specifies the position of the DC in the text (beginning of the statement, or middle of the statement), *Action* specifies the action that the DC has in determining the EDUs, *Relation* that specifies the relation that the DC signals, and finally *Regular expression* that contains the regular expression of the cue phrase. We note that in the case of a DC is followed by another (e.g. أن أي/>y >n/*so that*), authors tacked into account just the first DC. Finally, using a corpus containing 100 articles (each article ranges between 450 and 800 words), the presented discourse parser is used to automatic summarization Arabic texts where authors achieve a precision of 65% using human evaluations. Since any evaluations are presented for the discourse parser. Example 14 presents an output of the parser where sentence has been segmented into three EDUs and two relations are identified: *Confirmation*(1,2) and *Justification*(2,3).

(14) [لم يذهب خالد إلى السوق هذا اليوم،]$_1$ [بل لم يخرج من البيت]$_2$ [بسبب الأمطار الغزيرة.]$_3$

*[lm y*hb xAld <lY Alswq h*A Alywm,]$_1$ [bl lm yxrj mn Albyt]$_2$ [bsbb Al>mTAr Algzyrp.]$_3$*

[Khalid did not go to the market today,]$_1$ [but did not come out of the house]$_2$ [because of the heavy rains.]$_3$

2.3.2. Khalifa et al.'s work

Khalifa et al. (2012) proposed a taxonomy of Arabic discourse relation based on studying cue phrases and the different Arabic rhetoric structures respectively. This taxonomy is able to detect explicit and implicit Arabic discourse relation. Authors used a comparison between Arabic and English DCs relying on Arabic DCs to classify a group of explicit Arabic coherence relations similar to English relations, Arabic rhetoric literature for additional DCs and their corresponding explicit coherence relations and implicit relations from among the different Arabic rhetorical structures. We note that The English relation taxonomy of the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is considered a reference in comparing Arabic and English relations.

To produce the Arabic taxonomy, Khalifa et al. (2012) built a four-step algorithm. First, they selected a primary list of Arabic cue phrases by translating a list of English cue phrases (taken from (Knott, 1996)) into Arabic, using the Google translator tool. Then, they looked for instances of this list in their corpus, discarded unseen cues and collected new cues. After that, they related each Arabic cue phrase into their corresponding English relations and translated those English relations into Arabic. Finally, for those Arabic connectives that have no corresponding English relations, they added new Arabic explicit discourse relations. This procedure resulted in a flat taxonomy of 47 Arabic relations (see Table 1.1). A comparison between Arabic and English cue phrases has shown that all English coherence relations are also contained in the Arabic coherence relation's set. Additionally, extra 12 Arabic explicit coherence relations (relations in bold in Table 1.1) and 4 implicit relations were recognized (underlined relations in bold in Table 1.1). To our knowledge, these relations were not used in any annotation campaign and no available corpus annotated with discourse information has been build. Example 15 presents two EDUs linked with the relation implicit Arabic discourse إحتباك /Ehtibak[8](1,2).

(15) [هناك طلاب مجتهدون في الدراسة]$_1$ [وهناك طلاب متفوقون في الرياضة.]$_2$

[hnAk TlAb mjthdwn fy AldrAsp]$_1$ [whnAk TlAb mtfwqwn fy AlryADp.]$_2$

[Some students are distinguished in school]$_1$ [and other students are distinguished in sport.]$_2$

[8] "إحتباك /Ehtebak" is an Arabic implicit relation usually found in robust rhetoric texts, as in the Holly Quran. It connects two adhesive sentences in a way such that each sentence has two adjectives, one of them is explicit, and the other is hidden, but can be guesstimated from the other sentence. In turn, the two adjectives in the second sentence are in contrast with; or opposite to; the two adjectives in the first sentence.

1. Antithesis	15. Conjunction	29. Motivation	39. **Alert**
2. Evaluation	16. Volitional Result	30. Enablement	40. **Uncertain**
3. Justify	17. Unconditional Unless	31. Otherwise Purpose	41. **Not-Jumping to conclusions**
4. Volitional Cause	18. Evidence	32. **Indifference**	42. **n-Tuple condition**
5. Otherwise Purpose	19. Preparation	33. **Exclusion**	43. **Cascaded questioning to get an answer about one of many events**
6. List	20. Restatement	34. **Bidirectional condition**	44. **Narration change**
7. Background	21. Contrast	35. **For fear that**	45. **Cascaded questioning**
8. Concession	22. Unconditional Unless	36. **Conjunction of uncommon event**	46. **Impossible condition**
9. Condition	23. Non volitional Cause	37. **Choosing oneout of many alternatives**	47. **Ehtebak /إحتباك**
10. Solutionhood	24. Elaboration	38. **Exclude one of two opposite events**	
11. Means	25. Disjunction		
12. Interpretation	26. Non volitional Result		
13. Joint List	27. Circumstance		
14. Sequence	28. Summary		

Table 1.1. Taxonomy according to Khalifa et al. (2012)

2.3.3. Al-Saif and Markert's work

The closest research to our work is the one done by (Al-Saif and Markert, 2010) that aims at building the Arabic Discourse Treebank, the Leeds Arabic Discourse Treebank (LADTB) and automated modeling of discourse relations for Arabic. The Leeds Arabic Discourse Treebank LADTB is a news corpus where all discourse connectives are identified, and annotated with the discourse relations they convey as well as with the two adjacent arguments they relate. This corpus contains 5,651 annotated discourse connectives in 537 news texts. Authors defines DCs as lexical expressions that relate two text segments expressing abstract objects such as events, beliefs, facts or propositions (see Section 2.2). They extract frequently used DCs in MSA. In the discourse connective collection phase, Al-Saif and Markert (2010) were mostly interested in the nature of the discourse connective, where it occurs in the sentence, and what relation it typically signals. The syntactic sentence/clause boundaries were used initially to determine the argument boundaries. The properties of the DC describe the type, possible position, the discourse relations the connective usually signals, and its syntactic category. To build the final list of DCs that will be used to automatically predict Arabic discourse relations, authors follow two steps: collect an initial list of potential connectives and check for each connective its discourse usage. Authors analyzed around 50 random raw texts from the Penn Arabic Treebank (Penn ATB Part1) six articles from well-known Arabic websites (such as educational, political and social affairs) which were on average 600 words, and extracted all discourse connectives and their modified forms according to our definition of discourse connective. We can cite examples of discourse connectives with their frequency in LADTB: بالتالي/bAltAly/*consequently* (14 occurrences), جراء/jrA'/*because* (10 occurrences), على الرغم/ElY Alrgm/*yet* (9 occurrences), نظرا ل/nZrA l/*because* of (9 occurrences), انما /AnmA/*but* لو/lw/*if* (6 occurrences), فى ظل/fy Zl/*under* (6 occurrences) and كذلك/k*lk/*and that* (6 occurrences). Two annotators are used to annotate DCs in LADTB with an inter-annotator agreement 0.83 of Kappa. The gold standard LADTB contains 6,328 DCs: 1,276 are simple, 4,779 are clitic and 273 are more than token.

Additionally, Al-Saif and Markert (2011) built a supervised learning model to predict the discourse usage of DCs. Authors used 18,798 potential DCs for training and 5,880 DCs used for test. As features, authors used surface features of the potential connective, lexical features of surrounding words, part of Speech features, syntactic category of related phrases and morphological features. After 10-fold cross-validation, authors obtained an F-measure of 86%.

For Arabic discourse relations, the set of relations is the same as the hierarchy used in the English PDTB (Prasad *et al.*, 2008) except that the number of relations was reduced (from 33 to 17) and two new Arabic relations ("Expansion.Background" and "Comparison. Similarity") were added. The taxonomy used in LADTB is presented in Figure 1.9.

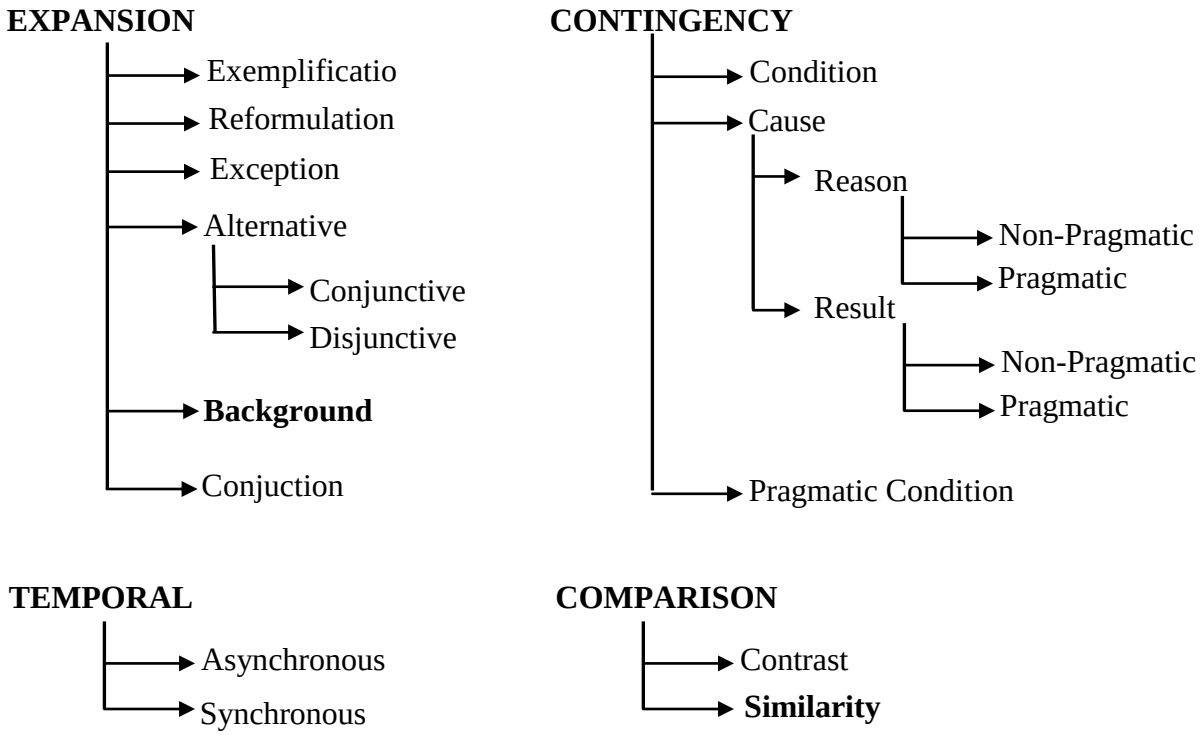

Figure 1.9. The LADTB relations**.**

Two annotators are asked to annotate Arabic discourse relations in LADTB with an inter-annotator agreement 0.710 of Kappa. The gold standard LADTB contains 6,039 explicit Arabic discourse relations where Conjunctive relation represents 54%. Then, Al-Saif and Markert (2011) built a supervised learning model to predict the Arabic discourse relations. As features, authors used connective features, words and POS of arguments, al-masdar, tense and negation, length and distance, argument order, argument parent and production rules. After 10-fold cross-validation, authors obtained an accuracy of 0.783.

3. Our approach

This section aims to compare our study to the one elaborated by Al-Saif and Markert (2010). We choose to present our approach by referring to this study for two main reasons: we share the same goal of the discourse analysis and we use the same kind of corpus (ATB).

Firstly, for discourse connectives, it is noteworthy that Al-Saif and Markert (2010) have built a lexicon of DCs. They described the connectives found in the LADTB corpus. We constructed a lexicon that includes DCs identified by Al-Saif and Markert (2010) (91 connectives), triggers of discursive relations as well as DCs that help to identify the discursive relations extracted from Arabic Treebank (ATB v3.2 part3), Elementary School Textbooks (EST), and from the Arabic literature. In total, our lexicon contains 174 discourse connectives.

In the context of similarity, we adopted the connectives types used by Al-Saif and Markert (2010): clitic, simple, compound. However, the only difference is that we ignored the type "modified" and the type "paired". Indeed, we do not see the usefulness of the type "modified" because such DC can present information completely different from the target DC. There is no link between target and its modified DCs. For the type "paired", all markers identified in LADTB and that are found in our corpus are composed of two non adjacent DCs and not two words (the two words are independent DCs that help to identify the same relation). In the same context, we used all POS tags used by Al-Saif and Markert (2010) by adding the reporting verbs, given that we have associated all possible grammatical functions to each connective.

For the difference points, Al-Saif and Markert (2010) use a classification based on discursive or non discursive usage. In contrast, since our lexicon describes a DC in different possible contexts, we do not have a DC with a non discursive usage. Again, we use three classes: discursive, lexical (as ومن نتائج كل ذلك/wmn ntA}j kl *lk / *the results of all this* and على اساس ان/ ElY AsAs An/*on the basis of*) and punctuation. For the punctuation class, we assume that these markers can have a rhetorical sense and can help us to identify discourse relations (further the discursive function of segmentation). In addition, we have associated an information to each DC that indicates whether it is strict (always reports the same relation), or ambiguous (may indicate numerous relations according to the context). In addition, we denote that the set of DCs that can be associated to each DC (two adjacent connectives as رغم ان/rgm An/*although*). Finally, we added further information: the lemma of each DC, its English translation, an example and a comment, if necessary.

In our approach, we choose to go 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. Indeed, we propose a semantically driven approach following SDRT where a document is represented by an oriented acyclic graph, which captures explicit and implicit relations as well as

complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups, and crossed dependencies. In fact, we choose to not reuse the LADTB relation set. Instead, we choose to start with the set of relations that is already defined within past SDRT-like annotation campaigns (cf. Discor (Reese *et al.*, 2007) for English and Annodis (Afantenos *et al.*, 2012) for French) and to refine 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.

Moreover, we extend Al-Saif and her colleague's study 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 Arabic Discourse Treebank corpus (ADTB) 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 4,963 EDUs, linked by 3,184 relations. 25% of these relations are implicit while 15% link non adjacent EDUs.

In addition, we investigate how Arabic discourse analysis can improve the NLP application results (e.g. summarization systems, translation systems, Question/Answering systems, etc.). Indeed, we propose an automatic Arabic text summarization based on discourse information (discourse relations and discourse structure). We use the semantic of the discourse relations and the discourse structure to extract the most important Elementary Discourse Units (EDUs) in the text. The selected EDUs for a summary must contain the main information, event, object, ideas, etc. of the text. This tool is useful for judging the adequacy of the text with the information requested by the user. Moreover, we propose many algorithms according to discourse criteria (coordinate/subordinate relations, Complex Discourse Units (CDUs), discourse level, etc.) and we evaluate these algorithms using two different corpora that have two different frameworks: ADTB (cf. Chapter 2), annotated according to the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) and Arabic Discourse RST corpus AD-RST (100 texts selected from the journal "Dar Al Hayat") (Keskes *et al.*, 2012d), annotated according to the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). As conclusion, the presented results confirm that discourse structure and discourse relation nature have a positive impact on the content selection.

Conclusion

In this chapter, we first introduced some backgrounds about discourse analysis (discourse connectives, discourse units and discourse structures), then we presented main existing discourse theories. We also presented the specificities of Arabic and the main difficulties that we need to overcome to automatically annotate Arabic texts with the discourse information. We finally gave an overview of the main studies on Arabic discourse processing.

Compared to related work, we propose the first approach that explicit the interactions between the semantic content of Elementary Discourse Units and the global pragmatic structure of Arabic discourse. The first step of this approach is to study the feasibility of the manual annotation of full discourse structure, as described in the Chapter 2.

Chapter 2: Manual Annotation For Arabic Discourse Analysis

Table of contents

Introduction 51
1. The data 51
2. Discourse segmentation manual 53
2.1. Annotation scheme 53
2.2. Inter-annotators agreement study 63
3. Manual annotation of discourse relations 64
3.1. Arabic rhetoric 64
3.2. Building a new hierarchy of Arabic discourse relations 68
3.3. Annotation campaign 82
3.4. Results 85
Conclusion 91

Introduction

In this chapter, we focus on the manual annotation of the Arabic Discourse Treebank (ADTB) corpus which is composed of newspaper documents collected from the syntactically annotated Arabic Treebank (ATB v3.2 part3) (Maamouri *et al.*, 2010b). The annotation starts by segmenting documents into Elementary Discourse Units (EDUs) that have to be linked by discourse relations, to form Complex Discourse Units or CDUs, which in turn may be linked via discourse relations to other discourse units. Herein, we define three segmentation principles: basic segmentation principles, segmentation principles into clauses, and segmentation principles into EDUs. Since EDU does not exceed the clause boundaries, we choose to define segmentation principles into clauses to be used by annotators as segmentation constraints. In addition, given our semantic-driven approach on discourse, we choose to not reuse the LADTB relation set. Instead, we choose to start with the set of relations that is already defined within past SDRT-like annotation campaigns and to refine 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.

This chapter is organized as follows: Section 1 presents the corpora. Section 2 details the segmentation manual for Arabic documents. Finally, Section 3 describes our hierarchy of discourse relations, the annotation scheme, a quantitative (in terms of discourse connectives, relation frequencies, proportion of implicit relations, etc.) and a qualitative analysis (inter-annotator agreements and error analysis) of the annotation campaign.

1. The data

In order to build the gold standard corpus ADTB, we use two different corpora: Elementary School Textbooks (EST) to carry out the manual segmentation into EDUs, and the syntactically annotated Arabic Treebank (ATB v3.2 part3) to build the manual annotation of discourse relations. These two manuals have been used to build the gold standard ADTB.

The Elementary School Textbooks (EST) is composed of 250 documents (1,095 paragraphs and 29,473 words). Some researchers from our ANLP research group have collected these EST documents. They have first randomly selected a set of texts from Tunisian Elementary School Textbooks (level 4^{th}, 5^{th}, 6^{th}, 7^{th} and 8^{th}), and then they have manually introduce them into a text file format. Three linguists manually segmented the corpus. The annotation relies on consensus. Table 2.1 gives more details on EST.

The EST documents are usually well structured. Sentences are short (around 5.6 words per sentence) with a quite simple syntactic structure. They are characterized by the presence of punctuation marks. Document length is also short (around 10 sentences per document).

	EST	
	Texts	**EDUs**
4th EST	47	944
5th EST	43	810
6th EST	50	701
7th EST	53	856
8th EST	57	975
Total	**250**	**4,286**

Table 2.1. EST details

Example 1 presents a sentence extracted from EST.

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

ElY $AT} AlHm~AmAt, AntSb HSn qdym, ydxlh Alz~A}r mn bw~Abp mqw~sp, tfDy bh <lY >rwqp msq~fp, k>swAq Almdynp AlEtyqp.

On Hammamet beach, an old fort is erected, in which a visitor can enter it from an arched gate, that leads him to wrapped corridors that resemble ancient city markets.

Arabic Treebank ATB v3.2 part3 (Maamouri *et al.*, 2010b) consists of 599 newswire stories from Annahar News Agency. There are 339,710 words/tokens before clitic[9] 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 part of speech level and then the syntactic Treebank annotation that characterizes the constituent structures of word sequences, provides categories for each non terminal node, and identifies null elements, co-reference, traces, etc. Comparing to EST, ATB documents are longer (around 25 sentences per document) and sentences are syntactically more complex. Example 2 presents a short sentence extracted from an ATB document along with the morphological analysis of its first two words (cf. Figure 2.1) and its syntactic tree (cf. Figure 2.2).

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

An swryA >SbHt AbtdA' mn mTlE Alsnp AljAryp EDwA gyr dA}m fy mjls Al>mn lmdp sntyn.

Since the beginning of the year, Syria has become a non permanent member of the Security Council for two years.

[9]A clitic has the syntactic characteristics of a word but depends phonologically on another word or phrase. Clitics include prepositions, conjunctions, and pronouns. For instance, the preposition (like ف/f/*then*), conjunctions (like و/w/*and*), articles (like ال/Al/*the*) and pronouns (like ه/h/*he*) can be affixed to nouns, adjectives, particles and verbs, which causes several lexical ambiguities. For example, the word فهم / fhm can be a noun (that means understanding) or a conjunction (ف/f/*then*) followed by the pronoun (هم/hm/*they*).

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

Figure 2.1. Morphological analysis of the two first words of Example 2 as given by ATB manual annotations.

In Figure 2.1, the annotation includes: the Arabic word, its transliteration (IS_TRANS), its position in the sentence (INDEX), its offsets, its corresponding unvocalized and vocalized words, its part-of-speech (POS) and its English translation (Gloss).

```
(ROOT
  (S
    (VP (VBP ان)
      (NP (NNP سوريا))
      (S
        (VP (VBD اصبحت)
          (NP
            (NP (NN ابتداء))
            (PP (IN من)
              (NP (NN مطلع)
                (NP (DTNN السنة) (DTJJ الجارية)))))
          (NP
            (NP (NN عضوا))
            (NP (NN غير)
              (NP (NN دائم))))
          (PP (IN في)
            (NP (NN مجلس)
              (NP (DTNN الامن))))
          (NP (NN لمدة)
            (NP (NNS سنتين))))))
    (PUNC .)))
```

Figure 2.2. Syntactic analysis of Example 2 as given by ATB manual annotations.

2. Discourse segmentation manual

We begin this section by defining our annotation scheme. Then, we present the inter-annotators agreement.

2.1. Annotation scheme

The annotation scheme defines a set of segmentation principles to guides the segmentation process. Our scheme is inspired from an already existing manual elaborated within the Annodis[10] project that focused on the selective annotation of multi-level discourse structures of French

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

documents following SDRT (Afantenos *et al*., 2012). Annodis manual provided annotators with an intuitive introduction to discourse segments, including the fact that discourse segments can be embedded in one another. Detailed instructions were provided describing how to handle segmentation for most of the cases that could naturally arise.

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

We give in this section basic segmentation cases, main segmentation principles into clauses as well as main segmentation principles into EDUs.

2.1.1. Basic principles

EDUs are delimited by square brackets. Discourse Connectives (DCs) are always at the beginning of a segment whereas punctuation marks that delimit segment frontiers always appear before the end of a segment. EDUs cannot overlap but they can be embedded in another (double square brackets are not allowed), as in Example 3.

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

[nAq$ Al>stA AlAmtHAn, [Al*y >jrAh AltlAmy* Al>sbwE AlmADy,] w Aldrs AlHAly.]*

[The teacher explained the exam the students sat for last week,] and the current lesson.]

An EDU is basically a verbal (cf. Example 4) or a nominal clause (مبتدأ/mbtd> and خبر/xbr) (cf. Example 5). A cutting point can neither separate a verb from its complement nor a subject from its verb. In addition, segment frontiers can never occur within a chunk or a named entity.

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

[qSft TA}rAt >myrkyp mjmEAt mn Alkhwf.]

[American aircrafts bombed a set of caves.]

(5) [كانت الطفلة جميلة.]

[kAnt AlTflp jmylp.]

[The girl was beautiful.]

2.1.2. Main segmentation principles into clauses

During the corpus analysis, three different segmentation principles were identified: (p1) use punctuation marks only, (p2) use the DCs only, and (p3) use both the principles (p1) and (p2) when the DCs are ambiguous.

- **Punctuation marks principles**

Punctuation marks, which are used today in Arabic writings, are the same ones utilized for the European writing system, but they do not necessarily have the same semantic functions. For example, the origin of the comma is to be found in the Arabic letter "و/w", which represents the conjunction "and" for English. Borrowed by the Italian typographers, the comma becomes mute in the Latin alphabet. The point is often used in Arabic to mark the end of a paragraph whereas the comma, in addition to its coordination function, can also be used to announce the end of a sentence (Belguith *et al.*, 2005).

In Arabic, parentheses, exclamation point, question mark, three points, etc. have the same values as those of European languages (Belguith, 2009). However, it should be noticed that some punctuation marks in Arabic look different from the European ones. Indeed, the Arabic comma points to the opposite way (،), the semi-colon is inverted (؛) and it is written on top of the line and the Arabic question mark looks to the opposite side (؟).

The punctuation marks are not widely used in current Arabic texts (i.e., at least not regularly) and when they are used, they do not respect the typography rules[11]. Therefore, their presence can not guide the segmentation process as for other languages such as English or French, which make segmenting Arabic text harder.

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

We give below Example 6 and Example 7 that introduce strong indicators:

(6) [ألقيت كلمة مازالت أحفظها إلى هذا اليوم :][«وطني. أحبّك يا وطني. »]

*[>lqyt klmp mAzAlt >HfZhA <lY h*A Alywm:] [«wTny. >Hb~k yA wTny. »]*

[I said a word that I still remember until today:] [«My country. I love you dear country. »]

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

(7) [طرد خليل من المدرسة ؛] [لأنه غش في الامتحان.]

[Trd xlyl mn Almdrsp;] [l>nh g$ fy AlAmtHAn.]

[Khalil was expelled from school;] [because he cheated in the exam.]

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

- If the full stop is part of a named entity, it does not represent the end of a segment, as in Example 8 and Example 9.

(8) [د. طارق سويدان عالج أمراض مختلفة.]

[d. TArq swydAn EAlj >mrAD mxtlfp.]

[Dr. Tarak Swiden has treated various diseases.]

(9) [يعتبر فيتامين ب.2 و ب.12 من اكثر الفيتامينات التي تساعد على مقاومة الزهايمر.]

[yEtbr fytAmyn b.2 w b.12 mn Akvr AlfytAmynAt Alty tsAEd ElY mqAwmp AlzhAymr.]

[The vitamins B.2 and B.12 are considered as the most effective to fight against Alzheimer illness.]

- If the dot is preceded by one word and this word is not a verb, then dot does not represent the end of a segment, as in Example 10.

(10) [وطني. أحبّك يا وطني.]

[wTny. >Hb~k yA wTny.]

[My country. I love you dear country.]

- If the comma is followed by a verb or اسم اشارة /Asm A$Arp/*demonstrative pronoun*, then it represents the end of segment, as in Example 11.

(11) [ترك بيروت ،] [لذلك كانت زوجته ليست دائماً إلى جانبه]

*[trk byrwt,][l*lk kAnt zwjth lyst dA}mAF <lY jAnbh.]*

[He leave Beirut], [so his wife was not always on his side]

- If an apposition contains only a named entity, then it does not represent the end of a segment, like shown in Example 12.

(12) [كتب الشاعر الكبير، نزار القباني، أشعار كثيرة عن المرأة.]

[ktb Al$AEr Alkbyr, nzAr AlqbAny, >$EAr kvyrp En Almr>p.]

[The great poet, Nizar Qabani, wrote many poems about woman.]

- For the other weak indicators, i.e. quotes, parenthesis, brackets, braces, and underscores, they usually indicate the beginning of a segment in the case they contain a verbal clause, as in Example 13 and Example 14.

(13) [طرق المدير باب القسم][(حيّانا ببشاشة)] [وتقدّم إلى معلّمنا.]

[Trq Almdyr bAb Alqsm][(Hy~AnA bb$A$p)][wtqd~m <lY mEl~mnA.]

[The director knocks the door of the classroom][(he smiles)][and then he comes to talk to our teacher.]

(14) [قال المدير"تحيّة العلم"][فانقطعت كل حركة.]

[qAl Almdyr "tHy~p AlElm"][fAnqTEt kl Hrkp.]

[The director said, "flag salutations"][and then all movements have stopped.]

Although Arabic language includes punctuation marks, written Arabic rarely contains these punctuations. Indeed, Arabic discourse intends to use long and complex sentences, so we can easily find an entire paragraph without any punctuation marks. Therefore, segmenting according to *p1* is not enough.

- **Discourse connective principles**

Using DCs could be a solution to further segment sentences into clauses, as in Example 15 where we have a contrast discourse relation.

(15) [سيعرف الجميع متى نبدأ][لكن لن يعرفوا متى سننتهي]

[syErf AljmyE mtY nbd>][lkn ln yErfwA mtY snnthy]

[They will know when we start][but they won't know when we will finish]

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

- If one of the DCs {ل/l/*for/to*, من أجل أن/mn <jl <n/*in order to*, حتى/htY/*to/until* كي/ky/*for/to*, etc. } is followed by a verb, it indicates the end of a segment, as in Example 16.

(16) [فبعض الكتاب يستخدمون كلمات سهلة في مقالاتهم] [من أجل أن يفهمها القراء.]

[fbED AlktAb ystxdmwn klmAt shlp fy mqAlAthm] [mn >jl >n yfhmhA AlqrA'.]

[Some authors use in their articles simple words] [in order to be understood by readers]

- If one of the DCs {إلا/<lA/*except*, بحيث/bHyv/*in fact*, لكن/ lkn/*but*, غير أن/gyr >n/*however*, بيد أن/ byd >n/*however*, etc.} is followed by a verb or if these cues are proceeded by the conjunction و/w/*and* or ف/f/*so*/*then*, then it indicates the end of a segment, as in Example 17 and Example 18.

(17) [يمكنك أن تستغني عن مالك] [ولكن ابتعد عن التبذير.]

*[ymknk >n tstgny En mAlk] [wlkn AbtEd En Altb*yr.]*

[You can spend your money] [but avoid to fritter frittering.]

(18) [نحرص على نظافة المطبخ][بحيث يتم التخلص من أي بقايا طعام]

[nHrS ElY nZAfp AlmTbx] [bHyv ytm AltxlS mn >y bqAyA TEAm]

[We keen to clean the kitchen] [so as we get rid of any food rest]

On the other hand, ambiguous DCs do not always mark the beginning of a segment, as the connective و /w/*and* and the particles (ثم/vm/*then*, ف/f/*so*/*then*, etc.). For example, the particle و/w can express either a new clause (cf. Example 19), a conjunction between NPs (cf. Example 20), or it can be a part of a word (cf. Example 21).

(19) [فنظر إليّ،] [وقال:]

[fnZr <ly~,] [wqAl:]

[Then he looked at me,] [and he said:]

(20) [فلاحظ البائع و الحريف يتناقشان على أسعار البضاعة]

[flAHZ AlbA}E w AlHryf ytnAq$An ElY >sEAr AlbDAEp.]

[Then he remarked the seller and the client were discussing the products' prices.]

(21) [كانت كل ورشة عمل تشكو من افتقار أجهزة العمل.]

[kAnt kl wr$p Eml t$kw mn AftqAr >jhzp AlEml.]

[Each workshop has suffered from a lack of equipment.]

During the annotation process, we observed that the DC principles could not resolve some ambiguities related to weak indicators (49 ambiguous DCs were identified). In addition, we have

also observed that some connectives, in some cases, can be easily disambiguated using punctuation marks. We need therefore to use both punctuation marks and DCs in order to better identify the right segment frontiers.

- **Mixed principles**

We give here, some rules to illustrate the mixed principles:

- If a comma is followed by the conjunction و/w/*and* or ف/f/*so*/*then*, and then by a localization preposition {على/ElY/*upon*, في/fy/*in*/*into*, عن/En/*about*, من/mn/*from*, إلى/<lY/*to* }, then it indicates the end of a segment, as in Example 22.

(22) [كان أهله على عادة كثير من العائلات التّونسيّة يتخلّعون ببلدة المرسى،] [وعلى شاطئها البديع بدأ اللّقاء حميما بينه وبين الطّبيعة.]

[kAn >hlh ElY EAdp kvyr mn AlEA}lAt Alt~wnsy~p ytxl~Ewn bbldp AlmrsY,] [wElY $AT}hA AlbdyE bd> All~qA' HmymA bynh wbyn AlT~byEp.]

[Like many of Tunisian families, his parents spend their summer holidays in Marsa city,] [and it's on its wonderful beach that they warmly meet nature.]

- If a comma is followed by the conjunction و/w/*and* or ف/f/*so*/*then* and then by a possessive noun {له/lh/*him*, لها/lhA/*her*, لهما/lhmA/*them*, لهنّ/lhn~/*them*, لهم/lhm/*them*, لي/lY/*me*, لك/lk/*you*, لكم/lkm/*you*, لكما/lkmA/*you* }, then it indicates the end of a segment, as in Example 23.

(23) [رأيت أختي في الخارج،] [لها دمية تتكلم.]

[r>yt >xty fy AlxArj,] [lhA dmyp ttklm.]

[I saw my sister outside,] [with a talking doll.]

- If a comma is followed by a demonstrative pronoun {تلك/tlk/*this*, هذه/h*h/*this*, هذا/h*A/*this*, ذاك/*Ak/*this*, ذلك/*lk/*this*, هم/hm/*these*, etc.} and then by a word that is not a verb, then we do not have a segment frontier, as in Example 24.

(24) [وقف معلّمنا سي حامد، هذا اليوم، أمامنا ينظر في وجوهنا مليّا.]

*[wqf mEl~mnA sy HAmd, h*A Alywm,>mAmnA ynZr fy wjwhnA mly~A.]*

[Mr. Hamed, our teacher, was standing up, looking at us.]

2.1.3. Main segmentation principles into EDUs

- **Al-masdar** (المصدر/AlmSdr): They are segmented only in indefinite accusative case (منصوب/mnSwb) because this construction generally signals discourse relations. For example, in Example 25, al-masdar بحثا/bHvA/*looking for* explains why Ahmed went to the library:

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

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

[Ahmed went to the library][looking for the mathematic book]

We do not segment sentence in other cases (like البحث/ *AlbHv/search)*, as in Example 26.

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

[Astmr fy AlbHv Enh fy kl AlmktbAt.]

[He keeps looking for it in all libraries.]

- **Conditionals** (شرط/$rT): They are always segmented, as in Example 27.

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

[>A >SbH AlTqs jmyl,][s>xrj >tnzh.]*

[If the weather is nice,] [I'll go for a stroll.]

- **Correlatives** (تلازم/tlAzm): They are always segmented, as in Example 28.

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

[klmA >TAlE Alktb,][klmA ttHsn vqAfty AlEAmp.]

[The more I read books,] [the more I learn.]

- **Coordinations** (ربط/rbT): In Arabic, a coordination is introduced by DCs such as و/w/*and*, ف/f/*so*/*then*, ثم/vm/*then*, أو/>w/*or*… which are highly ambiguous. For instance, the conjunction و/w/*and* can have six different senses (Khalifa *et al.*, 2011): (a) والقسم /w Alqsm that means testimony, (b) ورب /w rb that means few or someone, (c) والاستئناف /wAlAst}nAf which simply joins two unrelated sentences, (d) والحال/w AlHAl that introduces a state (cf. Example 29), (e) والمعية /w AlmEyp that means the accompaniment and (f) والعطف/w AlETf meaning the conjunction of related words or sentences (cf. Example 30).

(29) دخل الولد الفصل وهو يبتسم.

dxl Alwld AlfSl whw ybtsm.

The child enters to the classroom smiling

(30) انتهت العطلة وبدأت الدراسة.

Antht AlETlp wbd>t AldrAsp.

The holidays are over and classes begin.

Our treatment of coordination goes beyond discourse segmentation proposed in (Khalifa *et al.*, 2011), since we do not only deal with the DC و/w/*and* but also with other DCs. Therefore, we segment coordination in four cases: (i) coordination of independent clauses, (ii) coordination of subordinating clauses, (iii) when two verbal phrases share the same object or the same subject, as in Example 31, and finally (iv) coordination of prepositional phrases that introduce events, as in Example 32. We do not segment in all the other cases, such as the conjunction between two objects of the same verb.

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

[AstEAd Alr}ys Altwnsy EAfyth][wqAm bAstqbAl AlmwATnyn.]

[The Tunisian President has regained his health] [and has received the citizens.]

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

[<Elnt AlHkwmp Edm mwAfqthA ElY mHDr Aljlsp] [lEdm twfr Al$rwT Al>zmp]

[The government announced its refusal to open the session] [because of a lack of good conditions]

- **Subordinations** (صلة/Slp): They are always segmented. Relative clauses are introduced by the relative pronouns الذي/Al*y/ and التي/Alty/ that correspond in English to the pronouns which, who, whom and that (cf. Example 33). Some conjunction of subordinations (like أن/>n/*that*, أنّ/>n~/*that*, إن/<in/*if*, إذا/<i*A/*if-whether*, حتى/HtY/*so* that and طالما/TalamA/*as long as*) are generally used after a verb of communication or a reported speech verb (cf. Example 34). Other markers introduce temporal and causal subordinations such as قبل أن/qbl >n/*before that*, لأنّ/l>n~/*because*, حين/Hyn/*when* and غير أن/gyr >n/ *nevertheless*.

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

*[w fy ktAb Altklyf [Al*y wjhh AlY AlHkwmp Aljdydp ,] tm AtxA* kl AltrtybAt wAlAstEdAd AlkAml .]*

[In the book of reference [which has been sent to the new government,] all the arrangements have been taken.]

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

[wqAl wzyr AldfAE] [An nHw stp ms&wlyn Amyrkyyn wSlwA AlY AlblAd.]

