UNIVERSITY OF LEEDS

# Human and Automatic Annotation of Discourse Relations for Arabic

by

**Amal Alsaif**

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

**The University of Leeds**
**School of Computing**

**August 2012**

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated overleaf. The appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# Abstract

This thesis describes the first, inter-disciplinary, study on human and automatic discourse annotation for explicit discourse connectives in Modern Standard Arabic (MSA). Discourse connectives are used in language to link discourse segments (arguments) by indicating so-called discourse relations. Automating the process of identifying the discourse connectives, their relations and their arguments is an essential basis for discourse processing studies and applications. This study presents several resources for Arabic discourse processing in addition to the first machine learning algorithms for identifying explicit discourse connectives and relations automatically. First, we have collected a large list of discourse connectives frequently used in MSA. This collection is used to develop the READ tool: the first annotation tool to fit the characteristics of Arabic, so that Arabic texts can be annotated by humans for discourse structure. Second, our analysis of Arabic discourse connectives leads to formalize an annotation scheme for connectives in context, based on a popular discourse annotation project for English, the PDTB project. Third, we used this scheme to create the first discourse corpus for Arabic, the Leeds Arabic Discourse Treebank (LADTB v.1). The LADTB extends the syntactic annotation of the Arabic Treebank Part1 to incorporate the discourse layer, by annotating all explicit connectives as well as associated relations and arguments. We show that the LADTB annotation is reliable and produce a gold standard for future work. Fourth, we develop the first automatic identification models for Arabic discourse connectives and relations, using the LADTB for training and testing. Our connective recogniser achieves almost human performance. Our algorithm for recognizing discourse relations performs significantly better than a baseline based on the connective surface string alone and therefore reduces the ambiguity in explicit connective interpretation. At the end of the thesis, we highlight research trends for future work that can benefit from our resources and algorithms on discourse processing for Arabic.

# Acknowledgments

First of all, thanks be to Allah (GOD), who gave me the wellbeing and skills to finish one of the most important studies and experiences in my life for Arabic NLP. Then, the biggest thank you goes to my parents Mrs. Hellah Alsaif and Mr. Suliman Alsaif: without your motivation and prayers I would not have completed this study. Thank you also to my wonderful supervisor Dr. Katja Markert; I learnt from you a lot: being focused, precise, proactive and a researcher of high standard. Your guidance in making hard decisions during the annotation tasks in the study inspired me. Thanks for understanding my circumstances as a mother and the difficult times that I faced.

I would like also to thank all my colleagues in the NLP group for sharing ideas and working together. Thanks to Dr. Eric Atwell, Dr. Serge Sharoff, Andrew McKinlay, Dr. Majdi Sawalha, Abdul-Baquee Sharaf, Saman Hina, Dr. Claire Brierley and Kais Dukes. Thanks to Dr. Claire Brierley again for helping to proofread my thesis and showing her interest in getting involved in future research of Arabic discourse processing. A big thank you for the patient intelligent annotators: Dr. Latifa Alsulaiti, Abdul-Baquee Sharaf, Boshra Al-shyban, Maryam Algawi, and Basmah Al-Soli. We have achieved significant results together. A special thank you goes to Dr. Hussein Abdul-Raof for the linguistic advice on the collection of discourse connectives. Another special thank you goes to the PDTB team for valuable discussions. Final academic acknowledgement is due to Imam Muhammad Ibn Saud University (Saudi Arabia) for sponsoring the study, to the British Academy

for additional funding for annotators and creating the LADTB corpus, and to Denise DiPersio, the manger of the LDC, to distribute the corpus.

Lastly, but not least, I thank my dear husband Suleiman Alturki. You are my sweet heart that I cannot live without. You were with me at each moment in the research encouraging and helping me. I will not forget our discussion over some of the disagreements between annotators and annotation results, and remember all our conversations. You have done a lot for me. May Allah bless your support in all aspects and forgive any derelictions by both of us. My kids Saffanah, Anas and Ammar, I love you more than anyone in the world. I did my best to not let my study affect your growing up to be such great children as you are. I hope you will find living in the UK a nice memory for all of you in the future.

To my sisters and brothers in Saudi Arabia (Ahmad, Mohammad, Hithm, Maha, Asma, Manal, Mohanad, Ebtihal and Alaa) and in the UK who always stand beside me support me, trust me and love me: I appreciate all what you did for me, thank you so much.

# Declarations

Some parts of the work presented in this thesis have been published in the following papers. For each paper, the contributions by me and other authors are specified as well as the parts of the thesis that related to the paper.

**Alsaif, A, Markert, K and Abdul-Raof H. Corpus-based study: Extensive Collecting of Discourse Connectives for Arabic. The Third Saudi International Conference SIC2009, Surrey, 2009.**
**Contributions:** Amal Alsaif is the lead author, responsible for research design and all experiments. Hussein Abdul-Raof provided linguistic guidance and feedback. Katja Markert provided supervision, feedback and general guidance with regard to experiment design and write-up.

*Chapter 4 is mainly based on the work presented in this paper.*

**Alsaif, A and Markert, K. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic in: The Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Research Association. 2010.**
**Contributions:** Amal Alsaif is the principle author. Amal conducted all annotation experiments and contributed to research design and write-up. Katja Markert provided supervision, feedback, general guidance and contributed to paper write-up and research design (evaluation methodology).

*The work here is a summary of Chapter 6 and parts of Chapter 7.*

**Alsaif, A. Annotating Discourse Connectives in MSA: Disagreement Cases in the LADTB. Corpus Linguistics 2011, Discourse and Corpus Linguistics, Birmingham, Uk, 2011.**
Contributions: Lead author, responsible for research design and all experiments.

*Section 7.5 is based on the work presented in this paper.*

**Alsaif, A. and Markert, K. Modelling Discourse Relations for Arabic, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, July 2011, Edinburgh.**
**Contributions:** Equal lead authors. Both authors contributed equally to research design and paper write-up. Amal Alsaif conducted all experiments and interpreted the results. Katja Markert provided supervision, feedback and general guidance.

*Chapter 8 (apart from the error analysis) is based on the work presented in this paper.*

# Table of Contents

# List of Tables

# List of Figures

# Glossary of main terms and abbreviations used in the thesis

| | |
|---|---|
| **Discourse Connective (DC)** | A lexical marker used to link two abstract objects in a text. |
| **Abstract Object (AO)** | Abstract objects in discourse are things like proposition, events, facts and opinions. |
| **Argument (Arg)** | A text expressing an abstract object and linked by a DC. |
| **Human discourse Annotation** | Labelling discourse connectives and their arguments and relations in context by a human. |
| **MSA** | Modern Standard Arabic |
| **PDTB/ PDTB2** | The Penn Discourse Treebank |
| **ATB/PATB** | The Penn Arabic Treebank |
| **RST** | Rhetorical Structure Theory |
| **LADTB** | Leeds Arabic Discourse Treebank |
| **QAD/KQC** | Quranic Arabic Dependency Treebank/Kais Quranic Corpus |
| **LDM** | Linguistic Discourse Model |
| **RST-DT** | RST Discourse Treebank |
| **DU/DS** | Discourse Unit/ Discourse Segmant |
| **DCU** | Discourse Constituent Unit |
| **PCC** | Postsdam Commentary Corpus |
| **PADT** | Prague Arabic Dependency Treebank |
| **CATiB** | Columbia Arabic Treebank |
| **SDRT** | Segmented Discourse Representation Theory |
| **PA** | Percentage Agreement |
| **Ambiguous DC** | The ambiguous discourse connective can be (i) a potential connective which does not always have a discourse function in a context, or (ii) a connective which always has a discourse function in context but might signal more than one relation. The usage of the term differs according to the section topic. |
| **LDC** | Linguistic Data Consortium |
| **POS** | Part of Speech |

**The Common POS tags in the Penn TB and the Penn Arabic TB (Part1 v.2)**

| PTB tag | Description | PATB tag |
|---|---|---|
| CC | Coordinating conjunction | CONJ |
| CD | Cardinal number | NUM |
| DT | Determiner | DET |
| IN | Preposition or subordinating conjunction | FUNC_WORD, PREP |
| JJ | Adjective | ADJ |
| NN | Noun, singular or mass | NOUN+NSUFF |
| NNS | Noun, plural | NOUN+NSUFF_PL/DUL |
| NNP | Proper noun, singular | NOUN_PROP |
| NNPS | Proper noun, plural | NOUN_PROP_PL/DUL |
| PRP | Personal pronoun | IVSUFF_DO, PRON |
| PRP$ | Possessive pronoun | POSS_PRON |
| RB | Adverb | ADV |
| RP | Particle | PART |
| VBD | Perfect verb, past tense | VERB_PERFECT |
| VBN | Passive verb,old past participle | VERB_PASSIVE |
| VBP | Imperfect verb, non-3rd person singular present | VERB_IMPERFECT |
| WP | Wh-pronoun | REL_PRON |
| WRB | Wh-adverb | REL_ADV |

# Chapter 1
# Introduction

In the last two decades, discourse structure studies have become an attractive but challenging field for the NLP community. A text is not only a sequence of sentences or clauses, but rather it is a coherent object that has many cohesive devices linking its units (words, clauses and sentences). One of the critical aspects of such coherence concerns theoretical relations, or *discourse relations* as they are also known.

*Discourse relations* are semantic relations such as causality, contrast and temporality, that connect two textual units, typically clauses or sentences (Asher 1993a; Halliday and Hassan 1976). The textual units connected should express *abstract objects* (AOs) such as events, actions, facts or beliefs. They are also called *arguments* (Asher 1993a). There are two types of discourse relations: (i) relations that are signalled explicitly via so called *discourse connectives* (explicit relations), and (ii) relations that can be inferred from the context without any explicit signaling (implicit relations).

**Ex. 1-1**

(a) John didn't go to the party$_{cl1}$ **because** he was tired$_{cl2}$. **Instead**, he went to bed$_{cl3}$.
(b) John didn't go to the party. He was tired.

In Ex. 1-1 (a) the connective *because* in the second clause cl2 establishes explicitly that the reason for John being absent from the party, cl1, is that he was tired: a causal relationship. However, the connective *instead* in the third clause cl3 contrasts *going to bed* with *going to the party*; a contrast relation. The connective *because* therefore takes cl1 and cl2 as its arguments, whereas *instead* takes the non-adjacent units cl1 and cl3 as its arguments. Both relations are explicit relations. By contrast, in Ex. 1-1 (b) the second sentence in the example gives a potential reason for the event in the

first sentence: there is a causal relationship between the two arguments. This relation is inferred from the context without using any connectives.

Discourse relations are widely studied in theoretical linguistics (Halliday and Hassan 1976; Hobbs 1985), where a number of different relational taxonomies have been derived (Knott and Sanders 1998; Hobbs 1985; Mann and Thompson 1988; Marcu 2000c; Prasad *et al.* 2008a; Webber and Prasad 2006). As a result of these, different inventories have been used in annotating English corpora for discourse relations (Marcu 2000c; Marcu 2000a; Webber and Prasad 2006; Hobbs 1985; Carlson *et al.* 2002), these also can differ in other respects, such as whether they prescribe a tree, a graph or a flat structure for discourse annotation (more details are discussed in Chapter 2). In addition, the English discourse corpora have been used as a basis for the automatic discovery of discourse relations (Lin, Kan and Ng 2009; Pitler, Louis and Nenkova 2009; Pitler *et al.* 2008; Wang, Su and Tan 2010; Prasad *et al.* 2005; Marcu, Lynn and Maki 2000; Marcu 2000b). In contrast, for many other languages, neither corpora annotated with discourse relations nor automatic methods exist.

This study presents the first effort to annotate a corpus with discourse relations for Arabic, and the first corpus study to develop automatic models for the recognition of Arabic discourse relations and connectives. The next section describes what motivated this study for Arabic, our claims and goals. Then, we summarize the contributions of the work (Section 1.2) and describe the thesis structure (Section 1.3).

## 1.1 Motivation and Research Statement

Arabic remains a challenging language in many respects for computational linguistic studies. Arabic has a complex morphology, a free word order in addition to the possibility of constructing a full clause or sentence using only one token. Sentences in Arabic writing are often long, using punctuations but not always in a systematic way such as in other languages. That makes the automatic determination of clause and sentence boundaries another challenge for Arabic studies. The language uses both letters and other symbols such as Hamzah (ء) and diacritics. These symbols are often not used in modern Arabic writing such as newswire. That leads to a higher ambiguity level in automatic recognition/tagging of words in Arabic. Moreover, there is a wide variety of lexical expressions in Arabic to link discourse parts such as

discourse connectives. Section 3.1 describes more characteristics of MSA, together with their impact on this this thesis.

Discourse connectives are mostly unambiguous in English (Pitler and Nenkova 2009), so that their relations are easily identified automatically on the basis of the connective string. Discourse connectives therefore are intensively studied in theoretical linguistics, and offer a wide range of applications in computational linguistics as well. For example, in automatic text generation, it is necessary to use the right connectives in the right places in the generated text (Hovy 1993). Moreover, for text summarization, text segments offering mainly elaboration of related text segments might be ignored (Marcu 2000c). Discourse connectives are also used in improving machine translation, in essay marking and in question answering systems (Popescu-Belis and Zufferey 2006; Marcu, Lynn and Maki 2000; Pitler and Nenkova 2008; Girju 2003; Taboada and Mann 2006a). More details about these applications are discussed in Section 2.7.

To date, theoretical studies as well as studies on applications have tended to focus on English. Despite the fact that natural languages have elements in common, each has a special flavour, and different characteristics. The interest in discourse relations has recently crossed from English into other languages such as Turkish (Zeyrek and Webber 2008), Hindi (Prasad *et al.* 2008b) and Chinese (Xue 2005). This led to annotation of corpora with their own inventory of discourse relations and connectives. But neither corpora, nor inventories of discourse relations and discourse connectives have been developed for Arabic.

The existing Arabic corpora mainly include raw text/spoken scripts such as the Arabic Gigaword corpus (Graff 2003), syntactic/morphological annotation (Maamouri *et al.* 2004) (Dukes and Buckwalter 2010; Habash and Roth 2009), lexical and semantic relationships (WordNet) (Elkateb *et al.* 2006). However, there are as yet no theoretical or empirical attempts to annotate Arabic text for discourse features in a large scale study.

As far as we are aware the existing small scale studies of discourse relations for Arabic (Seif, Mathkour and Touir 2005a; Al-Sanie, Touir and Mathkour 2005) do not formalize discourse annotation by collecting potential discourse connectives and relations, nor do they annotate a corpus to be used for automatic annotation for

discourse in Arabic. This lack of studies and resources affects the growing language technology for Arabic in many applications that were improved by using discourse analysis of English.

This thesis is the first large-scale discourse annotation study for MSA, using newswire texts. The study claims that:

- Arabic uses explicit connectives frequently to link discourse units. This is especially true for newswire texts, due to genre conventions. Therefore, it is very important, for Arabic discourse processing, to annotate explicit connectives manually and automatically.

- Arabic has a great variety of discourse connectives with a wide range of syntactic types such as conjunctions, prepositions, nouns, adverbial and prepositional phrases and other expressions (not phrases). The connectives can be clitics attached at the begninnig of words.

- Arabic discourse connectives have a high ambiguity level. The clitics and preposition connectives do not always have discourse function in context. In addition, the connectives can signal more than one discourse relation.

- The annotation principles designed to annotate discourse connectives in English in the PDTB2 (Prasad *et al.* 2008a; Prasad *et al.* 2007b), can be adapted and applied to reliably annotate discourse connectives in Arabic newswire. This allows bilingual comparative corpus-linguistic studies, and also might allow sharing algorithms for discourse connective recognition and disambiguation.

- Machine learning models can be used to identify discourse connectives and relations in Arabic newswire. In particular, the automatic tagging can be used to extract useful syntactic features. This model can achieve good results for text that do not have a manual gold-standard tagging.

- Supervised machine learning models can identify Arabic discourse connectives and their relations with high reliability. This is especially true for discourse connective recognition, which reaches almost human performance and for which high performance is even possible with automatic pre-processing only. This is promising for texts that do not have any manual morphological or syntactic annotation.

**The Objectives**

The study aims:

1. To identify the most common explicit discourse connectives in Modern Standard Arabic (MSA).
2. To design reliable and high-coverage discourse annotation guidelines to annotate explicit discourse connectives, the relations they convey and their arguments.
3. To construct the first reliable Arabic discourse corpus, manually annotated for explicit connectives, their relations and arguments.
4. To develop the first discourse annotation tool for Arabic.
5. To develop algorithms that automatically recognize discourse connectives in the text, and identify the relations the connectives convey.
6. To draw a research plan for future studies and encourage researchers to contribute in this important field.

## 1.2 Contributions of this Work

The main contributions of this first large scale empirical study of Arabic discourse connectives are summarized below.

**The first collection of discourse connectives in MSA.** To the best of our knowledge, our connectives list is the first large scale attempt to identify discourse connectives. We used a combination of manual and automated techniques to analyse a range of MSA texts, to ensure a high coverage for discourse connectives in Arabic news.

**A discourse annotation tool for English and Arabic.** The READ tool has been developed in response to the need to manipulate specific features of Arabic. This is the first tool that can be used to annotate explicit discourse connectives for Arabic and English, by pre-highlighting potential discourse connectives. The annotator makes a decision for each highlighted connective by marking its arguments and relations. The READ tool can also be adapted to work for other languages as long as

they use Unicode format. The tool will be available online free of charge for non-commercial use.

**A novel, reliable, discourse annotation scheme for explicit discourse connectives in Arabic.** The annotation scheme covers guidelines for human annotation of explicit connectives, their relations and their arguments. It is based on annotation principles similar to the English Penn Discourse Treebank, the PDTB (Prasad *et al.* 2008a; Prasad *et al.* 2007b). It has been adapted to fit the characteristics of Arabic.

Reliability of the scheme was tested by human annotation on the newswire corpus Penn Arabic TB Part 1 v.2 (Maamouri and Bies 2004). A large scale human annotation and agreement study has been conducted by two native Arabic speakers, who (i) disambiguated potential discourse connectives, (ii) recognised the relations indicated by the connectives (iii) also marked the argument boundaries. The study measures inter-annotator agreement on all three components. The results were reliable and highly encouraging for the three tasks.

**The first discourse corpus for Arabic: The Leeds Arabic Discourse Treebank (LADTB v.1).** This new corpus has been constructed after manual and automatic post-processing of all types of disagreements in the human annotation (connectives, relations and arguments). The corpus contains 6,328 annotated explicit discourse connectives in 534 files, including 80 connective types and 55 discourse relation types. The current discourse annotation, the first discourse annotation effort for Arabic, annotates all explicit relations that exist in the ATB. However, it does not annotate other coherence devices such as attributions or implicit relations.

**The first computational models for recognising discourse connectives for Arabic**. Several supervised machine learning models using a rule-based classifier were developed to recognize connectives that have discourse usage. The models achieve significant improvements over a baseline, that uses the connective string only. The best models use the gold-standard ATB tokenization and syntactic annotation, and perform well with an extremely high accuracy of 92.4%. Our models also managed to generalise well regardless of individual connectives. Promising results were also recorded from an experiment with a model that assumes no gold standard tokenisation and syntactic annotation.

**The first computational models for discourse connective disambiguation for Arabic.** We developed the first models for relation recognition, using rule-based classifiers. We used features related to the explicit discourse connective and the arguments annotated in the LADTB. The best model achieves an accuracy of 78.8% over a baseline that always assigns the majority relation *Conjunction*, achieving 52.5%. The model also achieves a significant improvement over the baseline of using the connective string only, the latter performing at 77.2%.

## 1.3 Structure of the Thesis

The rest of the thesis is organised as follows.

Chapter 2 reviews the literature on discourse coherence and discourse structure theories. Historical definitions of basic concepts in our work, such as discourse connectives and relations, are presented, in addition to a discussion of previous attempts at human and automatic discourse annotation, and the work done so far on Arabic.

Chapter 3 presents the main characteristics of Arabic that impact on our work and what is available for discourse annotation studies for Arabic. It also describes the methodologies employed to achieve our contributions.

Chapter 4 describes our collection of discourse connectives. This chapter ends with a sizable list of 107 discourse connectives in MSA.

Chapter 5 describes our scheme for annotating explicit discourse connectives. We focus on the modifications we made when adapting the English scheme of the PDTB for Arabic. The full version of the scheme, which was given to the annotators, is attached in Appendix B.

Chapter 6 discusses the proposed discourse annotation tool for English and Arabic, READ v.1. The chapter describes in detail the annotation procedure that we follow in our annotation of discourse connectives.

Chapter 7 describes how we created the first discourse corpus for Arabic, the Leeds Arabic Treebank (LADTB). The human annotation involves three main tasks: recognizing discourse connectives, defining the argument boundaries, and assigning

appropriate relations. We also describe the inter-annotator agreement studies we conducted for each task to verify the reliability of the annotation. The gold standard corpus was derived after automatic and manual resolution of the disagreements. The chapter also presents a statistical analysis of the gold-standard and ends with a comparison of the two discourse Treebanks, the LADTB and the PDTB, as both were created using similar annotation principles.

Chapter 8 proposes supervised machine learning models to automatically detect discourse connectives and their relations. The rule-based classification produces results significantly better than good baselines for both tasks, using features including surface-based, tagging and parsing features. At the end, the chapter summarises our error analysis and discusses suggested features and ideas for further computational work in discourse processing for Arabic.

Chapter 9 concludes the thesis. A summary of the work is presented in addition to the reflections on decisions taken in the study. The chapter also draws some directions for further discourse studies for Arabic.

## 1.4 Notation Conventions

Examples in this thesis are presented according to the following conventions: (i) explicit discourse connectives are bold-faced and underlined, (ii) the text span which is introduced by the discourse connective and expresses an *abstract object* (Arg2) is marked in bold and colored in yellow, (iii) the text span which expresses the other *abstract object* (Arg1) is colored in blue (and marked in italics in the English translation). The examples of non-discourse annotation would not follow these conventions.

Arabic examples in all sections of the thesis are given in a four lines format: (1) an Arabic text (read right-to-left), (2) a left-to-right transliteration per token, (3) a gloss of each token under the transliteration tokens, and (4) a freer standard English translation (to be read from left to right). The first and last lines will show our annotation conventions of the discourse connective, Arg1 and Arg2.

Ex. 1-2 shows an example of our convention of the examples used throughout the thesis. For long examples, line 2 and 3 (transliteration and gloss) might be split into another two lines. Note that for a technical reason, Arabic clitic connectives are sometimes marked in Arg2.

**Ex. 1-2**

سيفعل دور الحكومة **<u>في حال</u>** **انتصار الجيش الأمريكي في العراق.**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| syfEl | dwr | AlHkwmp | fy | HAl | AntSAr | Aljy$ | Al>mryky | fy | AlErAq |
| Will be activated | role | governme nt | in | case | win | army | American | in | Iraq |

*The role of government will be activated* **<u>if</u>** **the American army wins in Iraq.**

# Chapter 2

# Literature Review

## 2.1 Introduction

*Discourse* usually refers to a form of written text or spoken language used to communicate ideas or beliefs to be recognised by the hearer/reader (Asher a, 2005b). People use this language as part of more complex social events, for instance, in specific situations such as encounters with friends, a phone call, a job interview or when writing or reading any kind of article. The concept of discourse deals with three dimensions (Halliday and Hassan 1976; Dijk 1997): (a) language use, (b) communication of beliefs, and (c) interaction in social situations. Given these dimensions, it is not surprising that several disciplines are involved in the study of discourse including: linguistics, psychology (study of beliefs), social sciences (analysis of interaction in social situations), and computational linguistics (to enhance language technology).

Discourse is not just a random sequence of sentences and clauses; rather, it is a coherent, understandable text for the reader or the hearer. In the last two decades, discourse studies have tended to agree on the notion that discourse has a genre-based structure which formalizes how discourse is constituted; thus the structure of academic writing/speech differs from that of story, political, or news texts. The structure is taking into account lexical items, grammatical and morphological features, and semantic and pragmatic features such as intention and attention of propositions and the relations between them. Consequently, discourse studies in computational linguistics attempt theoretically to specify the relationships between the discourse units in a way that can be applied empirically in language applications such as text generation, summarization, argument evaluation, machine translation, speech recognition, essay scoring and question answering systems (Taboada and

Mann 2006a); (Marcu, Lynn and Maki 2000); (Marcu 2000c); (Hovy 1993)..etc). For example, automatic text generation systems benefit from recognising the structure of discourse by applying suitable paragraphing or segmentation, correct punctuation and cue phrases between the text parts (sentences and clauses) in order to generate a coherent discourse.

This chapter provides an overview, from a computational linguistic view point, of discourse, its properties, and theories of how it is constructed, directed by the theme of this study which focuses on a critical discourse coherence device: *discourse relations*. The properties of discourse are reviewed in Section 2.2. Types and properties of discourse relations are described in Section 2.3. Details of discourse connectives are discussed in Section 2.4, as the study concentrates on explicitly signalled discourse relations. The common theories of discourse structure are reviewed in Section 2.5. The next two sections 2.6 2.7and 02.7 present the potential data resources, annotation tools, and applications for identification of discourse connectives and relations. Then, the automatic attempts for recognising discourse connectives and disambiguating their functions for English are reviewed in Section 2.7. The chapter ends with a summary of what relevant to our study.

## 2.2 Properties of Discourse

### Discourse Cohesion

The concept of discourse structure is the answer to the question: *What makes a discourse cohesive/coherent?*[1] In the late 20th century, linguists such as (Halliday and Hassan 1976) (hereafter, H&H) began to express cohesion through the lexicogrammatical system of the language (grammar and vocabulary). There are five types of cohesion associated with grammatical and lexical elements: (i) *reference* cohesion, when elements express referential identities via anaphora such as the pronoun in Ex. 2-1 (a). (ii) *substitution* cohesion, a replacement of one element by another such as *one* to be replaced by *axe* in Ex. 2-1 (b). (iii) *ellipsis* cohesion, a replacement of elements by nothing. The text is still understandable from prior

[1] Cohesion (adj. *cohesive*) and coherence (adj. *coherent*) are both properties of text related to the understanding of the whole text in a logical manner. The distinction between the two is not always clear. However, some linguists such as Yeh (2006) have identified text coherence as the fact that a particular text is coherent and sensibly understood whether or not it has cohesive devices.

elements, such as in the *nominal ellipsis* in Ex. 2-1 (c), (iv) *lexical* cohesion, as the reiteration/repetition of the same element via s synonym or hyponym. (v) *conjunction* cohesion where propositions in discourse are systematically related to prior propositions using lexical items (e.g. coordinating and subordinating conjunctions such as *or* and *but*, adverbials such as *besides,* and prepositional phrases such as *in contrast,* see Ex. 2-1 (d)). The fifth type of cohesion is the sole source of discourse relations, the concern of the presented study.

**Ex. 2-1**

(a) Wash **six apples**. Put **them** into a dish. (Reference, H&H, p.3)
(b) My axe is too blunt. I must get a sharper **one**. (Substitution, H&H, p.9)
(c) Would you like to hear another verse? I know **twelve more**. (Ellipsis, H&H, p.143)
(d) Mary won't come to school. **Because** she is not very well. (Conjunction)

Cohesion, as defined by H&H, has no constraints on theoretical locality, and on how many and what parts of the text can be linked (Webber 2006). However, H&H rejected explicitly any notion of structure in discourse in many places in their book, for example:

"*Whatever relation there is among the parts of a text- the sentences, the paragraphs, or turns in dialogue- it is not the same as structure in the usual sense, the relations which links the parts of a sentence or a clause*."

(Halliday and Hassan 1976, p.6)

## Bases for discourse structure

Webber and her colleagues, in (Webber, Egg and Kordoni 2011), specified several bases of structure and organisation of a text which had been studied in the literature. Firstly, discourse is structured by *entities under discussion*; thus a sequence of expressions that refer to the same entity can make an *entity chain* (this corresponds to H&H refential cohesion). The movement in entity chains presents a change in *topics* of the text segments. These topic changes mostly follow a second base of structure, the so called, *topical structure*. This structure is understood when defining the question/s that each part of the text addressed (which might be expressed by several sentences or paragraphs), lexical cohesion in each part highlights the topic. Thirdly, people in each field tend to use similar functional structures for their writing, which leads to what is called *genre-specific convention*. This convention represents the functions of different parts in the text. For example, the articles in Wikipedia about chemical elements should display a similar structure.

The last basis of discourse structure discussed by (Webber, Egg and Kordoni 2011) is *cohesion relations*, which are also called *discourse relations* in the literature (Moser and Moore ; Webber *et al.* 1999; Hutchinson 2004a) or *rhetorical relations* (Mann and Thompson 1988; Marcu 2000a; Asher 1993b; Hovy and Maier 1993). These relations link either the *content* of text segments (informational discourse relations), or the *speaker's intention* in the segments (intentional discourse relations). The former are the main focus of the current study for Arabic. Further explanation about discourse relations is presented in the next section.

## 2.3 Discourse Relations

It has been argued in the early studies of discourse, such as by Hobbs (1985), that most writers point out the existence of cohesion relations and list some of them but without a complete theoretical justification or framework. However, studies of discourse over the last three decades did formalize the concepts of common discourse relations and classified them into different categories such as Mann and Thompson (1986); Hovy (1988); Hobbs (1985) and Knott (1996). They dealt with a set of important questions regarding discourse relations such as: what exactly do the discourse relations relate? How many discourse relations are allowed to relate two segments? Can we define a standard definition for each discourse relation? Should the segments to be linked be adjacent? or non-overlapping? Is it permissible to cross the dependencies in discourse? What are the lexical items that signal discourse relations? What is the best structure to be constructed using these relations?

It is presumed in the literature that primary discourse segments are *clauses/sentence*s that express abstract entities such as events, facts or propositions (Marcu 1999b; Webber *et al.* 1999; Hovy and Maier 1993; Asher 1993a). A longer *text span* can be constituted when two discourse segments are discovered to be linked by one or more discourse relations. This is the key for building a structure of the whole discourse recursively (Hobbs 1985), although theories differ in formalizing the definition of discourse relations as different targets are desired. The following sections give an overview of discourse relations, their types and features, followed by brief descriptions of theories of discourse structure.

### 2.3.1 Intentional vs. Informational Relations

There are two types of relations, namely *intentional* (presentational) and *informational* (subject-matter) relations. The informational relations are semantic relations that can be recognized by a reader/hearer to *relate different content or meaning of text segments*. These segments represent abstract objects such as propositions, facts, events, or situations to be arguments for such relations (Asher 1993a). In Ex. 2-2, sentence 1 expresses an event; Jack gave Sarah a red rose, and sentence 2 expresses the writer's opinion, while sentence 3 presents a fact that the colour red indicates love. A reader can understand this discourse as that the argument in (2) gives a *reason* for the argument in (1), and the argument in (3) *elaborates* the writer's opinion and the conclusion in (2) that there is a love relationship between Jack and Sarah. Other examples of informational relations are Elaboration, Causal, Condition and Summary (Nicholas 1995).

**Ex. 2-2**

1) Jack gave Sarah a red rose.
2) He loves her so much.
3) The red colour often indicates love.

On the other hand, the *intentional relations* relate intentions or discourse segment purpose (DSP). The segmentation of discourse here is based on grouping the text/dialogue according to different intended purposes; that the writer/speaker wants to enable the hearer/reader to perform some action, or to increase his belief in some proposition (Moore and Paris 1993). The DSPs are the basic components of the *intentional discourse structure* as defined in (Grosz and Sidner 1986). The *intentional* relations are not limited to mere reader recognition; they can influence the reader. For example, there is a Justification relation between the two segments in Ex. 2-3 which increases the reader's inclination to accept what the writer asserts.

**Ex. 2-3**

Dr. John is serving a 7-year jail sentence for medical errors. Two nurses saw him mixing up drugs with names that sound alike.

Therefore, the literature proposed different taxonomies of relations which use one or both types of relations. For example, only two intentional relations are allowed to construct a discourse structure in the Intentional Discourse Model by Grosz and Sider (1986). On the other hand, Rhetorical Structure Theory (RST) by Mann &

Thompson (1987) used both informational and intentional relations but does not allow for more than one representation for a discourse. Later, (Moore and Pollack 1992) discussed the possibility of having two levels of representations (one informational and one intentional) for the same discourse in the RST framework. In fact, Mann and Thompson evaluated such potential ambiguity by considering only the *intentional representation*, since the intentions are what most directly express the speaker/writer's purpose For example, both Evidence (intentional relation) and Cause (informational relation) are applied for the relation between the two segments in Ex. 2-4. As a result, RST would consider only the Evidence relation. Moore and colleagues argue that a complete model of discourse must maintain both levels of relations (Moore and Pollack 1992). Section 2.5 will provide more details of different discourse structure theories.

**Ex. 2-4 (Moore and Pollack 1992)**

a) George Bush supports big business.
b) He is sure to veto House Bill 1711.

### 2.3.2 Explicit vs. Implicit Relations

It was discovered in early studies of discourse that discourse relations are often signalled explicitly for more readability using lexical elements called cue phrases, discourse markers (Marcu 2000c; Walker and Moore 1997; Fraser 1999; Schourup 1999) or *discourse connectives* (Webber, Knott and Joshi 1999; Xue 2005). The latter term is preferred in this thesis. Fraser (1999) categorises discourse connectives as conjunctions (*and, or, but*), adverbs (*because, instead* and *since*) and prepositional phrases *(in contrast)*. The examples, in Ex. 2-5, show different discourse connectives in different locations in the sentence. Section 2.4 presents a detailed review of discourse connectives, as the study focus is on discourse relations explicitly signalled in Arabic.

**Ex. 2-5 (Fraser 1999, p.8, p.9 and p.10).**

a) We left late**. However**, we arrived home on time.
b) Jack played tennis**, and** Mary read a book.
c) We don't have to go. I will go, **nevertheless**.
d) **While** she is pregnant, Martha will not take a plane.

It is true that not all discourse relations are explicitly signalled in the text; in many cases there are no lexical elements identifying the discourse relations between

arguments. The relations in Ex. 2-3 and Ex. 2-4 are not signalled and the discourse is still meaningful. That is because *a discourse should be as informative as required but no more informative than required* (Knott and Sanders 1998; Knott 1996). The discourse producer therefore should think about the features of a relation that can be easily inferable by the receiver from the context or his background without using extra lexical items such as connectives to avoid redundancy.

Such inferred relations are very frequently used and they are considered in (Wolf and Gibson 2005; Taboada and Mann 2006b; Webber *et al.* 2003; Prasad *et al.* 2008a; Miltsakaki *et al.* 2005a; Hobbs 1985). Recently, Miltsakaki and colleagues (Miltsakaki *et al.* 2006; Prasad *et al.* 2008a) annotated inferred relations (called here *implicit relations*) in the Penn Discourse Treebank by inserting the most suitable discourse connectives, called *implicit connectives.* For example, a Causal relation is inferred in Ex. 2-6, between the arguments *raising cash positions to record levels* and *high cash positions helping to buffer a fund,* even though there is no explicit connective expressing this relation. A label (Implicit = BECAUSE) is inserted in the PDTB annotation. Note that 53% of all discourse relations annotated in PDTB2 (34683, the explicit plus implicit relations only) are explicit[2] while 47% are implicit relations. However, this distrbuation of implicit and explicit relations does not necessary reflect the distrbuation in English news, as not all explicit connectives were in the scope of the PDTB annotation. Moreover, news corpora in different languages such as Arabic may also have different distrbuations.

**Ex. 2-6 (Prasad et al. 2007, p.22)**

> But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels*. (Implicit = BECAUSE) **High cash positions help buffer a fund when the market falls.** (WSJ text 0983)

### 2.3.3 Adjacency and Cross-dependency

There is an important debate among researchers centring on whether discourse relations link only non-overlapping adjacent text spans. Applying such an adjacency constraint in discourse representation, such as is done in RST (Mann and Thompson 1987), raises problems of cross-dependency of relations. As an example, in Ex. 2-7 and the corresponding Figure 2-1 (i) it is clear that clause 3 *he went to bed* is linked

[2] These relations only use discourse conectives in the PDTB list, and do not inlude AltLex annotations which use other lexical expressions to link adjacent arguments explicitly.

via a Contrast relation to the non-adjacent clause 1 *John didn't go to the party*. However, clause 1 is also linked to the adjacent clause 2 *he was tired* via a Causal relation.

**Ex. 2-7**

John didn't go to the party$_{c1}$, he was tired $_{c2}$. **Instead**, he went to bed $_{c3}$.

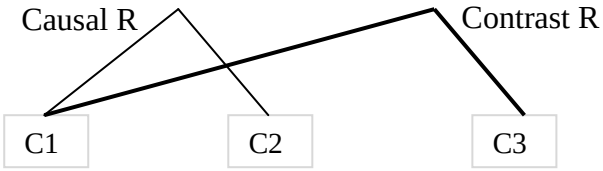



**Figure 2-1: Adjacent and non-adjacent clauses in Ex. 2-7 linked via two discourse relations.**

The cross-dependency is basically caused by crossing multiple semantic relations between non-adjacent segments (Webber 2006). Samples of these crossings are shown in Wolf and Gibson (2005), who found a large number of crossed dependencies of nodes with more than one parent in the RST-tree representation of some discourse. They proposed to use an undirected graph – a chain graph – to tackle this problem instead of trees as in RST to allow multi-parents nodes and cross dependency relations. Some samples of the crossed-dependency relations are shown in Figure 2-2.

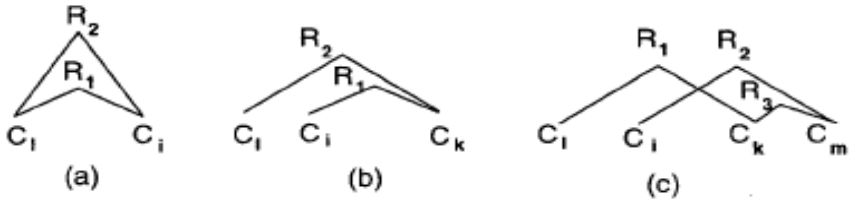



**Figure 2-2: Multiple semantic links (Rj) between discourse clauses (Ci) (Webber et al. 1999). The relations in (a) link the same discourse clauses; (b) are back to different discourse clauses; (c) are back to different discourse clauses, with crossing dependencies**

## 2.3.4 Taxonomies of Discourse Relations

A discourse relation taxonomy is a hierarchical structure that expresses hyponym relationships among a variety of coherence relations, with different levels depending on the theory applied (Marcu 2000a; Mann and Thompson 1988; Hobbs 1985; Hovy

and Maier 1993; Marcu 2000c). Hovy (1990) collected the discourse relations defined in the literature and classified them into a hierarchy of increasingly specific semantic relations. He argued also that discourse relation taxonomy is open-ended in one dimension and can be expanded with other relations if such are discovered later.

Most theories of discourse structure tend to use similar relations. However, the terminology for discourse relations is not standardised and that it is not always easy to map different terminologies into each other. Mann and Thompson (1988) posit 24 relations that are classified into: informational relations (e.g. Elaboration, Circumstance, Cause, Restatement) and intentional relations (e.g. Motivation, Background, Justify, Concession). They also proposed another classification based on where the locus of effect is (nucleus or satellite). Further details are discussed in Section 0.

Grosz and Sidner (1986) restricted their relation taxonomy to only two structural relations, *dominance and satisfaction-precedence* in their intentional-level organization. In contrast, Hovy and Maier (1993) proposed a comparison study and merged the 400 proposed relations in the literature into 70 frequent relations in new definitions; a sample is shown in Figure 2-3. While the majority of taxonomies of coherence relations are theory or task dependent, a new theory-neutral approach in discourse annotation in the PDTB project (Prasad *et al.* 2008a) defined 57 relations (called *senses*) using concepts from logic in a hierarchical manner, under four main classes: Temporal, Contingency, Expansion and Comparison; with a possibility of combining multiple relations from different levels as appropriate. Their relations are shown in Figure 2-4.

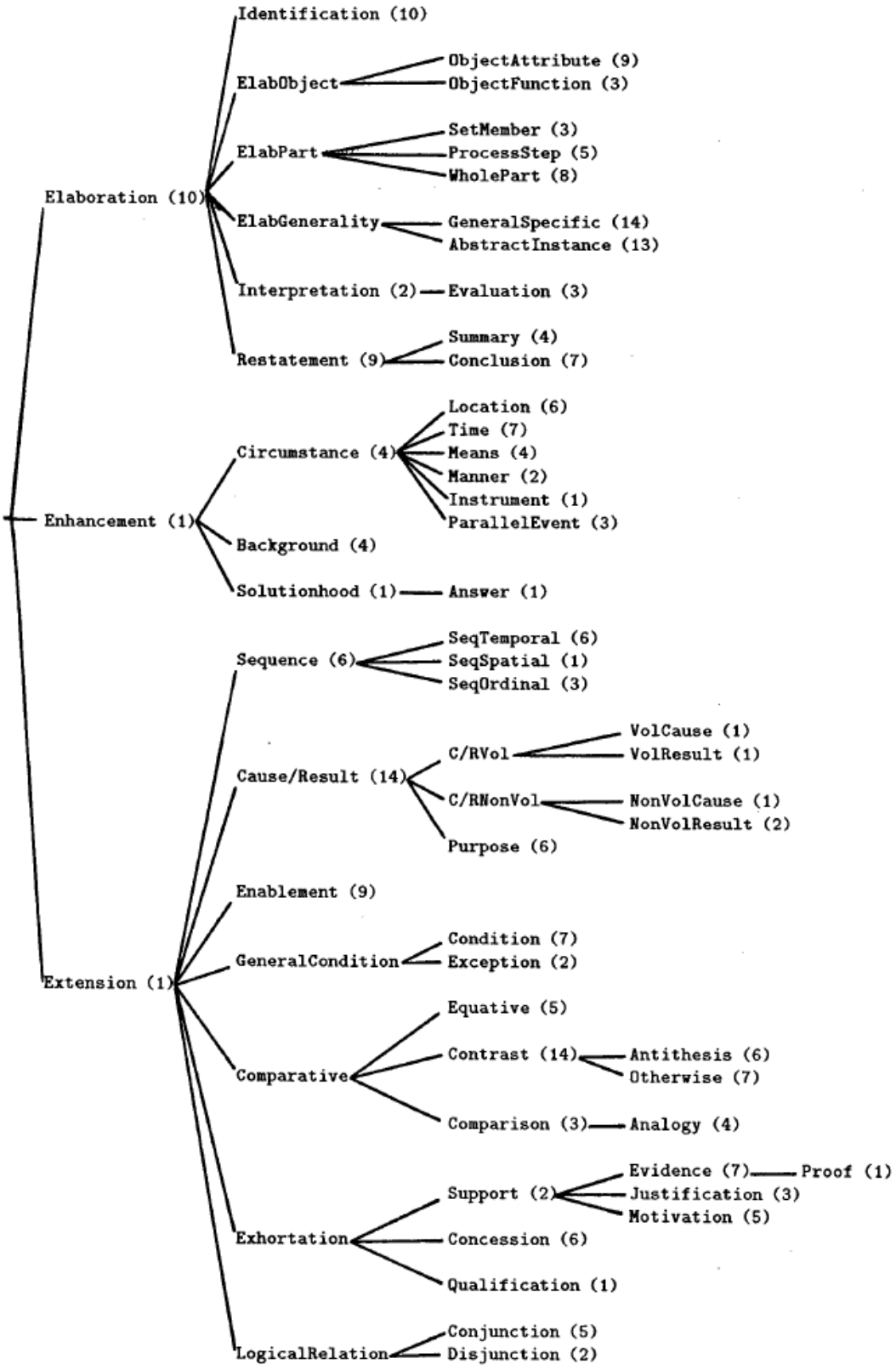



**Figure 2-3: A hierarchy of discourse relation taxonomy (Hovy 1990)**

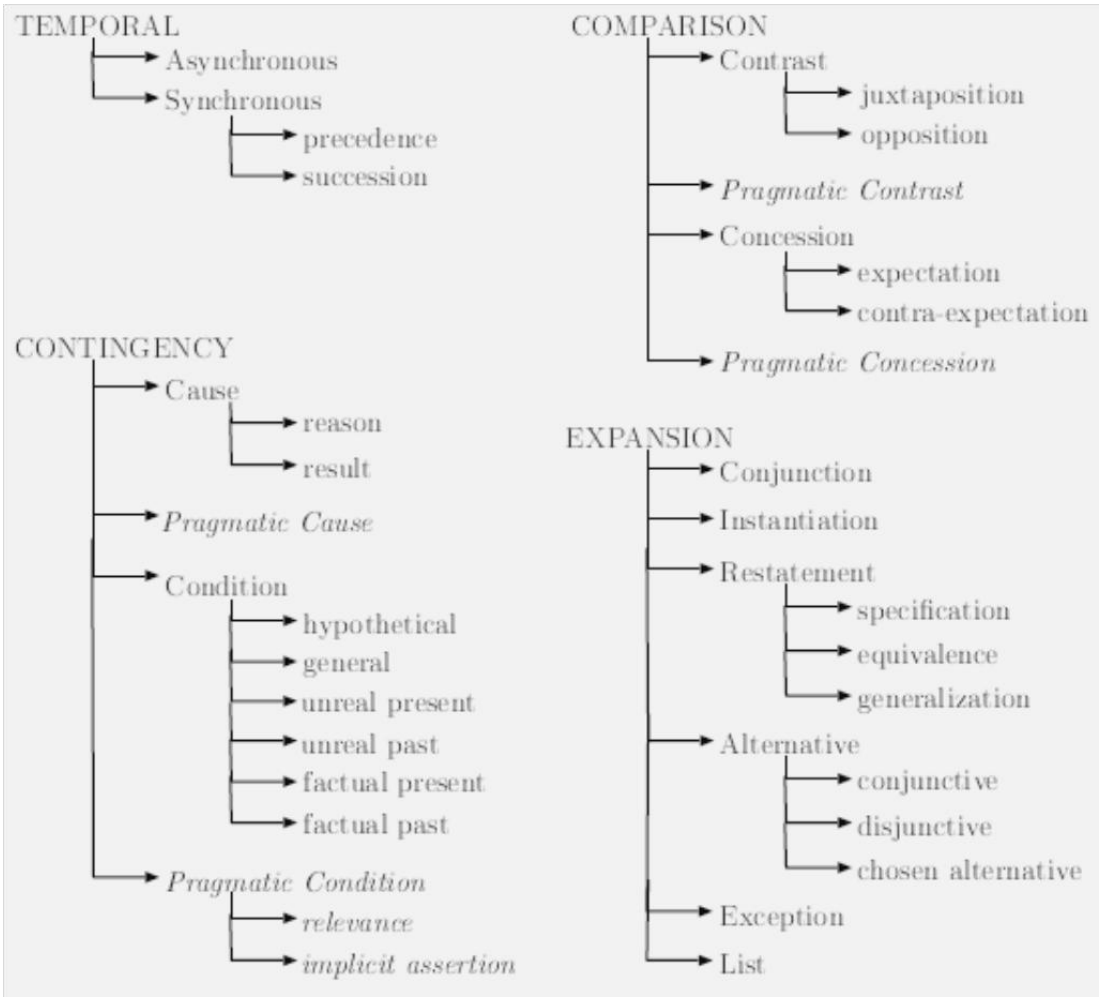



**Figure 2-4: The relation hierarchy of the PDTB (Prasad *et al.* 2008a)**

We use relation taxonomy similar to the PDTB in the current study of discourse annotation for Arabic, making adaptations as required. This decision was motivated by this taxonomy being theory-neutral and due to it covering informational as well as pragmatic discourse relations. In addition, a hierarchical structure of the fine-grained taxonomy allows for a more flexible annotation whose reliability can be tested on fine and coarse-grained levels. It also allows addition of new relations at any level by inserting a new branch in an appropriate position. The taxonomy is also mostly language independent; it has already been applied to English, Chinese, Hindi and Turkish (Xue 2005; Zeyrek and Webber 2008; Prasad *et al.* 2008a; Prasad *et al.* 2008b).

## 2.4 Discourse Connectives

The interest in studies of discourse connectives has increased rapidly in computational linguistics as they are recognised as informative, explicit cohesion devices used to tie parts of discourse together. A variety of labels were used in the literature for words with a similar function to that of the discourse connectives: cue phrases (Knott and Dale,1994), discourse connectives (Blakemore, 1987, 1992), discourse operators (Redeker, 1991), discourse particles (Schorup, 1985), discourse signalling devices (Polanyi and Scha, 1983), pragmatic connectives (Stubbs, 1983), pragmatic markers (Fraser, 1988), semantic conjuncts (Quirk et al., 1985), and sentence connectives (Halliday and Hasan, 1976). This section explores the role of discourse connectives in the text, the arguments they relate, and their grammatical status from a computational linguistic view point. These charactristies of discourse connectives cross languages with slight language-dependent changes such as more or less grammatical status. This is also true for Arabic, this study provides a large collection of discourse connectives and their features.

Discourse connectives have two distinct functions as distinguished by Cohen (1984): (i) enabling faster recognition of discourse relations by the reader (the hearer), and (ii) allowing the recognition of discourse relations which could not be inferred in the absence of a connective. Formalising the connective types and the potential arguments they relate might differ, depending on the task and genre of the study. In computational linguistics, discourse connectives are considered as important explicit, frequent indicators for discourse relations, reducing ambiguity in establishing discourse relations such as in (Mann and Thompson 1987; Hobbs 1985; Fraser 1999; Hovy and Maier 1993; Marcus, Santorini and Marcinkiewicz 1993; Sanders 1992; Miltsakaki *et al.* 2006; Pitler *et al.* 2008).

Redeker (1991), who worked on speech, defined a discourse connective (discourse operator) as:

"*a word or phrase that is uttered with the primary function of bringing the listener's attention to a particular kind of linkage of the upcoming utterance (clausal unit) with the immediate discourse context*" (Redeker, 1991, p.1168)

The discourse connectives in (Redeker, 1991) include clausal indicators of discourse structure (e.g. *let me tell you a story*), deictic expressions (e.g. *now*, *here* and *today*), and anaphoric pronouns.

Blakemore (1987, 1992) proposed that discourse connectives have a procedural meaning, which consists of instructions about how to manipulate the conceptual representation of the utterance. On a different note, Asher (1993) stated that discourse connectives are *the adverbials including only those which convey a relation between two abstract objects such as events or states*. Not too distant from the previous explanation, Fraser defined discourse markers as

"*a class of lexical expressions drawn primarily from the syntactic classes of conjunctions, adverbs, and prepositional phrases. With certain exceptions, they signal a relationship between the interpretation of the segment they introduce, S2, and the prior segment, S1*" (Fraser 1999, p.950).

These definitions and others share two factors, (i) words or expression that relate two abstract objects, and (ii) one of the two abstract objects is introduced by the connective. Therefore, our preferred definition for Arabic discourse connectives in the current study includes these factors following (Miltsakaki *et al.* 2006): any lexical expressions that relate discourse segment with any prior discourse segment where both segments express abstract objects such as events, facts, propositions and beliefs.

From the definition, it is clear that it cannot be determined without context whether a potential connective such as *while*, *and* or *until* has indeed discourse usage in a given text. For example, the potential connective *while* in Ex. 2-8 is a discourse connective in (a) but not in (b). Similarly, the potential connectives in the examples in Ex. 2-9 are not discourse connectives, as they do not relate two abstract objects.

**Ex. 2-8**

**a)** Schools in the north tend to be better equipped, **while** those in the south are relatively poor. (BNC) **(a discourse function)**
**b)** I have not seen you for a **while**. **(not a discourse function)**

**Ex. 2-9 (non-discourse usage of potential connectives)**

a) Judy **and** Sara went to the cinema last night.
b) We will walk **until** the sunset.
c) I'm not available Tuesday to Friday **except** Thursday morning.

### 2.4.1 The Order of Discourse Connectives and their Arguments

We call the two discourse segments, a discourse connective relates, its arguments. Studies of discourse processing consider arguments to be non-overlapping text spans of clauses or sentences (Polanyi 1988; Grosz and Sidner 1986; Webber and Prasad 2006; Webber 2006; Miltsakaki *et al.* 2004; Mann and Thompson 1987). In addition, these arguments can be more that one sentence/clause that express a proposition with other necessary complements (Prasad *et al.* 2008a). The PDTB annotation (Prasad *et al.* 2008a) also considers nominal expressions/noun phrases as valid arguments when they express abstract objects such as nominalizations that express an eventuality.

Fraser (1999) represented a range of canonical forms to specify the position of a discourse connective DC and its arguments Arg1 and Arg2 in texts, such as <Arg1. DC+Arg2>. <Arg1, DC+Arg2>, <Arg1. Arg2+DC > and < DC+Arg2, Arg1 >. Ex. 2-10 shows examples of those forms. Discourse connectives in English may also occur in the middle of an argument. For instance, the connective *for example* occurs in the middle of Arg2 in Ex. 2-11. We determine the possible orderings for discourse connectives and arguments for Arabic in Section 5.2.

**Ex. 2-10 (Fraser 1999, p.9 and p.10)**

| | | |
|---|---|---|
| a) | We left late**. However**, we arrived home on time. | <Arg1. DC+Arg2> |
| b) | Jack played tennis**, and** Mary read a book. | <Arg1, DC+Arg2> |
| c) | We don't have to go. I will go, **nevertheless**. | <Arg1. Arg2+DC > |
| d) | **While** she is pregnant, Martha will not take a plane. | < DC+Arg2, Arg1 > |

**Ex. 2-11 (Williams and Reiter 2003. p.1)**

Sometimes you did not pick the right letter. You did not, **for example**, click on the letter ‘d’.

### 2.4.2 The Grammatical Status of Discourse Connectives

Discourse connectives do not fall into a unique syntactic category (Fraser 1999; Webber and Prasad 2006; Prasad *et al.* 2008a; Taboada 2006). There are three main syntactic categories of discourse connectives in English: (i) coordinating or subordinating conjunctions, (ii) adverbials, (iii) prepositional phrases (Fraser 1999; Asher 1993a). However, not all conjunctions, adverbials and prepositional phrases always function as discourse connectives as they also need to relate abstract entities in discourse.

**Coordinating conjunctions**. Two clauses can be joined by a *coordinating conjunction* such as *and, or* and *but* (see Ex. 2-12 (a)). Frequent functions of these connectives are the discourse relations Conjunction, Alternative and Contrast, respectively.

**Subordinating conjunctions**. Those conjunctions introduce clauses that are syntactically dependent on the main clause. Examples are *because*, *although*, and *if*, which express discourse relations Causal, Contrast and Condition respectively. An example is given in Ex. 2-12 (b).

**Ex. 2-12**

a) Jack played football**, and** Mary read a book. **(<Arg1, DC+Arg2>, Conjunction)**
b) **Although** she joined the company only a year ago, she's already been promoted twice. **(< DC+Arg2, Arg1>, Contrast)**

**Adverbial connectives.** Sentence-modifying adverbs can express a discourse relation between two abstract entities (Miltsakaki *et al.* 2006). Examples are *therefore* and *then* which express discourse relations such as Causal and Conditional relation respectively in Ex. 2-13 (a, b).

**Prepositional phrases.** Such as *in contrast* and *as a result* can also express discourse relations. Contrast and Consequence relations are expressed respectively in Ex. 2-13 (c, d).

Discourse connectives can consist of two parts. These are called *paired connectives* where each connective's part introduces an argument of the connectives such as the paired connective *if…then* in Ex. 2-13 (b).

**Ex. 2-13**

a) John did not finish the report. **Therefore**, we will postpone the meeting.
b) **If** you want to answer the questions, **then** you have to read the book.
c) Math lectures are understandable. **In contrast**, I find Chemistry lectures are quite hard.
d) Peter has not studied very well. **As a result**, he failed in the exam.

Although, the syntactic classification of connectives so far was for English connectives, they are generalizable to other languages such as Hindi and Turkish (Prasad *et al.* 2008b; Zeyrek and Webber 2008; Oza *et al.* 2009). However, they are not necessarily the only syntactic categories possible for connectives in all languages. Some extra syntactic categories of discourse connectives in English either not yet annotated as connectives (such as prepositions) or not allowed (such as

nouns). For example, in Hindi (Prasad *et al.* 2008b) included *sentential relatives* such as (*which of reason/because of which*) as valid discourse connectives. In this study, we collected potential discourse connectives for Arabic (Chapter 4) and formalized their syntactic categories. In addition to the English categories, we found that prepositions and nouns can relate two valid abstract entities in Arabic.

### 2.4.3 Substitutability of Discourse Connectives

More than one discourse connective can signal the same discourse relation. As a result, discourse connectives can be swapped without affecting the structure of the discourse (Hutchinson 2005a; Knott 1996). Similarity and substitutability of discourse connectives has been studied early when Knott (1996) built up a taxonomy of 150 discourse connectives based on features of discourse relations that use discourse connectives as indicators. He addressed a set of features between discourse connectives that indicate similar discourse relations. The two connectives are: (i) *synonymous* when the two phrases can be used in the same context and have exactly the same features; (ii) *exclusive* when the phrases cannot be used in the same context without obvious change in the meaning and structure; (iii) *hypernym and hyponym* when one of phrases ph1 can be used in the context of the other phrase ph2 but ph2 cannot be used in all contexts of ph1; (iv) *contingently substitutable* when both phrases ph1 and ph2 can be substituted in some contexts of ph1 and ph2, but not in all contexts of ph1 and ph2. The four substitutability relationships are demonstrated in diagrams a, b, c and d respectively in Figure 2-5.

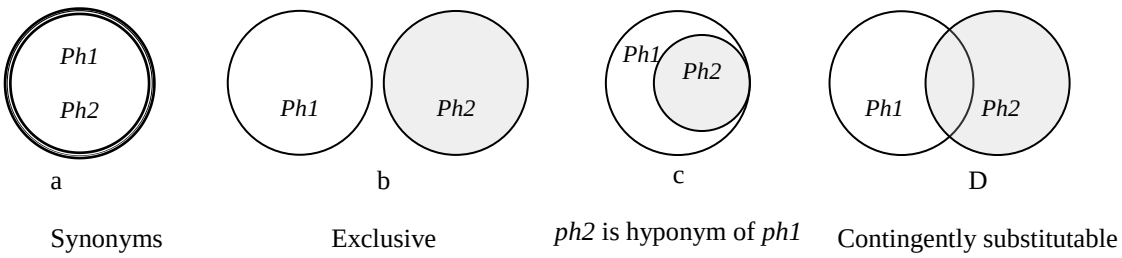



**Figure 2-5: Venn diagrams of different substitutability relationships between two discourse connectives. (Hutchinson 2005b) with a slight modification.**
***Ph1= the first phrase, Ph2= the second phrase.***

### 2.4.4 Ambiguity of Connectives

Discourse connectives can be ambiguous in two ways. First, a potential discourse connective can have discourse usage or not in a given context. For example, some discourse connectives in English are almost unambiguous in this respect such as many adverbial connectives: almost all their occurrences are discourse connectives (Pitler *et al.* 2008). Nevertheless, some potential connectives, such as *while* and *since* and *conjunctions* might have only sentential usage, or discourse usage as well in a given context. The syntactic categories of the potential connective and the words around it, and their positions in the sentence might help in distinguishing these functions in English (Pitler *et al.* 2008). For example, the conjunction *and* is not a discourse connective when it joins non-abstract nouns such as in (Mary *and* Jack left the country).

Second, discourse connectives might be ambiguous in terms of their interpretations, as they can signal more than one discourse relation. For example, the discourse connective *since* in Ex. 2-14 signals a temporal relation in (a), a causal relation in (b), and both relations in (c).

**Ex. 2-14**

a) This mark is the best ever mark I got **since** the exams were conducted in our department. (Temporal)
b) The suspect man in the next door was arrested **since** he stole a car. (Causal)
c) She could not sleep **since** her father died. (Temporal and Causal)

In fact, a part of this ambiguity problem is strongly related to the definitions of discourse relations. For example, the ambiguous connectives in one relation inventory (e.g. RST) are not necessarily ambiguous in another inventory (e.g. SDRT). For example, SDRT does not distinguish Explanation and Evidence, and therefore, the connective *because* is ambiguous in RST, but it is unambiguous in SDRT (Sporleder and Lascarides 2006). One to one mapping between discourse connectives and the discourse relations they signal, such as in RST, does not represent all potential discourse annotations (Taboada and Mann 2006). In current study for Arabic we tackle ambiguity problem in our manual annotation (Chapters 5 and 7) and how this affects on the computational modelling (Sections 8.2 and 8.4.3).