[The Minister of Defense said] [that six U.S. officials had arrived to the country.]

- **Appositions** (بدل/bdl). They are segmented in most cases. Appositions can be:

- adjectival phrases,
- adverbial phrases. They are introduced either by relative adverbs (such as متى/mtY/*when*, كيف/kyf/*how*, لماذا/lmA*A/*why*, حيث/Hyv/*where*) or by regular adverbs (such as حينذاك/Hyn*Ak/*at that time*, وقتذاك/wqt*Ak/*by then* and ربما/rbmA/*perhaps*) as in Example 35,
- nominal or verbal phrases introduced by pseudo-verbs like إن/<n/*that*, ليت/lyt/*hope that*, لعل/lEl/*may be*, or by non inflectional verbs like حيا/HyA/*come to*, سرعان/srEAn/*soon*,
- Prepositional phrases (introduced by إلى/<lY/*until*, عن/En/*about*, فى/fY/*in*, من/mn/*from* and على/ElY/*on*) that appear at the end of a clause are not segmented.

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

[An Aljnwd, [Hyv sykwnwn mslHyn,] ystTyEwn AldfAE En Anfshm.]

[The soldiers, [once they are armed,] they will be able to defend themselves.]

- **Adverbials** (ظرفية/Zrfyp): In some cases, an adverbial can be an EDU. This concerns adverbials that introduce an event or a state, as in Example 36 where we have a *Goal* relation, and adverbials that are at the beginning of the sentence, as in Example 37 where we have a *Frame* relation. Example 38 gives an example of a temporal adverbial introduced by البارحة على الساعة الرابعة والنصف مساء/AlbArHp ElY AlsAEp AlrAbEp wAlnSf msA'/*yesterday at four thirty in the afternoon* that does not indicate a cutting point.

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

[rjEt msrEA <lY Albyt][Hyv kAn AlmTr ythATl.]

[I returned quickly to home][while it was raining.]

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

[EndmA twfy jdy,][knt SgyrA jdA.]

[When my grand-father died,][I was very young]

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

*[AjtmE Almjls AlbArHp ElY AlsAEp AlrAbEp wAlnSf msA'][lmnAq$p h*A AlqAnwn]*

[The council assembled yesterday at four thirty in the afternoon][in order to discuss this law]

- **Other cases.** We segment reported speech sentences between quotes (this case indicates the *Attribution* relation). We also segment modifiers that begin with possessive pronouns that detail a previously introduced entity (cf. Example 39) since this case indicates the *Entity-Elaboration* relation. We do not segment in case of transliteration, Latin characters and abbreviations, as well as in case of demonstrative pronouns (هذا/h*A/*this*, هذه/h*h/*this* and هذان/h*An/*these*).

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

[wqd~mt lnA SHnA SgyrA][fyh mqrwDAt $hy~p.]

[She gave us a small dish] [containing tasty Makrouts.]

2.2. Inter-annotators agreement study

Two Arabic native speakers (undergraduate students in Arabic linguistics) were asked to doubly annotate a set of documents from our corpora following the guidelines given in the annotation scheme. First, annotators were trained on 4 EST documents (75 sentences) and 4 ATB documents (110 sentences). The training phase for ATB last longer compared to EST since ATB documents contain more complex. This phase allowed for revising the annotation guidelines. Then, each annotator was asked to annotate separately 5 EST documents and 2 ATB documents which correspond respectively to 71 and 63 sentences (documents used for training were discarded).

Agreements were computed by counting how often each annotator classifies each token as being an EDU boundary. We got an average Cohen's kappa of 0.830 for ATB and 0.890 for EST. We observe five cases of disagreement: (a) lexical ambiguities, especially for discourse connectives that appear as clitics (cf. Chapter 1), (b) long sentences with more than 5 words (cf. Example 2 in Section 2), (c) the absence of punctuation marks, especially when clauses are not separated using punctuation marks within a sentence (cf. Example 31 and Example 32 in Section 3.1.3) and (d) al-masdar constructions (cf. Example 40). Cases (b) and (c) are more frequent in ATB documents.

(40) [تشكر أحمد جارته][وفاء لعملها.]

[t$kr >Hmd jArth][wfA' lEmlhA.]

[Ahmed thanks his neighbor][for being loyal to her work.]

In Example 40, one annotator considers that the word وفاء/wfA' is a cutting point because this word is al-masdar in an indefinite accusative case of the verb وفى/wfY. Hence, the second EDU explains why Ahmed thanks his neighbor. On the other hand, the second annotator cut at the word لعملها/lEmlhA' because he considered the words جارته وفاء /dyArh wfA' as a named entity (the name of the neighbor). For him, the second EDU explains why Ahmed thanks his neighbor Wafa.

Of course, this is an error because, in our example, the word وفاء/wfA' is al-masdar construction and not a named entity.

Given the good inter-annotator agreements results, annotators were asked to build the gold standard by consensus by discussing main cases of disagreements, as discussed earlier. Table 2.2 gives statistics about the data in the gold standard. The column WORD+PUNC indicates the number of tokens.

	Texts	Size	Sentences	EDUs	Embedded EDUs	Word+PUNC
EST	25	67ko	442	924	86 (10.74%)	6437
ATB	50	267ko	1 272	2 788	372 (7.49%)	28 288
Total	**75**	**334ko**	**1 714**	**3 712**	**458 (8.10%)**	**34 725**

Table 2.2. Characteristics of our data in the gold standard

3. Manual annotation of discourse relations

3.1. Arabic rhetoric

The corresponding translation for the word rhetoric in Arabic is البلاغة /AlblAgp, which is derived from the root verb بلغ/blg that means "to reach, attain, arrive at, or to get to a destination". Arabic rhetoric علم البلاغة/Elm AlblAgp, presents then the art of reaching the perfection in speech or writing style. It is a discipline that deals with clarity, eloquence, correctness, beauty and purity in Arabic writing or oral expression. Although the birth of Arabic rhetoric started from the pre-Islamic period, its development was strongly related to Islam as religion and culture since the concept of البلاغة/AlblAgp, was introduced to enable the understanding of the unique style of the Holy Quran (Sloane, 2001). Among the major earlier Arab rhetoricians, we cite[12]: الجاحظ /AljAHZ/Al-Jahiz (d. 255/868), ابن المعتز/Abn AlmEtz/Ibn Al-Mu'tazz (d. 296/908), عبد القاهر الجرجانى/Ebd AlqAhr AljrjAnY/Abd Al-Kader Al-Jarjeni (d. 471/1078), الزمخشري/Alzmx$ry/Al-Zamakhchari (d. 538/1143) and السكاكي/AlskAky/Al-Sakaki (d. 626/1229).

Arabic rhetoric is divided into three sub-disciplines: علم البيان/Elm AlbyAn, or science of clarity, علم المعاني/Elm AlmEAny, or science of ideas, and علم البديع /Elm AlbdyE, or science of embellishment. These disciplines have provided a rhetorical analysis of Arabic at three different levels: الكلمة/Alklmp/*the term*, by focusing on the constituent features of eloquence of words (Owens, 2006), الجملة/Aljmlp/*the sentence*, in order to establish the theoretical framework of Arabic rhetoric and finally النص/AlnS/*the text /the discourse level*, by the study of literary texts such as poetry and the Holy Quran. This section provides a quick overview of rhetorical senses on each level within these three sub-disciplines[13]. For a detailed analysis of rhetorical senses see Hussein Abdul-Raof's book (Abdul-Raof, 2012) which explores the history, disciplines, order

[12] For each rhetorician, we provide the date of death both in the Islamic calendar and in the Gregorian calendar.
[13] Note however that only rhetorical senses at the sentence and the discourse level are important for our task.

and pragmatic functions of Arabic figures of speech. See also (Abubakre, 1989) (Al-Jarim and Amine, 1999) (Sloane, 2001) (Musawi and Muhsin, 2001) and (Owens, 2006) for additional readings.

The first sub-discipline علم البيان/Elm AlbyAn, known as figure of speech, is the art of expressing a thought with clarity. It concerns "the eloquent discourse that uncovers the emotional feelings of the communicator and exposes them to the addressee" (Abdul-Raof, 2012). It enables the speaker to express figurative and not literal usages through which we can discern a single meaning by expressing it clearly in different ways. Figure 2.3 presents the major constituents of Arabic figures of speech. It is not demonstrated in this figure but each constituent is further decomposed into sub-constituents. Among the main figures, we cite *simile* (تشبيه/t$byh), which is an imaginative comparison which is usually introduced by مثل/mvl/*like* or ك/k/*as* (cf. Example 41), *metaphor* (الاستعارة /AlAstEArp) (cf. Example 42) and *metonymy* (المجاز المرسل/AlmjAz Almrsl).

(41) حل الرئيس مثل القمر في الاجتماع العام

Hl Alr}ys mvl Alqmr fy AlAjtmAE AlEAm

The president comes to the main meeting like a moon

(42) بيت احمد كثير الجرذان

*byt AHmd kvyr Aljr*An*

Direct translation: Ahmed's house contains many rats

Meaning: Ahmed's house is untidy and unclean

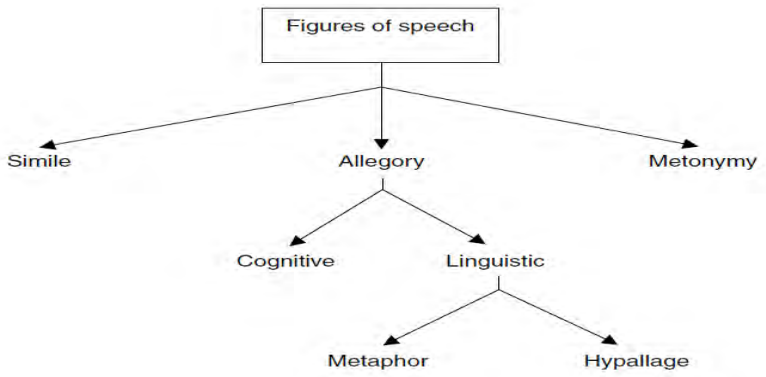

Figure 2.3. Figures of speech (**علم البيان**/Elm AlbyAn) in Arabic rhetoric (Abdul-Raof, 2012).

The second discipline concerns the syntax-semantic interface and discourse analysis. It is "the juxtaposition of sentence constituents in various word orders that leads to distinct pragmatic significations" (Abdul-Raof, 2012). It is divided into 17 sub-disciplines as shown in Figure 2.4.

For example, *restriction* is generally realized by coordination particles such as: إلا/<lA/*except*, غير/gyr/*unless*, عدى/EdY/*unless*, etc., as in Example 43. *Conjunction* aims at preserving the cohesion process through conjunction between individual words, as in Example 44 and between phrases of more than one lexical item, as in Example 45.

(43) خرج جميع التلاميذ من القسم إلا أحمد

xrj jmyE AltlAmy mn Alqsm <lA>Hmd*

All the students have left the classroom except Ahmed

(44) تعطى القروض حسب الاحتياجات والأولويات

tETY AlqrwD Hsb AlAHtyAjAt wAl>wlwyAt

Loans are given according to needs and priorities

(45) أكلت سمكة وشربت عصيرا

>klt smkp w$rbt ESyrA

I ate a fish and drank juice

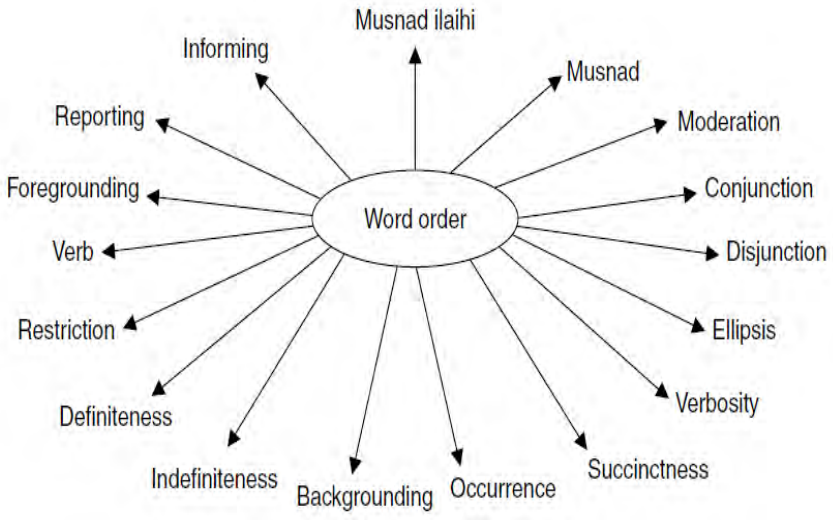

Figure 2.4. Word order (**علم المعاني**/Elm AlmEAny) in Arabic rhetoric (Abdul-Raof, 2012).

Finally, the last discipline refers to the linguistic and stylistic mechanisms that aim to provide ornamentation to Arabic discourse. We distinguish both semantic and lexical embellishments. Semantic embellishment includes around 30 mechanisms such as *antithesis, asterism, observation, quotation* and *rhetorical question*, as shown in Figure 2.5. For instance, *antithesis* refers to the combination of two opposite things whether they are allegorical or non allegorical

(Abdul-Raof, 2012), as in Example 46 where the non negated antithesis is achieved by the antonyms (متحمس/mtHms/*enthusiastic*) and (متهاون/mthAwn/*indifferent*). *Exordium* on the other hand sets the scene for the addressee by referring to the major areas he is going to speak about, as in Example 47 where the first two sentences describe the background of the commentary. *Scholastic approach* is related to the argumentation and debate where the communicator attempts to provide substantiating cognitive evidence to prove his point of view, as in Example 48. Finally, lexical embellishment includes 16 subcategories, among which we cite *alliteration*, where the communicator uses a number of words which initial letters are successively identical, and *assonance* which refers to the agreement in the last letter(s) of two propositions.

(46) احمد متحمس في دراسته و متهاون في امتحاناته.

AHmd mtHms fy drAsth w mthAwn fy AmtHAnAth.

Ahmed is enthusiastic about his studies and indifferent about his exams.

(47) ذهبت إلى السينما أمس. شاهدت "الياسمين الأزرق ". أنه شيء رائع.

**hbt <lY AlsynmA >ms. $Ahdt "AlyAsmyn Al>zrq". >nh $y' rA}E.*

I went to the movie theater yesterday. I saw 'Blue Jasmine'. It was awesome.

(48) لو حافظ على حماسه في دراسته، لتفوق في امتحاناته.

lw HAfZ ElY HmAsh fy drAsth, ltfwq fy AmtHAnAth.

If he has maintained his enthusiasm about his studies, he would have succeeded in his exams.

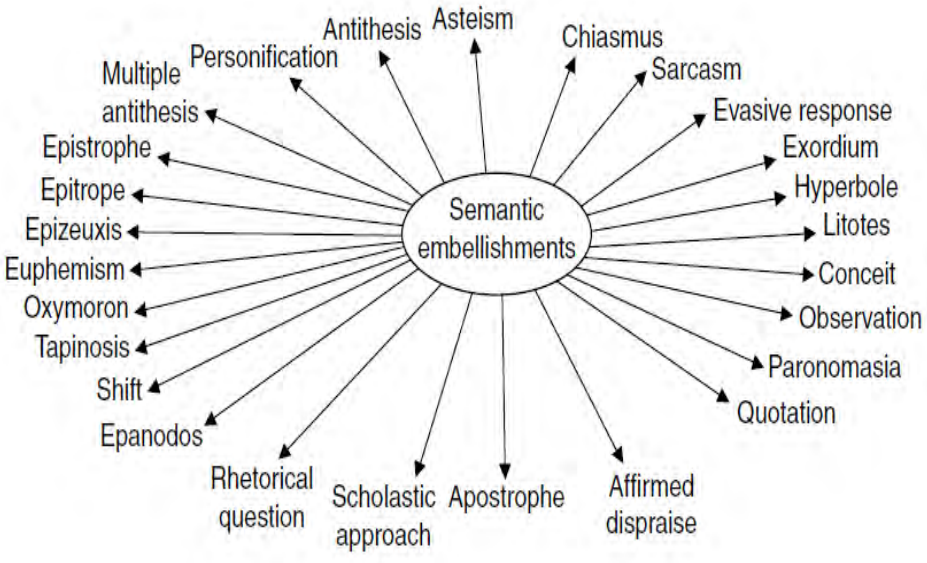

Figure 2.5. Semantic embellishment (**علم البديع** /Elm AlbdyE) in Arabic rhetoric (Abdul-Raof, 2012).

3.2. Building a new hierarchy of Arabic discourse relations

3.2.1. General methodology

Each theory defines its own inventory of discourse relations. There is no consensus neither on the number of these relations nor on their classification. Hence, the characterization of a unique set of relations is both suitable to accurately describe all attachments in a corpus and to granularity appropriate for manual annotation. This may explain why there is no standardized taxonomy of discourse relations to be applicable across languages (see (Zufferey *et al.*, 2012) for a discussion on multilingual annotation schemes for discourse relations). What seems to be undeniable however relations have a certain semantic or interpretive effects? But most theories do not individuate relations on the basis of these effects. SDRT insists on a semantic characterization of relations, which provides a method to verify whether two relations are similar, one entails the other, are independent or are incompatible. We adopt here this approach in the annotation manual to describe a relation independently from its possible DCs, (too often ambiguous, especially in the Arabic language), and to focus on what distinguishes relations that are often confused.

In this chapter, we rely on the previous set of 19 relations defined within the Annodis project. They are grouped into seven categories: *Causation, Structural, Logic, Reported Speech, Exposition/Narration, Elaboration,* and *Commentary*. Among these relations, we focus on semantic relations between entities from the propositional content of the clauses (we discarded meta-talk (or pragmatic) relations that link the speech acts expressed in one unit and the semantic content of another unit that performs). Table 2.3 summarizes these relations along with their definitions.

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 choose to reduce the number of these classes and, at the same time, to adapt Annodis relations to the Arabic specificities. Therefore, we decided to build a new classification by flattening the Annodis hierarchy so as to analyze the semantic of each relation relying on Arabic rhetoric literature and corpus analysis. Our new hierarchy is composed 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 the Figure 2.6.

Three experts in Arabic linguistics built our 4 levels hierarchy. 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 (cf. Table 2.3). We name this initial set *Annodis_set*. We have also provided a description of Arabic rhetorical senses as previously defined in earlier studies in Arabic rhetoric (cf. Section 4.1). We will refer to this set by *Arabic_set*. We asked the experts to collapse these two sets using corpus analysis focusing on both explicit and implicit marked rhetorical relations. The data used by experts is composed of 10

newspaper documents (706 EDUs) extracted from the syntactically annotated Arabic Treebank (ATB v3.2 part3) as well as 25 documents (924 EDUs) extracted from Tunisian Elementary School Textbooks (EST) built by our own. The main goal behind exploiting two corpus genres in this stage is to enable experts to better capture the semantic of discourse relations. Indeed, EST documents are usually well structured with simple style of writing. Rhetorical relations are often marked. Sentences are short (around 5.6 words per sentence) with a quite simple syntactic structure. Document length is also short (around 10 sentences per document). Contrary to EST, ATB documents are longer (around 25 sentences per document) and sentences are syntactically more complex (cf. Section 2 for a detailed description of ATB documents).

Annodis Relations	**Definitions**
Causation	
Explanation (S)	-The main eventuality of β is understood as the cause of the eventuality in α.
Goal (S)	-β describes the aim or the goal of the event described in α.
Result (C)	-The main eventuality of α is understood to cause the eventuality given by β.
Structural	
Parallel (C)	-α and β have similar semantic structures and requires α and β to share a common theme.
Continuation (C)	-α and β elaborate or provide background to the same segment.
Contrast (C)	-α and β have similar semantic structures, but contrasting themes or when one constituent negates a default consequence of other.
Logic	
Conditional (C)	-α is a hypothesis and β is the consequence. It can be interpreted as: if α then β.
Alternation (C)	-α and β are related by a disjunction.
Reported Speech	
Attribution (S)	-Relates a communicative agent stated in α and the content of a communicative act introduced in β.
Exposition/Narration	
Background (S)	-α constituent β provides information about the surrounding state of affairs in which the eventuality mentioned in α occurs.
Narration (C)	-α and β introduce an event and the main eventualities of α and b occur in sequence and have a common topic.
Flashback (S)	-Is a equivalent to Narration(β,α). The story is told in the opposite temporal order.
Frame (S)	-α is a frame and β is on the scope of that frame.
Temporal Location (S)	- β contains a temporal localization of the event described in α.
Elaboration	
Elaboration (S)	-β provides further information (a subtype or part of) about the eventuality introduced in α.
E-Elaboration (S)	-β gives more details about an entity introduced in α.
Commentary (S)	-β provides an evaluation of the content associated with α.

Table 2.3. SDRT relations in Annodis project. α and β stand respectively for the first and the second arguments of a relation. (S) and (C) correspond respectively to subordinating and coordination relations.

The collapsing procedure works as follows: For each relation R in the *Annodis_set*, experts look for its corresponding rhetorical senses in the *Arabic_set.* Five situations may occur:

-There is an *exact* correspondence between the semantic of R and its equivalent in *Arabic_set*. Then the relation R is selected and experts analyze how R is marked in the corpus in order to

give a preliminary list of its discourse connectives. 9 relations feat in this case. They are dotted and underlined in Figure 2.6.

-There is *only a partial* correspondence between the semantic of R and its equivalent in *Arabic_set.* Then, the relation R is selected and experts specify its semantic according to the particularities of the Arabic language. There are two relations in this case. They are followed by a double star (**) in Figure 2.6.

-The semantic of R covers different senses in the *Arabic_set* and each sense has its own realization in the corpus. R needs then to be *specialized.* New relations are added and experts were asked to define their semantics along with their corresponding discourse connectives. Consequently, we obtained 4 relations that are underlined in Figure 2.6.

-A group of relations from the *Annodis_set* correspond to one sense in the *Arabic_set* and in addition these relations are often not differentiated in the corpus. In this case, experts are asked to *generalize* these relations and create a new top-level relation. One relation corresponds to this case. It is underlined in bold font in Figure 2.6.

-There is *no correspondence* of R in the *Arabic_set* and no instance of R in the corpus. R is discarded.

After applying this algorithm, experts were asked to identify new relations. Only one relation was added. It is in underlined twice in Figure 2.6.

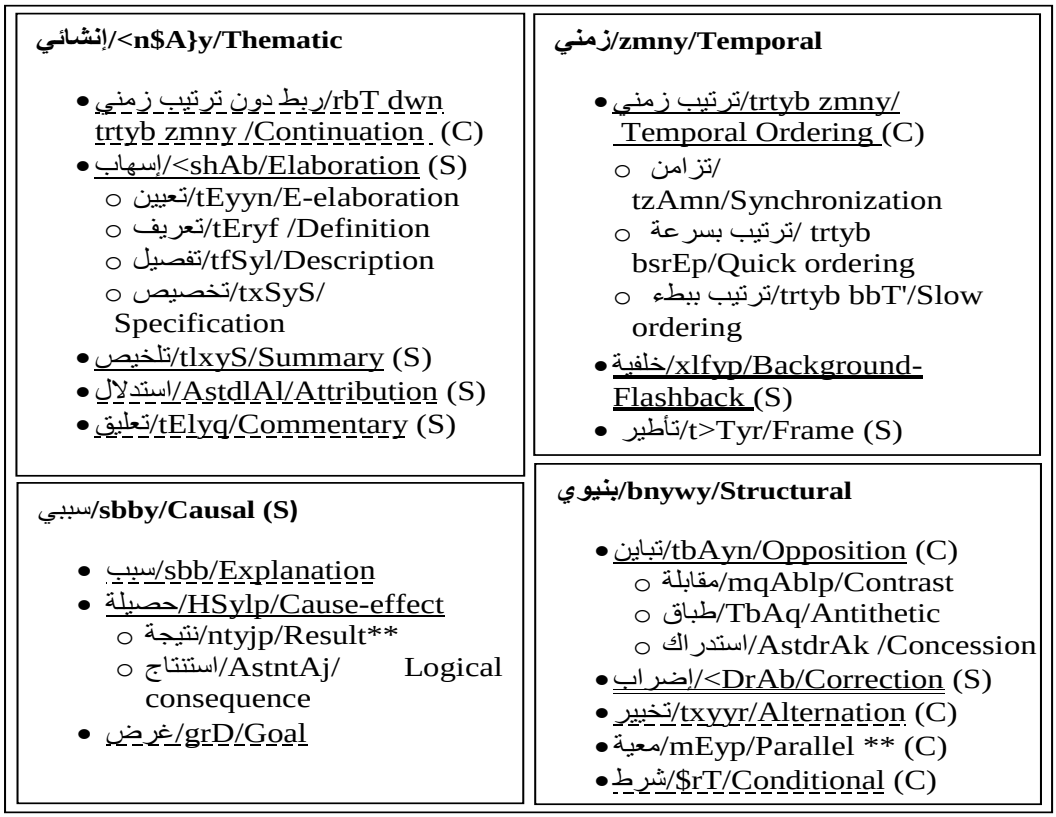

Figure 2.6. Hierarchy of Arabic discourse relations used in the ADTB corpus. (S) and (C) correspond respectively to subordinating and coordination relations.

3.2.2. A detailed description of our hierarchy

In this section, all relations are given following the Arabic reading order, from the right to the left, i.e. the notation:

(b,a)R

Indicates that *a* is the first argument and *b* is the second argument of the relation *R*. Complex segments (CDU) are between square brackets, i.e. the notation:

(c,[b,a])R

Indicates that CDU [*a*,*b*] is the first argument of *R*. Finally, the notation [*a*-*d*] indicates that CDU [*a*-*d*] is composed of four segments: *a*, *b*, *c* and *d*.

The انشائي/<n$A}y/*Thematic* class.

This class groups one relation that have a coordination function (ربط دون ترتيب زمني/rbT dwn trtyb zmny/*Continuation*) and where arguments are of equal importance, three subordinating relations (تلخيص/tlxyS/*Summary*, استدلال/AstdlAl/*Attribution*, and تعليق/tElyq/ *Commentary*), and one subordinating subclass (إسهاب/<shAb/ *Elaboration*). It is composed of eight discourse relations:

- ربط دون ترتيب زمني/rbT dwn trtyb zmny/*Continuation*. Literally, it means coordination without temporal order. This relation has the same semantic as *Continuation* in SDRT and imposes that its two arguments share the same topic and generally realize the same rhetorical function with a preceding segment (for instance, in case of تفصيل/tfSyl/*Description* or سبب/sbb/*Explanation*, (cf. above)). It is a veridical relation and it is usually signaled in Arabic by commas, as in Example 49 or by the DCs و/w/*and*, as in Example 50.

(49) [تغيب المعلم عن الدرس،]$_1$ [قدم شهادة طبية]$_2$ [وبرر غيابه.]$_3$

[tgyb AlmElm En Aldrs,]$_1$ [qdm $hAdp Tbyp]$_2$ [wbrr gyAbh.]$_3$

[The teacher was absent from his course,]$_1$ [he presented a medical certificate]$_2$ [and justified his absence.]$_3$

([3,2],1)*Background-Flashback* /xlfyp/خلفية

(3,2)*Continuation*/rbT dwn trtyb zmny/ربط دون ترتيب زمني

(50)[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 entertaining movie.]$_1$ [My brother laughed]$_2$ [and had a good time.]$_3$

([3,2],1) *Description*/tfSyl/تفصيل

(3,2) *Continuation*/rbT dwn trtyb zmny/ربط دون ترتيب زمني

- The إسهاب/<shAb/*Elaboration* class refers to a group of discourse relations that connect utterances describing the same state of affairs: reformulation (restatement), specification (particularization), generalization, etc. This class is equivalent to the relation *Elaboration* in SDRT. However, we have further specialized this class into 4 relations:

- تعيين/tEyyn/*E-elaboration* is equivalent to *Entity Elaboration* in SDRT. In Arabic, it is 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* …, as in Example 51.

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

[qAmt qwAt Aljy$, [Alty AqtHmt Almnzl,]$_2$ bAEtqAl jmyE AlAfrAd]$_1$

[The army troops, [that broke into the house,]$_2$ have arrested all the family members]$_1$

(2,1) *E-elaboration* /tEyyn/تعيين

- تعريف/tEryf/*Definition*. It holds when the second argument defines an entity or a concept introduced in the first argument. Some DCs include: هو/hw/*he*/*him*/*it*, هي/hy/*she*/*her*/*it* …, as in Example 52.

(52) [كنت ألعب بالكرة،] $_1$ [هي عبارة عن مطاط دائري مملوء بالهواء.] $_2$

[knt >lEb bAlkrp,]$_1$ [hy EbArp En mTAT dA}ry mmlw' bAlhwA'.]$_2$

[I was playing with the ball,]$_1$ [it is a spherical rubber filled with air.]$_2$

(2,1) *Definition*/tEryf/تعريف

- تفصيل/tfSyl/*Description* indicates that the second argument gives further information or details about the situation or the event presented in the first argument, as in Example 53. This relation is generally implicit.

(53) [جميع أفراد العائلة متعبون:] $_1$ [الأب متعب من شغله،] $_2$ [والأم منهكة من الأعمال المنزلية .] $_3$

[jmyE >frAd AlEA}lp mtEbwn:]$_1$ [Al>b mtEb mn kvrp Al$gl,] $_2$ [wAl>m mnhkp mn Al>EmAl Almnzlyp.]$_3$

[All family members are tired:]$_1$ [the father is tired because of his job,]$_2$ [and the mother is exhausted because of housework.]$_3$

([3-2],1) *Description*/tfSyl/تفصيل

(3,2) *Continuation*/rbT dwn trtyb zmny/ربط دون ترتيب زمني

- تفصيل/tfSyl/*Description* also covers cases of تمثيل/tmvyl/*Illustration* and تشبيه/t$byh/*Simile* where authors provide examples to illustrate his idea. Main DCs are: ك/k/*like*, كأن/k>n/*as*, كما/kmA/*as* …, as in Example 54 and Example 55.

(54) [أكل الطفل المربى بشراهة] $_1$ [كأنه لم يذقه قط.] $_2$

*[>kl AlTfl AlmrbY b$rAhp]$_1$ [k>nh lm y*qh qT.]$_2$*

[The child eat jam greedily]$_1$ [as if he did never taste it before.]$_2$

تفصيل/tfSyl/ *Description* (2,1)

(55) [حدث ذلك بالضبط] $_1$ [كما حدث في استراليا العام الماضي] $_2$

*[Hdv *lk bAlDbT]$_1$ [kmA Hdv fy AstrAlyA AlEAm AlmADy]$_2$*

[This happened exactly]$_1$ [as it did in Australia last year]$_2$

تفصيل/tfSyl/ *Description* (2,1)

- تخصيص/txSyS/*Specification* indicates that the second argument elaborates on a portion or a part of the first argument. This relation is generally implicit, as shown in Example 56. When it is marked, it is signaled by خاصة/xASp/*especially*, بالخصوص/bAlxSwS/*in particular*, خصوصا/xSwSA/*especially*, وبالأخص/wbAl>xS/*in particular*…, as in Example 57.

(56) [قامت الدولة بطرح برامج جديدة]$_1$ [مشاريع تربوية ورياضية.] $_2$

[qAmt Aldwlp bTrH brAmj]$_1$ [m$AryE trbwyp w ryADyp.]$_2$

[The government has proposed new programs]$_1$ [educational and sport projects]$_2$

تخصيص/txSyS/*Specification* (2,1)

(57) [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* (2,1)

– تلخيص/tlxyS/*Summary* indicates that the second segment summarizes the story introduced in previous segments. In Arabic, it generally holds between blocs of EDUs and an EDU that concludes all information presented in this bloc. This relation has the same semantic as combining the relations *Description or Continuation* and *Commentary*. However, we choose to add a new relation to take into account the complexity of the discourse structure. Main DCs are: وخلاصة القول/wxlASp Alqwl/*in sum*, الخلاصة/AlxlASp/*the summary*, خلاصة الأمر/xlASp Al>mr/*in sum*, نلخّص/nlx~/*to summarize*, نستخلص أنّ/nstxlS >n~/*to summarize*…, as in Example 58.

(58) [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.]$_1$ […]$_x$ [And in sum, all his adventures were exciting.]$_{x+1}$

(x,1) *Description*/tfSyl/تفصيل

(x+1,[x-1]) *Summary* /tlxyS/تلخيص

— استدلال/AstdlAl/*Attribution*. It is equivalent to *Attribution* in SDRT. It is generally marked by typographical signs like ':', '«', and '»' or by lexical triggers which are mainly reporting speech verbs, such as قال/qAl/*say*, أكد/>kd/*confirm*, صرح/SrH/*say*/*assert*, أوضح/<AwDH/*explain*, اعلن/AEln/*announce*, …, as in Example 59.

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

[qAl >Hmd:]$_1$ [«<n AlmbArAp kAnt SEbp»]$_2$

[Ahmed said:]$_1$ ["the match was difficult"]$_2$

(2,1) *Attribution* /AstdlAl/استدلال

— تعليق/tElyq/*Commentary* corresponds to *Commentary* in SDRT[14]. Commentary can be تفضيل/tfDyl/*preference*, مدح/mdH/*praise* or ذم/*m/*vitriol*, as in Example 60.

(60) [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 under the expectations.]$_2$

(2,1) *Commentary*/tElyq/تعليق

The زمنية/zmnyp/*Temporal* class.

It groups relations that impose a temporal ordering between the events introduced in their arguments. It is composed of three main subclasses: ترتيب زمني/trtyb zmny/*Temporal Ordering*, خلفية/xlfyp/*Background-Flashback*, and تأطير/t>Tyr/*Frame*, with a total of five relations.

— ترتيب زمني/trtyb zmny/*Temporal Ordering*. In this sub-class, arguments need to share the same topic. In addition, it requires a temporal precedence of the eventualities e1 and e2 introduced in the two segments. It is a coordinating relation close to *Narration* in SDRT. However, according to the duration or the time interval *t* between the events e1 and e2, we distinguish 3 cases:

[14] Note that this relation does not figure in the Annodis relation set. However, it was already defined in Discor (the SDRT English annotation campaign).

- تزامن/tzAmn/*Synchronization*. This relation holds when the events e1 and e2 occurs at the same time and the two events are triggered by different subjects. Main DCs are: فى نفس الوقت/fY nfs Alwqt/*at the same time*, حينها/HynhA/*meanwhile*, في تلك الحظة/fy tlk AllHZp/*at that moment*, كل ما/kl mA/*whenever*, متى ما/mtY mA/*whenever*, في غضون ذلك/fy gDwn *lk/*meanwhile*, في هذا الأثناء/fy h*A Al>vnA'/*meanwhile*, …, as in Example 61.

(61) [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,]$_1$ [meanwhile the teacher arrived]$_2$

(2,1) *Synchronization*/tzAmn/تزامن

- ترتيب بسرعة/trtyb bsrEp/*Quick ordering*. It holds in two main situations: (1) the event e2 occurs at a short interval time *t1* after the event e1, i.e. an immediate time without delay (cf. Example 62) and (2) the pre-state of the eventuality e2 overlaps with the post-state of the eventuality e1 (cf. Example 63). This relation is mainly signaled by the DCs ف/f/*so/then/just/after*, كاد/>w$k/*nearly*, أوشك/>w$k/*nearly*, حتى/HtY/*until*, قبيل/qbyl/*just before*, etc.

(62) [أكمل المعلم الدرس،] $_1$ [فرن الجرس.] $_2$

[<kml AlmElm Aldrs,]$_1$ [frnn Aljrs.]$_2$

[The teacher has finished the lesson,]$_1$ [just after, the bell rang.]$_2$

(2,1) *Quick ordering*/trtyb bsrEp/ترتيب بسرعة

(63) [أوشك الفريق التونسي على الفوز،] $_1$ [حتى سجل الفريق المنافس هدفا.] $_2$

[>w$k Alfryq Altwnsy ElY Alfwz,]$_1$ [HtY sjl Alfryq AlmnAfs hdfA.]$_2$

[The Tunisian team almost won,]$_1$ [when the opposing team has scored a goal.]$_2$

(2,1) *Quick ordering*/ trtyb bsrEp/ترتيب بسرعة

- ترتيب ببطء/trtyb bbT'/*Slow ordering*. It holds when the event e2 occurs at an interval time *t2>t1* after the event e1, i.e. there is a temporal gap between the events denoted by the verbs in the arguments. This relation is mainly signaled by the DC ثم/vm/*afterward*, as in Example 64.

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

[<kml AlmElm Aldrs]$_1$ [vm xrj jmyE AltlAmy mn Alqsm]$_2$*

[The teacher has finished the lesson]$_1$ [afterward all the students have leaved the classroom]$_2$

(2,1) *Slow ordering*/ trtyb bbT'/ ترتيب ببطء

Temporal ordering relations can also hold in case of several co-occurring events, as in Example 65.

(65) [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 burned public institutions,]$_1$ [then shops,]$_2$ [then houses]$_3$

(2,1) *Slow ordering*/trtyb bbT'/ ترتيب ببطء

(3,2) *Slow ordering*/trtyb bbT'/ ترتيب ببطء

— خلفية/xlfyp/*Background-Flashback*. The Arabic word خلفية/xlfyp means the scene or the event that forms a setting for a main event or a main state. Thus, it covers the semantic of *Background* (which is often signaled by aspectual shift, i.e., a shift from an event to a state, or a state to an event) as well as the semantic of *Flashback* (an interruption of chronological sequence by interjection of events of earlier occurrence). In Arabic, it is mainly triggered by clauses introduced by subordinating conjunctions such as على الرغم/Alrgm ElY/*although*, وبالرغم/wbAlrgm/*although*, مع أنّ/>n~ mE/*even if*, as in Example 66, or by DCs like من قبل /mn qbl/*previously*, سابقا/sAbqA/ *previously*, as in Example 67.

(66) [كنت أركض بصعوبة،] $_1$ [كانت الساحة ممتلئة] $_2$

[knt >rkD msrEA,]$_1$ [kAnt AlsAHp mmtl}p]$_2$

[I ran hardly,]$_1$ [the place was crowded]$_2$

(2,1) *Background-Flashback* /xlfyp/خلفية

(76) [لن أعود لشرح الدرس مرة أخرى.] $_1$ [لقد شرحته من قبل.]$_2$

[ln >Ewd l$rH Aldrs mrp >xrY.]$_1$ [lqd $rHth mn qbl.]$_2$

[I won't explain this lesson again.]$_1$ [I had explained it previously.]$_2$

(2,1) *Background-Flashback*/xlfyp/خلفية

— تأطير/t>Tyr/*Frame*. This relation is similar to the relation *Frame* in SDRT. It is a subordination relation that indicates that an event which is introduced in the second argument occurs in the scope of a temporal frame تأطير زماني/t>Tyr zmAny, a spatial frame تأطير مكاني/t>Tyr mkAny (cf. Example 68) or a topic frame الجوهر/Aljwhr (cf. Example 69). Some DCs are: من/mn/*from*, إلى/<lY /*to*, على/ElY/*on*, في /fy/*in*, etc.

(68) [في ركن من البيت،] $_1$ [قمت بتلوين هذه الصورة.] $_2$

*[fy rkn mn Albyt,]$_1$ [qmt btlwyn h*h AlSwrp.]$_2$*

[In a corner of the house,]$_1$ [I painted this picture.]$_2$

(2,1) *Frame*/t>Tyr/ تأطير

(69) [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. system,]$_1$ [the student studies a three years Bachelor degree,]$_2$ [then two years Master degree,]$_3$ [then three years Doctorate.]$_4$

([4-2],1) *Frame*/t>Tyr/ تأطير

(3,2) *Slow ordering*/trtyb bbT'/ ترتيب ببطء

(4,3) *Slow ordering*/trtyb bbT'/ ترتيب بطء

The سببي/sbby/*Causal* class.

This top-level class covers relations, which semantic is to specify why and how an event happens. It groups three subclasses: *Explanation*, *Cause-effect*, and *Goal* with a total of four subordinating relations. Moreover, this class includes relations where the second utterance gives "support" to the first one, including causal explanation, justification, motivation, etc. (Mann and Thompson, 1988; Asher and Lascarides, 2003; Danlos and Gaiffe, 2004). It is composed of three subclasses:

- سبب/sbb/*Explanation*. This relation is similar to *Explanation* in SDRT. It indicates that the event or the state in the second argument is the cause of the event or a state in the first argument. *Explanation* can be explicitly marked using DCs such as: لما/lmA/*whereas*, لأن/l>n/*because*, والسّبب/wAls~bb/*as a cause of*, بسبب/bsbb/*because*, as in Example 70. It can also be implicit, as in Example 71, where Al-Masdar[15] بحثا/bHvA/*looking for* explains why Ahmed went to the library.

(70) [رجعت مسرعا إلى البيت]$_1$ [بسبب تهاطل الأمطار.]$_2$

[rjEt msrEA <lY Albyt]$_1$ [bsbb thATl Al>mTAr.]$_2$

[I returned quickly at home]$_1$ [because it was raining.]$_2$

(2,1) *Explanation* /sbb/سبب

[15] Al-masdar is a verbal noun construction, frequent in Arabic. It names the action denoted by its corresponding verbs.

(71) [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]$_1$ [looking for the book of mathematic.]$_2$

(2,1) *Explanation* /sbb/سبب

— حصيلة/HSylp/*Cause-effect*. This sub-class groups relations that relate a cause to its effect and thus, is the dual of the relation *Explanation*. The experts have identified 3 relations here:

- نتيجة/ntyjp/*Result*. It is close to *Result* in SDRT, i.e. it covers cases where an event e2 in the second argument is the consequence of an occurring event e1 in the first argument. Main DCs in Arabic are: نتيجة لذلك/ntyjp l*lk/*for this result*, والنتيجة/wAlntyjp/*and the result*, إذن/<*n/*so*, لهذا/lh*A/*therefore*, as in Example 72. In addition to this definition, this relation also indicates a change or an evolution (تحول/tHwl) of a state introduced in the second argument after an occurring event in the first argument. This case is specific to the Arabic language and has no equivalence in French or English. It is usually lexically marked by specific Arabic verbs such as: صار/SAr/*become*, أصبح/>SbH/*become*, بات/bAt/*become*, أمسى/>msY/*become*, أضحى/>DHY/*become*, غدت/gdt/*become* …, which have no exact translation in English and means roughly *to become*, as in Example 73.

(72) [معظم الناس لا يدركون تماما هذه المعلومات،] $_1$ [والنتيجة هم تحصلو على تغطية جزئية في نهاية المطاف.] $_2$

*[mEZm AlnAs lA ydrkwn tmAmA h*h AlmElwmAt,]$_1$ [wAlntyjp hm tHSlwn ElY tgTyp jz}yp fy nhAyp AlmTAf.]$_2$*

[Most people are not fully aware about this information,]$_1$ [as a result, they have only a partial coverage of the situation]$_2$

(2,1) *Result*/ntyjp/نتيجة

(73) [جاع القط] $_1$ [فصار يموء] $_2$

/jAE AlqT]$_1$ [fSAr ymw']$_2$

[The cat was hungry,]$_1$ [he started meowing]$_2$

(2,1) *Result*/ntyjp/نتيجة

- استنتاج/AstntAj/*Logical consequence*. This relation indicates that the result introduced in the second segment is an evidence, a justification or a logical consequence on which a judgment of a conclusion may be based. Main DCs are: نستنتج/nstntj/*we conclude*, الاستنتاج هو/AlAstntAj hw/*the conclusion is*, ومن هنا/wmn hnA/*hence*, ذلك هو/*lk hw/*this is* …, as in Example 74.

(74) [دخل السجن] $_1$ [وترك عائلته] $_2$ [وترك دراسته.] $_3$ [ذلك هو من يتعاطى المخدرات.]$_4$

*[dxl Alsjn]$_1$ [wtrk EA}lth]$_2$ [wtrk drAsth.]$_3$ [*lk hw mn ytEAT AlmxdrAt.]$_4$*

[He went into prison]$_1$ [left his family]$_2$ [and abandoned his studies.]$_3$ [This is what happens to those who take drugs]$_4$

([3-1],4) *Logical consequence* /AstntAj/استنتاج

(2,1) *Continuation*/rbT dwn trtyb zmny/ربط دون ترتيب زمني

(3,2) *Continuation*/rbT dwn trtyb zmny/ربط دون ترتيب زمني

- غرض/grD/*Goal*. It has the same semantic as *Goal* in SDRT. This relation has common discourse markers with the other previous relations of the تبرير/tbryr/*Causal* class. For instance, the DCs ل/l/*for*/*to* can be combined with other DCs such as لأن/l>n/*inasmuch*, لذلك/l*lk/*given*..., as in Example 75.

(75) [اضرب الباحثون]$_1$ [ليُظهروا استياءهم] $_2$

[ADrb AlbAHvwn]$_1$ [lyuZhrwA AstyA'hm]$_2$

[The researchers are on strike]$_1$ [to show their dissatisfaction]$_2$

(2,1) *Goal* /grD/غرض

The بنيوي/bnywy/*Structural* class.

We have here five subclasses with a total of seven relations.

- تباين/tbAyn/*Opposition* contains three relations whose semantic is that the two arguments are in opposition.
- مقابلة/mqAblp/*Contrast*. It is equivalent to *Contrast* in SDRT. In Arabic, it is introduced by specific DCs 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 Example 76.

(76) [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.]$_2$

(2,1) *Contrast*/mqAblp/مقابلة

- طباق/TbAq/*Antithetic* means that the two arguments are diametrically opposed. In Arabic, it holds when there is a verb in the first argument and its negation in the second argument (cf. Example 77) or when the two verbs are antonyms as in Example 78.

(77) [يأكل رقائق البطاطا المحمرة]$_1$ [ولا يأكل البطاطا المقلية]$_2$

[y>kl rqA}q AlbTATA AlmHmrp]$_1$ [wlA y>kl AlbTATA Almqlyp]$_2$

[He eat chips]$_1$ [and he does not eat fried potatoes]$_2$

طباق/TbAq/*Antithetic* (2,1)

(78) [يضحك أخي]$_1$ [ويبكي.]$_2$

[yDHk >xy]$_1$ [wybky.]$_2$

[My brother laughs]$_1$ [and cries.]$_2$

طباق/TbAq/*Antithetic* (2,1)

- استدراك/AstdrAk/*Concession*. It indicates that the second argument is contrary to the expectation of the first argument. Main DCs are: لكن/lkn/*but*, غير أنّ/gyr >n~/*but*, إلا أنّ/<lA>n~/*but* …, as in Example 79.

(79) [حضر جميع الطلاب،]$_1$ [لكنّ سعيدٌ غائبٌ.]$_2$

[HDr jmyE AlTlAb,]$_1$ [lkn~ sEydN gA}bN.]$_2$

[All the students come,]$_1$ [but Said is absent.]$_2$

استدراك/AstdrAk/*Concession* (2,1)

- إضراب/<DrAb/*Correction*. It is similar to *Correction* in SDRT. Indeed, it links two segments that have common topics such that the focus of the second segment is inconsistent with the focus of the first argument, i.e. the second argument corrects the information given in the first argument. Main DCs include: إنّما/<n~mA/*however*, بل/bl/*however/but* …, as in Example 80.

(80) [لا يعد الفلم المسئ للرسول اهانة للمسلمين فقط ،]$_1$ [بل يعد اهانة لحرية التعبير]$_2$

[lA yEd Alflm Alms} llrswl AhAnp llmslmyn fqT,]$_1$ [bl yEd AhAnp lHryp AltEbyr]$_2$

[The movie that humiliates the Prophet does not only insults the Muslims,]$_1$ [but also it insults freedom of expression]$_2$

إضراب /<DrAb/ *Correction* (2,1)

- تخيير/txyyr/*Alternation*. This is a non veridical relation that has the same semantic as *Alternation* in SDRT, which is of a disjunction. It is a coordinating relation and is generally introduced in Arabic by إما/<mA/*either*, أو/>w/*or*, أم/>m/*or*, سواء/swA'/*either*, …, as in Example 81.

(81) [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]$_2$

(2,1) *Alternation*/txyyr/تخيير

— معية/mEyp/*Parallel*. It indicates that two segments share the same event and they have semantically similar constituents, as in Example 82. It is a coordinate relation close to *Parallel* in SDRT. However, in addition to this definition, this relation also holds in Arabic when each argument introduces two different events triggered by the same subject, and when these events must happen. This point is illustrated in Example 84 in which the events of *repairing the care* and *painting it* must occur before *selling it*. Main DCs include: و/w/*and*, معا/mEA/*together*, أيضا/>yDA/*too*, كما/kmA/*also*, etc.

(82) [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 on this solution,]$_1$ [and you also agree to apply it.]$_2$

(2,1) *Parallel* /mEyp/معية

— شرط/$rT/*Conditional*. It is equivalent to *Conditional* in SDRT. Its main DCs are: 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 Example 83 and Example 84.

(83) [إذا أصبح الطقس جميل،] $_1$ [سأخرج أتنزه.]$_2$

[>A >SbH AlTqs jmyl,]$_1$ [s>xrj >tnzh.]$_2$*

[If the weather will be nice,]$_1$ [I'll go for a stroll.]$_2$

(2,1) *Conditional* /$rT/شرط

(84) [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]$_3$

(3,[2,1]) *Conditional* /$rT/شرط

(2,1) *Parallel*/mEyp/معية

3.3. Annotation campaign

Two experts in discourse analysis[16] were asked to annotate our corpus. We provide them with a precise definition of the meaning of discourse relations (cf. section 4.2) and asked to them to insert relations between constituents. When appropriate, EDUs can be grouped to form complex discourse units. The relations were defined in semantic terms in the manual. The goal of the manual was the development of an intuition for each relation, suitable for the level of annotators. Occasional examples were provided, and we gave a list of few possible connectives 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 connectives, very frequent in Arabic.

Since our goal is to evaluate the feasibility of full discourse analysis of Arabic documents, our annotation manual details clearly what are 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 examines how well SDRT predicts the intuition of subjects, regardless of their knowledge of discourse theories. Main SDRT constraints concern: segment attachment (no isolated segment in the graph, attachment mainly follows the reading order of the document), right frontier principle (Asher and Lascarides, 2003) (cf. Figure 2.7), and structural constraints including accessibility, complex segments, no cycles, etc. (cf. Figure 2.8).

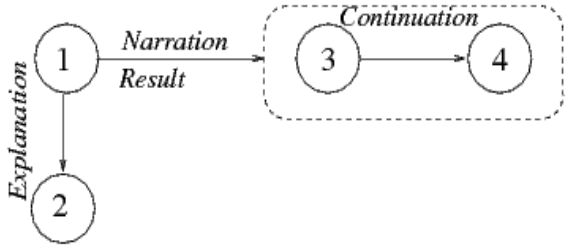

Figure 2.7. Right frontier principle. In this example, open attachment sites are the segment 4 and the CDU [3,4].

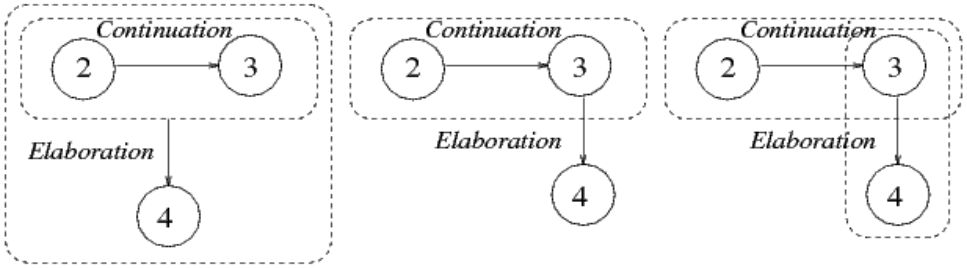

Figure 2.8. An example of a CDU constraint. Figures in the left and in the middle are correct configurations whereas the one in the right is not allowed because CDUs cannot overlap.

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

3.3.1. The corpus

We have randomly selected 90 documents from ATB. In order to avoid errors in determining the basic units (which would thus make the inter-annotator agreement study tedious), we have decided to discard the 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 approach using the Stanford classifier which 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 uses a rich lexicon (with more than 174 connectives) and a combination of punctuation, morphological and lexical features. It achieved an average F-score of 0.847, an average accuracy of 0.949 on token boundary recognition and an average accuracy of 0.769 on EDUs recognition after a post-processing step that corrected wrong end bracketing. See Chapter 3 for a detailed description of our segmentation principles of Arabic texts and for a presentation of our learning method.

3.3.2. Annotation procedure

We performed a three-step annotation where an intermediate analysis of agreement and disagreement between the two annotators were carried out. Annotators were first trained on 13 documents (911 EDUs). During the training phase, we noticed that the document length was a handicap since the annotation of a document can take two days given that making the task of connecting all the EDUs in the same whole structure is very tedious (we recall that each document has around 26 sentences[17] and 8 paragraphs). To overcome this problem, we decided to annotate separately the discourse structure of each paragraph in a document, and then to link these structures with the top-level relation إسهاب/<shAb/*Elaboration* by convention, in order to guarantee the connectivity of the resulting graph. After training, annotators were asked to double annotate the same 7 documents (462 EDU). The time needed to annotate the entire text was about 8 hours. This step allows computing inter-annotator agreements both in terms of attachment points and relation labeling. Given the good agreements reached in this second step (cf. Section 3.4), the experts were asked to annotate the rest of the corpus (70 documents) by consensus. Table 2.4 summarizes the characteristics of our gold standard corpus.

	Texts	**Size**	**Sentences**	**EDUs**	**Embedded EDUs**	**Words+ Punctuations**
ADTB	70	381ko	1 832	4 963	542 (9.16%)	39 746

Table 2.4. Characteristics of our gold corpus.

[17] Arabic discourse tends to use long and complex sentences, so we can easily find an entire paragraph without any punctuation mark.

Example 85 presents an annotated paragraph taken from the document ANN20020115.0003.

(85) [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 kilometers Southwest of Khost.]$_9$ [And it said:]$_{10}$ ["the bombing has lasted 48 hours."]$_{11}$

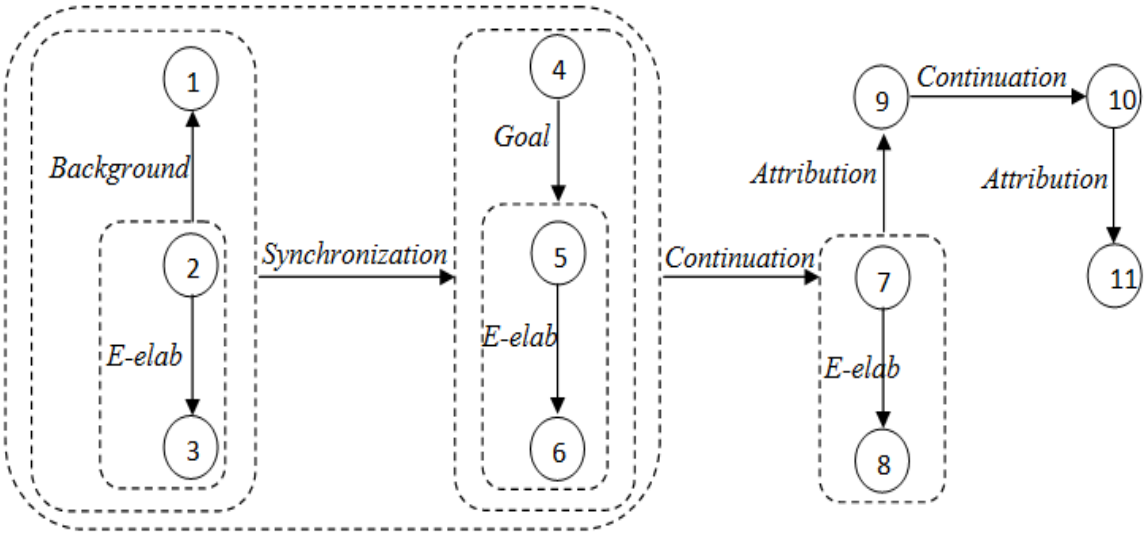

Figure 2.9. The discourse annotation for Example 85.

3.4. Results

3.4.1. Qualitative analysis

Discourse annotation consists in two stages: linking attachment points and labeling of the attachment arcs via discourse relations. Two inter-annotator agreements have 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 and obtained an F-measure of 0.890, which is good. Main disagreements came from non adjacent EDUs. Indeed, one annotator has tended to form CDUs more frequently while the other often produces "flat" structures. Figure 2.10 shows two discourse annotations for Example 86. We observe that the annotator on the left used to form less CDUs than the other annotator on the right, which causes one attachment error.

(86) [قال وزير الد فاع]$_1$ [ان نحو ستة جنود اميركيين وصلوا الى البلاد]$_2$ [وان الجنود، [حيث سيكونون مسلحين،]$_3$ يستطيعون الد فاع عن انفسهم.]$_4$

[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 Defense said]$_1$ [that six U.S. soldiers arrived in the country]$_2$ [and the soldiers, [when they will be armed,]$_3$ will be able to defend themselves.]$_4$

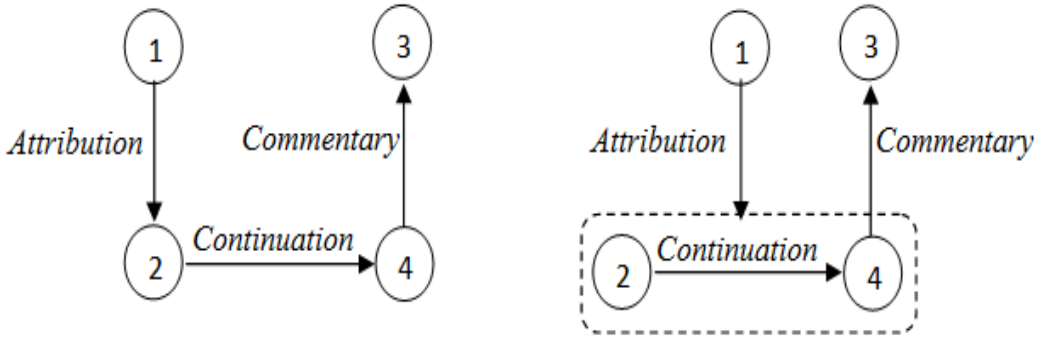

Figure 2.10. Two discourse annotations for Example 86

The used algorithm 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 example in Example 87, the annotation *Frame(1,2)* and *Continuation(2,3)* is equivalent to *Frame(1,[2,3])* and *Continuation(2,3)*.

(87)[في نظام التعليم الجامعي أ.م.د.،] 1 [يدرس الطالب ثلاث سنوات إجازة،] 2 [و يدرس 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 [and he studies two years Master's degree.]3

When commonly attached pairs are considered, we get a Cohen kappa of 0.750 for the full set of 24 relations, which is good. Here again, this kappa is computed without an accurate analysis of the equivalence between rhetorical structures. Our results are very encouraging. This proves that our hierarchy of discourse relations has an appropriate level of granularity and the definition of each relation in terms of its semantic effect, independently from its possible discourse markers, is adequate to avoid confusions. However, it should be noted that some relations are difficult to distinguish because they are triggered by the same discourse markers. We give below the most frequent cases of confusions.

– غرض/grD/*Goal* vs. سبب/sbb/*Explanation* (cf. Example 88).

(88) [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 to his patient a set of drugs]1 [to treat his pain and injury.]2

One annotator puts سبب/sbb/*Explanation* (1,2) while the other one puts غرض/grD/*Goal* (1,2). Here, the right annotation is the second one. Indeed, the intention introduced in the segment 2 explains why the doctor prescribed drugs to his patient. This does not mean that the patient will effectively take his treatments. Hence, 2 cannot explain 1.

– تأطير/t>Tyr/*Frame* vs. استدلال/AstdlAl/*Attribution* (cf. Example 89).

(89) [من قبل السيد احمد،] 1 [نزل الفريق الى دوري الدرجة الثانية.] 2

[mn qbl Alsyd AHmd,]1 [nzl Alfryq AlY dwry Aldrjp AlvAnyp.]2

[According to Mr Ahmed,]1 [the team down to the second division.]2

Annotator 1: تأطير/t>Tyr/*Frame*(1,2)

Annotator 2: استدلال/AstdlAl/*Attribution*(1,2)

In this example, the confusion between annotators comes from the word قبل/qbl. The first annotator considered that this word is a spatial-temporal preposition (*قَبل*/qabl), so he used

تأطير/t>Tyr/*Frame* relation. However, in the context of Example 89, this word (قِبَل/qibal) introduces a reported speech, which means *according to*.

– سبب/sbb/*Explanation* vs. تفصيل/tfSyl/*Description* (cf. Example 90)

(90) [لقد استغنيت عن هذا الكتاب،]$_1$ [انه لا يحتوي على معلومات قيمة،]$_2$

*[lqd Astgnyt En h*A AlktAb,]$_1$ [Anh lA yHtwy ElY mElwmAt qyymp,]$_2$*

[I don't need this book,]$_1$ [it don't contain important information,]$_2$

Annotator 1: تفصيل/tfSyl/*Description* (1,2)

Annotator 2: سبب/sbb/*Explanation* (1,2)

In this example, the first annotator considers that the second argument provides a description of the book whereas the second annotator considers that this segment explains why the book was not needed, which is the right interpretation. Example 90 presents an example of an implicit relation (سبب/sbb/*Explanation*).

It is mandatory to note that traditional confusion, which often holds between the relations *Explanation* and *Elaboration*, as observed in past SDRT-like annotation campaigns, is very rare in our case. This shows that our refinement of the relation إسهاب/<shAb/*Elaboration* into 4 relations seems to be useful for better disambiguation between this two cases. Overall, our results are higher compared to those obtained by Annodis (0.660 F-measure for attachment and a Cohen Kappa of 0.400 for relation labeling) mainly for three reasons: (1) our annotation manual is more constrained since we have provided annotators with a detailed description of how discourse structures should be, (2) our annotations were done by experts and (3) we restricted the full discourse structure annotation to one paragraph (around 20 EDUs) which implies less long distance attachments than in news texts or Wikipedia documents used for the Annodis campaign (around 60 EDUs).

3.4.2. Quantitative analysis

Our gold corpus is composed of 70 documents (39,746 words and punctuation). These documents have different sizes, varying from two paragraphs (12 sentences) to 10 paragraphs (88 sentences). The total number of EDUs is 4,963. Three months were needed to annotate our gold corpus by the two experts, by consensus. The total number of annotated discourse relations is 3,184. The distribution of these relations is presented in Table 2.5.

In the statistics presented in Table 2.5, 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%). Infrequent relations (less than 1%) are: تخيير/txyyr/*Alternation*, استنتاج/AstntAj/Logical *consequence*, تلخيص/tlxyS/*Summary*, مقابلة/mqAblp/*Contrast* and طباق/TbAq/*Antithetic*.

Discourse relations		Frequency	Percentage
إنشائي/<n$A}y/*Thematic*	ربط دون ترتيب زمني/rbT dwn trtyb zmny /*Continuation*	673	21.14%
	إسهاب/<shAb /*Elaboration*	727	22.83%
	تعيين/tEyyn/E-*Elaboration*	482	15.14%
	تعريف/ tEryf /*Definition*	50	1.57%
	تفصيل/ tfSyl /*Description*	147	4.62%
	تخصيص/txSyS/*Specification*	48	1.51%
	تلخيص/tlxyS/*Summary*	14	0.44%
	استدلال/AstdlAl/*Attribution*	412	12.94%
	تعليق/tElyq/*Commentary*	44	1.38%
	Total	**1,870**	**58.74%**
زمني/zmny/*Temporal*	ترتيب زمني/trtyb zmny/*Temporal Ordering*	195	6.12%
	تزامن/ tzAmn/*Synchronization*	82	5.58%
	ترتيب بسرعة/ trtyb bsrEp/*Quick ordering*	52	1.63%
	ترتيب ببطء/ trtyb bbT'/*Slow ordering*	61	1.92%
	خلفية/xlfyp/*Background-Flashback*	124	3.90%
	تأطير/t>Tyr/*Frame*	44	1.38%
	Total	**363**	**11.40%**
سببي/sbby/Causa	سبب/sbb/*Explanation*	111	3.49%
	حصيلة/HSylp/*Cause-effect*	158	4.96%
	نتيجة/ntyjp/*Result*	143	4.50%
	استنتاج/AstntAj/*Logical consequence*	15	0.47%
	غرض/grD/*Goal*	289	9.08%
	Total	**558**	**17.53%**
بنيوي/bnywy/*Structural*	تباين/tbAyn/*Opposition*	128	4.02%
	مقابلة/mqAblp/*Contrast*	27	0.85%
	طباق/TbAq/*Antithetic*	12	0.38%
	استدراك/AstdrAk /*Concession*	89	2.80%
	إضراب/<DrAb/*Correction*	44	1.38%
	تخيير/txyyr/*Alternation*	17	0.53%
	معية/mEyp/*Parallel*	93	2.92%
	شرط/$rT/*Conditional*	111	3.49%
	Total	**393**	**12.35%**

Table 2.5. Discourse relation distribution in the gold corpus**.**

Table 2.6 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.

Total number of relations	3,184
Argument type	
EDU	5,798 (91%)
CDU	570 (9%)
Discourse relation and EDU position	
Relations between adjacent EDUs	2,706 (85%)
Relations between non adjacent EDUs	478 (15%)
Discourse relation and Argument type	
R (EDU,EDU)	2,682 (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	2,382 (74.8%)
Implicit relations	802 (25.2%)

Table 2.6. Discourse relation and argument type in the gold corpus.

Moreover, Figure 2.11 presents the distribution of the top-level classes according to their argument types. The بنيوي/bnywy/*Structural* class contains the most number of CDUs in their arguments.

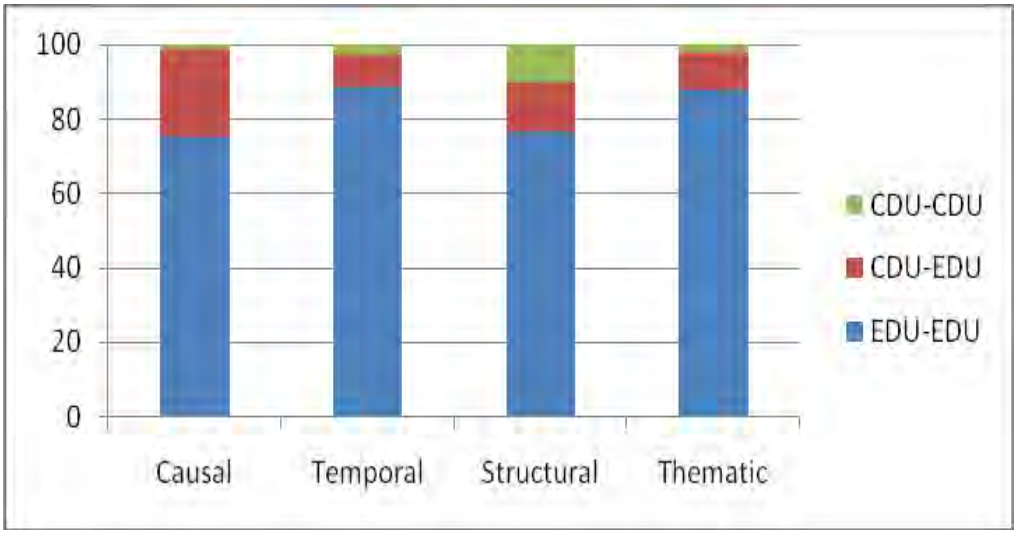

Figure 2.11. The distribution of our top-level classes according to their argument types.

We have also analyzed the distribution of discourse relations according to whether they are lexically triggered or not. For example, the relation طباق/TbAq/*Antithetic* is usually implicit whereas the relations تخيير/txyyr/*Alternation*, استنتاج/AstntAj/*Logical consequence*, تلخيص/tlxyS/*Summary* and تعيين/tEyyn/*E-elaboration* are usually explicit. We observe that among the 3,184 relations, more than 25% of relations (802) are *implicit*, i.e. signaled by any connectors. Concerning *explicit* relations, 941 are signaled by *strong discourse markers* that are non ambiguous and generally indicate the same relation (around 35% in our gold corpus). For example, the marker بل/bl/*however* for the relation إضراب/<DrAb/*Correction*, لكن/lkn/*but* for the

relation استدراك/AstdrAk/*Concession* and لذلك/l*lk/*given* for the relation غرض /grD/*Goal*. On the other hand, 1,441 explicit relations are triggered by *weak discourse markers* that are highly ambiguous and can trigger 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* 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*. Table 2.7 presents a list of some weak makers along with the relations they can signal. We use "NONE" to indicate that weak discourse markers can not indicate a discourse relation.

Weak discourse connectives	**Discourse relations signaled**
وهو/whw/*he/is it/is this/ which*	تعريف/ tEryf/*Definition*, نتيجة/ntyjp/*Result*, معية/mEyp/*Parallel*, تعليق/tElyq/*Commentary*, تلخيص/tlxyS/*Summary*, تفصيل/ tfSyl/*Description*, تعيين/tEyyn/*E-elaboration*, and NONE
الا/AlA/*except/but*	استدراك/AstdrAk/*Concession*, شرط/$rT/*Conditional*, تخصيص/txSyS/*Specification*, مقابلة/mqAblp/*Contrast*, إضراب/<DrAb/*Correction*, and NONE
وهذا/wh*A/*this/that*	نتيجة/ntyjp/*Result*, تلخيص/tlxyS/*Summary*, تعليق/tElyq/*Commentary*, استنتاج/AstntAj/*Logical consequence*, and NONE
وذلك/w*lk/*so/that/since/for*	نتيجة/ntyjp/*Result*, سبب/sbb/*Explanation*, غرض/grD/*Goal*, ربط دون ترتيب زمني/rbT dwn trtyb zmny /*Continuation*, and تأطير/t>Tyr/*Frame*
اذ/A*/*even/if/so*	ترتيب بسرعة/ trtyb bsrEp/*Quick ordering*, نتيجة/ntyjp/*Result* and شرط/$rT/*Conditional*
كما/kmA/*as/like*	تزامن/tzAmn/*Synchronization*, معية/mEyp/*Parallel*, خلفية/xlfyp/*Background-Flashback*, and NONE
ولو/wlw/*if/though*	شرط/$rT/*Conditional*, سبب/sbb/*Explanation*, خلفية/xlfyp/*Background-Flashback*, and NONE
اما/AmA/*either/else/or*	استدراك/AstdrAk/*Concession*, تخيير/txyyr/*Alternation*, and NONE
فيما/fymA/*with/while*	تزامن/tzAmn/*Synchronization*, مقابلة/mqAblp/*Contrast*, and NONE
لكي/lky/*in order to*	سبب/sbb/*Explanation* and غرض/grD/*Goal*

Table 2.7. Some weak discourse connectives and the possible relations that can signal

Conclusion

In this chapter, we presented the Arabic Discourse Treebank corpus (ADTB)[18], the first resource that explicit the interactions between the semantic content of Elementary Discourse Units and the global, pragmatic structure of the discourse. The corpus is composed of documents extracted from the syntactically annotated Arabic Treebank v3.2 part 3 where each document is represented by an oriented acyclic graph that provides a recursive and a complete discourse structure of the document. We studied the segmentation principles to segment text into clauses and EDUs as well as the rhetorical relations from a semantic point of view by focusing on their effect on meaning and not on how they are lexically triggered. We built a new hierarchy of relations relying on Arabic rhetoric literature and corpus analysis. Our new classification is organized around 4 top-level classes with a total of 24 relations. The results of the annotation campaign show that full discourse annotation is feasible in Arabic where a good inter-annotator agreement has been reached.

Our corpus contains 70 documents with a total of 4,963 EDUs and 3 184 relations which is comparable to the Annodis corpus (3,199 EDUs and 3,355 relations). 25% of the relations are implicit and 15% of them relate non adjacent EDUs. The next step is to automatically learn discourse segmentation (cf. Chapter 3) and Arabic discourse relation recognition (cf. Chapter4). As future work, we plan to extend this corpus by annotating more documents.

[18] We thank our experts in Arabic linguistics: Fathi Boujelben, Atef Ktari and Monji Châaben for their efforts and their feedback during the elaboration of the annotation manual.

Chapter 3: Automatic Discourse Segmentation

Table of contents

Introduction 93
1. Related work 94
1.1. EDU segmentation: main approaches 94
1.2. Arabic EDU segmentation 95
2. Rule-based approach 97
2.1. The data 97
2.2. Proposed approach 97
2.3. Experiments and results 100
3. Learning approach 102
3.1. The data 102
3.2. Proposed approach 103
3.3. Experiments and results 106
Conclusion 112

Introduction

Discourse segmentation aims at splitting texts into Elementary Discourse Units (EDUs) which present non overlapping units that serve to build the discourse structure of a document. Indeed, EDUs are the entities that have to be linked by coherent relations that have to be grouped together if a set of EDUs is, as a whole, an argument of a coherent relation. Thus, identifying EDU boundaries is an important step in discourse parsing, since a wrong segmentation degrades the discourse parser performances. For instance, Soricut and Marcu (2003) have pointed out that perfect segmentation reduces the number of parser errors by 29%.