### 2.4.5 Classification of Discourse Connectives

The literature contains many different classifications of discourse connectives, depending on whether the research concentrates on either written and/or dialogue discourse or according to what type of relation they signal. For example, the classification might be based on external/internal textual cohesion (Halliday and Hasan, 1976), cognitive plausibility (Sanders et al., 1992) or substitutability (Knott 1996). In addition, Webber and her colleagues (2003) classified the connectives according to their dependency into either discourse adverbials (including *then, also, otherwise, nevertheless,* and *instead*) and structural connectives between adjacent discourse units (including *coordinating and subordinating conjunctions* and *paired connectives*). In the more recent work on annotating discourse connectives for English in the PDTB (Prasad *et al.* 2008a), 100 distinct discourse connectives were annotated and classified into associated discourse relations.

Some studies dealt with a subset of connectives to acquire their meaning empirically. For example, Hutchinson (2004) used only three features to classify connectives: polarity, veridicality and type; where the latter corresponds to a very coarse-grained set of relations such as Additive, Temporal and Causal.

It is not clear how big the connective taxonomy for Arabic is. Up until now, there has not been a large scale study to collect and classify the discourse connectives for Arabic. The current study will propose the first inventory of Arabic discourse connectives, and a taxonomy for their relations.

## 2.5 Theories of Discourse Structure

Linguists and computational linguists have over the last three decades attempted to produce reasonable generalised theories to represent discourse structure. Theories of discourse structure differ in their focus according to the type of discourse such as written text or dialogue, the type of organization such as intentional organization (speaker's plan) or informational organization (semantic and pragmatic), their background and objectives. The ability to test and apply the theory empirically is an important factor of how representative these theories are. This section discusses popular theories of discourse structure that have impacted on the field and their bases.

Webber (2006) stated that theories of discourse structure such as in RST (Mann and Thompson 1987), Linguistic Discourse Model - LDM (Polanyi 1998), D-LTAG (Webber *et al.* 2003; Webber *et al.* 1999) and GraphBank (Wolf and Gibson 2005) take *constituency* and *anaphoric dependency* as sources of defining their discourse relations. Constituency refers to constructing a part of a text by joining smaller parts, where each part has a specific role or function in the text. Anaphoric dependency refers to dependency relations between words and phrases in that a part of an element's interpretation depends on prior concepts in the discourse context (Halliday and Hassan 1976; Webber 2006).

Before describing each theory, an example of a text and one possible discourse structure derived from it is presented in Ex. 2-15 and Figure 2-6. The discourse consists of propositions in clauses (a, b, c, and d). A Temporal relation obviously exists between propositions 1 (a, b and c) and 2 (d) which is indicated explicitly by the adverbial *then*. Clause (a) states the topic sentence, and clauses (b and c) elaborate on this by breaking it into two subtopics that are discussed in sequence. In addition, a Joint relation links the two clauses (b and c), and is indicated by the conjunction *and*. A reader obviously can recognise these discourse relations between discourse propositions while reading without extra effort.

**Ex. 2-15 (Hobbs 1985, p.1)**

1) **(a)** I would now like to consider the so-called "innateness hypotheses", **(b)** to identify some elements in it that are or should be controversial, **(c) and** to sketch some of the problems that arise as we try to resolve the controversy.
2) **(d) Then**, we may try to see what can be said about the nature and exercise of the linguistic competence that has been acquired, along with some related matters.

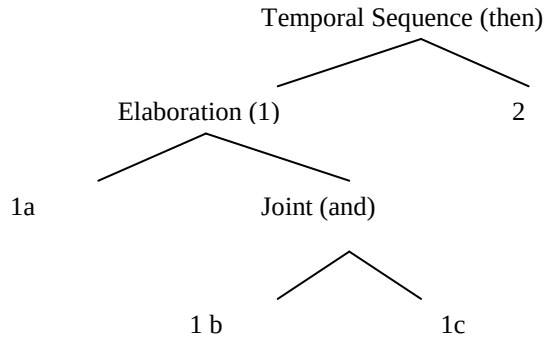



**Figure 2-6: One possible discourse structure of the discourse in Ex. 2-15 (Hobbs 1985, p.2)**

The study presented in this thesis focuses on local relations and does not address any global relations that construct a complete structure for discourse in Arabic.

### 2.5.1 Rhetorical Structure Theory (RST)

RST is a theory of how coherence in text is achieved. It is one of the most popular discourse theories, especially within the area of computational linguistics. RST was developed in the 1980s by a group of researchers interested in Natural Language Generation (Mann and Thompson 1988). Originally, the central tenet of RST is the notion of rhetorical relations (discourse relations), which exist between two *adjacent* and *non-overlapping* text spans (discourse units).

RST considers both informational and intentional relations in its relation taxonomy. However, RST, in fact, takes into account the intention of the writer for all relations by defining two nuclearity levels of text spans: *Nuclei* (N), the most important parts of a text and essential to the writer's purpose, and *Satellites* (S), the elements less important to the writer's purpose. Satellite contributes to the nuclei understanding, but the text is still understood when the satellites have been deleted. Using this principle the discourse relations in RST are divided into: *multinuclear* relations (both spans related by a discourse relations are important for a complete meaning) and *nucleus-satellite* relations. For example, Contrast is a *multinuclear* relation, while Concession is a *nucleus-satellite* relation.

The discourse structure according to RST can be achieved by analyzing the text via linking non-overlapping adjacent text spans recursively using five RHS (Right-hand sisters) *structural schemas* to produce a top-down binary tree structure- RS-Trees (Mann and Thompson 1988). Figure 2-7 displays the five schemas; the arrows in the schemas represent a direction from satellite to nucleus units. Each span, except for the span that contains the entire text, is either a minimal unit or a constituent of another schema application.

RST only allows for a single analysis of a discourse. A judgment must be made in case of ambiguity when more than one applicable scheme between sisters exists. This constraint, along with others such as the stipulation of adjacency between relation arguments, led to heated discussions in the discourse community about the suitability of RST to represent a general organisation of discourse (Moore and Pollack 1992).

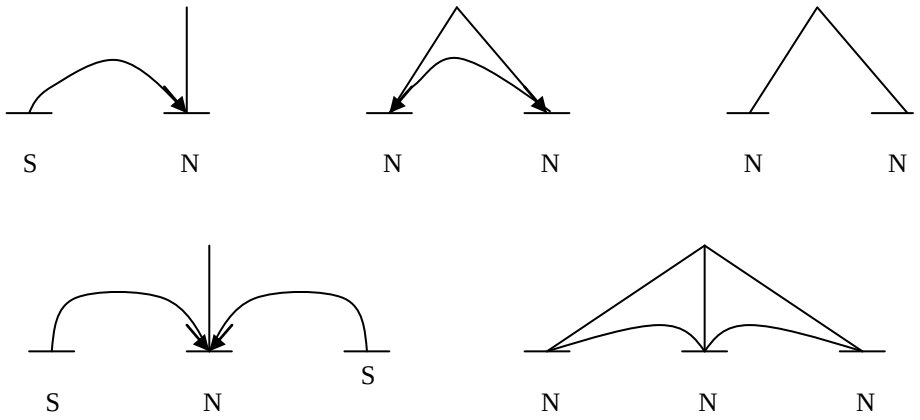



**Figure 2-7: The structural schemas in RST (Mann and Thompson 1988). N = nuclei and S = satellites. The direction of arrows is from S to N.**

Over the years, RST has been adopted for different purposes (Taboada and Mann 2006b; Hovy 1990); (Hovy and Maier 1993). RST was also practically tested via annotation of the RST Discourse Treebank corpus (Carlson, Marcu and Okurowski 2001; Carlson *et al.* 2002; Taboada and Mann 2006a). The corpus has been used in developing language applications such as summarization (Marcu 2000c), question answering (Girju *et al.* 2003), and text generation (Williams and Reiter 2003).

### 2.5.2 Discourse GraphBank Theory: Wolf and Gibson

Wolf and Gibson (2005) present a view of discourse related to RST but rather than analyzing a text as a binary tree structure of discourse spans built recursively via discourse relations between adjacent segments, they represent discourse as a *chain graph* (a graph of directed and undirected arcs between nodes to represent the RST discourse relations between one or more previous, adjacent or non-adjacent discourse segments). In this approach, a text is analysed by grouping the segments into topic and sub-topic segments, linking the non-adjacent segments or groups, if possible, using any of eleven broad classes of binary relations: Same, Condition, Attribution, Cause-Effect, Contrast, Similarity, Example, Expectation, Temporal sequence, Generalisation and Elaboration. This representation allows multiple parents and crossing arcs between nodes. Figure 2-8 shows two representations of the text in Ex. 2-16: one by RST and the other following Wolf and Gibson (W&G). The RST representation does not annotate an Expectation relation between 2-3 and 4-5 in contrast to the graph representation by W&G, because the tree constraint does not allow for crossed dependencies (Wolf and Gibson 2005).

**Ex. 2-16 (Wolf and Gibson 2005, p.18)**

1) Mr. Baker's assistant for inter-American affairs, Bernard Aronson,
2) while maintaining
3) that the Sandinistas had also broken the cease-fire,
4) acknowledged:
5) "It's never very clear who starts what."

On the other hand, there is no guarantee in W&G's approach that whole text segments are linked in one framework, which limits the benefits as no complete structure emerges, especially in computational applications. Wolf and Gibson (2005) also studied how frequent the multiple-parent nodes and crossed dependencies are in 135 texts that were annotated according to their approach. Their results showed that such cases are not rare (12.5% of arcs in a coherence graph have to be deleted in order to make the graph free of crossed dependencies) and cannot be avoided to produce tree structures.

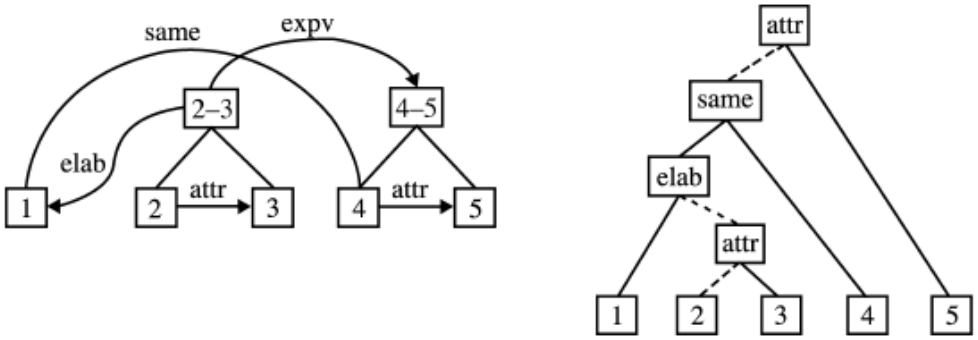



**Figure 2-8: A graph structure by W&G (left) and RST tree structure by Carlson, Okurowski, and Marcu, 2002 (right) for a discourse in Ex. 2-16. elab=elaboration, attr=attribution, expv= violated expectation and same= same segment but separated by intervening discourse segments. Broken lines represent the start of asymmetric/directional relations; continuous lines represent the end of asymmetric coherence relations; symmetric/in directional coherence relations have two continuous lines. Graphs reproduced from (Wolf and Gibson 2005).**

### 2.5.3 The Linguistic Discourse Model (LDM)

The Linguistic Discourse Model -- LDM (Polanyi 1998; Polanyi and Berg 1996; Polanyi *et al.* 2004) -- is a theory of discourse interpretation and parsing to build a structural and semantic representation of text. The main components of LDM are discourse constituent units (DCUs- carrying propositional information such as events, facts and states), and discourse operators (DOs – carrying non-propositional information such as logical operator and connectives). The discourse parsing consists of two parts. First, the discourse units (sentences or clauses) are parsed using traditional syntactic theories. Second, these discourse units are then combined using semantic context-free relations (discourse grammar) into a tree structure. There are only three discourse grammar rules in the LDM:

- Discourse coordination is an N-ary branching rule where all RHS (Right-hand-sister) nodes have the same relationship to the common parent such as a list of elements and narratives.

- Discourse subordination is a binary elaboration relationship between a subordinate node (one sister) and dominant nodes (other sisters). The interpretation of the parent is the interpretation of the dominant daughter.

- Logical or rhetorical relations are derived between RHS sisters in an N-ary branching rule. The interpretation of the parent derives from the interpretation of each daughter and the relationship between them.

Polanyi and colleagues in (Polanyi *et al.* 2004) proposed an implementation of a parser based on the LDM. Nevertheless, LDM is a syntactically informed, semantically driven model, thus adopting this parser to work with other languages is a complex process (Polanyi *et al.* 2004).

### 2.5.4 Intentional Discourse Model: Grosz & Sidner (G&S)

The intentional discourse model concentrates on the role of discourse purpose and the speaker's plan, developed mainly for Task Oriented Dialogue (Grosz and Sidner 1986). Their main claim was

"*discourse is coherent only when its discourse purpose is shared by all the participants (speaker and hearer) and when each utterance of the discourse contributes to achieving this purpose, either directly or indirectly, by contributing to the satisfaction of a discourse segment purpose*" (Grosz and Sidner 1986, p.28).

Discourse structure here is composite of three interacting constituents: *a linguistic structure, an intentional structure,* and *attentional state*. Each component deals with different aspects of the utterances in a discourse.

The *linguistic structure* is a structure of utterance sequences that make up a discourse segment; these utterances have similar roles to that of words in phrases. The interpretation of a linguistic expression in discourse is affected by the discourse segmentation process. G&S pointed out that the availability of some linguistic cues assists in detecting discourse segment boundaries such as *but, yah, and so*. These linguistic markers explicitly indicate changes in the *intentional structure* and in the *attentional state* as well.

Intentional relations between intentions, *discourse segment purposes* (DSPs), are the basic components of *intentional structure*. They also distinguish between intentions that are intended to be recognized and those intentions that are associated with discourse. The discourse segment purpose is always intended to be recognised. Two structural relations are introduced to represent intentional structure of discourse:

*dominance* and *satisfaction-precedence*. Thus DSP1 *contributes* to DSP2, and DSP2 *dominates* DSP1, when the intention DSP1 may be intended to provide part of the satisfaction of DSP2. The dominance relation invokes a partial ordering on DSPs, the dominance hierarchy. Also, DSP1 satisfaction-precedence DSP2 is true whenever DSP1 must be satisfied – recognized- before DSP2. There is no finite list of discourse purposes as there is of syntactic categories.

The third component is the *attentional state*, which contains information about the objects, properties, relations and discourse intentions that are most salient at any given point. The *attentional state* is modelled by a set of focus spaces, defined as:

"*a set of transition rules that specify the conditions for adding and deleting spaces*"
(Grosz and Sidner 1986, p.5)

G&S's theory had an important impact on discourse studies of dialogue. (Litman and Allen 1990) were concerned about the relationship between plan recognition in discourse and the underlying commonsense structures that are necessary to support the discourse. They provided an implementation of discourse structure that originated in G&S's theory. Grosz and Sidner (1986) also argued the compatibility of proposed relations with other rhetorical relations such as Elaboration, Summarization and Justification, which had been investigated in other discourse structure theories. These rhetorical relations incorporate implicitly a form of intentions (the intention to summarize, the intention to justify and so on). As discussed previously in Section 2.3.1, a complete model of discourse structure should maintain both organisation levels (Moore and Pollack 1992).

### 2.5.5 Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG)

Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG) is a lexicalized approach to discourse relations (Webber *et al.* 2003; Forbes-Riley, Webber and Joshi 2006; Webber 2004). The main belief here is that establishing relations between discourse units is based on a similar concept as establishing relations within the clause. LTAG is a tree representation of syntactic and lexical items of part of a text. However, Lexicalization in D-LTAG means that each elementary tree in D-LTAG is anchored by a discourse connective which indicates a discourse relation, and links other trees

for other parts of the text (arguments), using two language independent composition operations, namely substitution and adjunction. These predicate-argument trees are recursively linked to present the discourse structure. However, LTAG trees are not annotated to be linked with left and right adjacent trees, as RST does (Webber 2006).

The PDTB (Webber and Prasad 2006; Prasad *et al.* 2008a) annotates semantic and pragmatic relations (almost informational relations) held between two not necessarily adjacent arguments, following the approach of D-LTAG. They introduced also so called implicit connectives between adjacent arguments. Both explicit and implicit discourse connectives are annotated to link arguments via discourse relations. However, the PDTB approach did not annotate global relations to build a structure for discourse. More details about the PDTB are presented in Section 2.6.2. We based our discourse annotation for Arabic in current study on similar approach of the PDTB.

### 2.5.6 Segmented Discourse Representation Theory (SDRT)

Rhetorical relations are also a fundamental aspect of Segmented Discourse Representation Theory - SDRT (Asher & Lascarides, 2003). The logical form of discourse, according to their perspective, consists of a set of labels (which label the content of clauses, or of text spans in terms of truth conditions), and a mapping of those labels to *logical forms*, which can consist of rhetorical relations between the labels (arguments). A hierarchical structure is then created over the labels, allowing rhetorical relations to relate the contents of individual clauses or extended text spans. Figure 2-9 shows SDRT representation of text segments in Ex. 2-17. SDRT's rhetorical relations are less fine-grained than those used, for example in RST. The SDRT's Rhetorical relations must connect propositions, questions or requests. The contents of text spans can participate in more than one rhetorical relation unlike in RST (see Section 0).

**Ex. 2-17 (Sporleder and Lascarides 2006)**

a) The high-speed Great Western train hit a car on an unmanned level crossing yesterday.
b) It derailed.
c) Transport Police are investigating the incident.

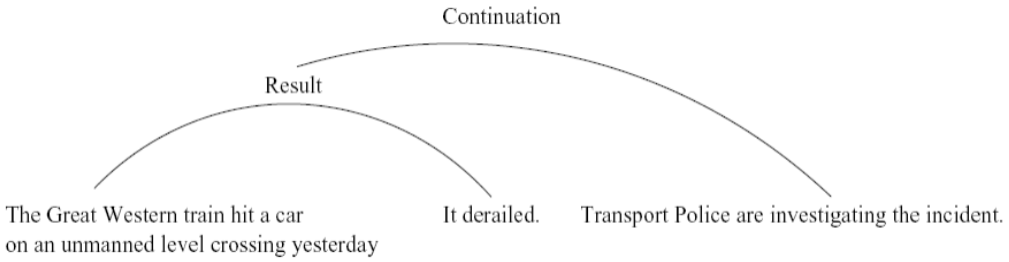



**Figure 2-9: The SDRT representation of Ex. 2-17. (Sporleder and Lascarides 2006, p.2)**

## 2.6 Resources for Discourse Studies

The demand for data resources such as corpora annotated with some form of discourse structure is growing as a result of the variety of potential applications that will be discussed in Section 2.7. However, the number of annotated corpora is still small given the extent of research interest in discourse structure (Webber, Egg and Kordoni 2011). While several resources have been annotated for English, only a few were constructed for other languages such as German, Danish, Czech, Hindi, Turkish, Chinese and Japanese. However, before the current study, no corpora were annotated for Arabic at the discourse level. One of the aims of this research is to produce the first corpus annotated for discourse properties in Arabic. The following sections describe available textual resources in other languages for discourse processing.

### 2.6.1 RST-based Corpora

As a result of the increased interest in RST theory, the first discourse resources have been annotated according to its principles. The RST Discourse Treebank (RST-DT) (Carlson *et al.* 2002; Carlson, Marcu and Okurowski 2001) comprises 385 articles from the Wall Street Journal corpus whose syntax has been annotated in the Penn Treebank. For German, the Postsdam Commentary Corpus (Stede 2004) consists of 170 commentaries from the German Regional daily newspaper Markische Allgemeine Zeitung. The PCC has annotation of both the syntactic and discourse levels, the latter again according to RST. The Discourse GraphBank (Wolf and Gibson 2005) is an English corpus that consists of 135 texts from the AP newswire

and Wall Street Journal, annotated according to W&G's theory which is an adaptation of RST (see Section 2.5.2). However, unlike RST corpora, annotators were not required to link all segment structures to have a full structure for a text. Thus the resulting annotation is a flat structure rather than hierarchical, with many cross-dependencies which were mainly related to the Elaboration relation (Webber 2006).

### 2.6.2 PDTB and Related Corpora

The PDTB project began with the D-LTAG representations in mind, as described in Section 2.5.5. However, the annotation guidelines were subsequently made as theory independent as possible so that the corpus would be usable by a wide range of users (Webber and Prasad 2006; Prasad *et al.* 2008a). The latest version of the Penn Discourse Treebank PDTB2 contains annotations of discourse relations and their arguments on the one million words syntactically annotated of the Wall Street Journal in the Penn Treebank. The annotation contains mostly informational discourse relations with a few pragmatic relations yielding for low-level discourse structure. The relations are mainly elementary predicate-argument relations whose predicates come mainly from discourse connectives and whose arguments come from units of discourse expressing abstract objects (AOs).

Discourse relations in the PDTB might be signalled explicitly by discourse connectives such as subordinating or coordinating conjunctions or discourse adverbials. Implicit relations are also annotated, but only between adjacent text spans. For the latter, the implicit inferable relations are annotated by inserting a so-called *implicit connective* that best expresses the inferred relation.

In Ex. 2-18, the subordinating conjunction *since* is an Explicit connective indicating a Temporal relation between the event of the earthquake hitting and a state where no music is played by a certain woman.

**Ex. 2-18**

*She hasn't played any music* **since the earthquake hit**. (WSJ text 0766)

An example of a relation inferred due to adjacency is given in Ex. 2-19, where the Causal relation between the AOs denoted by the two adjacent sentences is annotated with *because* as the Implicit connective.

**Ex. 2-19**

*Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda*. Implicit=BECAUSE (CAUSE) **As a former White House [...], he is savvy in the ways of Washington**. (WSJ text 0955)

Arguments in the PDTB do not have to be phrases at the syntactic level but rather all linked text spans must meet the conditions of relation arguments. In addition, annotators are allowed to annotate relations signalled by expressions not defined as discourse connectives such as AltLex (Alternative Lexicalization relations which use non-connective lexical expressions to link adjacent sentences), Entity and Attribuation.

The PDTB annotation principles of discourse relations are almost theory-neutral, with clear definitions of relations that link adjacent and non-adjacent arguments, and allowing for crossing dependencies. Good inter-annotator agreement was reported when annotating discourse relations for English in the PDTB2 and other languages such as the METU Turkish Discourse Bank (Zeyrek and Webber 2008), the Hindi Discourse Relation Bank (Prasad *et al.* 2008b) and the Chinese Treebank (Xue 2005), all of which were annotated using similar annotation principles as the PDTB. However, no attempt has yet been made to test these annotation principles on Arabic.

In the first discourse corpus creation project for Arabic, we annotate explicitly signalled discourse relations following similar annotation principles as the PDTB after applying all required Arabic-specific adaptations.

### 2.6.3 Dependency Treebanks

The Copenhagen Dependency Treebank (Buch-Kromann and Korzen 2010) consists of 480 annotated parallel texts in Danish and English, and 300 annotated parallel texts for German, Italian, and Spanish. Both syntactic and discourse annotation were done in the form of a tree dependency structure, linking up the top dependency node of each sentence with those of other sentences and labelling the relation between them.

The Prague Dependency Treebank, PDT 3.0 has a layer of annotation which captures discourse relations. The difference between the PDTB and the PDT is that the

annotator links the megatree of sentences (a tree structure of syntactic dependency in the PDT 2) as arguments of an inter-sentential relation. For intra- sentential relations, such as clausal coordination, the syntactic annotation is already annotated in the PDT 2 and should be transformed automatically into the discourse layer.

### 2.6.4 Annotation Tools

Large scale annotation projects require a software tool-kit to make the annotation process a more reliable and faster task. The available tools for discourse annotation are theory-oriented, namely they are developed with one theory of discourse structure in mind and provide options that fit with its requirements and relation taxonomy. The RST Annotation Tool, is an extension of Mick O'Donnell's *RSTTool*[3], a graphical interface for marking up the structure of text based on RST theory and for implementing required tasks such as automatic text segmentation. The Java tool *annotator* (Wolf *et al.* 2003) was used to annotate text in the Discourse Graph Bank by linking discourse units with an arc in graph representation (this tool is for lab use only and not available to the public).

Some tools use stand-off annotation methodology that allows the annotator to mark-up all potential cases. This might handle overlaps and crossings among relations. For example, in the first stage of the PDTB project, the *WordFreak* annotation tool (Morton and LaCivita 2003) was used to annotate discourse relations and arguments. However, in the second stage of the annotation PDTB2, a Java tool *annotator*[4] was developed especially for their discourse annotation tasks. For creating the METU Turkish Discourse Bank, *DATT* (Discourse Annotation Tool for Turkish) was developed and the tool produces XML files as annotation data (Aktaş, Bozsahin and Zeyrek 2010).

In previous work (Seif, Mathkour and Touir 2005a), I have designed a shallow annotation tool based on RST concepts for Arabic. The tool used rules to segment a text into units, to identify the discourse connectives and then links units via unambiguous relations and builds all valid RST trees for the text. However, this tool is very limited in functionality and did not generalize well to annotate unseen text as

[3] http://www.wagsoft.com/RSTTool/, the download page of the RST tool is http://www.isi.edu/licensed-sw/RSTTool/

[4] The download page of the *annotator* tool is http://www.seas.upenn.edu/~pdtb/tools.shtml#annotator.

it used a very small inventory of relations and connectives. The purpose of this tool was to test the applicability of the RST concept to Arabic on a sample of 4 articles, as a part of my master dissertation. Apart from this RST-tool, no available annotation tools can be used to annotate Arabic discourse connectives, their relations and arguments. Further discussion about tools for Arabic discourse is presented in Section 3.2.3.

### 2.6.5 Inter-annotator Agreement Coefficients

To test the reliability of an annotation scheme and annotation process, different measures can be used test the agreement between several annotators. These measures are also used to evaluate the performance of automatic systems. The appropriate agreement measure depends on the coding task and number of labels. The coding task might code data with two labels (binary coding). For example, for a given potential connective, an annotator marks the instance as either a discourse connective or not a discourse connective in context. The coding task might also mark the instance with one or more labels from a pool of labels specified in the task such as annotating discourse relations for discourse connectives. In addition, the coding task can mark instances with no pre-defined labels such as marking the boundaries of the argument or discourse unit.

The most common agreement coefficient for a finite number of lables is *percentage agreement*. It is defined as the proportion of times that the coders agree (1 means they agreed on all data instances, 0 means they never agreed). However, this measurement might be misleading, in that the overuse of very common labels by one or more coders will produce high agreement by chance. The kappa coefficient (K) was developed to factor in chance agreement.

In Equation 2-1, P(A) is observed agreement or percentage agreement. P(E) is the percentage of agreement expected by chance. The kappa coefficient has two versions: $K_{Co}$(Cohen 1960) and $K_{S\&C}$ (Siegel and Castellan 1988). They differ only in the way of measuring chance agreement. K is 1 when there is perfect agreement among the coders. In contrast, when k is zero, this means the agreement is equal to chance. The content analysis researchers assume the annotation is highly reliable when $K > 0.8$, that there are tentative conclusions to be drawn when $0.67 < K < 0.8$,

and that the annotation and the scheme are not reliable when k < 0.67. For more details about K refer to (Artstein and Poesio 2008).

$$K = \frac{P(A)-P(E)}{1-P(E)}$$

**Equation 2-1: Kappa coefficient. P(A) is observed agreement, P(E) is agreement expected by chance.**

K is not a very appropriate measure for annotation tasks, where labels might partialy overlap. A weighted agreement measure α was developed to tackle partial agreement among coders in such cases by using a distance metric between two labels A and (Artstein and Poesio 2008). The distance is 0 when A and B are identical, 1 when there is no overlap between A and B, or a certain fraction in between that depends on the overlap and the distribution of the labels.

For open-ended set of labels such as the agreement on words of text spans, it is not possible to use kappa or α metrics. In such cases, *exact match* and *agr* measures can be used. *Exact match* is a metric used to measure how often two annotators marked exactly the same text; it is 1 when both coders mark the same text spans, 0 when not. *Agr* is a metric used to measure partial agreement among coders ann1 and ann2, *agr*(ann1||ann2). It is a directional measure of agreement using Equation 2-2 that measures what proportion of text marked by coder ann1 was also marked by coder ann2. The first usage of *agr* was by (Wiebe, Wilson and Cardie 2005) to measure agreement on opinion and emotion expressions. The overall agreement is the average of the *agr* measure for both directions *agr*(ann1||ann2) and *agr*(ann2||ann1).

$$agr(ann1||ann2) = \frac{|\, token\ of\ \text{ann1}\ \boldsymbol{Match}\ tokens\ of\ \text{ann2}|}{|tokens\ of\ \text{ann1}|}$$

**Equation 2-2: The agr measure for two text span marked by coder 1 (ann1) and coder 2 (ann2). Modified from (Wiebe, Wilson and Cardie 2005).**

## 2.7 Algorithms and Applications for Discourse Structure

In order to use discourse structure in developing computational applications, it is an elementary prerequisite to develop algorithms for detecting the structure of a discourse based on one of the theoretical viewpoints discussed in Section 2.5. This section presents a brief overview of the algorithms that are used for recognizing and generating various forms of discourse structure, and the common applications of the discourse structure in literature. A complete recent survey of the algorithms and applications is reviewed in (Webber, Egg and Kordoni 2011). According to this survey, the common algorithms can be classified into three types: *discourse segmentation, chunking and parsing.*

*Discourse segmentation* segments the text into adjacent *topically-coherent* or *functionally-coherent* segments such as the TextTilling approach in (Hearst 1997). In this approach the segment boundaries are determined by a threshold of similar initial fixed-length spans using a *cosine similarity* for the frequent word stems of adjacent spans. Discourse chunking identifies the text segments that convey informational discourse relations. One method of discourse chunking is by identifying the lexical signals for discourse relations in a text such as connectives, and then identifying their arguments (Prasad *et al.* 2008a; Pitler and Nenkova 2009). Discourse parsing is the process of constructing a complete structured cover of a text such as a tree structure whose leaves are the elementary discourse units linked by local and global relations.

Prior work in both discourse chunking and discourse parsing is strongly related to our computational modeling that attempts to identify discourse units (arguments), their signals (discourse connectives), and the discourse relations conveyed. Section 2.7 provides more details with regards to other works for detecting discourse structure in English.

One of the earliest applications influenced by weighted (such as the nuclearity principle in the RST) and un-weighted discourse structure theories, is *automatic document summarization*. The nucleus-satellite classification of discourse relations in RST led to the view that in summaries, satellite arguments can be omitted without affecting text readability. Satellites represent in general extra information for more elaboration only (Marcu and Echihabi 2002 ; Marcu 2000c). Summarization could

also have other objects such as genre-specific summarization. To summarize scientific papers, Teufel and Moens (2002) assumed that most papers consist of similar functional parts (*aim, outline, methods, results, discussion, and related work*). Most summarization efforts use news and scientific papers as a source, thus their texts usually follow a specific structure. Barzilay and Elhadead (1997) devised another approach to summarization, where only sentences with strong *lexical chains* are extracted to represent a summarized text.

The most frequent use of RST has been in Natural Language Generation (NLG). Discourse relations are used in discourse modules to find appropriate discourse markers. The types of text generated in the literature include instruction manuals (dialogue and text), administrative forms, user documentation, descriptions of tourist sites and descriptions of concepts (see (Taboada and Mann 2006a) for a summary).

Another common application of discourse structure is *information extraction* (IE). The systems here extract entities, relations between them, and event structure that plays a role in the text. Event structure is often defined by a template to be filled by extracted entities. Flat and hierarchical discourse structures can be used to identify relevant regions for a specific piece of information. For example, Mizuta (2006) uses discourse segmentation of topics (zones) to extract the novel contribution of scientific articles. Maslennikov and Chua's (2007) extract semantic relations between entities such as *x is located in y* using a full hierarchical discourse structure.

Essay scoring and analysis use the organizational structure of an essay (a crucial feature of quality) to automatically identify thesis statements (Burstein et al 2001). In their approach, decision-trees and probabilistic classifiers are trained on annotated data and evaluated against unseen data using features extracted from RST parsing and lexical items.

Question-answering is a well known application that can use discourse relations to answer complex queries about the content of a discourse which goes beyond the content of its individual clauses (Girju *et al.* 2003; Marcu 1999b). Also, (Pitler and Nenkova 2008) used discourse relations for predicting text readability and ranking the readability of essays.

More details about the applications of discourse processing are in (Webber, Egg and Kordoni 2011).

## 2.8 Computational Modeling of Discourse

Discourse structure and relations have in recent decades enjoyed growing interest among NLP researchers. They share the main objective to create a reliable discourse parser that can build a structure for a whole text. The empirical studies focused on different parts of this problem statement. We will now discuss approaches to the identification of discourse units (arguments), discourse connectives and discourse relations. These approaches were mainly developed for English. The automatic models differ on the theory of discourse structure they rely on, type and size of the training/testing sets for the supervised models (whether they are manually or automatically extracted datasets), and the feature sets they used.

### 2.8.1 Identification of Discourse Units

Because the definition of discourse units in RST differ slightly from the definition of the arguments in the PDTB annotation, or discourse segment purposes in G&S, different automatic models were developed to identify these elementary discourse units. Marcu (1999) addressed in his first attempt to develop a RST-based parser that the quality of identifying elementary discourse units strongly affects the performance of identifying discourse relations between the units in the parser. He identified the discourse units using a decision tree model with surface features such as potential connectives, position of verbs and punctuation in addition to part of speech features. His parser, then, was trained with another decision tree model on these automatically identified discourse units. However, the parser achieved very low accuracy 15-45% compared to the human accuracy of 70-80%. The same parser had a high accuracy of 50-60% when it was trained on manually identified discourse units.

Soricut and Marcu (2003) improved the parser by using lexicalised syntactic parse trees in a probabilistic model to identify discourse units and relations. The syntactic trees were produced from two sources: the manually annotated ones in the Penn Treebank and ones created automatically by Charniak's parser (Charniak 2000). The model was trained on the RST Discourse Treebank and the error reduction was around 15-20% over the parser in (Marcu 2000c; Marcu 1999a). However, these high results were only for discourse units of intra-sentential discourse relations (both units

are in the same sentence). Thus, discourse units of inter-sentential relations, such as for the majority of adverbials, were not addressed in this parser.

The second trend found in the literature when identifying discourse units or arguments of explicit connectives, is identifying the head of arguments in a dependency annotation, rather than identifying full argument spans. (Wellner and Pustejovsky 2007) approach is the first study that proposed a practical evaluation of using this methodology. They trained ranker models on the PDTB for Arg1 and Arg2 identification for a given discourse connective, and then a joint re-ranking model for the proposed pair. Their features include the dependency parse path, constituency parse path, connective type (coordinating/subordinating conjunctions or adverbials) and lexical-syntactic features for attributions. They demonstrated that dependency parse features were very significant and their model achieved an accuracy of 74.2 % with gold-standard parses, and 64.6% accuracy with automatic parses (Charniak's parser). Recently (Wang, Su and Tan 2010) used also sub-trees as features rather than using the path between a connective and a potential argument, and achieved a significant improvement on identifying arguments and explicit and implicit discourse relations in one go.

Rather than using a single general classifier to identify arguments (Arg1 and Arg2) of different explicit connectives in the PDTB, Elwell and Baldridge (2008) trained separate models for each connective and connective type. They had noted that connectives differ in their distribution and behaviours, so there would be conflicting effects on the feature weights in a general model. A proposed mixture of general and connective specific models was used to identify the arguments of discourse connectives. The performance of this model exceeds the ones of (Wellner and Pustejovsky 2007) by 3.6% when using features from gold-standard parses, and by 9.0% when using automatically produced parses.

Recently, work in (Prasad, Joshi and Webber 2010a) assumed that identification of Arg2 is relatively trivial in that it is syntactically associated with the connective in the PDTB. Therefore, the challenging task is the identification of the Arg1 argument; it may or may not be adjacent to the connective. The interesting idea here is to identify the *sentence* containing Arg1, rather than the exact argument span, for inter-sentential connectives which occur on non-initial position of the paragraph (ParaNonInit). In the PDTB, 91% of the time, Arg1 of ParaNonInit connectives is the

previous sentence, and only 49% of the time Arg1 of ParaInit connectives is the previous sentence. They claimed, therefore, that the automatic identification of Arg1 sentence for ParaInit connectives is a harder task, and so was not addressed in this paper. They were filtering the potential candidate Arg1 sentences (all prior sentences in the paragraph) using co-reference-based rankers to evaluate manually the candidate sentences. They achieved, on a set of 743 tokens, an overall accuracy of 86.3%, with an improvement of 3% over the baseline (choosing a sentence immediately preceding the sentence hosting the connective).

The identification of arguments of Arabic discourse connectives is beyond the scope of the current work but will be a main task to be addressed in the future (Section 9.3).

## 2.8.2 Modeling Discourse Connectives

### 2.8.2.1 Recognition of Discourse Connectives

The majority of (potential) discourse connectives in English are unambiguous in terms of having discourse usage in text (Pitler *et al.* 2008). Most potential connective strings (such as *because* or *in contrast*) are always discourse connectives, independent of context. However, some discourse connectives such as the conjunction *and* or the connectives *while* and *once* might occur in a text with a non-discourse function, for example, as a different part of speech (*while* is a noun in *I have not seen you for a while*) or sentential (Mary *and* John). Thus, the detection of the discourse usage of potential connectives is a task required to discover discourse relations.

The only comprehensive empirical study to classify given potential connectives into discourse connectives or not discourse connectives in context was conducted by (Pitler and Nenkova 2009). The authors used syntactic and pair-wise interaction features between the connective and each syntactic feature plus the connective string itself. Applying a maximum entropy classifier on PDTB explicit connectives and non-annotated potential connectives in the corpus, they achieve 96% accuracy over the high performance baseline (86%) of using the connective string alone. However, this classifier was based on the gold standard parses only, and there are no studies

available that compare its results to models that use automatic parsers such as the Stanford[5] or Charniak parsers.

#### 2.8.2.2 Prediction of Discourse Connectives

Lapata and Lascarides worked on determining *temporal connectives* and their relations for the growing interest of event order in language applications such as text generation, summarisation and question answering (Lapata and Lascarides 2004). The authors developed Naïve Bayes models for inferring temporal connectives. For that, they extracted the training data automatically from the BLLIP corpus (30M words), a Treebank-style machine-parsed version of the Wall Street Journal. They identified *temporal connectives*, with respect to the temporal relations they signal and then removed the connectives. The task was to recover the discourse connective itself using lexical and grammatical features. The best model acquired up to 70.7% of connectives correctly. Some of the connectives are ambiguous in terms of relations they signal, but the authors did not address the task of disambiguation.

On the other hand, a different classification task for discourse connectives was conducted by (Hutchinson 2005). He investigated empirically how well one discourse connective could be substituted for another by modeling substitutability and similarity of discourse connectives as in (Knott 1996).

### 2.8.3 Modeling Discourse Relations

As discussed earlier in Sections 2.1 and 2.3.2, discourse relations might be inferred from the context (implicit relations) or signaled by discourse connectives (explicit relations). Although discourse connectives in English are almost unambiguous, in that each connective indicates almost only one discourse relation (Pitler *et al.* 2008; Pitler and Nenkova 2009), there are connectives such as *since* which can signal several relations such as temporal, causal relations or both as shown respectively in the examples (a, b and c) in Ex. 2-20.

**Ex. 2-20**

d) This mark is the best ever mark I got **since** the exams were conducted in our department. (Temporal)

[5] http://nlp.stanford.edu/software/lex-parser.shtml.

e) The suspect man in the next door was arrested **since** he stole a car. (Causal)
f) She could not sleep **since** her father died. (Temporal and Causal)

Models for recognizing discourse relations differ in their definitions for relations, the theory the developers consider, dataset for training and evaluation, and types of relations (explicit, implicit or both with no clear distinction). The main task of these models is, given two discourse units/arguments, to discover what discourse relation(s) relate them. We will start with models that treat both relation types with no distinction. As they did not distinguish the two types of relations, any improvement might result from recognizing relations explicitly signaled which are almost unambiguous.

Soricut and Marcu (2003) showed that the strong connection between lexical and syntax features can benefit automatic discourse parsing (see Section 2.7 for more details about discourse parsing). The authors used supervised probabilistic models using surface and syntactic features on data from the RST Discourse Treebank (RST-DT) to detect 18 coarse granularity RST relations classifed by (Carlson, Marcu and Okurowski 2003) such as *Attribution, Background, Cause, Comparison, Condition, Contrast, Elaboration, Enablement, Evaluation, Explanation, Joint, Manner-Means, Topic-Comment, Summary, Temporal, Topic-Change*. The relations were local (between terminal nodes) and global (to link subtrees), and were almost all explicitly signaled. The parser recorded a good performance 75.5% with syntactic and lexical features, better than using lexical features only as for the parser in (Marcu 2000b), but the performance dropped when using automatic identification of discourse units instead of the gold-standard segmentation from the Penn Treebank.

An improved faster parser using RST relations was developed later by (Duverle and Prendinger 2009) to build RST trees using support vector machine models in a bottom-up tree building approach on gold standard segmentation data. The features included syntactic, lexical and features from previously classified sub-trees. The approach proved that words on the edge of discourse segments are the most meaningful for signaling relations, as they include discourse connectives.

Baldridge and Lascarides (2005) developed a dialogue parsing system using SDRS discourse relations. Because the SDRS-representation scheme uses graph structures at the sentential level, it does not propose a structure for the whole discourse. They designed a head-driven probabilistic parsing model using sentential parsing (Collins,

2003) to parse discourse of the Verbmobil appointment scheduling and travel planning dialogs from the Redwoods Treebank, annotated with SDRT rhetorical relations. In addition to lexical and syntactic features, the mood of each sentence, discourse connectives and dialogue-specific features are used. Their best model performs well (67.9%) on unlabeled data over the baseline of assigning the most frequent relations (53.3%).

The first parsing system using Graph Bank representation was developed by (Wellner *et al.* 2006). They used a variety of lexical, syntactic, and semantic features based on relationships between words inferred from the Brandeis Semantic Ontology (Pustejovsky *et al.* 2006) and word similarity. The best model achieved 81% accuracy which out-performed the baseline of the majority relation (45.7%). A further improvement was reported when using dependency features, with accuracy of 82.3%.

Other studies concentrate on identifying specific relation types such as temporal discourse relations (Mani *et al.* 2006; Lapata and Lascarides 2006). Lapata and Lascarides (2006) used temporally annotated corpora (using the TimeMl annotation scheme) that annotate temporal features manually within the main and subordinate clauses. Models are generated using features including temporal discourse connectives (e.g. *before, after* and *while*), tensed verbs, aspects, adjectives, time expressions and world knowledge. The best model achieved F-score of 69.1% on inferring temporal relations when trained and tested on the BLLIP corpus. They found also that syntax trees encode sufficient information for recognizing temporal relations.

There are interesting attempts in the literature to avoid the time and cost of human labeling for discourse studies. Their training and testing data are automatically generated using either unambiguous discourse connectives and/or structural patterns for specific relations (Marcu and Echihabi 2002 ; Sporleder and Lascarides 2005; Hutchinson 2004a; Hutchinson 2004b; Lapata and Lascarides 2004; Sporleder and Lascarides 2008). The connectives then are removed to simulate implicit relation instances. The task is then to regain the original connective (Lapata and Lascarides 2004) or to identify the relation (Marcu and Echihabi 2002 ; Sporleder and Lascarides 2005).

An advantage of this method is the possibility of collecting a large amount of the data that models require for specific infrequent relations. However, the studies concluded that the good performance achieved by models on artificial data, did not carry over when tested on manually annotated data of implicit relations (Sporleder and Lascarides 2008). That clarifies that the two assumptions that these studies rely on are not quite correct. The first assumption is that sentence/clause features are the same whether the discourse relations between them are signalled explicitly or implicitly. The second assumption is that the distribution of implicit relations is the same as that of signalled relations.

#### 2.8.3.1 Recognition of Explicit Discourse Relations

Models recognizing explicit discourse relations (senses) of discourse connectives mostly treated the problem as a classification task. Studies in (Hutchinson 2003, 2004, 2005) provided empirical evidence for the correlations between discourse relations and certain linguistic features such as lexical and syntactic features in the context. For instance, Hutchinson (2004) automatically classified 140 unambiguous discourse connectives using the definitions in (Knott 1996; Knott and Sanders 1998) with regard to three classes: polarity (negative or positive), veridicality (veridical or non-veridical) and type (additive, temporal or causal). The last class represents theory-neutral discourse relations signaled by a given connective in its context. The data for the experiments was collected and parsed automatically from the British National Corpus and the World Wide Web for the targeted connectives.

To avoid annotating the data, he distinguished the tokens of the discourse connectives using predefined syntactic patterns such as (SBAR (IN after) (S..)) (Hutchinson 2004b). Features such as part-of-speech, verb tense, temporal expression and the discourse connectives themselves were used to run two models. The k-nearest neighbor model was used based on a hypothesis that connectives at the same class will have similar co-occurrence patterns. The Naïve Bayes model was also applied which takes the overall distribution of each class into account. The best model achieved over 90% accuracy on all three classes.

(Haddow 2005) treated the disambiguation of discourse connective functions as a form of word sense disambiguation. Only six ambiguous discourse connectives

(*after, as soon as, before, once, since* and *while*) were considered and disambiguated according to the SDRT relations. He used maximum entropy models with features such as collocations (words or POS tags occurring in a particular position in a window of defined size centered on the connective), co-occurrences, structural features using punctuation pattern of the sentence. The best model achieved an average of 70.4% accuracy across all the connectives, with a good improvement over the most frequent sense baseline of 57.2%.

Regarding PDTB–based discourse parsing, Miltsakaki and colleagues (Miltsakaki *et al.* 2005a) proposed a first step at disambiguating the senses of a small subset of connectives (*since, while*, and *when*). They used syntactic features derived from the uncompleted Penn Discourse Treebank and a MaxEnt model to distinguish between temporal, causal, and contrastive usages of these connectives. An improvement of 15-20% was achieved over the baseline (most frequent sense per connective).

Studies by (Pitler *et al.* 2008; Pitler and Nenkova 2009) disambiguate all explicit discourse connectives at the class level in the PDTB2. They concluded that by using only the connective string, discourse relations between known arguments can be predicted with a high accuracy of 93.67 for the four main class relations (see the relation hierarchy in Figure 2-4, p.20). Adding syntactic features that were extracted from gold standard parse trees in the Penn Treebank plus surface based features, the model achieved almost human performance, 94%. However, they did not address instances when a connective signals more than one relation. In addition, they did not investigate how automatic parsing would affect the results. For the best of our knowledge, these two issues have not been investigated for (class or fine-grained) explicit relations in the PDTB.

#### 2.8.3.2 Recognition of Implicit Discourse Relations

Recognizing discourse relations that can be inferred from context, without explicit signaling, attracted many researchers and is a challenging task when developing a discourse parser. In fact, roughly half of the sentences in the British National Corpus do not contain any discourse connectives (Sporleder and Lascarides 2005)[6]. Also 47.5% of the discourse relations annotated in the PDTB2 are implicit relations (refer

[6] Note that these sentences might have other lexical expressions to link discourse segments such as AltLex annotations in the PDTB.

to Section 2.3.2). Another challenge here is that supervised machine learning models require a reasonably large amount of annotated data for such relations, and this is hard to be achieved automatically since there are no explicit signals such as connectives that can be used to collect data.

Most of the recent work on recognizing implicit relations is based on the PDTB (Pitler, Louis and Nenkova 2009; Blair-Goldensohn, McKeown and Rambow 2007; Lin, Kan and Ng 2009; Wang, Su and Tan 2010; Louis and Nenkova 2010; Zhou *et al.* 2010). Pitler and colleagues (2009) used surface, lexical, POS tags, word-pairs of non-function words, immediate preceding explicit relations, and modality to classify adjacent arguments in the PDTB into their class level relations. The best combination of features for the four classes in a Naïve Bayes model led to improvement by 4% for Comparison and 16% for Contingency over the baseline of randomly assigning classes.

Lin and his colleagues (Lin, Kan and Ng 2009) used similar features as in (Pitler, Louis and Nenkova 2009) and added constituency parse features such as production rules and dependency parse features to classify 12 fine-grained relations. Their maximum entropy classifier achieved a 14% improvement over the baseline (26.1%) of the majority class (Cause). In 2010, they developed the first PDTB end-to-end parser (Lin, Ng and Kan 2010). Zhou and his colleagues (2010) addressed implicit relation recognition via two classification tasks: first predicting a discourse connective that should be inserted between two adjacent arguments (implicit connectives in the PDTB2) using a language model, and then recognizing the relation by using the predicted connectives as features. In addition to the connectives, the supervised model used other features such as lexical, and syntactic features that were useful in prior work (Lin, Kan and Ng 2009; Pitler, Louis and Nenkova 2009). Similar to Pitler and her colleagues (2009), Zhou and his colleagues used four binary classifiers, one for each relation type at class level. Their approach achieved an average F-score improvement of 3% over the baseline by (Pitler, Louis and Nenkova 2009).

### 2.8.4 Discussion and Influence on This Work

Few studies have been conducted for discourse connective identification in English. Using only simple lexical, surface-based and syntactic features, the models can achieve almost human performance. However, this might be not the case for identifying discourse connectives in other languages.

With regard to relation identification, most discourse connectives in English are unambiguous in term of the relations they indicate. Therefore, few successful approaches have dealt with the ambiguity problem of connectives such as the connective *since*. None of these attempts have used automatic tagging/parsing to extract features, or tried to detect more than one relation per connective.

The challenge in the field is identifying implicit discourse relations where there are no discourse connectives signalling the relations explicitly. In general, little improvements (3-16%) have been achieved over the baseline by different models using lexical, surface-based, syntactic, semantic and parse features. Understanding context might sometimes not be enough, (Lin, Kan and Ng 2009) suggests using world knowledge to understand the relations between arguments in the absence of explicit connectives.

We tackled in this human and automatic annotation study only the explicit discourse connectives and their relations in MSA.We claim that explicit connectives are highly frequent used in MSA, with a highly ambiguity level in terms of having discourse function and signalling relations (See Section 7.7 for more discussion). Therefore, this study will develop models for identifying explicit discourse connectives and their relations using insight from previous experiments for English. We also use additional Arabic-specific features that might improve the performance for some connectives such as Al-maSdar nouns. Our expertments and a full discussion are presented in Chapter 8.

## 2.9 Summary

This chapter presented an overview of discourse structure, a way of formalizing discourse coherence. The explored discourse structure studies and theories consider

discourse relations between arguments as a central base. The relations might be signalled explicitly by discourse connectives. We described the types of connectives, and relations and their taxonomies for English.

The discourse structure theories represent either intentional or informational organisations (or both) of a discourse. RST seems to be the most popular theory used in computational studies and applications such as text generation, automatic summarisation and machine translation. It is simply the case that trees are convenient, easy to represent, and easy to process. However, RST does not allow conveying relations between non-adjacent discourse segments, which prohibit many necessary cross dependencies (Wolf and Gibson (2005) and (Webber 2006)). Graph representation of discourse structure was assumed in W&G going beyond a tree structure of discourse. However, a graph structure would not solve all problems; they often do not cover a whole discourse.

A new wave of discourse studies focuses on local relations between arguments in a theory-neutral approach. The PDTB is a famous well-established project following this approach. To date, it is not possible to generalize one representation of discourse structure for written and spoken language, leaving discourse structure a genre-based attractive research field which requires further study and investigation.

We also reviewed the existing resources such as corpora and annotation tools for discourse studies. The corpora such as the PDTB are used in building models to recognise discourse relations automatically. While the explicitly signalled relations are much easier for automatic identification, less progress has been achieved for implicit relations.

While the discourse studies and resources discussed focused on English, other natural languages seem to share many of the basic concepts. They also might benefit at least partially from those studies. To date, there is no large scale study of discourse processing in Arabic nor corpora and tools to be used as basis for the studies. Our work in this thesis would establish a new generation of discourse processing and resources for Arabic. This study has two strong targets, namely creating reliable resources for Arabic discourse following the PDTB approach, and then using them to model discourse relations automatically.

# Chapter 3 Object of Investigation and Research Methodology

This study promises substantial contributions to the field of Arabic discourse processing. In this part of the thesis, we summarise the main characteristics of Modern Standard Arabic (MSA) that have impacted the study methodology, and the rigorous methods that were used to obtain the results.

## 3.1 Characteristics of Modern Standard Arabic

**Arabic** is the sixth most populoua language in the world, with up to 246 million native speakers and is an official language in 25 countries. The Arabic script has 28 letters; most of them are fully connected when writing. A few letters are connected only to preceding letters. In such cases, there will be small white spaces between letters of a single word, for example (كتاب/book), which require special manipulation in character recognition systems, for example. In addition to (constant and vowels) letters, other phonological symbols are used in Arabic such as short vowels, vocalic length, Hamza (glottal stop), shadda (consonantal length) and optional diacritics.

The contemporary written Arabic is called **Modern Standard Arabic** (MSA). It is derived from Classical Arabic - CA (Quranic Arabic and the language used in 6$^{th}$ century by Arabs). MSA is the language used nowadays in education, news, press, books, but not always used in spoken language due to the effects of dialects of different Arab regions (Habash 2010; Ryding 2005). Most modern Arabic NLP studies, including the study in this thesis, use MSA as source of their data. However, they also learn from linguistic studies on CA as both sharing the same language characteristics with slight differences.

Arabic has a complex root-based morphology where a complete sentence can be expressed in one white-space word. Three types of concatenative morphemes exist: stems (the core), affixes (prefixes, suffixes and postfixes) and clitics (proclitics and enclitics). Clitics attach to the stem after affixes and both are optional. Distinguishing clitics from affixes is a confusing task for the Arabic researchers in the field (Attia 2007). Affixes have morpho-syntactic features such as tense, person, gender or number, while clitics have syntactic functions such as negation, definition, conjunction or preposition (Attia 2007; Habash 2010; Ryding 2005). For example, the sentence '*then they will read it'* is presented in Arabic as one white-space word 'فسيقرأونها'. The cliticization and the gloss translation of this word are presented in Figure 3-1, to show the affixes and clitics (one proclitic, one prefix, one postfix and one enclitic) attached to the stem.

| | |
|---|---|
| Arabic word: | فسيقرأونها |
| Cliticization: | فـ + سـ +يقرأ + ون + ها |
| | **enclitic + postfix + stem + prefix + clitic** |
| Gloss translation and syntactic analysis: | |
| It (object)+ they(subject)+**read** (present verb)+will (tense)+then (connective) | |
| English translation: | **then they will read it** |

**Figure 3-1: The clitiziation and a syntactic analysis of one word in Arabic that represents a complete sentence, to be read from right-to-left (apart from English translation).**

In addition, more than one stem can be produced from a root of 3 or 4 letters using different derivations of internal structure (patterns). For example, from the consonantal root كتب */ktb/write* several forms can be derived that indicate different grammatical features such as the verbs كتب*/ktb /to write,* verbal sentences {كتبتُ*/katab-tu/I wrote*, كتبتَ*/ktbt/you wrote* (masculine singular), كتبتِ*/ktbt/you wrote* (feminine singular), اكتب*/Aktb/I write*, nouns (كُتب*/ktb/books* (plural), كتاب*/ktAb/book* (singular) *and* مكتبة*/mktbp/library* (object))}. One of the morphological derivations that plays a critical role in our study is the *al-maSdar noun*.

**Al-maSdar** is a well-known noun category that expresses events without tense. The events can be related via discourse relations, which can be indicated by cohesive devises such as explicit discourse connectives, the subject of this study. Prepositions are often followed by al-maSdar nouns. That makes prepositions potential discourse connectives, and al-maSdar nouns potential arguments for them in our discourse annotation for Arabic.

Al-maSdar nouns are generated by using well-defined morphological patterns (أوزان) for 3 or 4 letter-roots. The patterns can attach suffixes to the root and insert consonant/vowel letters or diacritics in the root. More than 60 morphological patterns can be used to generate al-maSdar nouns (M. Abdl al latif, Zahran and Al-Arabi 1997; Ryding 2005; Alansari 1985). Some patterns of the 3-letter roots use only diacritics, without addition of any letters. A list of common al-maSdar patterns is provided in Appendix A. Figure 3-2 describes the steps of using the pattern إنفعال to generate an al-maSdar noun *انعكاس/reflection* from a root of three letters س ك ع */reverse* or *reflect*. In contract to al-maSadar generation, detecting al-maSdar nouns automatically is not a trivial task in MSA due to the absence of diacritic and al *hamzah* symbols in contemporary writing.

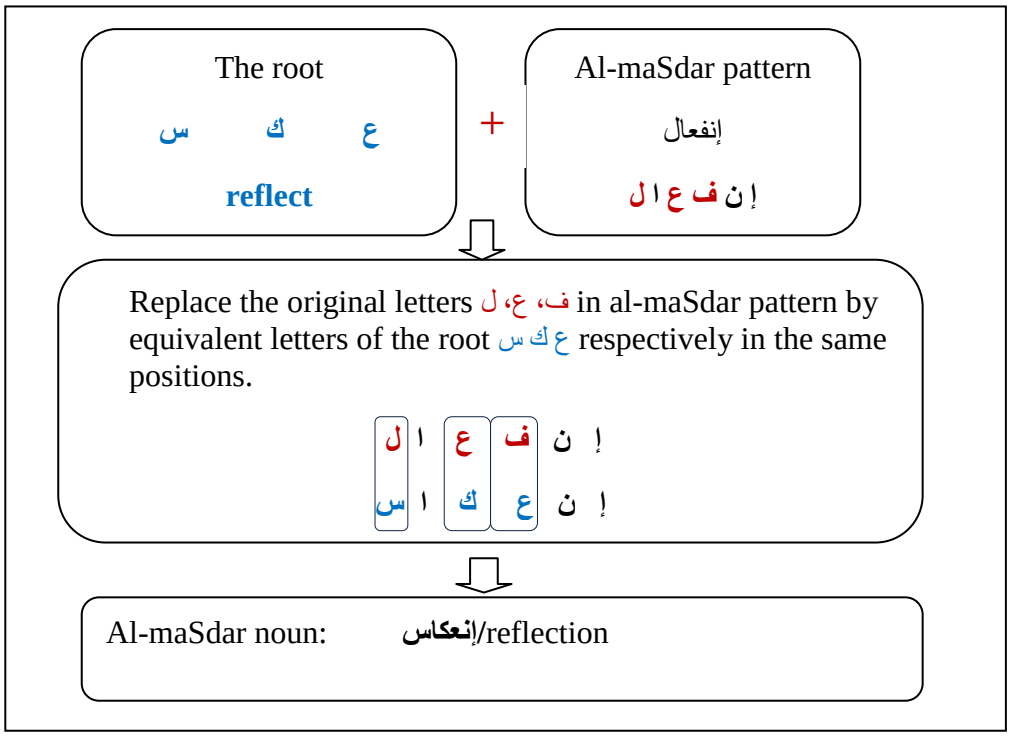



**Figure 3-2: The derivation of the al-maSdar noun انعكاس/reflection from a 3 letter root عكس/reverse.**

Al-maSdar nouns do not fit into one grammatical or morphological category in English; they might correspond to a gerund, nominalization or not nominalized nouns. Table 3-1 shows examples of al-maSdar nouns translated into different categories in English.

**Table 3-1: Examples of al-maSdar nouns, roots and patterns with English correspondences.**

| Root | Morph. Pattern | Al-maSdar noun | English |
|---|---|---|---|
| سبح/sbh | فعالة | سباحة | swimming |
| عكس/eks | انفعال | انعكاس | reflection |
| جرب/jrb | تفعلة | تجربة | experiment |
| حَرَبَ/hrb | فَعْلْ | حَرْبْ | war |
| دفع/dfe | فعال | دفاع | defence |

**Word order in Arabic.** Although the canonical order of Arabic sentences is VSO (verb –subject -object), a range of other orders are possible in specific grammatical constructions (Al-Sughaiyer and Al-Kharashi 2004).