Several works on automatic discourse segmentation have been undertaken using rule-based (Le Thanh *et al.*, 2004; Tofiloski *et al.*, 2009) or learning techniques (Fisher and Roark, 2007; Sporleder and Lapata, 2005). Most studies have focused on English. We note, however, some efforts for other languages such as French (Afantenos *et al.*, 2010), Thai (Charoensuk *et al.*, 2005), German (Lungen *et al.*, 2006), Spanish (Da Cunha *et al.*, 2010), and Brazilian Portuguese (Pardo *et al.*, 2004). As far as we know, there is no work developed for Modern Standard Arabic (MSA) that has investigated EDU segmentation. This chapter is an attempt to carry out discourse segmentation task using the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003).

Due to the morphological and syntactic properties of MSA, discourse segmentation poses different set of challenges. In particular, what are the principles that guide the segmentation process of Arabic texts? How can discourse segmentation deal with Arabic complex morphology where words, notably, discourse connectives (DCs), are highly ambiguous? What kind of suitable morphological analysis, that is, shallow versus deep? Are morphological features sufficient to achieve good results? What is the added value of shallow syntactic features? To answer these questions, we propose to first build a rule-based approach for Arabic discourse segmentation into clauses. Given the important number of discourse segmentation principles, we choose to implement only the clause-based segmentation principles.

In the second step, we propose a supervised learning approach to be applied on two corpora (Elementary School Textbook and ADTB) using the Stanford classifier which is based on the Maximum Entropy model (Berger *et al.*, 1996), where segments are EDUs. We use state-of-the-art features whose efficiency has been empirically determined such as punctuation, morphological, lexical, and syntactic features (Afantenos *et al.*, 2012; Fisher and Roark, 2007; Soricut and Marcu, 2003; Sporleder and Lapata, 2005). Their use in Arabic discourse segmentation is, nonetheless, novel. We investigate how each feature contributes to the learning process. In particular, we analyse the impact of shallow and extensive morphological features as well as chunks. We report our experiments on boundary detection, which presents, the ability of the system to classify each token into the right class, as well as on EDU recognition, namely, the ability of the system to identify EDU boundaries. We show that an extensive morphological

analysis is crucial to achieve the best results for both corpora. Similarly, we show that adding chunks does not boost the performance of our system.

The first section of this chapter introduces the most known researches on discourse segmentation and their theoretical frameworks. Section 2 describes the rule-based approach to automatic discourse segmentation into clauses and details its results. Finally, Section 3 introduces the supervised learning approach to automatic discourse segmentation into EDUs ended by the obtained results.

1. Related work

Two parts have been explored in this section: the main approaches of discourse segmentation into EDUs for different languages and Arabic discourse segmentation into EDUs.

1.1. EDU segmentation: main approaches

Several works have been undertaken on automatic discourse segmentation for different languages. They can be basically classified into two broad categories: rule-based approach and machine learning-based approach. In the first approach, handcrafted rules aim at identifying potential cutting points relying on a combination of surface cues (punctuation and lexical markers) and syntactic patterns that encode syntactic categories and parts-of-speech. In the English language, we can mention Le Thanh et al. (2004) who reported an F-measure of 0.869 when evaluating their segmenter against the boundaries in the RST Discourse Treebank (RST-DT) (Carlson *et al*., 2003). Tofiloski et al. (2009) built the *SLSeg* system on top of an automatic syntactic parser and showed that their approach outperforms those of other approaches by achieving an F-score of 80–85% in segment boundary. Symbolic approaches have also been used in other languages like German (Lungen *et al*., 2006), Spanish (Da Cunha *et al*., 2010), Brazilian Portuguese (Pardo *et al*., 2004), and Japanese (Sumita *et al*., 1992). Most of these systems are based on the RST framework.

In addition, learning approaches, usually exploit lexical and syntactic features to classify each token in a sentence as being an EDU boundary or not. Within the RST framework, Soricut and Marcu (2003) proposed a sentence-level discourse parser. They made an extensive use of the syntactic tree in which each token is modeled by taking into account syntactic dominance features (the token itself, its parent, and its siblings). Sporleder and Lapata (2005) used the RST-DT corpus and labeled each token with four different tags: B-NUC and B-SAT for nucleus and satellite initial tokens, and I-NUC and I-SAT for non initial tokens. For the segmentation task, they performed a binary classification, where each span (and not a token) can have a Begin or an Inside label. Span boundaries are given by the gold standard. Using this method, they showed that employing lexical and low-level syntactic information (such as parts-of-speech and syntactic chunks) is sufficient to achieve good performance. Their approach is comparable to Soricut and

Marcu (2003). Fisher and Roark (2007) proposed various improvements using finite-state analysis. Subba and Di Eugenio (2007) used a neural network (multilayer perceptron) while Hernault et al. (2010a) used conditional random fields to train a discourse segmenter on the RST-DT corpus. For other languages, we cite Charoensuk et al. (2005) who proposed a hybrid approach for Thai using a decision-tree learning system and some heuristic rules.

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

Current state-of-the-art approaches in discourse segmentation make an extensive use of syntactic information going from chunking to deep syntactic parsing, including dependencies. However, some languages are lack reliable deep syntactic parsers. Sporleder and Lapata (2005) have shown that good results can be reached only by chunking and their approach can be portable to languages for which deep parsers are not available. We plan here to go further by analyzing what extent EDU segmentation is feasible without using shallow syntactic information. We adopt a multiclass classification approach as done by Afantenos et al. (2012). We use a combination of state-of-the-art features to predict nesting. To the best of our knowledge, the use of these features for Arabic discourse segmentation is novel.

1.2. Arabic EDU segmentation

Little work has been done on the discourse level. Among them, let us cite Belguith et al. (2005) who proposed a rule-based approach to segment non voweled Arabic texts into sentences. The approach consists of a contextual analysis of the punctuation marks, the coordination conjunctions, and a list of particles considered as boundaries between sentences. The authors defined 183 rules to segment texts into paragraphs and sentences. These rules were implemented in the STAr system, a tokenizer based on the proposed approach. Touir et al. (2008) proposed a rule-based approach to segment Arabic texts using connectors without relying on punctuation marks. Segmentation principles did not follow any discourse theory. They performed an empirical study of sentence and clause connectors and introduced the notion of active connectors, which indicate the beginning or the end of a segment and the notion of passive connectors that do not imply any cutting point. Passive connectors are useful only when they co-occur with active connectors since this might imply the beginning or the end of a segment. Finally, Khalifa et al. (2011) proposed a learning approach to segment Arabic texts by only exploiting the six rhetorical

functions of the DC و/w/*and* (cf. Chapter 1). A set of 22 syntactic and semantic features was then used in order to automatically classify each instance of the DC و/w/*and* into these two classes. The authors reported that their results outperform those of Touir et al. (2008) when considering the DC و/w/*and*.

The closest research to ours is the one done by Al-Saif and Markert (2010, 2011) that, respectively, described how to recognize DCs and how to automatically identify explicitly marked discourse relations within the Penn Discourse Treebank (PDTB) framework (Prasad *et al.*, 2008). Discourse segmentation in PDTB tends to larger units than to EDUs since arguments can be as small as a nominalization or as large as several sentences. Segmentation in PDTB requires three main steps: (1) identifying DCs, (2) identifying the locations of Arg1 and Arg2, and (3) labeling their extent. Arg1 can be located within the same sentence as the DC or in some previous sentences of the connective. When Arg1 and Arg2 are in the same sentence, we can have several cases: Arg1 coming before Arg2 as in Example 1, Arg1 coming after Arg2 as in Example 2, and Arg2 embedded within Arg1 as in Example 3 (cf. Chapter 2).

(1)[تعرضت لأضرار]$_{arg1}$ [نتيجة الاصطدام.]$_{arg2}$

[tErDt lADrAr]$_{arg1}$ [ntyjp AlASTdAm]$_{arg2}$

[Suffered damages]$_{arg1}$ [as a result of the collision.]$_{arg2}$

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

[fy Hyn AnhA HAml,]$_{arg2}$ [<ynAs lA t>x AlTA}rp.]$_{arg1}$*

[While she is pregnant,]$_{arg2}$ [Ines did not take the plane.]$_{arg1}$

(3)[ناقش الأستاذ الامتحان،[الذي أجراه التلاميذ الأسبوع الماضي،]$_{arg2}$ كما ناقش الدرس الحالي .]$_{arg1}$

[nAq$ Al>stA AlAmtHAn, [Al*y >jrAh AltlAmy* Al>sbwE AlmADy,]$_{arg2}$ kmA nAq$ Aldrs AlHAly.]$_{arg1}$*

[The teacher explained the exam, [that was passed by the students last week,]$_{arg2}$ and the current lesson.]$_{arg1}$

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

2. Rule-based approach

2.1. The data

We used the Elementary School Textbooks (EST) (1,095 paragraphs, 29,473 words). The distribution of the number of texts and clauses per genre is shown in Table 3.1. We get a total of 4,186 segments. 60% of these segments were used for building our segmentation patterns. The rest of the corpus was left for test.

	Training corpus		**Test corpus**	
	Texts	**Clauses**	**Texts**	**Clauses**
4th EST	30	604	17	340
5th EST	28	550	15	260
6th EST	30	400	20	301
7th EST	31	541	22	315
8th EST	32	630	25	345
Total	**151**	**2,625**	**99**	**1,561**

Table 3.1. Training and test distrubition.

2.2. Proposed approach

During the corpus analysis step, three different segmentation principles were identified: (p1) using punctuation marks only, (p2) using DCs only, and (p3) using both the principles (p1) and (p2). To build a rule-based approach for automatic text segmentation into clauses, we implement each principle as a rule. Then, we designed three discourse segmenters. The first two are based respectively on the principles (p1) and (p2), while the last one is based on the principle (p3). To build the third segmenter, we propose a three steps segmentation algorithm. First, texts are segmented according to (p1). This leads to a first segmentation level, which is refined according to the principle stated in (p2). The final segmentation is obtained by applying the principle (p3). Each step has its own patterns coupled with linguistic resources (Mesfar, 2008) as the dictionaries of verbs, nouns, adjectives as well as morphological and surface syntactic analysis in order to resolve the agglutination problem. These dictionaries are used to recognize the type of indicators as well as their right and left contexts. Figure 3.1 describes the general architecture of our system. The output of our process is an XML file that contains the segmented text (cf. Figure 3.4).

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

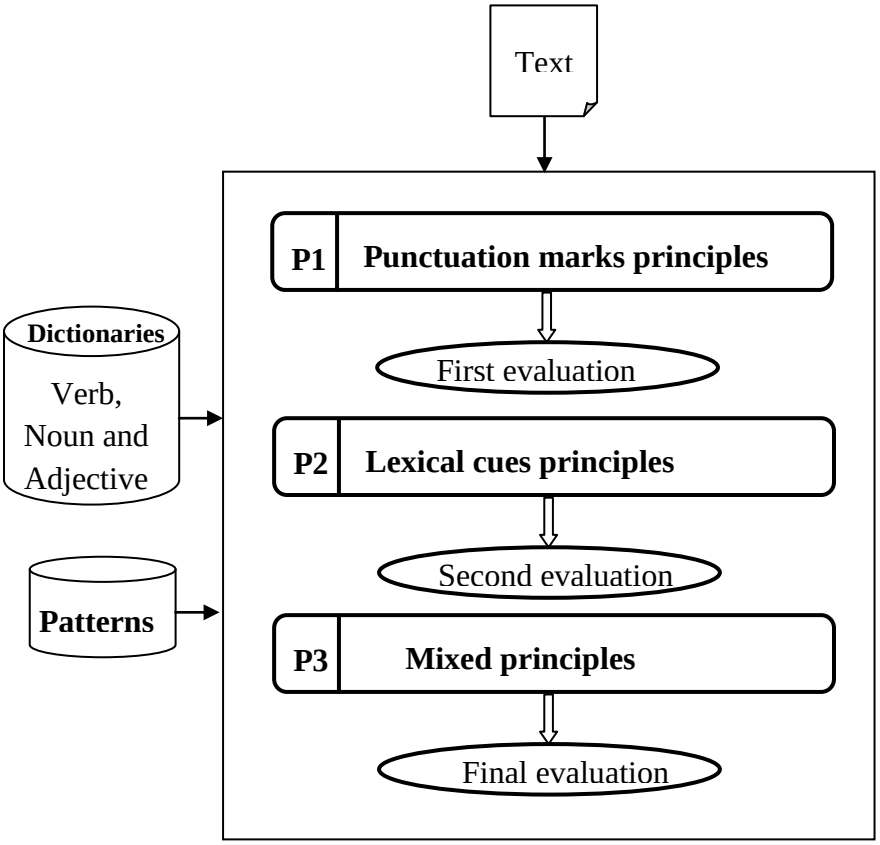

Figure 3.1. A rule-based approach for discourse segmentation.

Figure 3.2 presents an example of a NooJ local grammar for the segmentation using dots: if there is an abbreviation in the beginning or in the middle of a sentence, the dot does not represent the end of a segment. To more explain our segmentation process, we give another local grammar for segmentation using punctuation as well as DCs (cf. Figure 3.3). The output is an XML file that contains the segmented text (cf. Figure 3.4).

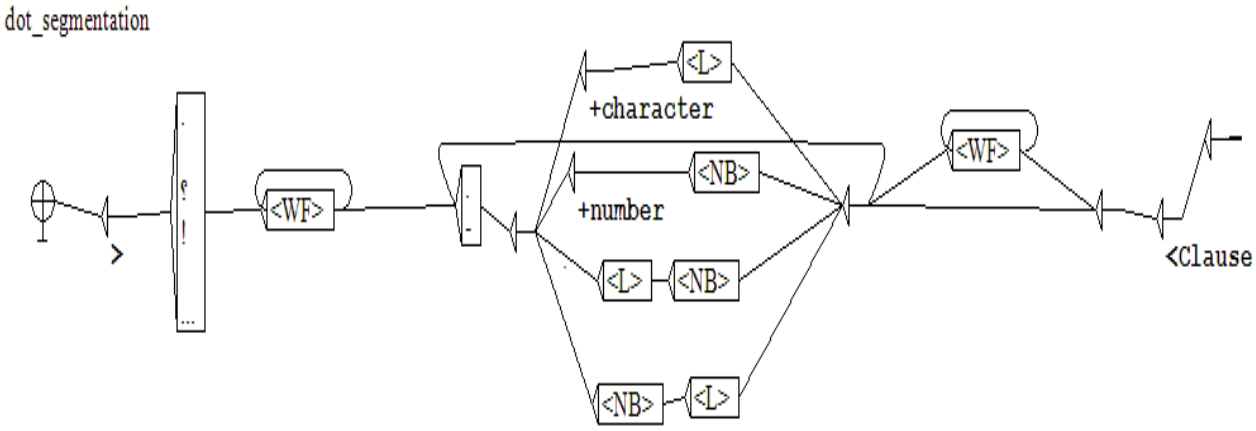

Figure 3.2. NooJ local sub-grammar for the dot marker.

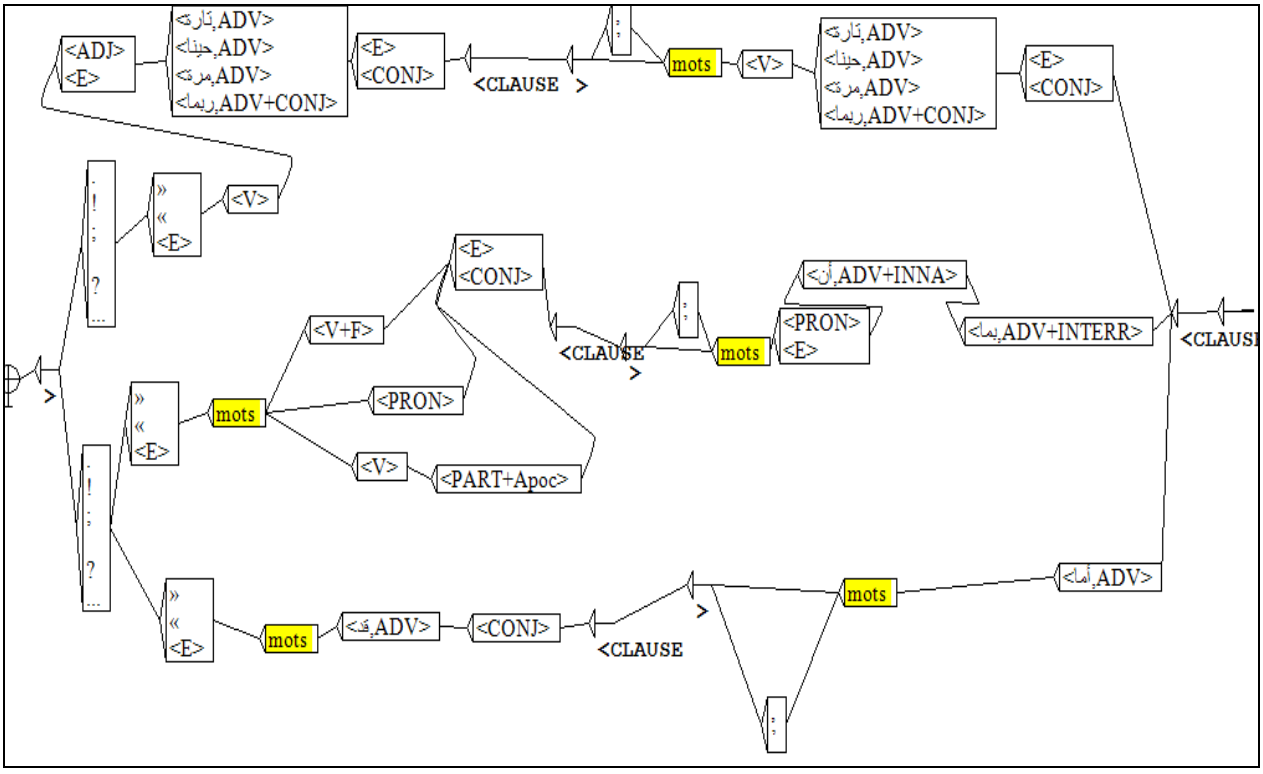

Figure 3.4. NooJ local sub-grammar for DCs and punctuation marks patterns.


```
<TEXT>
  <PH>
    <CLAUSE>كان ديّان قطبيّان نائمين على مقربة من برج المراقبة</CLAUSE>
    <CLAUSE>عندما وصلت مع عالم الطبيعة «جون كروغر».</CLAUSE>
  </PH>
  <PH>
    <CLAUSE>وما إن سمعا صوت الطوّافة</CLAUSE>
    <CLAUSE>حتّى انطلقا يعدوان مذعورين.</CLAUSE>
  </PH>
  <PH>
    <CLAUSE>
      تتجمّع مئات من الدببة القطبيّة كلّ خريف قرب رأس تشرشل على الشاطئ الغربيّ لخليج هدسون في كندا.
    </CLAUSE>
  </PH>
  <PH>
    <CLAUSE>
      وهي تقتات في الشّتاء والرّبيع وأوائل الصيف بحيوانات الفقمة
    </CLAUSE>
    <CLAUSE>التي تصطادها فوق الجليد.</CLAUSE>
  </PH>
  <PH>
    <CLAUSE>علما أنّها أضخم الحيوانات اللاّحمة</CLAUSE>
    <CLAUSE>التي تدبّ على وجه الأرض.</CLAUSE>
    <CLAUSE>ومع تقاصر ساعات النّهار تنتقل الدببة شمالا ببطء</CLAUSE>
    <CLAUSE>
      وتتجمّع على مقربة من رأس تشرشل ورؤوس أخرى في انتظار تراكم الجليد من جديد.
    </CLAUSE>
  </PH>
  <PH>
    <CLAUSE>أمضيت سبعة فصول أرصد الدببة القطبيّة في رأس تشرشل:</CLAUSE>
    <CLAUSE>
      راقبت خمسة منها من عربات ضخمة مصفّحة واثنين من حجرة صغيرة في قمّة برج ترتفع حوالي 13 مترا.
    </CLAUSE>
    <CLAUSE>كنت أنظر إليها في البداية</CLAUSE>
    <CLAUSE>على أنّها حيوانات مهيبة رائعة.</CLAUSE>
  </PH>
```


Figure 3.5. The XML output of our segmentation process.

Our approach is novel in three ways: first, it relies on an extensive analysis of a large set of DCs as well as punctuation marks. Thus, it goes beyond the method proposed by (Touir *et al.*, 2008) since we handle both a greater number of DCs and punctuation marks. Our approach goes also beyond the work of (Khalifa *et al.*, 2011) since their method relies only on one DC.

A second aspect of our research is that our analysis was carried out on two different corpus genres: news articles and Elementary School Textbooks. Corpus analysis allows us to group connectors into different categories depending on whether they are (or not) a good indicator to begin or end a segment.

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

Finally, we have proposed a clause-based segmentation algorithm that requires three steps: first by using only punctuation marks, then by relying only on DCs and finally by using both typology and DCs. The results obtained by our rule-based approach will be compared to manual segmentations elaborated by experts (cf. Chapter 2).

2.3. Experiments and results

Our three discourse segmenters, that follow respectively the principles (p1), (p2) and (p3), have been evaluated on the test data of EST. Table 3.2 summarizes the obtained results.

	Segmentation level	**Precision**	**Recall**	**F-measure**
EST	P1	46%	44%	45%
	P2	68%	64%	66%
	P3	**86%**	**85%**	**85,5%**

Table 3.2. Evaluation results of the rule-based approach.

As expected, the first level segmentation (i.e., based on punctuation marks) performs badly. For instance, our rules, for dots, do not perform well in case of the presence of abbreviations at the end of the segment, since this does not imply a cutting point (cf. Example 4).

(4) [حصل على جائزة البنك الإسلامي للتنمية في الاقتصاد الإسلامي لعام 1411 هـ.]

[HSl ElY jA}zp Albnk Al<slAmy lltnmyp fy AlAqtSAd Al<slAmy lEAm 1411 h.]

[He obtained the Islamic bank award for the development in the Islamic economy for the year 1411 H.]

We also observed that our rules for commas often fail mainly because our system do not correctly handle lexical ambiguities, as in Example 5, where the adverb بعد / (after) was identified as a verb بعد/ (to move away).

(5) [أكل الولد تفاحة،][بعد غسلها]

[>kl Alwld tfAHp,][bEd gslhA]

[The boy ate an apple,] [after washing it]

The second level segmentation achieves better results compared to the first level, which shows that DCs are good indicators to segment sentences into clauses. As for segmentation principle (p1), main errors come from lexical ambiguities, as in Example 6, where the word جرحه/jrHh is recognized as a verb (to injure). We have a segmentation error since this word is a noun in the context of Example 6. The cutting point here should be the word لمعالجة/lmEAljp, because the DC ل/l/*to* is a good indicator of the relation Goal.

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

wSf AlTbyb llmryD mjmwEp mn Al>dwyp lmEAljp >lmh wjrHh

The doctor recommended to the patient a set of drugs to treat his pain and injury.

Errors come also from the syntactic parser, as in Example 7, where, the named entity فضل/fDl is parsed as a conjunction ف/f/*then* and a verb ضل/Dl/*lost* which implies a beginning of a segment.

(7)استقبلت عائلة مصطفى فضل البارحة.

Astqblt EA}lp mSTfY fDl AlbArHp.

I received Mustapha Fadhl's family yesterday.

Finally, segmentation using both punctuations and DCs gives the best results. This demonstrates that using morphological and syntactic information is helpful to disambiguate some DCs as well as weak punctuation marks. Of course, mixed principles present some limits, because in some cases, both punctuation marks and DCs are omitted, as in Example 8, where we have two segments related by the rhetorical relation goal.

(8) *فأخذنا نقرأ بعض الصّفحات معا نناقش محتواها*

*f>x*nA nqr> bED AlS~fHAt mEA nnAq$ mHtwAhA*

We have read together some pages and we have discussed about their content

The main challenge in Arabic discourse segmentation remains the disambiguation of DCs. Given that Arabic is an agglutinative language, we have to go beyond standard morpho-syntactic analysis, in order to deal with lexical ambiguities. Thus, we need semantics. Interesting efforts in this direction include the work of (Khalifa *et al.*, 2011) on the DC و /w/*and* that can be used efficiently in our framework to improve the results of our system when using the principle (p3).

In this section, we have proposed a rule-based approach for Arabic text segmentation into clauses. Our main goal was to automatic prove the validity of our segmentation manual on a new corpus and that our segmentation principles are independent from the empirical data used in the manual building step. We evaluate our three segmentation principles (the first based on the exclusive use of punctuation marks, the second relies on DCs and the last one is based on a combination of the first two principles) using EST. Our results show that the third principle corresponds to the best segmentation algorithm.

In the next section, we proposed a supervised learning approach to automatic discourse segmentation into EDUs according to the SDRT framework (Asher and Lascarides, 2003).

3. Learning approach

3.1. The data

Our data comes from two different corpus genres: Elementary School Textbooks (EST) and ADTB, newspaper documents extracted from the syntactically annotated Arabic Treebank (ATB v3.2 part3) (cf. Chapter 2).

We randomly selected a set of 34 documents from EST. These documents contain a total of 622 sentences, which corresponds to 8,704 tokens (words and punctuations). Contrary to ADTB documents, it is important to note that EST documents are not associated to any kind of manual annotation.

Again, we have randomly selected 56 documents from ADTB for a total of 1,427 sentences and 31,682 tokens (words and punctuations). Table 3.3 gives statistics about the data in the gold standard. The column WORD+PUNC indicates the number of tokens.

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

Table 3.3. The gold standard corpus characteristics.

3.2. Proposed approach

Current state of the art in discourse segmentation makes an extensive use of syntactic information going from chunking to deep syntactic parsing, including dependencies. However, some languages present a lack of reliable deep syntactic parsers. (Sporleder and Lapata, 2005) have already shown that good results can be reached only by chunking and that their approach can be portable to languages for which deep parsers are not available. We wanted here to go further by analyzing at what extent EDU segmentation is feasible without using shallow syntactic information. We performed a supervised learning on the gold standard data basing on the Maximum Entropy model (Berger *et al*., 1996) which is Stanford classifier. Each token can belong to one of the following three classes: *Begin*, if the token begins an EDU, *End*, if it ends an EDU, or *Inside*, if a token is none of the above[19].

To identify EDU boundaries, we used four groups of features: *punctuation*, *lexical*, *morphological* and *syntactic* features. A feature vector is associated to each token. The features were designed after analyzing the documents used for training as well as the documents used to compute inter-annotator agreements (which correspond to 6 ATB documents (181 sentences) and 9 EST documents (138 sentences)). Our set of features is given below.

3.2.1. Punctuation features

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

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

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

3.2.2. Lexical features

We consider here both DCs such as حيث/Hyv/*where*, بينما/bynmA/*while*, عندئذ /End}*/*at that time*, and a set of specific words, called indicators that are important for the segmentation process. Indicators can be reporting verbs and propositional attitude verbs (e.g. قال/qAl/*say*, أعلن/>Eln/*announce*, إعتقد/<Etqd/*believe*), 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*, المفروض من/mn AlmfrwD/*normally*, فقط/fqT/*only*), conjunctions (e.g. حالما/HAlmA/*the moment that* and طالما/TAlmA/*so/often*), and particles (e.g. لم/lm/*not* and لن/ln/*never*). Like punctuation marks, we have two DCs categories (henceforth DCC): *strong* and *weak*. Strong connectors are usually followed by a verb which indicates the beginning of a segment. Some of these DCs are: كي/ky/*to*, ل/l/*for*, لكن/lkn/*but*, لكنَّ/lkn~a/*but*, غير أن/gyr >n/*nevertheless*, بيد أن/byd >n/*however*, أجل من أن/mn >jl>n/*in order to*. On the other hand, ambiguous DCs do not always mark the beginning of a segment, as the DC و/w/*and* and the particles ثم/vm/*then*, ف/f/*so/then*, etc. For example, the particle و/w/*and* can express a new clause, a conjunction between NPs, or it can be a part of a word, as in ورشة/wr$p/*atelier*.

We have explored four lexical features: (1) TOKEN_LEX, the current token DCC; (2) BEFORE_LEX, the DCC of the token that precedes the current token; (3) AFTER_LEX, the DCC of the token that follows the current token; and (4) TOKEN_BeginLex, a Boolean feature to indicate whether the current token begins with an indicator or with a DCs. This last feature treats the agglutination. DCC can take five values: 0, if the token is not a DC; 1, if the token is a strong DC; 2, if the token is a weak DC; 3, if the token is a strong indicator; and 4, if the token is a weak indicator.

To handle both punctuation and lexical features, we built a lexicon of segmentation indices where each entry is characterized by its type (a punctuation mark, a discourse cue or an indicator), its nature (strong or weak) and a list of its possible parts of speech (POS). We have also indicated if the lexical entry is composed of other words, such as خلاصة القول/Alqwl xlASp/*in summa*ry and باختصار/bAxtSAr/*briefly*. If so, we detail each element of the composition. Finally, we matched each entry with its English translation and an example of its usage in context. Our lexicon contains 174 entries: 11 punctuation marks (4 strong: the exclamation mark, the question mark, the colon and the semi-colon, and 7 weak: the full stop, the comma, quotes, parenthesis, brackets, braces and underscores) and 163 lexical cues (83 DCs and 80 indicators) among which 76.4% are strong and 23.6% are weak.

3.2.3. Morphological features

Our main goal is to identify what kind of morphological analysis is suitable for Arabic discourse segmentation, that is, shallow versus extensive. To this end, we propose to use two contextless parsers that provide different morphological information: Alkhalil (Boudlal *et al.*, 2011), a shallow parser, and the Standard Arabic Morphological Analyzer SAMA version 3.1

(Maamouri *et al.*, 2010a), an extensive analyzer. We have thus designed two sets of morphological features, one for each parser output.

Alkhalil gives each token a non ordered list of all its possible forms (by default, the first form of this list is chosen) (Boudlal *et al.*, 2011). More precisely, it generates the stem, its grammatical category, and its possible roots, where each root is associated to its corresponding patterns, proclitics, and enclitics. Alkhalil does not take into account the context and the punctuation marks. In addition, it does not provide affixes information, and its database does not contain information about the closed nouns except their fully diacritized form and their Arabic class name, along with the allowed proclitics and enclitics. For each token, we investigated six features provided by Alkhalil: (1) STEM, the token stem; (2) POS, the token parts-of-speech; (3) CATEGORY, the token grammatical category; (4) HAS_PREFIX and (5) HAS_SUFFIX that, respectively, indicate if the token has a prefix or a suffix; and (6) PATTERN, the token pattern. All the features are encoded into strings (in Arabic script).

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

3.2.4. Syntactic features

To evaluate the added value of syntactic features to discourse segmentation of Arabic texts, we propose to take into account the chunks. To determine these chunks, we rely on manual annotations instead of using a shallow syntactic parser such as AMIRA (Diab, 2009). Indeed, our aim is to test the upper bound for shallow syntax features. If we do not find useful chunks, it is not necessary to use a parser to predict them. Syntactic features concern only the ATB corpus (we recall that EST documents do not contain any manual annotations (cf. Chapter 2)).

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

3.3. Experiments and results

In order to measure the impact of the morphological and syntactic features on the performance of our segmenter, we designed three classifiers: (C1) that uses punctuation marks, lexical, and Alkhalil features; (C2) that relies on punctuation, lexical and SAMA features; and (C3) that uses punctuation, lexical, SAMA features; and syntactic features. (C1) and (C2) were run on EST and ATB while (C3) concerns only ATB. Punctuation features are the same for all classifiers. Lexical features are obtained by checking whether the current token lemma (as given by SAMA) or the current token stem (as given by Alkhalil) is an entry in our lexicon. Our first experiment showed that best results are achieved when using SAMA lemmatization. We have thus decided to use the token lemma as given by SAMA.

For each corpus, we have performed a ten-fold cross-validation where 10% of the corpus was left for test. For all experiments, we have used both n-gram character and n-gram word as features. Best results were achieved with n=4. Because we have few EDU boundaries, our data set is skewed (see Table 3.3, Section 3.1 for an overview of our data characteristics). Note that we did not observe any problem related to the class imbalance in the training set with the parameters we used when building the classifier.

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

— token boundary detection, which is the ability of the system to classify each token into the correct class (*Begin*, *End*, and *Inside*);

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

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

3.3.1. Token boundary detection

3.3.1.1. Analyzing the impact of punctuation, lexical and morphological features.

Unlike Tofiloski et al. (2009) and Soricut and Marcu (2003) who measure only the score of their segmenter on boundaries inside sentences (to avoid artificially boosting the performance), the evaluation of our system takes into account sentence boundaries. Indeed, end-of-sentence or end-of-paragraph boundaries are not given automatically but are predicted by our segmenter. Table 3.4 gives (C1) and (C2) overall performances in terms of precision, recall, F-score, and accuracy, averaged over the three classes *Begin*, *End*, and *Inside*. Best performances are marked

in boldface. We first start with punctuation features to demonstrate which several features are progressively added; this is marked by the "+" sign in the table. We have also compared the performance of each classifier against two baselines: (B1) that only uses the current token punctuation category (TOKEN_PUNC); and (B2) that uses both the current token punctuation and lexical category (i.e., TOKEN_PUNC and TOKEN_LEX).

Our first baseline (B1), that checks if the current token is a punctuation mark (from the strong or the weak type) or not, performs badly for both corpora. Taking into account both right and left context (by adding BEFORE_PUNC and AFTER_PUNC features) improves the F-score by, respectively, 0.074 for EST and 0.037 for ATB. However, punctuation features alone are not sufficient to achieve good results for both corpora, for three main reasons: the absence of regular punctuation marks, especially for ATB, the high frequency of weak punctuation marks (cf. Example 9), and the presence of named entities.

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

[kAnt rAfEp ydhA AlT~wylp, fAtHp fmhA AlwAsE,]

[She was raising her long arms, opening her wide mouth,]

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

When adding lexical features, EST results remained stable while at the same time ATB results (in terms of accuracy) improved significantly over (B1) + BEFORE_PUNC + AFTER_PUNC by more than 0.300. We assume that the absence of improvement for EST can be explained by the fact that EST is characterized by regular punctuation marks, which seems to be adequate to reach an accuracy of 0.686. The good results obtained for ADTB show that our lexicon is a useful resource for discourse analysis. In addition, we observe that adding contextual lexical features, mainly lexical type (strong or weak) of the left (BEFORE_LEX) and the right token (AFTER_LEX) improves ADTB results. Indeed, unlike rule-based approach where the adverb بَعْد/baEud/*after* was identified as a verb بَعُد/baEod/*to move away* (cf. Example 5), these features were able to disambiguate cases as in Example 10. However, lexical features cannot deal with other types of ambiguities, like named entities (cf. error analysis at the end of this Section).

(10)[أكل الولد تفاحة، بعد غسلها]

[<kl Alwld tfAHp, bEd gslhA]

[The boy ate an apple, after washing it]

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

Table 3.4. Results of the baselines, (B1) and (B2); and the classifiers, (C1) and (C2) in terms of Precision (P), Recall (R), F-score (F), and Accuracy (Acc).

We note that using the McNema's test, the difference between (C1) and (C2) is significant at $p<0.05$ for both EST and ADTB.

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

Overall, both corpora achieved good F-scores that are comparable to human results (cf. Chapter 2). An interesting observation comes from punctuation features, they perform badly when they are used alone, removing them from the feature vector has a negative impact on the results for both classifiers. For instance, we get an F-score of 0.840 for EST and 0.837 for ADTB when running the classifier with SAMA features. Another interesting point is that using morphological features alone are not sufficient. Indeed, we get an F-score of 0.713 for ADTB and 0.772 for EST when running (C1) and (C2) without punctuation and lexical features. Moreover,

when comparing (C1) and (C2), only the Begin class is biased (the F-score decreases from 0.899 to 0.540) while the results of the End and the Inside classes remain stable. Finally, the overall evaluation on EST documents gets similar results compared to those obtained for ADTB documents. As expected, we can conclude that discourse segmentation does not rely only on punctuation marks and that text length has no impact on the segmentation. Our results demonstrate that our first intuition is wrong when stipulating that segmenting EST documents will be more simple and will achieve better results compared to other corpora. This shows that combining punctuation, lexical, and extensive morphological features is necessary to achieve good segmentation results.

We finally give in Table 3.5 the results of our best configuration (C2) per class a. For both corpora, the End class gets lower results compared to the Inside and the Begin class (in terms of F-score).

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

Table 3.5. Results of the (C2) classifier with SAMA features on each class.

The error analysis of the outputs of classifier (C2) on the ATB documents shows that our classifier successfully distinguishes between the Begin and the End classes. In addition, the prediction of embedded EDUs is good in terms of precision (about 0.920, 0.900, and 0.700 for, respectively, Inside, Begin, and End classes). As we can see in the confusion matrix (see Table 3.6), main confusions (in bold font) are between End class and Inside class.

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

Table 3.6. Confusion matrix of the (C2) classifier on ADTB.

The analysis of these confusions shows that most errors come from the presence of named entities and from some weak punctuation marks. Examples 11.1 and 11.2 show, respectively, a gold-standard annotation and the output of our classifier. Our system predicts that the word و/w/*and* is a cutting point because the word أكرم/>krm/*Akram* has been analyzed as the verb أكرم/>krm/*to honor*, which is, of course, wrong since this word is a named entity.

(11.1) [حصل خالد وأكرم على جائزة.]

[HSl xAld w>krm ElY jA}zp.]

[Khalid and Akram obtained an award.]

(11.2) [حصل خالد][وأكرم على جائزة.]

[HSl xAld][w>krm ElY jA}zp.]

[Khalid][and Akram obtained an award.]

Similarly, Examples 12.1 and 12.2 illustrate that our classifier fails to deal with weak punctuation marks. In Example 12.2 our classifier predicts an EDU boundary after the comma.

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

[ln >Ewd l$rH Aldrs, mrp >xrY.]

[I won't explain this lesson, again.]

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

[ln >Ewd l$rH Aldrs,][mrp >xrY.]

[I won't explain this lesson,][again.]

3.2.1.2. Analyzing the impact of syntactic features.

We have assessed the reliability of syntactic features on discourse segmentation of ADTB documents (refer to Table 3.7) by adding chunk information to the feature vector that achieved best performance in (C2). We observe that adding chunks does not really boost the results. The only improvement (in bold font in Table 3.7) concerns the recall of the *Inside* class (+ 0.003) and the precision of the *End* class (+ 0.011). The overall F-score of the (C3) classifier is 0.847, which corresponds to a marginal improvement of 0.010 compared to (C2). Similar observations go for the accuracy measure. We can thus conclude that shallow syntactic features are not useful for Arabic discourse segmentation.

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

Table 3.7. Results of the (C2) classifier with SAMA features and the (C3) classifier with syntactic features.

3.3.2. EDU recognition

An EDU is correctly recognized if, for each begin bracket, there is a corresponding end bracket. Otherwise, we have to perform a post-processing to ensure correct bracketing. Since the *End* class is the one that performs badly (cf. Table 3.7), we have decided to correct only end bracketing. Post-processing consists in adding an end bracket for each opening bracket that has no corresponding end. Table 3.8 presents our results on both corpora in terms of Accuracy (Acc), before and after post-processing. For this experiment, we have run the classifier (C2) with all the

features described in Table 3.4 except for the SAMA feature GLOSS (this feature corresponds to the penultimate line in Table 3.4).

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

Table 3.8. Accuracy (Acc) of EDUs recognition before and after post-processing.

As expected, we observe that post-processing boosts the results for both ADTB and EST with more than 0.390 for EST and 0.130 for ADTB. The results are more impressive for EST (characterized by regular punctuation marks) because using punctuation features biased the EDUs' recognition results. For the embedded EDUs (present in around 11% in the EST corpus and 8% in ADTB corpus), we have also observed the same tendencies. The obtained results are, however, lower compared to the ones obtained for non embedded EDUs. This may be explained by the low frequency of embedded EDUs in each test data (around 8 for the EST test and 37 for the ADTB test). Finally, we have observed that the performance of our segmenter is sensitive to the length of EDUs in terms of the number of tokens. Indeed, when this length is less than or equal to 3, we get an accuracy of 1.

3.3.3. The learning curve

In order to analyze how the learning procedure can be influenced by the number of annotated documents, we have computed a learning curve by dividing our corpus into 10 different sets. For each set, we performed a tenfold cross validation, using the features set of the classifier (C2). The learning curve is shown in Figure 3.6. As we can see, the curve grows regularly between 0 and 5,000 tokens (that is, 10 documents, i.e., around 255 sentences) while it seems to plateau between 5,000 and 25,000 tokens (that is, 50 documents). We can thus conclude that the addition of more than 10 ADTB documents will slightly increase the performance of the segmenter.

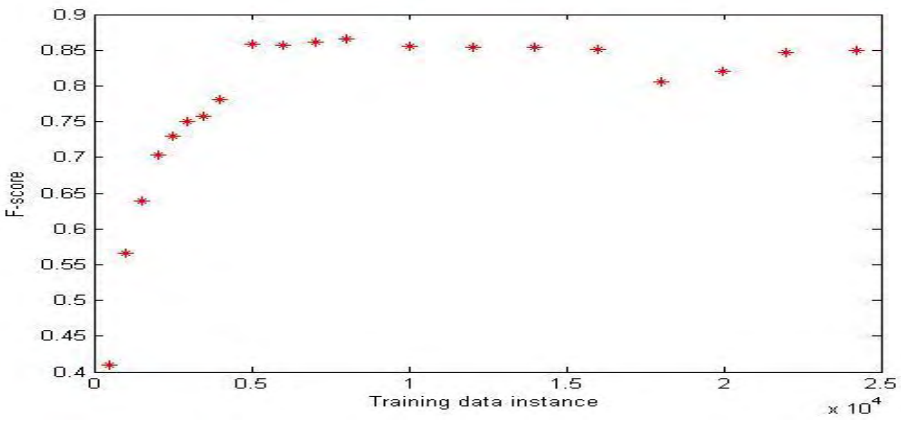

Figure 3.6. The learning curve of (C2) for ADTB.

Conclusion

In this chapter, we have presented the first work that fully addresses the Arabic discourse segmentation. We proposed a rule-based approach to segment Arabic texts into clauses and the first multi-class supervised learning approach that predicts EDUs boundaries in Arabic texts.

The rule-based approach uses EST to validate our segmentation manual and to show that these segmentation principles are independent from the empirical data used in the manual building step. In other words, we validate our discourse segmentation principles before building ADTB.

After building the Arabic Discourse Treebank corpus (ADTB), we performed a multi-class supervised learning approach that predicts EDUs boundaries and not only discourse connectives as in (Al-Saif and Markert, 2011). Our approach uses a rich lexicon (with more than 174 connectives) and relies on a combination of punctuation, morphological and lexical features. Our results showed that EST segmentation is very sensitive to punctuation features contrary to ADTB where punctuations are not widely used. In addition, contextual lexical features have a positive effect on the results especially for ADTB which shows that ADTB documents tend to use more complex words than EST documents. For both corpora, we have shown that extensive morphological features are more suitable than shallow morphological analysis since best scores were obtained when adding information of the root, the prefix and the suffix. Finally, we have shown that Arabic discourse segmentation is feasible on both corpus genres without any use of shallow syntactic information (chunks).

Another main contribution in this chapter is the recognition of EDU frontiers even in case of the absence of discourse markers (that is, in case of implicit relations), which represent 25% of cases in our data. Note that Al-Saif and Markert (2011) have treated only the cases of explicit markers.

For the moment, we have run our experiments by considering Alkhalil features and SAMA features separately. It would be interesting in the future to run our classifiers by combining features form both sets (cf. Chapter 4).

Discourse segmentation is the first step towards discourse analysis. The second step presented in the next chapter will be the automatic recognition of discourse relations in ADTB. We will propose the first work that fully addresses learning implicit and explicit Arabic discourse relations by proposing a multi-class supervised learning approach that predicts discourse relations between EDUs in Arabic texts. Our approach uses the same lexicon (174 connectives) but is enriched by discourse relation information and relies on a combination of lexical, morphological, syntactic and lexico-semantic features. We will compare the proposed approach to three baselines that are based on the most frequent relations, discourse connectives and the features used by (Al-Saif and Markert, 2011).

Chapter 4: Automatic Discourse Relation Recognition

Table of contents

Introduction 114
1. Related work 115
2. The features 117
2.1. Al-Saif et al.'s features 119
2.2. New features 120
3. Experiments and results 124
3.1. Overall results 125
3.2. Fine-grained classification 127
3.3. Mid-level classification 132
3.4. Coarse-grained classification 132
3.5. The learning curves 133
Conclusion 134

Introduction

Automatic identification of coherent relations is a crucial step in discourse parsing. This task automatically labels the attachment between the two discourse units with discourse, rhetorical or coherence relations such as Elaboration, Explanation, Cause, Concession, Consequence, Condition, etc (see Chapter 1). It 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) and (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 Arabic Discourse Treebank corpus (ADTB) 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 4,963 EDUs, linked by 3,184 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 used 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 first Section of this chapter gives an overview of the related work and our theoretical framework. Section 2 details the used features. Finally, Section 3 presents our experiments and obtained results.

1. Related work

We present in this section main known studies on discourse relation recognition, by grouping them according to their corresponding theoretical frameworks. We end this section by presenting our theoretical framework and highlighting the main contributions of this work.

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 (2007) further extended this unsupervised model using syntactic filtering and topic segmentation. Several authors have also proposed supervised approaches based on 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[20]. 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 text units. 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 the best of our knowledge, there are two SDRT-like parsers. The first one has been developed for appointment scheduling dialogues (Baldridge and Lascarides, 2005) and the second 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,

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

syntactic and linguistically informed features. Lin et al. (2009) implemented an implicit discourse relations model using the same features as in (Pitler *et al*., 2009) and 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 Market (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.

We proposed the first model for the Arabic language that fully addresses both explicit and implicit relations that link adjacent or non adjacent units within the Segmented Discourse Representation Theory framework. We used several kinds of features and analyzed how each feature contributes to the learning process. We first experimented with morphological and syntactic features, as already done by (Al-Saif and Market, 2011). Our results show that these features are primordial for discourse relation recognition but they are not sufficient for achieving good results. When adding contextual, lexical and lexico-semantic features, the results have been boosted for all configurations (Level1, Level2 and Level3).

2. The 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. The discourse structure of Example 1 is shown in Figure 4.1. 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 build as many instances as relations.

(1) [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.]$_4$

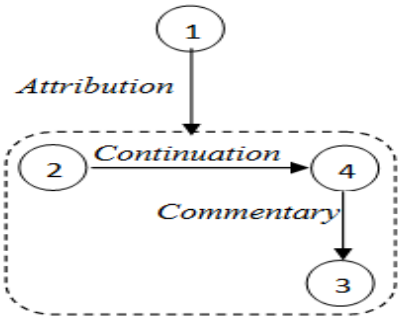

Figure 4.1. Discourse annotations for Example 1.

We designed thirteen groups of features. The first five groups (connectives, arguments, al-masdar, tense and negation, length and distance) follow (Al-Saif and Market, 2011)[21]. However, compared to (Al-Saif and Market, 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, 2000a). 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.

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

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

2.1. Al-Saif et al.'s features

(F1) Connectives. We have 6 string features that encode the connective string, the connective lemma, the connective POS, the connective position (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 Example 2, 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)))".

(2)[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 3, explains why Ahmed went to the library.

(3) [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)

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 4). 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 5). 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.

(4) [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]$_2$

(5) [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.

2.2. 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[22]) which can be helpful for identifying relations as خلفية/xlfyp/*Background-Flashback* and تأطير/t>Tyr/*Frame* (cf. Example 6) 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

[22] 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.

تلخيص/tlxyS/*Summary* (cf. Example 7) where the second argument usually occurs at the end of paragraphs.

(6) [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 degree,]$_2$ [two years Master 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)

(7) [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. Example 8). For each unit, we use 12 features that test for the presence of specific punctuations (!, ?, ., comma and :) as well as of typographical markers (“”, ( ), [], {}, _ and -). 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 doesn’t contain any specific punctuations or typographical markers.

(8) [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 to identify some relations such as تعليق/tElyq/*Commentary* and تعيين/tEyyn/*E-elaboration* (cf. Example 9).

(9) [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. Example 10). Moreover, the presence of pronouns and anaphora in the same argument can help identify the relation معية/mEyp/*Parallel* (cf. Example 11).

(10) [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)

(11) [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 (2,181 entries), people (2,309 entries) and organizations (403 entries). To test the presence of anaphora, we manually built a lexicon of 60 Arabic most frequent pronouns and anaphora (e.g. 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 (e.g. أكد/Akd/*confirm* يرى/yrY/*see*, يعتقد/yEtqd/*think*, أوضح/<AwDH/*explain*, لاحظ/lAHZ/*remark*). Modality can help detect relations like استدلال/AstdlAl/*Attribution* (cf. Example 12).

(12) [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 missing concepts and 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 (e.g. 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 4.1 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.

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)

Table 4.1. Examples of concepts related by AWN relations and some discourse relations that they can trigger.

(F12) Polarity. To deal with polarity information, we use the translated MPQA subjectivity lexicon (Elarnaoty *et al.*, 2012) that contains more than 8,000 English words and their corresponding Arabic translations[23]. 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

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

objective or a subjective sense depending on the context, such as الأحكام/Al>HkAm/*judgments*. Polarity can be of four types: positive, negative, both, and neutral.
We associate to each argument two string features: one for subjectivity that checks 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 training (i.e. 20 documents, 1,400 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 أجل أن من/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 ربط دون ترتيب زمني/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.

3. Experiments and results

The classifier aims to predict both explicit and implicit adjacent and non adjacent discourse relations. To this end, we carried out supervised learning on ADTB, based on the Maximum Entropy model (Berger *et al.*, 1996), as implemented in the Stanford MaxEnt package[24]. For all

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

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, Level3 with 24 relations), in mid-level classes (henceforth, Level2 with 15 relations) and also in the top-level classes (henceforth, Level1 with 4 relations). For each level, we have the same number of instances, i.e. 3,184 vectors. See Table 4.3 (cf. Section 2) 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 Level3 and Level2 and to the relation إنشائي/<n$A}y/*Thematic* for Level1. The second baseline (B2) is based only on lexical cues features (i.e. (F13), as described in the last section). Finally, the third baseline (B3) groups the features of (Al Saif and Market, 2011), 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 (Level1, Level2 and Level3). We finally conclude by presenting the learning curves.

3.1. 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 4.2 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:

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

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

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

	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**

Table 4.2. Overall results for the fine-grained classification.

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 Level3 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 (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) features are partially taken into account by Al-Saif et al.'s morphological and syntactic features.

Once we have empirically demonstrated the effectiveness of each feature group 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). Table 4.6 (the last row) 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 Level3, we have run the same model for mid-level relation classification (Level2) and for top-level classification (Level1). Table 4.3 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.

	Level2		Level1	
	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**

Table 4.3. Overall results for the mid-class (Level2) and coarse-grained (Level1) classification.

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 Level3, Level2 and Level1. 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 Market (2011) on identifying fine-grained discourse relations (accuracy=0.700, F-score=0.690) and on class-level relations (accuracy=0.835, F-score=0.750). 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.

Finally, it is interesting to note that our features alone (cf. (F6) to (F13)) lead to lower results compared to (B3) for all configuration levels. For example, on Level3, we obtain an F-score of 0.370 and an accuracy of 0.500. These results show that using only semantic features (e.g. modality, AWN, MPQA, etc.) can not outperform the baseline (B3) and that morphological and syntactic features are *primordial* for our task.

3.2. Fine-grained classification

In this section we analyze the impact of each feature group ((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. Figures 4.2, 4.3, 4.4, and 4.5 present respectively how F-scores evolve when adding each feature group.

Figure 4.2 shows that textual organization (F6) doesn't 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 Example 13.

(13) [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.]$_2$

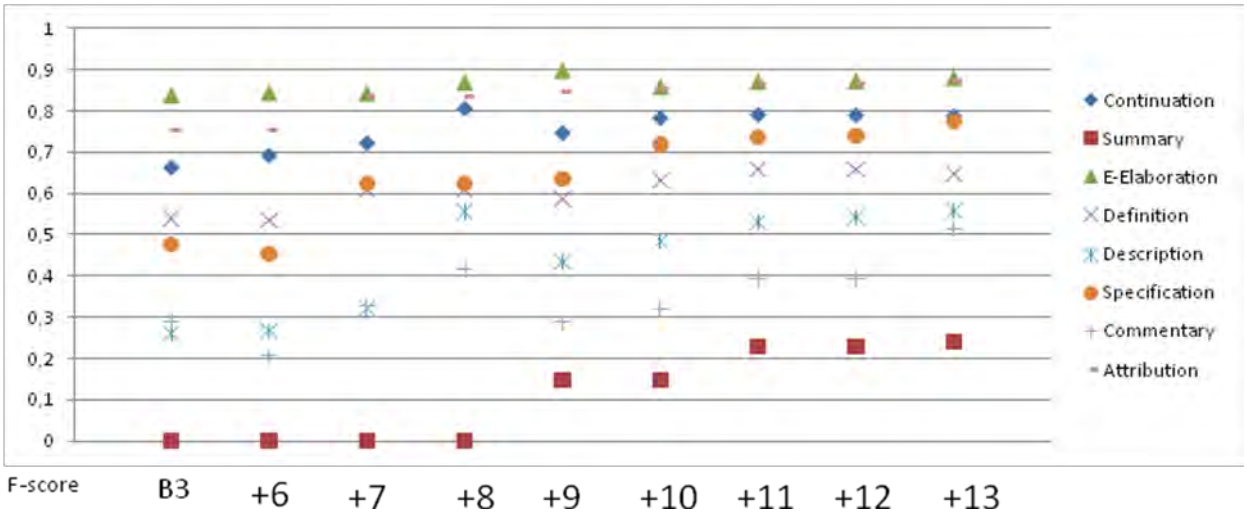

Figure 4.2. Feature impact on the **إنشائي**/<n$A}y/*Thematic* relations in terms of F-score.

In Figure 4.3, 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 Example 14. Embedding features (F8) do not seem to improve the results for all relations. Named entity and anaphora features (F9) boost the scores of all relations. This is very salient for تأطير /t>Tyr/*Frame* with an improvement of more 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 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*.

(14) [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]$_3$

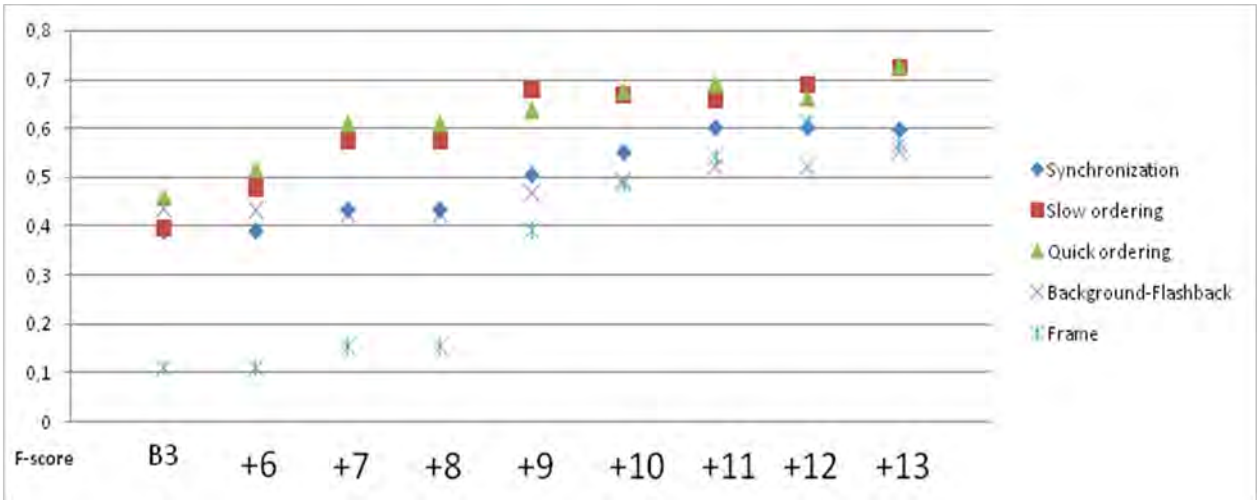

Figure 4.3. Feature impact on the **زمني**/zmny/*Temporal* relations in terms of F-score.

Figure 4.4 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/$rT/*Conditional* and إضراب/<DrAb/*Correction*, which are often signaled in Arabic by specific markers like إما/<mA/*either*, أو/>w/*or*, أم/>m/*or* and سواء/swA'/*either* for تخيير/txyyr/*Alternation* (cf. Example 15), س/s/*so*, لو/lw/*if*, إذا/<*A/*if* and لولا/lwlA/*except* for شرط/$rT/*Conditional*, and بل/bl/*however* for إضراب /<DrAb/*Correction*.

(15) [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

Example 16. 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 Example 17. 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.

(16) [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.]$_2$

(17) [يضحك أخي] $_1$ [ويبكي.] $_2$

[yDHk>xy]$_1$ [wybky.]$_2$

[My brother laughs]$_1$ [and cries.] $_2$

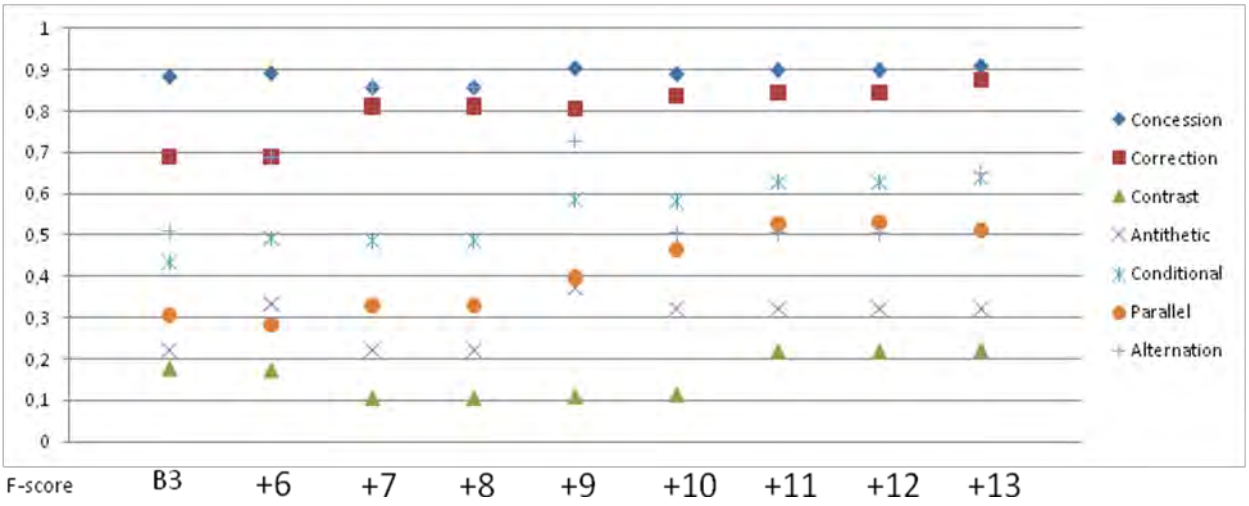

Figure 4.4 Feature impact on the بنيوي/bnywy/*Structural* relations in terms of F-score.

Finally, Figure 4.5 shows that our model fails to predict infrequent relations, such as 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.

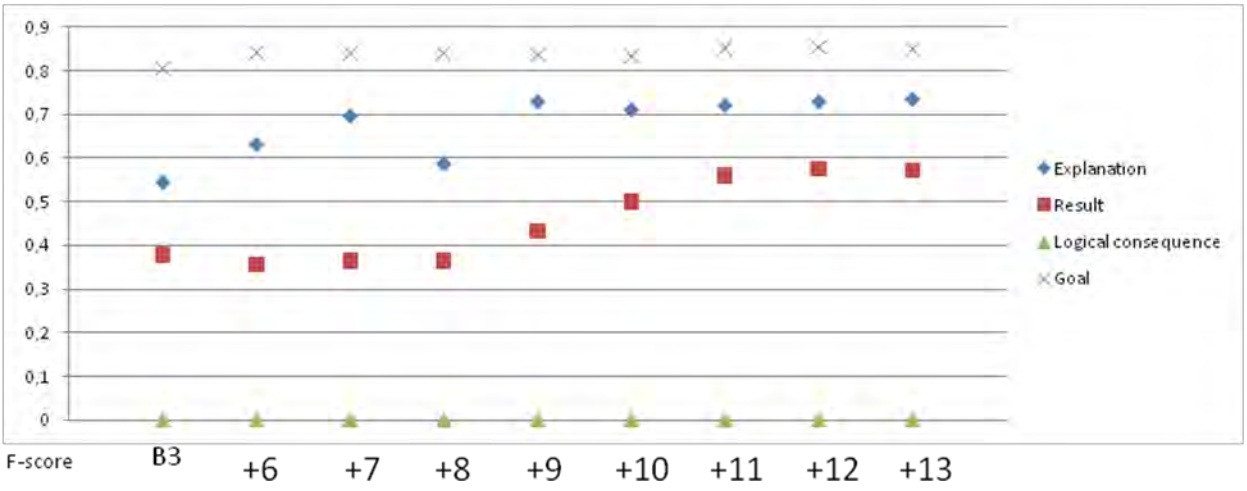

Figure 4.5. Feature impact on the **سببي**/sbby/*Causal* relations in terms of F-score.

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 (e.g. 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 Level3 shows that our model fails to discriminate between the relations غرض/grD/*Goal* and سبب/sbb/*Explanation* (cf. Example 18), the relations استدلال/AstdlAl/*Attribution* and تعيين/tEyyn/*E-Elaboration*, and the relations تعيين/tEyyn/*E-Elabor*ation and تفصيل/tfSyl/*Description*.

(18) [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)

3.3. Mid-level classification

Table 4.4 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).

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**

Table 4.4. Detailed results for the mid-level classification (Level2).

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

(19) [لقد استغنيت عن هذا الكتاب]$_1$ [انه لا يحتوي على معلومات قيمة،]$_2$

*[lqd Astgnyt En h*A AlktAb,]$_1$ [Anh lA yHtwy ElY mElwmAt qy-ymp,]$_2$*

[I don't need this book,]$_1$ [it doesn't contain any important information,]$_2$

Gold corpus: سبب/sbb/*Explanation* (1,2)

Predicting relation: إسهاب/<shAb /*Elaboration* (1,2)

3.4. Coarse-grained classification

Table 4.5 presents our results on the coarse-grained 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. The frequency of

each class in ADTB 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 Section 1).

Level 1	Precision	Recall	F-score	Accuracy
انشائي/<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**

Table 4.5. Detailed results for the top-level classification (Level1).

Table 4.6 shows major confusions. Main errors (in bold font) are between إنشائي/<n$A}y/*Thematic* and سببي/sbby/*Causal* classes.

	Thematic	Causal	Structural	Temporal
Thematic	1 727	**112**	52	45
Causal	**82**	422	21	27
Structural	38	34	261	33
Temporal	32	37	34	227

Table 4.6. Confusion matrix for the coarse-grained classification.

3.5. The 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 Figure 4.6. For Level1, the curve grows steadily between 0 and 2,000 discourse relations (that is 45 documents, i.e. around 1,200 sentences) while it seems to plateau between 2,000 and 3,184 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 Level2 seems to plateau between 2,400 and 3,184 discourse relations while the curve of Level 3 seems to plateau between 2,800 and 3,184 discourse relations.

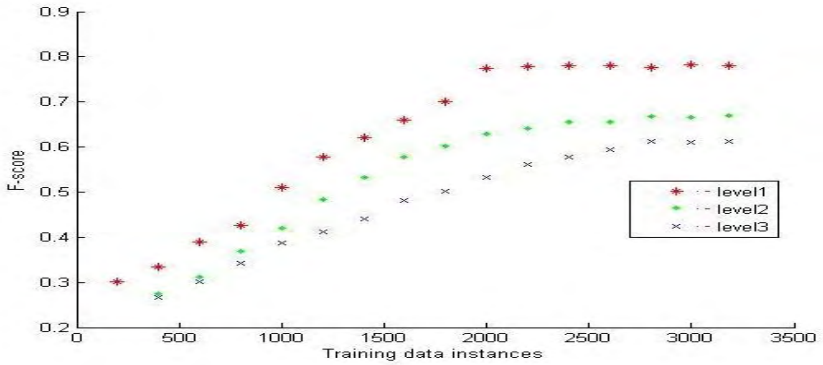

Figure 4.6. The learning curve of our three level models.

Conclusion

In this chapter, we presented the first work that fully addresses learning implicit and explicit Arabic discourse relations by proposing a multi-class supervised learning approach that predicts discourse relations between Elementary Discourse Units in Arabic texts.

Our approach used a rich lexicon (174 connectives) and relied on a combination of lexical, morphological, syntactic and lexico-semantic features. 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 outperform all the baselines. However, we note that attachment level has not been resolved in this chapter. This complex task needs more resources as used in discourse relation recognition task and more annotated documents. On the other hand, attachment task still has poor results for other languages, such as English.

To our knowledge, there has been little work that has so far been investigated how Arabic discourse analysis can improve the NLP application results (e.g. text summarization system, text translation system, Question/Answering system). In Chapter 5, we will investigate the performances of our discourse parser to efficiently perform Arabic text summarization. Indeed, we will propose a novel approach to automatic Arabic text summarization based on SDRT graph. Moreover, we will use the discourse relation semantics to extract the most important information from the Arabic text.

Chapter 5: Automatic text summarization using SDRT framework

Table of contents

Introduction 136
1. Related studies 138
1.1. Numerical and symbolic approaches 138
1.2. Main studies for Arabic 144
2. The data 148
2.1. ADTB corpus 149
2.2. AD-RST corpus 150
3. Content selection algorithms 151
3.1. Tree-based content selection algorithm (A1) 151
3.2. Graph-based content selection (A2) and (A3) 152
4. Examples 156
4.1. Example from AD-RST corpus 156
4.2. Example from ADTB corpus 157
5. Experiments and results 158
Conclusion 162

Introduction

In this chapter, we show how our partial SDRT discourse parser can be used in NLP applications. We focus in particular on automatic text summarization which aims at shortening a document or a set of documents by providing only the most relevant information. In the literature, many genres of summaries have been proposed (Hahn and Mani, 2000; Barzilay and McKeown, 2005; Jezek and Steinberger, 2008; Carenini and Cheung, 2008; Haghighi and Vanderwende, 2009; Wang *et al.*, 2009; Shen and Li, 2010; Qazvinian *et al.*, 2013; Cheung and Penn, 2013; Pai, 2014). We can cite the classification used in (Varghese and Saravanan, 2014): extractive/abstractive, generic/query-based, single-document/multi-documents and monolingual/multilingual/crosslingual.

Extractive summaries consider a document as a set of words, sentences or paragraphs and then select the most appropriate subsets that better summarize the original document. To produce abstractive summaries we need first to convert the document into a non linguistic representation (such as logical formulas) then to use natural generation techniques to generate natural language summaries from these formal representations. Abstractive (non extractive) summarization involves a deeper understanding of the input text, and is therefore limited to small domains. Query-based summaries are produced in reference to a user query (e.g., summarize a document about an international summit focusing only on the issues related to the environment) while generic summaries attempt to identify salient information in the text without taking into account the context of a query. The difference between single and multi-document summarization is quite obvious. Some multi-document summarization problems are qualitatively different from the ones observed in single-document summarization (e.g., addressing redundancy across information sources and dealing with contradictory and complementary information). This chapter focuses on generic Extractive Summaries of single Arabic documents (ExS).

ExS is the process of identifying the most salient information in a document or set of related documents. Salience can be defined in different ways because users may have different backgrounds, tasks, and preferences. Salience also depends on the structure of the source document. In addition, information which is salient for one user, may not be important for another. Therefore, it is very difficult to give consistent judgments about summary quality from human judges. This fact has complicated the evaluation (and hence, improvement) of automatic summarization.

ExS has received a great attention in the literature. Many types of extractive summaries have been proposed such as (Minel, 2002; Saggion and Lapalme, 2002; Jagadeesh *et al.*, 2005; Chatterjee and Mohan, 2007; Gupta and Lehal, 2010; Elsner and Santhanam, 2011; Cheung and Penn, 2014):

- *Indicative summary*. It aims at selecting from the source document a set of passages (sentences, paragraphs, etc.) to represent the whole document. This kind of summary helps users getting a general idea of a text without taking into account further details.
- *Informative summary*. It aims at representing all the relevant information of the original text. All major subjects or themes should be included in the summary.
- *Opinion or evaluative summary*. It focuses on summarizing user's judgments, evaluations and opinions.
- *Conclusion summary*. It is also known as recap summary or result summary. It provides only the results and the conclusions that are presented in the source text.

In this chapter, we propose a discourse-based approach to produce indicative summaries of Arabic documents. Our goal is to select the most relevant Elementary Discourse Units (EDUs) in the text that must contain the main information, events, objects, ideas, etc. For this purpose, we design several content selection algorithms that take as input the document discourse structure and produce as output a subset of EDUs which better summarizes the original document. The selection process is guided by three discursive criteria: the semantics of discourse relations, their nature (coordinating vs. subordinating) and the document discourse structure (tree vs. graph). To measure the impact of discourse structure on producing indicative summaries, we evaluate our algorithms by comparing their performances against gold standard summaries which are manually generated from two different corpora that have been annotated according to two different frameworks: the ADTB corpus (cf. Chapter 2), annotated according to the Segmented Discourse Representation Theory (SDRT), where each document is represented by an acyclic oriented graph, and the Arabic Discourse RST corpus AD-RST (Keskes *et al.*, 2012d), annotated according to the Rhetorical Structure Theory (RST) where each document is represented by an oriented tree. For each corpus, we perform two evaluation settings. The first one evaluates the automatic content selection algorithms when inputs are given by gold standard discourse structures. The second one is an end-to-end evaluation that takes as input the outputs generated by the partial discourse parser (described in Chapter 4).

This chapter is organized as follows. Section 1 gives an overview of indicative summarization approaches in general and on Arabic indicative summarization in particular. Section 2 presents our corpora. Section 3 details the proposed content selection algorithms. Section 4 reports on our experiments and details the results.

1. Related studies

1.1. Numerical and symbolical approaches

The headline of this sub-section tackles main existing work on ExS, by grouping them according to two main categories: numerical approaches which are based on statistics and machine learning methods and symbolical approaches which are based on linguistic rules. Besides these approaches, we notice an orientation towards hybrid approaches which combine numerical and statistical approaches.

1.1.1. Numerical approaches

Before presenting studies that tackle text summarization task using numerical approaches, we detail below the most used features:

- ***Word frequency.*** This method is based on the fact that the author uses some important words to express main ideas. Indeed, this suggestion focuses on the assumption that an author usually repeats certain words that are related. High frequency words present indicative elements to select the most relevant information in the document. In addition to word frequency, some studies propose to use the notion of "proximity" that aims at studying the distance, in terms of words, between the most frequent words in the text (Ellouze, 2004; He *et al*., 2008; Rene and Yulia, 2009; Maaloul, 2012).
- ***Title words***. This method uses the words present in the title to extract the most relevant sentences. Some studies have already shown that titles can have two types of word (Douzidia, 2004; Zhang *et al*., 2008; Pallavi and Mane, 2014): "full", for title words that introduce important information in the text and "empty" for the other words. The selected sentences must contain the maximum of "full" word.
- ***Sentence position***. This method stipulates that the relative position of a sentence in a paragraph or a text determines the degree of its importance. Usually, the first and last sentences of a paragraph are included in the summary (Canasai and Chuleerat, 2003; Yeh *et al*., 2008; Gupta and Lehal, 2010; Suanmali *et al*., 2011).
- ***Lexical co-occurrences***. This method uses the lexical co-occurrences to calculate the frequency of each word in the text and to assign a score to each sentence. For instance, the sentence which contains the most frequent word gets the highest score. The final summury contains the set of sentences with the highest scores (Ellouze, 2004; Alguliev and Aliguliyev, 2005; Zamanifar *et al*., 2008; Gupta and Lehal, 2010; Maaloul, 2012).
- ***Indicative expressions.*** Two types of expression have been defined (Saggion, 2000; Zhanq *et al*., 2005; Osminin, 2014): (1) *bonus*, are mainly superlatives ("biggest", "bravest", "coldest", "easiest", "quickest", etc.) and indicative expressions (such as "this article presents", "summarizing", "in conclusion", etc.) which indicate that the author announces

the general theme of the text. (2) *stigma* are mostly anaphora and words that introduce secondary information (such as "for example" "indeed", "other", "in other words", etc.). Bonus expressions increase the score of a sentence whereas stigma expressions decrease its score.