**Punctuations in Arabic.** Unlike English, no capital letters exist in Arabic, the full stops and commas are used instead in modern Arabic books. However, the conventions for Arabic punctuations are less standardized and systematic than those in English (Dickins, Higgins and Hervey 2002). They claimed that the length of an orthographic sentence in English is almost equivalent to a single spoken sentence. However, the one orthographic sentence in Arabic is equivalent most of the time to two or more spoken sentences (Dickins, Higgins and Hervey 2002). This factor increases the challenge of defining the boundaries of sentences automatically in Arabic NLP studies. In the absence of proper punctuations, the connectives such as coordinating/subordinating conjunctions are used also for defining the sentences' boundaries. Figure 3-3 presents one orthographic full-stop ended sentence that contains more than one spoken sentence[7]. The punctuations such as (, / : /! ), and connectives such as (*و/w/and* and *بل/bl/but*) are used to present the boundaries of sentences. However, such punctuation usage is not systematic and not widely used in MSA.

[7] From an article written by Dr. Abdul-karem Bakar عبدالكريم بكار, one of the famous writers in contemporary Arabic literature, http://islamtoday.net/nawafeth/artshow-40-147981.htm

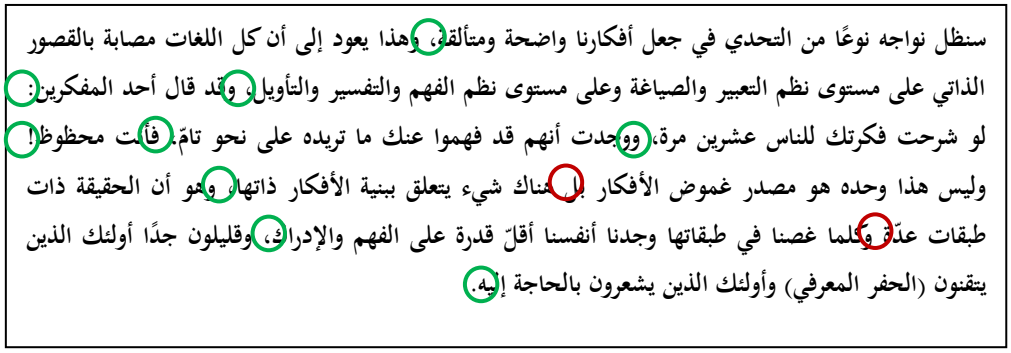



**Figure 3-3: Multiple sentences/clauses exist in one orthographic full-stop sentence. Other punctuations and connectives are used to separate sentences and clauses.**

**Arabic Discourse Connectives**: In the absence of a large categorized list of discourse connectives for Arabic, we noted that discourse connectives are not limited to the basic syntactic categorization of discourse connectives in the English PDTB (conjunctions, adverbial and prepositional phrases). For instance, prepositions also can link discourse segments when one or both arguments are al-maSdar nouns. Prepositions in English also have discourse functions in context but they were not annotated in the PDTB2. In addition, some nouns such as (*نتيجة/ntyjp/result*, *خشية/ks.yp/fear and رغبة/bqyp/desire*) are used as discourse connectives in Arabic. This is unlike English, nouns alone never have discourse function.

In addition, the discourse connectives in Arabic might occur: (i) individually such as (*لكن/lkn/however*), (ii) in conjunction with other connectives using the coordinating conjunction *و/w/and* such as (*لكن و قبل/lkn w qbl/however and before*), or (iii) as multiple connectives without conjunction such as (*الا بعد/AlA bEd/ except after*). More explanation about our collection of Arabic discourse connectives is given Chapter 4.

## 3.2 Discourse Processing for Arabic

Arabic is one of the challenging languages in front of the NLP community. The majority of Arabic language processing dealt with character, word and sentence levels: character recognition systems (Khorsheed 2003), semantic relations between words in WordNet systems (Elkateb *et al.* 2006), syntactic tagging (Maamouri, Bies and Kulick 2008), morphological analyzing (Al-Sughaiyer and Al-Kharashi 2004), stemming (Harmanani, Keirouz and Raheel 2006), spell checkers (Shaalan 2005). phrase chunking, sentence parsing and grammar checkers (Cavalli-Sforza and

Zitouni 2007; Chiang *et al.* 2006; Shaalan 2005). All of these processing tasks in Arabic NLP require cliticization, stemming or segmentation to strip clitics and suffixes as pre-processing steps (Harmanani, Keirouz and Raheel 2006). It is worth noting that huge efforts are still required to improve the performance in such studies for Arabic in order to achieve similar performance than for other languages such as English.

In contrast, almost no corpus linguistic studies have been dealt with regarding to the discourse level and how discourse segments are connected in Arabic. Few studies (Al-Sanie, Touir and Mathkour 2005; Seif, Mathkour and Touir 2005b; Khalifa and Farawila 2012) presented small non-corpus based studies on a number of RST-relations and discourse connectives. It is shown in these studies that discourse connectives play a critical role in linking discourse units and signalling discourse relations. Up to the study date, no annotated corpus, and no large list of discourse connectives and their relations exist for Arabic.

Discourse processing therefore remains a challenging field for the Arabic NLP community due to a lack of required resources such as annotated corpora and tools on the one hand, and reliable resources and algorithms for Arabic syntax and parsing, on the other hand.

### 3.2.1 Arabic Corpora

Collections of plain spoken/written data such as the Arabic Gigaword corpus[8], the Corpus of Contemporary Arabic[9] and Arabic Broadcast News Transcripts[10], are important resources for corpus-based studies in NLP. For more advanced studies such as building and evaluating statistical parsers, such as Standard Arabic Morphological Analyzer (SAMA 3.1)[11], special tokenization and syntactic analysis of sentences are required. However, due to the cost and time required for such annotation with long guidelines, only few small morphologically and syntactically annotated corpora exist for Arabic: the Penn Arabic Treebank (Maamouri *et al.* 2004; Maamouri, Bies and Kulick 2008; Maamouri, Bies and Kulick 2006), the Prague

[8] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T30
[9] http://www.comp.leeds.ac.uk/eric/latifa/research.htm
[10] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T20
[11] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010L01

Arabic Dependency Treebank (PADT) (Hajic *et al.* 2004), the Columbia Arabic Treebank (CATiB) (Habash and Roth 2009) and the Quranic Arabic Dependency Corpus (QAD) (Dukes and Buckwalter 2010).

Each of these treebanks has its own form of representation for modelling Arabic syntax. The QAD is CA, while the PATB, PADT and CATiB are MSA newswire corpora. Therefore, none of these treebanks are fully representative for MSA. The original newswire text annotated in these treebanks, does not show the diacritics and the hamzah symbols, and does not show the proper usage of punctuations. This increases the complexity and ambiguity of the automatic text processing using the raw text alone. However, the manually added POS tags and the tokenization in the treebanks can tackle such problems.

Syntactic tagging can make discourse connectives easier to identify, as they often belong to specific parts of speech such as conjunctions. In addition, the parse trees provide informative features for identifying discourse connectives, relations and argument boundaries automatically (see Chapter 8). Also, the additional discourse features can be used with the other syntactic and morphological features for different applications and studies in Arabic NLP.

Dukes and Buckwalter (2010) compare the four Arabic treebanks, as shown in Table 3-2. The column *feature* indicates if features such as *gender, lemma and verb moods* are included in the mark-up. The last column indicates whether the syntactic annotation considered the traditional Arabic grammar, which leads to minimize the training efforts for human annotation. Unlike the PATB, the PADT and the CATiB use dependency grammars for the newswire texts. In fact, both treebanks have used the PATB or some of its tools to develop their new treebanks and for annotating additional data (Habash and Roth 2009; Hajic *et al.* 2004). Also, the PATB's tokenization is considered standard for most Arabic treebanking efforts (Habash 2010).

Because of all these characteristics of the PATB and for its avaibility at the study time, it was chosen to be a base corpus for our discourse annotation[12].

[12] The PADT is smaller in size than the PATB.

**Table 3-2: A comparison of syntactic Arabic corpora. (Dukes and Buckwalter 2010, p.2).**

| Treebank | Dependency | Features | Traditional |
|---|---|---|---|
| Penn | no | yes | no |
| Prague | yes | yes | no |
| Columbia | yes | no | yes (subset) |
| Quran | yes (hybrid) | yes | yes |

### 3.2.2 The Penn Arabic Treebank - Part1 v.2

Among the few existing annotated treebanks, we decided to use the first part of the PATB (Maamouri *et al.* 2004) in this first effort to annotate discourse connectives and their relations in newswire text. At the end of the study, an additional discourse layer of the PATB (Part1 v 2.0) will be realised. It is named the Leeds Arabic Discourse Treebank (LADTB v.1).

The PATB uses syntactic annotation guidelines similar to the PTB for English (Marcus, Santorini and Marcinkiewicz 1993) after performing all necessary adaptations. It is a continuous project by the team at the University of Pennsylvania for annotating Arabic newswire corpora using Tim Buckwalter's lexicon and morphological analyzer. They generate an appropriate part of speech (POS) for each word in the corpus as well as a parse tree structure for each sentence (Maamouri *et al.* 2004a; Maamouri and Bies 2004). The PATB has many released parts Part1, 2, 3, and 4 (almost 650K words in total), with different versions through the Linguistic Data Consortium - LDC. Each version has a degree of improvement in the syntactic analysis.

The PATB has been used in different studies and applications in Arabic NLP such as tokenization, diacritization, part-of-speech (POS) tagging, morphological disambiguation, base phrase chunking, and semantic role labelling (Habash and Rambow 2004; Habash and Roth 2009; Dukes and Buckwalter 2010; Sadat and Habash 2006; Chiang *et al.* 2006). The treebanks are also used to provide empirical evidence for the frequency of Arabic linguistic constructions (Dukes and Buckwalter 2010).

The first version of the PATB (Part1) was released in January 2003. It consists of 734 files with roughly 166K words of written Modern Standard Arabic newswire text

from the Agence France Press (AFP) [13]. Most of the PATB sentences have been translated to English. Some have also been treebanked in English, creating a unique parallel resource.

### 3.2.3 Discourse Annotation Tools for Arabic

There is a need for an annotation tool to mark three components in discourse annotation (discourse connectives, their two arguments and relations) to ensure a reliable annotation. The few existing annotation tools, that at the study time could be used for discourse annotation, such as WordFreak (Morton and LaCivita 2003), GATE (Wilcock 2009) [14] and a prototype of a discourse annotation tool used for English in the PDTB annotation[15], did not fulfil the requirements for Arabic discourse annotation. One of the main reasons is that the discourse connectives in Arabic can be clitics attached to nouns, verbs, pronouns or adjectives. Also, the arguments, the second argument in particular, might start from the middle of a word. However, none of the available tools allow highlighting/marking parts of words.

Using the existing annotation tools to annotate the whole word that has the clitic connective might confuse the annotators, in which the rest of the word might play important role in annotating the right arguments and relations for the connective, on one hand. On the other hand, this method requires extra post-processing to expand the argument boundary to cover the rest of the word having the clitic connective. Unlike in Turkish discourse annotation (Zeyrek and Webber 2008), the connective clitic can be attached to verbs, nouns or pronouns. Thus, the post-processing might require another manual annotation effort.

In addition, the layout of the text in these tools is from left-to-right, which reflects wrong indices of the right-to-left Arabic text for connectives and arguments.

Moreover, the tool also should use the Arabic relation hierarchy for annotating the sense of a connective, which has some new relations not included in the tools used

[13] A new release of ATB Part1 was distributed at the summer of 2010. However, the collection study and the discourse annotation began in 2007 and was based on the older version, v. 2. Later, the University of Leeds was no longer a member of the LDC. Thus, we could not re-conduct the study on the new version.
[14] http://gate.ac.uk/
[15] It was thankfully provided by Alan Lee, the PDTB team's member.

for the PDTB annotation or other languages. Also, it is not possible to build a hierarchy structure for relations in Gate.

One option to overcome the shortcoming of existing tools is to expand the features of the annotation tool used to annotate syntactically the text in the PATB project to cover the discourse annotation requirements. However, this option was not possible in the study time.

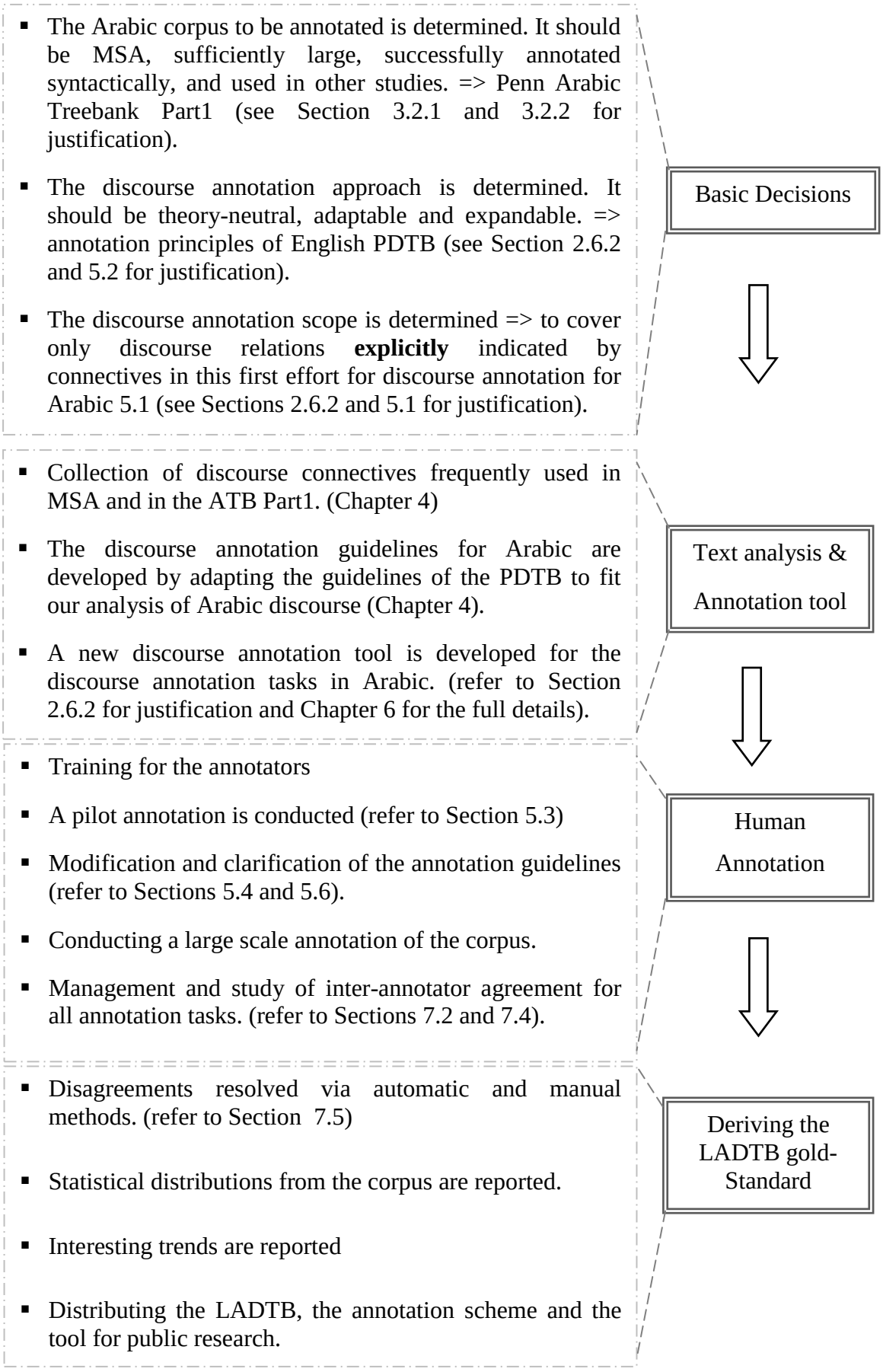
In response to all these special requirements, and to ensure a reliable annotation, we developed a dedicated annotation tool for Arabic discourse (READ), as one of the new resources this study provides to the community. The annotation is a stand-off style (based on the raw texts only), similar to the PDTB annotation. The syntactic annotation of the ATB is not displayed to the annotator in the tool, to ensure more flexibility and reliability (it can be used to annotate any new text with no syntactic annotation available).

## 3.3 Research Methodology

The objectives of this study can be grouped into two main targets: (i) creating the first Arabic Discourse Treebank, the Leeds Arabic Discourse Treebank (LADTB) and (ii) automated modelling of discourse relations for Arabic. For each target, we will use flowcharts to illustrate the process pipeline of the work, and to show the required integrated processes with justification of the major decisions we made.

### 3.3.1 Creating a Discourse Corpus for Arabic

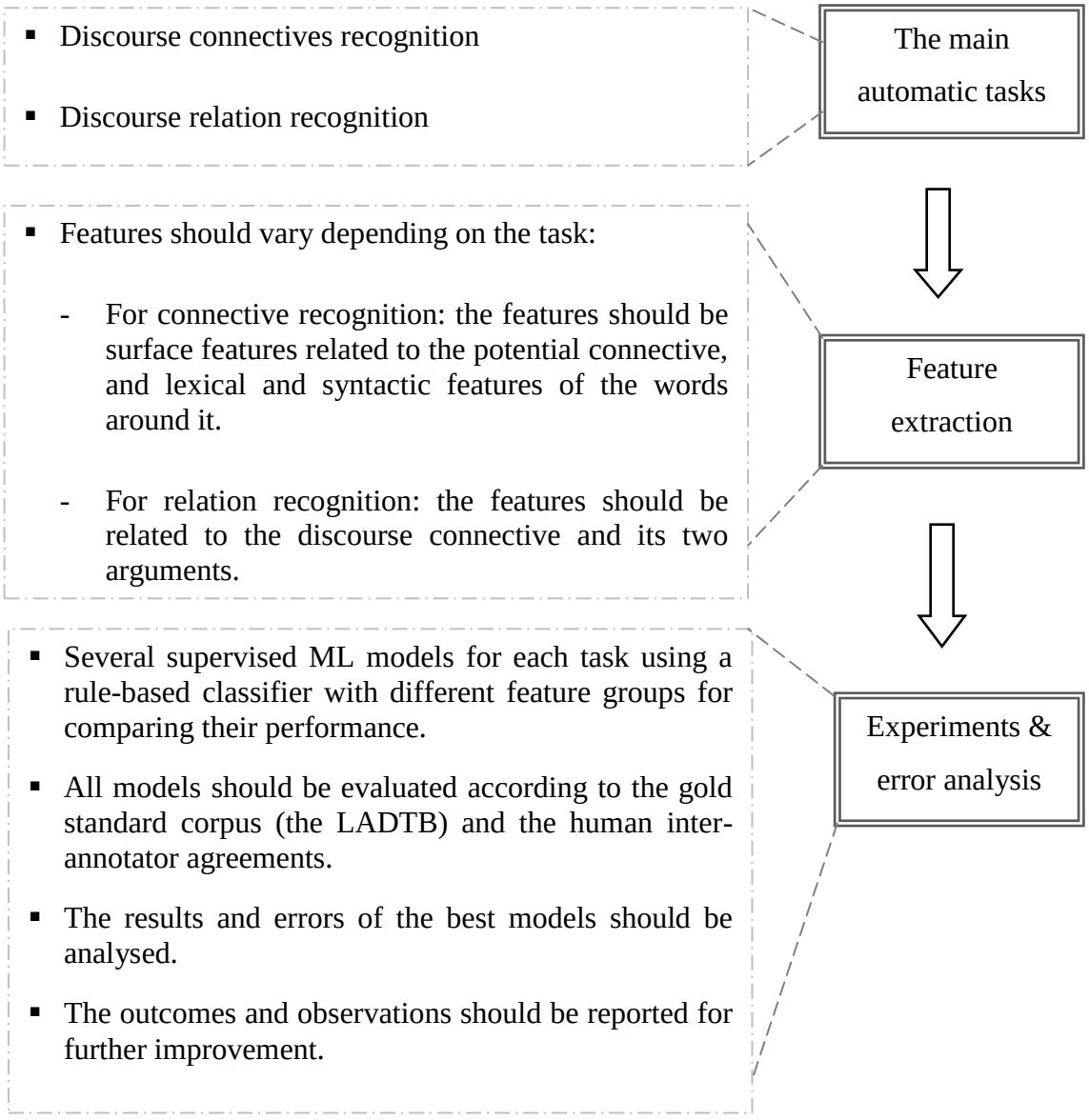
The Leeds Arabic Discourse Treebank (LADTB) is the first discourse corpus for Arabic that would enhance corpus linguistic studies as well as computational studies. It will be used as gold-standard for modeling discourse relations automatically. The flow chart below presents a general pipeline of the procedure of creating the LADTB. The details of the work are discussed in the relevant Chapters 4, 5, 6 and 7.


The Arabic corpus to be annotated is determined. It should be MSA, sufficiently large, successfully annotated syntactically, and used in other studies. => Penn Arabic Treebank Part1 (see Section 3.2.1 and 3.2.2 for justification).
The discourse annotation approach is determined. It should be theory-neutral, adaptable and expandable. => annotation principles of English PDTB (see Section 2.6.2 and 5.2 for justification).
The discourse annotation scope is determined => to cover only discourse relations explicitly indicated by connectives in this first effort for discourse annotation for Arabic 5.1 (see Sections 2.6.2 and 5.1 for justification).
Basic Decisions
Collection of discourse connectives frequently used in MSA and in the ATB Part1. (Chapter 4)
The discourse annotation guidelines for Arabic are developed by adapting the guidelines of the PDTB to fit our analysis of Arabic discourse (Chapter 4).
A new discourse annotation tool is developed for the discourse annotation tasks in Arabic. (refer to Section 2.6.2 for justification and Chapter 6 for the full details).
Text analysis & Annotation tool
Training for the annotators
A pilot annotation is conducted (refer to Section 5.3)
Modification and clarification of the annotation guidelines (refer to Sections 5.4 and 5.6).
Conducting a large scale annotation of the corpus.
Management and study of inter-annotator agreement for all annotation tasks. (refer to Sections 7.2 and 7.4).
Human Annotation
Disagreements resolved via automatic and manual methods. (refer to Section 7.5)
Statistical distributions from the corpus are reported.
Interesting trends are reported
Distributing the LADTB, the annotation scheme and the tool for public research.
Deriving the LADTB gold-Standard

### 3.3.2 Modeling of Discourse Relations for Arabic

The second group of the objectives of this thesis is to develop the first algorithms to detect automatically explicit Arabic discourse connectives and their relations. Since we created the LADTB as an informative discourse layer on top of the syntactic ATB, we are able to use supervised machine-learning models. Therefore, we chose rule-based classifiers for the two tasks (recognition of discourse connectives and disambiguating their functions). The rule-based classifier is a good technique to monitor the behaviour of the extracted features, and the rules across different models and data.

The features were extracted from the gold-standard tagging and tokenization in the ATB. However, for discourse connective recognition, we also use an automatic tagger and a simple tokenizer to record the performance in case of a new text, which does not have gold-standard syntactic annotation. The flowchart below presents our pipeline for the development of models for detecting discourse connectives and relations for Arabic. Argument boundaries identification is beyond the scope of this study but should be one of the first tasks to be addressed in future studies. The full details of the automatic modelling work are discussed in Chapter 8.


The main automatic tasks
Discourse connectives recognition
Discourse relation recognition
Feature extraction
Features should vary depending on the task:
For connective recognition: the features should be surface features related to the potential connective, and lexical and syntactic features of the words around it.
For relation recognition: the features should be related to the discourse connective and its two arguments.
Experiments & error analysis
Several supervised ML models for each task using a rule-based classifier with different feature groups for comparing their performance.
All models should be evaluated according to the gold standard corpus (the LADTB) and the human inter-annotator agreements.
The results and errors of the best models should be analysed.
The outcomes and observations should be reported for further improvement.

# Chapter 4 Collection of Discourse Connectives for Arabic

## 4.1 Introduction

Discourse connectives in Arabic such as لان/*lAn/because*, لكن/*lkn/but*, اذا/*A*A/if* and بعد/*bEd/after* are often used to improve text coherence. The most appropriate and readable discourse connectives are used by the author or speaker (Pitler and Nenkova 2008). Such discourse connectives have been used in studies for English, Turkish, Hindi, and Chinese as the anchors for discourse relations in human and automatic annotation (see Section 2.4 for a full discussion). Arabic also uses connectives frequently (see Section 7.7 for frequency study), studying connective types is an essential starting point for discourse studies for Arabic.

**Ex. 4-1**

أحمد لم يذهب الى الحفلة **لأنه كان متعبا**. **على النقيض لقد خلد الى النوم**.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| >Hmd | lm | y*hb | AlY | AlHflp | **l>nh** | kAn | mtEbA. | ElY | AlnqyD | lqd | xld | AlY | Alnw |
| Ahmad | not | go | to | the-party | **But-he** | was | tired. | In | contrast | was | stay | to | Sleep |

*Ahmad didn't go to the party*, **because** ***he was tired***. **Instead**, **he went to bed**.

In Ex. 4-1 the connective لان/*lAn/because* in the second clause establishes explicitly that the reason for Ahmad being absent from the party is that he was tired (Causal relation), whereas the connective على النقيض/*ElA Alnqyz/instead* in the third clause contrasts going to bed with going to the party (Contrast relation). The connective لان/*lAn/because*, therefore, takes clause 1 and clause 2 as its arguments. However, the second connective على النقيض/*ElA Alnqyz/instead* takes clause 1 and clause 3 as its arguments. It can be seen that there is no need for arguments to be adjacent, and they may differ in length and structure (see also the discussion in Section 2.3.3).

As mentioned in Section 3.2, there is no well-defined list of discourse connectives available for Arabic, nor does a corpus exist where the discourse connectives are annotated in context with regard to their discourse relations or arguments. The absence of such corpora and related studies for Arabic motivated our work in collecting potential discourse connectives.

This chapter describes our initial empirical efforts towards the first, extraction, and analysis of the frequently used discourse connectives in MSA. Thereafter, the proposed inventory of discourse connectives is used to create the first annotation scheme for annotating discourse connectives and associated discourse relations and arguments (Chapter 5), and develop the first discourse annotation tool for Arabic (Chapter 6). In addition, this inventory of Arabic discourse connectives is promising to enhance the discourse processing studies for Arabic theoretically and empirically. Bilingual studies of discourse will also benefit from the well-established inventory of discourse connectives for Arabic to compare the discourse features of different languages. Clarifying the differences and similarities of discourse connectives of Arabic and other languages will enhance computational applications such as machine translation from/to Arabic. We use the PATB Part1 to base our study on (Section 3.2.2 describes the corpus and justifies this decision).

The rest of this chapter is organized as follows: Section 4.2 describes the manual and automatic techniques of the collections work. Different types of connective and their grammatical categories are discussed in Section 4.3. The most common cases of ambiguity that arose in extracting discourse connectives and their relations automatically are reported in Section 4.4. Finally, we present our final inventory of discourse connectives in Section 4.5 which covers a wide variety of potential discourse connectives in MSA. Section 4.6 explores a comparison between Arabic and English discourse connectives using our collection and the connectives in the PDTB. A summary is then offered for the collection process of Arabic discourse connectives.

## 4.2 Collecting Arabic Discourse Connectives

First of all we have to define what a discourse connective is. As mentioned earlier in Sections 2.5.5 and 3.3.1, it was decided to use the same definition as was used in the PDTB and follow-on work for other languages. Thus we follow Miltsakaki, Prasad et al. (2006) in that we define discourse connectives as *lexical expressions that relate two text segments expressing abstract objects such as events, beliefs, facts or propositions*. The text segments are called *arguments* (Arg1 and Arg2) of a specific connective. This connective should indicate one or more discourse relations such as Elaboration, Exemplification, Contrast, Temporal, Exception, Causal or simply Conjunction.

It is claimed in (Prasad, Joshi and Webber 2010b) that discourse connectives in English are not a closed set and can be expanded to cover all expressions used to link discourse arguments. Thus the syntactic categories of discourse connectives in Arabic might exceed the predominant syntactic categories of English connectives (conjunctions, adverbial and prepositional phrases). Therefore, our discourse connective list should not be limited to the small set of connectives defined in the literature (see the first stage of our collection process), and this requires further discourse analysis to collect potential connectives in MSA.

**Table 4-1: Canonical forms of ordering arguments and discourse connectives in Arabic**

| | | |
|---|---|---|
| <Arg1. DC+Arg2> | <Arg1, DC+Arg2> | <Arg1+DC+Arg2> |
| <DC+Arg2, Arg1> | <DC+Arg2+Arg1> | <Arg1+DC+Arg2+Arg1> |
| <DCP1+Arg2+DCP2+ Arg1> | <DCP1+Arg2, DCP2 +Arg1> | |

We found from our analysis that the order of the connective DC and its arguments Arg1 and Arg2 might occur in the text following one of the canonical forms in Table 4-1. For example, the connective بعد*/bEd/after* in Ex. 4-2 is following the order <DC+Arg2, Arg1>. In the table, DCP1 and DCP2 are the first and second parts of the connective if it is a paired connective such as *if..then*…. The second argument Arg2 is syntactically introduced by the connective DC or DCP1, while the first argument Arg1 can occur prior (often) or after (rare) the second argument Arg2 in the text. In addition, it is not essential to have punctuation as clause-separators to determine the

argument boundaries. The argument is a proposition that includes necessary complements such as temporal adverbs.

**Ex. 4-2 (canonical form <DC+Arg2, Arg1>)**

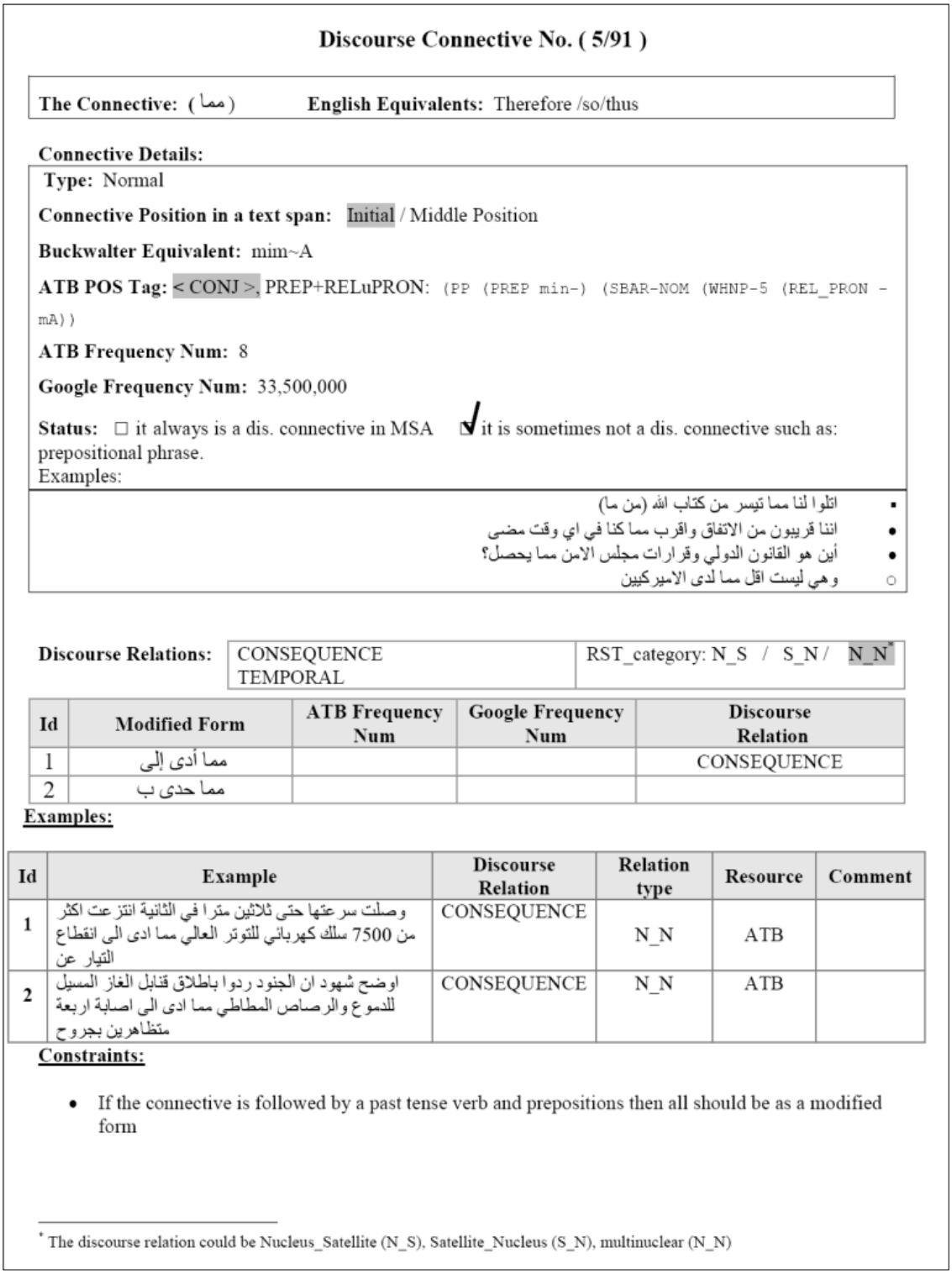
**بعد رحيلي عن القرية**، لم اشعر بالسعادة مجدداً

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| bEd | rHyly | En | Alqryp, | lm | A$Er | bAlsEAdp | mjddAF |
| after | Leaving-I | from | The-village, | Dont | feel | happiness | Again |

After **I left my home village**, *I never was happy again*.

In the discourse connective collection phase we were mostly interested in the nature of the discourse connective, where it occurs in the sentence, and what relation it typically signals. A template shown in Figure 4-1 is used to collect potential features of each connective. The syntactic sentence/clause boundaries were used initially to determine the argument boundaries. Therefore, the recording features do not specify all potential boundaries of the arguments. It is ensured that at least two examples are recorded in each form per connective. The properties of the connective describe the type, possible position, the discourse relations the connective usually signals, and its syntactic category in the ATB (POS tag) and in Arabic traditional syntax. At this stage, we did not restrict our analysis to connectives and relations of the PDTB. We started from Arabic itself and how the reader understands the discourse connections between abstract objects, with the basic annotation principles in mind.

The list of potential Arabic discourse connectives was collected by me – the researcher- in different stages without agreement measurements but with a subsequent check by a second native speaker – the supervisor Dr. Hussein Abdul-Raof. This list was later enhanced in a pilot annotation study. I used four main stages for collecting the discourse connectives:



**Discourse Connective No. ( 5/91 )**

**The Connective:** ( مما ) **English Equivalents:** Therefore /so/thus

**Connective Details:**

**Type:** Normal

**Connective Position in a text span:** Initial / Middle Position

**Buckwalter Equivalent:** mim~A

**ATB POS Tag:** < CONJ >, PREP+RELuPRON: (PP (PREP min-) (SBAR-NOM (WHNP-5 (REL_PRON -mA)))

**ATB Frequency Num:** 8

**Google Frequency Num:** 33,500,000

**Status:** ☐ it always is a dis. connective in MSA ☑ it is sometimes not a dis. connective such as: prepositional phrase.

Examples:

- اتلوا لنا مما تيسر من كتاب الله (من ما)
- اننا قريبون من الاتفاق واقرب مما كنا في اي وقت مضى
- أين هو القانون الدولي وقرارات مجلس الامن مما يحصل؟
- وهي ليست اقل مما لدى الاميركيين

**Discourse Relations:** CONSEQUENCE TEMPORAL | RST_category: N_S / S_N / N_N*

| Id | Modified Form | ATB Frequency Num | Google Frequency Num | Discourse Relation |
|---|---|---|---|---|
| 1 | مما أدى إلى | | | CONSEQUENCE |
| 2 | مما حدى ب | | | |

**Examples:**

| Id | Example | Discourse Relation | Relation type | Resource | Comment |
|---|---|---|---|---|---|
| 1 | وصلت سرعتها حتى ثلاثين مترا في الثانية انتزعت اكثر من 7500 سلك كهربائي للتوتر العالي مما ادى الى انقطاع التيار عن | CONSEQUENCE | N_N | ATB | |
| 2 | اوضح شهود ان الجنود ردوا باطلاق قنابل الغاز المسيل للدموع والرصاص المطاطي مما ادى الى اصابة اربعة متظاهرين بجروح | CONSEQUENCE | N_N | ATB | |

**Constraints:**

- If the connective is followed by a past tense verb and prepositions then all should be as a modified form

* The discourse relation could be Nucleus_Satellite (N_S), Satellite_Nucleus (S_N), multinuclear (N_N)

**Figure 4-1: An example of the template used in the discourse connective collection stage**

### 4.2.1 First Stage: Discourse Connectives in the Arabic Literature

I established the initial list by collecting all potential discourse connectives from different Arabic resources (Ryding 2005; Alansari 1985; Alfarabi 1990). In most literature books that I have reviewed (Ryding 2005; Alansari 1985; Alfarabi 1990; M. Abdl al latif, Zahran and Al-Arabi 1997; Dickins, Higgins and Hervey 2002), the discourse usage of some connectives such as conjunctions and adverbials are discussed alongside other usages such as the syntactic, semantic and theoretical usages. In Ex. 4-3 the connective ب*/b/by* in (Alansari 1985) has 14 functions with only one discourse usage but also without any clear distinction of the usages. In addition, most of the examples in the traditional literature books are from classic Arabic text (mostly from text of the 12th century and earlier); some usages are no longer used in contemporary MSA.

**Ex. 4-3 (Alansari 1985)**

**من كتاب مغني اللبيب عن كتب الأعاريب (النسخة الالكترونية)**

**الباء المفردة حرف جر لأربعة عشر معنى:**

**أولها الإلصاق قيل وهو معنى لا يفارقها فلهذا اقتصر عليه سيبويه ثم الإلصاق حقيقي ك أمسكت بزيد ....**

**الثاني التعدية وتسمى باء النقل أيضا وهي المعاقبة للهمزة في تصيير الفاعل مفعولا ..قوله تعالى * ولو شاء الله لذهب بسمعهم وأبصارهم * ....**

**الثالث الاستعانة وهي الداخلة على آلة الفعل نحو كتبت بالقلم ....**

**الرابع السببية نحو * إنكم ظلمتم أنفسكم باتخاذكم العجل * ...**

**الخامس المصاحبة نحو * اهبط بسلام * ....**

**والسادس الظرفية نحو * ولقد نصركم الله ببدر * * نجيناهم بسحر * ...**
**والسابع البدل كقول الحماسي * فليت لي بهم قوما إذا ركبوا ... شنوا الإغارة فرسانا وركبانا * ...**
**والثامن المقابلة وهي الداخلة على الأعواض نحو اشتريته بألف ....**

**والتاسع المجاوزة قيل تختص بالسؤال نحو * فاسأل به خبيرا *....**
**العاشر الاستعلاء نحو * من إن تأمنه بقنطار *....**

**الحادي عشر التبعيض * عينا يشرب بها عباد الله *....**

**الثاني عشر القسم وهو أصل أحرفه ولذلك خصت بجواز ذكر الفعل معها نحو أقسم بالله لتفعلن ...**

**الثالث عشر الغاية نحو * وقد أحسن بي * أي إلي وقيل ضمن أحسن معنى لطف ...**
**الرابع عشر التوكيد وهي الزائدة ....**

As early as this stage we noticed the occurrence of so-called *modified forms* of a connective, similar to English. The modified form connective consists of one of the basic connectives and an extra token which could be a pronoun, an adverb, or

another connective. For example, *في حين/حينها/when/while/at the same time* are modified forms of *حين/hyn/when,* and *بالرغم/برغم ان/برغم/although* are modified forms of *رغم/rgm/although*. These modified form connectives perform similarly in structure and functionality to the original connective. We therefore include in the connective list for Arabic all modified forms that we came across in our reading.

### 4.2.2 Second Stage: Manual Discourse Analysis of the ATB and the Internet

We have analysed around 50 random raw texts from the Penn Arabic Treebank (Penn ATB Part1), and have extracted all discourse connectives and their modified forms according to our definition of discourse connective. All new potential discourse connectives then were added into the list. Our aim was to build an extensive list of discourse connectives for MSA, not just from news only. Therefore, we analyzed an additional six articles from well-known Arabic websites (such as educational, political and social affairs) which were on average 600 words long.

Moreover, to ensure that frequently occurring discourse connectives were not missed, the English discourse connectives and modified forms in the PDTB were translated into Arabic. This process yielded 8 new connectives not yet in the Arabic list such as *in the meantime/في هذه الأثناء*, *in fact/في الواقع* and *in sum/باختصار* , which were added after a manual verification in context by using the internet to collect real examples or making-up acceptable Arabic examples.

### 4.2.3 Third Stage: Automatic Extraction of DCs from the ATB

Discourse connectives, in English for example, share properties such as syntactic category (conjunctions, adverbials and prepositional phrases). Thus to extract automatically unseen connectives for Arabic, we extracted automatically from the ATB all tokens that have similar syntactic categories (POS tags) to the discourse connectives in our list from Stage 2. For example, tokens that have CONJ tag were automatically extracted from the ATB and a random small set, around 5 instances on average, were manually examined in context to include connectives that have discourse function which were not yet in the list. In fact, we found as a result of this

process some discourse connectives which were not in the list such as طالما/*TAlmA/as long as*. Note, not all discourse connectives are annotated in the PDTB (Sections 2.3.2 and 7.7)

### 4.2.4 Fourth Stage: Ambiguity Status Estimation of DCs

Like other languages, not all Arabic connectives in our list always function as discourse connectives. Therefore, we extracted from the ATB examples of the connectives using the Buckwalter transliteration and examined manually how frequent the connectives have discourse usage in context on a random subset per each connective. We found that clitic and conjunction connectives are the most ambiguous connectives in terms of signalling discourse relations. Thus, we should conduct an agreement study of recognizing the discourse connectives in the human discourse annotation. Moreover, labelling the discourse connectives automatically using simple surface-based rules would probably not work for Arabic. This task requires a further study to determine the useful features that can be used, indeed. Refer to Sections 8.2 and 8.4 for more discussion about our experience in this study.

The collection process ended with a list of 107 discourse connectives overall including modified forms. The following sections describe in details properties of the Arabic discourse connectives and the main differences between Arabic and English discourse connectives.

## 4.3 Types of Discourse Connectives

As mentioned in Section 3.1, Arabic discourse connectives do not belong to only one syntactic category. Instead, they can be coordinating conjunctions, subordinating conjunctions, adverbials, prepositional phrases, nouns or prepositions. Moreover, the connective types might be simple (a single white space separated token), clitic (attached at the beginning or end of another token), or consist of more than one token (syntactical/non-syntactical phrase). Clitic and nouns connectives do not exist in English. In the following sections, we discuss common categories of discourse connectives, and provide examples for each category.

### 4.3.1 Coordinating Conjunctions

Two clauses or sentences can be joined by a co-ordinating conjunction such as *لكن/lkn/but*, *أو/Aw/or* or *و/w/and*. These conjunctions (ATB POS: CONJ) indicate respectively the discourse relations Contrast (Ex. 4-4), Alternative (Ex 4-5) and simply Conjunction (Ex 4-6).

**Ex. 4-4 (Contrast)**

السيارة متطورة جدا . **لكنها باهضة الثمن**

| AlsyArp | mtTwrp | jdA. | lknhA | bAhDp | Alvmn |
|---|---|---|---|---|---|
| The-car | modern | very | But-it | too-high | cost |

*The car is very modern*. **But it is too expensive.**

**Ex 4-5 (Alternative)**

**اما** ان تذهب الى البيت الآن **او تنتظرني ساعة واحدة**

| AmA | An | t*hb | AlY | Albyt | Al\|n | Aw | tntZrny | sAEp | wAHdp |
|---|---|---|---|---|---|---|---|---|---|
| either | that | You-go | to | home | now | Or | Wait-for-me | hour | One |

*You can go home now* **or you wait for me one hour**

**Ex 4-6 (Conjunction)**

أحمد يلعب كرة القدم، **و مريم تقرأ كتابا**

| >Hmd | ylEb | krp | Alqdm. | w | mrym | tqr> | ktAbA |
|---|---|---|---|---|---|---|---|
| Ahmad | play | ball | foot | and | Mary | read | book |

*Ahmad is playing football*, **and Mary is reading a book**

### 4.3.2 Subordinating Conjunctions

Subordinating conjunctions introduce clauses that are syntactically dependent on the main clause. In Arabic there are two kinds of subordinating conjunctions (similar to English, Chinese and Turkish):

**Simple subordinating conjunctions**: the subordinating clause is introduced by a subordinating conjunction such as *لان/lAn/because*, which indicates a Causal relation as in Ex. 4-7. The connectives *بينما/bynmA/while* and *حيث/Hyv/where/since* are also simple subordinating conjunctions.

**Ex. 4-7 (Causal)**

تم رفض الخطة المقترحة للمشروع **لأنها غير مستوفية للشروط**

| tm | rfD | AlxTp | AlmqtrHp | llm$rwE | l>nhA | gyr | mstwfyp | ll$rwT |
|---|---|---|---|---|---|---|---|---|
| done | denied | the-plan | the-suggested | for-project | because-it | not | comply | for-conditions |

*The proposed plan of the project has been rejected* **because it does not comply with the agreed terms**.

**Paired subordinating conjunctions**: Paired subordinating conjunctions consist of two non-adjacent lexical parts: the first introduces the subordinate clause Arg2 and the other introduces the main clause Arg1. Interestingly, these connectives are frequent in MSA. But they also occur sometimes as simple subordinating conjunctions (without using the second part). In Ex. 4-8 and Ex. 4-9, the paired connectives (.. رغم أن...الا ان /*although/despite*), and (..اذا ..فـ /*if…then*) indicate the discourse relations Contrast and Condition respectively. Note that they sometimes are translated with simple connectives in English, as seen in the examples.

**Ex. 4-8 (Contrast)**

**رغم ان الطائرات كانت تحلق باستمرار في السماء ،الا ان** *الحياة المدنية لم تتأثر*

| rgm | An | AlTA}rAt | kAnt | tHlq | bAstmrAr | fy | AlsmA', | AlA | An |
|---|---|---|---|---|---|---|---|---|---|
| Although | that | The-planes | were | flying | continously | in | the-sky, | but | that |
| AlHyAp | Almdnyp | lm | tt>vr | | | | | | |
| The-life | civilian | not | affected | | | | | | |

**Although planes were flying continuously in the city sky**, *civilian life was not affected*

**Ex. 4-9 (Condition)**

**اذا كان الجو صحواً ، فـ**لنلعب في الحديقة

| A*A | kAn | Aljw | SHwAF, | flnlEb | fy | AlHdyqp |
|---|---|---|---|---|---|---|
| if | was | weather | Clear, | Lets-play | in | the-garden |

**If the weather is fine**, *let's play in the garden*

### 4.3.3 Adverbial and Prepositional Phrase Connective

As in English, adverbial and prepositional phrase connectives in MSA are sentence-modifying connectives which express a discourse relation between two abstract entities. For example, the prepositional phrase connective *بالتالي/bAltaly/consequently* indicates a Consequence relation, while the adverbial connective *نتيجة ل/ntyjp l/as a result of* indicates a Causal relation. Adverbials also can be simple or paired, for example the connective *طالما..ف/TAlmA.. f../as-long-as* is a paired adverbial connective in Arabic, as can be seen in Ex. 4-10, but it is not paired connective in English.

**Ex. 4-10 (Pragmatic Condition)**

**طالما ان المؤتمر لم يحقق اهدافه ف**لن نجد من يثق بنتائجه لاحقا

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TAlmA | An | Alm&tmr | lm | yHqq | AhdAfh | Fln | njd | mn | yvq | bntA}jh | lAHqA |
| As long as | that | the-conference | not | achieve | its-objectives | then-not | find | from | trust | On-its-results | later |

**As long as the conference does not achieve its objectives,** *nobody will trust its findings later*

### 4.3.4 Preposition Connectives

Prepositions usually relate concrete objects, however, they might relate events or propositions. Prepositional connectives are often attached to al-maSdar nouns which express events or actions without indicating tense. Al-maSdar is a well-defined noun category in Arabic literature (Ryding 2005; Alansari 1985) and in some ways it corresponds to nominalization in English. For example, the al-maSdar noun *تبليغ/informing* in Ex. 4-11 is a valid argument for the preposition connective *ل/l/due to/for*. More details about al-maSdar have been given in Section 5.4.1. Appendix A also presents the common morphological forms of al-maSdar nouns. We consider al-maSdar nouns as arguments in our annotation guidelines for Arabic (Section 5.2).

Prepositional clitic discourser connectives such as *ل/l/due to/for* and *ب/b/by* are usually attached to al-maSdar nouns. However, not all prepositional connectives are clitics in Arabic. Some subordinating conjunctions in English such as *بعد/bEd/after*,

*قبل/qbl/before* and *منذ/mn*/since* corespond to prepositions in Arabic followed by, but not attached to, Al-maSdar nouns such as in Ex. 5-11 and Ex. 7-1. Table 4-7 lists the common prepositional connectives in the ATB.

**Ex. 4-11 (Causal)**

ذهبنا الى مركز الشرطة **للتبليغ عن فقدان وثائق الشركة الرسمية**

| *hbnA | AlY | mrkz | Al$rTp | lltblyg | En | fqdAn | wvA}q | Al$rkp | Alrsmyp |
|---|---|---|---|---|---|---|---|---|---|
| went -we | to | centre | police | For-informing | about | loss | documents | company | Official |

*We went to the police station* **in order to report the loss of the company official documents.**

## 4.3.5 Noun Connectives

One of the interesting findings of our analysis is that nouns in Arabic can function as discourse connectives. They occur as (i) simple nouns such as *بغية/bgyp/desire* and *نتيجة/ntyjap/result*, or (ii) combined nouns with a preposition such as *فضلا عن/fdlA En/as well as* or attached to the function word *ان/An/that* such as *بيد ان/byd An/but*. Both the noun connective *بيد ان/byd An/but* and the conjunction connective *لكن/lkn/but* are subordinators and can be swapped in many cases. However, the usage of *بيد ان/byd An/but* is very formal.The noun connectives *بغية/bgyp/desire* and *نتيجة/ntyjap/result* have also a semantic content themselves.

Ex. 4-13 shows the ATB syntactic annotation of the example of the noun connective *بغية/bgyp/desire.* The connective is introduced with a mark –PRP which represents a modifier showing purpose or cause in the syntactic analysis. However, the syntactic analysis does not always show the semantic function of the connective. For instance, the ATB analysis (NP-ADV (NP (NOUN+NSUFF_FEM_SG ntyjp) (NP (NP (NOUN Drb) for the discourse connective *نتيجة/ntyjap/result* in (..نتيجة طرد/ result of expulsion of ..) introduced a adverbial NP but does not show any semantic function. The noun connectives require a special corpus-linguistic study on more data to define the relation between their syntactic and discourse functions. This study is out of scope of this thesis.

**Ex. 4-12 (Contrast)**

كانت حياته مستقرة **بيد ان الظروف لم تسمح له ان يكون تاجرا**

| kAnt | HyAth | mstqrp | byd | An | AlZrwf | lm | tsmH | lh | An | ykwn | tAjrA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| was | his-life | stable | but | that | circumstances | not | allow | him | that | Be | Businessman |

*His life was stable* **but circumstances did not allow him to be a businessman**

**Ex. 4-13 (Causal)**

في ١٢ حزيران نشرت الحكومة لائحة ب ٨٠٤ مزارع **بغية نزع الملكية عنها**

| fy | HzyrAn | n$rt | AlHkwmp | lA}Hp | b | 804 | mzArE | bgyp | nzE | Almlkyp | EnhA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| in | July | announced | government | list | of | 804 | farmer | desire | Taking-out | ownership | From-it |

*In July, the government announced a list of 804 farmers* **in order to remove the possession from them**

(S (PP-TMP (PREP fy) (NP (NUM 12) (NP (NOUN_PROP HzyrAn)))) (VP (VERB_PERFECT+PVSUFF_SUBJ:3FS n$rt) (NP-SBJ (DET+NOUN+NSUFF_FEM_SG AlHkwmp)) (NP-OBJ (NP (NOUN+NSUFF_FEM_SG lA}Hp)) (PP (PREP b) (NP (NUM 804) (NOUN mzArE)))) (NP-PRP **(NOUN+NSUFF_FEM_SG bgyp)** (NP (NP (NOUN nzE) (NP (DET+NOUN+NSUFF_FEM_SG Almlkyp))) (PP (PREP En) (NP (PRON_3FS hA))) ))) .

## 4.4 Ambiguity Problems

In this first effort to collect Arabic discourse connectives in the ATB, the text analysis was based mainly on manual recognition of discourse connectives but enhanced by automatic process, as discussed in Section 4.2. Some problems, however, arose in this automatic process and highlighted the complexity of recognising Arabic discourse connectives. Arabic has a complex morphology; connectives do not have to correspond only to a separate word or a well-defined phrase as in English. The Arabic discourse connective can occur as a prefix clitic to a verb or noun, such as *ف/f/then*, *لكن/lkn/but*, and *بعدما/bEdmA/after that*, or a sequence of words that is not a syntactic phrase such as *فضلاعن/fDlAEn/as well as* and *نظرا لـ/nZrA l/because of*.

In addition, the connective could introduce an al-maSdar noun phrase (discourse connective) and other nouns (non-discourse connective) as well. Thus, we recognised

at an early stage that a strong linguistic competence is essential to distinguish the type of nouns after the potential connectives. This task is not trivial and is confusing especially for nouns having three or four letters (similar to the root but with different diacritic marks). Making correct decisions in annotation requires intensive practice plus the linguistic experience as well. For example, the preposition عند/*End/when* is rarely used to signal a discourse relation, but it is a discourse connective when followed by al-maSdar noun such as انفجار/*explosion* in Ex. 4-14, where it indicates a Cause relation.

**Ex. 4-14 (Causal relation)**

لقي *18* شخصا مصرعهم **عند** **انفجار انبوب نفط في نيجيريا**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| lqy | 18 | $xSA | mSrEhm | End | AnfjAr | Anbwb | nfT | fy | nyjyryA |
| faced | 18 | person | their-death | when | explosion | tube | oil | in | Nigeria |

*18 people were killed* **when** **an oil pipeline was blown up in Nigeria**

Furthermore, considerable ambiguity related to surface formation arose when we collected the instances of connectives from the ATB automatically. For instance, the absence of the hamzah (ء) and diacritics (ً,ٌ,ٍ,ّ,ْ,َ,ُ,ِ) in the ATB and in the raw text led to ambiguity whether for example الا is the connective الا/*AlA/except* or the question word ألا. Also, the connective اذا/*A*A/if* can be confused with the non-connective اذاً.

In addition, the Arabic TB Part1 v.2 which we used in our study, has several annotation mistakes such as frequently assigning wrong POS tags or inconsistent Buckwalter transliterations. This lack of consistency reduced the benefit of using the POS tag as a good indicator to find similar discourse connectives. For example, the connective حيث/*Hyv/where/since* has two POS tags in the ATB: CONJ and REL_ADV. The connective حيث/*Hyv/where/since* could not be a conjunction.

## 4.5 Final Inventory of Arabic Discourse Connectives

The discourse connectives collection process resulted in a list containing 91 basic Arabic discourse connectives, enhanced with 16 modified forms, yielding 107

discourse connectives overall. This number is comparable to the number of 100 distinct English connectives in the PDTB. We noted that MSA reflects greater variety in usage than in English, where a few connectives are very common, and many more are much less common. See Section 4.6 and Section 7.7 for more discussion on distribution and frequency. The connectives are categorized by the syntactic status as annotated in the ATB and presented in Table 4-2 to Table 4-8. The position of the connective at third column is either at beginning of a sentence (BOS) or at middle of a sentence (MOS). Note: the POS tags in the last column are according to version 2 of the ATB Part1. They might be modified slightly in the new version of the ATB. The Arabic connectives are ordered alphabetically in the tables. Their frequency in the LADTB is presented in Appendix B.