Some studies have used the previous methods as features to build learning-based approaches. These approaches include binary classifiers (Kupiec *et al.*, 1995; Zhu and Penn, 2006), Markov models (Conroy *et al.*, 2004; Dunlavy *et al.*, 2007), Bayesian methods (Aone *et al.*, 1998; Maskey and Hirschberg, 2005; Daume III and Marcu, 2005; Wang *et al.*, 2008), and heuristic methods that determine feature weights (Schiffman, 2002; Lin and Hovy, 2002). We highlight below main existing work on machine learning approaches to automatic summarization.

Minel (2002), Amini and Gallinar (2003), and Amini and Usunier et al. (2007) have adopted supervised learning to extract the most relevant information. In addition to the previously cited numerical methods, the authors used morpho-syntactic features as well as other common features (i.e. sentence length, word length, word position, etc). Again, Amini and Gallinari (2003) have used both semi-supervised and unsupervised learning based on neuron networks to summarize a corpus of one million dispatches from Reuters News Agency[25]. Finally, Aliguliyev (2006), Alguliev and Aliguliyev (2008) and Aliguliyev (2010) have used sentences clustering for automatic document summarization.

Wang et al. (2008) proposed a new framework based on sentence-level semantic analysis (SLSS) and symmetric non negative matrix factorization (SNMF). The authors construct the similarity matrix (the sentence-sentence similarities) using semantic analysis. They used semantic roles parsing to describe the relationship that a constituent plays with respect to the verb in the sentence. This semantic analysis is basing on PropBank semantic annotation (Palmer *et al.*, 2005). Then, they calculate the similarity between each two sentences using the symmetric matrix factorization to conduct the clustering (group sentences into clusters). The similarities are computed using the semantic relations of terms in WordNet (Fellbaum, 1998). Finally, the most informative sentences are selected from each cluster to form the summary. In the same context, Lee et al. (2009) presented an unsupervised generic document summarization method using SNMF. Authors have benefited from the advantages of the unsupervised method (i.e. does not require training summaries for the summarizer and the training step) and provide better performance in identifying subtopics of a document, as compared to the methods using SLSS. Indeed, authors can more intuitively find comprehensible semantic features used for determining subtopics of documents.

Gupta and Lehal (2010) have used a cluster-based method. The authors built a set of triplets (subject, verb, objects related to each sentence) to capture and express the semantic nature of a

[25] http://boardwatch.internet.com/mag/95/oct/bwm9.html

given document. Then, the authors clustered these triplets (considered as the basic unit in the process of summarization) using term frequency-inverse document frequency (TF-IDF) (Yongzheng *et al.*, 2005). Term frequency used in this context is the average number of occurrences (per document) over the cluster. IDF value is computed based on the entire corpus. The summarizer takes already clustered documents as input. Each cluster is considered a theme. The theme is represented by words with top ranking term frequency, inverse document frequency (TF-IDF) scores in that cluster.

Binwahlan et al. (2010) and Suanmali et al. (2011) used a *Fuzzy logic* for the Text Summarization task. *Fuzzy logic* uses decision module to compute the importance sentence degree based on its rated features. Decision module is designed using a fuzzy inference system. It works in four steps: (1) text preprocessing, (2) feature extraction of both words and sentences, (3) Fuzzy logic scoring, and (4) extracting sentences of higher ranks to generate summary. During the third step, the sentence features are divided into five fuzzy set (very low, low, Medium, high, and very high). The important part in this step is the definition of fuzzy IF-THEN rules. The important sentences are extracted from these rules according to eight feature criteria. The last step in fuzzy logic system is the defuzzification to convert the fuzzy results from the inference engine into a crisp output for the final score of each sentence. Suanmali et al. (2011) used further genetic algorithm and semantic role labelling to improve the quality of summary. The authors exploited the benefits of the genetic algorithm in the optimization problem for feature selection. Fuzzy IF-THEN rules were used to balance the weights between important and unimportant features. Binwahlan et al. (2010) used further a model based on Particle Swarm Optimization (PSO) to obtain the weights of the sentence features. To extract sentences for the final summary, they used an objective function composed of cohesion, readability and relationship with the title.

Abuobieda et al. (2012) presented an hybrid approach for feature selection using a genetic algorithm and probabilistic theory extractive-base single document summarization. The authors selected a random set of features using a (pseudo) Genetic concept as an optimized trainable features selection mechanism. To test the ability of the proposed model while doing feature selection rather than investigating the features themselves, the features are represented and encoded using the structure of binary genes, while their appearance is governed using probability. Indeed, each gene refers to a feature represented in binary format level. If the gene position (bit) holds a value of 1, it means that the corresponding feature is active and counted in the final score, otherwise, if the bit contains zero, it means that the corresponding feature is inactive and shall not be considered in final score.

Mendoza et al. (2014) proposed a method of extractive single-document summarization based on genetic operators and guided local search. The authors addressed the summarization task of a single document as a binary optimization problem where the quality (fitness) of the solutions is based on the weighting of individual statistical features of each sentence (such as position,

sentence length and title words). Two fitness functions are proposed to allocate a score to each sentence in the document: the first function is based on individual statistical features of each sentence and the second function is based on similarity features between sentences. Finally, the authors used a memetic algorithm (evolutionary algorithms with local search heuristics) to integrate guided local search strategy. Memetic algorithm contributed to the successful resolution of different combinatory optimization problems (Cobos *et al*., 2010; Neri and Cotta, 2012).

1.1.2. Symbolical approaches

The symbolical approaches are mainly based on the representation of document into tree or graph structure. There are two kinds of representation approaches: (1) the discourse structure representation approaches that use coherence discourse relations identified in the text to represent discourse structure and (2) hierarchical structure representation approaches that use topics and themes to represent the document into hierarchical structure or graph structure. In (1), approaches differ with respect to what kind of discourse structure they are intended to represent. Most accounts of discourse coherence assume tree structures (Mann and Thompson, 1988). Some accounts do not allow crossed dependencies but appear to allow nodes with multiple parents (Lascarides and Asher, 1991). Other accounts assume that less constrained graphs allow crossed dependencies as well as nodes with multiple parents (Wolf and Gibson, 2005). In (2), the first step is identifying the issues or topics addressed in the document. After the common preprocessing steps, namely, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph. There is a node for each sentence. Two sentences are connected with an edge if the two sentences share some common words. The nodes with high cardinality (number of edges connected to that node) correspond to the relevant sentences.

We detail below the two main approaches used for the summarization task: tree-based approaches and graph-based approaches.

1.1.2.1. Tree-based approaches for ExS

Marcu (1998) showed the importance of using discourse segments (and not sentences) for ExS. Given that a discourse segment is generally smaller than a sentence, it helps to select the most pertinent information in a sentence. Besides, the author used the concept of Nucleus/Satellite to identify the most important segments in the text (cf. Chapter 1). Indeed, the nucleus segments are crucial to achieve the coherence of the text, so they are potentially useful for the summary. A satellite must be associated with a nucleus to be intelligible. Each parent node identifies its nuclear children as salient. Sentences are penalized according to their rhetorical role in the tree. A weight of 0 is given to nuclei units and a weight of 1 is given to satellite units. The final score of sentences is given by the sum of weights from the root of the tree to the sentence.

Ono et al. (1994) used the same concept where segments are penalized according to their rhetorical role in the tree; a score of 1 is assigned to each nucleus segment and a score of 0 to each satellite segment. The final score of a sentence is calculated by summing the score from the root of the tree up to the sentence. Bosma (2005) has proposed a Query-Based Summarization using RST. The author shows how answers to questions can be improved by extracting more information about the topic with summarization techniques for a single document extracts. RST is used to create a tree representation of the document – a weighted tree in which each node represents a sentence and the weight of an edge represents the distance between two sentences. If a sentence is relevant to an answer, a second sentence is evaluated as relevant too, based on the weight of the path between the two sentences. The result is an answer that is more informative than an 'exact answer' (as returned by traditional QA systems), and more concise than a full document (as returned by IR systems). Additionally, Yong-dong et al. (2007) have proposed Multi-document Rhetorical Structure (MRS) for the summarization task. This structure represents multiple relationships between text units at different levels of granularity (sentences, paragraphs, sections and documents) including rhetorical relationships, semantic relationships and temporal relationships. Moreover, it can describe simultaneously the change of various events. MRS simplifies traditional multi-document representation in cross structure theory and supplement change and distribution information of events topics which cannot be obtained in information fusion theory. Concretely, a series of algorithms including building MRS, multi-document information fusion based MRS and summarization generation are proposed.

The reported experiments using RST to produce a summary are promising (Da Cunha *et al.*, 2007). However, the lack of efficient automatic discourse parser for long texts, which identify the structural composition of documents, present a major problem.

1.1.2.2. Graph-based methods for ExS

Using an empirical study of 135 texts from the Wall Street Journal and the AP Newswire, Wolf and Gibson (2005) showed that trees are not a descriptively adequate data structure for representing discourse structure. In coherence structures, authors found many different kinds of crossed dependencies, as well as many nodes with multiple parents. The authors proposed to use graph discourse structures rather than trees. They used informational-level-based taxonomies (Hobbs, 1985) to build the text graph structure. Then, the authors used this structure to calculate the importance of segments

Kruengkari and Jaruskulchai (2003) proposed a graph-theoretic method to identify the important sentences in a document. There is a node for each sentence. Two sentences are connected with an edge if the two sentences share some common words, or in other words, their similarity (cosine or such) is above some threshold. This representation yields two results: the partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub

graphs) form distinct topics covered in the document. The nodes with high cardinality (number of edges connected to that node) are the important sentences in the partition, and hence carry higher preference to be included in the summary.

Mihalcea and Tarau (2005) proposed a language independent extractive summarization that relies on iterative graph-based ranking algorithms. In these algorithms, the importance of a vertex within the graph is iteratively computed from the entire graph. A graph is constructed by adding a vertex for each sentence in the text, and edges between vertices are established using sentence inter-connections. These connections are defined using a similarity relation. The similarity is measured as a function of content overlap. The overlap of two sentences can be determined as the number of common tokens between two sentences. The execution of ranking algorithms on the graph provides sorted sentences in reversed order according to their score. The final summary contains just the top ranked sentences.

Banu et al. (2007) proposed a semantic graph approach by identifying triples of Subject Object Predicate from sentences of source document. Then, authors applied a syntactic analysis to compress sentences. Authors also used the triples of SOP for reducing the frequency of nodes of semantic graph of source document.

Qazvinian et al. (2013) proposed C-LexRank, a graph-based summarization method. This method models a set of citing sentences as a network in which vertices are sentences and edges represent their lexical similarity. The authors identified vertex communities (clusters) in this network to generate summaries, by extracting representative sentences from the citation summary network. Therefore, a good sentence selection from the citation summary network will include vertices that are similar to many other vertices and which are not very similar to each other. On the other hand, a bad selection can include sentences that represent only a small set of vertices in the graph. Finally, the authors compared C-LexRank with the state-of-the-art summarization systems where this method outperforms leverage diversity method (Mei *et al.*, 2010), random summaries method (Erkan and Radev, 2004) and LexRank method (Zajic *et al.*, 2007).

Zhang et al. (2008) proposed an adaptive model for summarization (AdaSum), under the assumption that the summary and the topic representation can be mutually boosted. AdaSum aims at optimizing the topic representation as well as extracting effective summaries. A graph-based subtopic partition algorithm for summarization (GSPSummary) is proposed by ranking sentence importance with the "personalized" LexRank and removing redundancy with sub-topic partition, where the global features are taken as the "personalized" vector for LexRank.

Wan (2010) used graphs for the automatic generation of extractive summaries. The author carried out simultaneously the summaries of a single document as well as multiple documents. He used the local importance that indicates the relevance of a sentence within a document to generate the summary of a single document; and of a global importance, that indicates the

relevance of the same sentence. However, this relevance is related to the entire set of documents to generate the summary of multiple documents.

Cheng et al. (2013) introduced a single document summarization method based on a triangle analysis of dependency graphs. The authors proposed an algorithm, called TriangleSum that built a dependency graph for the underlying document based on co-occurrence relation and syntactic dependency relations. Indeed, nodes represent words or phrases of high frequency, and edges represent dependency-co-occurrence relations between them. Moreover, the authors computed the clustering coefficient from each node to measure the strength of connection between a node and its neighbors in a dependency graph. By identifying triangles of nodes in the graph, a part of the dependency graph can be extracted as key of sentences. As results, TriangleSum extracted a set of key sentences that represent the main document information.

A comparative study proposed by Louis et al. (2010) aimed at analyzing which discourse structure provides the strongest indication for text content selection. First, the authors examined the benefits of both the discourse structures and the semantic sense of discourse relations. Their result showed that the discourse structure information is the most robust indicator for measuring the importance of segments. However, semantic sense of discourse relation complements the discourse structure information and leads to improve the performance. Second, the authors gave a comparison between graph vs. tree discourse structure for content selection. The discourse graph structures turn out as strong indicators of segment importance. In fact, the better performance of graph structures comes from higher recall score compared to tree structure; their precision score is comparable. Finally, given that building graph structure is more challenging, authors proposed a general text graph method. It focused on lexical similarity (lexical overlap information) to build the text structure instead of discourse relations. The authors used *cosine similarity* to link sentences in the lexical graph. Links with similarity less than 0.100 were removed to filter out weak relationships. The lexical graph gives the best results, with an F-score of 0.530 (an F-score of 0.480 for graph structure and an F-score of 0.420 for tree structure). Finally, we can cite Webber et al. (2011) who gave a survey of text summarization applications that use discourse structure analysis.

1.2. Main studies for Arabic

For the Arabic language, Douzidia (2004) proposed a generic extractive summarization system called “Lakhas” based on numerical approaches. The objective was to identify the features characterizing relevant contents in a document and extract the linguistic marks which can express pertinent information. The author has introduced compression technique to enhance the quality of summaries produced by “Lakhas”. This tool is composed of different modules. The first module focuses on the segmentation of a text into different levels (paragraphs, sentences, and words). It first segments a text into paragraphs and sentences, and then each sentence is tokenized into

words according to spaces and punctuation marks. The second module concerns the normalization of the target document in a standard format for easy manipulation. This normalization includes the suppression of special characters, the replacement of some Arabic letters such as (أ or إ with ا, ة with ه, and ي with ى). The third module focuses on the suppression of stop words based on an anti-dictionary. Then, a lemmatization is applied to each word and a score is associated to each word in order to generate the summary. This score is computed according to the frequency of the word in the sentence. This score will be increased in case of indicative expressions (cf. Section 1.1.1). Also, another score is computed for the sentences using Formula 1.

$$(1) \quad \text{Score}_{\text{hybride}}(S) = a_1 * \text{Score}_{\text{tf*idf}}(S) + a_2 * \text{Score}_{\text{lead}}(S) + a_3 * \text{Score}_{\text{cue}}(S) + a_4 * \text{Score}_{\text{titre}}(S)$$

were the tf*idf score is computed using term frequency-inverse document frequency, the lead score is extracted from leading sentences up to the given threshold, the cue score is computed according to sentence cues, and the title score is computed according to title words in the sentence.

Thanks to its flexibility, the different modules of this summarizer tool can communicate together. The comparison of this tool with the Arabic summarizer of Sakhr (Chalabi, 2001) and the Pertinence summarizer (Lehmam, 2000) reported that "Lakhas" is a competitive tool. Also, to evaluate the "Lakhas" summarizer, the author has participated in the Document Understanding Conference campaign[26] DUC 2004 and the tool was ranked at the fifth position using the ROUGE-1 measure (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003). ROUGE-1 is a metric used to evaluate the similarity between produced summaries and reference summaries. It is a 1-gram recall between a candidate summary and a set of reference summaries.

Alrahabi et al. (2004) proposed a semantic filtering system of Arabic texts, based on the contextual exploration method. Its principle is based on linguistic knowledge and allows to find the relevant information using linguistic markers (e.g. thematic segments, definition utterances, titles, underlining, summing ups, and conclusions). Using the same method (the contextual exploration method), Alrahabi and Desclés (2009) proposed a platform for semantic annotation, called "EXCOM" that enables, across a great range of languages, to perform automatic annotations of textual segments by analyzing surface forms in their context. Texts are approached through discursive "points of view", of which values are organized into a "semantic map". The annotation is based on a set of linguistic rules, manually constructed by an analyst, and that enables to automatically identify the textual representations underlying the different semantic categories of the map. The system provides through two sorts of user-friendly interfaces (analyst

[26] A workshop focuses on summarization and the evaluation of summarization with large-scale experiments.

or end-user) a complete pipeline of automatic text processing which consists of segmentation, annotation and other post-processing functionalities. Annotated documents can be used, for instance, for information retrieval systems, classification or automatic summarization. Alrahabi (2010) proposed a second version of this platform called "EXCOM-2". This new version adds an analysis of the linguistic markers of the enunciative modalities in direct reported speech in a multilingual framework concerning Arabic and French.

El-Haj and Hammo (2008) proposed a query-based Arabic text summarization system. The authors adapted the traditional Vector Space Model (VSM) and the cosine similarity measure to determine the most relevant passages extracted from Arabic document to produce a text summary. The system consists of two main modules: i) the Document Selector that selects relevant documents from a document collection based on a user query. This module is based on a concordance method, which simplifies the documents collection using an alphabetical index of all unique words in the collection along with their occurrences. It is used to locate documents based on simple matching techniques between the query's bag-of-words and the document collection. The user then selects the document to be summarized. ii) The Single Document Summarizer that extracts a set of the most relevant paragraphs from the original document. After paragraph splitting, the authors used a matching technique such as the cosine measure to match the paragraphs against the same query used to retrieve the documents.

Lehmam (2010) built an automatic text summarization system called Essential Summarizer, which takes into account discursive elements of the text. This system produces summaries in twenty languages, including Arabic. The system used five steps: 1) recognition of semantic cues called Semantic Extraction Markers (SEMs) to determine relevant sentences and of paragraphs to be selected for the summary; 2) Specialization by domain to better target the summary; 3) Consideration of expressions or concepts that are important for the user's needs; 4) Observation of manual summarization of representative texts and analysis of user feedback.

El-Haj et al. (2011) proposed an Arabic concept-based text summarization system. Unlike El-Haj and Hammo (2008) which used standard retrieval methods to map a query against a document collection and to create a summary, this system creates a query-independent document summary. Indeed, it takes a bag-of-words representing a certain concept as the input to the system instead of a user's query. The summary consists of sentences that best match the words in the query or concept. The sentence matcher module of the Arabic concept-based text summarization system ignores the user query that was used to select the documents. Instead, each sentence is matched against a set of keywords representing a given concept. On the other hand, El-Haj et al. (2011) discussed the results of the two summarization systems for Arabic by reporting on five groups of users from different ages and educational levels. The authors used Wikipedia text to test the two systems using a set of forty queries to retrieve a set of documents. The system generates a summary for each returned document. A group of 1,500 users

participated in evaluating the readability of the generated summaries. Finally, the authors claimed that the query-based summarizer performs much better than the concept-based summarizer.

Another work on Arabic text summarization was done by (Mathkour *et al*., 2008) who adopted a symbolic method based on Rhetorical Structure Theory (RST). They used discourse markers and frequently co-occurring word pairs to identify the discourse relations. The authors designed a rule-based discourse parser for Arabic and cues to identify the discourse relations. The proposed approach extracts the Arabic rhetorical relations based on studying the English relations, analyzing Arabic corpus and using an Arabic cue phrases. This approach is based on the translation of English relations and cue phrases into Arabic. Only English relations and cue phrases found in Arabic corpus are used for the text summarization (11 discourse relations). For text summarization, the authors pruned the suitable tree by selecting relevant segments relying only on the nucleus/satellite distinction. A comparison between their summarization tool outputs and a manual summarization gives an overall precision of 0.620. These results are very sensitive to the form of the rhetorical trees. Indeed, the trees that were the most balanced were the most suitable to generate summaries.

Azmi and Al-Thanyyan (2012) proposed an hybrid two-pass summarization. The first pass uses the RST tree first levels (Mathkour *et al*., 2008) to generate a primary summary, while the second pass uses the primary summary to produce a shorter version. The second pass computed the score sentences of the primary summary using formula (1). The authors claimed that the two-pass summarizer improves the basic RST summarizer.

In the same context, Keskes et al. (2012d) used the RST framework to build the final summary. Indeed, the authors tried to find the RST relations (Marcu, 2000b) in AD-RST (100 texts selected from the journal "Dar Al Hayet") using the translation of discourse markers of each discourse relation into Arabic. Referring to Arabic experts, 16 rhetorical relations have been determined and 4 new relations dedicated to the Arabic language have been identified (*Restriction* signaled by the markers إلا/<lA/*unless*, غير/gyr/*except* …, *Confirmation* signaled by the markers إن/<n/*that*, قد/qd/*have* …, *Specification* signaled by the markers خاصة/xASp/*especially*, بالخصوص/bAlxSwS/*in particular*, and *Affirmation* signaled by the markers لا/lA/*no*, لم/lm/*not*, etc.). For the content selection, authors used both the nucleus/satellite notion and the discourse relation semantics to prune the RST tree. Only 9 rhetorical relations, chosen by Arabic experts, are used for the summarization task. The results achieved an F-measure of 0.500. Belguith et al. (2014) extend this work using a machine learning method to predict the suitable discourse relations when these latter are implicit or present an ambiguous discourse marker. The authors performed an improvement of F-measure to reach 0.530.

Oufaida et al. (2014) proposed a statistical summarization system mRMR for Arabic texts. This system uses a clustering algorithm and an adopted discriminant analysis method of score

terms to ensure a minimum redundancy and a maximum relevance. Using mRMR system, terms are ranked according to their discriminant and coverage power, whose goal is to select a subset of features which significantly represents the whole space of features. It is based on mutual information[27] between pairs of features, which reflects the level of similarity between them. This system built different configurations on how to use the scoring method, depending on the requested summary size (Very Short: speed decrease, Short: slow decrease). Moreover, the scoring method uses minimum language-dependent processing, only at the root extraction level and does not use any structural or domain-dependent features. mRMR system selects sentences with top ranked terms and maximum diversity based on minimal language-dependant processing: sentence splitting, tokenization, and root extraction. Experimental results in The TAC MultiLing 2011 workshop (Giannakopoulos *et al*., 2011) showed that mRMR system is competitive to the state of the art systems.

In this thesis, we propose a novel discourse-based approach to summarize Arabic texts based on the Segmented Discourse Representation Theory. Our aim is to select the most relevant EDUs using the graph discourse structures and the semantic of discourse relation. We use our discourse parser that fully addresses both explicit and implicit relations to link adjacent as well as non adjacent units within the SDRT framework. For the evaluation, we use our ADTB corpus which has been manually summarized by two experts to compare 4 algorithms for content selection within different criteria. Moreover, we use AD-RST (Keskes *et al.*, 2012d) to study the difference in terms of the quality of the summary between using discourse graph and using discourse tree in Arabic texts. Experts will judge this difference.

2. The data

We use two different corpora that have two different frameworks: ADTB (cf. Chapter 2), annotated according to the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) and AD-RST (100 texts selected from the journal "Dar Al Hayet") (Keskes *et al*., 2012d), annotated according to the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). For each corpus, we ask two Arabic native speakers to manually select the most pertinent EDUs from each document, following the annotation guidelines already proposed in the literature (Belguith *et al*., 2014). In particular, we did not impose any restrictions on the number of selected EDUs, their position in the document or their length in terms of words count. Each annotator produces one summary per document. Gold standard summaries have been built by selecting for each document, in a given corpus, the EDUs commonly chosen by the two annotators. Our

[27] Mutual information aims to measure the information quantity that two features share. Therefore, if two features have a high mutual information quantity, then they are highly correlated and consequently, one can replace the other with minimum information loss.

algorithms have been evaluated by comparing their performances against gold standard summaries. We detail below our data.

2.1. ADTB corpus

As described in Chapter 2, ADTB corpus contains 70 documents with a total of 4,963 EDUs. 20 texts have been used to train our annotators, which correspond to a total of 1,432 EDUs. After training, two annotators were asked to manually generate two summaries for each text. The inter-annotator agreements have been computed on the entire corpus through of the *Kappa* measure. We obtained a Kappa of 0.770. Some statistics on ADTB are shown in Table 5.1.

	Number of Texts	**Size**	**Number of Sentences**	**Number of Words+Punctuations**	**Number of EDUs**	**Number of Selected EDUs for summaries**
ADTB corpus	50	267 ko	1,272	28,288	3,540	780

Table 5.1. ADTB characteristics.

In the gold standard, the average number of EDUs per summary is 15.6 and the average size of a summary is 22% of the source text. We finally note that 30% of the selected EDUs are from the beginning of paragraphs, among which 0.5% are embedded EDUs.

After the annotation campaign, the two annotators were asked to select a subset of discourse relations from our relation hierarchy (cf. Chapter 2, Table 5.5) which are considered to be useful for the summarization task. Indeed, the annotators chose the discourse relations that potentially contain relevant EDUs as arguments. The selection criteria are given according to the semantics and the definitions of the discourse relations. Among the 24 relations, annotators selected 15 relations, as shown in Table 5.2.

Selected Discourse relations	**Type**
شرط/$rT/*Conditional*	Coordinate
سبب/sbb/*Explanation*	Subordinate
نتيجة/ntyjp/*Result*	Subordinate
استنتاج/AstntAj/*Logical consequence*	Subordinate
تلخيص/tlxyS/*Summary*	Subordinate
غرض/grD/*Goal*	Subordinate
تعريف/tEryf/*Definition*	Subordinate
تزامن/tzAmn/*Synchronization*	Coordinate
خلفية/xlfyp/*Background-Flashback*	Subordinate
مقابلة/mqAblp/*Contrast*	Coordinate
طباق/TbAq/*Antithetic*	Coordinate
استدراك/AstdrAk/*Concession*	Coordinate
إضراب/<DrAb/*Correction*	Subordinate
استدلال/AstdlAl/*Attribution*	Subordinate
ربط دون ترتيب زمني/rbT dwn trtyb zmny/*Continuation*	Coordinate

Table 5.2. SDRT discourse relations selected for the summarization task.

It is interesting to note that annotators considered the discourse relation ربط دون ترتيب زمني/rbT dwn trtyb zmny/*Continuation* as being an important relation for the summarization task even if this relation has a weak semantics. This can be justified by two reasons: this relation can link paragraphs and it often links Complex Discourse Units (CDUs).

2.2. AD-RST corpus

In ADTB, documents are represented by an acyclic oriented graph. In order to compare the impact of different discourse structures on the content selection, we also evaluate our algorithms against summaries generated from tree-based discourse representations. To this end, we use the Arabic Discourse RST corpus AD-RST (Keskes *et al.*, 2012d) which contains 100 documents selected from Dar Al Hayat news paper. Each document has been annotated according to the Rhetorical Structure Theory (RST). The annotation of this corpus proceeded as follows. First, annotators segmented each document into spans[28] (cf. Chapter 1), using only explicit discourse markers and punctuation marks. Hence, there are no embedded segments. Then, they were asked to connect adjacent spans by means of RST discourse relations. Only one discourse relation can be used to link two spans. The set of relations used in this annotation campaign has been inspired from three main sources: a translation of the 7 English discourse relations defined by (Marcu, 2000b) into Arabic, the set of 11 Arabic discourse relations defined in Mathkour et al. (2008)[29], and the analysis of discourse relations in our corpus. This procedure resulted in a set of 19 Arabic discourse relations, such as condition, evidence, concession, and ordering). For each discovered relation, we built a list of rhetorical frames that contain the Arabic discourse markers (Keskes *et al.*, 2012d). Table 5.3 presents an example of a rhetorical frame for the relation تخصيص/txSyS/*Specification*. Finally, annotators built the document discourse structure (RST tree) following the RST guidelines (Marcu, 2000b), after training on 20 Arabic texts. To build summaries, annotators adopted the same annotation procedure as for ADTB (cf. last Section). Table 5.4 presents the characteristics of the gold standard corpus AD-RST.

Discourse relation	تخصيص/txSyS/*Specification*
Constraints on EDU$_1$	Contains one or more indicators[30]: لا /lA/no,لم /lm/no,ليس /lys/not, etc.
Constraints on EDU$_2$	Contains the discourse marker لاسيما/lAsymA/especially
Discourse marker position	Middle
Nucleus	EDU$_2$

Table 5.3. Rhetorical frame of the relation **تخصيص**/txSyS/*Specification*.

[28] The discourse segmentation principles used in the RST framework is different from the SDRT framework (size, markers, punctuations, etc.). Using the RST framework, we segment text into spans and not EDUs.

[29] We note that all the discourse relations of Mathkour et al. are translated from the English discourse relations defined by (Marcu, 2000b).

[30] The indicator is not a discourse marker, it help annotators to select the suitable discourse relation (Keskes et al., 2012d).

	Number of Texts	Size	Number of Sentences	Words+Punctuations	Number of spans	Number of Selected spans for summaries
AD-RST	100	521 ko	2 098	61 021	3 894	1 212

Table 5.4. AD-RST characteristics.

In AD-RST, a summary has an average of 12 EDUs per document and the average size of a summary is 31% of the source text, which is larger than the summaries produced in ADTB by 9%. This difference is due to segmentation principles. Indeed, the size of spans is longer compared to the EDU length.

Like ADTB, two annotators were asked to select a sub-set of RST discourse relations from the main list (19 rhetorical relations) which are useful for the summarization task. 8 discourse relations have been chosen as shown in Table 5.5.

RST discourse relations selected for summarization task	
Condition	Evidence
Concession	Ordering
Restriction	Affirmation
Confirmation	Definition

Table 5.5. RST discourse relations selected for the summarization task.

3. Content selection algorithms

Our algorithms have as input a document discourse structure (a graph or a tree), prune it according to discursive criteria and output a subset of EDUs[31] that are deemed to be relevant. We have mainly used three pruning criteria: the semantics of discourse relations (which correspond to the subset of relations selected by our annotators (cf. Table 5.2 for ADTB and Table 5.5 for AD-RST), their nature (coordinating vs. subordinating) and the document discourse structure (tree vs. graph). We designed six algorithms. The first one takes a tree as an input while the five others a graph. The next sections will detail our approach.

3.1. Tree-based content selection algorithm

Let D be a document such as D= {EDU_1, …, EDU_i}, let Relation(D) be the set of discourse relations of D such as Relation(D)= {$R_1(EDU_i, EDU_j)$, …, $R_w(EDU_x, EDU_z)$}, let Nuclei(D) be the set of nuclei segments of D, let Rel_RST the set of relevant RST discourse relations (cf. Table

[31] To refer to the text unit generated in the RST framework, we use the same notion used in the SDRT framework: Elementary Discourse Unit (EDU).

5.5), and let Sum= {} be the set of relevant segments that have to be included in the final summary. The tree-based content selection algorithm (A1) requires four steps:

1- For each EDU_i , $EDU_j \in$ D, if $R_w \in$ Relation(D) and $R_w(EDU_i, EDU_j)$ and $R_w \notin$ Rel_RST, remove EDU_i and EDU_j from D.
2- For each EDU_i , $EDU_j \in$ D, if $R_w \in$ Relation(D) and $R_w(EDU_j, EDU_i)$ and $R_w \notin$ Rel_RST, remove EDU_i and EDU_j from D.
3- For each removed EDU_i, $EDU_j \notin$D, if $R_w(EDU_i, EDU_j)$ and ($R_w(EDU_x, EDU_j)$ or $R_w(EDU_j, EDU_x)$), remove EDU_x from D.
4- For each EDU_i, $EDU_j \in$ D, if ($EDU_i \in$ Nuclei(D)), add EDU_i to Sum.

Let us illustrate the algorithm (A1) on a concrete example extracted from AD-RST. In Example 1, underlined words refer to the discourse markers. Figure 5.1 shows the tree structure of this example.

[1]تشتهر مدينة صفاقس بتقديم أطباق ثمار البحر على أنواعها. (1) عندما يرتاد زوار مدينة صفاقس،(2) فإنهم يطلبون باستمرار أطباق ثمار البحر (3) وخاصة طبق المحار والإخطبوط المشوي على الفحم. (4)

[2] Sfax city is famous for all kinds of seafood dishes. (1) If visitors come to Sfax city, (2) then they are constantly asking for seafood dishes (3) in particular dish oysters and octopus grilled over charcoal. (4)

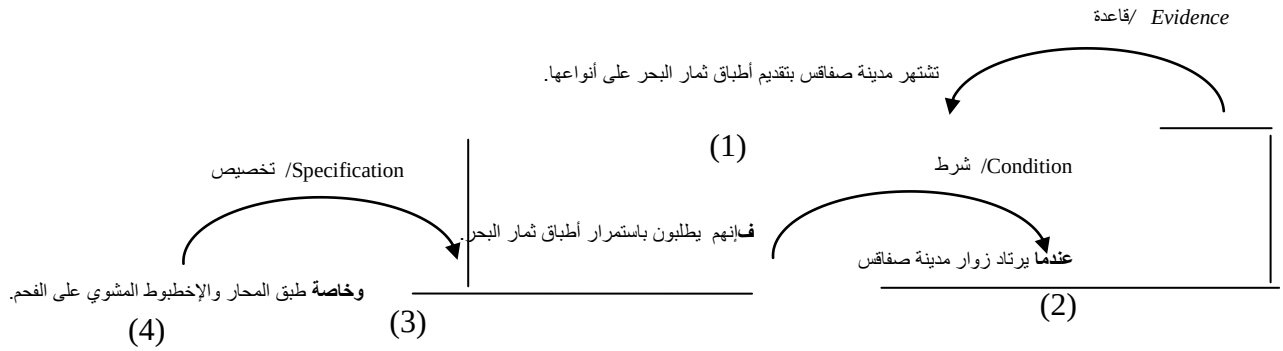

Figure 5.1. RST tree for Example 1.

When applying our algorithm, the relation تخصيص /txSyS/*Specification* will be removed, since it is non selected relation for summarization task. Then, we remove the satellite EDU_2.The final summary will contain the nuclei EDU_1 and the nuclei EDU_3, Sum= {EDU_1, EDU_3}.

3.2. Graph-based content selection algorithms

We propose two types of algorithms: (A2) "s*trict pruning*" that flatten CDUs by taking into account only their head (that is the first EDU) and (A3) "e*asy pruning*" that recursively apply the

same algorithm for each EDU in the CDU. Moreover, for each type of algorithms, we perform two types of pruning: one based on the distinction between coordination and subordinating relations (henceforth WithDistinction) and the other one does not take into account the nature of relations (henceforth WithoutDistinction). It is mandatory to note that for a given document, our algorithms are applied to each paragraph. The final summary is composed of the union of all the relevant EDUs extracted from each paragraph.

3.2.1. Strict pruning

This algorithm doesn't take into account the CDUs. If a document discourse structure contains CDUs, we perform a pre-treatment process that aims at flattening each CDU by selecting its head (the first EDU of the CDU) and removing its body (the other EDUs). Table 5.6 presents examples of all possible cases of pre-treatment:

Before pre-treatment	**After pre-treatment**
R(1,2)	R(1,2)
R([1-3],4)	R(1,4)
R(1,[2-4])	R(1,2)
R([1-3],[4-6])	R(1,4)

Table 5.6. Pre-treatment cases.

We note that this pre-treatment step automatically remove all relations that hold between the EDUs of the CDU body. For example, in case we have R1([1-3],4) and R2(1,[2,3]), R2 will be automatically removed from the discourse structure.

After pre-treatment, two main pruning strategies may be applied, as described below.

3.2.1.1. WithoutDistinction (A2.1)

In this strategy, we do not make any distinction between a subordinating and a coordinating relation. Only the discourse relation semantics is used. Let Rel_SDRT be the set of relevant SDRT discourse relations (cf. Table 5.2). The algorithm works as follows:

1- For each EDU_i, $EDU_j \in$ D, if $R_w(EDU_i, EDU_j)$ and $R_w \notin$ Rel_SDRT, remove EDU_i and EDU_j from D.
2- For each EDU_i, $EDU_j \in$ D, if $R_w(EDU_j, EDU_i)$ and $R_w \notin$ Rel_SDRT, remove EDU_i and EDU_j from D.
3- For each removed EDU_i, $EDU_j \notin$ D, if $R_w(EDU_i, EDU_j)$ and ($R_w(EDU_x, EDU_j)$ or $R_w(EDU_j, EDU_x)$), remove EDU_x from D.
4- For each EDU_i , $EDU_j \in$ D, add EDU_i to Sum.

Figure 5.2 presents a discourse structure where all the discourse relations are selected for the summarization task: R1, R2, and R3 $\in$ Rel_SDRT.