Moreover, our analysis of connectives recorded their discourse relations as indicated in the examined instances. In consequence, we can develop our relation taxonomy as discussed in Chapter 5. Table 4-9 lists the discourse connectives we collected from resources other than the ATB Part1 (refer to Section 4.2.2). Two connectives (listed in Table 4-6) consist of preposition and a relative pronoun, do not fit on any syntactic classes in Section 4.3

**Table 4-2: The coordinating conjunction connectives in the LADTB.**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| إذ/A*/as | Simple | B/MOS | CONJ |
| أو/Aw/or | Simple | MOS | CONJ |
| ف/f/then | Clitic | B/MOS | CONJ |
| لكن/lkn/but | Simple, Clitic | B/MOS | CONJ, NO_FUNC |
| و/w/and | Simple, Clitic | B/MOS | CONJ |

**Table 4-3: The subordinating conjunction connectives in the LADTB.**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| إذا/A*A/if | Simple | B/MOS | CONJ |
| إلا/AlA/except | Simple | MOS | EXCEPT_PART |
| إلا إذ/AlA A*A/except if | MoreThanToken | MOS | EXCEPT_PART+CONJ |

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| *الا ان/AlA An/but* | MoreThanToken | MOS | EXCEPT_PART+Func_word |
| *الا بعد/AlAbEd/expect after* | MoreThanToken | MOS | EXCEPT_PART+PREP, PREP+PREP |
| *اما/AmA/while* | Simple | BOS | PREP |
| *انما/AnmA/but* | Simple | B/MOS | CONJ |
| *حيث/Hyv/where/since* | MoreThanToken | MOS | PREP+CONJ |
| *بسبب/bsbb/because of* | Simple | B/MOS | PREP,PREP+NOUN |
| *بعدما/bEdmA/after that* | Simple | B/MOS | CONJ, RELuADV |
| *بل/bl/but* | Simple | B/MOS | CONJ |
| *بمعنى آخر/bmEnYxr/in other words* | MoreThanToken | B/MOS | PREP+NOUN |
| *بينما/bynmA/while* | Simple | B/MOS | CONJ,REL_ADV |
| *عندما/EndmA/when* | Simple | MOS | CONJ,REL_ADV |
| *غير ان/gyr An/however* | MoreThanToken | B/MOS | NEG_ PART+FUNC_WORD |
| *حيث/Hyv/where/since* | Simple | MOS | CONJ, REL_ADV |
| *كأن/k<n/as* | Simple | MOS | CONJ |
| *كلما/klmA/when ever* | Simple | B/MOS | CONJ |
| *كما/kmA/as* | Simple | B/MOS | CONJ |
| *كي/ky/to* | Simple | MOS | CONJ |
| *لذا/l*A/for this* | MoreThanToken | B/MOS | CONJ |
| *لاسيما/lAsymA/particularly* | Simple | B/MOS | NEG_PART+ADV |
| *لان/lAn/because* | Simple, Clitic | B/MOS | CONJ |
| *لكي/lky/for/in order to* | Simple | B/MOS | CONJ |
| *لو/lw/if (in past)* | Simple | MOS | CONJ |
| *لولا/lwlA/if not* | Simple | B/MOS | PREP |
| *طالما/TAlmA/as long as* | Simple | BOS | CONJ |
| *وقبل/wqbl/and before* | MoreThanToken | BOS | NONE |

**Table 4-4: The noun connectives- single and modified nouns in the LADTB**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| *اضافة الى/ADAfpAlY/in addition to* | MoreThanToken | MOS | NOUN+PREP |
| *بغية/bgyp/desire/to* | Simple | MOS | NOUN, PREP |
| *بيد/byd/but* | Simple | B/MOS | NOUN |
| *بيد ان/byd An/but* | MoreThanToken | B/MOS | NOUN+FUNC_WORD |
| *فضلا عن/fDlAEn/as well as* | MoreThanToken | B/MOS | NOUN+PREP |
| *حينها/HynhA/when that* | MoreThanToken | B/MOS | NOUN+POSS_PRON |
| *نتيجة/ntyjp/result of* | Simple | MOS | NOUN |
| *قبيل/qbyl/shortly before* | Simple | MOS | NOUN, PREP |
| *رغم/rgm/though* | Simple | B/MOS | NOUN, PREP |
| *رغم ان/rgm An/although* | MoreThanToken | B/MOS | NOUN+FUNC_WORD, PREP+FUNC_WORD |
| *خلافا ل/xlAfA l/unlike* | MoreThanToken | B/MOS | NOUN+PREP |
| *نظرا ل/nZrA l/because of* | MoreThanToken | B/MOS | NOuFUNC+PREP, NOUN+NO_FUNC, NOUN+PREP |

**Table 4-5: The Adverbial connectives in the LADTB**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| *ايضا/AyDA/also* | Simple | B/MOS | ADV |
| *حال/HAl/when* | Simple | B/MOS | NONE |
| *حتى/HtY/until* | Simple | B/MOS | ADV, CONJ, PREP |
| *حتى لو/HtYlw/even if* | MoreThanToken | B/MOS | ADV+CONJ |
| *حين/Hyn/when* | Simple | B/MOS | ADV |
| *كذلك/k*lk/and that* | Simple | B/MOS | ADV, NOUN |
| *لذلك/l*lk/for that* | MoreThanToken | B/MOS | ADV |
| *من ثم/mn vm/then* | MoreThanToken | MOS | PREP+ADV, PREP+NOUN |
| *ثم/vm/then* | Simple | MOS | ADV |
| *خصوصا/xSwSA/specially* | Simple | B/MOS | ADV |

**Table 4-6: The (preposition + relative pronoun) connectives in the LADTB.**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| *فيما/fy mA/while* | MoreThanToken | B/MOS | CONJ, PREP+REL PRON |
| *مما/mmA/which (+ past verb)* | MoreThanToken | MOS | CONJ, PREP+REL_PRON, REL_PRON |

**Table 4-7: The preposition connectives in the LADTB.**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| *أثر/Avr/after* | Simple | MOS | PREP |
| *ب/b/by* | Clitic | B/MOS | PREP |
| *بعد/bEd/after* | Simple | B/MOS | PREP |
| *عقب/Eqb/shortly after* | Simple | B/MOS | PREP |
| *جراء/jra/because* | Simple | MOS | PREP |
| *ل/l/for* | Clitic | MOS | EMPHATIC_PARTICLE, PREP, RuCuP, SUBJUNC |
| *منذ/mn*/since* | Simple | B/MOS | CONJ, NOuFUNC, PREP |
| *قبل/qbl/before* | Simple | B/MOS | PREP |
| *قبل ان/qbl An/before that* | MoreThanToken | B/MOS | PREP+FUNC_WORD |
| *خلال/xlAl/during* | Simple | MOS | PREP |

**Table 4-8: The prepositional phrase connectives in the LADTB.**

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| *بالمقابل/bAlmqAbl/in contrast* | MoreThanToken | B/MOS | PREP+NOUN |
| *بفضل/bfDl/thanks to* | Simple | MOS | PREP+NOUN |
| *بهدف/bhdf/in order to* | MoreThanToken | MOS | PREP+NOUN |
| *برغم/brgm/although* | Simple | B/MOS | PREP+NOUN |
| *بالاضافة الى/bAlADAfpAlY/in addition to* | MoreThanToken | B/MOS | PREP+NOUN |

| Dis. Conn | Type | Position | ATB POS |
|---|---|---|---|
| بالرغم من*bAlrgm mn/although* | MoreThanToken | B/MOS | PREP+NOUN |
| بالتالي*bAltAly/consequently* | MoreThanToken | B/MOS | ADV,PREP+NOUN |
| على الرغم*ElY Alrgm/although* | MoreThanToken | B/MOS | PREP+NOUN |
| في المقابل*fy AlmqAbl/in contrast* | MoreThanToken | B/MOS | PREP+NOUN |
| في حال*fy HAl/in case* | MoreThanToken | B/MOS | PREP+NOUN |
| في حين*fy Hyn/while* | MoreThanToken | B/MOS | PREP+ADV, PREP+NOUN |
| في ظل*fy Zl/under* | MoreThanToken | B/MOS | PREP+NOUN |

**Table 4-9: Discourse connectives in MSA that do not occur in the ATB Part1.**

| Dis. Conn | Type | Position | Syntactic Class |
|---|---|---|---|
| *على العموم/ElY AlEmwm/in general* | MoreThanToken | BOS | Adverbial |
| *مثلا/mvlA/for example* | Simple | B/EOS | Adverbial |
| *باختصار/bAxtSAr/briefly/in sum* | MoreThanToken | BOS | Adverbial, prepositional phrase |
| *بالاساس/bAlAsAs/basically* | MoreThanToken | M/EOS | Adverbial, prepositional phrase |
| *بالاضافة/bAlADAfp/in additionto* | MoreThanToken | BOS | Adverbial |
| *بالفعل/bAlfEl/in deed* | MoreThanToken | B/M/EOS | Adverbial, prepositional phrase |
| *بحجة أن/bHjp<n/because of* | MoreThanToken | B/MOS | Subordinating conj |
| *بعد ذلك/bEd *lk/after that* | MoreThanToken | BOS | Subordinating conj |
| *جدير بالذكر/jdyr bAl*kr/ it should be noted* | MoreThanToken | BOS | Subordinating conj |
| *ختاما/xtAmA/finally* | Simple | BOS | Adverbial |
| *خلاصة/xlASp/to sum up* | Simple | BOS | Adverbial |
| *دليلا على/dlylA ElY/evidence for* | MoreThanToken | MOS | Adverbial |
| *ذلك ان/*lk An/that because* | MoreThanToken | BOS | Subordinating conj |
| *علاوة على/ElAwp ElY/in addition to* | MoreThanToken | BOS | Adverbial |
| *على العكس/ElY AlEks/by opposite* | MoreThanToken | BOS | prepositional phrase |
| *على النقيض/ElY AlnqyD/In contrast* | MoreThanToken | BOS | prepositional phrase |
| *على سبيل المثال/ElY sbyl AlmvAl/for example* | MoreThanToken | BOS | prepositional phrase |
| *عموما/EmwmA/generally* | Simple | BOS | Adverbial |

| Dis. Conn | Type | Position | Syntactic Class |
|---|---|---|---|
| فعلا/fElA/indeed | Simple | M/EOS | Subordinating conj |
| في الواقع/fy AlwAqE/of course/ in fact | MoreThanToken | BOS | Subordinating conj |
| في أعقاب/fy <EqAb/after all | MoreThanToken | MOS | prepositional phrase |
| في هذه الاثناء/fy h*h AlAvnA /in the meantime | MoreThanToken | BOS | Subordinating conj |
| كدليل/kdlyl/as an evidence | MoreThanToken | EOS | Adverbial |
| لاجل/lAjl/for | MoreThanToken | B/MOS | Subordinating conj |
| لهذا السبب/lh*AAlsbb/for this reason | MoreThanToken | BOS | Subordinating conj |
| لئلا/l>lA/for not | MoreThanToken | MOS | Subordinating conj |
| نتيجة ل/ntyjp l/resulted by | MoreThanToken | B/MOS | Subordinating conj |
| وفي الختام/fy AlxtAm/finally | MoreThanToken | BOS | prepositional phrase |

## 4.6 Comparison with English

We conducted a comparison of Arabic and English discourse connectives using our collection of Arabic discourse connectives and the English connectives in the PDTB2. We defined a set of similarities and differences. Overall, both languages share basic discourse characteristics including the connectives (function, position and type), discourse relations and arguments (type and order in the text). However, Arabic has more variety in nature of its explicit connectives. For instance, clitics and nouns were considered as discourse connectives for Arabic, as they, according to our definition of discourse connective, link two valid propositions. Prepositions are discourse connectives in both languages but they are not annotated in the PDTB2.

Some connectives in Arabic do not have equivalent connectives in English. For instance, the connective اثر/*Avr/after* is translated always into *after* but it has an additional causal meaning over the usual temporal connective بعد/*bEd/after*. It is rarely translated into the connective *since.* The connective اثر/*Avr/after* has a causal function more than a temporal function. Similarly, some Arabic connectives lose their function as connectives when translated into English such as اما/*AmA* and

*و/w/and* at BOS. Also, it is not required in English to use the second part *ف/f/then* of some paired connectives such as *اذا/A*A/if*, but it is often used in Arabic.

On the other hand, there are different connectives in Arabic that are translated into the same connective in English. For example, the connectives *الا ان/AlA An/but*, *انما/AnmA/but*, *بل/bl/but*, *بيد ان/byd An*, *بيد/byd/but*, *لكن/lkn/but* and *غير ان/gyr An/however/but* are translated into *but/however* in English. This diversity might reflect the different strength of the discourse relation (Contrast) that connectives indicate. A deep bilingual corpus-study would be needed in order to prove such a hypothesis, and could be very useful for translation studies. Also, it might be required sometimes to add other adverbs to the connective in English such as *only* and *rather* to get the same usage of only the connective in Arabic, as in Ex. 4-15.

**Ex. 4-15 (Contrast)**

ان قضية فلسطين ليست قضية وطنية **بل** **مسالة تهم العالم الاسلامي اجمع**

| An | qDyp | flsTyn | lyst | qDyp | wTnyp | bl | msAlp | thm | AlEAlm | AlAslAmy | AjmE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| that | problem | Palestine | not | issue | national | but | issue | concern | the-world | Islamic | all |

*The Palestine problem is not only a national problem* **but rather** **a matter of concern for the entire Islamic world**

Interestingly, all fine-grained Conditional relations (General, Unreal_Past, Factual_Past, Unreal_Present and Factual_Present) in the English PDTB are indicated by just the basic conditional connective *if* or one of its modified forms. However, there is a wide range of connectives in Arabic ( لولا- طالما - في حال - اذا - ما - حال دام- لو) which can signal different fine-grained conditional relations. For example, the relation Unreal_Past is signalled often by *لو/lw/if (in the past)* in Arabic and not by *اذا/A*A/if*. Again, a deeper comparison study is needed to generalise this finding linguistically.

## 4.7 Summary

We described in this chapter the first large-scale collection of Arabic discourse connectives, resulting in a large repository of 107 potential discourse connectives for

Arabic. The total of Arabic discourse connectives in our list is comparable to the number of 100 distinct English connectives in the PDTB. This first discourse connective repository for Arabic was collected using manual and automatic techniques to ensure a high coverage of frequently used connectives in MSA.

The collection was enhanced by mining the properties of the connectives and by to including discourse relations they might signal using a detailed template that list real-life examples from the ATB and contemporary articles from the Internet. We have also described the ambiguity problems that we faced during the automatic discourse analysis which shows the difficulty of identifying discourse connectives in Arabic text automatically.

Although Arabic and English share many discourse features, there are also interesting differences shown in our analysis which can be used to enhance language studies and applications. We would encourage other linguistic researchers to recognise and study further these similarities and differences, in order to foster understanding of the two languages, and develop further empirical applications.

The collection of discourse connectives for Arabic was subsequently used for discourse annotation in context, which formed the next stage in this study pipeline. Firstly, the text analysis needed to be integrated with the discourse annotation principles of the PDTB (Prasad et al. 2008) in a manner compatible with the properties of Arabic. The result of that was the creation of new discourse annotation guidelines for Arabic, as discussed in Chapter 5. Secondly, a new discourse annotation tool for Arabic was developed to annotate our collection of Arabic discourse connectives, their relations and arguments in context (see Chapter 6).

# Chapter 5
# Discourse Annotation Guidelines for Arabic

## 5.1 Introduction

We present the first discourse annotation guidelines for Arabic in this research. The annotation scheme is based on similar discourse annotation principles as in the PDTB project for English (Prasad *et al.* 2008a). We first developed the scheme according to our analysis of discourse features in MSA using the basic definitions of discourse connectives and relations as described in Chapter 4. Then we mapped our analysis to the annotation guidelines of the PDTB, adding all necessary adaptations to produce the final discourse annotation guidelines for Arabic (Appendix B).

The most attractive features of the PDTB are that its developer designed a theory–neutral approach for annotating local discourse relations, with few restrictions as to the position of discourse connectives and related arguments. Section 2.6.2 presents more details. In addition, the annotation scheme of the PDTB can be adapted by adding more restrictions or annotation layers to fit with other existing discourse structure theories (i.e. RST-tree or graph) that have many successful applications in computational linguistics. The PDTB annotation guidelines have been also successfully adapted and tested in recent years for other languages such as Hindi ((Prasad *et al.* 2008b), Turkish (Zeyrek and Webber 2008) and Chinese (Xue 2005). Using similar principles in annotating discourse in different languages has the potential to improve bilingual studies and applications, and generalize theories and discourse properties across language barriers.

In this chapter, we will demonstrate in Section 5.2 the basic annotation principles in our scheme that are similar to the English ones. The adaptations and the new principles in annotating discourse connectives, arguments and relations in Arabic,

which resulted from our discourse analysis and the pilot annotation, are presented in Sections 5.4 and 5.6. In Section 5.6.3, we have designed some techniques to help annotators disambiguate discourse connectives. Some special cases are described in Sections 5.6.3 - 5.8 to overcome frequent disagreements in this first effort for annotating Arabic discourse connectives and relations. Section 5.5 presents the finalized hierarchy of discourse relations for Arabic which is tested practically in the pilot annotation. The chapter concludes with a summary of our work in developing the first discourse annotation scheme for Arabic in addition to recommendations for expanding the scheme.

## 5.2 Basic Annotation Principles

The discourse annotation in our study concentrates on annotating *explicit* discourse connectives and associated arguments and relations they convey. Definitions of our terms, following the terminology in the PDTB (Prasad *et al.* 2008a) are repeated here for a complete view of the annotation principles. Discourse connectives are lexical expressions that relate two text segments that express abstract objects (AOs) such as events, beliefs, facts or propositions. We refer to the text segments as arguments (Arg1 and Arg2). Figure 5-1 shows a diagram of the definition of discourse connectives. The discourse connectives can be simple (لكن/*lkn/but*), paired ( إذا.. ...ف/*a*a..f../if..then*), modified forms (بالرغم من/*bAlrqm mn/although*) and have different syntactic categories. Types of connectives are described with examples in Section 4.3 and in the annotation guidelines in Appendix B. Similar to the PDTB, we annotate multiple connectives such as و لكن */w lkn/and but* separately as two independent connectives, although they might share one or two arguments. Both arguments must express AOs and be related explicitly via a connective. If this is not the case, we do not annotate the connectives as discourse connectives.

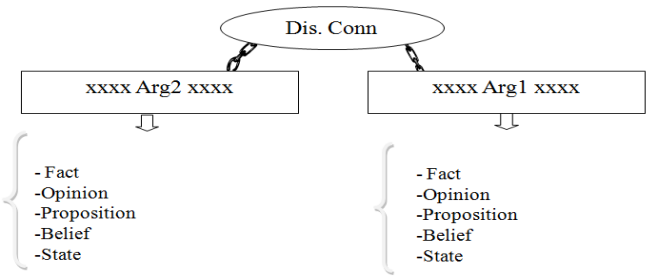



**Figure 5-1 The annotation definition of discourse connectives**

**Arguments** can be simple clauses or sentences, sequences of sentences, or nominalizations. they are also be adjacent or non-adjacent, with no restrictions on position or order. The only restriction is that Arg2 is always the argument that is introduced by the connective. We also apply the so-called minimality principle introduced by (Prasad *et al.* 2008a) in our annotation scheme, in that only the text representing the AO is considered as a valid argument. However, the argument should also include any necessary complements to the AO.

**Discourse relations** are grouped into four main classes: Temporal, Expansion, Contingency and COMPARISON, similar to (Prasad *et al.* 2008a). Each class has at maximum two levels of fine-grained relations (see Section 5.8 for the Arabic relation taxonomy). An instance of a connective can indicate more than one relation, and if so they should all be annotated.

We do not annotate attributions or implicit relations in this first discourse annotation analysis for Arabic as this is beyond the scope of a single thesis. Concentrating on only explicit connectives was also the theme of the very first version of the PDTB (Webber and Prasad 2006).

## 5.3 The Pilot Annotation

We test the initial annotation scheme with the basic principles in a pilot annotation on 121 texts from the ATB in stages by two native speakers having a good linguistic knowledge. At this early stage we had used the annotation tool designed for the

PDTB[16]. However, this tool was not compatible with Arabic because, for example, the highlighting considers only white-space-tokens rather than part of the words as required in Arabic, a language with high morphological complexity (see Section 3.1). Thus, set of preprocessing and post-processing tools were developed to tackle these problems. We decided later to develop a proper discourse annotation tool for Arabic, (see Chapter 6). Although we made progress in improving the inter-annotator agreement on connectives and relations over the annotation stages, the average agreement for connectives was still low, only 90%, and the average agreement for relations did not exceed 60%.

We realized that achieving a highly reliable annotation for Arabic discourse connectives is not a straightforward task. Therefore, we discussed intensively the adaptations required in the annotation scheme for Arabic and tested them practically in the latest stages of the pilot annotation.

## 5.4 Adaptations for Identifying Discourse Connective and Arguments

The required adaptations and additions were made in order to tackle the special characteristics of Arabic. Some connectives may operate either with or without a discourse function in the text. Thus, the identification of discourse connectives is directly related to the identification of the correct arguments. Firstly here, the new guidelines for identifying arguments are discussed, then those that concern the identification of discourse connectives.

### 5.4.1 Al-maSdar nouns

**Al-maSdar** is a well-known noun category that expresses events without tense. These events are eligible for being arguments of discourse relations. Al-maSdar patterns and their construction procedure are discussed earlier in Section 3.1. Al-maSdar nouns can be the full argument alone, or with additional complements. They can be arguments for any connective type. In particular, preposition connectives are

[16] Alan Lee thankfully provided us a prototype for the new discourse annotation tool for the PDTB project.

always followed by al-maSdar nouns or their negation. The al-maSdar argument is usually located at the first or second place in Arg2. It is also allowed to have al-maSdar nouns on both arguments Arg1 and Arg2. In Ex. 5-1, *تبليغ/informing* is the al-maSdar form of *بلغ/inform*, which acts as argument for the preposition connective *ال/l/for*. InEx. 5-2, *انعدام/lack* is the al-maSdar form of *عدم/reduce* and the argument of the prepositional phrase connective *بسبب/bsbb/because of*.

Identifying al-maSdar nouns requires the linguistic ability to check whether a noun after the potential connective fits one of the al-maSdar patterns in Appendix A. Section 8.4.1 describes an algorithm for detecting al-maSdar nouns automatically.

**Ex. 5-1 (Causal)**

ذهبنا الى مركز الشرطة **للتبليغ عن فقدان وثائق الشركة الرسمية**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *hbnA | AlY | Mrkz | Al$rTp | lltblyg | En | fqdAn | wvA}q | Al$rkp | Alrsmyp |
| gone | to | centre | police | inform | that | loss | documents | company | official |

*We went to the police station* **in order to inform about the loss of the company official documentsz**

**Ex. 5-2 (Causal)**

أن كبسولة الانقاذ لم تتمكن من الالتحام بالغواصة **بسبب انعدام الرؤية**.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| >n | kbswlp | AlAnqA* | Lm | ttmkn | mn | AlAltHAm | bAlgwASp | bsbb | AnEdAm | Alr&yp |
| that | capsule | rescue | not | could | from | attach | submarine | because of | lack | vision |

*The rescue capsule could not be attached to the submarine* **because of the lack of visibility**

## 5.4.2 The Order of Arguments

In Arabic, discourse connectives and their arguments follow different canonical forms in text. Figure 5-2 summarises the potential ordering of Arabic discourse connectives (DCs) and their two arguments (AOs) Arg1 and Arg2. This was also discussed ealier in Section 4.1. The two main canonical forms are the linear orders <Arg1+DC+Arg2> and <DC+Arg2+Arg1>, which are the sequences used mainly for simple connectives. On the other hand, there is only one possible canonical form for paired connectives: <DCP1+Arg2+ DCP2+Arg1> where DCP1 and DCP2 stand respectively for the first and second parts of the paired connective. It is often the case

in Arabic news that Arg2 and the connective divide Arg1 into two parts. We see this in the final sequence in Figure 5-2. Ex. 5-3 and Ex. 5-4 present examples of different sequences of discourse connectives, and their two arguments. More examples are presented in the actual annotation scheme which is attached in Appendix B.

The discourse connective might occur at the beginning of a sentence/clause or at the middle, but not at the end. Unlike English, we did not come across any case of sentence-final connectives in our text analysis and the pilot annotation.

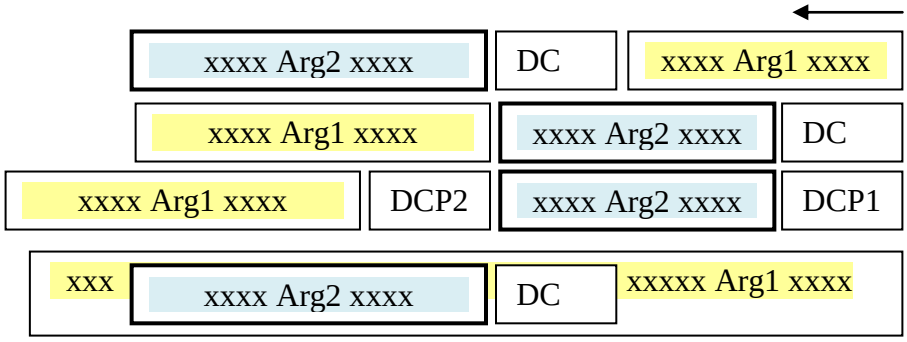



**Figure 5-2: Different sequences of discourse connectives, and their two arguments in Arabic text (to be read from right-to-left).**

**Ex. 5-3 (a canonical form <Arg1+DC+Arg2>)**

تم رفض الخطة المقترحة للمشروع **لأنها غير مستوفية للشروط**

| tm | rfD | AlxTp | AlmqtrHp | llm$rwE | l>nhA | gyr | mstwfyp | ll$rwT |
|---|---|---|---|---|---|---|---|---|
| done | denied | the-plan | the-suggested | for-project | Because-it | not | comply | for-conditions |

*The proposed plan for the project has been rejected* **because it does not comply with the agreed terms.**

**Ex. 5-4 (a canonical form <DCP1+Arg2+ DCP2+Arg1>)**

**رغم ان الطائرات كانت تحلق باستمرار في السماء ،الا ان** الحياة المدنية لم تتأثر

| Rgm | An | AlTA}rAt | kAnt | tHlq | bAstmrAr | fy | AlsmA' | AlA | An |
|---|---|---|---|---|---|---|---|---|---|
| Although | that | The-planes | were | flying | continously | in | The-sky | but | that |
| AlHyAp | Almdnyp | lm | tt>vr | | | | | | |
| The-life | ceivelian | not | affected | | | | | | |

**Although planes were flying continuously in the city sky**, *civilian life was not affected*

### 5.4.3 New Potential Discourse Connectives

During the pilot annotation, annotators came across some new potential connectives to be added to our connective list, such as the nouns *عقب/Eqb/shortly after*, *قبيل/qbyl/shortly before*, *بغية/bgyp/desire to*, *جراء/jra/because* and a prepositional phrase *في أعقاب/in the following*. The new potential connectives were added to our connectives list for Arabic after a double manual verification of several examples retrieved from the internet.

### 5.4.4 The Connective و /w/and

The conjunction *و/w/and* is the most frequent potential connective in Arabic texts. It is a very flexible conjunction, used in Arabic to join nouns, numbers, adjectives, prepositional phrases, clauses, sentences, paragraphs and other connectives as well. It also introduces almost every paragraph and sentence in newswire text in order to produce a coherent report. It can also signal any discourse relation. Thus, unsurprisingly, the connective *و/w/and* is the most ambiguous of all connectives, presenting the most difficulty when it comes to determining discourse function or discourse relations.

The annotators on the project were encouraged to pay more attention when dealing with the connective *و/w/and*, in order to distinguish discourse and non-discourse connective instances, and to identify arguments correctly. In particular, when *و/w/and* occurs at the beginning of a paragraph (BOP) in news text such our corpus, all prior propositions could be valid arguments to be linked with the argument introduced by the connective *و/w/and*. Therefore, it was decided that those instances of the connective *و/w/and* at BOP should be seen as relating to the closest potential proposition and a Conjunction relation was assigned, unless clearer discourse relations were explicitly indicated.

### 5.4.5 The Connective حيث/hyv/where-since-when

The potential connective *حيث/hyv/where-since-when* is often used to refer to a place or time in prior text, such as in Ex. 5-5. In these cases, it is a relative pronoun

without discourse function. However, it sometimes has a discourse function by relating two AOs such as in Ex. 5-6, where it relates the change and how this change happened. In order to attempt to distinguish between the two functions of this connective, the syntactic annotation was considered. However, the ATB annotates the discourse connective *حيث/hyv/where-since-when* inconsistently with different POS tags and analysis. A special case study was designed for this potential discourse connective, consisting of several examples to show how the connective *حيث/hyv/where-since-when* should not refer to time or places in prior proposition when it has discourse usage.

The connective *حيث/hyv/where-since-when* is similar to *when* in English, which can function as a relative pronoun as in *the time in May when I visited Leeds* or a subordinating conjunction *when I visited Leeds, I stopped at the Art Gallery* or a complementizer (*I know when I should go home*.). Note that *when* is not always translated into *حيث/hyv/where-since-when* in Arabic.

**Ex. 5-5**

كان محتشمي شغل في الثمانينات منصب سفير ايران في دمشق **حيث** *اصيب بجروح خطيرة في انفجار استهدفه* عام ١٩٨٢

| | | | | | | |
|---|---|---|---|---|---|---|
| kAn | mHt$my | $gl | fy | AlvmAnynAt | mnSb | sfyr |
| Was | Mohteshmi | held | in | eighties | position | ambassador |
| AyrAn | fy | dm$q | Hyv | ASyb | bjrwH | xTyrp |
| Iran | in | Damascus | where | injured | wounded | serious |
| Fy | AnfjAr | Asthdfh | EAm | 1982m | | |
| In | explosion | attack-him | year | 1982 | | |

Mohteshmi held a position 'Iran's ambassador' in Damascus in the eighties, **where** he was seriously wounded in bomb attack on him in 1982

**Ex. 5-6 (Reformulation)**

طرأ تعديل على نادي اللاعبات **حيث ارتقت الاسبانية ارانشا مرتبة واحدة و تبادلت المركزين التاسع و العاشر مع الالمانية انكه**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Tr> | tEdyl | ElY | nAdy | AllAEbAt | Hyv | Artqt | AlAsbA-nyp | ArAn$A | mrtbp |
| occur | change | on | club | players | where | raised | Spanish | Arancha | position |
| wAHdp | w | tbAdlt | Almrkzyn | AltAsE | w | AlEA$r | mE | AlAlmAnyp | Ankh |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| one | and | exchange | positions | nineth | and | thinth | with | German | Anke |

*There was a change to the club of female players* **where the Spanish Arancha rose one rank and swapped ninth and tenth places with German Anke**

### 5.4.1 The Clitic Connectives

Arabic has many clitics functioning as discourse connectives in context. The clitics can be attached to pronouns such as *لكن/lkn/but* in *لكنه/lknh/but-he*, to verbs such as *ف/f/then* in *فقال /then-said*, or to nouns such as *ل/l/for* in *للحد/llhd/for-limiting*. The clitic connectives have different syntactic categories, which determines what words they can be attached to. For example, *ف/f/then* is a conjunction while *ل/l/for is a preposition*. The prepositions cannot be attached to a verb.

The successful identification of clitic discourse connectives is strongly affected by correctly determining whether the token attached to the clitic is part of a valid argument. For instance, the prepositional clitic connectives *ل/l/for* and *ب/b/by* must be attached to al-maSdar nouns in order to act as discourse connectives.

## 5.5 Hierarchy of Discourse Relations

In common with the English PDTB and projects based on other languages, our discourse relation taxonomy has a hierarchical structure for more flexibility and reliability. We share with others (Prasad *et al.* 2008a; Prasad *et al.* 2008b; Zeyrek and Webber 2008; Xue 2005) the same main four classes: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. Each class has a number of fine-grained relation types, and some of them have further subtypes for more detailed relations. From the text analysis that we had done in the first place to collect Arabic discourse connectives, we realised that most of the discourse relations in the PDTB also exist in Arabic text (see Section 2.3.4). Thus after running a pilot annotation, we determined the frequently used relations in our news corpus. For example, we merge the very rarely used fine-grained relations that would confuse annotators and lead to low agreement among them. The hierarchy of Arabic discourse relations is shown in

Figure 5-3 after applying the adaptations and addition of relations that will be discussed in the next section.

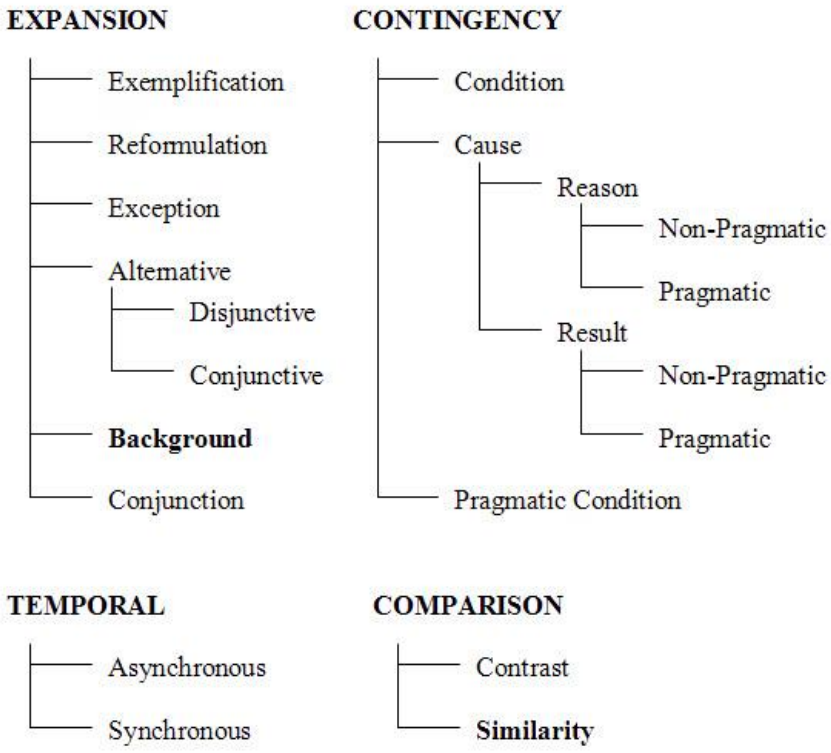



**Figure 5-3: The hierarchy of discourse relations for Arabic**

## 5.6 Adaptations for Discourse Relation Annotation

Overall, the definitions of the majority of relations from the PDTB were taken over unchanged. This section presents two types of adaptations that were made for discourse relations in Arabic: simplification of relations and adding new relations. These adaptations were made in the scheme before the final run of the discourse annotation on which agreement was measured (see Section 7.2).

### 5.6.1 Relation Hierarchy Simplification

#### Expansion.List Relation

Expansion.List is defined in the PDTB scheme as follows:

"*List applies when Arg1 and Arg2 are members of a list, defined in the prior discourse*." (Prasad *et al.* 2007b, p.42).

However, annotators in the pilot annotation study often disagreed on the Expansion.List relation, and argued that only the Expansion.Conjunction relation can be applied correctly, especially when the list theme is absent. Thus, it was decided to exclude List relation from our EXPANSION relations inventory in this study, and use a sequence of Conjunction relations instead.

### Fine-grained Relations under EXPANSION.Reformulation

EXPANSION.Reformulation has three fine-grained relations (Specification, Generalization and Equivalence) in the PDTB hierarchy. However, although they all seemed to occur in Arabic, annotators often disagreed when it came to distinguishing between relations of EXPANSION.Reformulation. A decision was therefore made to merge them in this study and retain the more general relation EXPANSION.Reformulation in our taxonomy. More detailed, deeper annotation would be required in future, as we expect these relations to be important for some applications such as automatic summarization. Louis and Nenkova (2011) have used Expansion.Specification to devise a classifier for 'general' vs. 'specific' sentences in English, which they claim will be useful for work in automated extractive summarization.

### Pragmatic Contrast Relations

COMPARISON.Contrast, CONTINGENCY.Condition, CONTINGENCY.Reason and CONTINGENCY.Result relations might be indicated pragmatically with an indirect relation. However, the annotators did not often capture Pragmatic Contrast relations, and there was an argument about them in the majority of its instances in the pilot study. A decision therefore was made in this study to merge direct and indirect contrast relations into one relation COMPARISON.Contrast.

### General Conditional Relation

There are not enough instances of the PDTB fine-grained relations of CONTINGENCY.Condition such as General, Unreal_Past, Factual_Past, Unreal_Present and Factual_Present in our analysis and pilot annotation. Thus, in

this first discourse study for Arabic, we merge them into the upper-level relation CONTINGENCY.Condition. Inclusion of text from, for example, instruction books would be useful to increase the variety of the conditional discourse usage.

### 5.6.2 Introduction of Novel Relations

Two new relations, EXPANSION.Background and COMPARISON.Similarity, were introduced during our analysis of discourse connectives for Arabic.

#### EXPANSION.Background

The type "Background" applies when Arg2 describes a situation related to a prior situation in Arg1 by giving background information in order to give the reader a wider view of the situation in Arg1. For example, Arg2 in Ex. 5-7 presents information about the war in Iraq and how it began. Similarly in Ex. 5-8, Arg2 gives information about the task of the Lebanese delegation. In both examples, the relation is more than a combination of Temporal. Asynchronous and Contingency.Cause.Reason. Arg2 gives background information for a full understanding of the argument in Arg1.

**Ex. 5-7**

غادر الرئيس جورج بوش العراق بخيبة أمل من إيجاد حل سياسي للحرب على الإرهاب في العراق. **وقد بدأت الحرب في العراق عام *2005* اثر مزاعم امريكية بنية العراق امتلاك سلاح نووي**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| gAdr | Alr}ys | jwrj | bw$ | AlErAq | bxybp | >ml | mn |
| left | president | George | Bush | Iraq | disappointed | political | from |
| <yjAd | Hl | syAsy | llHrb | ElY | Al<rhAb | fy | AlErAq. |
| having | solution | political | War | on | terrorism | in | Iraq |
| wqd | bd>t | AlHrb | fy | AlErAq | EAm | 2005 | Avr |
| where | starts | war | In | Iraq | year | 2005 | after |
| mzAEm | Amrykyp | bnyp | AlErAq | AmtlAk | slAH | nwwy | |
| Allegations | American | intention | Iraq | acquiring | weapon | nuclear | |

*President George W. Bush, left Iraq disappointed not to have found a political solution to the war in Iraq*. **(and) The war in Iraq began in 2005 after U.S. allegations that Iraq had the intention of acquiring nuclear weapons**

**Ex. 5-8**

ان الطائرة التي تقل الوفد اللبناني الرسمي وصلت اليوم الثلاثاء الى طرابلس. **وكان قد اتى الوفد لاصطحاب الرهينة اللبنانية ماري ميشال معربس المحتجزة في الفيليبين**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| An | AlTA}rp | Alty | tql | Alwfd | AllbnAny | Alrsmy | wSlt |
| that | plane | which | carrying | delegation | Lebanese | official | arrive |
| Alywm | AlvlAvA' | AlY | TrAbls. | wkAn | qd | AtY | Alwfd |
| today | Tuesday | to | Tripoli | was | that | come | delegation |
| lASTHAb | Alrhynp | AllbnAnyp | mAry | AlmHtjzp | fy | alflbyn | |
| to-accompany | hostage | Lebanese | Marie | hostage | In | Philippines | |

*The plane, which was carrying the official Lebanese delegation, arrived in Tripoli on Tuesday*. **(and) The delegation came to accompany the Lebanese hostage Marie, who was held in the Philippines**.

## COMPARISON.Similarity

The type *Similarity* applies when the connective indicates that the two arguments express similar abstract objects. It is therefore a complement to the contrast relation. The two arguments in Ex. 5-9 are presenting a similar action in how one feel when miss (home-country in Arg1) and (a small child in Arg2).

**Ex. 5-9**

انك تتألم من فراق الوطن **كما تتألم الأم على فقد رضيعها**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ank | tt>lm | mn | frAq | AlwTn | kmA | tt>lm | Al>m | ElY | fqd | rDyEhA |
| You | suffering | from | leaving | home-country | as | suffering | mother | on | losing | her-child |

*You are suffering from leaving your home country* **as a mother suffers from losing her child**

Our identification of a Comparison.Similarity relation led the PDTB group to notice that this was also a gap in the set of senses for English discourse relations and that instances of "*just as*" in the corpus had been annotated incorrectly: They should have been annotated with this sense[17].

[17] This comment was by Bonnie Webber in person, 2012.

### 5.6.3 Special Case: Conjunction Relation

The Conjunction relation was often assigned in the pilot study as a second relation in combined relations, due to the conjunction function of the majority of discourse connectives in news texts. This leads to an increase in annotator bias, and so to over estimate of partial agreement in the inter-annotator agreement study. A decision was therefore made to prevent the combination of a Conjunction relation with other relations in the scheme. As a result, Conjunction relation is only assigned if and only if there is no another relation indicated by the connective.

### 5.6.4 Special Case: Entity-based Relation and Conjunction

An argument might express information about one or more entities in prior discourse but not the AOs. This is a case of entity-based coherence (annotated in the PDTB with the label EntRel). Unlike in the PDTB, annotating entity relations is beyond the scope of this first discourse study for Arabic. However, in Arabic Arg2 in such relation instances are often introduced by an explicit connective such as *و/w/and* (see example Ex. 5-10). If so, we treat these entity relations in a similar way to discourse relations. In the majority of the cases, the entity relation is assigned a Conjunction relation and the arguments should cover almost the entire sentences/clauses such as in Ex. 5-10.

**Ex. 5-10 (Conjunction)**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| وصل رئيس الوزراء من رحتله الى الشرق الأوسط. **و التي خصصت لبحث مفاوضات السلام المتوقفة.** | | | | | | | |
| wSl | r}ys | AlwzrA' | mn | rHtlh | AlY | Al$rq | Al>wsT. |
| arrive | President | minister | from | trip | to | The- East | Middle |
| w | Alty | xSSt | lbHv | mfAwDAt | AlslAm | Almtwqfp | |
| And | which | allocated | find | negotiations | peace | expicted | |
| *The Prime Minister arrived from his trip to the Middle East*, **(and) which was allocated to discuss the stalled peace negotiations** | | | | | | | |

### 5.6.5 Special Case: Temporal and Causal relations

Causal relations, whether to do with reason or result, imply a temporal sequence of their abstract objects. Thus, there is normally no need to annotate both temporal and

causal relations when annotating causal connectives. However, connectives, that are usually used to indicate the temporal order of AOs such as قبل/*qbl/before*, بعد/*bEd/after* and عقب/*Eqb/shortly after*, should be dealt with differently if they can indicate causal relations as well. In these cases, both relations should be assigned to those instances as multiple relations. In Ex. 5-11, travelling away from the person's home village is the (implied) reason for never being happy again. The relation here is a combination of TEMPORAL.Asynchronous and CONTINGENCY.Reason.Non-Preagmatic.

The same situation occurs in the annotation of the PDTB, with the subordinating conjunction *since*, which is ambiguous between Temporal.Succession (but not causal), Causal.Reason (but not temporal) and both. In English, Causal.Reason does not imply a temporal sequence, as in "*I am unhappy since I am not with you*". Only causal connectives لان/*lAn/because* can used for this example in Arabic " *انا لست سعيدا لأني لست معك*". It may have both senses in English and Arabic, as in "*I have been unhappy since you left/انا لست سعيدا منذ رحيلك*".

**Ex. 5-11**

**بعد رحيلي عن القرية**، لم اشعر بالسعادة مجددا .

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| bEd | rHyly | En | Alqryp | lm | A$Er | bAlsEAdp | mjddAF |
| after | leaving | from | The-village | not | feel | happiness | again |

**After I left my home village**, *I was never happy again*.

## 5.7 Techniques for Disambiguating Discourse Connectives

We have developed some techniques to assist annotators disambiguating discourse connectives in context correctly according to our annotation scheme.

### 5.7.1 Connective Substitution

In the pilot study, annotators disagreed more on assigning discourse relations than on the identification of connectives. A *substitution technique* was therefore developed, to be applied to instances of non-Conjunction and non-Background relations. The technique is based on the substitution of a connective that is ambiguous with regards to the relation it signals with a less ambiguous connective indicating a clear relation. The stronger connective with the same relational function was substituted temporarily in order to test the function of the original, and to make it possible to determine its function correctly. The two connectives should indicate the same relation, not change the writer's intention in the discourse.

The technique can be applied many times with different, less ambiguous connectives, as it was permitted to annotate more than one discourse relation (multiple relations). Thus the connectives of discourse relations should be tested in order as presented, in Table 5-1.

For example, the annotator replaces the original connective with the first connective في المقابل*/fyAl mqAbl/in contrast*.

- ⇨ If the connective fits smoothly with the context and gives a roughly similar meaning that the author intends to present, then the relation COMPARISON.Contrast is the correct relation to assign to the original connective.
- ⇨ If the meaning is only partially complete, try other substitutions. It could be a combined relation.
- ⇨ If the first substituted connective does not express the right meaning, try the next suggested connective in the table, and so on.

This technique is useful for connectives of low ambiguity in terms of relations. Thus, Conjunction and Background relations are excluded from the substitution technique as they are often signalled by softer ambiguous connectives such as و*/w/and*, which can indicate any relation in our taxonomy.

**Table 5-1: A sequence of substitutions for disambiguating discourse connectives in terms of relations**

| | Substituted connective(s) | Discourse Relation | Further examination |
|---|---|---|---|
| 1 | *في المقابل/fyAl mqAbl/in contrast* | COMPARISON.Contrast | |
| ٢ | *لذا/l*A/for that*<br>*نتيجة لذلك/ntyjp l*lk/as a result*<br>*بالتالي/bAlfEl/consequently*<br>*So, Thus* | CONTINGENCY. Result | Try also No. 4 if the original connective has temporal meaning |
| ٣ | *بسبب/bsbb/because of*<br>*لان/lAn/because* | CONTINGENCY. Reason | |
| ٤ | *بعد/bEd/after*<br>*ثم/vm/then* | TEMPORAL.Asynchronous | |
| ٥ | *خلال/xlAl/during*<br>*بالتزامن/bAltzAmn/at the same time* | TEMPORAL.Synchronous | |
| ٦ | *باستثناء/b<stvnA/except*<br>*الا/<lA/except* | EXPANSION.Exception | |
| ٧ | *او/Aw/or*<br>*كبديل/kbdyl/as alternative* | EXPANSION.Alternative | |
| ٨ | *على سبيل المثال/Ela sbyl AlmvAl/for example* | EXPANSION.Exemplification | |
| ٩ | *خصوصا/xSwSA/specially*<br>*عموما/EmwmA/generally*<br>*بعبارة أخرى/bEbArp Axra/in other words* | EXPANSION.Reformulation | |

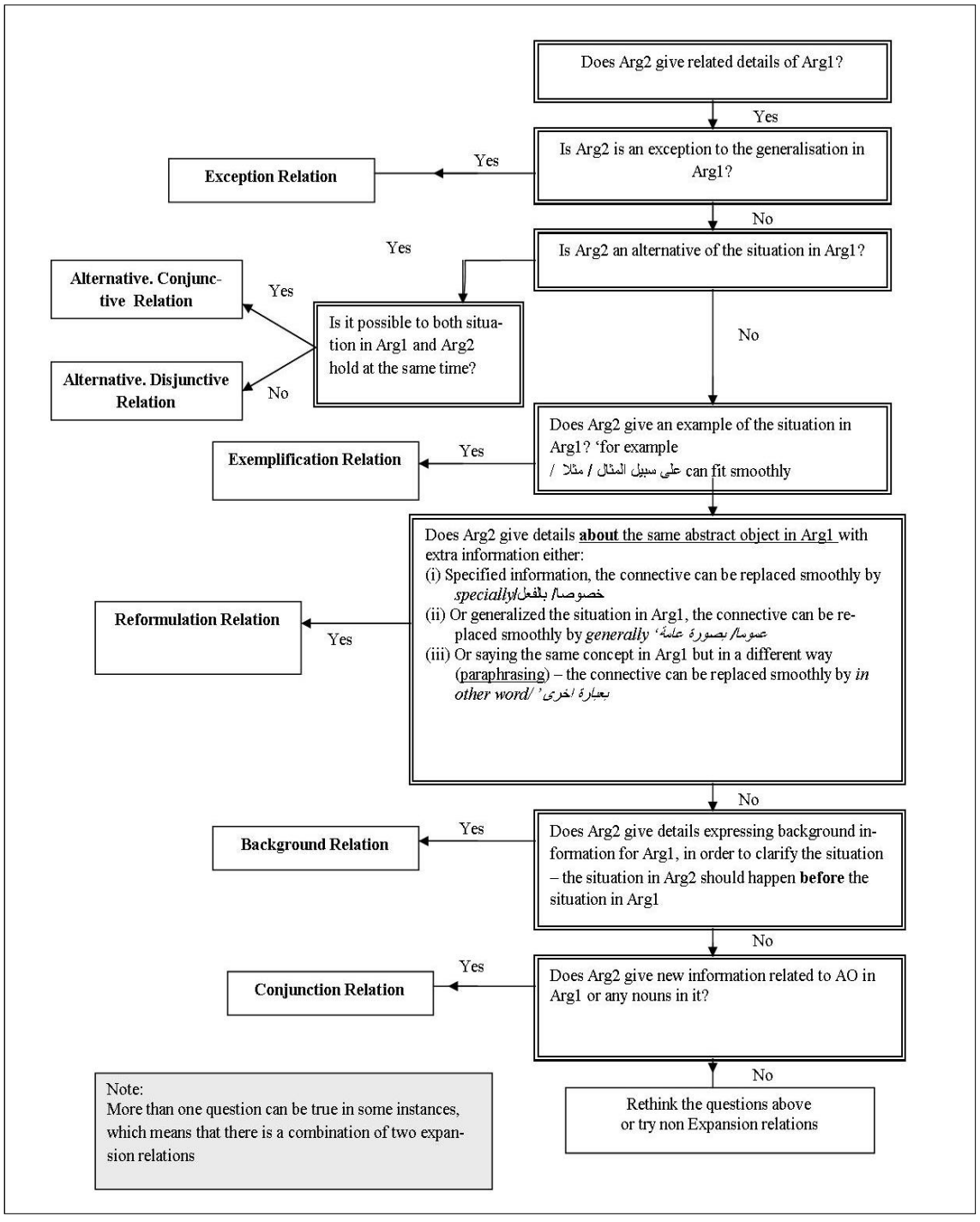



**Figure 5-4: A decision tree for disambiguating Expansion relations**

## 5.7.2 Decision Tree for Expansion Relations

The most ambiguous instances in the pilot annotation were those of Expansion relations. The annotators agreed on the class level Expansion, but were confused

when it came to distinguishing the lower level relations, especially Exemplification, Conjunction, Reformulation and Background relations. Thus, we proposed a decision flowchart just for Expansion relations; a sequence of questions to help clarify how the annotator should think before making a decision concerning this kind of relation. The flowchart in Figure 5-4 starts with the easily identifiable relation EXPANSION.Exception. Of course, the assignment of Conjunction should be the last alternative.

## 5.8 Summary

The discourse annotation manual for Arabic is based on similar annotation principles as the one for English in the PDTB. However, we have made the required adaptations regarding discourse connectives, relations and their arguments, to fit with the specific features for Arabic. The most important adaptations are that we consider prepositions and nouns as valid discourse connectives, and al-maSdar nouns as valid arguments, and that we intrduced novel relations for Arabic. In this first version of discourse annotation guidelines for Arabic, we limited the scope of the annotation to strongly agreed discourse relations in the pilot annotation. Thus we ended with an expandable taxonomy of 17 fine-grained discourse relations under 4 main classes similar to English sense classification in the PDTB.

Although a few long articles from the internet were annotated in the initial discourse analysis for Arabic, the scheme is developed and used to annotate mainly news text from the ATB. However, the scheme can be used to annotate longer texts from different genres with further improvements, if required.

Although the discourse annotation in the present study focused on the annotation of explicit connectives and their relations, we also came across other discourse devices during our analysis such as implicit connectives (inferred relations), entity relations, attribution and anaphora. But they are not reported in the scheme as they are beyond the study target. In fact, we annotated a special case of entity relations that are introduced by explicit connectives, which are manily assigned the Conjunction relation. In addition, one more restriction is implented for the Conjunction relation to avoid confusion; it is not allowed to combine Conjunction relation with other

relations. Future studies in discourse annotation in Arabic would be able to take this research further, by using this thesis as a base, and developing a complete scheme of discourse annotation for Arabic.

# Chapter 6
# READ: An Annotation Tool for Arabic and English Discourse Relations

## 6.1 Introduction

The discourse annotation tasks in our study should identify three components for each annotation: the explicit discourse connective, its arguments Arg1 and Arg2, and associated relations. Thus, we need a tool that can be easily used to annotate these components with basic functions such as pre-highlighting of potential Arabic discourse connectives (our collection in Chapter 4), and use our discourse relation hierarchy (see Section 5.5). The existing annotation tools, at the study time, did not fulfil the requirements of discourse annotation for Arabic such as marking clitics as connectives and the possibility of starting the argument from the middle of a word. Refer to Section 3.2.3 for more discussion.

We decided to conduct the annotation in a stand-off style (based on the raw texts only), similar to the PDTB annotation. This allows wider ability of using the tool to annotate text without syntactic annotation. Therefore, no syntactic annotation of the ATB is displayed to the annotator in the tool, or used for the highlighting of the potential connectives.

This chapter presents the user guidelines and features of our discourse annotation tool (READ: Relation annotation for English and Arabic Discourse). Section 6.2 illustrates the language setting of the interface and the annotation text. The tool provides useful features that are described in Section 6.3. The text preparation before the annotation phase is presented in Section 6.4, followed by the procedure of discourse connective annotation in Section 6.5. The output of the tool is a text file

following the format described in Section 6.6. The chapter ends with a summary of the main features of the READ tool.

## 6.2 Language Setting

The tool firstly offers a language option of either Arabic or English for the interface as well as the text to be annotated, which also affects the layout of the tool (see Figure 6-1). The text is in Unicode format, and the layout of the text is based on the selection of the 'Files Language' as either Arabic (عربي) or English. The setting of the files language is very important, as the tool will highlight the appropriate potential connectives of the selected language.


**Figure 6-1: Language setting of the READ's interface and the text display**

## 6.3 Features of the READ Tool

### Function menu

The tool has four drop-down functional menus, as shown in Figure 6-2:

- File: to open, save and close the annotation file
- Connectives: to modify the list of potential connectives supplied with the tool
- Align: to change the alignment of the text appearing in the text box
- Help: to show the annotation manual and information about this version of the READ tool.

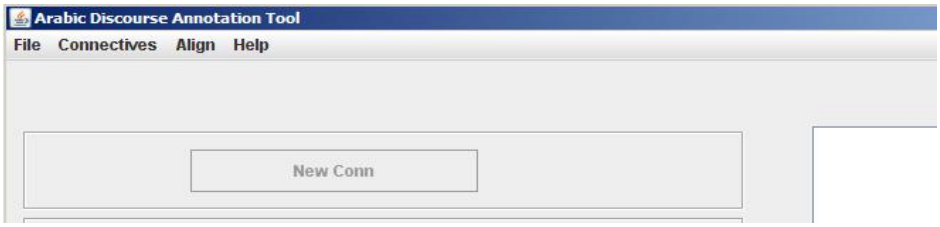



**Figure 6-2: The menu bar of the READ tool (File, Connectives, Align, and Help drop-down submenus).**

## Potential discourse connectives

The READ tool is supplied with two modifiable lists of potential discourse connectives, one for Arabic (our collection described in Section 4.7) in a file 'conn.txt', and one for English (PDTB2 collection) in a file 'Eng_conn.txt'. The user can simply add or remove potential connectives directly from the text files in the tool package. Alternatively, they can use the menu Connectives>Add/Remove to update the connective list, and then restart the tool, to configure the new list of potential connectives.

## Discourse relations

The relation hierarchy in the READ tool considers the discourse relations in the Arabic taxonomy, in this version of the tool. If a connective is deemed to express two relations at the same time, the annotator is enabled to pick up one or more relation from the drop-down list, by holding the CTRL key while selecting relations from the list. Figure 6-3 shows a screenshot of the hierarchy of Arabic discourse relations in the READ tool. Two relations are selected in this screenshot.

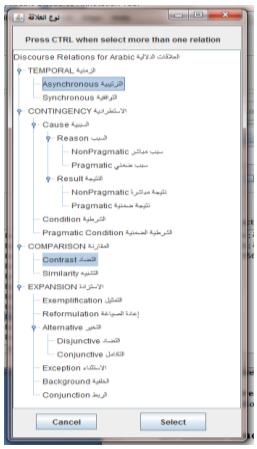



**Figure 6-3: The hierarchal structure of discourse relations in the READ tool**

### Comment Box

Annotators are allowed to make comments or suggestions in the comment box, such as the occurrence of new connectives which are not highlighted, or new relations which are not listed in the tool. These comments will be valuable for creating the next generation of discourse annotation guidelines for Arabic (see Figure 6-4).

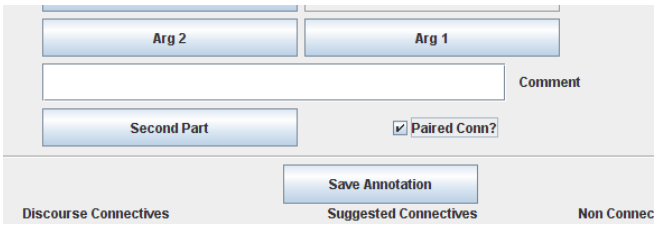



**Figure 6-4: The comment box and paired connective annotation options**

### Paired Connectives

Although the majority of discourse connectives are either simple (one token), or a phrase (more than one token), Arabic frequently uses connectives with two separated parts, where each one introduces an argument of the connective (a paired connective). Refer to Section 4.3 for a full description. Thus, the READ tool allows the user to mark a second part of the connective as well by ticking the checkbox 'Paired Conn?' and thus enabling 'Second Part'. Figure 6-4 shows a snapshot of the section of the tool that concerns paired connective annotation.

## 6.4 Pre-annotation Text Preparation

The text to be annotated is prepared by highlighting all potential discourse connectives from our discourse connective list for Arabic (Section ). As READ is a stand-off tool and not linked to any syntactic annotation or segmentation, potential clitic connectives will be highlighted when appearing at the beginning of words using string matching only.

To do that, the annotator simply selects the raw text file from the menu File>Open. The name of the file will appear at the top of the text box. The tool will automatically highlight all potential discourse connectives in pink, using our pre-defined

connectives list (see Section 4.5). A snapshot of the initial status of the tool after opening a text file is shown in Figure 6-5. The output of the annotated file will have the same name as the original file with a different extension (.ann), and will be stored at the same location.

The highlighted potential connectives are also presented in an ordered list of suggested discourse connectives (the list in the middle in Figure 6-5), with *starting and ending* indices of the connective. In this phase all functional buttons are disabled, and the two lists 'Discourse Connectives' and 'Non-connectives' are empty. The highlighted colour of a potential connective will be switched to blue once it is selected by the user from the Suggested Connectives list.

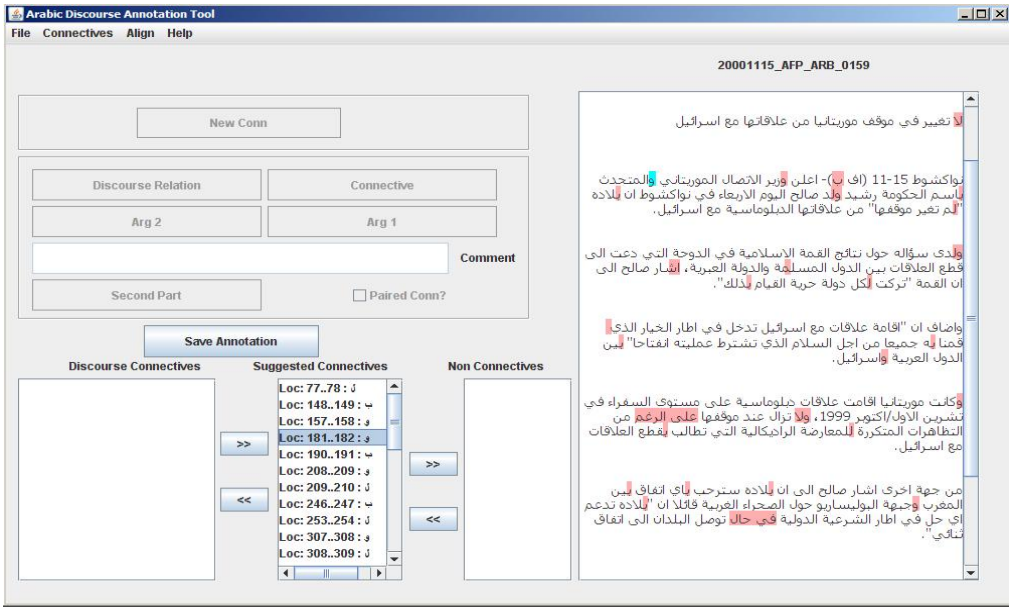



**Figure 6-5: Initial status of the READ tool after opening a desired text for annotation**

## 6.5 Connective-based Annotation

First of all, the annotator should read the entire text to achieve an overall understanding of the discourse and whatever knowledge or information is conveyed by the text. Then, s/he should make a series of context-based decisions for each potential connective in the Suggested Connectives list, using the following procedure for each raw file:

1) Using the mouse, point to a desired highlighted potential connective in the Suggested Connectives list, and decide whether it is a discourse connective or not in this context by using the arrows. Figure 6-7 shows a description of the arrows that are used to annotate the potential connectives in the Suggested Connectives list. The decision is made by answering the question 'Does this potential connective have a discourse function in context', according to our annotation guidelines in Appendix B:

   - If yes, use the arrows to move the highlighted connective into the Discourse Connectives list on the left. The text is then free from any highlighting except the selected connective. Then, go to Step 2.
   - If no, use the arrows to move the highlighted connective into the Non Connectives list on the right. Then, Jump to Step 1 for the next highlighted connective.

2) Mark the first argument (Arg1) and press the Arg1 button.

3) Mark the second Argument (Arg2) and press the Arg2 button

4) Select one or more suitable discourse relation(s) from a drop-down hierarchy of Arabic discourse relations that appears when the Discourse Relations button is clicked. The user can select more than one relation by holding the Ctrl key on the keyboard.

5) If the connective is paired, the user should tick the checkbox and mark the second part, then click on the Second Part button.

6) The user can record any comment or suggestion about this annotation in the comment box, if necessary.

7) Save the annotation, and go to Step 1 for the next highlighted connective.

At the end, there should be no potential connectives in the Suggested Connectives list, as in Figure 6-6. Save the annotation and open another raw file for the next annotation, if any.

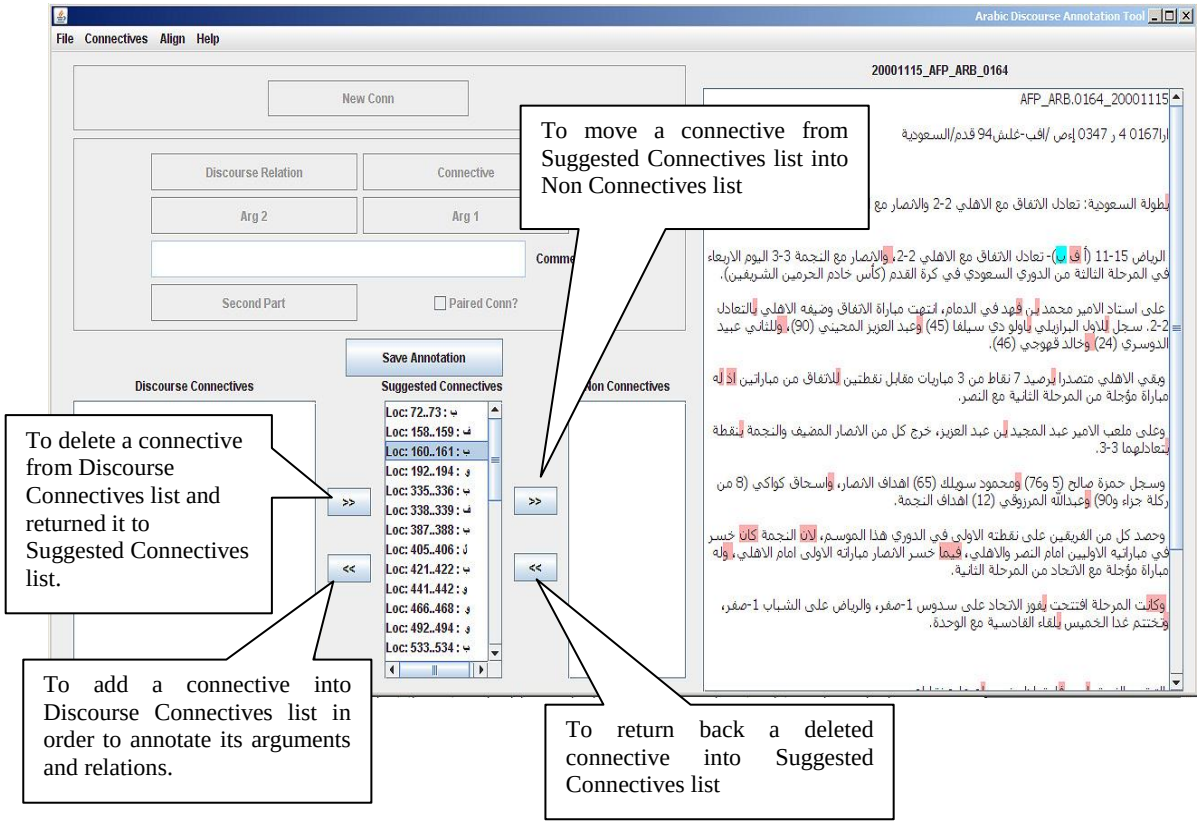



**Figure 6-7: A description of the arrows on the annotation tool READ**

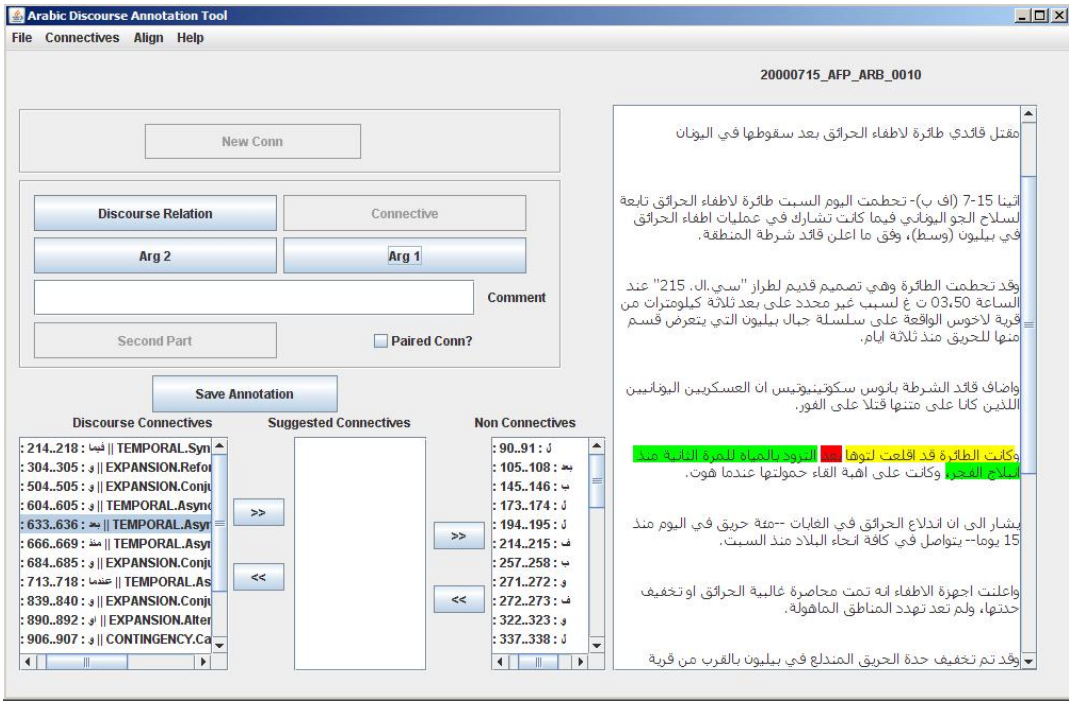



**Figure 6-6: The final status of the tool after annotating all potential discourse connectives.**

## 6.6 Output Format

The READ tool saves the annotation in a text file using the indices of: (i) start and end of the connectives/non-connectives, and (ii) start and end of the text spans representing Arg1 and Arg2. Also it saves the annotation of discourse relations, paired connectives and comments the annotator has entered. Each connective's annotation is saved on a sigle line, and the annotation parts for one connective are seperated by vertical bars. The annotations are separated by bars. For example, the connective *و/w/and* at BOP at the index 220, Arg1 (100..220) and Arg2(223..400) would appear in the output file as follows:

EXPLICIT|و|220..221|||100..220||223..400||||EXPANSION.CONJUNCTION|C|BOP|P||

Or, NONCONN|و|220..221| ,if it was annotated as non-discourse connective.

We use |C| to introduce a comment and |P| to introduce the second part of a paired connective, if any. The comment in the above example is 'BOP' and the connective و/w/and is not paired connective. A snapshot of the output file is shown in Figure 6-8.

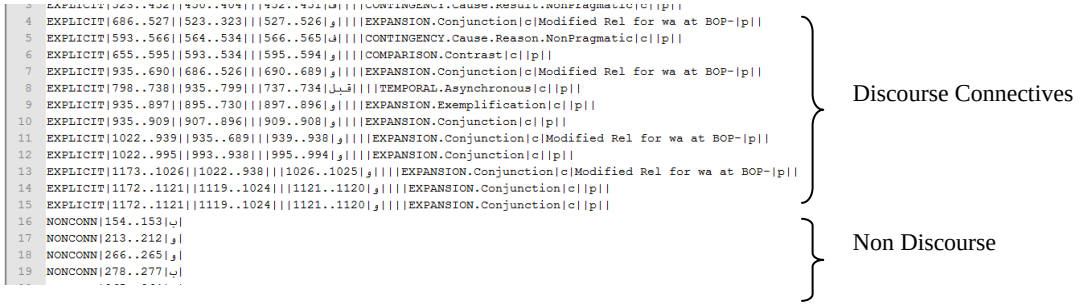



**Figure 6-8: A snapshot of the output of an annotated file showing the text format.**

There might be a need for a post processing step to exclude final punctuation or mistakenly included function words that from any argument. In Section 7.3 we discuss more details about our post-processing in the current study.

## 6.7 Summary

The READ tool is a discourse annotation tool for manual disambiguation of the potential discourse connectives for Arabic and English. It can, however, be used for annotating discourse connectives in any language that uses Unicode format. As long as the discourse connective list in the file 'conn.txt' is updated with a new list for the language.

The READ tool is a very useful annotation tool for annotating discourse connectives for Arabic. It solved problems that arose when using tools that were not compatible with Arabic, such as annotating newly introduced discourse relations and clitic connectives. It was developed and tested to enhance annotation reliability, and have an enjoyable annotation process compared with purely manual annotation.

The tool was then used to annotate raw texts from the Penn ATB Part1, to produce the first discourse annotated Treebank for Arabic, the LADTB. The tool is distributed free of charge for non-commercial purposes. It can be downloaded from the Arabic Discourse Treebank website[18], or can be ordered personally by emailing the authors. All copyrights are reserved by the University of Leeds, the British Academy and the Imam University[19].

[18] The LADTB website is www.arabicdiscourse.net
[19] The licence of the READ tool is shown in Appendix E.

# Chapter 7

# Creating the Leeds Arabic Discourse Treebank

## 7.1 Introduction

Discourse corpora are elementary but essential components for discourse processing studies. Such corpora are annotated for cohesive devices, for example, anaphora and discourse relations. In this chapter, we show that Arabic can be reliably annotated for explicit discourse relations following our adaptation of the PDTB guidelines (Chapter 5). The READ tool (Chapter 6) was used to annotate discourse connectives, their relations and arguments in the Penn Arabic Treebank Part1 v.2 (Maamouri and Bies 2004). As stated in Section 3.3, the target is to expand the level of annotation in the treebank to include a discourse layer. This extension annotation is the first discourse corpus for Arabic – the Leeds Arabic Discourse Treebank (LADTB v.1).

The discourse annotation covers three main tasks:

**Task 1:** identification of explicit Arabic discourse connectives.

**Task 2:** disambiguating discourse connectives by annotating discourse relations they convey.

**Task 3:** Annotating the two arguments, the abstract objects linked by a particular connective.

In this first discourse annotation effort for Arabic, we concentrate on explicit discourse relations that are signalled by one of the discourse connectives in our inventory for Arabic. We do not annotate implicit relations, attribution, entity relations and anaphora; they are out of scope of this study.

The human annotation was conducted by two well-trained Arabic native speakers, who have a good linguistic background, on 537 news files from the Penn Arabic Treebank Part1 including 126,394 tokens after the treebank clitic segmentation. The

gold-standard of the LADTB includes 6,328 annotations of 80 explicit connective types, and 55 distinct discourse relations (17 single relations).

The LADTB is one of the main contributions of my study which is promising to be a rich resource for corpus-based discourse studies. The corpus will be distributed to the public via the LDC - in 2012.

The corpus creation steps of the first Arabic discourse corpus starting from raw text untill the gold-standard LADTB are discussed in the following sections. Section 7.2 presents the human annotation process and inter-annotator agreement studies for the three annotation tasks. The annotation was then filtered by semi-automatic post-processing to drive towards a gold standard for easily-solved disagreements (Section 7.3). After the post-processing, the inter-annotator agreement studies were repeated to examine the effects of post-processing (Section 7.4). In addition, the common disagreement cases of all annotation tasks are reported in Section 7.5 for future development. The first gold standard was derived by manual adjudication of remaining disagreement cases. The statistics of discourse connectives and relations in the LADTB and their frequency are presented in Section 7.6. Complete distributions of discourse connectives and relations in the LADTB gold standard are shown in Appendix C and D, respectively. When producing the first discourse corpus for Arabic (LADTB), it is very useful to explore the similarities and differences of discourse properties of Arabic (LADTB) and English (PDTB2) corpra that are using similar annotation principles; a statistical comparison study is described in Section 7.7. At the end of the chapter, A summary of the creation of the LADTB and how reliable our annotation of explicit discourse relations was, is presented.

## 7.2 Human Annotation

Two independent native speakers of Arabic, who were not involved in the tool or scheme development or pilot annotation, were trained on the first 150 texts in the ATB. Agreement studies were conducted on a regular basis for the discourse annotation tasks on the next 387 texts. Once the annotation reached a stable agreement, the training texts (150) were re-annotated and then included in the overall agreement studies. We measure in the first task whether annotators agree on the

binary decision whether an item constitutes a discourse connective in context. For the second task, we measure whether annotators agree which discourse relation an identified connective expresses. In addition, we measure whether annotators agree on the text spans that constitute arguments, the third task.

We have used percentage agreement and kappa/alpha for measuring the agreement on discourse connectives and relations. Alpha is used to measure a partial agreement of multiple relations such as TEMPORAL.Asynchronous/COMPARISON.Contrast and TEMPORAL.Asynchronous/EXPANSION.Reformulation. In contrast, the agreement on argument boundaries is measured by two different metrics (i) exact match and (ii) word overlap (see Section 2.6.5 for more details about agreement measurement).

### 7.2.1 Agreement Studies for Annotating DCs and Relations

The inter-annotator agreement studies of Task1 (discourse connective identification) and Task2 (discourse relation identification) were conducted approximately on a weekly basis for in average 22 texts over six months, on two different datasets: (i) Set 1 of all instances of potential connectives in the files and (ii) Set 2 of instances of potential connectives excluding *و/w/and* at beginning of paragraph (BOP). As we noticed during the pilot annotation (see Section 5.4.45.3) the connective *و/w/and* introduces almost each paragraph without a specific discourse relation conveyed. Thus, the second study on Set 2 is conducted to observe the behavior of inter-annotator agreement when excluding the most ambiguous connective *و/w/and* at beginning of a paragraph.

Disagreement cases in discourse connective and relation identification were discussed at each turn of independent annotation, to learn from the mistakes, for the next annotation phase. However, no major adaptations were made to the annotation scheme at this stage. The inter-agreement studies are always conducted on the data before the discussion.

**Table 7-1: The inter-annotator agreement for two annotation tasks: discourse connective recognition and identification of fined-grained and class level relations. PA = percentage agreement.**

| **Human Annotation** | **Set 1 – all conn** | **Set 2 – excluding *و/w/and* at BOP** |
|---|---|---|
| **Number of files** | | 537 |
| **Number of potential connectives** | 23331 | 21200 |
| **Agreement on discourse connective recognition** | | |
| **Agreed discourse connectives** | 5586 | 3500 |
| PA | 95% | 95% |
| Kappa | 0.88 | 0.83 |
| **Agreement on discourse relation disambiguation on agreed connectives – fine-grained relations** | | |
| PA | 66% | 74% |
| Kappa | 0.57 | 0.69 |
| Alpha | 0.58 | 0.71 |
| **Agreement on discourse relation disambiguation on agreed connectives – class level relations** | | |
| PA | 80% | 86% |
| Kappa | 0.67 | 0.75 |
| Alpha | 0.69 | 0.77 |

The statistics of overall inter-annotator agreement, merging the data from 6 months, are presented in Table 7-1. The annotation of discourse connectives is highly reliable, with a percentage agreement of 95%/95% and kappa of 0.88/0.83 on Set 1 and Set 2 respectively. These significant results on both datasets show that our annotation guidelines are clear on identifying discourse connectives.

On the other hand, the agreement on annotation of fone-agreed discourse relation recognition does not exceed 67% percentage agreement, 0.57 kappa and 0.58 alpha on Set 1. This result highlights the difficulty of achieving good agreement for a language with highly ambiguous connectives in terms of the discourse relations they signal. However, the agreement rises to 74%, kappa 0.69 and alpha 0.71 to be at an acceptable level on Set 2 when tokens of *و/w/and* at BOP were excluded. These differences highlight the expectation of the behavior of the connective *و/w/and* at BOP, the most ambiguous connective. We can consider the instances of the connective *و/w/and* at BOP to have a similar discourse function as implicit connectives in English. Therefore it is essential to arrange a special manipulation in

the current corpus for the connective *و/w/and* at BOP and also do comprehensive studies on this particular connective in future work.

The agreement for discourse relation recognition is measured also for relations at class level in order to examine how often the annotators disagree on the upper level relations. Relevant results in Table 7-1 show that annotators have agreed on 13% more relations when using the four main classes only. They agree on 80%/86% with a kappa of 0.67/0.75 on Set 1 and Set 2 respectively instead of 66%/74% and kappa 0.57/0.69 of the tokens for fine-grained relations.

### 7.2.2 Agreement Studies for Argument Identification

Unlike the limited binary judgments in discourse connectives recognition or discourse relation identification among a relatively small number of categories, measuring the agreement of two unrestricted judgments such as text spans is a difficult task. Generally speaking, the annotator can mark any text prior to the connective as a first argument, and any text after the connective as a second argument as long as it starts in the same sentence that is introduced by the connective. Both arguments can span more than one single sentence. In addition, the annotation is conducted on raw text so the sentence and clause boundaries are not defined.

For these reasons, ordinary evaluation metrics such as accuracy and kappa are not suitable. Therefore, we measure the agreement of argument text spans Arg1 and Arg2 separately, using two special measurement metrics. The first is the *exact match* of white-space-tokenized words of argument spans, as used for the English PDTB study as well (Miltsakaki *et al.* 2004). The second metric is *agr* which takes into account the word overlap in the two judgments rather than the exact boundaries only. The *agr* metric is a directional measure of agreement between two judges (ann1 and ann2) (see Section 2.6.5 for a full explanation). We will compute both directions of *agr* and consider the average of the two *agr*.

Argument agreement on the 5586 agreed connective tokens is shown in Table 7-2. Overall, the agreement for Arg2 is more reliable than for Arg1. 13% of the tokens are without any overlap at all on Arg1 and only 0.3% on Arg2. This difference is

influenced by the annotation principles that restrict Arg2 to the sentence/clause introduced by the connective; while Arg1 might be any discourse unit prior to the connective in the usual order Arg1_DC_Arg2 or after the connective in the order DC_Arg2_Arg1. However, for 32% of the connectives Arg2 does not produce an exact match. That is due to, on the one hand, differences in inclusion of punctuations, attributions or function words and, on the other hand, the exclusion of some necessary complements in verb sentences by one of the annotators. More details will be discussed in Section 7.5.1. The majority of cases without overlap for Arg1 are for the connective *ﻭ/w/and* at BOP.

**Table 7-2: Inter-annotator reliability for arguments Arg1 and Arg2 using two different measurements (a) exact match and (b) agr.**

| **Total agreed connectives** | 5586 | |
|---|---|---|
| **a) Exact match metric** | **Arg1** | **Arg2** |
| exact match =1 | 2361 (42%) | 3803 (68%) |
| exact match =0 | 699 (13%) | 18 (0.3%) |
| 0 < exact match < 1 | 2526 (45%) | 1765 (32%) |
| **b) Agr metric** | **Arg1** | **Arg2** |
| agr(ann1//ann2) | 78% | 93% |
| agr(ann2//ann1) | 74% | 93% |
| Average agr | 76% | 93% |

The second metric *agr* measures word overlap on arguments Arg1 and Arg2 individually. We report high word overlap (93%) for Arg2 and lesser, but still a substantial agreement for Arg1 (76%). Disagreement of arguments will be discussed with examples in Section 7.5.2.

## 7.3 Automatic Post-processing

We automatically corrected easily made annotator mistakes with regard to annotating connectives, arguments and relations, and made any defensible automatic modifications which might reduce the amount of manual work needed in the gold standard production. While the annotators annotated the raw text, post-processing and regularization made use the syntactic analyses provided in the ATB. They involved:

**Removal of easily identifiable mistakes**

- We deleted all annotation of connectives that do not have syntactic annotation in the Arabic Treebank such as those in titles or footers. (This action will affect the number of potential connectives and agreed connectives).

- We excluded punctuation, the function word *ان/that* and connectives (outside of the scope of the annotations and the sentences) from argument boundaries.

- We converted some modified connectives into only the original connectives. For example, the modified connective *وقد/wqd/and it had* was converted into the single connective *و/w/and* alone, and *قد/qd/was* was included in Arg2. Similarly the modified connective *وكان/wkan/and (it/he/she) was* is converted into *و/w/and* alone, and *كان/kan/(it/he/she) was* was included in Arg2. The same conversion took place for modified connectives with similar properties such as the inclusion of the function word *ان/An/that*.The reason behind that is to match the ATB syntactic annotation of the sentence. In fact, it was a mistake to include these function words in the connectives as modified forms in our initial collection of the discourse connectives, as these function words are syntactically parts of the argument. These modifications do not affect the inter-annotator agreement, as they have been done for both annotations.

- We converted some multiple connectives, that include *و/w/and*, into different annotations for each connective. They do not share the same parent in the syntactic annotation of ATB. Thus, it is hard technically to combine them as one set when they have different syntactic features. For example, the connective *ولكن/wa lkn/and but* is converted into two connectives *و/w/and* and *لكن/lkn/but* independently. Both annotations have almost the same arguments, apart from including *لكن/lkn/but* in the second argument Arg2 of the connective *و/w/and*. We assign EXPANSION.Conjunction relation if the first connective is *و/w/and* and keep the agreed relation for the second connective. (This action will affect the number of agreed discourse connectives and relations in the study).

- We included all obligatory complements in VP and NP arguments by expanding the boundary of the argument to cover tokens in their trees. An exception is the expansion of Arg1 when the order of the arguments is Arg1-Conn-Arg2-Arg1, because the syntactic annotations of connective and Arg2 are included in the annotation of Arg1 (in one parse tree). Ex. 7-1 presents the ATB annotation of Arg1 showing that the connective بعد*/bEd/after* and Arg2 انقطاع دام يومين*/cutting of two days* are both within the Arg1 tree.

**Ex. 7-1 (file: 20000915_AFP_ARB.0023)**

استأنف الرئيس القبرصي غلافكوس كليريديس اليوم الجمعة في الامم المتحدة **بعد** **انقطاع دام يومين**، محادثاته غير المباشرة حول مستقبل جزيرة قبرص

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ast>nf | Alr}ys | AlqbrSy | glAfkws | klyrydys | Alywm | fy | AlAmm | AlmtHdp | bEd |
| resume | president | Cypriot | Glafcos | Clerides | today | in | nations | united | after |
| AnqTAE | dAm | Ywmyn, | mHAdvAth | gyr | AlmbA$rp | Hwl | mstqbl | jzyrp | qbrS |
| cut | last | two-days | negaiation | not | direct | about | future | island | Cyprus |

*The Cypriot President Glafcos Clerides resumed today at the United Nations,* **after** **a lapse of two days**, *the indirect talks on the future of the island of Cyprus*

The ATB: (S (VP (VERB_PERFECT Ast>nf_استأنف) (NP-SBJ (NP (DET+NOUN Alr}ys_الرئيس) (DET+ADJ AlqbrSy_القبرصي)) (NP (NOUN_PROP glAfkws_غلافكوس) (NOUN_PROP klyrydys_كليريديس))) (NP-TMP (NP (NOUN Alywm_اليوم)) (NP (DET+NOUN_PROP+NSUFF_FEM_SG AljmEp_الجمعة))) (PP-LOC (PREP fy_في) (NP (DET+NOUN AlAmm_الامم) (DET+ADJ+NSUFF_FEM_SG AlmtHdp_المتحدة))) (PP-TMP (PREP bEd_بعد) (NP (NP (NOUN AnqTAE_انقطاع)) (SBAR (WHNP-1 (-NONE- *0*)) (S (VP (VERB_PERFECT dAm_دام) (NP-SBJ-1 (-NONE- *T*)) (NP-TMP (NOUN+NSUFF_MASC_DU_ACCGEN ywmyn_يومين)))))) (PUNC ,_،) (NP-OBJ (NP (NOUN+NSUFF_FEM_PL mHAdvAt_محادثات) (POSS_PRON_3MS h_ه)) (ADJP (NEG_PART gyr_غير) (DET+ADJ+NSUFF_FEM_SG AlmbA$rp_المباشرة)) (PP (PREP Hwl_حول) (NP (NOUN mstqbl_مستقبل) (NP (NP (NOUN+NSUFF_FEM_SG jzyrp_جزيرة)) (NP (NOUN_PROP qbrS_قبرص)) (ADJP (NO_FUNC Almqsmp_المقسمة) (PP-TMP (PREP mn*_منذ) (NP (NUM 26_26) (NOUN+NSUFF_MASC_SG_ACC_INDEF EAmA_عاما)))))))) (PUNC ._.))

## Provisional decisions in the first discourse corpus for Arabic

With regard to discourse relation assignment, a relation EXPANSION.Conjunction is assigned automatically to all disagreed instances of *و/w/and* at BOP[20]. As mentioned previously this type of *و/w/and* functions generally as a junction tool between newswire paragraphs without other clear discourse usages. This action of assigning EXPANSION.Conjunction automatically for such disagreements is clearly reported in our publications and any documentation of the LADTB. We encourage establishing intensive linguistic studies of discourse connectives such as *و/w/and* at BOP. (As we have many disagreements on instances of *و/w/and* at BOP, this action will clearly affect the agreement figures on discourse relations).

**Table 7-3: The inter-annotator agreement after the automatic post-processing for two annotation tasks: discourse function of the potential connectives and discourse relations at fined-grained and class levels.**

| **Human Annotation** | **Set 1 – all conn** | **Set 2 – excluding *و/w/and* at BOP** |
|---|---|---|
| **Number of potential connectives** | 20312 | 18080 |
| **Agreement on discourse connective recognition** | | |
| **Agreed connectives** | 5541 | 3170 |
| **PA** | 94% | 93% |
| **Kappa** | 0.88 | 0.83 |
| **Agreement on discourse relation disambiguation only on agreed connectives – fine-grained relations** | | |
| **PA** | 86% | 76% |
| **Kappa** | 0.8 | 0.71 |
| **Alpha** | 0.81 | 0.73 |
| **Agreement on discourse relation disambiguation only on agreed connectives – class level relations** | | |
| **PA** | 90% | 83% |
| **Kappa** | 0.81 | 0.76 |
| **Alpha** | 0.83 | 0.78 |

[20] No change was made for agreed relations for *و/w/and* at BOP.

## 7.4 Agreement after the Automatic Post-processing

We measure the agreement again after automatic correction (Table 7-3). The number of agreed discourse connectives is changed slightly after automatic correction; 5541 instead of 5586. The overall agreement of discourse connective identification remains high at 94% percentage agreement and 0.88 kappa for all connectives in Set 1 but it dropped slightly to 93% percentage agreement and 0.83 kappa when tokens of *ﻭ/w/and* at BOP were excluded in Set 2. However, on both sets connective recognition is still highly reliable.

As expected, on the other hand, the agreement of discourse relation recognition increased on Set 1 to 86% and kappa 0.8 due mainly to the automatic assignment of EXPANSION.Conjunction to the disagreed instances of *ﻭ/w/and* at BOP in the automatic post-processing. At the same time, a slightly higher agreement is recorded for fine-grained discourse relation assignment on Set 2 after the automatic post-processing with a percentage agreement 76% and kappa 0.71. This result is due to converting some multiple-connectives in the automatic post-processing into two connectives and assigning EXPANSION.Conjunction to the first connective.

Similarly, the percentage agreement at class level relations rises to 90% on Set 1 instead of only 80% without automatic correction, while it is lower but still substantial at 83% on Set 2 with a higher kappa of 0.76.

**Table 7-4: Inter-annotator reliability for arguments Arg1 and Arg2 after applying the automatic post-processing using two different measurements (a) exact match and (b) agr.**

| **Total agreed tokens** | | 5541 |
|---|---|---|
| **b) Exact match metric** | **Arg1** | **Arg2** |
| exact match =1 | 2478 (45%) | 4186 (76%) |
| exact match =0 | 677 (12%) | 4 (0.1%) |
| 0 < exact match < 1 | 2386 (43%) | 1351 (24%) |
| **b) Agr metric** | **Arg1** | **Arg2** |
| *agr*(ann1//ann2) | 80% | 94% |
| *agr*(ann2//ann1) | 75% | 96% |
| Average *agr* | 78% | 95% |

**Argument agreement:** the automatic inclusion of complements in arguments helped increase the exact match annotations, and at the same time reduce the non-overlap annotations for Arg1 and Arg2, as shown in Table 7-4. These higher agreement figures will definitely reduce the manual effort in producing the gold-standard annotation.

The next section will describe the common disagreement cases on discourse connective recognition, relation assignment and argument boundaries identification.

## 7.5 Disagreement Cases

We present the common disagreement cases during our discourse annotation experiment, which is the first effort for Arabic. Hopefully, our observations provide a good basis for improving future discourse annotation studies. Ideally, we would like to give an estimate of the frequency of each disagreement or error type. However, as the annotation was conducted in stages with discussions in-between, a frequent error in an early annotation stage might become less frequent after discussion so that any accumulated frequencies can be misleading.

### 7.5.1 Ambiguity in Identification of DCs and Arguments

Identifying discourse connectives and their arguments is closely related; if there are no valid arguments that a potential connective relates then most likely this potential connective has no discourse function. Therefore, the obvious approach is to deal with their disagreement cases in one go.

#### Semantic vs. discourse function

Annotators were sometimes confused whether the connective has a semantic or a discourse function in the sentence. For example, the potential discourse connective *ب/b/by* expresses 14 meanings according to the literature (Alfarabi 1990) (see Section 4.2.1). Some of which have a discourse function such as Causal usage (i.e. حصل على المركز الأول **ب**حصوله على الدرجة الكاملة في الإختبار/he got the first position **by** gaining a full mark in exam). However, the majority of its meanings have non-discourse usage

such as a *ظرف مكان/preposition* (for example, *الكتاب بالبيت/the book is in the home*) or a meaning of *المعية/المصاحبة/with* (for example, *نم بسلام/sleep in peace*).

Annother example, the potential connective *اذا/A*A/if* is almost always a discourse connective with a conditional function. However, there are exceptions such as in Ex. 7-2; the potential connective *اذا/A*A/if* here is a relative pronoun *whether* with only one argument, and so it is not a discourse connective.

**Ex. 7-2** (incorrect: discourse connective (Rel: Condition), correct: not a discourse connective)

ليس واضحا **اذا** كانت معدات الإنقاذ الأمريكية يمكن استخدامها على الفور.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| lys | wADHA | A*A | kAnt | mEdAt | Al<nqA* | Al>mrykyp | ymkn | AstxdAmA | ElY | Alfwr |
| not | clear | if | was | equipments | rescue | US | can-be | used | on | now |

It is not clear **whether** the U.S. rescue equipments can be used immediately.

## Missing discourse relations

In some cases, a connective might have a discourse function but signal a discourse relation that is not in our taxonomy. Annotators disagreed on whether to not annotate this connective at all or whether to assign a relation that does not fully fit. In Ex. 7-3, the connective *ب/b/by* has a discourse function expressing a Mean or Method relation (a meaning of *بواسطة/via/by)*; which is not in the current relation taxonomy. This is leading to annotator disagreement. For example, *including extra countries* is not a reason of *seeking to expand the OPEC cartel*, as it was annotated by one of the annotators. This new relation can be considered in the advanced annotation.

**Ex. 7-3** (incorrect: discourse connective (Rel: Reason), correct: not a discourse connective)

يسعى لتوسيع اوبك بانضمام دول أخرى اليها

| | | | | | | |
|---|---|---|---|---|---|---|
| ysEY | ltwsyE | Awbk | bAnDmAm | dwl | >xrY | AlyhA |
| seek | for-expanding | OPEC | by- including | countries | other | to-it |

It is seeking to expand the OPEC cartel **by** including extra countries

## Syntactic ambiguity

The connectives might signal a syntactic and discourse link at the same time. The discourse annotation of those connectives is strongly affected by the syntactic analysis. For example, the preposition connective *لل/for* in Ex. 7-4 is followed by an al-maSdar noun which is a valid argument. However, the confusion arose from the first argument; two legitimate syntactic attachments are possible for the preposition connective *لل/for*. First, it could be attached to the concrete object قدرات نووية / *nuclear capability*, then the connective does not have a discourse function. Second, it could be attached to the al-maSdar noun *حصول/acquiring*, where the connective *لل/for* is a discourse connective indicating a causal relation.

However, in our post-processing we considered such cases of syntactic ambiguity as non-discourse connectives as the ATB syntactic annotation always uses the first analysis.

**Ex. 7-4** (incorrect: discourse connective (Rel: Reason), correct: not a discourse connective)

| ان اسرائيل تعتبر ان ايران ستبدأ بالحصول على قدرات نووية **لـ**اغراض عسكرية | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| An | AsrA}yl | tEtbr | An | AyrAn | stbd> | bAlHSwl | ElY | qdrAt | nwwyp | lAgrAD | Eskryp |
| that | Israel | consider | that | Iran | Will-start | gaining | on | capability | nuclear | for-purposes | military |
| Israel believes Iran begins to acquire a nuclear capability **<u>for</u>** military purposes | | | | | | | | | | | |

## Verb Ellipsis

Recognising verb phrase ellipsis is not clear for the annotators when the phrase that is introduced by a potential connective is a prepositional phrase. In Ex. 7-5, the prepositional phrase *في احدى الحالات الثلاث/in one of the three cases* is part of the main argument and not verb ellipsis. In contrast, the prepositional phrases من الغرق/*from drowning* and *من الجفاف/from dehydration* in Ex. 7-6 are subject to be valid arguments in our discourse annotation due to the verb phase ellipsis توفوا من الجفاف/*they have died by dehydration*. Thus, the connective *او/Aw/or* is a discourse connective indicating the alternative relation.

**Ex. 7-5 (incorrect: discourse connective (Rel: Exception), correct: non-discourse connective)**

لن تستطيع المشاركة في دورة الالعاب الاولمبية **الا** في احدى الحالات الثلاث

| ln | tstTyE | Alm$Arkp | fy | dwrp | AlAlEAb | AlAwlmbyp | AlA | fy | AHdY | AlHAlAt | AlvlAv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| not | able | participate | in | circle | games | olympic | except | in | one | cases | three |

You will not be able to participate in the Olympic Games **except** in one of the three cases.

**Ex. 7-6 (incorrect: non-discourse connective, correct: discourse connective (Rel: Alternative))**

ان نحو ٤٠٠ مكسيكي توفوا غرقا **او** من الجفاف

| An | nHw | 400 | mksyky | twfwA | grqA | Aw | mn | AljfAf |
|---|---|---|---|---|---|---|---|---|
| that | around | 400 | Mexicans | died | drowning | or | from | dehydration |

*About 400 Mexicans have died by drowning* **or by dehydration**

## Al-maSdar Recognition

Although al-maSdar is a well-defined morphological category in the Arabic literature with more than 60 morphological patterns, annotators do not always recognise the al-maSdar nouns after a potential connective. That is a frequent case with al-maSdar patterns that have only three letters, and are therefore exactly similar to the root of three letters (فَعَلَ) but with different sounds/diacritics (فَعَلْ، فُعْلْ، فَعْلَ). For instance, the noun *طَلَبْ/request* after a potential connective *ب/b/by* in Ex. 7-7 is an al-maSdar noun derived from the verb *طَلَبَ/to order* using the form فَعَلْ.

**Ex. 7-7 (incorrect: non-discourse connective, correct: discourse connective (Rel: Reason))**

وصلت قوات بريطانية الى الأردن **بطلب من الملك حسين**

| wSlt | qwAt | bryTAnyp | AlY | Al>rdn | bTlb | mn | Almlk | Hsyn |
|---|---|---|---|---|---|---|---|---|
| arrive | forces | British | to | Jordan | By-request | from | king | Hussein |

*British forces arrived in Jordan* **due to a request by King Hussein**

The annotators were sometimes confused between a conjunction of al-maSdar nouns and a conjunction of non al-maSdar nouns. This might again be the result of not recognizing al-maSdar nouns. For example, the connective *و/w/and* indicates a conjunction of the non-al-maSdar nouns (*وجهة/perspective* and *مقاربة/approach*) in Ex. 7-8 (a) and (*اللاعنف/non-violence* and *العصيان/disobedience*) in Ex. 7-8 (b).

**Ex. 7-8 (incorrect: discourse connective (Rel: Conjunction), correct: not discourse connective)**

(a)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| انطلاقا من وجهة نظر مختلفة و مقاربة جديدة تتسم بالليونة | | | | | | | | | |
| AnTlAqA | mn | wjhp | nZr | mxtlfp | w | mqArbp | jdydp | ttsm | bAllywnp |
| going | from | view | point | different | and | comparision | new | looks | In-flexibility |
| Starting from a new perspective and a new approach based on flexibility | | | | | | | | | |

(b)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| المظاهرة تعتنق عقيدة غاندي في اللاعنف و العصيان المدني | | | | | | | | |
| AlmZAhrp | tEtnq | Eqydp | gAndy | fy | AllAEnf | w | AlESyAn | Almdny |
| demonstration | take | belief | Gandhi | in | nonviolence | and | disobedience | civil |
| The demonstration embraces the doctrine of Gandhi on nonviolence and civil disobedience. | | | | | | | | |

## 7.5.2 Disagreements in Argument Boundaries

### Both arguments are in a relative clause

The main clause of the sentence might be erroneously included in Arg1 when both arguments are within a relative clause; this mistake and the correct annotation are exemplified in Ex. 7-9.

**Ex. 7-9**

**(incorrect)**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | ***وضمت جزء كبير منها عام ١٩٦٧ مستقبل الفلسطينين*** | | | | تمثل قضية مستقبل القدس الشرقية التي احتلتها اسرائيل |
| tmvl | qDyp | mstqbl | Alqds | Al$rqyp | Alty | AHtlthA | AsrA}yl | wDmt |
| represent | issue | future | *Jerusalem* | *East* | which | *Occupied-her* | *Israel* | and-annexed |
| jz' | kbyr | mnhA | EAm | 1967 | mstqbl | AlflsTynyn | | |
| part | large | From- | year | 1967 | future | Palestinians | | |
| *The issue of the future of East Jerusalem, which Israel occupied* **and annexed a large part of in 1967, is the future of the Palestinians** | | | | | | | | |

**(correct)**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| تمثل قضية مستقبل القدس الشرقية التي احتلتها اسرائيل **وضمت جزء كبير منها عام ١٩٦٧** مستقبل الفلسطينين | | | | | | | | |
| The issue of the future of East Jerusalem, which *Israel occupied* **and annexed a large part of in 1967**, is the future of the Palestinians | | | | | | | | |

## Annotation of the order Arg1_DC_Arg2_Arg1

Annotators sometimes failed to distinguish the boundaries of arguments Arg1 and Arg2. The rest of Arg1 might be included in Arg2 by mistake such as in Ex. 7-10, or even be missed and not marked as part of any arguments.

**Ex. 7-10**

**(incorrect)**

طلبت كوريا الجنوبية **خلال القمة التاريخية بين كيم داي جونغ وكين جونغ ايل في يونيو الماضي فتح الحدود للعوائل**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Tlbt | kwryA | Aljnwbyp | xlAl | Alqmp | AltAryxyp | byn | Kym dAy jwng |
| request | Korea | Southen | during | summit | historic | between | Kim Dae Jung |
| Wkyn jwng Ayl | fy | HzyrAn | | AlmADy | ftH | AlHdwd | llEwA}l |
| And- Kim Jong Il | in | June | | last | open | border | For-family |

*South Korea had requested* **during the historic summit between Kim Dae Jung and Kim Jong Il last June to open the border for families**

**(correct)**

طلبت كوريا الجنوبية **خلال القمة التاريخية بين كيم داي جونغ وكين جونغ ايل في يونيو الماضي** فتح الحدود للعوائل

*South Korea had requested* **during the historic summit between Kim Dae Jung and Kim Jong Il last June** *to open the border for families*

## The argument is more than one sentence/clause

However, only one sentence is marked as an argument. In Ex. 7-11, Arg2 consists of two abstract objects expressed in two sentences.

**Ex. 7-11**

**(incorrect)**

الجنرال سوموزا يتخلى عن الحكم **بعد معارك استمرت خمسة اسابيع ضد الساندينيين** واسفرت عن مقتل ٤٠ الف مدني

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AljnrAl | swmwzA | ytxlY | En | AlHkm | bEd | mEArk | Astmrt | xmsp |
| *General* | *Somoza* | *resigns* | from | power | after | batels | lasting | five |
| AsAbyE | Dd | AlsAndynyyn | wAsfrt | En | mqtl | 40 | Alf | mdny |
| weeks | against | Sandinistas | and-result | on | kill | 40 | thousand | civilian |

*General Somoza resigns from power* **after fighting lasted five weeks against the Sandinistas**, (and) killed around 40 thousand civilians

**(correct)**

الجنرال سوموزا يتخلى عن الحكم **بعد معارك استمرت خمسة اسابيع ضد الساندينيين واسفرت عن مقتل ٤٠ الف مدني**

General Somoza resigns from power **after fighting lasted five weeks against the Sandinistas, (and) killed around 40 thousand civilians**

## Conjunction of noun/verb phrases and relative clauses

Recognizing the boundaries of phrases which are under a conjunction relation is sometimes difficult for annotators. For example, an annotator might include the matrix clause in Arg1 as in Ex. 7-12. However, this inclusion is against the minimality principle in our scheme. Ex. 7-13 and Ex 7-14 are also examples of such disagreement.

**Ex. 7-12**

**(incorrect)**

اضاف الشاهد ان عسكريين طلبوا من سكان الحي الإحتفاظ بالهدوء ***و البقاء في منازلهم***

| ADAf | Al$Ahd | An | Eskryyn | TlbwA | mn | skAn | AlHy | Al<HtfAZ | bAlhdw' | w | AlbqA' | fy | mnAzlhm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| added | witness | that | *army* | asked | from | residents | area | keep | calm | and | stay | in | their-homes |

*The witness added that the army asked residents to keep calm* **and stay at their homes.**

**(correct)**

اضاف الشاهد ان عسكريين طلبوا من سكان الحي الإحتفاظ بالهدوء **و البقاء في منازلهم**

The witness added that the army asked residents *to keep calm* **and stay at their homes.**

**Ex. 7-13**

**(incorrect)**

ان التدريبات للايام القادمة ستكون حول الخطوة العسكرية **و الزحف على البطن**

| An | AltdrybAt | llAyAm | AlqAdmp | stkwn | Hwl | AlxTwp | AlEskryp | w | AlzHf | ElY | AlbTn |
|---|---|---|---|---|---|---|---|---|---|---|---|
| that | *exercises* | For-days | next | Will-be | about | The-step | *military* | and | **crawl** | on | **belly** |

*The exercises of the day will be on the military stepping* **and the crawl on a belly**

**(correct)**

ان التدريبات للايام القادمة ستكون حول الخطوة العسكرية **و الزحف على البطن**

The exercises of the day will be on *the military stepping* **and the crawl on a belly**

**Ex 7-14**

**(incorrect)**

هذا البرنامج الذي يجدد كل ستة اشهر **والذي دخل في يونيو مرحلته الثامنة**

| h*A | AlbrnAmj | Al*y | yjdd | kl | stp | A$hr | wAl*y | dxl | fy | ywnyw | mrHlth | AlvAmnp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| this | program | which | renewed | every | six | months | And-which | enter | in | June | Its-stage | eighteenth |

*This program, which is renewed every six months*, (and) **which he entered in June eighth stage**

**(correct)**

هذا البرنامج الذي يجدد كل ستة اشهر **والذي دخل في يونيو مرحلته الثامنة**

This program, *which is renewed every six months*, (and) **which he entered in June eighth stage**

### Connectives at BOP and the minimality principle

In news articles, the common usage of connectives at the beginning of paragraph is a conjunction among discourse units. However, since the first argument could be any abstract object prior to the connective, it is subject to wide confusion as to which paragraph/sentence is most closely conjoined to the sentence introduced by the connective. In many cases, several prior discourse units are legitimate annotations. One proposed solution is to limit the annotation of Arg1 to the closest potential discourse unit.

### Attribution and function words

We do not annotate attribution and our guidelines only give very short guidelines that are not sufficient to cover in all instances whether attribution should be included or not. Our annotation guidelines given to the annotators are in Appendix B. Therefore, in various cases annotators disagreed on argument length and attribution inclusion. A later version of the LADTB should handle attribution in more principled way, following discussion in (Prasad *et al.* 2007a) and how attributions apply in Arabic.

## 7.5.3 Ambiguity in Discourse Relations

The common disagreement cases between annotators with regard to annotating discourse relations are presented in the following sections:

### Different relations for *و/w/and* at beginning of paragraph

Annotators often assigned different relations because of the different Arg1 boundaries they marked.

### Entity relations and Exemplification

The conjunction *و/w/and* introduces arguments of an entity relation as well. Arg2 might describe entities in prior discourse such as people, locations and organizations and not abstract objects. We deal with such entity relations that appear as conjoined clauses in MSA, as conjunction relations in our annotation scheme (see Section 5.6.4). Therefore, we annotate the connective *و/w/and* with the EXPANSION.Conjunction relation. However, these entity relations are sometimes understood by the annotators as exemplification relations between two discourse segments, such as in Ex. 7-15, where Arg2 is linked to *الأحلام/the dreams* which is not an abstract object and not to *تتبدد الأحلام/disappearing the dreams* which is an abstract object. This kind of relation might be translated as complement in English with no use of any connectives such as in '*their dreams might disappear which are to win the cup and regain control of the continent of Asia'* and in *'their dreams that they win the cup and regain control of the continent of Asia might disappear'.*

**Ex. 7-15 (incorrect: Exemplification, correct: Conjunction)**

يمكن ان تتبدد الأحلام **و هي احراز الكأس و استعادة السيطرة على قارة آسيا**.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ymkn | An | ttbdd | Al>HlAm | w | hy | AHrAz | Alk>s | w | AstEAdp | AlsyTrp |
| possible | that | lost | the-dreams | and | it | get | the-cup | and | regain | power |
| ElY | qArp | \|syA | | | | | | | | |
| on | continent | Asia | | | | | | | | |

*It possible that the dreams disappear* **(and) they are to win the cup and regain control of the continent of Asia.**

### TEMPORAL relations: Synchronous or Asynchronous

Determination of the overlap period between the events expressed by the two arguments is not very clear in some cases. For example, which temporal period should be considered in Arg2 in Ex. 7-16: *الحرب/the war* or *اندلاع الحرب الأهلية/starting of the war*. The relation should be TEMPORAL.Asynchronous if the latter is annotated.

**Ex. 7-16 (incorrect: TEMPORAL.Synchronous, correct: TEMPORAL.Asynchronous)**

توفي نصف مليون طفل **منذ اندلاع الحرب في افغانستان**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| twfy | nSf | mlywn | Tfl | mn* | AndlAE | AlHrb | fy | AfgAnstAn |
| died | half | *million* | child | since | starting | the-war | in | Afghanistan |

*Half a million children have died* **since starting the war in Afghanistan**

## Pragmatic vs. non-pragmatic relations

Pragmatic/indirect relations are easily missed by the annotators. That might be because they are less frequent in our corpus. The connective *إذ/A*/as* in Ex. 7-17 indicates a Reason relation but because Arg2 expresses an evidence of 'being unable to impose control over the events in the match' and is not a direct reason, it should be Pragmatic reason.

**Ex. 7-17 (incorrect: CONTINGENCY.Reason.NonPragmatic , correct: CONTINGENCY.Reason.Pragmatic)**

عجزوا عن فرض السيطرة على مجريات المباراة **إذ اهتزت شباكهم مبكر**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| EjzwA | En | frD | AlsyTrp | ElY | mjryAt | AlmbArAp | A* | Ahtzt | $bAkhm | mbkr |
| unable | on | impose | control | on | actions | the-match | as | moved | Their-net | early |

*They were unable to impose control over the events in the match*, **as their goal's net was hit earlier**

## Reason or Result relations

The basic guidance in distinguishing between Reason and Result relations is based on what Arg2 expresses to Arg1, reason or result. However, this was not always clear for annotators. For example in Ex. 7-8, the *اصطدام/collision* in Arg2 is a reason for the damages in Arg1. But one annotator was confused by the meaning of the connective *نتيجة ل/natyjp li/resulting for*, thus he annotated it as Result relation.

Moreover, the connective *ل/l/for* usually indicates a Reason relation but this is not the case in Ex. 7-19; where Arg2 *تجديد عقود ابرز اللاعبين/renewing contracts of famous players* describes how *استفاد منها الفريق كثيرا/they got a huge benefit* in Arg1. One annotation was Reason and the other was Result. However, it is a Reformulation relation instead of causal.

**Ex. 7-18 (incorrect: CONTINGENCY.Cause.Result.NonPragmatic, correct: CONTINGENCY. Cause. Reason.NonPragmatic)**

تعرضت لاضرار **نتيجة للاصطدام**

| | | | |
|---|---|---|---|
| tErDt | lADrAr | ntyjp | llASTdAm |
| had | damage | result | For-collision |

*It has been damaged* **as a result of the collision**

**Ex. 7-19 (incorrect: CONTINGENCY.Cause.Result.NonPragmatic or CONTINGENCY.Cause. Reason.NonPragmatic, correct: EXPANSION. Reformulation)**

استفاد مدير الفريق من المنحة الكثير **لتجديد عقود ابرز اللاعبين**.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| AstfAd | mdyr | Alfryq | mn | AlmnHp | Alkvyr | ltjdyd | Eqwd | Abrz | AllAEbyn. |
| benefit | manager | The-team | from | scholarship | huge | For-renew | contracts | important | players |

*The team's manager got a huge benefit from the scholarship* **by renewing contracts of famous players.**

## 7.6 The Gold standard LADTB

Deriving a gold standard version requires extra annotation for the remaining disagreements at all levels {discourse connectives (1013), relations (775) and arguments (Arg1: 3063, Arg2: 1355)} by an adjudicator not initially involved in the annotation. The adjudicator was me (the main researcher) as I have conducted all discussions and am an expert in discourse annotation following our guidelines for Arabic. In addition, a decision was made to include annotation of 5 new potential connective types not in our initial connective list but commented on by the annotators during the annotation process. These new annotations were done by me and not included in any agreement studies. Disagreements of connectives and relations were grouped by their occurrence in files and I re-annotated them according to the results of previous discussions with the annotators during the agreement studies on those instances. Three files were removed as well from the corpus because they contain no discourse connectives.

Regarding the disagreements of arguments, we have three situations: first, non-overlapping arguments with zero *exact match* (Arg1: 677, Arg2: 4). Second, arguments with up to 80% overlap (Arg1: 1829, Arg2: 944). Third, arguments with

more than 80% overlap (Arg1: 557, Arg2: 407). For the latter case, the disagreements were manipulated automatically by keeping only the overlapping tokens. For no overlap cases, one of the annotations was chosen with slight modification if necessary.

The heaviest work in the post-processing stage was for arguments with agreement up to 80%. Our guidelines of the correction focus on the common cases which were discussed in the disagreements of argument boundaries in Section 7.5.2. This ensured consistent correction for these cases. Other individual cases were also manipulated as required.

The final discourse treebank we produced has 6,328 annotated explicit connectives in 534 files. 68 connective types were found, rising to 80 connective types if we include all modified forms of a connective as distinct types such as *بالرغم/bAlrgm* and *رغم ان/rgm An* which are modified forms of *رغم/rgm/although*. 27 Arabic connective types from our initial discourse connective collection (Section 4.5) are not used on the LADTB.

All 17 discourse relations in our relation taxonomy appear in the LADTB. Most of the discourse connectives (95%) were annotated with a single relation and 5% were annotated with two relations. These statistics are summarized in Table 7-5.

**Table 7-5: Statistics of the final gold standard corpus LADTB**

| | | |
|---|---|---|
| **Total tagged Tokens** | | 126,394 |
| **Files** | | 534 |
| **Total Paragraphs** | | 3312 |
| **Total Sentences** | | 3607 |
| **Total potential discourse connectives** | 20312 | 100% |
| - Discourse connectives | 6,328 | 31% |
| - Not a discourse connective | 13984 | 69% |
| **Discourse connective types** | | 80 |
| **Discourse relation types** | 55 | 100% |
| - Single relations | 17 | 31% |
| - Combined relations | 38 | 69% |
| **Total discourse connective tokens** | 6,328 | 100% |
| - Single Relation tokens | 6039 | 95% |
| - Combined relation tokens | 289 | 5% |

## Discourse connectives

Our categorization of discourse connectives is based on the status of the connective in raw text rather than in the ATB. The syntactic annotation of the Arabic Treebank does not consider the discourse function of the connectives, for example, some phrasal discourse connectives are not syntactically phrases. Therefore, it is better not to base our categorization of connectives on the ATB annotation.

The types of our connectives and their position in the sentence are shown in Table 7-6. The majority of discourse connectives in the LADTB are clitics (76%) including the conjunctions *و/w/and*, *ف/f/then* and the prepositions *ل/l/for* and *ب/b/by*. Table 7-7 lists the most frequent discourse connectives and their POS tags in the LADTB, consisting almost exclusively of conjunctions and prepositions. Only 4% of the tokens are MoreThanToken connectives presenting 24 connective types, some of which are syntactically not phrases. 20% of the connective are simple, one token not attached to other words.

40% of the discourse connectives are located at the beginning of a sentence (BOS) and 60% are in the middle of a sentence or a clause (Moser and Moore 1996). Unlike English, there are no connectives in the LADTB located at the end of sentences. If we exclude the instances of *و/w/and* at BOS (around 2400), we reach the very interesting result that only 147 (3%) of non *و/w/and* connectives are located at BOS and the remainder including *و/w/and* is 3741 (60%) connectives are at MOS, mostly relate two arguments located at the same sentence (intra-sentential connectives). This result might not apply for other genres in Arabic. The promising hypothesis here, it is possible to automatically identify arguments of majority of Arabic connectives in the LADTB with a high performance apart from *و/w/and* at BOS. A special discourse study is strongly needed for *و/w/and* at BOS and BOP to check whether this kind of connectives behaves like implicit connectives in English.

Table 7-7 shows the 18 most frequent discourse connectives in the LADTB. The table shows the total occurrences of each connective as discourse and non-discourse predicate. The last two columns show the ambiguity status of a connective in terms of the number of relations the connective signals and the most frequent relation. A full distribution of Arabic connectives is shown in the Appendix C.

**Table 7-6: Discourse connective types and location in the LADTB.**

| **Types of discourse connectives** | 6,328 | 100% |
|---|---|---|
| Simple | 1276 | 20% |
| Clitic | 4779 | 76% |
| MoreThanToken | 273 | 4% |
| **Connective position in a sentence** | | |
| Beginning of sentence - BOS | 2587 | 41% |
| و*/w/and* at BOS | 2440 | 38.6% |
| Non و*/w/and* at BOS | 147 | 2.4% |
| End of sentence - EOS | 0 | - |
| Middle of sentence - MOS | 3741 | 59% |

Two types of ambiguity arose to the surface when analysing the distribution of connectives, which highlight the difficulty of recognizing discourse connectives and identifying the relations automatically. First, the ambiguity of having a discourse function, only few connectives appear more than 90% of the time as discourse connectives in the LADTB. For instance, the connective ل*/l/for* has a discourse function only 11% of the time it appears. Second, ambiguity with regard to which discourse relations the connective conveys. For example, the connective فبينما*/fy mA/while* is indicating a Contrast relation 36% of the time, leaving the rest for six other relations. The ambiguity problems will be discussed in more detail in Sections 8.28.5 and 8.5.

**Table 7-7: The most frequent discourse connectives in the LADTB v.1**

| **Connective** | **Total** | **Non Dis.Conn** | | **Dis.Conn** | | **#Rel** | **The most frequent relation** |
|---|---|---|---|---|---|---|---|
| و*/w/and* | 7375 | 3376 | 46% | **3999** | 54% | 31 | {76%:EXPANSION.Conjunction |
| ل*/l/for* | 4306 | 3838 | 89% | **468** | 11% | 4 | {93%:CONTINGENCY.Cause.Reason. NonPragmatic (437)} |
| لكن*/lkn/ however* | 207 | 3 | 1% | **204** | 99% | 5 | {97%:COMPARISON.Contrast (198)} |
| بعد*/bEd/after* | 315 | 121 | 38% | **194** | 62% | 7 | {51%:TEMPORAL.Asynchronous |
| ف*/f/then* | 1525 | 1426 | 94% | **99** | 6% | 13 | {29%:CONTINGENCY.Cause.Result. |
| ب*/b/by* | 4168 | 4072 | 98% | **96** | 2% | 4 | {89%:CONTINGENCY.Cause.Reason. NonPragmatic (86)} |
| منذ*/mn*/since* | 220 | 151 | 69% | **69** | 31% | 5 | {69%:TEMPORAL.Asynchronous (48)} |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| كما/kmA/asl | 105 | 36 | 34% | **69** | 66% | 11 | {57%:EXPANSION.Conjunction (40)} |
| عندما /EndmA /whenl | 55 | 1 | 2% | **54** | 98% | 10 | {51%:TEMPORAL.Synchronous (28)} |
| الا ان//AlA An/but | 41 | 0 | 0% | **41** | 100% | 4 | {92%:COMPARISON.Contrast (38)} |
| ثم/vm/then | 48 | 12 | 25% | **36** | 75% | 4 | {91%:TEMPORAL.Asynchronous (33)} |
| فيما/fy mA/while | 41 | 5 | 12% | **36** | 88% | 7 | {36%:COMPARISON.Contrast (13)} |
| حيث /Hyv/ where/since | 96 | 64 | 67% | **32** | 33% | 10 | {40%:CONTINGENCY.Cause.Reason. NonPragmatic (13)} |
| حتى/HtY/until | 75 | 46 | 61% | **29** | 39% | 12 | {20%:CONTINGENCY.Cause.Reason. NonPragmatic (6)} |
| في حين/fy Hyn/while | 28 | 1 | 4% | **27** | 96% | 4 | {44%:COMPARISON.Contrast (12)} |
| خصوصا /xSwSA/speciall | 64 | 41 | 64% | **23** | 36% | 7 | {39%:EXPANSION.Reformulation (9)} |
| بعدما/bEdmA /after that | 23 | 0 | 0% | **23** | 100% | 4 | {52%:CONTINGENCY.Cause.Reason. NonPragmatic/TEMPORAL.Asynch-ronous(12)} |
| اذ/A*/as | 22 | 0 | 0% | **22** | 100% | 8 | {45%:CONTINGENCY.Cause.Reason. NonPragmatic (10)} |

## Discourse relations

Although we have in the LADTB 38 combined relations, 95% of the annotated tokens signal one of the 17 single discourse relations. We report that distribution of distinct relations togather with the frequency that each discourse connective conveys the relation. For example, Table 7-8 presents details of the Condition relation: it is used 77 times in the LADTB with 10 different discourse connectives for indicating the relation in context. For each connective we present the following data: (i) how often the relation is signalled by the connective (e.g. 45.5% of the instances of the relation Condition are signalled by the connective *في حال/fy HAl/in case*), (ii) the discourse connective frequency out of the total of the discourse connective occurrences in the LADTB and its percentage. For example, the connective *في حال/fy HAl/in case* signals a Condition relation 35 times out of the 42 times the connective occurs in the LADTB, thus signalling Condition 83% of the time. The two most common connectives signalling the Condition relation in the LADTB are {45.5%: *في حال/fy HAl/in case* (35 OutofConnTotal 42⁄83%)} and {41.6%: *اذا/A*A/if* (32 OutofConnTotal 49⁄65%)}. Therefore, around 13% of Condition instances are

signalled by other connectives, see Table 7-8. The full distribution of relations is in the Appendix D.

**Table 7-8: A distribution of only one relation CONTINGENCY.Condition. The full distribution of other relations is shown in Appendix D.**

| Discourse Relation | Total | Discourse Connectives | #Dis. Conn |
|---|---|---|---|
| CONTINGENCY.Condition | 77 | {45.5%: في حال/fy HAl (35,OutofConnTotal: 42/ 83 %)}<br>{41.6%: اذا/A*A (32, OutofConnTotal: 49/ 65%)}<br>{2.6%: لو/lw (2, OutofConnTotal: 14/ 14%)}<br>{2.6%: طالما/TAlmA (2, OutofConnTotal: 4/ 50%)}<br>{1.3%: و/w (1, OutofConnTotal: 7375/ 0.0%)}<br>{1.3%: لولا/lwlA (1, OutofConnTotal: 1/ 100%)}<br>{1.3%: عندما/EndmA (1, OutofConnTotal: 55/ 2%)}<br>{1.3%: حتى/HtY (1, OutofConnTotal: 75/ 1 %)}<br>{1.3%: حال/HAl (1, OutofConnTotal: 2/ 50%)}<br>{1.3 %: الا اذا/AlA A*A (1, OutofConnTotal: 2/ 50%)} | 10 |

Apart from the EXPANSION.Alternative relation, which is signalled by only one connective *او/Aw/or*, all relations are signalled explicitly by different connectives. Table 7-9 lists the most frequent relations and the number of discourse connectives that are used to indicate the relation. The most frequent relations in the LADTB are Conjunction, Reason, Contrast and Temporal.Asynchronous. This is not surprising because in news it is normal to provide more justifications and to report events in temporal order. On the other hand, Condition and pragmatic relations are used less frequently in the LADTB. This might differ for different genres in Arabic.

**Table 7-9: List of the most frequent relations ordered by the number of distinct discourse connective types signalling the relation in the LADTB**

| Discourse Relation | #Dis. Conn | Total |
|---|---|---|
| CONTINGENCY.Cause.Reason.NonPragmatic | 26 | 806 |
| COMPARISON.Contrast | 25 | 440 |
| EXPANSION.Conjunction | 19 | 3167 |
| TEMPORAL.Asynchronous | 17 | 417 |
| TEMPORAL.Synchronous | 15 | 219 |
| CONTINGENCY.Cause.Reason.NonPragmatic/ TEMPORAL.Asynchronous | 11 | 157 |
| CONTINGENCY.Cause.Result.NonPragmatic | 10 | 228 |
| CONTINGENCY.Condition | 10 | 77 |
| EXPANSION.Reformulation | 10 | 331 |
| CONTINGENCY.Cause.Reason.Pragmatic | 8 | 28 |

| Discourse Relation | #Dis. Conn | Total |
|---|---|---|
| EXPANSION.Exemplification | 8 | 47 |
| CONTINGENCY.Cause.Result.Pragmatic | 7 | 33 |
| COMPARISON.Contrast/ TEMPORAL.Asynchronous | 6 | 11 |
| CONTINGENCY.Cause.Result.NonPragmatic/ TEMPORAL.Asynchronous | 6 | 22 |
| COMPARISON.Contrast/ TEMPORAL.Synchronous | 5 | 19 |
| CONTINGENCY.Cause.Reason.Pragmatic/ TEMPORAL.Asynchronous | 5 | 14 |
| CONTINGENCY.PragmaticCondition | 4 | 6 |
| EXPANSION.Exception | 4 | 5 |
| CONTINGENCY.Cause.Reason.NonPragmatic/ TEMPORAL.Synchronous | 4 | 14 |
| EXPANSION.Background | 3 | 186 |
| CONTINGENCY.Cause.Reason.Pragmatic/ TEMPORAL.Synchronous | 3 | 3 |
| COMPARISON.Similarity | 2 | 14 |

## 7.7 LADTB and PDTB in Comparison

We compare our annotation outcomes for Arabic newswire in the LADTB with the recent version of the PDTB for English news. There are several reasons why any comparison between the PDTB and the LADTB can only lead to approximate conclusions for bilingual studies for English and Arabic. First, the PDTB is three times larger than the LADTB. Second, there is only an approximate match in genre as the LADTB contains newswire reports whereas the PDTB contains a wider range of news texts (including letter to the editor, ..etc). Third, and most importantly, both corpora reflect the discourse proprieties of the language only through the mirror of annotation decisions made by its developers. An example, in the PDTB some subordinate such as *'in order to'* and '*so that*' are not yet annotated as discourse connectives. Therefore, counts of, for example, intra-sentential connectives are an underestimate of intra-sentential explicit discourse relations in English news. Therefore, all following comparisons yield only hypotheses on language similarities and differences, that need further linguistic and corpus-linguistic in future work. We still believe that the overall annotation principles used are similar enough to yield hypotheses and observations worth pursuing.