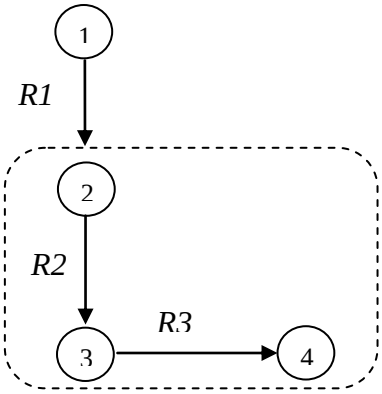

Figure 5.2. Example of a discourse structure.

When we apply this algorithm on the discourse structure presented in Figure 5.4, we obtain the relation R1(1,2) after the pretreatment step. After the pruning step, EDU_1 and EDU_2 are selected for the final summary, Sum= {EDU_1, EDU_2}.

3.2.1.2. WithDistinction (A2.2)

Unlike the algorithm (A2.1), this strategy takes into account the nature of discourse relations. Let Rel_SDRT_C the set of relevant coordinate SDRT discourse relations and let Rel_SDRT_S the set of relevant subordinate SDRT discourse relations (cf. Table 5.2). Content selection works as follows:

1- For each EDU_i , $EDU_j \in$ D, if $R_w(EDU_i, EDU_j)$ and $R_w \notin$ Rel_SDRT_C and $R_w \notin$ Rel_SDRT_S, remove EDU_i and EDU_j from D.
2- For each EDU_i , $EDU_j \in$ D, if $R_w(EDU_j, EDU_i)$ and $R_w \notin$ Rel_SDRT_C and $R_w \notin$ Rel_SDRT_S, remove EDU_i and EDU_j from D.
3- For each removed EDU_i, $EDU_j \notin$ D, if $R_w(EDU_i, EDU_j)$ and ($R_w(EDU_x, EDU_j)$ or $R_w(EDU_j, EDU_x)$), remove EDU_x from D.
4- For each EDU_i , $EDU_j \in$ D, if $R_w(EDU_i, EDU_j)$ and $R_w \in$ Rel_SDRT_C, add EDU_i and EDU_j to Sum.
5- For each EDU_i , $EDU_j \in$ D, if $R_w(EDU_j, EDU_i)$) and $R_w \in$ Rel_SDRT_C , add EDU_i and EDU_j to Sum.
6- For each EDU_i , $EDU_j \in$ D, if $R_w(EDU_i, EDU_j)$ and $R_w \in$ Rel_SDRT_S, add EDU_i to Sum.
7- For each EDU_i , $EDU_j \in$ D, if $R_w(EDU_j, EDU_i)$) and $R_w \in$ Rel_SDRT_S , add EDU_j to Sum.

When we apply this algorithm on the discourse structure presented in Figure 5.4, we obtain the relation R1(1,2) after pre-treatment step. After the pruning step, only EDU_1 is selected for the final summary, since R1 is a subordinate relation, Sum= {EDU_1}.

3.2.2. Easy pruning

Unlike (A2), the proposed algorithm (A3) takes into account the CDUs. There is thus any pre-treatment step since all the EDUs within a CDU are candidate for pruning. In short, this algorithm recursively apply the algorithm (A2) to all the CDUs in the graph.

As for (A2), two main pruning strategies may be applied, as explained below.

3.2.2.1. WithoutDistinction (A3.1)

This strategy only relies on the discourse relation semantics, as follows:

1- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_i, EDU_j)$ and $R_w \notin Rel\_SDRT$, remove EDU_i and EDU_j from D.
2- For each EDU_i , $EDU_j \in D$,if $R_w(EDU_j, EDU_i)$ and $R_w \notin Rel\_SDRT$, remove EDU_i and EDU_j from D.
3- For each removed EDU_i, $EDU_j \notin D$, if $R_w(EDU_i, EDU_j)$ and ($R_w(EDU_x, EDU_j)$ or $R_w(EDU_j, EDU_x)$) , remove EDU_x from D.
4- For each EDU_i , $EDU_j \in D$, add EDU_i to Sum.

When we apply this algorithm on the discourse structure presented in Figure 5.4, all EDUs are selected for the final summary since all the discourse relations are relevant for the summarization task i.e. Sum= {EDU_1, EDU_2, EDU_3, EDU_4}.

3.2.2.2. WithDistinction (A3.2)

In this strategy, selected EDUs must be the first argument of a relevant subordinating relation. It works as follows:

1- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_i, EDU_j)$ and $R_w \notin Rel\_SDRT\_C$ and $R_w \notin Rel\_SDRT\_S$, remove EDU_i and EDU_j from D.
2- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_j, EDU_i)$ and $R_w \notin Rel\_SDRT\_C$ and $R_w \notin Rel\_SDRT\_S$, remove EDU_i and EDU_j from D.
3- For each removed EDU_i, $EDU_j \notin D$, if $R_w(EDU_i, EDU_j)$ and ($R_w(EDU_x, EDU_j)$ or $R_w(EDU_j, EDU_x)$), remove EDU_x from D.
4- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_i, EDU_j)$ and $R_w \in Rel\_SDRT\_C$, add EDU_i and EDU_j to Sum.
5- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_j, EDU_i))$ and $R_w \in Rel\_SDRT\_C$, add EDU_i and EDU_j to Sum.
6- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_i, EDU_j)$ and $R_w \in Rel\_SDRT\_S$, add EDU_i to Sum.
7- For each EDU_i , $EDU_j \in D$, if $R_w(EDU_j, EDU_i))$ and $R_w \in Rel\_SDRT\_S$, add EDU_j to Sum.

When we apply this algorithm on the discourse structure presented in Figure 5.2, only EDU_1 is selected for the final summary because R1 is a subordinate relation. Figure 5.3 presents another example where all relations are selected for the summarization task. In this example, R1 and R2 are coordinating relations while R3 is subordinating.

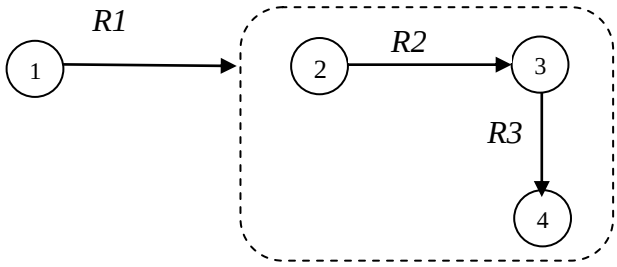

Figure 5.3. Example of a discourse structure.

Using (A3.2), the final summary contains the EDU_1, EDU_2, and EDU_3. EDU_4 is removed because it is the second argument of a subordinate relation.

4. Examples

4.1. Example from AD-RST corpus

We illustrate the algorithm (A1) proposed above on a concrete example. Example 2 is an annotated paragraph taken from the document ADC516. Table 5.7 presents the algorithm outputs.

(2)[ولعل اتفاق الدوحة مرآة النهج هذا وجوانبه المختلفة والمتصلة.] $_1$ [فهو نقل المعالجة،] $_2$ [أي التسكين والتعليق المؤقتين، من جامعة الدول العربية وأمينها العام "المصري"، من المال والسلاح] $_3$ [إلا من التكليف المعنوي، الى لجنة وزارية سباعية على رأسها قطر.]$_4$

*[wlEl AtfAq AldwHp mr|p Alnhj h*A wjwAnbh Almxtlfp wAlmtSlp.]$_1$ [fhw nql AlmEAljp,]$_2$ [>y Altskyn wAltElyq Almwqtyn, mn jAmEp Aldwl AlErbyp w>mynhA AlEAm "AlmSry", mn AlmAl wAlslAH]$_3$ [<lA mn Altklyf AlmEnwy, AlY ljnp wzAryp sbAEyp ElY r>shA qTr.]$_4$*

[Perhaps the Doha's agreement reflects this approach and its different related aspects.] $_1$ [So, it is the transfer processing,] $_2$ [that means temporary pacification and stopping, from the League of Arab States and its "Egyptian" secretary-general, of money and arms] $_3$ [except, moral assignment, to a heptagonal ministerial committee headed by Qatar.]$_4$

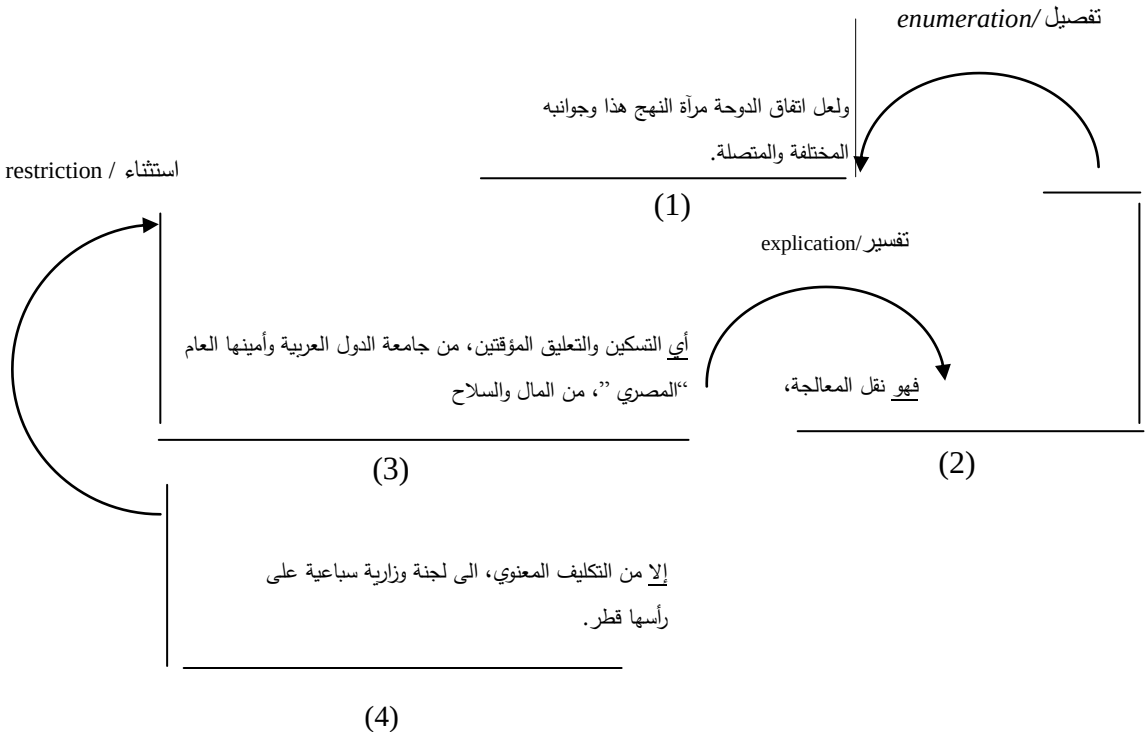

Figure 5.4. The discourse annotation for Example 2.

Algorithm	**Selected EDUs**
A1	EDU_1

Table 5.7. Algorithm outputs.

4.2. Example from ADTB corpus

We illustrate the algorithms (A2) and (A3) proposed above on a concrete example. Example 3 presents an annotated paragraph which is taken from the document ANN20020115.0003 of the ADTB corpus. Table 5.8 presents the algorithms outputs.

(3)[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 kilometers Southwest of Khost.]$_9$ [And it said:]$_{10}$ ["the bombing lasted 48 hours."]$_{11}$

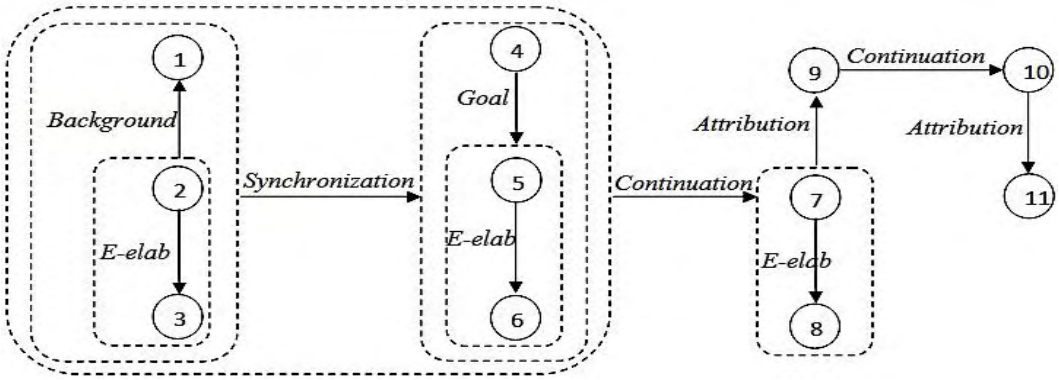

Figure 5.5. The discourse annotation for Example 4.

Table 5.8 presents the selected EDUs for each proposed algorithms.

Algorithms	**Selected EDUs**
(A2.1)	EDU_1, EDU_4, EDU_7, EDU_9, EDU_{10}, and EDU_{11}
(A2.2)	EDU_1 and EDU_7
(A3.1)	EDU_1, EDU_2, EDU_4, EDU_5, EDU_7, EDU_9, EDU_{10}, and EDU_{11}
(A3.2)	EDU_1, EDU_4, and EDU_7

Table 5.8. Algorithms outputs.

5. Experiments and results

The proposed five algorithms have been implemented and evaluated on ADTB and AD-RST gold standard summaries, (cf. Section 2). In each corpus, we compare the performance of the automatic content selection against two baselines: (B1) that selects the first two EDUs of each paragraph and (B2) that selects the first EDU from the first two sentences of each paragraph.

There are two ways to evaluate a summary: the evaluation of the summary content (using precision, recall, and F-measure) and the evaluation of the *linguistic quality* of summary (using

ROUGE, Recall-Oriented Understudy for Gisting Evaluation which are based on the similarity of n-grams (Lin and Hovy, 2003) and Pyramid, is a semi-automatic evaluation method (Nenkova and Passonneau, 2005)). There are several aspects of summary linguistic quality, we can cite: (1) Grammaticality, the summary should not contain non textual items (i.e., markers). (2) Non redundancy, the summary should not contain redundant information. (3) Reference clarity, the anaphora should be clearly referred to nouns and pronouns in the summary. (4) Coherence and structure, the summary should have good structure and the sentences should be coherent. In our case, we aim to evaluate the summary content to know the ability of our algorithms to select the relevant segments. Table 5.9 reports the results of the two baselines on each corpus in terms of precision, recall and F-measure.

Corpus	Baseline	Precision	Recall	F-measure
ADTB	(B1)	0.377	0.387	0.382
	(B2)	0.419	0.457	0.437
AD-RST	(B1)	0.485	0.309	0.377
	(B2)	0.495	0.352	0.411

Table 5.9. The baseline results.

As shown in Table 5.9, (B2) yields better results compared to (B1) on both corpora. This can be justified by the fact that annotators rarely chose two adjacent segments when they manually generate summaries. Overall, the results on ADTB are better compared to AD-RST for two reasons. First, segmentation principles in AD-RST are mainly based on explicit discourse markers. EDU are thus globally longer in AD-RST than in ADTB, which makes the EDU selection process more difficult. Second, there are no embedded EDUs in AD-RST. Consequently, segments may contain a lot of non pertinent information compared to ADTB. For example, the EDU_3 in Example 2 will be segmented within the framework of SDRT into two embedded EDUs, as illustrated in Example 4.

(4) [أي التسكين والتعليق المؤقتين، [من جامعة الدول العربية وأمينها العام" المصري "،] $_2$ من المال والسلاح]$_1$

[>y Altskyn wAltElyq Almwqtyn, [mn jAmEp Aldwl AlErbyp w>mynhA AlEAm "AlmSry",] $_2$ mn AlmAl wAlslAH] $_1$

[that means temporary pacification and stopping, [from the League of Arab States and its "Egyptian" secretary-general,] $_2$ of money and arms] $_1$

We then evaluate the performances of the tree-based content selection algorithm (A1) and the graph-based content selection algorithms ((A2.1), (A2.2), (A3.1), and (A3.2)) by conducting two evaluations settings. The first one evaluates the algorithms when inputs are gold standard discourse structure while the second takes as input automatically parsed documents. In this last setting, automatic parsing consists on automatic discourse relation labeling (henceforth partial discourse parser, as described in Chapter 4) relying on gold standard segmentations and gold

standard attachments. Given that the evaluation is based on EDU selection (i.e. checking whether an EDU selected by the annotators is also selected by our algorithms), we must have the same discourse units to compare the two final summaries (one generated automatically and the other one manually generated). For this reason, we use only the automatic discourse relation-labeling step of the parser. We use the partial RST parser described in Keskes et al. (2012d) (cf. Section 1.2). Similarly to the proposed parser using the SDRT framework (cf. Chapter 3 and Chapter 4), the RST parser does not treat the attachment problem task. Table 5.10 presents the results. Best performances are marked in boldface.

	Using manually annotated discourse structure			**Using automatic discourse structure**		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
A1	0.711	0.536	0.611	0.596	0.470	0.525
A2.1	0.501	0.396	0.442	0.482	0.378	0.424
A2.2	0.660	0.344	0.452	0.503	0.351	0.413
A3.1	0.625	0.698	0.659	0.544	**0.573**	0.558
A3.2	**0.742**	**0.707**	**0.724**	**0.688**	0.537	**0.603**

Table 5.10. The results of the proposed algorithms.

On the first hand, all proposed algorithms outperform the two baselines using manually annotated discourse structures and using the discourse parser. We first conclude that EDU position is not enough for content selection task. Moreover, compared to the strict pruning algorithm (that flattens CDUs and takes into account just the head of each CDU for content selection task), the easy pruning algorithm obtained better performances which show that the discourse structure information are more sensitive to the content selection task (+0.217 of F-measure using the gold corpus and +0.134 of F-measure using the partial parser output, without distinguishing between the nature of discourse relation). We then conclude that the discourse structure is needed for content selection task. Again, best results are obtained when we take into account the nature of discourse relation. For example, (A3.2) improves the F-measure by +0.065 using the gold standard discourse annotation corpus and by +0.045 using the partial parser output). We finally conclude that the nature of discourse relation (coordinate/subordinate) is important information for content selection since it can help to select the most relevant EDUs.

On the other hand, the use of the partial discourse parser stills more challenging. The results of the strict pruning algorithm decreased slightly (-0.018 for (A2.1) and -0.039 for (A2.2) in terms of F-measure) and the results of the easy pruning algorithm decreased significantly (-0.101 for (A3.1) and -0.121 for (A3.2) in terms of F-measure). This difference can be explained by the fact that the strict pruning algorithm does not use the full discourse structure (i.e. it does not takes into account the CDUs) since the easy pruning algorithm treats the full discourse structure. However, the use of the partial discourse parser in (A3.2) is more appropriate than the use of the gold corpus in (A2.2). This fact permits to confirm the efficiency of using such discourse parser to reach promising results.

Finally, our results confirm that both discourse structure and the nature of discourse relation have a positive impact on content selection. However, there is no indication that allows us to automatically compare the algorithm applied on AD-RST with the algorithms applied on ADTB. In other words, we cannot conclude which discourse structure (tree or graph) is more suitable for the content selection task. Moreover, given that we use two different discourse structures, we are not able to use a unique parser for both corpora. To better compare our algorithms, we asked the annotators to manually compare the quality of the summary generated by the algorithm (A1) and the algorithm (A3.2) when applied on the partial parser output. After this comparison, annotators observed that the summaries produced by (A1) and (A3.2) have almost similar quality. However, they observe that the best summary quality is provided by the Algorithm A3.2 that uses a graph as discourse structure (SDRT framework). The annotators justified their decision by four main reasons:

- The semantic of discourse relations (i.e. the list of relations that are deemed to be relevant for the summarization task) used in the two frameworks has no impact on the summary quality.
- The discourse relation nature (coordinate/subordinate) in the SDRT framework has its equivalent in the RST framework (nucleus/satellite). Hence, this notion has no impact on the summary quality.
- The notion of CDU in ADTB helps to provide a Non redundancy summary. CDU tends to group information (idea, events, etc.) by themes or topics.
- In some cases, embedded segments in ADTB help to select the part of sentence that contains just the relevant information. In fact, the selected segments for summary in ADTB contain less secondary information compared to the selected segments in AD-RST. Example 5 presents one segment from AD-RST. If we apply the discourse segmentation principles according to the SDRT framework, we obtain two EDUs. EDU_2 doesn't contain relevant information. Therefore, when we use SDRT framework, only EDU_1 will be selected for summary.

(5)[و في كتاب التكليف [الذي وجهه إلى الحكومة الجديدة ،]$_2$ تمت اتخاذ كل الترتيبات والاستعداد الكامل.] $_1$

*[w fy ktAb Altklyf [Al*y wjhh AlY AlHkwmp Aljdydp,]* $_2$ *tmt AtxA* kl AltrtybAt wAlAstEdAd AlkAml.]* $_1$

[In the book of reference [which has been sent to the new government,] $_2$ *all the arrangements have been taken.]* $_1$

Conclusion

In this chapter, we proposed an automatic Arabic text summarization based on discourse information. We used the semantic of the discourse relations and the discourse structure to extract the most important Elementary Discourse Units (EDUs) in the text. The selected EDUs for summary must have the main information, event, object, ideas, etc. in text.

To achieve this purpose, we have proposed five algorithms according to several discourse criteria (coordinate/subordinate relations, Complex Discourse Units (CDUs), discourse structures, etc.). We evaluated these algorithms using two different corpora that have been annotated according to two different frameworks: ADTB (cf. Chapter 2) annotated following the Segmented Discourse Representation Theory (SDRT) and the Arabic Discourse RST corpus (AD-RST) annotated following the Rhetorical Structure Theory (RST). In addition, we evaluated the difference between using discourse graphs and discourse trees based on annotator judgments. Our results show that discourse information is important for content selection. When comparing the quality of the produced summary, our results demonstrate that the best summary is the one produced when the discourse structure is a graph (thanks to the embedded segments and the notion of CDU).

As future work, we plan to investigate the performances of our partial discourse parser to improve the results of other NLP applications (e.g. generation systems, translation systems, Question/Answering systems, etc.). Therefore, we tend to extend some work done by our research group; mainly we plan to add our discourse parser as a module to the Arabic Question/Answering system (Trigi *et al.*, 2014).

General conclusion

In this dissertation, we proposed a semantically-driven approach to analyze Arabic discourse (Modern Standard Arabic (MSA)), following the SDRT framework. This discourse analysis fully addresses the discourse segmentation using both explicit and implicit discourse connectives and the discourse annotation of explicit and implicit Arabic discourse relations. Discourse relations permit to link adjacent as well as non adjacent units within the Segmented Discourse Representation Theory framework. Additionally, we built a Arabic Discourse Treebank corpus (ADTB), assessed the reliability of the framework on this corpus, and applied our discourse analysis on a practical application aiming to select the most relevant information in a text.

We started our dissertation by a background and an overview of the state of the art concerning discourse analysis in different languages. Then, we proposed a manual of Arabic discourse annotation. Herein, we described main discourse segmentation principles, listed the Arabic discourse relations, the hierarchy of Arabic discourse relations, the Arabic discourse connectives, and we defined the discourse attachment principles.

Discourse relations are organized around 4 top-level classes with a total of 24 relations. The annotation manual is used by annotators in order to build the gold standard ADTB, which presents the first resource that identifies the interactions between the semantic content of Elementary Discourse Units and the global pragmatic structure of the discourse. ADTB is composed of 70 documents extracted from the syntactically annotated Arabic Treebank (v3.2 part 3) where each document is represented by an oriented acyclic graph that provides a recursive and a complete discourse structure of the document. In addition, we built a discourse lexicon which contains 174 discourse connectives used to explicitly express discourse relations between discourse parts, to contribute to discourse coherence, and to mark discourse structure. The results of the annotation campaign show that full discourse annotation is feasible for Arabic where a good inter-annotator agreement has been reached.

After building ADTB, we performed a multi-class supervised learning approach that predicts EDUs and embedded EDUs boundaries. The approach uses our rich lexicon and relies on a combination of punctuation, morphological and lexical features. The evaluation results showed that extensive morphological features are more suitable than shallow morphological analysis since best scores were obtained when adding information of the root, the prefix and the suffix. Moreover, we have shown that Arabic discourse segmentation is feasible without any use of shallow syntactic information (chunks). Finally, we fully addressed the recognition of EDU frontiers even in case of lack of discourse markers (that is, in case of implicit discourse relations), which represents 25% of cases in our data. This task is the first step to build a partial Arabic discourse parser. As a second step, we built a multi-class supervised learning approach that

predicts both explicit and implicit Arabic discourse relations between EDUs in Arabic texts. To accomplish this task, we relied on a combination of lexical, morphological, syntactic and lexico-semantic features. We compared our approach to three baselines that are based on the most frequent relations, discourse connectives and the features used by (Al-Saif and Markert, 2011). Our experimental results are promising since we outperform all the baselines. However, attachment level has not been resolved. This complex task needs more resources and more annotated documents as used in discourse relation recognition task.

Finally, we proposed an automatic Arabic text summarization tool based on discourse information to show the positive impact of the partial discourse parser in NLP applications. Indeed, we used the semantic of the discourse relations and the discourse structure to extract the most important EDUs in the text. This tool is useful to measure the adequacy of the text according to the information requested by the user. For this purpose, we have implemented five algorithms corresponding to the defined discourse criteria (discourse segmentation, coordinate/subordinate relations, Complex Discourse Units (CDUs), and discourse structure). Afterwards, we evaluated these algorithms using two different corpora that have two different frameworks: ADTB and the Arabic Discourse RST corpus (AD-RST) (100 texts selected from the journal "Dar Al Hayat") (Keskes *et al*., 2012d), annotated according to the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). Furthermore, we evaluated the difference between using discourse graph and discourse tree based on annotator judgments. Annotators reported that all discourse information are useful for the content selection task and in turn improve the results of the automatic Arabic text summarization. However, a slightly best summary quality in terms of "selected EDUs contain only the relevant information" and "redundancy" using the SDRT framework thanks to the embedded segments and the CDU notion.

The future work of this dissertation can be regrouped in three main categories: theoretical future work, technical future work, and applicative future work.

- As theoretical future work, we intend first to handle long distance dependencies that exceed the paragraph boundaries. In other words, we will try to annotate Arabic discourse relations that link EDUs located in different paragraphs. Then, we plan to investigate the performances of our Arabic discourse framework that use SDRT to study other types of corpora. Given that our research group carried out many studies on Tunisian dialect (Graja *et al*., 2013; Karoui *et al*., 2013; Zribi *et al*., 2013), we intend to propose further improvements for the Arabic discourse framework by building an annotated discourse corpus for Tunisian dialect texts. We first aim at tackling manually and automatically discourse segmentation of the Tunisian dialect corpus. Then, we tend to handle a set of discourse connectives for this dialect and update our hierarchy of the discourse relations to take into account the Tunisian dialect specificities. The final goal is to build a discourse parser for the Tunisian dialect.

- As technical future work, we tend to annotate the whole ATB corpus (600 newspapers) according to the SDRT framework using a semi-supervised approach. Given a large corpus annotated with discourse information, we can tackle the attachment problems and develop a full Arabic discourse parser.
- As applicative future work, we plan to investigate the performances of our discourse parser to improve the results of other NLP applications. More precisely, we plan to exploit our discourse parser in the context of the DefArabicQA system (an Arabic definition question answering system that aims at dealing with the results returned by Web search engines to return the appropriate information to a user question) deplopped in our research group (Trigi et *al.*, 2014). The idea is to add our discourse parser as module to the DefArabicQA system in order to improve the selection process of the relevant answers returned by the system.

References

Abdul-Mageed M. and Diab M. (2012). AWATIF: A multi-genre corpus for Modern Standard Arabic subjectivity and sentiment analysis, In Proceedings of LREC, Istanbul, Turkey, 2012.

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

Abdulmuttalib H.M. (2003). Al-Nahu Al-Muiassar, Dar Al-Aafag Al-Arabiah.

Aboaoga M. and Ab-Aziz MJ. (2013). Arabic person names recognition by using a rule based approach. Journal of Computer Science, ISSN: 1549-3636, 9 (7): 922-927, 2013.

Abouenour L., Bouzoubaa K. and Rosso P. (2012). Idraaq: New arabic question answering system based on query expansion and passage retrieval. In: CLEF (Online Working Notes/Labs/Workshop)

Abubakre R.D. (1989). Bayan in Arabic Rhetoric: An analysis of the core of Balāgha. Ibadan: Intec Printer Limited.

Abu-Jbara A., King B., Diab M. and Radev D. (2013). Identifying Opinion Subgroups in Arabic Online, Discussions, The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers (ACL Short Papers 2013) Soifa, Bulgaria, August 4-9, 2013.

Abuobieda A., Salim N., Albaham A. T., Osman A. H. and Kumar Y. J. (2012), Text summarization features selection method using pseudo genetic-based model, International conference on information retrieval knowledge management, Kuala Lumpur, pp. 193–197.

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

Afantenos S., Asher N., Benamara F., Bras M., Fabre C., Ho-Dac M., Draoulec A. L., Muller P., Pery-Woodley M.-P., Prevot L., Rebeyrolles J., Tanguy L., Vergez-Couret M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC 2012.

AlAnsari I. H. (1985). Mogny Alabib En Kutb AlAEareb. Lebnan: Dar Alfekur.

AlAnsari I.H. (2003). Mugni Al-Labeeb An Kutub Al-Aareeb, Al-Maktabah Al-Asriah for publishing and printing.

Alfarabi H. (1990). Ketab Alhroof. Dar Almashreg, Lebnan.

Alhuqbani M.N. (2013). The English But and Its Equivalent in Standard Arabic: Universality vs. Locality. Theory and Practice in Language Studies, Vol. 3, No. 12, pp. 2157-2168, December 2013. ISSN 1799-2591. doi:10.4304/tpls.3.12.2157-2168.

Alguliev R. M. and Aliguliyev R. M. (2005). Effective summarization method of text documents, Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), France, pp. 264–271, 19–22 September 2005.

Alguliev R. M. and Aliguliyev R. M., (2008). Automatic text documents summarization through sentences clustering, Journal of Automation and Information Sciences, Vol. 40, No. 9, pp. 53–63, 2008.

Ali Mohammed M. and Omar N. 2011. Rule Based Shallow Parser for Arabic Language. Journal of Computer Science, (10): 1505-1514, 2011, ISSN 1549-3636.

Aliguliyev R. M. (2006). A novel partitioning-based clustering method and generic document summarization, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'06 Workshops) (WI-IATW'06), Hong Kong, China, pp. 626–629, 18–22 December 2006.

Aliguliyev R. M. (2009). Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization, Computational Intelligence. Volume 26, Number 4, 2010. DOI: 10.1111/j.1467-8640.2010.00365.x.

Al-Jarim A. and Amine M. (1999). البلاغة الواضحة/Al-Balagha al-Wadiha. Editor: Dar Al-maaref. ISBN: 977-02-5784-2.

Alrahabi M. (2010). Excom-2 : plateforme d'annotation automatique de catégories sémantiques : conception, modélisation et réalisation informatique : applications à la catégorisation des citations en arabe et en français. Thèse de doctorat en Informatique Linguistique.

Alrahabi M. and Desclés J.-P. (2009). EXCOM : Plate-forme d'annotation sémantique de textes multilingues. TALN 2009, Senlis, 24-26 juin 2009.

Alrahabi M., Mourad G. and Djioua B. (2004). Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique. Le traitement automatique de l'arabe, JEP-TALN 2004, Fès, 19-22 avril 2004.

Alrahabi M. (2010). EXCOM-2: plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe. PhD thesis. Paris-Sorbonne University.

Aloulou C. (2005). Un modèle multi-agent pour l'analyse syntaxique de la langue arabe, Thèse de doctorat en Informatique, Ecole Nationale des Sciences de l'Informatique de Tunis, Juin 2005.

Al-Saif A. and Markert K. (2010). The Leeds Arabic Discourse Treebank: Annotating Dis-course Connectives for Arabic, In Proceedings of the International Conference on Language Resources and Evaluation, (LREC 2010), Valletta, Malta.

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

Al-Sanie W., Touir A. and Mathkour H. (2008). Parsing Arabic texts using rhetorical Structure Theory text. Journal of Computer Science, Science Publication, Vol. 4, No. 9, 2008, pp. 713-720.

Al-Sughaiyer I.A. and Al-Kharashi I.A. (2004). Arabic morphological analysis techniques: A comprehensive survey. Journal of Americal Society of Information Science Technology 55(3), 189-213.

Aone C., Okurowski M. and Gorlinsky J. (1998). Trainable scalable summarization using robust nlp and machine learning. In Proceedings of the 17th COLING and 36th ACL.

Amini M., and Gallinari P. (2003). Semi-supervised learning with an explicit label-error model for misclassified data. IJCAI2003.

Amini M. and Usunier N. (2007). A Contextual Query Expansion Approach by Term Clustering for Robust Text Summarization. In Proceedings of the 7th Document Understanding Conference, pages 48–55, Rochester, USA, 2007. DUC.

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

Asher N. and Lascarides A. (2003). Logics of Conversation. Cambridge University Press.

Aubadah M.I. (1983). Al-Jumlah Al-Arabiah, Munshaat Al-Ma'aref.

Azmi A.M. and Al-Thanyyan S. (2012). A text summarizer for Arabic. Comput. Speech Lang., 26 (4) (2012), pp. 260–273.

Bahou Y. (2012). Automatic comprehension of spontanious Arabic speech : Integration in an interactif voval server/Compréhension automatique de la parole arabe spontanée : Intégration dans un Serveur Vocal Interactif. Defended on Marsh 15th 2012 at FSEGS, Sfax

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

Baldridge J. and Lascarides A. (2005). Probabilistic Head-Driven Parsing for Discourse Structure, In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNNL), Ann Arbor, 2005.

Barzilay R. and McKeown K. (2005). Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3):297– 328.Basha, A. Z. (1912).الترقيم وعلاماته في اللغة العربية. (Punctuation and its marks in Arabic Language).

Banu M., Karthika C., Sudarmani P. and Geetha T.V. (2007).Tamil Document Summarization Using Semantic Graph Method, International Conference on Computational Intelligence and Multimedia Applications, IEEE, pp. 128-134, 2007.

Belguith Hadrich L. (1999). Traitement des erreurs d'accord de l'arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritère pour la correction, Thèse de doctorat en Informatique, Faculté des Sciences de Tunis, Février 1999.

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

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

Belguith Hadrich L., Baccour L. and Mourad G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. 12th Conference on Natural Language Processing (TALN'2005), Dourdan.

Belguith Hadrich L., Ellouze M., Maâloul M. H., Jaoua M., Kallel J. F. and Blache P. (2014). Automatic summarization in Natural Language Processing for Semitic Languages, Series: Theory and application of Natural Language Processing. (Imed Zitouni editor), Springer 2014. pp 371-403, ISBN-10: 3642453570 ISBN-13: 978-3642453571.

Benajiba Y. Rosso P. Abouenour L. Trigui O. Bouzoubaa K. and Belguith, HL. (2012). Question Answering for Semitic Languages' in the book. in 'Natural Language Processing Approaches to Semitic Languages' edited by Pr. Imed Zitouni and published by Springer, 2012.

Benajiba Y., Rosso P. and Benedi J. M. (2007). ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. CICLing, Springer-Verlag, Berlin, Heidelberg, pp. 143-153.

Benveniste E. (1971). Problems in General Linguistics, University of Miami Press, Florida (first published 1966).

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

Binwahlan M.S., Salim N. and Suanmali L. (2010). Fuzzy swarm diversity hybrid model for text summarization Information Processing and Management, 46 (2010), pp. 571–588

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

Blakemore D. (1992). Understanding utterances. Oxford: Blackwell.

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

Blais A., Desclés J.-P., Djioua B. (2006). Le résumé automatique dans la plate-forme EXCOM, Digital Humanities, Paris (2006).

Bosma W. (2005). Query-Based Summarization using Rhetorical Structure Theory. LOT Occasional Series, volume 4, pp. 29 – 44.

Boudabous M. M., Chaâben N., Khedher N., Belguith Hadrich L. and 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. and Bebah M. (2011). Alkhalil morpho sys: A morphosyntactic analysis system for Arabic texts.

Boujelben I. Jamoussi S. and Ben Hamadou A. (2013). Enhancing machine learning results for se-mantic relation extraction, NLDB, Manchester, UK, pp337 – 342, 2013.

Borisova I. and Redeker G. (2010). Same and Elaboration relations in the Discourse Graphbank. In Proceedings of the 11th annual SIGdial Meeting on Discourse and Dialogue, Tokyo, 2010.

Buch-Kromann M. and Korzen I. (2010). The unified annotation of syntax and discourse in theCopenhagen Dependency Treebanks, In Proceedings of the Fourth Linguistic Annotation Workshop, pages 127–131, July.