A general statistical comparison of the LADTB and PDTB is shown in Table 7-10. We have used white space separated tokens to collect the potential discourse connectives in English, as this figure is not reported in any published works. However, this is not the case for Arabic, as we also include the possibility of having clitics as connectives. Only half as many of the relation types of the PDTB are used in Arabic due to a less fine-grained taxonomy at the lowest level. In addition, in English, any combination of different relations at (potentially) different levels is allowed whereas we only allow relation combinations at the most fine-grained level. 95% of the annotations in both corpora are for single discourse relation usages.

**Table 7-10: General comparison statistics of discourse annotation for Arabic (LADTB) and for English (PDTB)**

| | **LADTB** | | **PDTB** | | **LADTB: PDTB** |
|---|---|---|---|---|---|
| **Total tagged Tokens** | 126394 | | 1253013 | | 10% |
| **Files** | 534 | | 2159 | | 25% |
| **Potential discourse connectives** | 20312 | 100% | 55601 | 100% | 37% |
| - Explicit Discourse connectives | 6,328 | 31% | 18459 | 33% | 34% |
| - Non-discourse connectives | 13984 | 69% | 37142 | 67% | 38% |
| **Discourse connective types** | 80 | | 100 | | 80% |
| **Distinct discourse relation types** | 55 | | 111 | | 50% |
| - Single relation types | 17 | 31% | 32 | 29% | 53% |
| - Combined relation types | 38 | 69% | 79 | 71% | 48% |
| **Single relation tokens** | 6039 | 95% | 17490 | 95% | 35% |
| **Combined relation tokens** | 289 | 5% | 969 | 5% | 30% |

In general, coordinating conjunctions and prepositions are frequently used connectives in the LADTB, while coordinating/subordinating conjunctions are the most frequently used connectives in the PDTB, as shown in Table 7-11. Prepositions are not yet annotated in the English PDTB as potential discourse connectives. For example, prepositions such as *to/for/during* and *ب/b/by* are considered as potential discourse connectives in Arabic only. The extremely high usage of *و/w/and* (63%) affects the distribution of the connectives in the LADTB. This is due to genre specific properties in Arabic. In addition, unlike English, the conditional connective *إذا/A*A/if* does not appear in the list of frequent Arabic discourse connectives in Table 7-11. The common POS tags in the PDTB and LADTB are given in p.xvii.

**Table 7-11: The most frequent explicit discourse connectives in the LADTB and the PDTB**

| **Total annotation tokens in the LADTB** | | 6,328 | | **Total annotation tokens in the PDTB** | | 18419 | |
|---|---|---|---|---|---|---|---|
| **Conn** | **ATB POS** | **Total** | **%** | **Conn** | **POS** | **Total** | % |
| و/w/and | ABBREV, CONJ | 3999 | 63.2% | But | CC, IN | 3308 | 18% |
| ل/l/for | EMPHATIC_ PARTICLE, PREP, SUBJUNC | 468 | 7.4% | and | CC, NN, JJ | 3000 | 16.3% |
| لكن/lkn/but | CONJ, NO_FUNC | 204 | 3.2% | also | RB | 1746 | 9.5% |
| بعد/bEd/after | PREP | 194 | 3.1% | if | IN | 1158 | 6.3% |
| خلال/xlAl/duri | PREP | 102 | 1.6% | when | WRB | 945 | 5.1% |
| ف/f/then | CONJ | 99 | 1.6% | as | RB, IN | 861 | 4.7% |
| ب/b/by | PREP | 96 | 1.5% | because | IN, RB | 783 | 4.3% |
| قبل/qbl/before | PREP | 84 | 1.3% | while | IN, | 778 | 4.2% |
| لان/lAn/becau | CONJ | 80 | 1.3% | after | IN, RB | 487 | 2.6% |
| كما/kmA/as | CONJ | 69 | 1.1% | however | RB | 485 | 2.6% |
| منذ/mn*/since | CONJ, NO_FUNC, PREP | 69 | 1.1% | Although | IN | 328 | 1.8% |
| اثر/Avr/after | PREP | 67 | 1.1% | so | IN,RB, | 295 | 1.6% |
| عندما/EndmA/ when | CONJ, REL_ADV | 54 | 0.9% | before | IN, RB | 283 | 1.5% |

Regarding the location of arguments, 3741 (60%) of the connectives in the LADTB have connectives in middle of sentence, most of them are intra-sentential (having both arguments in the same sentence). See Table 7-6 and and Section 8.6.1 that we use position of arguments as a feature in our modeling of discourse relations. This number is a comparable with the 11236 (61%) intra-sentential annotated tokens in the PDTB2 (Prasad *et al.* 2008a). Next section will discuss the number of tokens when arguments are located in different sentences, inter-sentential tokens.

### 7.7.1 Inter-sentential Relations

Discourse coherence can be a result of having relations across sentences or so called inter-sentential discourse relations. Thus, we examine the strength of inter-sentential

discourse relations in both languages by counting the explicit relations between adjacent sentences in the PDTB and the LADTB. It is important to note that adjacent sentences might be related via non-discourse relation such as Entity relations (PDTB: EntRel, 5210) as well as discourse relations. Also, some sentences might be linked via non-connective lexical expressions (PDTB: AltLex, 624) (Prasad *et al.* 2008a). Both types were not annotated for Arabic in the LADTB. Therefore, a comparison of the explicit inter-sentential relations is a rough estimate of how adjacent sentences linked in the news of English and Arabic, using the available resources the PDTB2 and the LADTB.

We count all two adjacent trees with S tag in the treebank (excluding trees with X tags) as an adjacent sentence pair (ASP). There are 44,470 ASP in the PDTB and 3,073 ASP in the LADTB. Among these, each pair has two arguments located in a different S tree linked via (Explicit relations or AltLex) in the PDTB, and Explicit relations in the LADTB is counted as an explicit inter-senential relation. In particular, the focus was on connectives of argument orders Arg1_DC_Arg2 and DC_Arg2_Arg1. The tree might represent the whole argument or with text beyond the argument boundaries. The question here is whether Arabic follows English in its frequency of explicit inter-sentential discourse relations between adjacent sentences.

**Table 7-12: Inter-sentential adjacent sentences linked explicitly in the LADTB compared to the PDTB**

| **Inter-sentential relations** | **LADTB** | **PDTB** |
|---|---|---|
| Adjacent sentence pairs (ASP) | 3,073 | 44,470 |
| AltLex | NA | 624 (1.5%) |
| ASP linked via explicit DCs | 2,140 (70%)<br>Non-و/w/and: 948 (30%) | 5,549 (12.5%) |
| Total | 2,140 (70%) | 6,173 (14%) |
| ASP not linked via explicit DCs | 933 (30%) | 38,297 (86%) |

Table 7-12 shows that 70% of adjacent sentence pairs in Arabic are linked via explicit connectives comparing to only 12% of ASPs in English. Moreover, even if we exclude *و/w/and* at beginning of sentences, still 30% of adjacent pairs are linked via an explicit connective in the LADTB. Adding all types of explicit discourse links between ASP in the PDTB (Explicit +AltLex), makes only 14% linked explicitly in

English news. This interesting result stresses the importance of the explicit connectives for Arabic discourse processing.

### 7.7.2 Discourse Relation Comparison

The discourse relations taxonomy in the PDTB, the so called sense hierarchy, has more fine-grained relations than the current relations taxonomy for Arabic (see Section 5.5). Thus, in discourse relation comparison, we exclude connectives that do not have equivalent relations in both LADTB and PDTB taxonomies. For example, we exclude the tokens annotated with EXPANSION.Background and CONTINGENCY.Similarity as there are no corresponding relations in PDTB. On the other hand, as the PDTB has deeper fine-grained relations, we combined all lower level relations in the PDTB into one upper level relation that has an equivalent description in the LADTB.

Table 7-13 shows a statistical comparison of discourse relations in the LADTB and the PDTB. Two different sets of LADTB are examined: Set 1 includes all connectives, and Set 2 excludes tokens of *و/w/and* at BOP, as the disagreed instances of this connective are annotated automatically with Conjunction relation in the LADTB. In the most sensible comparison dataset of the PDTB, Set 2, the majority of relations in both corpora are single relations, ~95%. Although the distribution of relations is very similar in both languages, Causal and Reformulation relations are used in Arabic more than double the frequency than in English. On the other hand, Contrast relations are more frequently used in English news than in Arabic.

It is not completely clear whether these differences are due to (i) intrinsic differences between how discourse is structured in the two languages or (ii) differences in how the news genre is realized in the different cultural settings. We also remind the reader that the genre in the two corpora is not completely identical (newswire vs. news, see Section 7.7). Future work looking also at journalistic connectives should address this question.

| **LADTB v.1** | **Set 1** | | **Set 2** | | | **PDTB2** | | |
|---|---|---|---|---|---|---|---|---|
| **Single relations** | 6039 | 95% | 3814 | 93% | | **Single relations** | 17450 | 95% |
| **Combined relations** | 289 | 5% | 285 | 7% | | **Combined relations** | 969 | 5% |
| **Total relations** | 6,328 | 100% | 4099 | 100% | | **Total relations** | 18419 | 100% |
| **A comparison of only equivalent single relations in the LADTB and PDTB** | | | | | | | | |
| **LADTB v.1** | **Set 1** | | **Set 2** | | | **PDTB2** | | |
| CONTINGENCY | 1178 | 20.2% | 1162 | 30.8% | | CONTINGENCY | 3104 | 19.9% |
| CONTINGENCY.Cause | 1034 | 17.7% | 1019 | 27.0% | | CONTINGENCY.Cause | 1725 | 11.0% |
| - CONTINGENCY.Cause.Reason.NonPragmatic | 806 | 13.8% | 804 | 21.3% | | - CONTINGENCY.Cause.Reason | 1135 | 7.3% |
| - CONTINGENCY.Cause.Result.NonPragmatic | 228 | 3.9% | 215 | 5.7% | | - CONTINGENCY.Cause.Result | 590 | 3.8% |
| CONTINGENCY.Condition | 77 | 1.3% | 77 | 2.0% | | CONTINGENCY.Condition | 1307 | 8.4% |
| CONTINGENCY.Cause.Pragmatic | 61 | 1.0% | 60 | 1.6% | | CONTINGENCY.Cause.Pragmatic | 7 | 0.0% |
| - CONTINGENCY.Cause.Result.Pragmatic | 33 | 0.6% | 33 | 0.9% | | - | | |
| - CONTINGENCY.Cause.Reason.Pragmatic | 28 | 0.5% | 27 | 0.7% | | - | | |
| CONTINGENCY.PragmaticCondition | 6 | 0.1% | 6 | 0.2% | | CONTINGENCY.PragmaticCondition | 65 | 0.4% |
| TEMPORAL | 636 | 10.9% | 618 | 16.4% | | TEMPORAL | 2922 | 18.7% |
| TEMPORAL.Asynchronous | 417 | 7.1% | 401 | 10.6% | | TEMPORAL.Asynchronous | 1835 | 11.7% |
| TEMPORAL.Synchronous | 219 | 3.8% | 217 | 5.8% | | TEMPORAL.Synchronous | 1087 | 7.0% |
| COMPARISON | 440 | 7.5% | 425 | 11.3% | | COMPARISON | 3786 | 24.2% |
| COMPARISON.Contrast | 440 | 7.5% | 425 | 11.3% | | COMPARISON.Contrast | 3786 | 24.2% |
| COMPARISON.Similarity | - | - | - | - | | - | | |

| **LADTB v.1** | **Set 1** | | **Set 2** | | | **PDTB2** | | |
|---|---|---|---|---|---|---|---|---|
| EXPANSION | 3585 | 61.4% | 1566 | 41.5% | | EXPANSION | 5817 | 37.2% |
| EXPANSION.Conjunction | 3167 | 54.2% | 1341 | 35.6% | | EXPANSION.Conjunction | 4968 | 31.8% |
| EXPANSION.Reformulation | 331 | 5.7% | 142 | 3.8% | | EXPANSION. Restatement | 153 | 1.0% |
| EXPANSION.Exemplification | 47 | 0.8% | 43 | 1.1% | | EXPANSION.Exemplification | 302 | 1.9% |
| EXPANSION.Background | - | - | - | - | | - | | |
| EXPANSION.Exception | 5 | 0.1% | 5 | 0.1% | | EXPANSION.Exception | 14 | 0.1% |
| EXPANSION.Alternative | 35 | 0.6% | 35 | 0.9% | | EXPANSION. Alternative | 190 | 1.2% |
| - EXPANSION.Alternative.Disjunctive | 7 | 0.1% | 7 | 0.2% | | - EXPANSION.Alternative.Disjunctive | 143 | 0.9% |
| - EXPANSION.Alternative.Conjunctive | 28 | 0.5% | 28 | 0.7% | | - EXPANSION.Alternative.Conjunctive | 47 | 0.3% |
| **Total** | **5839** | **100%** | **3771** | **100%** | | **Total** | **15629** | **100%** |

**Table 7-13: A full statistical comparison of single relations in the LADTB and PDTB2 (only equivalent relations at similar and lower levels) – Set 1 all connectives, Set 2 excluding *s/w/and* at BOP.**

**Table 7-14: A statistical comparison of equivalent class level discourse relations in the LADTB (Set 1- all tokens, Set 2 excluding و/w/and at BOP) and the PDTB2.**

| | LADTB v.1 | | | | PDTB2 | |
|---|---|---|---|---|---|---|
| | **Set 1** | **%** | **Set 2** | **%** | | **%** |
| **TEMPORAL** | 636 | 10.9% | 618 | 16.4% | 2922 | 18.7% |
| **CONTINGENCY** | 1178 | 20.2% | 1162 | 30.8% | 3104 | 19.9% |
| **EXPANSION** | 3585 | 61.4% | 1566 | 41.5% | 5817 | 37.2% |
| **COMPARISON** | 440 | 7.5% | 425 | 11.3% | 3786 | 24.2% |
| **Total** | **5839** | **100.0%** | | | **15629** | **100.0%** |

Table 7-14 presents a comparison of equivalent class level relations in both corpora. Figure 7-1 shows a graphical representation of this comparison of only Set 2 (excluding *و/w/and* at BOP in the LADTB), for a sensible argument. Interestingly, more EXPANSION and CONTINGENCY relations are in Arabic, in contrast to the more COMPARISON and TEMPORAL relations in English. As mentioned in Section 7.7, the size and the genre of the corpora might impact on the figures in Table 7-14. Therefore, for a more accurate comparison, a larger annotated discourse corpus is needed for Arabic that contains longer articles from different genres, similar to the Wall Street Journal corpus.

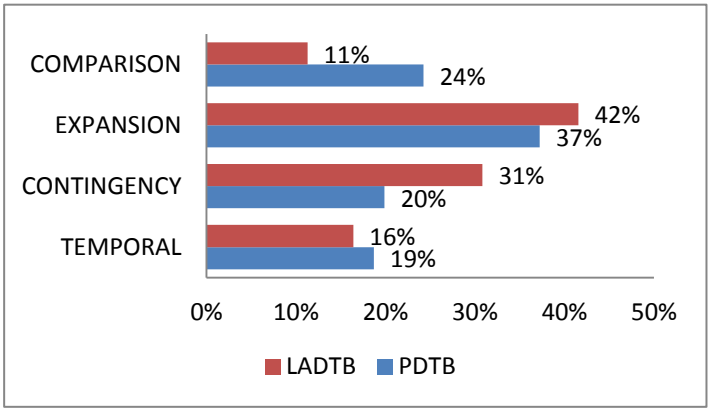



**Figure 7-1: A bar chart of relations in class level of the LADTB (Set 2, excluding *و/w/and* at BOP) and the PDTB2**

## 7.8 Summary

We present the first effort towards producing an Arabic Discourse Treebank, the LADTB v.1; the news corpus where all explicit connectives, associated relations and arguments are annotated.

The human annotation shows that the identification of discourse connectives, their arguments and the determination of the discourse relations they convey are reliable. Overall the annotation of the LADTB follows the annotation principles in the Penn Discourse Treebank for explicit connectives with necessary adaptations with regard to Arabic discourse connectives, relations and arguments. Similar annotation principles were used to annotate discourse connectives in other languages in addition to English such as Turkish, Hindi and Chinese.

We also discussed disagreement cases on the human annotation of connectives, relations and arguments. This discussion was used to derive the gold standard of the annotation using automatic correction for simple errors and manual correction for the rest as a post-processing step. Our current annotated corpus encompasses a final 6,328 annotated discourse connectives in 535 news texts, 80 distinct connective types and 55 discourse relations including single and multiple relations.

A statistical comparison study between discourse annotation in Arabic (the LADTB) and English (the PDTB) was conducted. This comparison in a rough estimate and could not be finial for news in the two languages for several reasons: the size, the genre, and annotation differences of discourse connective types and relation taxonomy. It was shown that the LADTB has more Expansion and Contingency relations than in English, in contrast to more Comparison and Temporal relations in English than in Arabic. However, differences between the PDTB and the LADTB in terms of discourse relations, might reflect how news is reported in English and Arabic, rather than of intrinsic differences of how discourse is structured in the two languages.

The increasing value of this study comes from the result that Arabic uese explicit connectives with high frequency for inter-sentential relations (30% of connectives excluding وَ/w/and at BOS, Section 7.7.1). Also, 60% of the connectives in the

LADTB are located in middle of sentences, most of them are intra-sentential (having both arguments in the same sentence). This will benefit identifying argument boundaries automatically in future work.

This first discourse corpus for Arabic will be used for training and testing automated methods for discourse connective and relation recognition. The LADTB will be released in 2012 via the LDC for people in Arabic NLP to establish advanced studies of discourse processing for Arabic. The corpus might be used to conduct studies for improving computational language applications such as machine translation, question answering, and readability scoring.

# Chapter 8
# Supervised Models for Discourse Processing

## 8.1 Introduction

Discourse modeling for explicit connectives, which is the focus of this study, should cover three main tasks: (i) explicit discourse connective recognition, (ii) interpretation and (iii) arguments assignment. In this first computational discourse study for Arabic, we propose supervised machine learning modeling using the newly built discourse corpus, the LADTB, for training and testing purpose for the first two tasks: recognising the discourse connectives and identifying their discourse relations. The second task focuses on identifying single relations at the fine-grained level (95% of the annotation in the LADTB), as there are very few instances for multiple-relations (289, 5%). Models were also developed to recognise relations at the class level. The automatic arguments assignment lies outside the scope of this study because of time constraints.

Regarding our concentration on explicit discourse connectives in Arabic, we are motivated by our observations in discourse annotation and the statistics of the gold standard LADTB (see Chapter 7). First, explicit discourse connectives are very frequently used in Arabic to relate arguments. As discussed in Section 7.8.1, almost 70% of adjacent sentences/clauses in the LADTB texts are linked explicitly via a connective, 30% were linked via non *وَ/w/and* connectives. In addition, intra-sentential relations (two arguments in the same sentence) tend to be marked by connectives anyway in Arabic. Second, potential Arabic discourse connectives are highly ambiguous in two respects: (i) whether they have a discourse usage or not in a given context and (ii) the discourse relations that they signal. Therefore, modeling of explicit discourse connectives is primary for Arabic discourse studies.

The two ambiguity aspects of connectives in the LADTB are described in detail in Section 8.2 and Section 8.5 respectively[21]. Models of connective recognition achieve very good results, in particular, the model that does not rely on full parsing or gold standard syntactic annotation (see Section 8.4). Full details of data setting, features and results of different models for connective recognition are discussed in Sections 8.3 and 8.3.

With regard to discourse connective disambiguation, we developed supervised learning models that use a wide feature set and that achieve significant improvements over the baseline of the most frequent relation per connective. Full details of data setting, features and results of different models are discussed in Section 8.6. We present in Section 8.6.4 our error analysis of the models to investigate how we could improve the models further. Our models use, in addition to Arabic-specific features, features inspired by prior work for discourse modeling of explicit discourse connectives and implicit relations in English (Marcu 2000; Pitler and Nenkova 2009; Miltsakaki, Dinesh et al. 2005; Lin, Kan and Ng 2009; Wang, Su and Tan 2010). We refer the reader to Section 2.7 for a brief survey of related works.

At the end of this chapter, a summary of our work and observations is presented together with notes on the limitations affecting the study and ideas for additional improvement of discourse modeling for Arabic.

## 8.2 Discourse Usage of Connectives

The potential Arabic discourse connectives do not always have a discourse function in their context. For example, the clitic preposition بـ/*b/by* is a discourse connective in فاز فريق ريال مدريد في التصفيات بتقدمه ٣ أهداف على فريق برشلونة/ *Madrid won its lead in the playoffs by recording 3 goals on Barcelona*, but it is not a discourse connective in الشنطة بالسيارة/*the bag is in the car*. Of the 80 discourse connective types occurring in the LADTB, 42 are almost unambiguous when it comes to discourse usage, i.e. at least 90% of their occurrences are indeed discourse connectives. However, they account only for 860 out of 6,328 discourse connective tokens in the LADTB, leaving 86% of tokens for the 34 discourse connective types with higher levels of ambiguity. Table 8-1 displays the details of unambiguous connectives; 17 of them

[21] The term 'ambiguous connective' varies in its usages, depending on the section's focus.

might only be unambiguous because they occur rarely in the LADTB (< 5 times). The last section in the table presents potential discourse connectives that almost always have discourse usage in context.

**Table 8-1: Unambiguous discourse connective types in terms of discourse function. The connectives in the lower part of the table are almost unambiguous.**

| Conn | | Freq | % Dis. Conn |
|---|---|---|---|
| الا ان/*AlA An/but* | | 41 | |
| بعدما/*bEdmA/after that* | | 23 | |
| اذ/*A*/as* | | 22 | |
| بينما/*bynmA/while* | | 16 | |
| جراء/*jra/because* | | 10 | |
| على الرغم/*ElY Alrgm/although* | | 9 | |
| نظرا ل/*nZrA l/because of* | | 9 | |
| في ظل/*fy Zl/under* | | 6 | |
| بيد ان/*byd An/but* | | 6 | 100% |
| رغم ان/*rgm An/although* | | 6 | |
| غير ان/*gyr An/however* | | 6 | |
| عقب/*Eqb/shortly after* | | 5 | |
| بفضل/*bfDl/thanks to* | | 5 | |
| قبيل/*qbyl/shortly before* | | 5 | |
| في المقابل/*fyAl mqAbl/in contrast* | | 5 | |
| بالرغم من/*bAlrgm mn/although* | | 5 | |
| بغية/*bgyp/desire/to* | | 5 | |
| طالما/*TAlmA/as long as* | من ثم/*mn vm/then* | <5 | 100% |
| لأن /*l>n/because* | قبل ان/*qbl An/before* | | |
| الا اذا/*AlA A*A/except if* | حتى لو /*HtY lw/even if* | | |
| بيد/*byd/but* | كأن /*k>n/as* | | |
| خلافا ل /*xlAfA l/unlike* | برغم /*brgm/although* | | |
| بمعنى اخر/*bmEnY xr /in other words* | الولا/*wlA/if not* | | |
| بالمقابل/*bAlmqAbl/in contrast* | بحيث /*bHyv/since* | | |
| حال /*HAl/when* | كلما/*klmA/when ever* | | |
| وقبل/*wqbl/and before* | | | |

| Conn | Freq | % Dis. Conn |
|---|---|---|
| **Potential discourse connectives often have discourse usage** | | |
| لكن/*lkn/but* | 204 (+3 not DC) | 99% |
| عندما/*EndmA/when* | 54 (+1 not DC) | 98% |
| اثر/*Avr/after* | 67 (+2 not DC) | 97% |
| في حين/*fy Hyn/while* | 27 (+1 not DC) | 96% |
| بسبب/*bsbb/because of* | 49 (+3 not DC) | 94% |
| بل/*bl/but* | 15 (+1 not DC) | 94% |
| بالتالي/*bAltAly/consequently* | 14 (+1 not DC) | 93% |
| اما/*AmA/while* | 24 (+2 not DC) | 92% |

The following list shows the most frequent (potential) discourse connectives and how often they have discourse function in context: و/*w/and* (54%), ل/*l/for* (11%), لكن/*lkn/but* (99%), بعد/*bEd/after* (62%), خلال/*xlAl/during* (81%), ف/*f/then* (6%), ب/*b/by* (2%), قبل/*qbl/before* (52%), لأن/*l>n/because* (73%), منذ/*mn*/since* (31%), كما/*kmA/as* (66%), اثر/*Avr/after* (97%), عندما/*EndmA/when* (98%), بسبب/*bsbb/because of* (94%), الا ان/*AlA An/but* (100%), فيما/*fy mA/while* (88%), ثم/*vm/then* (75%), او/*Aw/or* (38%), في حال/*fy HAl/in case* (83%), اذا/*A*A/if* (69%), حيث/*Hyv/where/since* (33%) and رغم/*rgm/though* (82%). Apart from لكن/*lkn/but* and الا ان/*AlA An/but,* these frequent connectives are ambiguous in terms of discourse usage, with several being highly ambiguous.

The clitics ب/*b/by*, ف/*f/then* and ل/*l/for* in addition to coordinating conjunctions such as و/*w/and*, او/*Aw/or* and كما/*kmA/as* are the most ambiguous discourse connectives (see Table 8-2). Some of them are mostly not discourse connectives, the potential connective ب/*b/by* is a discourse connective only (2%), and ف/*f/then* is a discourse connective only (6%) of the times they appear in the LADTB. The potential clitic connectives often occur as original parts of words, not as real clitics or connectives. For instance, the connective لان/*l>n/because* which at first sight always has discourse usage, is a discourse connective only 73% of the time. As an example, the first three letters (لان/*lAn/because*) form neither a connective nor a clitic in (لانهاء/*lAnhA/for finishing*).

**Table 8-2: A list of the most ambiguous, potential discourse connective types with regard to disourse function. The first two connectives are almost do not have discourse function.**

| Conn | Freq | % Dis.Conn |
|---|---|---|
| ب /b/by | 4168 | 2% |
| ف/*f/then* | 1525 | 6% |
| ل/*l/for* | 4306 | 11% |
| ايضا/*AyDA/also* | 102 | 17% |
| منذ/*mn*/since* | 220 | 31% |
| حيث/*Hyv/where/since* | 96 | 33% |
| او/*Aw/or* | 93 | 38% |
| قبل/*qbl/before* | 161 | 52% |
| و/*w/and* | 7375 | 54% |
| بعد/*bEd/after* | 315 | 62% |
| كما/*kmA/as* | 105 | 66% |
| لأن/*l>n/because* | 106 | 73% |
| خلال/*xlAl/during* | 126 | 81% |

## 8.3 Data Used in Experiments

Our experiments in discourse modeling use the data of all LADTB files (534) for training and testing with 20,312 potential discourse connective tokens and 6,328 real discourse connective tokens. A potential discourse connective is any string in our discourse connective list independent of its ATB annotation. Refer to Section 6.1 for a description of how we identify the potential discourse connectives in our annoatation of the raw texts in the LADTB. We called this overall dataset, Set 1. However, we noticed that there are some duplicated discourse connective tokens in Set 1. These repetitions result from (i) there being 4 texts entirely duplicated in the ATB Part1, and therefore in the LADTB too, (ii) some news are repeated in which the reporter reused the same sentences/arguments in different article. Thus, it is worth to examine the effect of those repetitions in our experiments by removing all repetitions from the training/testing dataset, Set 2, leaving 18,798 potential connectives tokens and 5,880 real discourse connective tokens.

For modeling discourse relation recognition, we examined the effect on single relations only in Set 1 (6039) and Set 2 (5880). Also, similar models were examined on the same two datasets after excluding the most frequently used connective

*w/and* at BOP, the majority of whose occurrences are assigned automatically to the Conjunction relation in the LADTB (see Section 7.6).

## 8.4 Automatic Recognition of Discourse Connectives

The task of the models here is to distinguish discourse vs. non-discourse usage for the potential connectives in datasets Set 1 and Set 2. Different types of features were used in our models in order to achieve a high performance. The features were extracted from different annotations of the texts. In the remaining parts of this section, we describe the features, the experimental setup and our analysis of the results and errors of the best model.

### 8.4.1 Features

Some prior work in English discourse modeling has ignored surface strings that are too ambiguous with regard to discourse usage (Marcu 2000c). However, recent work (Pitler and Nenkova 2009) used gold standard syntactic features as well as the connective surface string in a supervised model for discourse connective recognition in English. They achieved very high results with this approach: accuracy 91.1% and F-score 86.4 on the English PDTB. For further discussion of related work we refer the reader to Section 2.8.2.1. We will (i) show that similar features work well for Arabic, (ii) take into account Arabic-specific morphological properties that improve results further, and (iii) present a robust version of this approach that does not rely on full parsing or gold standard syntactic annotations and still has good results.

We include surface based, lexical and syntactic features in our models; the syntactic features (Syn) are inspired by (Pitler and Nenkova, 2009) and (Dipper and Stede 2006). However, Lexical/POS patterns of surrounding words, the clitic features and a morphological feature that captures whether the next noun is an *al-maSdar* or not, are novel in our study. Features are either extracted from raw files tokenized by white space only (M2) and tagged by the Stanford tagger[22] (Models M3, M4) or from

[22] The Stanford tagger is currently the only freely available tagger for Arabic; however, it requires ATB tokenization. http://nlp.stanford.edu/software/tagger.shtml

the Arabic Treebank (ATB) gold standard part-of-speech and syntactic annotation (models M5-M9).

Apart from the surface string of the potential connective, we use the following features:

**Surface Features of the Potential Connective (SConn).** These include the position of the potential connective (sentence-initial, medial or final). We also specify the type of the potential connective; it is SIMPLE when the potential connective is a single token not attached to other tokens, CLITIC when it is attached. Models where we use ATB or automated tagging (M3-M9) distinguish further between potential clitics that are assigned a POS and ones that are not (original part of a word in the raw text). Potential connectives containing more than one token have MoreThanToken type. Models that use ATB annotation also distinguish between potential connectives that correspond to a phrase in the ATB {MoreThanTokenPhrase} and the ones that do not {MoreThanTokenNonPhrase}.

**Lexical features of surrounding words (Lex).** We encode the surface strings of the two words before and three words after the connective, recording position. These features are especially useful for languages where no accurate parser or tagger is available as lexical patterns can capture discourse and non-discourse usage. For instance, if a potential connective is followed by *ان/An/that*, it most likely has a discourse function, as in Ex. 8-1. Note here that the English translation does not show that the two clauses are complete sentences in Arabic.

**Ex. 8-1**

ان الأطفال يمكن ان يصابوا بالتعب **و ان يشعروا بالنعاس** خلال الدراسة اذا لم يناموا جيدا.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| An | Al>TfAl | ymkn | An | ySAbwA | bAltEb] | w | [An | y$ErwA | bAlnEAs] | xlAl |
| that | children | may | that | they-got | in-tired | and | that | they-feel | In-sleep | during |
| AldrAsp | A*A | lm | ynAmwA | jydA | | | | | | |
| study | if | not | they-sleep | well | | | | | | |

Children might *be tired* **and feel sleepy** during school time if they did not sleep well.

**Part of Speech features (POS).** We include the pos tag of the potential connective via the ATB/Stanford Tagger. For potential connectives that consist of more than one token, we combined its ordered POS tags. Thus, the potential connective *في حال/fy HAl/in case* with its tags (fy PREP, Hal NOUN) will receive the pos PREP#NOUN.

If a potential connective does not receive a separated POS tag in the ATB standard tagger, the value 'NONE' is assigned. This allows clitics to be distinguished from letters at the start of a word.

To tackle problems when not having proper syntactic phrases in the ATB for connectives of more than token, we use a combination of POS of leaf nodes. For example, the potential connective في حال/*fy HAl/in case* is a prepositional phrase, but it has two different syntactic analyses in the ATB: (i) as prepositional phrase PP ( (fy PREP, Hal NOUN)) and (ii) introducing a prepositional phrase PP ( (fy PREP,Hal NOUN) (NP)…). The connective category of both cases would be PREP#NOUN. They also have accordingly two different types, MoreThanTokenPhrase and MoreThanTokenNonPhrase respectively.

The potential clitics connectives were separated from the beginning of words when using the Stanford Tagger, as there is no automatic tokenization included in the tagger and there is no freely available ATB tokenization tool.

We also record the POS of the three words before and after the connective (when using ATB/Stanford Tagger). Similar to lexical patterns, these can capture discourse and non-discourse usage. For instance, if a potential connective is soon followed by a modal such as قد/*qd/may/had* in the first three words after the connective, it is more likely to have a discourse function.

**Syntactic category of related phrases (Syn)**. We record the syntactic category of the parent of the potential connective in the ATB. For example, it is rare that cases where the parent of the potential connective is an adjective phrase correspond to discourse-usage. A typical example of a non-discourse usage of و/*w/and* ( المدرسة كبيرة و جميلة/*the school is very large and beautiful*) illustrates this. Unlike English, parents of true discourse connectives in Arabic often are noun phrases as nominalizations are frequent arguments of prepositional connectives. We also encode the left sibling category (preceding token) and right sibling category (following token) of the connective. The left sibling might be the syntactic category of a word, a phrase or 'NONE' if the connective is the first substring inside its parent category. For discourse connectives, the right sibling is normally S, SBAR, VP or an NP (if the connective is a preposition).

**Morphological features: Al-maSdar.** Potential connectives followed by or attached to Al-maSdar are more likely to have discourse usage (see Section 5.4.1). For instance, preposition connectives are normally followed by (for example, بعد *عمل/after doing)* or attached at the beginning of an al-maSdar noun (for example, *بإجراء/by processing*). If the prepositions are *followed by /attached to the beginning* of non al-maSdar nouns, then they are very unlikely to have a discourse function. The reader can refer to Sections 3.1 and 5.4.1 for more justification.

Al-maSdar information is not included in the ATB nor in the automatic Stanford Tagger. Thus, we constructed a binary al-maSdar feature from (tagged) text by examining the first noun after the potential connective. We developed an algorithm to judge such a noun as al-maSdar or not. This algorithm consists of a pipeline of text processing steps using a plural/singular list *Lex* provided by (Sawalha and Atwell 2010)[23] and a list of al-maSdar morphological patterns *Mas* from a documentation of **Alkulil Morpho Sys** by KACST and ALECSO [24].

[23] We acknowledge our colleague Mr. Sawalha in Leeds for letting us use his unpublished lexicon in this research.

[24] 'برنامج الخليل الصرفي' is the most comprehensive open source morphological analyser and was developed in 2010 by KACST and ALECSO. The downloading page: http://www.econtent.org.sa/Projects/InitiativeProjects/Lists/InitiativeProjects/DispForm.aspx?ID=25.

Input: *N*: a noun with more than three letters, and *its ATB pos tag*.

*Lex*: A list of plural/singular nouns.

*Mas*: A list of al-maSdar patterns; see Appendix A.

Step1: Stemming:

Use *the ATB pos tag* and *Lex* to:

1. Discard the determiners from *N*, if any.
2. Convert *N* from potential plural into singular, if *N* is plural.

Step2: an ordered sequence of surface-based filters

**Filter 1:** Filter al-maSdar patterns in *Mas* to keep only patterns with the same length of *N*. Go to Filter 2.

**Filter 2:** If *N* starts/ends with the suffix *تاءT or ألف/alf (A)*, keep only the patterns in *Mas* that also start/end with the suffix *تاءT or ألف/alf (A)*. Go to Filter 3.

**Filter 3:** for each pattern *p* in *Mas*, match the letters at the same positions in *N* and *p*. Keep patterns with maximum number of matching letters.

Output:

*N* is al-maSdar noun if *Mas* has at least one pattern left. Otherwise, N is not al-maSdar noun.

**Figure 8-1: Pseudo-code of surface-based al-maSdar detection.**

The pseudo-code in Figure 8-1 shows this pipeline of different surface-based filters of Mas. For example, the Mas list is filtered at each stage as appropriate to examine a noun إدمان/addiction in Figure 8-2. The algorithm is designed to examine nouns with at least four letters. The 3-letter nouns should at least have diacritics for al-maSdar detection using this surface-based method. Alternatively, generating all potential al-maSdar nouns from the root of the noun and examining them for a match with current noun, is another advanced automatic solution. However, this is a separate sizable project by itself.

The automatic algorithm has been used to examine 5586 nouns that follow the potential connectives in Set1, and are more 3 letters long. after excluding 3-letter nouns (1020). The algorithm achieved 92% accuracy (5152 out of 5586 nouns), with 434 wrong detections (8 false negative and 425 false positive). In addition to the 434

wrong detections by the algorithm, we have 1020 nouns of 3-letters were all examined and assigned the al-maSdar feature value manually.

***N*:** إدمان/addiction

***Mas* list- initial**: تفعال، تفعلة، علة، فاعلة، فعال، فعالة، فعالية، فعل، فعلان، فعلة، فعلوت، فعلولة، فعلى، فعلي، فعلية، فعول، فعولة، فعولية، فعيل، فعيلة، فعيلى، فيلولة، مفعل، مفعلة، مفعول، إفعال، افعال، إفعلة، افعلة، تفعيل، مفاعلة، انفعال، افتعال، افعلال، تفاعل، تفعل، استفعال، استفعلة، افعيعال، افعوال، افعيلال، فعللة، فعلال، تفعلل، افعنلال

*Mas* list- after Filter 1: تفعال، تفعلة، فاعلة، فعالة، فعلان، فعلوت، فعلية، فعولة، فعيلة، فعيلى، مفعلة، مفعول، إفعال، افعال، إفعلة، افعلة، تفعيل، تفاعل، فعللة، فعلال، تفعلل،

*Mas* list- after Filter 2: إفعال، افعال، إفعلة، افعلة

*Mas* list- after Filter 3: إفعال

**Result**: there is one pattern left in *Mas*. *N* (إدمان/addiction) is al-maSdar noun.

**Figure 8-2: Step by step al-maSdar examination of the noun *إدمان/admAn/addiction.***

## 8.4.2 Experimental Setup

The implementation JRip of the rule-based classifier Ripper is used in our experiments using the machine learning tool WEKA (Witten *et al.* 1999) with its standard settings. The rule-based classifiers are helpful in determining which features are more useful than others for discourse connectives recognition because of their readable output. However, the Ripper classifier produces rules applied in order (first match is used). Instances that are matched by a rule are excluded from the testing dataset when considering the next rules. This might impact on the results when instances fit more than one rule.

For training and testing purposes, the positive examples are the explicit discourse connectives annotated in the LADTB (6,328) and the negative ones are the same strings that were annotated as non-discourse connectives (13,984) in Set 1 (20,312). We also repeated the experiments removing any repetitions in the data, Set 2 (18,798). The 10-fold cross-validation is used throughout and significance tests are reported using the McNemar test on accuracy at the significance level of 0.01. We have run two types of experiments on both Set 1 and Set 2: (A) Auto-tag models

where the features were extracted from simple and freely available white space tokenization and an automatic tagger (Stanford tagger) without any manual preprocessing, and (B) ATB-tag models where features were extracted from the gold-standard tokenization, tagging and parsing in the Arabic Treebank annotation.

### 8.4.3 Results and Evaluation

The results do not vary very much between Set 1 (Table 8-3) and Set 2 (Table 8-4), thus we discuss only the results on Set 1. A baseline of the most frequent category would assign all potential connectives as *not discourse connective*, achieving an accuracy of 68.9% on Set 1, as only 6,328 of our potential 20,312 connectives actually have discourse usage. The results of further advanced models using different features are shown in Table 8-3. We use accuracy and kappa measurements in the table. For further comparison studies with similar models, we also calculate recall, precision and F-score for positive class (discourse connective) for the models, using automatic tagging (M2-M4) and gold-standard tagging (M5-M10).

A connective specific majority class model M1 that only uses the connective string improves significantly over the baseline of majority class with 75.7% accuracy and F-score of 0.67 but a kappa of only 0.48 on Set 1, showing that using only the connective string is not a reliable strategy. M1 will be used as baseline for the other models. Models M2-M4 do not rely on gold standard annotation or parsing (in contrast to the models for English in (Pitler and Nenkova 2009)). Using only surface and lexical features that can be extracted from white-spaced tokenized raw files in addition to a tokenization for clitic connectives (M2), gains a substantial improvement over using the connective string alone. This is further improved by using POS tags of connectives and surrounding words with an automatic tagger (M3) and by including the al-maSdar feature (M4), thus making good use of the morphological properties of Arabic. All differences are statistically significant (M1 < M2 < M3 < M4). The final model is reliable (kappa 0.70), an encouraging result given the absence of parsing and important for resource-scarce languages.

The model M4 recorded a precision of 86%, a recall of 75% and F-score of 80% on Set1 for the positive class (discourse connective). Removing the repetitions (Set2)

causes only slight change in precision (87%), recall (74%), and F-score (80%), see Table 8-4.

**Table 8-3: Performance of diffrent models for discourse connective recognition on Set 1.**

| | **Features** | **Set 1 of all conn (20312)** | | | | |
|---|---|---|---|---|---|---|
| | | **Acc** | **K** | **Prec** | **Rec** | **F-** |
| | Baseline – not conn | 68.9 | 0 | 0 | 0 | 0 |
| M1 | Conn only | 75.7 | 0.48 | 0.58 | 0.79 | 0.67 |
| **Auto-tag models: White space tokenization + auto tagger-based features** | | | | | | |
| M2 | Conn+SConn+Lex | 85.6 | 0.62 | 0.88 | 0.60 | 0.71 |
| M3 | Conn+SConn+Lex+POS | 87.6 | 0.69 | | | |
| **M4** | **Conn+SConn+Lex+POS+MaSdar** | **88.5** | **0.70** | **0.86** | **0.75** | **0.80** |
| **ATB-tag models: ATB tokenization, tagging and parsing features** | | | | | | |
| M5 | Conn+SConn+Lex | 86.2 | 0.65 | 0.87 | 066 | 0.75 |
| M6 | Conn+SConn+Lex+POS | 88.2 | 0.71 | 0.88 | 0.72 | 0.80 |
| M7 | Conn+SConn+Lex+POS/Syn | 91.2 | 0.79 | 0.90 | 0.81 | 0.85 |
| **M8** | **Conn+SConn+Lex+POS/Syn+ MaSdar** | **92.4** | **0.82** | **0.90** | **0.85** | **0.87** |
| M9 | Conn+SConn+ POS/Syn | 91.2 | 0.79 | **0.90** | **0.81** | 0.85 |
| M10 | SConn+Lex+ POS/Syn +MaSdar | 91.2 | 0.80 | 0.90 | 0.82 | 0.86 |

With ATB gold standard tokenization, tagging and parsing (ATB-tag models) in Set 1, our models (not surprisingly) improve further showing the same pattern of (M1 (75.7%) < M5 (86.2%) < M6 (88.2%) < M7 (91.2%) < M8 (92.4%)) with all differences being significant. The final best model (M8) achieves highly reliable results (accuracy 92.4% and kappa 0.82). It also records precision 90%, recall 85%, F-score 87% for positive class (discourse connective). Removing the repetitions (Set 2, Table 8-4) increases precision to 90%, recall to 90%, F-score to 87% for positive class of the same model (M8). This means that M8 classified more true positive connectives in Set 2 than in Set 1.

**Table 8-4: Performance of diffrent models for discourse connective recognition excluding repetitions (Set 2).**

| | **Features** | **Set 2 excluding repetitions (18798)** | | | | |
|---|---|---|---|---|---|---|
| | | **Acc** | **K** | **Pre** | **Rec** | **F-** |
| | Baseline – not conn | 68.8 | 0 | 0 | 0 | 0 |
| M1 | Conn only | 75 | 0.47 | 0.59 | 0.79 | 0.67 |
| **Auto-tag models: White space tokenization + auto tagger-based features** | | | | | | |
| M2 | Conn+SConn+Lex | 84.2 | 0.60 | 0.89 | 0.58 | 0.70 |
| M3 | Conn+SConn+Lex+POS | 86.4 | 0.67 | 0.86 | 0.68 | 0.76 |
| **M4** | **Conn+SConn+Lex+POS+MaSdar** | **88.6** | **0.73** | **0.87** | **0.74** | **0.80** |
| **ATB-tag models: ATB tokenization, tagging and parsing features** | | | | | | |
| M5 | Conn+SConn+Lex | 83.1 | 0.60 | 0.98 | 0.48 | 0.65 |
| M6 | Conn+SConn+Lex+POS | | | | | |
| M7 | Conn+SConn+Lex+POS/Syn | 90.6 | 0.78 | 0.90 | 0.81 | 0.85 |
| **M8** | **Conn+SConn+Lex+POS/Syn+ MaSdar** | **92.3** | **0.82** | **0.90** | **0.90** | **0.87** |
| M9 | Conn+SConn+ POS/Syn | 92.2 | 0.82 | 0.90 | 0.80 | 0.85 |
| M10 | SConn+Lex+ POS/Syn +MaSdar | 91.5 | 0.80 | 0.90 | 0.82 | 0.82 |

We also conclude that syntactic features are more useful than lexical patterns as model M9 (syntax with no lexical patterns) achieves equally good results as M7. However, lexical patterns are useful if syntactic features are not available. Note that removal of repetitions leads to decreased performance by models M5, M6 and M7 that use lexical patterns. This is because including lexical features leads to overfitting data which is not the case when we exclude the repetitions. In contrast, slight improvements in performance were recorded, when we exclude the repetitions, for models that do not use lexical patterns features such as M9.

Our models also manage to generalize well over individual connectives. If we leave out the connective string (M10), we still achieve a highly reliable result.

### 8.4.4 Error Analysis and Discussion

The focus of our analysis will be on the best model M8 on all connective tokens, Set 1. There are two main reasons for the improvement in results of M8 over the model

M1, which uses the connective string only: (i) generalization and (ii) disambiguating ambiguous connectives.

## Generalization

The model M8 succeeds in identifying 28% of the instances of true discourse connectives (1800 out of 6,328) without using the connective string; recording by that a good performance using only generalized rules. The general rules with accuracy of each rule are highlighted in Table 8-5. The rules are given in the same order as output by the classifier. For example, 87% of 481 tokens that have Simple preposition connectives and are followed by al-maSdar noun are discourse connectives regardless of what the connective strings are. Also, 23 out of 25 tokens are discourse connectives when the connective is Simple, at the middle of the sentence, and attached to a clause not starting with al-maSdar noun. Note that the classifer orders the rules according to which rule covering as many positive instances as possible, while covering as few negative instances as possible.

Al-maSdar, POS features and connective's parent category are the most used features in the generalized model. General rules can handle data with previous unseen potential connectives.

**Table 8-5: The ordered rules used in recognizing discourse connectives (M8). The highlighted rules do not use the connective string (general rules).**

| Rules | Total match | Correctness | Acc |
|---|---|---|---|
| (Parent_cat = S) and (Conn = w) = | 3309 | 3229 | 98% |
| (conn_type = Simple) and (Isalmasdar_w_after_conn = Yes_masdar) and (Conn_pos = PREP) = | 481 | 419 | 87% |
| (Conn_pos = CONJ) and (Parent_cat = S) and (Conn = lkn) = | 187 | 186 | 99% |
| (Conn_pos = CONJ) and (Left_sib = NONE) and (Parent_cat = SBAR) = | 259 | 231 | 89% |
| (Conn_pos = CONJ) and (Parent_cat = VP) = | 195 | 171 | 88% |
| (Conn_pos = CONJ) and (Right_sib = S) = | 153 | 114 | 75% |
| (Isalmasdar_w_after_conn = Yes_masdar) and (Parent_cat = NP) and (Second_w_after_conn_pos = NOUN) and (Parent_left_sib = PREP) and (Conn_pos = PREP#NOUN) = | 42 | 38 | 90% |
| (Parent_cat = SBAR) and (Isalmasdar_w_after_conn = Not_masdar) and (conn_type = MoreThanToken_NonPhrase) = | 95 | 88 | 93% |
| (Isalmasdar_w_after_conn = Yes_masdar) and (Parent_left_sib = PP) and (Conn = l) = | 163 | 128 | 79% |

| Rules | Total match | Correctness | Acc |
|---|---|---|---|
| (Isalmasdar_w_after_conn = Yes_masdar) and (Conn_pos = CONJ) and (Second_w_after_conn_pos = PREP) and (Right_sib = NP) = | 77 | 58 | 75% |
| (Isalmasdar_w_after_conn = Not_masdar) and (Parent_cat = SBAR) and (Conn = w) = | 102 | 89 | 87% |
| (conn_type = Simple) and (Parent_cat = SBAR) = | 202 | 131 | 65% |
| (Second_w_after_conn_pos = NOUN) and (Conn_pos = CONJ) and (Isalmasdar_w_after_conn = Yes_masdar) and (Parent_right_sib = NONE) = | 253 | 145 | 57% |
| (Isalmasdar_w_after_conn = Not_masdar) and (Right_sib = S) = | 91 | 81 | 89% |
| (Isalmasdar_w_after_conn = Yes_masdar) and (Second_w_after_conn_pos = POSSuPRON) and (Third_w_after_conn = ElY) = | 39 | 34 | 87% |
| (Second_w_after_conn_pos = NOUN) and (conn_type = MoreThanToken_NonPhrase) = | 56 | 44 | 79% |
| (Isalmasdar_w_after_conn = Not_masdar) and (Parent_cat = S) and (Left_sib = NONE) = | 139 | 123 | 88% |
| (conn_type = Simple) and (Parent_cat = PP) and (Conn = xlAl) = | 53 | 35 | 66% |
| (Isalmasdar_w_after_conn = Yes_masdar) and (Second_w_after_conn_pos = POSSuPRON) and (Right_sib = S) = | 14 | 12 | 86% |
| (Isalmasdar_w_after_conn = Not_masdar) and (conn_type = Simple) and (Right_sib = SBAR) and (Left_sib = NONE) and (conn_position_hostingS = MED) = | 25 | 23 | 92% |
| (Isalmasdar_w_after_conn = Not_masdar) and (Next_w_after_conn_pos = PREP) and (Conn_pos = CONJ) = | 152 | 88 | 58% |
| Classified by rules | 6087 | 5226 | 86% |
| Classified as not Dis. Conn (default value) | 14225 | 13364 | 94% |
| Total | 20312 | 18590 | 92.4% |

## Unambiguous Connectives: Discourse Usage

Only 850 (4%) instances of Set 1 belong to connectives that are unambiguous in discourse usage (see Section 8.2). Theoretically, these should be identified by the connective string alone (model M1). However, many of these are so rare that they appear only in the training or only the test data, making recognition by M1 impossible. Ripper will also want to create robust rules with good coverage and might judge a connective-string-only rule that holds for few instances worse than applying the default value that assigns not-a-connective to any instance.

Table 8-6 shows a table comparing M8 and M1; a total of 61 instances are not classified correctly using either the connective string or any further features in M8. This includes 9 very rare unambiguous connectives such as حال/*HAl/when* (2), حتى لو/*HtY lw/even if* (2), وقبل/*wqbl/and before* (1), بيد/*byd/but* (1), بحيث/*bHyv/since* (1) and خلافا لـ/*xlAfA l/unlike* (1). However, those results would most likely be improved with more annotated instances of such rare connectives in our corpus. In addition, the order of the rules generated by M8 incorrectly changes the results of 22 instances which are classified correctly by M1. In these cases, generalized rules fire before connective-specific rules.

**Table 8-6: The comparison matrix of the rich features model M8 and the baseline M1 for unambiguous connectives**

| M8-classifier / ConnOnly-classifier (M1) | Correct | Incorrect | Total |
|---|---|---|---|
| **Correct** | 629 | 22 | 651 |
| **Incorrect** | 138 | 61 | 199 |
| **Total** | 767 | 83 | 850 (4%) |

The generalization rules successfully identified 138 instances of 18 rarely occurring unambiguous connectives such as عقب/*Eqb/shortly after* (5), طالما/*TAlmA/as long as* (4), الا اذا/*AlA A*A/except if* (2), بيد ان/*byd An/but* (6), رغم ان/*rgm An/although* (6), غير ان/*gyr An/however* (6), بفضل/*bfDl/thank to* (5), قبيل/*qbyl/shortly before* (5), قبل ان/*qbl An/before that* (3), كلما/*klmA/when ever* (1), لولا/*lwlA/if not* (1).

The connective string alone is a sufficient feature for 629 instances of 10 unambiguous connectives in discourse usage: عندما/*EndmA/when* (55), لكن/*lkn/but (207)*, الا ان/*AlA An/but (41)*, بالإضافة/*b AlADAfp/in addition to (10)*, كان/*kAn/as (316)*, عموما/*EmwmA/generally (2)*, فعلا/*fElA/in deed (7)*, علاوة على/*ElAwp ElY/in addition (2)*, في الواقع/*fy AlwAqE/actually (2)* and بعد ذلك/*bEd *lk/after that (4)*. This advantage might be lost when a larger corpus is used; these connectives are unambiguous only in our data but they might be ambiguous if more instances were included.

## Ambiguous Connectives: Discourse Usage

The majority of our training and testing dataset Set 1 are tokens of 44 potential connectives which have different degrees of ambiguity in discourse usage (19462 out

of 20312, 96%). Table 8-7 shows the comparison of M8 and M1 (using the connective string alone) for these ambiguous connectives. 72% of the ambiguous connective tokens in Set 1 are classified correctly by both models (14114); the majority of them are not a discurse connective. In contrast, both models failed to classify correctly a set of 920 instances of potential connectives of ambiguous connective types, representing 5% of the ambiguous connectives in Set 1. The most frequent connective types that have more than 20 incorrectly classified tokens by both models are, in descending order, *و/w/and* (291), *ل/l/for* (276), *ب/b/by* (66), *كما/kmA/as* (36), *خلال/xlAl/during* (23) and *او/Aw/or* (20).

**Table 8-7: The comparison matrix of the rich features model M8 and the baseline M1 for connectives not always having discourse usage.**

| M8-classifier / ConnOnly-classifier (M1) | correct | incorrect | Total |
|---|---|---|---|
| **Correct** | 14114 | 572 | 14686 |
| **Incorrect** | 3856 | 920 | 4776 |
| **Total** | 17970 | 1492 | 19462 |

A set of 12 ambiguous connectives types, a total of 572 instances (3%), has a worse classification in M8 than using the majority class per connective (M1). This set involves the connectives *كما/kmA/as*, *حيث/Hyv/where/since*, *ب/b/by*, *اثر/Avr/after*, *ايضا/AyDA/also*, *الا/AlA/except*, *بسبب/bsbb/because of*, *رغم/rgm/though*, *بالاضافة الى/bAl ADAfp AlY/in addition to*, *اضافة الى/ADAfp AlY/in addition to*, *حين/Hyn/when*, and *لذلك/l*lk/for that*. This result might be improved using different classifiers, as in these cases Ripper's ordering play a decisive role. We leave the testing of this hypothesis to future work. See Section 9.3 for more suggestions for future work.

On the other hand, M8 gained an advantage on 3856 instances of 24 ambiguous connective types over using the majority class for each connective (M1). This set represents 20% of ambiguous connectives in Set 1 and were mostly recognized using only generalized rules, without using the connective string (the highlighted rules in Table 8-5). Table 8-8 also lists some of those connectives ordered according to how much they improved in M8 using the generalized rules. Interestingly, different generalized rules can be used to recognize instances of a particular connective. For example, the potential connective *قبل/qbl/before* (161; 77 Non-DisConn and 84

DisConn) is a discourse connective when one of the three rules is applied: (i) when it is followed by al-maSdar (43), (ii) when the parent category is SBAR (20) or (iii) when the word after the connective is not al-maSdar, and the left sibling is NONE but the right sibling is SBAR (10).

**Table 8-8: A list of ambiguous connectives which are improved using generalized rules using the full ATB-features model (M8).**

| Conn | Freq | Accuracy of ConnOnly | Accuracy of M8 |
|---|---|---|---|
| لكي/*lky/for/in order to* | 6 | 17% | 100% |
| كي/*ky/to* | 3 | 33% | 100% |
| في حين/*fy Hyn/while* | 28 | 18% | 68% |
| قبل/*qbl/before* | 161 | 48% | 89% |
| بهدف/*bhdf/in order to* | 27 | 44% | 85% |
| و/*w/and* | 7375 | 54% | 93% |
| الا بعد/*AlAbEd/except after* | 6 | 17% | 50% |
| انما/*AnmA/but* | 10 | 30% | 60% |
| اذا/*A*A/if* | 49 | 41% | 69% |
| اما/*AmA/while* | 26 | 8% | 35% |
| بالتالي/*bAltAly/consequently* | 15 | 7% | 33% |
| بعد/*bEd/after* | 315 | 62% | 87% |
| لأن/*l>n/because* | 109 | 73% | 98% |
| منذ/*mn*/since* | 220 | 69% | 89% |
| حتى/*HtY/until* | 75 | 61% | 76% |
| فضلا عن/*fDlA En/as well as* | 14 | 57% | 71% |
| ثم/*vm/then* | 48 | 58% | 71% |
| فيما/fy mA/while | 41 | 83% | 93% |
| لو/*lw/if (in the past)* | 14 | 57% | 64% |
| ف/*f/then* | 1525 | 94% | 98% |
| خصوصا/*xSwSA/specially* | 64 | 64% | 69% |
| او/*Aw/or* | 93 | 62% | 67% |
| ل/*l/for* | 4306 | 89% | 93% |

We found a few incorrect classifications which are results of wrong annotation in the LADTB. For example, there are 4 instances of the connective خلال/*xlAl/during* which were annotated as non-discourse connectives though in fact they relate valid abstract objects such as التزام/*commitment* in Ex. 8-2 (a), and الجولة/*tour* in Ex. 8-2 (b). These nouns are al-maSdar but they were missed in the LADTB annotation by both annotators. Thus, they were also not verified in the post-process.

**Ex. 8-2**

(a)

اطلق الجيش النار على المتظاهرين **خلال التزام الهدنة بين الجانبين**

| ATlq | Aljy$ | AlnAr | ElY | AlmtZAhryn | xlAl | AltzAm | Alhdnp | byn | AljAnbyn |
|---|---|---|---|---|---|---|---|---|---|
| hold | army | fire | on | demonstrators | during | commitment | truce | between | Two-sides |

*Army opened fire on the demonstrators* during **the commitment of a truce between the two sides**

(b)

زار الرئيس الأمريكي جورج بوش العراق **خلال الجولة الشرق أوسطية** ليقابل رئيس الحكومة المؤقتة

| zAr | Alr}ys | Al>mryky | jwrj | bw$ | AlErAq | xlAl | Aljwlp | Al$rq |
|---|---|---|---|---|---|---|---|---|
| visit | President | American | George | Bush | Iraq | during | the-tour | East |
| >wsTyp | lyqAbl | r}ys | AlHkwmp | Alm&qtp | | | | |
| middle | to-meet | head | government | temporary | | | | |

*The U.S. President George W. Bush visited Iraq,* **during the tour in the Middle East**, *to meet the President of the interim government.*

**Table 8-9: The ordered rules used in recognizing discourse connectives (M4) on Set 1. The highlighted rules do not use the connective string (general rules).**

| The rule | Total | Correctness | Acc |
|---|---|---|---|
| (conn_status = BOS) and (Conn = w) = | 2470 | 2439 | 99% |
| (conn_type = Simple) and (Conn_pos = NN) and (Isalmasdar_w_after_conn = Yes_masdar) = | 216 | 195 | 90% |
| (conn_type = Simple) and (Conn = lkn) = | 205 | 202 | 99% |
| (conn_type = Simple) and (Conn_pos = NN) and (First_w_raw_tag = NN) = | 221 | 156 | 71% |
| (conn_type = Simple) and (First_w_raw_tag = VBD) and (Conn_pos = IN) = | 92 | 70 | 76% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (w_before_conn_raw_tag = DTNN) = | 269 | 215 | 80% |
| (conn_type = Simple) and (Conn_pos = NNP) = | 281 | 205 | 73% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (First_w_raw_tag = VBP) = | 164 | 134 | 82% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (w_before_conn_raw_tag = NNP) = | 250 | 158 | 63% |
| (conn_type = Simple) and (First_w_raw_tag = VBD) and (Conn = EndmA) = | 44 | 43 | 98% |
| (conn_type = Simple) and (First_w_raw_tag = VBP) and (Conn_pos = IN) = | 27 | 26 | 96% |
| (conn_type = Simple) and (Conn_pos = CC) and (Isalmasdar_w_after_conn = NONE) = | 166 | 119 | 72% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (First_w_raw_tag = VBD) and (w_before_conn_raw_tag = DTJJ) = | 68 | 56 | 82% |
| (conn_type = Simple) and (Conn_pos = NN) and | 104 | 85 | 82% |

| The rule | Total | Correctness | Acc |
|---|---|---|---|
| (First_w_raw_tag = VBP) = | | | |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (First_w_raw_tag = RP) = | 50 | 40 | 80% |
| (Conn = w) and (Isalmasdar_w_after_conn = Yes_masdar) and (First_w_raw_tag = NN) and (Third_w_raw_tag = NN) = | 62 | 50 | 81% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (First_w_raw_tag = VBD) and (w_before_conn_raw_tag = DTNNP) = | 32 | 23 | 72% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (First_w_raw_tag = VBD) and (w_before_conn_raw_tag = JJ) = | 48 | 35 | 73% |
| (conn_type = Simple) and (conn_status = MOS) and (First_w_raw_tag = VBD) and (Third_w_raw_tag = NN) = | 30 | 22 | 73% |
| (Conn = w) and (w_before_conn_raw_tag = DTNN) and (First_w_raw_tag = NN) and (Isalmasdar_w_after_conn = Yes_masdar) = | 37 | 26 | 70% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (First_w_raw_tag = DT) = | 27 | 23 | 85% |
| (conn_type = Simple) and (w_before_conn_raw_tag = NN) and (Conn_pos = NN) = | 38 | 26 | 68% |
| (conn_type = Simple) and (Conn_pos = RP) and (Conn = AmA) = | 26 | 24 | 92% |
| (conn_type = MoreThanToken) = | 252 | 188 | 75% |
| (conn_type = Simple) and (conn_status = MOS) and (Isalmasdar_w_after_conn = Yes_masdar) and (Conn = mn*) = | 21 | 21 | 100% |
| (Conn = w) and (Sec_w_raw_tag = IN) and (First_w_raw_tag = DTNN) = | 77 | 45 | 58% |
| (conn_type = Simple) and (conn_status = MOS) and (First_w_raw_tag = VBP) and (Conn_pos = RP) = | 13 | 13 | 100% |
| (Conn = w) and (Sec_w_raw_tag = NNP) and (First_w_raw_tag = DTNN) = | 34 | 33 | 97% |
| (Conn = w) and (First_w_raw_tag = VBD) and (w_before_conn_raw2_tag = CD) = | 32 | 21 | 66% |
| (Conn = w) and (Isalmasdar_w_after_conn = NONE) and (w_before_conn_raw_tag = DTNNS) = | 26 | 18 | 69% |
| **Classified as discourse connective by rules** | 5382 | 4711 | 88% |
| **Classified as not discourse connective (default rule)** | 14930 | 13345 | 89% |
| **Total** | **20312** | | |

**Discussion of M4**

We have not conducted a complete error analysis for model M4 because we did not have access to an ATB-style automatic tokenization that is needed for the Stanford tagger[25]. Therefore, the POS features are less reliable than we would expect when using an automatic tagger. Apart from error chaining due to error in automatic tagging, M4 also has access to less syntactic information than M8 as parent and sibling categories are not known (M4 does not have access to parse tree). Therefore, M4 used fewer generalized rules than M8 as shown in Table 8-9. Note that the classifer orders the rules according to which rule covering as many positive instances as possible, while covering as few negative instances as possible. We discuss in the future work section in Chapter 9 that using proper tokenization will definitely improve the performance further.

## 8.5 Sense Ambiguity of Discourse Connectives

We investigate the ambiguity of Arabic discourse connectives with regard to their sense at class level (4 main relations) as well as the more fine-grained level (17 relations). Of 80 connective types, 52 are unambiguous at the class level and 45 at the fine-grained level: *خلال/xlAl/during,* *قبل/qbl/before,* *لأن/l>n/because,* *بسبب/bsbb/because of,* *في حال/fy HAl/in case,* *ثم/vm/then,* *رغم/rgm/though,* *مما/mmA/which lead a result of which,* *يهدف/bhdf/in order to,* *جراء/jra/because,* *على الرغم/ElY Alrgm/although,* *نظرا ل/nZrA l/because of,* *بعدما/bEdmA/after that,* *بيد ان/byd An/but,* *فضلا عن/fDlA En/as well as,* *غير ان/gyr An/however,* *كذلك/k*lk/and that,* *رغم ان/rgm An/although,* *بالرغم من/bAlrgm mn/although,* *بفضل/bfDl/thank to,* *بغية/bgyp/desire/to,* *في المقابل/fyAl mqAbl/in contrast,* *لكي/lky/for/in order to,* *قبيل/qbyl/shortly before,* *بالاضافة الى/bAl ADAfp AlY/in addition to,* *لأن/l>n/because,* *قبل ان/qbl An/before that,* *الا/AlA/except,* *عقب/Eqb/shortly after,* *حتى لو/HtY lw/even if,* *حينها/HynhA/when that,* *طالما/TAlmA/as long as,* *اضافة الى/ADAfp AlY/in addition to,* *كي/ky/to,* *بمعنى آخر/bmEnY xr/in other words,* *بيد/byd/but ,* *برغم/brgm/although,* *بحيث/bHyv/since,* *حين/Hyn/when,* *كأن/k>n/as,* *لذا/l*A/for this,* *لذلك/l*lk/for that,* *لولا/lwlA/if not,* *وقبل/wqbl/and before* and *خلافا ل/xlAfA l/unlike* (see Appendix D).

[25] The only available ATB tokenization tool is TOKEN which is included in a BAMA package, the Arabic syntactic analyser via the LDC. We were unable to get the package by the study time.

However, they account an only 574 of 6,328 (9%) discourse connective tokens. Thus, many of the most frequent connectives are highly ambiguous at class level and at the fine-grained level. Table 8-10 contains the most ambiguous connectives (in terms of how many relations they can signal) and specifies how often they occur with their predominant relations.

**Table 8-10: A list of the most ambiguous connectives in terms of how many single, fine-grained relations they signal in the LADTB. The full distribution is presented in Appendix C which also shows multiple relations.**

| Connective | Most frequent relations | #Sing.Rel |
|---|---|---|
| و*/w/and* | EXPANSION.Conjunction (3068, 77.5%), EXPANSION.Reformulation (287, 7.2%) CONTINGENCY.Cause.Result.NonPragmatic (134, 3.4%), EXPANSION.Background (183, 4.6%) | 14 |
| ف*/f/then* | CONTINGENCY.Cause.Result.NonPragmatic (29, 30.2%), CONTINGENCY.Cause.Reason.NonPragmatic (20, 20.8%), EXPANSION.Reformulation (18, 18.8%), EXPANSION.Exemplification (12, 12.5%), CONTINGENCY.Cause.Reason.Pragmatic (2, 6.7%) | 10 |
| حتى*/HtY/until* | COMPARISON.Contrast (6, 27.3%), CONTINGENCY.Cause.Reason.NonPragmatic (6, 27.3%), CONTINGENCY.Cause.Result.NonPragmatic (3, 13.6%), CONTINGENCY.PragmaticCondition (2 , 9.1%) | 8 |
| كما*/kmA/as* | EXPANSION.Conjunction (40, 61.5%), COMPARISON.Similarity (9, 13.8%) | 7 |
| منذ*/mn*/since* | TEMPORAL.Asynchronous (48, 70%), TEMPORAL.Synchronous (11, 16%) | 2 |
| اثر*/Avr/after* | TEMPORAL.Asynchronous (9, 50%), CONTINGENCY.Cause.Reason.NonPragmatic (9, 50%) | 2 |
| او*/Aw/or* | EXPANSION.Alternative.Conjunctive (28, 80%), EXPANSION.Alternative.Disjunctive (7, 20%) | 2 |

## 8.6 Recognition of Discourse Relations

Our discourse model disambiguated between 17 single relations for connective instances in the LADTB. Multiple relations are excluded from this study as they have few instances in the LADTB. We carried out the experiments on discourse connectives of the same datasets Set 1 and Set 2 (see data setting in Section 8.3). The

total of single relations in Set 1 is 6039 tokens and 5880 in Set 2 (without repetitions). In addition, the best models were run also on the same datasets but excluding tokens of *و/w/and* at BOP, leaving 3813 token in Set 1, and 3731 in Set 2. The reason behind these experiments is the fact that not all instances of *و/w/and* at BOP had proper human annotation in the LADTB, as a set of them were assigned the Conjunction relation automatically (see Section 7.3). The term *ambiguous connectives*, in this section, refers to discourse connectives that have more than one sense in discourse.

If we just assign the most frequent connective-specific reading to each of the 3813 connectives in Set 1 excluding *و/w/and* at BOP, we achieve an accuracy of 82.7% at the class-level and 74.3% at the more fine-grained level for relation assignment, leaving a substantial margin of error. This contrasts with the English PDTB, where at the class-level 92% can be achieved with this simple method and 85% at the second-level[26]. This shows the challenge of disambiguating explicit discourse connectives in Arabic. We assume in this task that the arguments of the connective are known, something which is well-established also for PDTB relation recognition (Wang, Su and Tan 2010; Lin, Kan and Ng 2009; Miltsakaki *et al.* 2005b).

Our models are the first algorithms to recognise Arabic discourse relations. We take into account Arabic specific features, in addition to features used in prior work for English. In the following sections, we describe our features regarding explicit connectives and their arguments for identifying the relations. We discuss the experimental setting as well as the results of our models with an intensive error analysis.

### 8.6.1 Features

Prior works in automatic disambiguation of explicitly signaled relations in English achieved good results using simple features (Pitler et al., 2008). A more comprehensive study on discourse connectives in the PDTB (Pitler *et al.* 2008; Pitler and Nenkova 2009) reveals that most connectives are *not ambiguous* in English, at least at the class level. Using syntactic features of the connective, they achieve only a

[26] The second level in the PDTB with its 16 relations corresponds roughly to our fine-grained inventory. This comparison can only be approximate due to slight differences in the lower-grained relation inventory.

very small improvement over a *most frequent relation per connective* baseline for which significance tests are not given[27]. However, a task specific study (Miltsakaki et al. 2005) concentrates on disambiguating only three connectives {since, while, when}, using a very small set of features indicating tense and temporal markers in arguments. They achieve good improvements over a *most frequent relation per connective* baseline. However, the case is different for Arabic where high ambiguity levels are recorded for discourse connective interpretation (see Section 8.5).

We build useful features used in prior work for disambiguating explicit connectives and recognizing implicit relations in English (Lin, Kan and Ng 2009; Wang, Su and Tan 2010; Pitler, Louis and Nenkova 2009). Some of these features are not widely used for automatic explicit connective interpretation and they are all novel for Arabic. In addition, we use novel Arabic specific features in our models. We mainly extracted the features from the ATB gold standard parses, and they involve:

**Connective features**. This includes the surface connective features and POS tag of the connective described in Section 8.4.1, in addition to the connective string, *Conn*. We also use the syntactic path to the connective which is a novel feature for explicit connective disambiguation.

**Words and POS of arguments.** The words and pos tags of the first three words in Arg1 and Arg2 are used to catch patterns in arguments. These features are novel for recognising explicit relations. For example, when the first word of Arg2 is *قد/qd/might-was* or *كان/kAn/had-was* which are often used to express a proposition in the past, the relation is likely to be EXPANSION.Background or EXPANSION.Conjunction (see Ex. 8-3). Out of 336 instances that their first word is *قد/qd/might-was* or *كان/kAn/had-was* in Set 1, there are 291 instances of EXPANSION.Background or EXPANSION.Conjunction. If the arguments are very short, the value NONE might be used. We also measure word overlap between the arguments, hoping to catch relations such as COMPARISON.Similarity.

**Ex. 8-3 (Rel: EXPANSION.Background)**

ان الطائرة التي تقل الوفد اللبناني الرسمي وصلت اليوم الثلاثاء الى طرابلس. **وكان قد اتى الوفد لاصطحاب الرهينة اللبنانية ماري ميشال معربس المحتجزة في الفيليبين**

[27] Some work does not make the distinction between implicit and explicit and/or treats them in a joint framework (Soricut 2003; Mani 2006; Wang 2010).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| An | AlTA}rp | Alty | tql | Alwfd | AllbnAny | Alrsmy | wSlt |
| that | The-plane | which | carry | delegation | Lebanese | offical | arrived |
| Alywm | AlvlAvA' | AlY | TrAbls | wkAn | qd | AtY | Alwfd |
| today | Tuesday | to | Tripoli | And-it | was | came | delegation |
| lASTHAb | Alrhynp | AllbnAnyp | mAry | AlmHtjzp | fy | alflbyn | |
| For-accompany | hostage | Lebanese | Marie | being-hold | in | Philippines | |

*The plane, which was carrying the official Lebanese delegation, arrived in Tripoli on Tuesday*. **(and) The delegation came to accompany the Lebanese hostage Marie, who held in the Philippines**.

**Al-maSdar.** This feature states whether the first or second word in Arg2 is an al-maSdar noun. 563 out of 830 instances of prepositional connectives followed by an al-maSdar indicate a CONTINGENCY.Cause relation in Set 1 (see Ex. 8-4). In addition, if both arguments start with al-maSdar nouns (1490 instances) as in Ex. 8-5, it might be linked by only Conjunction relation (431 instances).

**Ex. 8-4 (Rel: CONTINGENCY.Cause)**

لاحقته الشرطة مرارا **يتهم غير خطيرة**

| | | | | | |
|---|---|---|---|---|---|
| lAHqth | Al$rTp | mrArA | bthm | gyr | xTyrp |
| Follw-him | police | again | by-claims | non | serious |

*Police repeatedly prosecuted him* **because of non-serious charges**

**Ex. 8-5 (Rel: EXPANSION. Conjunction)**

شدد عرفات على الحاجة الى حشد كافة الجهود **و تنسيق المواقف**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $dd | ErfAt | ElY | AlHAjp | AlY | H$d | kAfp | Aljhwd | w | tnsyq | AlmwAqf |
| stressed | Arafat | on | need | to | collect | all | efforts | and | coordinate | situation |

Arafat stressed on the need *for mobilizing all efforts* **and coordinating positions**

**Tense and Negation.** Inspired by Miltsakaki (Miltsakaki et al. 2005), we stipulate that tense is useful for recognizing Temporal and Causal relations. For example, the arguments of the relation TEMPORAL.Synchronous may have the same tense. In contrast, Arg1 tense may be prior to Arg2 tense for TEMPORAL.Asynchronous and Cause relations. Each argument is assigned its tense as one of {perfect, imperfect, future or none}. We also indicate whether the tenses of Arg1 or 2 are the same and whether a negation is part of Arg 1 or 2; we use the value NONE for these two features if the argument is only a clause SBAR/ADVP or noun phrase.

**Length and Distance.** We use the length of arguments (in words), word distance between a connective and its arguments (-1: for Arg1_Conn if arguments occur in the order Arg1_Conn_Arg2_Arg1), tree distance of connective and arguments (0 if the connective and the argument are in the same tree) and a binary feature of whether Arg1 and Arg2 are in different sentences. In Set 1 (6039) of single relations, there are 3660 (61%) instances their Arg1 and Arg2 are in the same sentence, 2004 (33%) instances their Arg1 and Arg2 are in adjacent sentences, and 374 (6%) instances where Arg1 and Arg2 are in different not adjacent sentences. Some relations rarely cross sentences such as COMPARISION. Contrast (351/440), TEMPORAL.Synchronous (214/219) and CONTINGENCY.Cause.Reason (829/834) out of 3660 instances having the two arguments in the same sentence. If a tree distance between Arg1 and a connective is more than 1 (426), then the relation may be EXPANSION.Conjunction (318) or EXPANSION.Background (51).

**Argument Order.** This nominal feature encodes one of the three orders Arg1_Conn_Arg2, Conn_Arg2_Arg1 and Arg1_Conn_Arg2_Arg1, the latter being frequent in Arabic for TEMPORAL relations (83 out of 118 instances of Arg1_Conn_Arg2_Arg1). We noticed that it is a frequent practice in Arabic news to split the first argument by the connective and second argument. The order Conn_Arg2_Arg1 (90) is also frequent for CONTINGENCY.Condition instances (29).

**Argument Parent.** We record the syntactic parent of each Argument. However, not every argument corresponds to a complete tree in the ATB - in these cases we extract the category of the parent shared by the first and last word in the argument. We supposed that different combinations of S, VP, NP and SBAR would help in the recognition task.

**Production Rules.** We use all non-lexical production rules that occur more than 10 times in the arguments as binary features. This was inspired by (Lin, Kan and Ng 2009) who used production rules to good effect for implicit relations in English. Three features of production rules per instance were created (120 binary features: is_the_production_rule_applied_in_Arg1, 120 binary features: is_the_production _rule_applied_in_Arg2, and another 120 binary features is_the_production_rule _applied_in_both_Arg1andArg2).

### 8.6.2 Experimental Setup

Our models predict single discourse relations on two levels according to our relation taxonomy: (i) 17 fine-grained relations and (ii) the 4 main class relations. We examine their performance on four datasets: Set 1 (6039) and Set 2 (5880) without repetitions, Set_1_excluding_*و/w/and*_at_BOP (3813) and Set_2_excluding_*و/w/and* _at_BOP (3731). We use 10-fold cross-validation and JRip as well as a McNemar test at the 5% level for significance tests.

We use two baselines- the majority class baseline assigns the overall most frequent relation EXPANSION.Conjunction (just EXPANSION at the class level) to all instances. The second, most-frequent-sense per connective baseline corresponds to a supervised model using the connective string as the sole feature (ConnOnly).

Table 8-11 shows the performance of the two baselines, as well as a model using all features described in apart from Production rules (37f_model) and a model including the Production rules features (1237f_model).

**Table 8-11: Performance of different models for recognising single discourse relations at fine-grained level on two datasets (Set 1 all tokens and Set 2 without repetitions) with and without و/w/and at BOP.**

| **All single relation tokens** | | | | |
|---|---|---|---|---|
| | Set 1- all conn (6039) | | Set 1- excluding *و/w/and* at BOP (3813) | |
| | **Acc** | **kappa** | **Acc** | **kappa** |
| Majority baseline | 52.5 | 0 | 35 | 0 |
| ConnOnly baseline | 77.2 | 0.60 | 74.3 | 0.65 |
| Conn+Conn_f+Arg_f (37f_model) | 78.8 | 0.66 | 77 | 0.69 |
| Conn+Conn_f+ Arg_f+ Production rules (1237f_model) | 78.3 | 0.65 | 76.7 | 0.69 |
| **Single relation tokens without repetitions** | | | | |
| | Set 2 without repetitions- all conn (5880) | | Set 2 without repetitions- excluding *و/w/and* at BOP (3731) | |
| | **Acc** | **Kappa** | **Acc** | **Kappa** |
| Majority baseline | 52.3 | 0 | 35 | 0 |
| ConnOnly baseline | 77.1 | 0.61 | 74.2 | 0.65 |
| Conn+Conn_f+Arg_f (37f_model) | 78.6 | 0.65 | 76.8 | 0.69 |

### 8.6.3 Results and Evaluation

The baseline of assigning the most frequent relation EXPANSION.Conjunction to every connective performs with an accuracy of 52.5% on fine-grained relations of Set 1 *All connectives* and 35% on Set_1_excluding_*و/w/and*_at_BOP. If we use a model that relies on the string of the discourse connective alone (ConnOnly) we achieve results of 77.2%/74.3% respectively. As noted in the introduction of Section8.6, this is substantially lower than what the same model can achieve for English (Pitler and Nenkova 2009).

Including connective and argument features (apart from production rules) in 37f_model leads to a small but significant improvement. That is also true when we run the models on data without repetitions. The results of the 37f_model are almost the same; the accuracy is 78.6%/76.8% and kappa 0.65 and 0.69 on Set 2 and Set 2_excluding_*و/w/and*_at_BOP respectively. The most important fact is that the 37f_model again improves significantly over the ConnOnly model. Further incorporation of production rules (1237f_model) does not improve the results where its accuracy is 78.3% on Set 1 and 76.7% on Set 1_excluding_*و/w/and*_at_BOP. Thus, we did not run further experiments of this model on other datasets.

We use F-score per relation class in Table 8-12 to examine how well the 37f_model classified each relation compared to using the connective string alone. F-scores are particularly well-suited to look at individual classes in binary judgments, where as accuracy gives a good idea of performance on several classes. Although the 37f_model achieves an overall significant improvement, the F-score is zero for pragmatic relations and the less frequent relations such as EXPANSION.Exemplification and EXPANSION.Exception. Interestingly, the model performs very well in identifying CONTINGENCY.Condition, EXPANSION.Alternative.Conjunctive, CONTINGENCY.Cause.Reason.NonPrag-matic and COMPARISON.Contrast. In some cases (such as Condition relatins) this is due to highly informative connective strings as the ConnOnly also performs well on them. In addition, the 37f_model records better recognition than the ConnOnly model for COMPARISON.Similarity, EXPANSION.Alternative.Disjunctive, CONTINGENCY.Cause.Result.NonPrag-matic and EXPANSION.Background relations. Thus, in future work one should concentrate on improving the performance

of relations (5-12) and increasing the size of the data to cover more instances of the less frequent relations (13-17).

Regarding the main class level (4 relations), the results of the same models on the same four datasets are presented in Table 8-13. Here, surprisingly, using additional features over the connective string does not lead to significant improvements on all datasets with/out repetitions. The results are relatively high, but still less than what similar models achieved for the class level on the English PDTB, 92%.

**Table 8-12: F-score performance of the 37f_model for each relation on dataset Set 1- excluding *و/w/and* at BOP.**

| | Discourse Relation | Freq (Set 1) | 37f_model F-Measure | ConnOnly F-Measure |
|---|---|---|---|---|
| 1 | CONTINGENCY.Condition | 77 | 0.92 | 0.92 |
| 2 | EXPANSION.Alternative.Conjunctive | 28 | 0.9 | 0.89 |
| 3 | CONTINGENCY.Cause.Reason.NonPragmatic | 806 | 0.89 | 0.89 |
| 4 | COMPARISON.Contrast | 440 | 0.87 | 0.82 |
| 5 | EXPANSION.Conjunction | 3167 | 0.79 | 0.75 |
| 6 | TEMPORAL.Asynchronous | 417 | 0.79 | 0.78 |
| 7 | TEMPORAL.Synchronous | 219 | 0.75 | 0.74 |
| 8 | COMPARISON.Similarity | 14 | 0.72 | 0 |
| 9 | EXPANSION.Alternative.Disjunctive | 7 | 0.4 | 0 |
| 10 | CONTINGENCY.Cause.Result.NonPragmatic | 228 | 0.31 | 0 |
| 11 | EXPANSION.Background | 186 | 0.06 | 0 |
| 12 | EXPANSION.Reformulation | 331 | 0.02 | 0 |
| 13 | CONTINGENCY.Cause.Reason.Pragmatic | 28 | 0 | 0 |
| 14 | CONTINGENCY.Cause.Result.Pragmatic | 33 | 0 | 0 |
| 15 | CONTINGENCY.PragmaticCondition | 6 | 0 | 0 |
| 16 | EXPANSION.Exception | 5 | 0 | 0 |
| 17 | EXPANSION.Exemplification | 47 | 0 | 0 |

### 8.6.4 Error Analysis and Discussion

We concentrate our discussion on fine-grained classification on Set 1 excluding *و/w/and* at BOP (3813), the most sensible dataset without any extra modification. Our improvements in Conn+Conn_f+Arg_f model (37f_model) over the connective-only classifier (ConnOnly) are in two main areas. First, our model performs generalisation, i.e. outputs some rules that do not use the connective string at all. These achieve a somewhat surprising improvement of the 37f_model over ConnOnly

for unambiguous connectives which are too rare to classify via the connective string. In those cases, they either (i) have not been seen in the training data before and are therefore not classifiable when seen first time in the test set by the ConnOnly classifier, or (ii) have been seen in the training data too rarely for the rule-based classifier to develop a rule judged to be more reliable than the default EXPANSION.Conjunction classification.

**Table 8-13: Performance of different models of identifying class level single discourse relations on two datasets with/out repeated instances: a) all connectives, and b) excluding *و/w/and* at BOP.**

| **Class level single relation tokens** | | | | |
|---|---|---|---|---|
| | Set 1- all conn (6039) | | Set 1- excluding *و/w/and* at BOP (3813) | |
| | **Acc** | **Kappa** | **Acc** | **Kappa** |
| Majority baseline | 62.4 | 0 | 41.8 | 0 |
| Conn only baseline | 88.7 | 0.78 | 82.7 | 0.74 |
| Conn+Conn_f+Arg_f (37f_model) | 88.7 | 0.78 | 83.5 | 0.75 |
| **Class level single relation tokens without repetitions** | | | | |
| | Set 2 without repetitions- all conn (5880) | | Set 2 without repetitions- excluding *و/w/and* at BOP (3731) | |
| | **Acc** | **Kappa** | **Acc** | **Kappa** |
| Majority baseline | 62.2 | 0 | 41.7 | 0 |
| ConnOnly Baseline | 88.6 | 0.78 | 82.4 | 0.74 |
| Conn+Conn_f+Arg_f (37f_model) | 88.8 | 0.79 | 82.7 | 0.74 |

Our data includes 47 unambiguous connective types, accounting for 574 of the 3813 tokens. Of these 47 types, 30 are so rare that mistakes were reported in the connective-only classification, including جراء/*jra/because* (10: 70%), على الرغم/*ElY Alrgm/although* (9: 44%), نظرا ل/*nZrA l/because of*(9: 44%), بعدما/*bEdmA/after that* (7: 14%), بيد ان/*byd An/but* (6: 0%), غير ان/*gyr An/however* (6: 17%), رغم ان/*rgm An/although* (6: 17%), بالرغم من/*bAlrgm mn/although* (5: 0%), بفضل/*b fDl/thanks to* (5: 0%), بغية/*bgyp/desire/to* (5: 0%), في المقابل/*fyAl mqAbl/in contrast* (5: 0%), لكي/*lky/for/in order to* (5: 0%), قبيل/*qbyl/shortly before* (5: 0%), قبل ان/*qbl An/before that* (3: 0%), الا/*AlA/except* (2: 0%), عقب/*Eqb/shortly after* (2: 0%), حتى لو/*HtY lw/even if* (2: 0%), *Hyn hA/when that* (2: 0%), كي/*ky/to* (2: 0%), طالما/*TAlmA/as long as* (2:0%), *b rgm/although* (1: 0%), بحيث/*bHyv/since* (1: 0%), *bmEnY Axr/in other words* (1: 0%),

بيد/*byd/but* (1: 0%), حين/*Hyn/when* (1: 0%), كأن/*kAn/as* (1: 0%), لذا/*l*A/for this* (1: 0%), لذلك/*l*lk/for that* (1: 0%), لولا/*lwlA/if not* (1: 0%) and وقبل/*wqbl/and before* (1: 0%). The frequency and the percentage that represents the accuracy for the particular connective in the ConnOnly classifier are in brackets.

For 14 of these 30 connectives, the 37f_model was able to use generalized rules to improve relation assignment. These rules involve mainly connective surface and POS features. Thus, sentence-start adverbials consisting of more than one token such as (*بيد ان/byd An/but*, 6), (*غير ان/gyr An/however*, 6) and (*برغم/brqm/although*, 1) were correctly classified as Contrast, using GR3 in Table 8-14. For the other 16 connectives neither of the models was able to classify them correctly

This advantage of our model over the connective-only model might disappear if in a larger corpus more instances of those connectives are found and are still unambiguous. Therefore, we are more interested in how our classifier performs on truly ambiguous connectives (33 connective types accounting for 3239 tokens of 3813 overall tokens).

**Table 8-14: Generalized rules learnt by the model 37f_Model in discourse relation recognition**

| | **Generalized Rules** | **Predicted Relation (total/ incorrect classification)** |
|---|---|---|
| G1 | (First_w_arg1 = AlDrbp) | CONTINGENCY.Cause.Reason.Pragmatic (2.0/0.0) |
| G2 | (First_w_arg2 = qd) and (First_w_arg1_pos = NOUN) | EXPANSION.Background (7.0/3.0) |
| G3 | (conn_type = MoreThanToken_NonPhrase) and (sharing_parent_cat_arg2 = S) | COMPARISON.Contrast (64.0/6.0) |
| G4 | (Conn_pos = PREP#NOUN) and (conn_type = MoreThanToken_Phrase) | COMPARISON.Contrast (25.0/9.0) |
| G5 | (conn_type = Simple) and (First_w_arg2 = mn) | COMPARISON.Contrast (5.0/0.0) |
| G6 | (First_w_arg1 = AlAmr) and (Third_w_arg1_pos = VERBuIMPERFECT) | COMPARISON.Contrast (5.0/1.0) |
| G7 | (Conn_pos = PREP) and (conn_type = Clitic_in_raw_and_TB_has_pos) | CONTINGENCY.Cause.Reason.NonPragmatic (494.0/26.0) |
| G8 | (conn_type = Simple) and (Word_distance_arg1_conn = 0) and (arg1_sametime_arg2 = 0) and (First_w_arg2_pos = VERBuIMPERFECT) | CONTINGENCY.Cause.Reason.NonPragmatic (20.0/3.0) |
| G9 | (conn_type = MoreThanToken_NonPhrase) and (sharing_parent_cat_arg2 = NP) | CONTINGENCY.Cause.Reason.NonPragmatic (36.0/8.0) |

| | Generalized Rules | Predicted Relation (total/ incorrect classification) |
|---|---|---|
| G10 | (conn_type = Simple) and (Word_distance_arg1_conn = 0) and (Second_w_arg2 = h) | CONTINGENCY.Cause.Reason.NonPragmatic (8.0/1.0) |
| G11 | (conn_type = Simple) and (Word_distance_arg1_conn = 0) and (Conn_pos = NOUN) | CONTINGENCY.Cause.Reason.NonPragmatic (7.0/1.0) |
| G12 | (Second_w_arg2 = bnAA) | Rel=CONTINGENCY.Cause.Reason.NonPragmatic (6.0/2.0) |

We conducted a separate significance test on **ambiguous connectives** only and found that the 37f_model improves over ConnOnly classification significantly at the 1% level. How well we do on individual connectives depends on their frequency and on their level of ambiguity. If connectives are ambiguous and of low frequency (i.e. لو*/lw/if (in the past)*, انما*/AnmA/but* or حال*/HAl/when*), both ConnOnly and 37f_model do perform badly on them.

In contrast, if connectives are frequent (10 or more occurrences) and have relatively low ambiguity (majority reading accounts for more than 70% of their instances), the overall performance of both ConnOnly and 37f_model is equal, often both using the connective string only (see Table 8-14).

**Table 8-15: Frequent low ambiguity level connectives for which both models ConnOnly and 37f_model only use the connective string.**

| Conn | Freq | ConnOnly accuracy | 37f_model accuracy |
|---|---|---|---|
| لكن/lkn/but | 201 | 98.5% | 98.5% |
| بعد*/bEd/after* | 103 | 97.1% | 97.1% |
| اذا/A*A/if | 33 | 97.0% | 97.0% |
| الا ان*/AlA An/but* | 40 | 95.0% | 95.0% |
| ايضا*/AyDA/also* | 17 | 94.1% | 94.1% |
| ل*/l/for* | 468 | 93.4% | 93.2% |
| عندما*/EndmA/when* | 35 | 80.0% | 80.0% |
| اما*/AmA/while* | 24 | 75.0% | 75.0% |
| بل*/bl/but* | 15 | 73.3% | 66.7% |
| و*/w/and* | 1738 | 71.5% | 71.3% |

On the other hand, if connectives are frequent and have high ambiguity (i.e. no such clear majority reading), then the 37f_model normally improves (often substantially)

on ConnOnly. Examples of such connectives are كما/*kmA/as*, فيما/*fy mA/while* and اثر/*Avr/after* - the full list is in Table 8-16. Most of the successful rules use tense in some form, either via part of speech of verbs or via comparing the tense in the two arguments. This, for example, led to successful recognition of all 9 instances of Similarity for the connective كما/*kmA/as* (whose majority relation is EXPANSION. Conjunction in 40 out of 65 occurrences).

23% of the connective ف/*f/then* tokens are distinguished into EXPANSION.Exemplification, CONTINGENCY.Cause.Result and CONTINGENCY.Cause.Reason readings, depending on the lexemes around it, the parents of its arguments, and whether its argument 2 is tensed or not. Thus, non-tensed arguments are most often nominalizations which lead to a reason reading, whereas if Arg2 is a verb phrase and Arg1 is a sentence, a result reading is often used. However, it is worth reporting that in cases of connectives of very high ambiguity, 37f_model still does not yield high performance, such as for the connectives ف/*f/then* and اثر/*Avr/after*.

**Table 8-16: Improvements of 37F_model over the ConnOnly model for frequent highly ambiguous connectives.**

| Conn | Freq | ConnOnly Accuracy | 37f_model Accuracy |
|---|---|---|---|
| كما/*kmA/as*l | 65 | 61.5% | 72.3% |
| في حين/*fy Hyn/while* | 20 | 30.0% | 50.0% |
| بالتالي/bAltAly/consequently | 14 | 21.4% | 28.6% |
| فيما/fy mA/while | 27 | 18.5% | 59.3% |
| حيث/*Hyv/where/since* | 30 | 6.7% | 23.3% |
| اثر/*Avr/after* | 18 | 5.6% | 27.8% |
| اذ/A*/as | 19 | 5.3% | 21.1% |
| حتى/HtY/until | 22 | 4.5% | 27.3% |
| ف/*f/then* | 96 | 0.0% | 22.9% |
| بينما/*bynmA/while* | 14 | 21.4% | 14.3% |

Some improvements again come from generalized rules: there are some very high-coverage and high precision generalized rules that reduce dependency on the connective string. For example, clitic prepositions (such as ل/*l/for*) can without any further information be classified as CONTINGENCY.Cause.Reason.NonPragmatic covering 494 occurrences with only 26 mistakes. These are cases where the following argument is normally al-maSdar.

During the intensive error analysis that we have done, we noted that a few errors have resulted from incorrect annotation in the LADTB or in the ATB. For example, one instance of *بغية/bgyp/desire* is incorrectly classified because the connective POS is PREP rather than NOUN (which is an annotation mistake in the ATB). So this does not fit with the generalized rule for such instances (Conn_type = Simple) and (Word_distance_arg1_conn = 0) and (Conn_pos = NOUN) > Rel= Reason.NonPragmatic).

Also, there are 3 instances of the connective *لكن/lkn/but* that both models classified as COMPARISON.Contrast relation. However, they were annotated wrongly with EXPANSION relations in the LADTB instead of PragmaticContrast, which would have been the correct relation but is not in our relation taxonomy. Thus, both annotators made the same mistake and annotated them with EXPANSION relations, as in Ex. 8-6.

**Ex. 8-6**

اعتقد بان لقاءات ستعقد قريبا بين الاسرائيليين و الفلسطينيين **لكن** **لست متاكدا من ان يشارك احد كبار المسؤولين الاميركيين في هذه اللقاءات**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| AEtqd | bAn | lqA'At | stEqd | qrybA | byn | AlAsrA}ylyyn | w |
| I-think | that | meetings | Will-be-conducted | soon | between | Israelis | and |
| AlflsTnyyn | lkn | lst | mtAkdA | mn | An | y$Ark | |
| Palestinians | however | not | too-sure | from | that | particiapte | |
| AHd | kbAr | Alms&wlyn | AlAmyrkyyn | fy | h*h | AllqA'At | |
| one | senior | officials | American | in | these | meetings | |

*I think that the meetings will be held soon between the Israelis and Palestinians*, **but am not sure whether a senior American official will take part, in these meetings.**

## 8.7 Summary

Discourse modeling is an essential prerequisite for automatic discourse processing applications in computational linguistics. We presented in this chapter the first discourse modelling study for Arabic covering explicit discourse connective recognition and disambiguation. The models used a rule-based classifier, with 10-fold cross-validation on the LADTB v.1. We explored several experiments on different types of dataset for training and testing purposes: data of all tokens in the

LADTB, tokens excluding *و/w/and* at BOP, and both with and without repetitions. For connective recognition, a wide range of features is used and extracted from the available resources covering, in addition to surface-based features, tagging, parse and tokenization features, either extracted from simple automatic tagging or gold-standard annotated corpus, the ATB. A new Arabic specific feature was introduced by the al-maSdar feature for a noun next to the potential connective and became very useful for connective recognition.

The best performance is recorded for ATB-tag models which achieve highly reliable results (accuracy 92.4%, F-score 92.2% and kappa 0.82). Those, however, which were using features extracted from the simple automatic tagger performed very promisingly for discourse connective recognition; therefore with just an advanced tagger it is possible to identify explicit connectives automatically. The model proved that the good performance of discourse connective recognition is not a result from using only the connective string, since a high ambiguity exists in discourse usage of the connectives in Arabic. Thus, our models accomplished their good results by using generalized rules that recognize over 82% of the tokens including tokens of ambiguous connectives on discourse usage. The most useful features, after the connective string, are al-maSdar, POS and parent category.

For relation recognition, we used a wide variety of the features related to the explicit connectives and their arguments. We also used features which were inspired by prior work for recognising implicit relations for English such as distance between the arguments and production rules. Al-maSdar, lexical features, production rules and some surface-based features such as the type of the connective and word distance between the connective and their arguments are novel features in recognizing the sense of explicit discourse connectives. The best model for disambiguating discourse connectives reported 3% improvement in accuracy for tokens excluding *و/w/and* at BOP over the baseline of using the connective string alone. For both tasks, lexical features achieve very limited advantages over syntactic and parse features. We discussed in details the connective-based errors analysis for the models to distinguish the performance for ambiguous and unambiguous connectives in Arabic.

# Chapter 9
# Conclusions and Research Trends

Discourse relations play a critical role in linking discourse units and to make a discourse coherent. They can be signalled explicitly via discourse connectives, or can be inferred from the discourse segments without explicit signals. Studies of discourse structure paid great attention to both types of discourse relations theoretically and empirically, but were conducted on English and to a limited degree on Turkish, Hindi and Chinese). Discourse relations in Arabic have not yet been explored in large scale studies. The main goal of this study was to fill the gap between discourse processing investigations of Arabic compared to what has been achieved for other languages. Our research journey began with annotating explicit discourse relations manually and automatically. In fact, Arabic frequently uses discourse connectives explicitly to indicate discourse relations with a wide variety of connective types, as investigated in Chapter 7.

This chapter looks back on our claims and revisits critical decisions taken to achieve the promised contributions for discourse processing for Arabic. Section 9.1 summarizes three novel resources for Arabic discourse that have been developed and evaluated for corpus-based linguistic research: The first inventory of discourse connectives, the READ annotation tool for annotating explicit relations, and the LADTB, the first corpus annotated for discourse relations for Arabic. Section 9.2 discusses two sets of machine learning models that we developed to identify explicit discourse connectives and their discourse relations. These models benefit from the available syntactic resources for Arabic. For each contribution, we discuss its advantages and report the limitations that they have and how to be improved in future work in Section 9.3.

## 9.1 Resources for Arabic Discourse Processing

We presented the first effort towards producing an Arabic Discourse Treebank, the LADTB v.1. The corpus encompasses a final 6,328 annotated discourse connectives in 535 newswire texts, 80 distinct connective types and 55 different discourse relations including single and multiple relations. The LADTB has been annotated by two native Arabic speakers using the READ annotation tool, the first discourse annotation tool that can deal with Arabic characteristics to ensure a reliable annotation process (Chapter 6). The tool highlights all potential discourse connectives from a prespecified list, and allows the annotator to disambiguate the discourse connectives. It is possible to use the READ tool for annotating discourse connectives in any language supporting the Unicode format (after updating the discourse connective list in the tool package for the new language).

This study also offers the very useful resource of the first large inventory of discourse connectives in Arabic. The discourse connectives have been collected manually and automatically together with a list of their properties. The inventory contains 107 distinct potential discourse connectives for Arabic. This number is a comparable to the 100 distinct English connectives in the PDTB with a wider variety of syntactic types.

Our annotation scheme used similar annotation principles as the PDTB2, the well-established guidelines for annotating discourse connectives for English (Prasad *et al.* 2008a). We discussed the adaptations and the new principles for Arabic that have been considered on the top of the basic annotation principles in Chapter 5. The major adaptations were to allow prepositions and nouns to be discourse connectives, and allowing al-maSdar nouns to be arguments. Prepositions function as discourse connectives in English as well but have not been annotated in the PDTB2. In contrast, noun connectives are completely new in our annotation. The human annotation shows that both the identification of discourse connectives and the determination of the discourse relations they convey are reliable, apart from annotation of discourse relations for *wa/and* at BOP. The *wa/and* connective recorded the most disagreements in the LADTB; it is used to link arguments in 40% of adjacent sentences in the LADTB. This connective can indicate any relation in the Arabic taxonomy which caused lots of disagreements (see Appendix D). Our

annoation also shows that annotating both arguments (Arg1 and Arg2) are reliable after applying automatic post-processing correction for easily detectable mistakes using ATB annotation.

We also discussed the disagreement cases in the human annotation of connectives, relations and arguments. This discussion was used to derive the gold standard annotation using automatic correction for simple errors and manual correction for the rest. In this first study of discourse connectives in Arabic, disagreed tokens of *و/w/and at BOP* were assigned automatically to Conjunction relations, the most frequently annotated relation of the agreed tokens of *و/w/and at BOP* in the LADTB.

A statistical comparison study between discourse annotation of newswire text in Arabic (the LADTB) and in English (the PDTB) was conducted in Section 7.7. Unlike the PDTB, the LADTB has a wider syntactic variety of connectives and its connectives are more ambiguous between having discourse function or not. In addition, 70% of adjacent sentences in the LADTB are linked via explicit connectives. This highlights the importance of the usage of explicit discourse connectives in MSA and the promising impact of recognizing them with their discourse function automatically. With regard to discourse relations, Expansion and Contingency relations are used more frequently in Arabic than in English, whereas, more Comparison and Temporal relations are used in English than in Arabic. This might be due to the high usage of *و/w/and* and the automatic solution to the disagreement cases of *و/w/and* at BOP. In addition, the PDTB contains a wider array of genres which might contain more Condition and Contrast relations than in the LADTB.

**Reflections of Decisions Made when Creating the LADTB**

The LADTB is a discourse annotation of the newswire corpus ATB Part1. Using newswire text, on the one hand, affects on our collection of Arabic discourse connectives and their relations. On the other hand, the extreme usage of *و/w/and* at BOP in newswire text led to a higher inter-annotator disagreement on its function (relations). Annotating different genres will introduce more discourse connectives and relations.

We based our annotation on similar annotation principles as the English PDTB2 which annotates local relations only. Therefore, the LADTB does not show how

discourse is constructed in Arabic newswire. We only annotate explicit relations that are signaled by discourse connectives in the LADTB. However, we noticed other discourse linking devices and implicit relations during our annotation that need advanced studies. In addition, our adaptation of the annotation manual involved merging more fine-grained relations into their upper level relation such as subrelations of Reformulation (Section 5.6.1), and excluding fine-grained relations such as List from our relation taxonomy in order to get higher inter-annotator agreement. These relations should be included again with other fine-grained relations in an advanced annotation study, as they are very useful and not very rare in the LADTB.

Despite the advantages of using the stand-off annotation tool READ that we developed (Sections 3.2.3 and 6.1), the tool does not show the syntactic boundaries of clauses and sentences which led to high relatively inter-annotator disagreement on argument boundaries in our annotation (Section 7.5.2). The tool also does not do automatic post-processing to exclude punctuations at the end of sentences or function words at beginning of sentences. This increased the disagreement cases and the manual verification in the current annotation.

## 9.2 Modeling of Explicit Discourse Relations

This first discourse corpus for Arabic, the LADTB v.1, was used to develop the first algorithms to detect discourse connectives and their interpretations. Supervised machine learning models were trained and their results evaluated according to the discourse annotation in the LADTB. Because of the effect of *و/w/and* at BOP in our annotation, several experiments were explored on different datasets: for all annotated tokens and for tokens excluding *و/w/and* at BOP. A wide range of features has been extracted from the available resources covering, in addition to surface-based features, syntax, parse and tokenization features, which were extracted either from automatic tagging or the gold-standard ATB.

The best performance is recorded for models using ATB annotation which achieve highly reliable results (accuracy 92.4%, F-score 87% (positive class) and kappa 0.82) for discourse connective recognition and moderately reliable results (accuracy 78.8% and kappa 0.66) for disambiguating discourse connectives.

Because of the high ambiguity in discourse usage of the potential connectives in the LADTB, the connective string alone is not sufficient to identify discourse connectives. However, our best model accomplished very significant improvements by using generalized rules that recognize 28% of the tokens (including tokens of ambiguous connectives) without using the connective string. Very promising results in discourse connective recognition were also recorded for those models that use features extracted via automatic tagging (M4); thus, explicit connectives can be identified automatically when using an advanced tagger for Arabic.

The thesis also presented intensive connective-based error analysis of our models that classified connectives according to their ambiguity level in terms of having discourse usage (for identifying the connectives) and having more than one sense (for disambiguating the connective interpretations).

Models for disambiguating discourse connectives with regard to their sense reported a 3% improvement in accuracy for tokens excluding *ﻭ/w/and* at BOP over using the connective string alone. The most useful features in recognising discourse relations, after the connective string, are al-maSdar of the nouns after the connective, POS of the connective and of the words at the beginning of the arguments, parent category of the connective and word distance between the connective and its arguments. The novel features in recognizing the sense of explicit discourse connectives that we use are Al-maSdar, lexical features, production rules and some surface-based features such as the type of the connective (Clitic, Simple or MoreThanToken) and the word distance between the connective and its arguments.

For both tasks, lexical features reported very limited advantages over syntactic and parse tree features. We also faced some limitations in our experiments due to the lack of reliable resources for Arabic NLP. For example, we were unable to extract parse features by running the automatic Stanford parser for Arabic[28], because the parser requires a highly accurate pre-processing tokenization, and such a tokenizer was not available to us at the study time.

For a similar reason, we could not examine how semantic classes of frequent words would improve the results by using, for example, the Arabic WordNet (Elkateb *et al.*

[28] The only freely available parser for Arabic, http://nlp.stanford.edu/software/parser-arabic-faq.shtml

2006)[29] and RDB (the Arabic lexical semantics)[30] (Attia *et al.* 2008). The Arabic WordNet is an incomplete project and still a very small resource (≈ 12,038 entities) which would not cover many of the words in our news corpus. In addition, syntactic dependency features, which might be very useful for recognising discourse relations, require resources such as the Dependency Treebank which is also not available for our corpus ATB Part1[31].

**Reflections of Decisions Made for Modelling Discourse Relations**

As we based our annotation of the LADTB on the ATB, our models of identifying discourse connectives and relations, on the one hand, got a huge benefit from the syntactic and parse features in the ATB. On the other hand, the ATB annotation does not involve annotation of semantic or dependency features which might improve further the performance of our models. The ATB also has some repetitions in files and parts of the text. We, therefore, examined our models also on the datasets excluding all token repetitions. The models use the ATB annotation achieved significant improvement over using the connective string alone, but this benefit might disappear when there is roubust automatic ATB annotation for unseen text.

In addition, the extreme use of the most ambiguous connective *ﻭ/w/and* at BOP and BOS in the LADTB and, therefore, the decision made of assigning Conjunction relation to its frequent disagreements led us to conduct experiments of relation recognition on two datasets including and excluding these annotations.

Although we benefit from using rule-based classifier in our discourse modelling, we noticed that the order of the rules in JRip classifier might play an important role behind some results such as misclassification of less frequent unambiguous connectives and some frequent ambiguous connectives (Section 8.4.4).

[29] http://www.globalwordnet.org/AWN/
[30] This language resource is not available to the public.
[31] Nizar Habash thankfully has shared with us his convertor ATB-to-CATiB-style which is used to build the Columbia Arabic Treebank (Habash and Roth 2009). However, the convertor works only with the latest ATB annotation standards, and unfortunately not with the older version such as the one we used in this project (ATB Part1, 2003).

## 9.3 Future Research Trends

The new resources and models presented so far for Arabic discourse processing, will establish a reliable foundation for many interesting linguistic and corpus-based studies. The READ tool, the first discourse connective list for Arabic, and the discourse annotation scheme are available either via the LADTB website (www.arabicdiscourse.net) or through the authors for the public to use, improve and evaluate. The LADTB v.1 will be released in 2012 via the LDC. We encourage researchers in bilingual studies to run corpus-based studies using the LADTB and our collection of Arabic discourse connectives to investigate the similarities and differences in the newswire text of languages with regard to how connectives relate similar segments, and enhance further empirical applications such as machine translation. We discussed some differences between Arabic and English connectives (Sections 4.6 and 7.7) which can act as triggers for other studies and applications. We provided an estimate comparison between the LADTB and the PDTB2. As mentioned in Section 7.7, this comparison does not reflect discourse proprieties of newswire of Arabic and English due to the differences in size, genres and annotation guidelines of the two corpora.

Future studies of discourse processing for Arabic might be classified into (i) studies to improve the coverage and the quality of current discourse resources, (ii) studies to improve the performance of the automatic models, and (iii) studies to enhance language applications for Arabic such as machine translation, summarization, question answering, and readability scoring. The latter might build on insight for the applications for English.

**To improve the quality and coverage of current discourse resources**.

It would be good to overcome the mistakes in the syntactic annotation of ATB Part1 2003, which we used in the LADTB v.1, by using the new syntactic annotation of the same corpus that was distributed in 2010 via the LDC. This would lead to a new version, the LADTB v.1.1.

It is possible, in order to enlarge the LADTB, to identify discourse connectives and their arguments and relations automatically in other parts of the ATB, and then verify those manually. It also would be good to increase the size of the LADTB by

annotating more text from different genres. That is necessary to cover more instances of low frequency connectives and relations (see the discourse connective and relation distribution in Appendix C and D). For example, annotating instruction manuals would increase the number of instances of the Condition relations.

From our human annotation experience and the agreement studies, we also suggest adding some relations to the relation taxonomy in the scheme. Annotators often disagreed on the relations signalled by some discourse connectives, got confused with current relations and sometimes introduced new relations as comments. For example, the connective *ل/l/for* as in *لعمل/for doing* indicates almost always a Cause relation but is sometimes closer to the relation *purpose*, which is not in our relation taxonomy. In addition, we need to include the fine-grained relations such as List and Reformulation relations (Specification, Generalization and Equivalence), as we expect these relations to be important for some applications such as automatic summarization (see Section 5.6.1 for related discussion).

In addition, the guidelines of Arabic discourse annotation might be enhanced insight of our discussion and observations in our annotation (Section 7.5). In particular, our annotation guidelines contains special cases that need further annotation study in the next advanced version of the LADTB such as (i) we did not allow combining EXPANSION. Conjunction relation with any other relations in our taxonomy (Section 5.6.3), and (ii) we annotate Entity relations between conjoined clauses with EXPANSION. Conjunction relation (Section 5.6.4).

Moreover, an intensive linguistic study should address the connective *و/w/and* at BOP or at BOS. The connective *و/w/and* introduces 40% of sentences in the LADTB as a discourse connective (Section 7.7.1). It also introduces 30% of Quran verses using Kais Qurainic Corpus (Dukes and Buckwalter 2010) as a potential discourse connective. This connective can signal any relation in our relation taxonomy. In addition, it is very interesting to find out whether all implicit relations in English could be translated into the connective *و/w/and* in Arabic, and whether the connective *و/w/and* at BOS can always be omitted when translating from Arabic into English.

Although the discourse annotation in the present study focused on the annotation of explicit connectives and their relations, we also came across other discourse devices during our analysis such as implicit connectives (inferred relations), entity relations,

attribution and anaphora. An advanced version of the LADTB discourse corpus must annotate new cohesion devices.

The READ tool also could be improved by applying an automatic syntactic parser to show potential argument boundaries for the annotators and exclude automatically punctuations that were annotated mistakenly.

**To improve the performance of automatic models for recognising discourse connectives and relations, and to use them to improve language applications.**

The most mileage in modelling discourse relations is in further improvements on frequent ambiguous connectives, whether with regard to discourse usage such as في حين/*fy Hyn/while*, منذ/*mn*/since* and قبل/*qbl/before* (Section 8.4.4); or signalling more than one relation such as ف/*f/then*, منذ/*mn*/since* and او/*Aw/or* (Section 8.6.4). Moreover, one should concentrate on improving the performance of relations with less F-score in Table 8-12 such as EXPANSION.Alternative.Disjunctive, CONTINGENCY.Cause.Result. NonPragmatic, EXPANSION.Background and EXPANSION.Reformulation.

This can be achieved with, on the one hand, training connective-specific classifiers on larger data sets to cover more instances of the less frequent discourse connectives such as عقب/*Eqb/shortly_after*, بفضل/*bfdl/thank to*, كلما/*klma/when ever* (Section 8.4.4), and of the less frequent discourse relations such as pragmatic relations, EXPANSION.Exception and EXPANSION.Exemplification (Section.8.6.3).

On the other hand, the classifiers also need a wider feature base. In particular connective-based features such as a morphological pattern(s) (see the discussion of ambiguous connectives in Section 8.4.4). In addition, we think from our corpus study that lexico-semantic features such as word pairs and semantic classes of verbal/nominalised arguments are the most promising new features in recognising discourse relations. We were unable to use these features as they need either a larger corpus or a deeper semantic ontology than the existing one (the WordNet). Therefore, a further cooperation is required with specialists in semantic analysis to enhance the Arabic resources for a wide coverage semantic annotation.

As the results of the models using features from automatic tagging (Stanford Tagger) are promising for discourse connective recognition (Section 8.4.3), it is good to

examine the models also using a proper ATB-style tokenization or using more advanced automatic tagger and parser for Arabic when they are available. We also suggest that using semi-supervised methods for relation recognition to alleviate data sparseness might achieve better improvement for some connectives.

It is also worth conducting experiments using different classifiers to overcome any drawbacks caused by the rule-based classifier. In contrast, general rules generated by JRip classifier can handle data with previously unseen potential connectives. The reader can refer to our discussion of the generalization by the connective recognition model in Section 8.4.4. It might be true that some rules that do not use the type or pos tag of the connective, can also be used to predict *implicit connectives* (no connective string to indicate the relation) such as the rule {(Isalmasdar_w_after_conn = Not_masdar) and (Right_sib = S)} in Table 8-2, if we suppose that the implicit connective should introduce a sentence/clause.

As we focused in this study on recognising discourse connectives and relations, one important future task is to develop algorithms to detect argument boundaries automatically. By automating all three discourse parsing components for Arabic, we can move forward to use these models to enhance language applications. A similar discourse corpus to the LADTB, the PDTB, has been used so far for discourse parsing, content summarization, question generation, genre distinctions and readability scoring (see Section 2.7). One other potential application of our models is to annotate Arabic discourse connectives for other genres, for example, the classical Arabic corpora (ie. Kais Quranic Corpus).

# Bibliography


AKTAŞ, B., C. BOZSAHIN and D. ZEYREK. 2010. Discourse relation configurations in Turkish and an annotation environment. *In*: Association for Computational Linguistics, pp.202-206.

AL-SANIE, W., A. TOUIR and H. MATHKOUR. 2005. Towards a Rhetorical Parsing of Arabic Text. *In: The International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC'05)*: IEEE Computer Society.

AL-SUGHAIYER, I. A. and I. A. AL-KHARASHI. 2004. Arabic morphological analysis techniques: A comprehensive survey. *JASIST* 55(3), pp.189-213.

ALANSARI, I. H. 1985. *Mogny Alabib En Kutb AlAEareb.* Lebnan: Dar Alfekur.

ALFARABI, H. 1990. *Ketab Alhroof.* Dar Almashreg, Lebnan.

ARTSTEIN, R. and M. POESIO. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34 (4), pp.555-596.

ASHER, N. 1993a. *Reference to Abstract Objects in Discourse.* Boston MA: Kluwer.

ASHER, N. 1993b. *Reference to Abstract Objects in Discourse* Kluwer Academic Publishers

ATTIA, M., M. RASHWAN, A. RAGHEB, M. AL-BADRASHINY, H. AL-BASOUMY and S. ABDOU. 2008. A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields. *In: Lecture Notes on Computer Science (LNCS): Advances in Natural Language Processing, August, Verlag Berlin Heidelberg*. pp.65-76.

ATTIA, M. A. 2007. Arabic tokenization system. *In*: Association for Computational Linguistics, pp.65-72.

BLAIR-GOLDENSOHN, S., K. R. MCKEOWN and O. C. RAMBOW. 2007. Building and Refining Rhetorical-Semantic Relation Models. *Proceedings of NAACL HLT*, pp.428-435.

BUCH-KROMANN, M. and I. KORZEN. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. *In: Proceedings of the Fourth Linguistic Annotation Workshop*, July 15-16, Uppsala, Sweden. pp.127-131.

CARLSON, L., D. MARCU and M. E. OKUROWSKI. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *In: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16, Aalborg, Denmark*. Association for Computational Linguistics.

CARLSON, L., D. MARCU and M. E. OKUROWSKI. 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.

CARLSON, L., M. E. OKUROWSKI, D. MARCU and L. D. CONSORTIUM. 2002. *RST discourse treebank.* Linguistic Data Consortium, University of Pennsylvania.

CAVALLI-SFORZA, V. and I. ZITOUNI. 2007. An Arabic Slot Grammar Parser. *In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Association for Computational Linguistics (ACL)* Prague, Czech Republic.

CHARNIAK, E. 2000. A maximum-entropy-inspired parser. *In: The Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp.132-139.

CHIANG, D., M. DIAB, N. HABASH, O. RAMBOW and S. SHAREEF. 2006. Parsing arabic dialects. *In: Proceedings of EACL-06*, p.112.

DICKINS, J., I. HIGGINS and S. HERVEY. 2002. *Thinking Arabic Translation.* Routledge.

DIJK, T. A. V. 1997. *Discourse as Structure and Process: A Multidisciplinary Introduction: Discourse as Structure and Process* SAGE Publications Ltd.

DIPPER, S. and M. STEDE. 2006. Disambiguating potential connectives. *In: Proceedings of the Konvens-2006 Workshop on the Lexicon-Discourse Interface*, pp.167-173.

DUKES, K. and T. BUCKWALTER. 2010. A dependency treebank of the Quran using traditional Arabic grammar. *In: the 7th international conference on Informatics and Systems,* Cairo, Egypt.

DUVERLE, D. A. and H. PRENDINGER. 2009. A novel discourse parser based on support vector machine classification. *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*: Association for Computational Linguistics, pp.665-673.

ELKATEB, S., W. BLACK, P. VOSSEN, D. FARWELL, H. RODRIGUE, A. PEASE, M. ALKHALIFA and C. FELLBAUM. 2006. Arabic WordNet and the Challenges of Arabic. *In: The Challenge of Arabic for NLP/MT, International Conference at The British Computer Society (BCS), 23 October, London*

ELWELL, R. and J. BALDRIDGE. 2008. Discourse connective argument identification with connective specific rankers. *In: Proceedings of the IEEE International Conference on Semantic Computing,* Santa Clara, CA.

FORBES-RILEY, K., B. WEBBER and A. JOSHI. 2006. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *J Semantics,* 23 (1), pp.55-106.

FRASER, B. 1999. What are discourse markers? *Journal of Pragmatics,* 31 (7), pp.931-952.

GIRJU, R. 2003. Automatic detection of causal relations for question answering. *In: Proceedings of the ACL 2003, Workshop on "Multilingual Summarization and Question Answering - Machine Learning and Beyond",* Sapporo, Japan. pp.76-83.

GRAFF, D. 2003. Arabic Gigaword, LDC Catalog No. *LDC2003T12. Linguistic Data Consortium, University of Pennsylvania*.