Buch-Kromann M., Korzen I. and Muller H. H. (2009). Uncovering the 'lost' structure of ¨ translations with parallel treebanks. In Fabio Alves, Susanne Gopferich, and Inger Mees, editors, Copenhagen Studies of Language: Methodology, Technology and Innovation in Translation Process Research, Copenhagen Studies of Language, vol. 38, pages 199– 224. Copenhagen Business School.

Canasai K. and Chuleerat J. (2003). Generic Text Summarization Using Local and Global Properties of Sentences, Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03) , 2003.

Cantarino V. (1975). Syntax of Modern Arabic prose. Bloomington/London: Indiana University Press.

Carenini G. and Cheung J. K. (2008). Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. Fifth INLG 08, 8 pages, Salt Fork, OH.

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

Carpuat M. Marton Y. and Habash N. (2012). Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment. Machine Translation 26(1-2): 105-120 (2012).

Chaalal I. (2010). Foreign Learners' Difficulties in Translating the Arabic Discourse Marker 'Fa' into English The Case of Third Year Students of Translation at the University of Constantine. Applied Language Studies.

Chai Y., and Jin R. (2004). Discourse structure for context question answering. In Proceeding of the HLT-NAACL 2004 Workshop on Pragmatics in Question Answering, Boston, MA.

Chalabi A. (2001). Sakhr Web-based Arabic<>English MT engine, ACL/EACL 2001 Workshop on Arabic Language Processing, Toulouse July 2001.

Chardon B., Benamara F., Popescu V., Mathieu Y., and Asher N. (2013). Measuring the Effect of Discourse Structure on Sentiment Analysis. In Proceedings Computational linguistics and intelligent text processing 14th International Conference, CICLing 2013, Samos, Greece.

Chareonsuk J., Sukvakree T. and Kawtrakul A. (2005). Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information, NCSEC, 2005.

Chatterjee N. and Mohan S. (2007). Extraction-Based Single-Document Summarization Using Random Indexing. 19th IEEE International Conference on Tools with Artificial Intelligence

Cheng K., Yanting L. and Wang X. (2013). Single Document Summarization based on Triangle Analysis of Dependency Graphs. 16th International Conference on Network-Based Information Systems. 978-0-7695-5052-7/13 $26.00 © 2013 IEEE DOI 10.1109/NBiS.2013.9

Cheung J. and Penn G. (2013). Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1233–1242, August.

Cheung J. and Penn G. (2014). Unsupervised Sentence Enhancement for Automatic Summarization Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 775–786, October 25-29, 2014, Doha, Qatar.

Cristal D. (1987). The Cambridge Encyclopaedia of Language, Cambridge University Press, Cambridge.

Cobos C., Montealegre C., Mejía M., Mendoza M. and León E. (2010). Web document clustering based on a new niching memetic algorithm, term-document matrix and Bayesian information criterion. Proceedings of the IEEE congress on evolutionary computation (IEEE CEC), IEEE, Barcelona, Spain (2010), pp. 4629–4636.

Conroy J., Schlesinger J., Goldstein J. and O'Leary D. (2004). Left-brain/right-brain multi-document summarization. In DUC 2004: Document Understanding Workshop, May 6–7, 2004, Boston, MA, USA.

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

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

Danlos L. and Gaiffe B. (2004). Event coference and discourse relations in L. Kulda (éd), Language, Music and Cognition, Kluwer Academic Publishers, Amsterdam.

Danlos L., Antolinos-Basso D., Braud C. and Roze C. (2012). Vers le FDTB : French Discourse Tree Bank. TALN 2012 : 19ème conférence sur le Traitement Automatique des Langues Naturelles, Grenoble : France.

Darwish K. (2013). Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, August 4-9, 2013.

Daumé III H. and Marcu D. (2005). Bayesian multidocument summarization at mse. In Proceedings of MSE.

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

Dehkordi P.-K., Kumarci F. and Khosravi, H. (2009). Text summarization based on genetic programming. In Proceedings of the international journal of computing and ICT research (Vol. 3, pp. 57–64).

Diab M. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In 2nd International Conference on Arabic Language Resources and Tools.

Diab M. Hacioglu K. and Jurafsky D. (2007). Arabic Computational Morphology: Knowledge-based and Empirical Methods, chapter Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking. Kluwer/springer edition, 2007.

Diab M., Moschitti A. and Pighin D. (2008). Semantic Role Labeling Systems for Arabic using Kernel Methods. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2008).

Douzidia F. S. and Lapalme G. (2004). Lakhas, an Arabic summarization system, Proceedings of DUC2004, 2004.

Dunlavy D.M., O'Leary D.P., Conroy J.M. and Schlesinger J.D. (2007). QCS: A system for querying, clustering and summarizing documents Information Processing & Management, 43 (2007), pp. 1588–1605

DuVerle D. A. and Prendinger H. (2009). A novel discourse parser based on support vector machine classificaion. Proceedings of ACL, 2009.

Elarnaoty M., AbdelRahman S. and Fahmy A. (2012). A machine learning approach for opinion holder extraction in Arabic, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

El-Haj M. and Hammo B. (2008). Evaluation of query-based Arabic text summarization system. In: Proceeding of the IEEE International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2008, pp. 1–7. IEEE Computer Society, Beijing (2008).

El-Haj M., Kruschwitz U. and Fox C. (2011). Experimenting with automatic text summarisation for Arabic. In: Z. Vetulani (Ed.), Human Language Technology. Challenges for Computer Science and Linguistics, Springer, Berlin Heidelberg, pp. 490–499.

Ellouze M. (2004). Des schémas rhétoriques pour le contrôle de la cohérence et génération de résumés automatiques d'articles scientifiques. PhD thesis, Université de Manouba, Ecole Nationale des sciences de l'Informatique, 2004.

Elsner M. and Santhanam D. (2011). Learning to fuse disparate sentences. In Proceedings of the Workshop on Monolingual Text-To-Text Generation, pages 54–63. Association for Computational Linguistics.

Emara S.A. (2014). The Functions of 'or' and 'aw': Implications for Translation. International Journal of Linguistics ISSN 1948-5425 2014, Vol. 6, No. 5. doi:10.5296/ijl.v6i5.5961.

Erkan G. and Radev D. R. (2004). Lexrank: Graph-based centrality as salience in text summarization. Journal of Artificial Intelligence Research (JAIR).

Eskander R., Habash N., Bies A., Kulick S. and Maamouri M. (2013). Automatic Correction and Extension of Morphological Annotations. ACL 2013: 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Sofia, August 4-9.

Fareh S. and Hamdan J. (1999). The translation of arabic `wa' into english: Some problems and implications. Dirasat, Human and Social Sciences.

Farghaly A. and Senellart, J. (2003). Intuitive coding of the arabic lexicon. In: Proceedings of the MT Summit IX, the Association for Machine Translation in the Americas, AMTA'03

Fellbaum C. (1998). WordNet: An Electronic Lexical Database. MIT Press, 1998.

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

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

Foucault M. (1972). The Archaeology of Knowledge, trans. Sheridan Smith, A. M., Tavistock, London (first published 1969).

Fraser B. (1988). Types of English discourse markers. Acta Linguistica Hungarica 38(1-4): 19-33.

Gabawah F. (1972). Iraab Al-Jumal wa Ashbah Al-Jumal, Dar Al-Qalam Al-Arabi.

Geurts B. (1999). Presuppositions and Pronouns. Elsevier Science.

Geurts B. and van der S. (1999). Presuppositions and backgrounds. In Proceedings of the 11th Amsterdam Colloquium. University of Amsterdam.

Giannakopoulos G., El-Haj M., Favre B., Litvak M., Steinberger J., Varma V. (2011). TAC 2011 MultiLing pilot overview. Proceedings of the Text Analysis Conference (TAC).

Graja M., Jaoua M. and Belguith Hadrich L. (2013). Discriminative Framework for Spoken Tunisian Dialect Understanding. International Conference on Statistical Language and Speech Processing (SLSP 2013), Tarragona Spain, July 29-31, 2013.

Green S. and Manning C. (2010). Better Arabic Parsing: Baselines, Evaluations, and Analysis. In COLING 2010.

Gridach M. and Chenfour N. (2011). Developing a New System for Arabic Morphological Analysis and Generation. Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 52–57, Chiang Mai, Thailand, November 8, 2011.

Grosz B.J., Joshi A.K. and Weinstein, S. (1995). Centering: a framework for modelling the local coherence of discourse. Computational Linguistics 21(2) p. 203-225.

Gupta V. and Lehal G. (2010). A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence, Vol. 2, NO. 3, August 2010.

Habash N. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies, Graeme Hirst, editor. Morgan & Claypool Publishers. 187 pages.

Habash N., Owen R. and Ryan R. (2009). MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Haghighi A. and Vanderwende L. (2009). Exploring content models for multi-document summarization. In Proceedings of HLT-NAACL.

Hahn U. and Mani I. (2000). The challenges of automatic summarization. IEEE Computer, 33(11): 29-36.

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

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

Hardmeier C. (2012). Discourse in Statistical Machine Translation. Discours 11 | 2012, mis en ligne le 23 décembre 2012. URL : http://discours.revues.org/8726 ; DOI : 10.4000/discours.8726.

Hassan I., Mathkour A. and Waleed A. (2008). Parsing Arabic Texts Using Rhetorical Structure Theory. In Journal of Computer Science 4 (9): 713-720.

He T., Shao W., Li F., Yang Z. and Ma L. (2008). The automated estimation of contentterms for query-focused multi-document summarization. In Proceedings of the 2008 fifth international conference on fuzzy systems and knowledge discovery (FSKD 2008), October 18–20, Jinan, China, vol. 5 (pp.580–584).

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

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

Hemeida M. (1997). Nedhum Al Ertebat wa AlRabt Tarkeeb AlGomla Al Arabeia (Published in Arabic). The Egyptian International Company for Publishing (Longman), Egypt.

Hobbs J. (1979). Coherence and coreference. Cognitive Science (3) 8, 67-90.

Hobbs J. (1985). On the coherence and structure of discourse. Technical Report 85-37, Center for the Study of Language and Information (CSLI), Stanford, CA.

Huang H. and Chen H. (2011). Chinese Discourse Relation Recognition In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP). Pages 1442-1446. Chiang Mai, Thailand. November 2011.

Huang H. and 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), pages 261–269, Seoul, South Korea, 5-6 July 2012.

Hutchinson B. (2004). Acquiring the meaning of discourse markers. In the 42nd Annual Meeting of the Association for Com putational Linguistics (ACL 2004), p. 684-691, Barcelona, Spain.

Hussein M. (2008a). The discourse markers 'but' in English and Standard Arabic: One procedure and different implementation. http://www.students.ncl.ac.uk/miri.hussein/publication.html.

Hussein M. (2008b). Discourse markers and procedural meaning: The case of fa in Standard Arabic. http:// www.students.ncl.ac.uk/miri.hussein/publication.html (accessed 20/2/2009).

Jagadeesh J, Prasad Pingali, Vasudeva Varma, (2005). Sentence Extraction Based Single Document Summarization. In Workshop on Document Summarization, 19th and 20th March, 2005, IIIT Allahabad.

Jezek K. and Steinberger J. (2008). Automatic Text summarization, Vaclav Snasel (Ed.): Znalosti 2008, pp.1- 12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

Jiun-Shiung Wu. (2005). The Semantics of the Perfective LE and Its Context-Dependency: An SDRT Approach. Journal of East Asian Linguistics 14.4: 299-366.

Jirawan C., Thana S., and Asanee K. (2005). Element Discourse Unit Segmentation for Thai Discourse Cues and Syntactic Information, The 9th National Computer Science and Engineering Conference, 27-28 October, 2005.

Kamp H., Van Genabith J. and Reyle U. (2011). Discourse Representation Theory. D. Gabbay and F. Guenthner (eds.), Handbook of Philosophical Logic, Volume 15, 125–394. DOI 10.1007/978-94-007-0485-5_3, Springer.

Kamp H. (1981). A theory of truth and semantic representation. In Formal Methods in the Study of Language; Mathematical Centre Tracts 135, eds. J.A.G. Groenendijk et al., 277-322. Amsterdam: Mathematical Centre.

Kamp H. and Hans (2001a). The importance of presupposition. In Rohrer, Christian; Rossdeutscher, Antje; and Kamp, Hans, Liguistic Form and its Computation. CSLIPublications, Standord.

Kamp H. and Hans (2001b). Presupposition computation and presupposition justification. In Bras, Myriam and Vie, Laure, Pragmatic and Semantic Issues in Discourse and Dialogue. Elsevier.

Karoui J., Graja M., Boudabous M. M. and Belguith Hadrich L. (2013). Semi-automatic Domain Ontology Construction from Spoken Corpus in Tunisian Dialect: Railway Request Information, International Journal of Recent Contributions from Engineering, Science & IT (iJES), Vol. 1, No.1, pp. 35-38, August 2013.

Keskes I., Benamara F., Acher N., Belguith Hadrich L. and Boujelben I. (2015). The Discourse Arabic Treebank: Towards Building Recursive and Complete Discourse Structures of Arabic Texts. Forthcoming in Language Resources and Evaluation (LRE) (under revision).

Keskes I., Benamara F. and Belguith Hadrich L. (2014a). Splitting Arabic Texts into Ele-mentary Discourse Units. Journal ACM Transactions on Asian Language Information Processing (TALIP). Volume 13, Issue 2, June 2014, Article No. 9, doi:10.1145/2601401.

Keskes I., Benamara F. and Belguith Hadrich L. (2014b). Learning Explicit and Implicit Arabic Discourse Relations. Journal of King Saud University - Computer and Information Sciences, Issue Spéciale on Arabic NLP, Elsevier, Volume 26, Issue 4, December 2014, Pages 398–416, doi:10.1016/j.jksuci.2014.06.001.

Keskes I., Benamara F. and Belguith Hadrich L. (2013). Segmentation de textes arabes en unités discursives minimales. 20th International conference in NLP (TALN 2013), regular paper, France.

Keskes I., Benamara F. and Belguith Hadrich L. (2012a). Discourse Segmentation of Arabic Texts Based on Cascade Grammars», NooJ conference (NooJ 2012), Paris, 14-16 June 2012.

Keskes I., Benamara F. and Belguith, Hadrich L. (2012b). Clause-based Discourse Segmentation of Arabic Texts, The eighth international conference on Language Resources and Evaluation (LREC 2012), Istanbul, 21-27 may 2012.

Keskes I., Lhioui M., Benamara F. and Belguith, Hadrich L. (2012c). التلخيص الآلي للنصوص العربية بالاعتماد على نظرية البنية الخطابية المجزأة. Automatic summarization of Arabic texts based on SDRS graph, International Computing Conference in Arabic (ICCA 2012), Egypt, 26-28 December 2012.

Keskes I., Boudabous M. M., Maaloul M. H. and Belguith, Hadrich L. (2012d). Etude comparative entre trois approches de résumé automatique de documents arabes. 19th International conference in NLP (TALN 2012), Gronoble, 4-8 June 2012.

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

Khalifa I., Zakareya A., and Farawila A. M. (2012). A Comprehensive Taxonomy of Arabic Discourse Coherence Relations. The Third International Conference on Communications and Information Technology ICCIT. Beirut, Lebanon.

Knott A. (1996). A data driven methodology for motivating a set of coherence relations, PhD thesis, 1996.

Kruengkari C. and Jaruskulchai C. (2003). Generic Text Summarization Using Local and Global Properties of Sentences, Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03) , 2003.

Kupiec J., Pererson J., and Chen F. (1995). A trainable document summarizer. Research and Development in Information Retrieval, pages 68–73.

Koch B.J. (1983). Presentation as proof: The language of arabic rhetoric. Anthropological linguistics 25, 47-60.

Lascarides A. and Asher N. (1991) Discourse Relations and Defeasible Knowledge, in Proceedings to the 29th Annual Meeting of the Association of Computational Linguistics (ACL91), pp55--63, Berkeley USA, June 1991.

Le Thanh H., Abeysinghe G., and Huyck C. (2004). Generating discourse structures for written text. In Proc. of the 20th International Conference on Computational Linguistics (COLING), pages 329–335, Geneva/Switzerland.

Lee A., Prasad R., Joshi A. and Webber B. (2008). Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank. Proc. Constraints in Discourse III Workshop.

Lehmam A. (2000). Résumé de texte automatique: des solutions opérationnelles, La Tribune des Industries de la Langue, de l'Information Électronique et du Multimédia, OFIL, Paris, pp.50-58.

Lehmam A. (2010). Essential summarizer: innovative automatic text summarization software in twenty languages. RIAO '10: Adaptivity, Personalization and Fusion of Heterogeneous Information Publisher: le centre de hautes etudes internationales d'informatique documentaire.

Lin C.Y. and Hovy E. (2002). Automated multi-document summarization in neats. In Proceedings of the Human Language Technology Conference (HLT2002).

Lin Ch. and Hovy E. (2003). Automatic Evaluation of Summaries Using n-Gram CoOccurrence Statistics. In Proceedings of HLT-NAACL, Edmonton, Canada, 2003.

Lin Z., Kan M. and Tou H. (2009). Recognizing implicit discourse relations in the penn discourse treebank. In EMNLP, pages 343–351.

Lin Z., Tou H. and 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. and Nenkova A. (2010). Using entity features to classify implicit discourse relations. In SIGDIAL Conference, pages 59–62.

Lüngen H., Lobin H., Bärenfänger M., Hilbert M. and Puskas, C. (2006). Text parsing of a complex genre. In Bob Martens and Milena Dobreva, editors, Proc. of the Conference on Electronic Publishing (ELPUB 2006), Bansko, Bulgaria.

Maâloul H. M. (2012). Approche hybride pour le résumé automatique de textes. Application a la langue arabe. Document and Text Processing. Universit´e de Provence - Aix-Marseille French.

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

Maamouri M., Bies A., Kulick S. Krouma S., Gaddeche and 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. and Kulick, S. (2010a). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium, Catalog No.: LDC2010L01.

Mann W.C. and 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 theDiscourse Structures of Texts. Technical Report, University of Southern California, 1999.

Marcu D. and Echihabi A. (2002). An unsupervised approach to recognizing discourse relations. In ACL, pages 368–375.

Marcu D. (2000a). From discourse structures to text summaries in Workshop Intelligent Scalable Text Summarization ACL, p. 82-88, Madrid, Espagne, 2000.

Marcu D. (2000b). The Theory and Practice of Discourse Parsing and Summarization. The MIT Press, Cambridge, MA, USA.

Marcus M., Santorini B., and Marcinkiewicz M. (1993). Building a large annotated corpus of English: the Penn treebank. Computational Linguistics, 19:313-330.

Marton Y. Habash N. and Rambow O. (2013). Dependency parsing of modern standard Arabic with lexical and inflectional features. Journal Computational Linguistics archive Volume 39 Issue 1, Pages 161-194, March 2013.

Maskey S. and Hirschberg J. (2005). Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. Proceedings of Interspeech 2005. Lisbon, Portugal.

Mathkour H., Touir A. and Al-sanea W. (2008). Parsing Arabic Texts Using Rhetorical Structure Theory. Journal of Computer Science 4 (9): 713-720, 2008 ISSN 1549-3636

Mei Q., Guo J. and Radev D. (2010). Divrank: the interplay of prestige and diversity in information networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD-10), pp. 1009–1018.

Mendoza M., Bonilla S., Noguera C., Cobos C. and León E. (2014). Extractive single-document summarization based on genetic operators and guided local search. Expert Systems With Applications journal. Elsevier Science. DOI 10.1016/j.eswa.2013.12.042

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

Mihalcea R. and Tarau P., (2005). An Algorithm for Language Independent Single and Multiple Document Summarization, In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Korea, 2005.

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

Mladova L., Zikanova S. and Hajicova E. (2008). From sentence to discourse: Building an annotation scheme for discourse based on the Prague Dependency Treebank. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Minel J. L. (2002). Filtrage sémantique de textes. Problèmes, conception et réalisation d'une plate-forme informatique. Habilitation à diriger des recherches, Université ParisSorbonne.

Mohamed A.H. and Omer M.R. (1999). Syntax as a Marker of Rhetorical Organization in Written Texts: Arabic and English, International Review of Applied Linguistics in Language Teaching (IRAL) 37(4): 291–305.

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

Mourad A. and Darwish K. (2013). Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 55–64, Atlanta, Georgia, 14 June 2013.

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

Musawi A. and Muhsin J. (2001). "Arabic Rhetoric", in Thomas O. Sloane (Ed.), Oxford Encyclopaedia of Rhetoric, Oxford: Oxford University Press, pp. 29-33.

Nenkova A. and Passonneau R. (2005). Evaluating Content Selection in Summarization: The Pyramid Method. In Document Understanding Conference, Vancouver, Canada, 2005.

Neri F. and Cotta C. (2012). Memetic algorithms and memetic computing optimization: A literature review Swarm and Evolutionary Computation, 2 (2012), pp. 1–14.

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

Ono K., Sumita K. and Miike S. (1994). Abstract Generation Based on Rhetorical Structure Extraction. In the Proceedings of the International Conference on Computational Linguistic – Coling-94, pp 344-348, Japan, 1994.

Osminin P. G. (2014). A Summarization Model Based on the Combination of Extraction and Abstraction. DIALOG 2014.

Ostler S. (1987). Academic and ethnic background as factors affecting writing performance. In: A. Purves (Ed.), Writing across Languages and Cultures: Issues in Contrastive Rhetoric, pp. 261-272. Newbury Park, CA.

Othman W. (2004). Subordination and coordination in english-arabic translation. Al-Basaer 8(2), pp.12-33.

Oufaida H., Nouali O., Blache P. (2014). Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. Journal of King Saud University - Computer and Information Sciences. Special Issue on Arabic NLP. Volume 26, Issue 4, Pages 450-461 (December 2014) .

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.

Oza U., Prasad R., Kolachina S., Sharma D. M. and Joshi A. (2009). The hindi discourse relation bank. In Proc. 3rd ACL Language Annotation Workshop (LAW III), Singapore, August.

Pai A. (2014). Text Summarizer Using Abstractive and Extractive Method. International Journal of Engineering Research & Technology. Vol. 3 - Issue 5, e-ISSN: 2278-0181

Pallavi D. and Mane P. (2014). An Overall Survey of Extractive Based Automatic Text Summarization Methods. International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064. Volume 3 Issue 11, November 2014

Palmer M., Kingsbury P., and Gildea D. (2005). The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, pages 71–106, 2005.

Pardo T.A.S., Nunes M.G.V. and Rino, L.H.M. (2004). DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. Lecture Notes in Artificial Intelligence.

Park J. and Cardie C. (2012). Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. Proceedings of the 13th Annual SIGdial Meeting on Discourse and Dialogue, SIGDIAL 2012.

Pitler E., Louis A. and Nenkova A. (2009). Automatic sense prediction for implicit discourse relations in text. In 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, 2009.

Pitler E., Raghupathy M., Mehta H., Nenkova A., Lee A. and Joshi A. (2008). Easily Identifiable Discourse Relations, Proceedings of COLING, 2008.

Polanyi L. and Scha R. (1983). The syntax of discourse. Text 3: 261-270.

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

Qazvinian V., Dragomir R., Saif M., Bonnie D., David Z., Whidby M. and Moon T. (2013). Generating Extractive Summaries of Scientific Paradigms. Journal of Artificial Intelligence Research 46 (2013) 165-201.

Quirk R., Greenbaum S., Leech G. and Svartvik J. (1985). A comprehensive grammar of the English language. London: Longman.

Ratnaparkhi A. (1997). A simple introduction to maximum entropy models for natural language processing. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania. An easy-to-read introduction to maximum entropy methods in the context of natural language processing.

Redeker G. (1991). Review article: Linguistic markers of discourse structure. Linguistics 29(6):

1139-1172.

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

Reid J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. Journal of Second Language Writing 1(2), pp. 79-107.

René A. G. H. and Yulia L. (2009). Word Sequence Models for Single Text Summarization, ACHI, 2009, International Conference on Advances in Computer-Human Interaction, International Conference on Advances in Computer-Human Interaction 2009, pp. 44-48, doi:10.1109/ACHI.2009.58.

Riley F., Webber B. and Joshi A. (2006). Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. Journal of Semantics 23(1):55–106.

Roze C., Danlos L. and Muller P. (2012). LEXCONN: A French Lexicon of Discourse Connectives, Discours, November 2012, URL : http://discours.revues.org/8645 ; DOI : 10.4000/discours.8645.

Ryding K.C. (2005). A Reference Grammar of Modern Standard Arabic. Reference Grammars. Cambridge University Press, New York.

Sadek J., Chakkour F. and Meziane F. (2012). Arabic rhetorical relations extraction for answering "why" and "how to" questions. In: Natural Language Processing and Information Systems, Lecture Notes in Computer Science, vol. 7337, pp. 385{390. Springer Berlin Heidelberg.

Sadat F. and Mohamed E. (2013). Improved Arabic-French Machine Translation through Preprocessing Schemes and Language Analysis. Canadian Conference on AI 2013: 308-314, 2013.

Salman J. (2003). **موسوعة معاني الحروف العربية**. عمان، دار أسامه للنشر والتوزيع

Saggion H. and Lapalme G. (2002). Generating indicative-informative summaries with SumUM. Computational Linguistics, 28(4):497–526.

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

Sarjala M. (1994). Signalling of Reason and Cause Relations in Academic Discourse, Anglicana Turkuensia 13: 89–98.

Sawalha M. Atwell ES. and Abushariah M. (2013). SALMA: Standard Arabic Language Morphological Analysis. Proceedings ICCSPA International Conference on Communications, Signal Processing, and their Applications, pp.1-6. 2013.

Schiffman B. (2002). Building a resource for evaluating the importance of sentences. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).

Schorup (1985). Common discourse particle in English conversation: Like, well, y'know. New York: Garland.

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

Sharoff S. and Sokolova L. (1995). Analysis of Rhetorical Structures in Technical Manuals and their Multilingual Generation, Proceedings of the Workshop on Multilingual Generation (IJCAI'95), pp. 119–28, Montréal, Canada.

Shen C. and Li T. (2010). Multi-Document Summarization via the Minimum Dominating Set. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 984–992, Beijing, August 2010.

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

Sloane T. O. (2001). Encyclopedia of Rhetoric. New York: Oxford University Press. xii, 837 pp.

Soricut R. and Marcu D. (2003). Sentence level discourse parsing using syntactic and lexical information. In HLT/NAACL, Edmonton, Canada.

Soames S. (1984). Presupposition. In Gabbay, D. and Guenthner, F., editors 1984,

Handbook of Philosophical Logic. Reidel. 553–616. Volume IV.

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

Stede M. (2004). The Potsdam Commentary Corpus. In Proceeding of the ACL-04 Workshop on Discourse Annotation, Barcelona, July 2004.

Stubbs M. (1983). Discourse analysis. Chicago, IL: The University of Chicago Press.

Suanmali L., Mohammed S. B. and Naomie S. F. (2009). Sentence Features Fusion for Text Summarization Using Fuzzy Logic, 978-0-7695-3745-0/09, 2009 IEEE.

Suanmali L., Naomie S. F. and Mohammed S. B. (2011). Fuzzy Genetic Semantic Based Text Summarization, 978-0-7695-4612-4/11, 2011 IEEE.

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

Subba R. and 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, pages 566–574,Boulder, Colorado, June 2009.

Sumita K., Ono K., Chino T., Ukita T. and Amano S. (1992). A discourse structure analyzer for Japanese text. In Proceedings of the international conference on fifth generation computer systems, Tokyo, Japan ,1133–1140.

Taboada M. and Mann W.C. (2006). Applications of Rhetorical Structure Theory. Discourse Studies 8 (4): 567-588.

Taha K., Jarrah M. A. and Jarrah R. (2013). DISCOURSAL WA (AND). International Journal of English Language and Linguistic Research. Vol. 1, No. 1, pp.10-20, June 2013.

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

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

Trigui O., Belguith Hadrich L., Rosso P. Ben Amor H. and Gafsaoui B. (2012). IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval. In: Forner P., Karlgren J., Womser-Hacke C. (Eds.), Notebook Papers of CLEF 2012 LABs and Workshops, CLEF-2012, September 17-20, Rome, Italy.

Van der S. (1992). Presupposition projection as anaphora resolution. Journal of Semantics 9:333–377. Special Issue: Presupposition, Part 2.

Van der Vlieth N., Berzlanovich I., Bouma G., Egg M. and Redeker G. (2011). Building a Discourse-Annotated Dutch Text Corpus. The DGfS Workshop "Beyond Semantics", Bochumer Linguistische Arbeitsberichte .

Varghese V. and Saravanan J. (2014). A Systematic Approach for News Caption Generation. International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014). ISSN: 2347 - 8446 (Online). Vol. 2, Issue 2, Ver. 1.

Venant A., Asher N., Muller P., Denis P. and 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). P. 264-275.

Vincent A. (2010). MuLLinG: MultiLevel Linguistic Graphs for Knowledge Extraction. Proceedings of TextGraphs-5 - ACL-2010 Workshop on Graph-based Methods for Natural Language, Association for Computational Linguistics. pp.69-73.

Wan X., Li H. and Xiao J. (2010). Cross-language document summarization based on machine translation quality prediction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), pp. 917–926.

Wang D., Zhu S., Li T. and Ding C. (2008). Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in Proc. SIGIR, 2008, pp. 307–314.

Wang D., Zhu S., Li T. and Gong Y. (2009). Multi-document summarization using sentence-based topic models. In Proceedings of the ACLIJCNLP.

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

Webber B., Stone M., Joshi A. and Knott A. (2003). Anaphora and discourse structure. Computational Linguistics, 29:545–587, 2003.

Webber B. L. (2004). D-LTAG: extending lexicalized TAG to discourse. Cognitive Science 28(5): 751-779.

Webber B., Knott A. and Joshi A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In Bunt, Muskens, and Thijsse, editors, Computing Meaning, volume 2, pages 229–245. Kluwer, Dordrecht.

Webber B. (2006). Accounting for discourse relations: constituency and dependency. In M. Dalrymple, editor, Intelligent linguistic architectures, pages 339–360, 2006.

Webber B., Egg M. and Kordoni V. (2012). Discourse structure and language technology. Natural Language Engineering 18, 437-490.

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

Wolf F., Gibson E. Fisher A. and Knight M. (2003). A procedure for collecting a database of texts annotated with coherence relations. Documentation accompanying the Discourse GraphBank, LDC2005T08.

Wolf F. and Gibson E. (2005). Representing Discourse Coherence: A Corpus-Based Study. Computational Linguistics, 249–287.

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

Wright W. (1975). A Grammar of Arabic Language.

Xu Y., Wang X., Liu Tao, Xu Z. (2007). Multi-document Summarization Based on Rhetorical Structure: Sentence Extraction and Evaluation, IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 3034 – 3039.

Xue N. (2005). Annotating discourse connectives in the Chinese treebank. In ACL Workshop on Frontiers in Corpus Annotation II, Ann Arbor MI.

Yeh J. Y., Ke H. R. and Yang W. P. (2008). iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. Expert Systems with Applications, 35(3):1451 – 1462.

Yongzheng Z., Nur Z. and Evangelos M. (2005). Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora, WIDM'5, 51-57, Bremen Germany,2005.

Yong-dong X., Xiao-long W., Tao L. and Zhi-ming X. (2007). Multi-document Summarization Based on Rhetorical Structure: Sentence Extraction and Evaluation, IEEE International Conference on Systems, Man and Cybernetics, 2007,pp.3034 –3039.

Zajic D. M., Dorr B. J., Lin J. and Schwartz R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Information Processing and Management (Special Issue on Summarization).

Zamanifar A., minaei-Bidgoli B. and Sharifi M. (2008). A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of Text, In Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, 635-639, Iran, 2008.

Zeyrek D. and Webber B. (2008). A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In Proceedings of the 6th Workshop on Asian Language Resources (ALR6).

Zeyrek D., Demirsahin I., Sevdik C. A., Balaban H. O., Yalcınkaya I. and Turan U. D. (2010). The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In Proceedings of the 4th Linguistic Annotation Workshop (LAW III).

Zeyrek D., Turan U. D., Bozsahin C. and Cakıcı R. (2009). Annotating Subordinators in the Turkish Discourse Bank. In Proceedings of the 3rd Linguistic Annotation Workshop (LAW III).

Zhang J., Cheng X., and Xu H. (2008). AdaSum: an adaptive model for summarization. In Proceedings of the ACM 17th Conference on Information and Knowledge Management. ACM Press, New York, pp. 901 – 909.

Zhang J., Xu H. and Cheng X. (2008). GSPSummary: a graph-based sub-topic partition algorithm for summarization. In Proceedings of the 2008 Asia Information Retrieval Symposium. Springer-Verlag, Berlin, Heidelberg, pp. 321 – 334.

Zhanq J., Sun L. and Zhou Q. (2005). A Cue-based HubAuthority Approach for Multi-Document Text Summarization, in Proceeding of NLP-KE'05, IEEE,642-645, 2005

Zhou Y. and Xue N. (2012). Pdtb-style discourse annotation of chinese text. In Proc. 50th Annual Meeting of the ACL, Jeju Island, Korea.

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

Zribi I, Graja M., Ellouze K. M., Jaoua M. and Belguith Hadrich L. (2013). Orthographic Transcription for Spoken Tunisian Arabic , A. Gelbukh (Ed.): CICLing 2013, Part I, LNCS 7816, pp. 153–163, 2013.

Zhu X. and Penn G. (2006). Summarization of Spontaneous Conversations. Proceedings of Interspeech 2006. Pittsburgh, PA.

Zufferey S., Degand L., Popescu-Belis A. and Sanders T. (2012). Empirical validations of multilingual annotation schemes for discourse relations. In: Proceedings of the 8th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, p. 77-84.

Abstract: In this dessertation, we propose the first effort towards a semantically driven approach of Arabic texts following the Segmented Discourse Representation Theory. Our main contributions are:
-A study of the feasibility of building a recursive and complete discourse structures of Arabic texts. In particular, we propose:
*an annotation scheme for the full discourse coverage of Arabic texts, in which each constituent is linked to other constituents. A document is then represented by an oriented acyclic graph which captures explicit and implicit relations as well as complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups, and crossed dependencies.
*a novel discourse relations hierarchy. We study rhetorical relations from a semantic point of view by focusing on their effect on meaning and not on how they are lexically triggered by discourse connectives that are often ambiguous, especially in Arabic.
*a quantitative analysis (in terms of discourse connectives, relation frequencies, proportion of implicit relations, etc.) and qualitative analysis (inter-annotator agreements and error analysis) of the annotation campaign.
-An automatic discourse parser where we investigate both automatic segmentation of Arabic texts into elementary discourse units and automatic identification of explicit and implicit discourse relations.
-An application of our discourse parser in Arabic text summarization. We compare tree-based vs. graph-based discourse representations for producing indicative summaries and show that the full discourse coverage of a document is definitively a plus.

Keywords: Discourse analysis, Discourse connectives, Discourse relations, Discourse structure, Segmented Discourse Representation Theory, Automatic summarization.

Résumé : Dans cette thèse, nous proposons le premier effort vers une approche basée sur l'analyse sémantique de textes arabes selon la théorie de la représentation discursive segmentée. Nos principales contributions sont les suivantes :

-Une étude de la faisabilité de la construction d'une structure de discours récursive et complète de textes arabes. En particulier, nous proposons :
*un schéma d'annotation qui couvre la totalité d'un texte arabe, dans lequel chaque constituant est lié à d'autres constituants. Un document est alors représenté par un graphe acyclique orienté qui capture les relations explicites et les relations implicites ainsi que des phénomènes de discours complexes, tels que l'attachement, la longue distance du discours pop-ups et les dépendances croisées.
*une nouvelle hiérarchie des relations de discours. Nous étudions les relations rhétoriques d'un point de vue sémantique en se concentrant sur leurs effets sémantiques et non pas sur la façon dont elles sont déclenchées par des connecteurs de discours, qui sont souvent ambigues en arabe.
*analyse quantitative (en termes de connecteurs de discours, les fréquences de relations, proportion de relations implicites, etc.) et une analyse qualitative (accord inter-annotateurs et analyse des erreurs) de la campagne d'annotation.
-Un outil d'analyse de discours où nous étudions à la fois la segmentation automatique de textes arabes en unités de discours élémentaires et l'identification automatique des relations explicites et implicites du discours.
-L'utilisation de notre outil pour résumer les textes arabes. Nous comparons la représentation de discours en graphes et en arbres pour la production de résumés.

Mots clés : Analyse de discours, Connecteurs de discours, Relations de discours, Structures de discours, la Théorie de la Représentation Discursive Segmentée, Résumé automatique.