GROSZ , B. J. and C. L. SIDNER. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics,* 12 (3), pp.175-204.

HABASH, N. and O. RAMBOW. 2004. Extracting a tree adjoining grammar from the Penn Arabic Treebank. *In: Proceedings of Traitement Automatique du Langage Naturel (TALN-04),* Fez, Morocco. pp.277-284.

HABASH, N. and R. M. ROTH. 2009. *Catib: The columbia arabic treebank*. Technical Report CCLS-09-01, Center for Computational Learning Systems, Columbia: Columbia University.

HABASH, N. Y. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies,* 3 (1), pp.1-187.

HADDOW, B. 2005. *Acquiring a Disambiguation Model For Discourse Connectives*. Master of Science thesis, School of Informatics University of Edinburgh.

HAJIC, J., O. SMRZ, P. ZEMÁNEK, J. ŠNAIDAUF and E. BEŠKA. 2004. Prague Arabic dependency treebank: Development in data and tools. *In: Proc. of NEMLAR,* pp.110-117.

HALLIDAY, M. A. K. and R. HASSAN. 1976. *Cohesion in English.* London: Logman.

HARMANANI, H. M., W. T. KEIROUZ and S. RAHEEL. 2006. A rule-based extensible stemmer for information retrieval with application to Arabic. *The International Arab Journal of Information Technology,* 3 (3), pp.265-272.

HOBBS, J. R. 1985. 85-37. *On the Coherence and Structure of Discourse*. Center for the Study of Language and Information (CSLI), Stanford University.

HOVY, E. H. 1990. Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. *In: 5th ACL Workshop on Natural Language Generation,* Dawson, Pennsylvania.

HOVY, E. H. 1993. Automated Discourse Generation using Discourse Structure Relations. *In: Artificial Intelligence (Special Issue on Natural Language Processing)*. Elsevier, pp.341-385.

HOVY, E. H. and E. MAIER. 1993. *Parsimonious and Profligate: How Many and Which Discourse Structure Relations? Discourse Processes* University of Southern Claifornia.

HUTCHINSON, B. 2004a. Acquiring the meaning of discourse markers. *In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain*. Association for Computational Linguistics.

HUTCHINSON, B. 2004b. Mining the web for discourse markers. *In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal., pp.407–410.

HUTCHINSON, B. 2005a. Modelling the similarity of discourse connectives. *In: Proceedings of the the 27th Annual Meeting of the Cognitive Science Society (CogSci2005)*.

HUTCHINSON, B. 2005b. Modelling the substitutability of discourse connectives. *In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Ann Arbor, USA*.

KHALIFA, I. and A. M. FARAWILA. 2012. A Comprehensive Taxonomy of Arabic Discourse Coherence Relations.

KHORSHEED, M. S. 2003. Recognising handwritten Arabic manuscripts using a single hidden Markov model. *Pattern Recognition Letters*, 24 (14), pp.2235-2242.

KNOTT, A. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. thesis.

KNOTT, A. and T. SANDERS. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30 (2), pp.135-175.

LAPATA, M. and A. LASCARIDES. 2004. Inferring sentence-internal temporal relations. *In: Proceedings of the North American Chapter of the Assocation of Computational Linguistics*, pp.153-160.

LAPATA, M. and A. LASCARIDES. 2006. Learning Sentence-internal Temporal Relations. *Journal of Artificial Intelligence Research*, 27, pp.85-117.

LIN, Z., M. Y. KAN and H. T. NG. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. *In: Proceedings of EMNLP'09*, pp.343-351.

LIN, Z., H. T. NG and M. Y. KAN. 2010. A PDTB-styled end-to-end discourse parser. *Arxiv preprint arXiv:1011.0835*. Report No.TRB8/10.

LITMAN, D. J. and J. F. ALLEN. 1990. Discourse processing and commonsense plans *In: In P. R. Cohen J. L. Morgan and M E Pollack editors Intentions in Communication Cambridge*. MIT Press.

LOUIS, A. and A. NENKOVA. 2010. Creating local coherence: an empirical assessment. *In: Proc. of NAACL*, pp.313-316.

LOUIS, A. and A. NENKOVA. 2011. General versus specific sentences: automatic identification and application to analysis of news summaries.

M. ABDL AL LATIF, A. U., M. ZAHRAN and D. A. AL-ARABI. 1997. *Alnhw AlAsAsi*. CSLI.

MAAMOURI, M. and A. BIES. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. *In: Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, Stroudsburg, PA. ACL, pp.2-9.

MAAMOURI, M., A. BIES, T. BUCKWALTER and W. MEKKI. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus *In: NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

MAAMOURI, M., A. BIES and S. KULICK. 2006. Diacritization: A challenge to arabic treebank annotation and parsing. *Proceedings of the Conference of the Machine Translation SIG of the British Computer Society*.

MAAMOURI, M., A. BIES and S. KULICK. 2008. Enhancing the Arabic Treebank: A collaborative effort toward new annotation guidelines. *Proceedings of the LREC 2008*.

MANI, I., M. VERHAGEN, B. WELLNER, C. M. LEE and J. PUSTEJOVSKY. 2006. Machine learning of temporal relations. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp.753-760.

MANN, W. C. and S. A. THOMPSON. 1987. *Rhetorical Structure Theory: a theory of text organization*. Technical Report ISI/RS- Information Sciences Institute

MANN, W. C. and S. A. THOMPSON. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text,* 8, pp.243-281.

MARCU, D. 1999a. A decision-based approach to rhetorical parsing. *In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics,* Maryland, USA. pp.365-372.

MARCU, D. 1999b. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization, The MIT Press*, pp.123-136.

MARCU, D. 2000a. Extending a formal and computational model of Rhetorical Structure Theory with intentional structures by Grosz and Sidner. *In: Proceedings of the 18th conference on Computational linguistics,* Saarbr, Germany. ACL.

MARCU, D. 2000b. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational linguistics,* 26(3), pp.395-448.

MARCU, D. 2000c. *The Theory And Practice Of Discourse Parsing And Summarization.* A Bradford book.

MARCU, D. and A. ECHIHABI. 2002 An Unsupervised Approach to Recognizing Discourse Relations *In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002),* Morristown, NJ, USA. pp.368-375.

MARCU, D., C. LYNN and W. MAKI. 2000. The automatic translation of discourse structures. *In: Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, Seattle, Washington*. Morgan Kaufmann Publishers Inc.

MARCUS, M. P., B. SANTORINI and M. A. MARCINKIEWICZ. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics,* 19, pp.313--330.

MASLENNIKOV, M. and T. S. CHUA. 2007. A Multi-resolution framework for information extraction from free text. *In*, p.592.

MILTSAKAKI, E., N. DINESH, A. LEE, R. PRASAD, A. JOSHI and B. WEBBER. 2005a. Experiments in Sense Annotation and Sense Disambiguation of Discourse Connectives. *In: Proceedings of Fourth Workshop on Treebanks and Linguistic Theories (TLT-2005).*

MILTSAKAKI, E., N. DINESH, R. PRASAD, A. JOSHI and B. WEBBER. 2005b. Experiments on sense annotations and sense disambiguation of discourse connectives. *In*: Citeseer.

MILTSAKAKI, E., R. PRASAD, A. JOSHI and B. WEBBER. 2004. Annotating Discourse Connectives and their Arguments *In: the HLT/NAACL workshop on Frontiers in Corpus Annotation,* Boston, MA.

MILTSAKAKI, E., R. PRASAD, A. JOSHI and B. WEBBER. 2006. The Penn Discourse Treebank.

MIZUTA, Y., A. KORHONEN, T. MULLEN and N. COLLIER. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics,* 75(6), pp.468-487.

MOORE, J. D. and C. L. PARIS. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational linguistics,* 19(4), pp.651-694.

MOORE, J. D. and M. E. POLLACK. 1992. A problem for RST: the need for multi-level discourse analysis. *Computational linguistics,* 18(4), pp.537-544.

MORTON, T. and J. LACIVITA. 2003. WordFreak: an open tool for linguistic annotation. *In: HLT/NAACL 2003: demonstrations*: Association for Computational Linguistics, pp.17-18.

MOSER, M. and J. D. MOORE. 1996. Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics,* 22(3), pp.409-419.

NICHOLAS, N. 1995. Parameters for Rhetorical Structure Theory Ontology. *University of Melbourne Working Papers in Linguistics,* 15, pp.77-93.

OZA, U., R. PRASAD, S. KOLACHINA, D. M. SHARMA and A. JOSHI. 2009. The hindi discourse relation bank. *In: Proceedings of the Third Linguistic Annotation Workshop*, Suntec, Singapore.

PITLER, E., A. LOUIS and A. NENKOVA. 2009. Automatic sense prediction for implicit discourse relations in text. *In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, pp.683-691.

PITLER, E. and A. NENKOVA. 2008. Revisiting readability: A unified framework for predicting text quality. *In*: Association for Computational Linguistics, pp.186-195.

PITLER, E. and A. NENKOVA. 2009. Using syntax to disambiguate explicit discourse connectives in text. *In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Singapore. Association for Computational Linguistics, pp.13-16.

PITLER, E., M. RAGHUPATHY, H. MEHTA, A. NENKOVA, A. LEE and A. JOSHI. 2008. Easily identifiable discourse relations. *In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK.

POLANYI, L. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5-6), pp.601-638.

POLANYI, L. 1998. *The Linguistic Discourse Model: Towards a Formal Theory of Discourse Structure*. Cambridge: BBN Labs.

POLANYI, L. and M. V. D. BERG. 1996. Discourse structure and discourse interpretation. *In P. Dekker and M. Stokhof, editors, Proceedings of the Tenth Amsterdam Colloquium*, pp.113--131.

POLANYI, L., C. CULY, V. D. BERG, M. A. THIONE, G. LORENZO and D. AHN. 2004. A Rule Based Approach to Discourse Parsing. *In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, Massachusetts, USA. ACL2004, pp.108-117.

POPESCU-BELIS, A. and S. ZUFFEREY. 2006. *Automatic Identification of Discourse Markers in Multiparty Dialogues*. School of Translation and Interpretation, University of Geneva.

PRASAD, R., N. DINESH, A. LEE, A. JOSHI and B. WEBBER. 2007a. Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*, 47(2), pp.43-64.

PRASAD, R., N. DINESH, A. LEE, E. MILTSAKAKI, L. ROBALDO, A. JOSHI and B. WEBBER. 2008a. The Penn Discourse TreeBank 2.0. *In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

PRASAD, R., S. HUSAIN, D. M. SHARMA and A. JOSHI. 2008b. Towards an Annotated Corpus of Discourse Relations in Hindi. *In: In The Third International Joint Conference on Natural Language Processing, January 2008, India*.

PRASAD, R., A. JOSHI, N. DINESH, A. LEE, E. MILTSAKAKI and B. WEBBER. 2005. The Penn Discourse TreeBank as a Resource for Natural Language Generation *In: Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, Birmingham, U.K.

PRASAD, R., A. JOSHI and B. WEBBER. 2010a. Exploiting scope for shallow discourse parsing. *In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta*.

PRASAD, R., A. JOSHI and B. WEBBER. 2010b. Realization of discourse relations by other means: alternative lexicalizations. *In: Proc. COLING 2010*, Beijing. Association for Computational Linguistics, pp.1023-1031.

PRASAD, R., E. MILTSAKAKI, N. DINESH, A. LEE, A. JOSHI, L. ROBALDO and B. WEBBER. 2007b. The penn discourse treebank 2.0 annotation manual. *The PDTB Research Group (2007)*.

PUSTEJOVSKY, J., C. HAVASI, R. SAURI, P. HANKS, A. RUMSHISKY and J. CASTANO. 2006. Towards a generative lexical resource: The brandeis semantic ontology. *In: Proceedings of the Fifth Language Resource and Evaluation Conference, 2006,* ITALY, GENOA.

RYDING, K. C. 2005. *A reference grammar of modern standard Arabic.* Cambridge: Cambridge University Press.

SADAT, F. and N. HABASH. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. *In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,,* Sydney, Australia. Association for Computational Linguistics.

SANDERS, T. J. M. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes,* 15(1), pp.1-35.

SAWALHA, M. and E. S. ATWELL. 2010. Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. *In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10),* Malta. European Language Resources Association (ELRA), pp.282-287.

SCHOURUP, L. 1999. Discourse markers. *Lingua,* 107(3-4), pp.227-265.

SEIF, A., H. MATHKOUR and A. TOUIR. 2005a. An RST Computational Tool for the Arabic Language. *In: iiWAS, Malaysia, Kuala Lumpur.*

SEIF, A., H. MATHKOUR and A. TOUIR. 2005b. An RST Computational Tool for the Arabic Language. *In: iiWAS.*

SHAALAN, K. F. 2005. Arabic GramCheck: A grammar checker for Arabic. *Software Practice and Experience,* 35(7), pp.643-665.

SORICUT, R. and D. MARCU. 2003. Sentence level discourse parsing using syntactic and lexical information. *In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1,* Edmonton, Canada. pp.149-156.

SPORLEDER, C. and A. LASCARIDES. 2005. Exploiting linguistic cues to classify rhetorical relations. *In: Proceedings of Recent Advances in Natural Language Processing (RANLP-05),* Borovets, Bulgaria.

SPORLEDER, C. and A. LASCARIDES. 2006. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering,* pp.1-48.

SPORLEDER, C. and A. LASCARIDES. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering,* 14(03), pp.369-416.

STEDE, M. 2004. The Potsdam Commentary Corpus. *In: Proceedings of the ACL 2004 Workshop on Discourse Annotation.,* Barcelona,.

TABOADA, M. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics,* 38(4), pp.567-592.

TABOADA, M. and W. C. MANN. 2006a. Applications of rhetorical structure theory. *Discourse Studies,* 8(4), pp.567-588.

TABOADA, M. and W. C. MANN. 2006b. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies,* 8(3), pp.423-459.

WALKER, M. A. and J. D. MOORE. 1997. Empirical studies in discourse. *Comput. Linguist.,* 23(1), pp.1-12.

WANG, W. T., J. SU and C. L. TAN. 2010. Kernel based discourse relation recognition with temporal ordering information. *In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010),* Uppsala, Sweden. Association for Computational Linguistics, pp.710-719.

WEBBER, B. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science,* 28(5), pp.751-779.

WEBBER, B. 2006. Accounting for Discourse Relations: Constituency and Dependency. *Intelligent Linguistic Architectures CSLI Publications,* pp.339-360. .

WEBBER, B., M. EGG and V. KORDONI. 2011. Discourse Structure and Language Technology. *Natural Language Engineering,* 1, pp.1-49.

WEBBER, B., A. KNOTT and A. JOSHI. 1999. Multiple Discourse Connectives in a Lexicalized Grammar for Discourse *In: In Third International Workshop on Computational Semantics, , Tilberg, The Netherlands*.

WEBBER, B., A. KNOTT, M. STONE and A. JOSHI. 1999. Discourse relations: a structural and presuppositional account using lexicalised TAG. *In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland*. Association for Computational Linguistics.

WEBBER, B. and R. E. A. PRASAD. 2006. *The Penn Discourse TreeBank 1.0 Annotation Manual*. University of Pennsylvania: Institute for Research in Cognitive Science.

WEBBER, B., M. STONE, A. JOSHI and A. KNOTT. 2003. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4), pp.545-587.

WELLNER, B. and J. PUSTEJOVSKY. 2007. Automatically identifying the arguments of discourse connectives. *In: Proceedings of Emprical Methods in Natural Language Pro-cessing and the Conference on Natural Language Learning*, Prague, Czech Republic.

WELLNER, B., J. PUSTEJOVSKY, C. HAVASI, A. RUMSHISKY and R. SAURI. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. *In: SigDIAL '06 Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia. Association for Computational Linguistics, pp.117-125.

WIEBE, J., T. WILSON and C. CARDIE. 2005. *Annotating expressions of opinions and emotions in language*. Springer.

WILCOCK, G. 2009. Introduction to linguistic annotation and text analytics. *Synthesis Lectures on Human Language Technologies*, 2(1), pp.1-159.

WILLIAMS, S. and E. REITER. 2003. A corpus analysis of discourse relations for Natural Language Generation. *In: In the Proceedings of Corpus Linguistics,*, Lancaster University., pp.899-908.

WITTEN, I. H., E. FRANK, L. E. TRIGG, M. A. HALL, G. HOLMES and S. J. CUNNINGHAM. 1999. Weka: Practical machine learning tools and techniques with Java implementations. *Proc ICONIP/ ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences*, pp.pp. 192-196.

WOLF, F. and E. GIBSON. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2), pp.249-287.

WOLF, F., E. GIBSON, A. FISHER and M. KNIGHT. 2003. A procedure for collecting a database of texts annotated with coherence relations. *Documentation accompanying the Discourse GraphBank, LDC2005T08*.

XUE, N. 2005. Annotating Discourse Connectives in the Chinese TreeBank. *In: Proceedings of the ACL 2005 Workshop on Frontiers in Corpus Annotation: Pie in the Sky II.*, Ann Arbor, Michigan.

ZEYREK, D. and B. WEBBER. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. *In: Proceedings of IJCNLP-2008. ,* Hyderabad, India.

ZHOU, Z. M., Y. XU, Z. Y. NIU, M. LAN, J. SU and C. L. TAN. 2010. Predicting discourse connectives for implicit discourse relation recognition. *In: Proc. of Coling 2010,,* Beijing, Chaina. Association for Computational Linguistics, pp.1507-1514.

# Appendices

# Appendix A

# Al-maSdar Morphological Forms

We used the morphological patterns of Al-maSdar nouns that are developed by the best automatic Arabic morphological analyzer so far 'Alkulil Morpho Sys' by KACST and ALECSO. The analyzer and its manual are open-source and can be downloaded for free[1]. In this appendix, the part of their manual that describes al-maSdar morphological forms is presented; we rely on this list on our annotation and development of the algorithms.

## المصدر الأصلي (Basic Al-masdar)

### 1 مصادر الفعل الثلاثي المجرد (Al-maSdar derived from 3-letter Verbs)

أوزان مصادر الفعل الثلاثي المجرد كثيرة جداً. وقد جمع بعض النحاة عدداً من هذه الأوزان، فرأوا أنها تنقاد لضوابط محددة، ولكنهم لم يزعموا أن القياس فيها تام مطّرد، بل يلجؤون إلى القياس على هذه الضوابط ما لم يَرد له سماع يخالفها. ورأى آخرون أن أوزان مصادر الفعل الثلاثي كلها سماعية.

يبين الجدول الآتي أكثر مصادر الأفعال الثلاثية دوراناً:

| أمثلة<br>Example | الوزن<br>The form | أمثلة<br>Example | الوزن<br>The form |
|---|---|---|---|
| فرح، أسف، وجع... | **فَعَل** | ضرب، نوم، عدّ... | **فَعْل** |
| شجاعة، فصاحة، كرامة... | **فَعَالة** | قدوم، صعود، لصوق... | **فُعُول** |
| إباء، فرار، جماح ... | **فِعَال** | بقاء، ثراء، جلال... | **فَعَال** |
| صهيل، حفيف، زفير... | **فَعِيل** | جولان، غليان، دوران... | **فَعَلان** |
| سهولة، خشونة، صعوبة... | **فُعُولة** | جولة، حسرة، رحمة... | **فَعْلَة** |
| سعال، دوار، زحار... | **فَعَال** | حُسْن، نُبْل، جُبْن... | **فُعْل** |
| زراعة، تجارة، صناعة... | **فِعَالَة** | حُمْرَة، صُفْرَة، زُرْقَة... | **فُعْلَة** |

### 2 مصادر الفعل الثلاثي المزيد (Al-maSdar derived from 3-letter Verb with extra letters)

تختلف مصادر الأفعال الثلاثية المزيدة عما سبقها (أي مصادر الأفعال الثلاثية المجردة) في أنها قياسية مطّردة، ذات أوزان معلومة، يندر الخروج عليها.

[1] http://www.econtent.org.sa/Projects/InitiativeProjects/Lists/InitiativeProjects/DispForm.aspx?ID=25

يبيِّن الجدول الآتي أوزان مصادر الأفعال الثلاثية المزيدة:

| Type | وزن الفعل Verb form | المضارع Present tense | المصدر The form | أمثلة Example |
|---|---|---|---|---|
| الثلاثي المزيد بحرف One letter extra | أفْعَلَ | يُفْعِلُ | إفْعَال<br>إفْعَلَة (معتل العين) | إكرام، إخراج، إنقاذ، إيمان...<br>إقامة، إعادة، إرادة، إشادة... |
| | فَعَّلَ | يُفَعِّلُ | تَفْعِيل<br>تَفْعِلَة (مُعَلّ اللام)<br>تَفْعِلَة (مهموز اللام) | تعليم، تدريب، تطويل، تبيين...<br>توصية، تسمية، ترقية، تغطية...<br>تخطئة، تبرئة، توطئة، تنشئة... |
| | فَاعَلَ | يُفَاعِلُ | مُفَاعَلَة<br>فِعَال (لغير المثال اليائي) | مجادلة، مبايعة، مقاومة، محادّة...<br>قتال، دفاع، نقاش، مراء، عداء... |
| الثلاثي المزيد بحرفين Two letter extra | انْفَعَلَ | يَنْفَعِلُ | انْفِعَال | انطلاق، انحدار، انقطاع، انهيار... |
| | افْتَعَلَ | يَفْتَعِلُ | افْتِعَال | احترام، استماع، اعتداء، احتواء... |
| | افْعَلَّ | يَفْعَلُّ | افْعِلال | احمرار، ابيضاض، ارتجاج... |
| | تَفَاعَلَ | يَتَفَاعَلُ | تَفَاعُل | تجاهُل، تدافُع، تداعٍ، تضامّ... |
| | تَفَعَّلَ | يَتَفَعَّلُ | تَفَعُّل | تعلُّم، تجوُّل، تغذٍّ، تولٍّ، تعلُّل... |
| الثلاثي المزيد بثلاثة أحرف 3 letters extra | اسْتَفْعَلَ | يَسْتَفْعِلُ | اسْتِفْعَال<br>اسْتِفْعَلَة (مُعَلّ العين) | استخراج، استفهام، استحمام...<br>استعاذة، استقالة، استقامة... |
| | افْعَوْعَلَ | يَفْعَوْعِلُ | افْعِيعَال | اعشيشاب، احديداب، اخليلاق... |
| | افْعَوَّلَ | يَفْعَوِّلُ | افْعِوَّال | اجلوّاذ، اخروّاط، اعلوّاط... |
| | افْعَالَّ | يَفْعَالُّ | افْعِيلال | احميرار، ابييضاض، اشهيباب... |

**3 مصادر الفعل الرباعي المجرد والمزيد**

**(Al-maSdar derived from a 4-letter Verb with extra letters)**

مصادر الأفعال الرباعية المجردة والمزيدة قياسية مطّردة، وفي الجدول الآتي أوزان مصادر الأفعال الرباعية المجردة والمزيدة:

| Verb type | وزن الفعل Verb form | المضارع Present tense | المصدر The form | أمثلة Example |
|---|---|---|---|---|
| الرباعي المجرد | فَعْلَلَ | يُفَعْلِلُ | فَعْلَلَة<br>فِعْلال (للمضاعف) | دحرجة، طمأنة، بسملة، زلزلة...<br>زلزال، قلقال، وِسواس، زعزاع.. |
| الرباعي | تفَعْلَلَ | يَتَفَعْلَلُ | تَفَعْلُل | تجلبُب، تبهرُج، تبعثُر، تزلزُل... |

| | | | | |
|---|---|---|---|---|
| **المزيد** | افْعَنْلَلَ | يَفْعَنْلِلُ | افْعِنْلال | افرنقاع، احرنجام، اسحنفار... |
| | افْعَلَلَّ | يَفْعَلِلُّ | افْعِلاَّل | اطمئنان، اشمئزاز، اقشعرار... |

**4 المصدر الميمي(al-maSdar starting with extra M)**

هو اسم يدل على الحدث، وأوله ميم زائدة، وليس على وزن (مُفاعَلَة)؛ نحو: **مَذْهَب، مَعْشَق، مَغْفِرَة، مَساءَة، مَحْيا، مَرَدٌّ.**

وهو كالمصدر الأصلي في معناه واستعماله، ولا يخالفه إلاّ في صورته اللفظية.

**1-2-3 صوغه من الفعل الثلاثي المجرد (derived from 3-letter verb)**

يصاغ المصدر الميمي من الفعل الثلاثي المجرد وفق ما يلي:

| الوزن The form | نوع الفعل Verb type | أمثلة Example |
|---|---|---|
| **مَفْعِل** | مثال واوي، صحيح اللام، تسقط فاؤه في المضارع | مَوْعِد، مَوْرِد، مَوْقِف، مَوْضِع، مَوْلِد، مَوْسِم، مَوْقِد، مَوْصِل، ... |
| | أجوف يائي، مكسور العين في المضارع | مَبِيع (أصله: مَبْيِع)، مَسِير، مَغِيب، مَجِيء، مَشِيب، مَصِير، مَقِيل، مَزِيد، مَبِيت... |
| **مَفْعَل** | ما عدا النوعين السابقين | مَطْلَع، مَدْخَل، مَقْتَل، مَوْجَل، مَتاب (أصله: مَتْوَب)، مَقال، مَمات، مَنْجَى، مَجْرَى، مَهْوَى، مَفَرّ (أصله: مَفْرَر)، مَسَدّ... |

ملاحظة: قد يكون المصدر الميمي على وزن (مَفْعَلَة)؛ نحو: **مَفْسَدَة، مَسْألة، مَبْخَلَة، مَجْبَنَة، مَسْغَبَة، مَيْسَرَة، مَوَدَّة، مَحَبَّة، مَذَلَّة، مَشَقَّة، مَنْجاة، مَهانَة، مَلامَة، مَخافَة، مَقالَة، مَساءَة...**

| أمثلة Example | المصدر The form | المضارع Present tense | وزن الفعل Verb form | Verb type |
|---|---|---|---|---|
| إكرامة<br>إقامة | إفْعَالَة<br>إفْعَلَة (معتل العين) | يُفْعِلُ | أفْعَلَ | **الثلاثي المزيد بحرف** |
| تكذيبة<br>توصية<br>تخطئة | تَفْعِيلَة<br>تَفْعِلَة (مُعَلّ اللام)<br>تَفْعِلَة (مهموز اللام) | يُفَعِّلُ | فَعَّلَ | |
| مبايعة | مُفَاعَلَة | يُفَاعِلُ | فَاعَلَ | |
| انطلاقة | انْفِعَالَة | يَنْفَعِلُ | انْفَعَلَ | **الثلاثي المزيد بحرفين** |
| استماعة | افْتِعَالَة | يَفْتَعِلُ | افْتَعَلَ | |
| ارتجاجة | افْعِلاَلَة | يَفْعَلُّ | افْعَلَّ | |

| | | | | |
|---|---|---|---|---|
| | تَفَاعَلَ | يَتَفَاعَلُ | تَفَاعُلَة | تواعدة |
| | تَفَعَّلَ | يَتَفَعَّلُ | تَفَعُّلَة | توحُّمة |
| **الثلاثي المزيد بثلاثة أحرف** | اسْتَفْعَلَ | يَسْتَفْعِلُ | اسْتِفْعَالَة<br>اسْتِفْعَلَة (مُعَلّ العين) | استخراجة<br>استجابة |
| | افْعَوْعَلَ | يَفْعَوْعِلُ | افْعِيعَالَة | احديدابة |
| | افْعَوَّلَ | يَفْعَوِّلُ | افْعِوَّالَة | اجلوّاذة |
| | افْعَالَّ | يَفْعَالُّ | افْعِيلاَلَة | ازويرارة |
| **الرباعي المجرد** | فَعْلَلَ | يُفَعْلِلُ | فَعْلَلَة | دحرجة |
| **الرباعي المزيد** | تفَعْلَلَ | يَتَفَعْلَلُ | تَفَعْلُلَة | تزلزلة |
| | افْعَنْلَلَ | يَفْعَنْلِلُ | افْعِنْلاَلَة | احرنجامة |

# Appendix B

**Appendix B: Arabic Discourse Annotation Scheme**

UNIVERSITY OF LEEDS

# Annotation Guidelines for Arabic Discourse Connectives

# in

# The Leeds Arabic Discourse Treebank project


Amal Al-Saif, Dr. Katja Markert

**School of Computing, University of Leeds**

&

Dr. Hussein Abdul-Raof

**Arabic & Middle Eastern Studies - University of Leeds**


**March 2009**

# Table of Contents

# 1 Introduction

## 1.1 Arabic NLP

Arabic is one of the most popular languages in the world. It is a Semitic language spoken by up to 246 million native speakers and it is the official language in 25 countries. Arabic is written as a right-to-left script with 28 basic Arabic letters and eight diacritical marks.

It has a complex root-based morphology. For example, several inflected forms can be derived from the consonantal root كتب/*ktb/write*. Each one indicates different grammatical features, such as number, gender and tense. Examples are the *verb* "to write" (كَتَبَ/kataba), "I wrote" ( كتبتُ/katab-tu), "you wrote"( كتبتَ/katab-ta, masculine singular), "you wrote" ( كتبتِ katab-ti feminine singular), "I write/will write" ( اكتبُ/Aktubu), and also *nouns* "books"( كُتُب /kutub) and "book" ( كتاب/ketab). Moreover, most Arabic processing applications require lemmatization or stemming to strip clitics and suffixes as pre-processing to produce the stem/root of words. The canonical order of Arabic sentences is VSO (verb–subject-object), but a range of other orders are possible in specific grammatical constructions.

Current NLP research on Arabic deals with many different language levels. For example, Arabic character recognition systems are the basic applications for Arabic at the character level. Morphological analysis, WordNet systems, tagging, stemming and spell checkers are the most common Arabic processing applications at the word level. Research at the sentence level has involved phrase chunking, sentence parsing and grammar checkers. In contrast, there is very little research on Arabic at the discourse level. This issue remains challenging for the Arabic NLP community. Al-Sanie and Seif and their colleagues (Seif, Mathkour and Touir 2005; Al-Sanie, Touir and Mathkour 2005) discussed a limited set of rhetorical relations and discourse connectives. Their studies had a small empirical basis using only a limited number of Arabic texts. Thus, building discourse annotated corpora for Arabic is necessary for advanced Arabic NLP as well as for linguistic purposes such as teaching/learning Arabic as foreign language by conducting comparative discourse studies with other languages.

## 1.2 Importance of discourse connectives

Discourse connectives have two distinct functions as distinguished by Cohen (1984): (i) enabling faster recognition of *discourse relations* by the reader (the hearer) and (ii) allowing the recognition of *discourse relations* which could not be inferred in the absence of a connective. Discourse connectives are widely studied in theoretical linguistics (Mann and Thompson 1987) (Hobbs 1985) (Fraser 1999) (Hovy and Maier 1993) (Marcus, Santorini and Marcinkiewicz 1993; Sanders 1992; Miltsakaki *et al.* 2006; Pitler *et al.* 2008). They explicitly indicate *discourse relations* between their arguments. The connective لأن/*because* in Example 1 establishes explicitly that the reason for Kald being absent from the party is that he was tired (Cause relation), whereas the connective *instead* in the third clause contrasts going to bed with going to the party (Contrast Relation). The connective *because* takes clause *cl1* and clause *cl2* as its arguments whereas *instead* takes clause *cl1* and clause *cl3* as its arguments.

(1)

[خالد لم يذهب الى الحفلة]cl1 **لأن** [ه كان متعبا]cl2 **بل** [ذهب الى الطبيب]cl3

Doctor to go but tired was he because party to go did-not Kald

Kald didnt go to the party,]$_{cl1}$ **because**[ he was tired.]$_{cl2}$ **Instead**, [he went to bed.]$_{cl3}$]]

D*iscourse relations* such as Contrast, Temporal and Cause relations do not have to be signalled explicitly using discourse connectives. In Example 2, the second sentence gives a potential reason for the event in the first sentence - a Cause relation between the two sentences holds. However, no explicit connective is present.

(2)

خالد لم يذهب الى الحفلة. لقد كان متعبا.

[tired was party to go did-not Kald]

Kald didn't go to the party. He was tired.

Our focus in this first version of discourse annotation is on annotating *discourse relations signalled by explicit connectives, ignoring discourse relations that are not signalled.* This makes sense as the usage of explicit connectives is very frequent in written Arabic, especially the connective و/wa is used very frequently. In addition, annotating discourse connectives automatically offers a wide range of applications in computational linguistics. For example, in automatic text generation, it is necessary to use the right connectives in the right places in the generated text. Moreover, for text summarization, text segments offering

mainly elaboration of related text segments might be ignored. Developing machine learning algorithms to recognize discourse relations and connectives requires a discourse corpus where all discourse connectives are annotated with associated relations and arguments.

There is no list of discourse connectives available for Arabic. Nor does a corpus exist where these connectives are annotated in context with regard to their discourse relations or arguments. The Leeds Arabic Discourse Treebank project aims to develop a large scale corpus annotated with information related to discourse structure.

I started the LADTB project by collecting a comprehensive list of discourse connectives for Arabic, using several linguistic and text analysis methods. The process yielded 107 potential discourse connectives and 17 possible discourse relations. We used similar annotation principles as the PDTB project for English (Prasad *et al.* 2007)**.** The motivation behind considering their annotation approach is that their principles are theory-neutral and have already been successfully adapted to other languages such as Chinese, Turkish and Hindi (Prasad *et al.* 2008; Zeyrek and Webber 2008; Oza *et al.* 2009). We believe using similar discourse annotation standards will benefit bilingual studies in linguistics and computational linguistics as well. In this manual, we will describe all annotation principles for Arabic regarding discourse connectives, discourse relations and arguments. All necessary adaptations were made to fit with the characteristics of Arabic.

## 1.3 The Penn Arabic Treebank

We annotate the Penn Arabic Treebank corpus Part1 v.2 (ATB), a parsed and tagged corpus of Modern Standard Arabic (MSA). It was released in January 2003 through the Linguistic Data Consortium (LDC) (Maamouri *et al.* 2004) and consists of 734 files with roughly 166K words of written Modern Standard Arabic newswire text from the Agence France Press (AFP). Although we annotate only the raw articles in the corpus to not confuse the annotators with syntactic annotation, the syntactic annotation in the ATB has been used for different tasks such as collecting potential discourse connectives that have the same Part-Of-Speech tag as known connectives.

## 1.4 Main tasks of discourse annotation

The discourse annotation process consists mainly of three tasks for each potential discourse connective (DC) in the corpus. All potential Dcs are highlighted in the annotation tool prior to annotation.

Task 1: Decide whether the potential DC does indeed have discourse usage in context. If so, do Task 2 and Task 3.

Task 2: Annotate the arguments Arg1 and Arg2 of the DC. Arguments are the text spans expressing Abstract Objects (Aos) related via the DC.

Task3: Assign suitable discourse relations from a pool of 17 pre-defined relations to the DC.

Annotation principles and definitions are described in detail in Chapter 2, 3, 4, 5 and 6. The annotation tool instructions are presented in Chapter 7

## 1.5 Notation conventions

Examples in the remainder of the manual obey the following conventions: (i) explicit discourse connectives are underlined (ii) the text span which is introduced by the discourse connective and expresses an AO is marked in bold (Arg2). (iii) The text span which expresses the first AO is marked in italics (Arg1). Punctuations should be excluded from the selection. The examples marked with a star are examples of potential Dcs without discourse usage in the particular context given.

Arabic examples are given a close-to-source translation to be read from right to left and indicated within square brackets as well as a freer standard English translation (to be read from left to right).

(3)

*سيفعل دور الحكومة الإنتقالية* <u>في حال</u> **انتصار الجيش الأمريكي في العراق**

[Iraq in American the-army win **i**<u>n-case</u> transitional the-government roll will-activated]

{*A transitional government will be activated* <u>if</u> **the American army wins in Iraq**}

# 2 Discourse annotation principles

## 2.1 Overview: explicit discourse connectives, arguments and discourse relations

As there is no standard definition of discourse connectives or markers in the literature, we follow the discourse annotation principles of the PDTB (Miltsakaki, Prasad et al. 2006). Thus, we define discourse connectives as *lexical expressions that relate two text segments that express abstract objects such as events, beliefs, facts or propositions.* We also use the same terminology, calling text segments that are linked via a DC *arguments* (Arg1 and Arg2). The link between the two arguments should represent specific *discourse relations*. Figure 2.1 summarises these concepts.

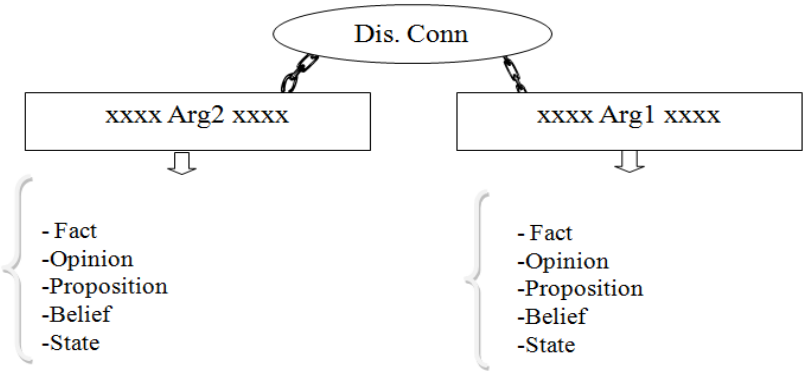



**Figure 2.2.1: Discourse annotation definition in the LADTB**

In Example 4, cl1 expresses an event that Jack gave Sarah a red rose, and cl2 expresses the writer's opinion. A causal relation is indicated by the connective *because th*at links cl1 (Arg1) and cl2 (Arg2). Although cl3 expresses a fact about the red colour and also gives a justification of the opinion in cl2, we do not consider this AO in our discourse annotation as an argument, because the relation is not indicated by an *explicit connective.*

(4)

[جاك أعطى سارة وردة حمراء]cl1.لأن[ه يحبها كثيرا]cl2. [اللون الأحمر يشير عادة الى الحب.]cl3

Love to often indicates red color. Much loves-her because-he.red rose Sarah gave Jack Jack gave Sarah a red rose]cl1. Because [he loves her so much]cl2. [The red colour often ] .indicates love]cl3

## 2.2 Order of discourse connectives and arguments

In Arabic, discourse connectives and their arguments follow different orders in texts. The two most frequent orders are <Arg1+DC+Arg2> and <DC+Arg2+Arg1>, which are mainly for simple connectives, i.e connectives consisting of adjacent lexical items only. Paired connectives are connectives which consist of non-adjacent lexical items, i.e. they have two parts DCP1 and DCP2. For paired connectives, only one order is possible, namely <DCP1+Arg2+DCP2+Arg1>. Figure 2.2 shows different orders of discourse connectives (DC) that relate two abstract objects (AOs) in Arg1 and Arg2.

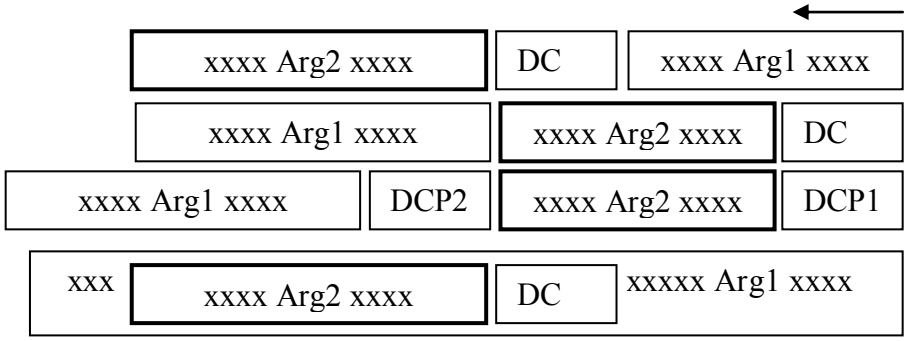



**Figure 2.2.2: Different orders of a discourse connective and its two arguments in Arabic text (to be read from right-to-left)**

Examples of the order <Arg1+DC+Arg2>:

(5)

*نصحه الطبيب أن يقلع عن التدخين*. و مع ذلك **استمر بالتدخين أكثر مما سبق**

[more in-smoking continued that with and .the-smoking of cease the-doctor advised previous than]

*The doctor advised him to cease smoking*. However, **he continued smoking more than** } {**before**

(6)

*ان قضية فلسطين ليست قضية اقليمية* او *وطنية* بل **مسالة تهم العالم الاسلامي اجمع**

[all Islamic the-world concern problem but national or regional issue not Palestine issue]

*The Palestine problem is not only a regional or national problem* but rather **a matter of** } {**concern to the entire Islamic world**

(7)

*تم رفض الخطة المقترحة للمشروع* لأن**ها غير مستوفية للشروط المتفق عليها**

[of-conditions compliant non because-it-is of-the-project the-proposed the-plan denied on agreed]

*The proposed plan of the project has been denied* because **it is non-compliant with the agreed terms**.}{

(8)

*تستطيع ان تذهب الى المنزل الآن* أو **تنتظرني لساعة واحدة**

[one for-hour wait-me or now home go you-can]

{*You can go home now* or **wait for me one hour**}

(9)

*أحمد يلعب كرة القدم*، و **مريم تقرأ كتاباً**

[a-book reads Mary and , football play Ahmad]

{*Ahmad is playing football*, and **Mary is reading a book**}

Examples of the order <DC+Arg2+Arg1>:

(10)

**بعد رحيلي عن القرية**، *لم اشعر بالسعادة مجدداً*

again in-happiness feel not, village from I-leave after

{After **I left home village**, *I never was happy again*}

Examples of the order <DCP1+Arg2+DCP2+Arg1>:

(11)

رغم **ان الطائرات كانت تحلق باستمرار في سماء المدينة** ، فإن *الحيا ة المدنية لم تتأثر*

affected not civilian the life **then**, city sky in in-continuous flying were planes although]

Although **the planes were flying continuously in the city sky,(**)** *civilian life was not affected*}{

Examples of the order < Arg1+DC+Arg2+Arg1>:

(12)

*استأنف الرئيس القبرصي اليوم الجمعة* **بعد انقطاع دام يومين**، *محادثاته غير المباشرة حول مستقبل جزيرة قبرص*

island future indirect talks two days lasting cutting after Friday today Cypriot President resumed

*The Cypriot President resumed on Friday,* **after** ***a lapse of two days****, the indirect talks on the future of the island of Cyprus*

# 3 Discourse connectives in Arabic

## 3.1 Syntactic categories of discourse connectives

Discourse connectives do not fall into a unique syntactic category. There are five main syntactic categories of discourse connectives in MSA: (i) coordinating conjunctions (ii) subordinating conjunctions, (iii) adverbials and prepositional phrases (iv) prepositions and (v) nouns. We have not noticed any significant differences in the behaviour of prepositional phrase connectives and adverbial connectives. Therefore, we deal with them as one category.

### 3.1.1 Coordinating conjunctions

Two independent clauses or sentences can be joined by a coordinating conjunction such as لكن/*but* , و/*and,* or أو/*or*. These conjunctions indicate discourse relations such as Contrast, Conjunction and Alternative as in Examples 12, 13 and 14 respectively.

(13)

*السيارة متطورة جدا* . لكنها **باهضة الثمن**

[ cost high but-it-is. very modern the-car ]

{*The car is very modern*. But **it is too expensive**.}

(14)

*أحمد يلعب كرة القدم*، **و مريم تقرأ كتاباً**

[a-book reads Mary **and ,** football play Ahmad]

{*Ahmad is playing footbal*l, and **Mary is reading a book**}

(15)

*تستطيع ان تذهب الى المنزل الآن* أو **تنتظرني لساعة واحدة**

[one for-hour wait-me or now home go you-can]

{*You can go home now* or **wait for me one hour**}

### 3.1.2 Subordinating conjunctions

Subordinating conjunctions introduce a clause that is syntactically dependent on the main clause. In Arabic, there are two kinds of subordinating conjunctions (similar to English, Chinese and Turkish):

### 3.1.2.1 Simple subordinating conjunctions

The subordinating clause is introduced by a subordinating conjunction such as لأن/because (see Example 7), بينما/while and حيث/since.

### 3.1.2.2 Paired subordinating conjunctions

Paired subordinating conjunctions consist of two non-adjacent lexical items: the first introduces the subordinate clause Arg2 and the other introduces the main clause Arg1. They are frequent in MSA. In Example 11 and 20, the paired connectives (... فان ورغم.. /*although/despite*, and (ف.. اذا .. /*if...then* indicate the discourse relations Contrast and Condition, respectively. The connective طالما/*as long as* indicates a Causal.Result/Condition relations in Example 17.

Note: M*ost paired connectives are translated to English with simple connectives.*

(16)

اذا **كان الجو صحواً** ، *فـلنلعب في الحديقة*

[the garden in let-us-play, clear atmosphere If

{If **the weather is fine**, *lets play in the garden*}

(17)

طالما **ان المؤتمر لم يحقق اهدافه** *فلن نجد من يثق بنتائجه لاحقا*

[later its-findings trust who find will-not its-objectives achieve not the-conference so-long-as]

{As long as *the conference has not achieved its objectives*, **nobody will trust its findings later**}

## 3.1.3 Adverbials and prepositional phrases

All sentence-modifying adverbials or prepositional phrases which express discourse relations between two abstract entities are discourse connectives. For example, the connectives لذلك /*therefore*, and (بالتالي/*consequently* often indicate a Result relation while نتيجة لـ/*as a result of* and بسبب/*because of* indicate a Reason relation, see Example 18. Theses connectives usually introduce Arg2.

(18)

*أن كبسولة الانقاذ لم تتمكن من الالتحام بالغواصة* بسبب **انعدام الرؤية**.

[ vision lack because-of to-the-submarine attaching from able non rescue capsule]

} t*he rescue capsule was unable to get attached to the submarine* because of **the lack of vision**}

### 3.1.4 Preposition connectives

There is a set of prepositions in Arabic that can relate AOs and indicate discourse relations. For example, the preposition *ل/due to/ in order to/for* in Example 19, often attached to AlmaSdar nouns. AlmaSdar nouns are a new argument category for Arabic, expressing AOs such as events, facts or propositions. More details about the AlMaSdar nouns are given in **Section xx.**

(19)

*ذهبنا الى مركز الشرطة* **للتبليغ عن فقدان وثائق الشركة الرسمية**

[the-official the-company documents loss of inform in-order-to police station to went]

*We went to the police station* for **informing about the loss of the company official documents**{ }

### 3.1.5 Noun connectives

Nouns in Arabic can function as discourse connectives. They occur as (i) simple nouns such as *بيد/byd/but, بغية/bgyp/desire/to* and *نتيجة/ntyjap/result*, or (ii) combined nouns with a preposition such as *فضلا عن/fdlA En/as well as*. The noun connectives *بغية/bgyp/desire/to* and *نتيجة/ntyjap/result* have also a semantic content themselves. See Examples: 20 and 21.

(20)

*كانت حياته مستقرة،* **بيد ان الظروف لم تسمح له ان يكون تاجرا**

[businessman be allow did-not circumstances **but** stable life was]

*His life was stable* **but circumstances did not allow him to be a businessman**

(21)

*قدم مراد طلب تقاعد مبكر* **بغية الاندماج في مشاريع الإصلاح**

[reformation projects in integrating in order to early retire request Murad apply]

*Murad put in an early retirement request* **in order to organise reformation projects**

## 3.2 Types of explicit discourse connectives

### 3.2.1 Simple Connectives

The simple connectives are discourse markers from any grammatical categories: coordinating/subordinating conjunctions, adverbials, or prepositions. They might be a single token (e.g. . رغم*although*, لأن*because*,بعدما*after* or a common conjunction و*and* ) or a phrase (such as some adverbials: على النقيض*in contrast*, نتيجة ل*as result of*, من جهة أخرى*besides*). Examples 14, 15 and 21.

### 3.2.2 Paired connectives

As mentioned above, some connectives consist of two parts. The first part of the connective introduces the first argument and the other introduces the second argument. They fall into one syntactic category, subordinating conjunctions. Examples: 16 and 17.

Note: Some paired connectives are not translated as paired connectives in English, see Example 22.

(22)

ما لبثت **المحاضرة أن تبدأ** حتى *دخل الطلاب في نوم عميق*

[sleeping deep in the-students enter then began the-lecture Once ]

{ Once **the lecture began**,(xx) *all students fell into a deep sleep* }

### 3.2.3 Clitic connectives

Almost preposition discourse connectives are clitics. The clitic connectives are attached to tokens such as almaSdar nouns, pronouns and verbs. Examples: 7, 13, and 19.

### 3.2.4 Modified connectives

Connectives might be modified by attaching lexical items expressing additional semantic/pragmatic meaning on top of the meaning of the connective. For example:

1) The connective is connected with non-pronoun clitics such as بالرغم من

2) The connective occurs always with function words such as أن/**that** for an emphasizing purpose or adverbs such as (أيضاً/also) ( **حتى لو**/even if) to add extra semantic information.

These modified connectives share the main features of the head connective: position, discourse relation and arguments. The second token here could not relate the arguments alone. We annotate modified connectives as one connective.

In Example 23, the temporal connective (بعد/ after is modified by clitic (ما) to generate a modified connective (بعدما/ after) which behaves exactly the same as the head connective (بعد/ after).

(23)

*اتخذ الشقيقان قرارهما* **بعدما وجدا تجاهلا تاما من قبل ادارة الكرة في نادي الاهلي**

[administration from complete ignoring found they **after** their-decision the-brothers made Alahli club in the-football]

{The brothers made their decision **after** they were completely disregarded by the football department in the Alahli Club}

### 3.2.5 Multiple connectives

In contrast, we do not consider any token that indicates a different discourse relation than the head connective does as a modified form of that connective. The two connectives are multiple connectives. If they relate different arguments, they should be annotated separately as new connectives. However, if the multiple connectives relate the same arguments, they should be considered as new connective. In Example 24, two connectives appear next to each other (**الا بعد**/*except after*) sharing exactly the same arguments and were annotated as one connective. The new connective indicates Exception/Temporal.Asynchronous relations.

(24)

*لم تشعر ليلى بطعم الراحة* الا بعد **ان سمعت خبر عودة ابيها من السفر**

[travel from her-father back news heard after except relax taste Laila Feel not]

{*Laila did not feel relax* except after **she heard news that her father is back from travel**}

# 4 Associated arguments

Each lexical expression/text span (whatever its length: clause, sentence and multiple sentences) that expresses one or more abstract objects is possible as an argument of a discourse connective. Arguments should include all complements necessary to understand the AO completely.

## 4.1 Adjacent and non-adjacent arguments

While the connective introduces Arg2, Arg1 might occur (i) in the same sentence as the connective occurs, such as in Example 7 and also all examples of paired connectives, (ii) in the previous sentence such as in Example 20 (iii) in previous non-adjacent sentences such as in Example 25 or (iv) in sentences following the sentence containing the connective and Arg2 such as in Example 10.

(25)

*لم يحضر أحمد الإمتحان* لأنه كان يعاني من ألم في معدته **بل ذهب الى المستشفى لإجراء الفحوصات**

his-stomach in pain from suffering was because-he the-exam Ahmad attend not examinations to-get the-hospital to went instead

*Ahmad did not attend the exam* because he had stomach pains. Instead **he went to hospital** } **{ to get examinations**

## 4.2 Types of arguments

The arguments of discourse connectives in Arabic can be simple clauses/sentences or a sequence of them, VP coordinations, almaSdar nouns or anaphoric expressions denoting abstract objects.

### 4.2.1 Simple clauses and sentences (or sequences of sentences )

Arabic sentences are divided in traditional Arabic grammatical theory into two categories: *jumla ismeia* nominal/equational/verb-less sentences *and jumla filia* verbal sentences depending on the nature of the first word in the sentence. The verbal sentences (verb, subject

and object) are definitely expressing an abstract object. One or more verb sentences can be annotated as arguments for a discourse connective such as Arg1 in Example 26.

(26)

تفاجأت أم خالد لما سمعت خبر وفاة ابنها خالد و *سقطت مغميا عليها* لذا **نقلوها الى أقرب مستشفى كحالة طارئة**

fell and Khaled her-son death news heard she when Khaled Mum surprised]

[emergency as-situation nearest to carried-off so dizzy

{Khaleds Mum was surprised when she heard the news of the death of her son Khaled and *fell dizzy*. Therefore, **she has been carried off to the nearest hospital as an emergency situation**.}

Equational sentences (subject and predicate) often express abstract objects as well (state, fact or belief). The following examples are from (Ryding 2005).

| | |
|---|---|
| Noun– adjective phrase: | **العالم قرية صغيرة** |
| | **The word is a small village** |
| Noun phrase – adjective: | **قصر الملك ضخم** |
| | **The kings palace is huge** |
| Pronoun – adjective phrase: | **انت صديقي** |
| | **You are my friend** |
| Demonstrative pronoun- noun: | **هذه تجربة مهمة** |
| | **This is an important experiment** |
| Noun – noun phrase: | **الزراعة لغة عالمية** |
| | **Agriculture is a world language** |
| Clause – equational sentence: | ***المسيحية و الاسلام*** **أصلهما واحد** |
| | ***Christianity and Islam*** **are from one source** |
| Negation of verbless sentences: | **ليست صديقتنا** |
| | **She is not our friend** |
| equational sentences (with Kan): | **Past: كان قصر الملك ضخم** |
| | **The king palace was huge** |
| | **Future: ستكون زوجتي طبيبة** |
| | **My wife will be a doctor** |
| Expression of possession | **عندي مشكلة** |
| (Predicate – subject ) | **I have a problem** |
| Existential predication (there is/are) - (Predicate – subject ) | **هناك عوامل كثيرة** |
| | **There are many factors** |

The subordinate clause (Arg2) in Example 27 has an equational clause structure (noun – adjective) which represents the cause for removing the building.

(27)

*تم ازالة مبنى البلدية في وسط المدينة* لأن **المبنى متهالك و آيل للسقوط في أي لحظة**

[The-building because The-town middle in municipal building removing finished time any in fall could-be and old]

*The municipal building in the town was removed* because **the building is old and it could** } { **fall at any moment**

### 4.2.2 Verb ellipsis

Verb ellipsis is defined in Wiktionary in the following way: *"To* remove *a verb from a phrase which is* grammatically *needed, but which is clearly understood without having to be stated"*[1]. Sometimes verb ellipsis is an essential process to avoid redundancy in the writing. The verb usually appears in prior discourse. Therefore, the clause involving verb ellipsis is usually considered as the second argument. Examples 28 and 29 show cases of verb ellipsis as arguments of a DC.

(28)

*سجل عبدالله الجمعان هدفي الهلال*، و**محمد مصطفى هدفي جبلة**

Jebelah goal Mustafa Mohamad and, Alhelal golas JamaanAbdulAlah record

*Abdullah Jumaan recorded two goals for Alhilal*, and **Mohamed Mustafa two goals for Jebelah**.

(29)

*اشرك المدرب الروماني بيلاتشي النيجيري مانجوت بدلا من السنغالي داين فاين*، ثم **الكاتو بدلا من محمد الشلهوب**

*Romanian coach Bilache replaced the Nigerian Manjut instead of the Senegalese Dane Fine*, then **Alcato instead of Mohammad Al Shlhoub**

### 4.2.3 Al-maSdar nouns

Al-maSdar is a noun denoting an action/state without indicating tense. They are derived from corresponding verbs. For example, *وصول/arrival* is a noun derived from the verb وصل/to arrive and محاولة/attempt is a noun derived from the verb *حاول/to try*. In the Arabic grammatical tradition, this noun category is well-defined with at least 60 common morphological patterns of al-maSdar[2]. Al-maSdar nouns do not fit into one grammatical category in English; they might correspond to a gerund (*swimming*), a nominalization (*reflection)* or a noun not normalization (Wolf *et al.*). Table 1 displays several Al-masdar nouns, the patterns with which they are derived and an English translation.

---

1 http://en.wiktionary.org/wiki/ellipse

2 Some linguistics argue that there is an unpredictable list of morphological patterns of al-maSdar M. ABDL AL LATIF, A. U., M. ZAHRAN and D. A. AL-ARABI. 1997. *Alnhw AlAsAsi.* CSLI..

**Table 4-1: al-maSdar examples with corresponding morphological pattern and English equivalent**

| Root | Morph. Pattern | Al-maSdar noun | English gloss |
|---|---|---|---|
| سبح/sbh | فعالة | سباحة | swimming |
| عكس/eks | انفعال | انعكاس | reflection |
| جرب/jrb | تفعلة | تجربة | experiment |
| حَرَبَ/hrb | فَعْلْ | حَرْبْ | war |
| دفع/dfe | فعال | دفاع | defence |

Al-maSdar noun can be considered on its own or with a clause's complements as an argument of a DC. Al-maSdar nouns frequently express an event after prepositions. In Example 30, we consider the clause ( **إجتياح فيضانات عنيفة شملت البلاد مؤخرا**/*strong flooding over the country recently* as the Arg2 of the connective نتيجة ل/*as a result of* where the stem of a head noun اجتياحis an al-maSdar noun using the pattern افتعال. The morphological patterns of al-maSdar are listed in Appendix B**.** Examples 31 and 32 are further examples with al-Masdar as DC arguments.

(30)

*تم ازالة مبنى البلدية القديم* الذي يعد أقدم مبنى حكومي في المدينة نتيجة ل **اجتياح فيضانات عنيفة شملت البلاد مؤخراً**

*The old municipal building was removed*, which is the oldest governmental building in the city as a result of **the strong flooding over the country recently**

(31)

قررنا تأجيل رحلة الصيد **بسبب** المطر

the-rain **because-of** hunting trip postpone to we-decided

{We have decided to postpone the hunting trip **because of** the rain}

(32)

*نجح مصطفى في ترجمة ركلة الجزاء الثانية لفريقه* **اثر خطأ من فهد المفرج**

*Mustafa succeeded in converting a second kick penalty for his team* after **a mistake by Fahd Almofreej**

### 4.2.4 Anaphoric expressions denoting abstract objects

Anaphoric expressions can be annotated as arguments of DCs as long as their antecedent is an abstract object. Therefore, anaphoric expressions such as ذلك/that in Example 33 which refers to (الاستيلاء على شاحنة البترول/ *stealing of oil truck* is annotated as Arg2 of the connective بعد/ after.

(33)

استولى عدد من االإرهابيين على شاحنة لنقل البترول. **بعد ذلك** *وضع المهاجمون الشاحنة في منتصف الطريق لوقف حركة السير*

A number of terrorists have stolen a truck for transporting oil, after *that* **they placed the truck in the middle of the road to stop traffic and then killed three people**.

## 4.3 What can not be considered as an Argument?

### 4.3.1 Conjunction of simple verbs and nouns

We do not assume the conjunction of simple verbs, nouns, proper nouns, adjectives and prepositional phrases as arguments for DCs such as in Examples 34 to 38.

(34) *- verbs:

رأيت الأطفال يلعبون **و** يصرخون في حديقة المستشفى

[ the-hospital garden in shouting **and** playing children I-have-seen]

I have seen children playing **and** shouting in the hospital garden

(35) *– prepositional phrases:

ذهبت **الى المكتبة** ثم **الى المدرسة** ثم **الى الحديقة**

I went **to the library** and then **to school** and then **to the park**

(36) * - nouns:

ذهب أحمد **و** فاطمة إلى سوق المجوهرات لشراء هدية لأمهما

[for-their-mum gift buy to jewelry shop to Fatima and Ahmad went]

{Ahmad **and** Fatima went to jewelry shop to buy a gift for their mum}

(37) * - adjectives:

الرياض مدينة كبيرة **و** جميلة في المملكة العربية السعودية

beautiful **and** large city Riyadh

Riyadh is a large **and** beautiful city in Saudi Arabia

(38) *-adverbs:

حضر المحامي الى قاعة المحكمة مسرعاً **و** مرتبكاً

nervously **and** quickly the-court room to the-lawyer came

The lawyer came to the court room quickly **and** nervously

### 4.3.2 Relative clause الذين/الذي/التي/... **who/ that/which**

We establish rules for three possible cases of relative clauses.

a) A relative clause that is introduced by a connective should be considered as an argument of entity relation which is annotated in our scheme with a Conjunction relation.

b) A relative clause that is not introduced by a connective but is a necessary complement clause to an argument *a,* should not be considered as an argument on its own but should be included in the argument *a.* I*n* Example 39 the visiting event includes the relative clause *which was built in 1985*.

(39)

زرنا متحف فكتوريا، الذي بني عام 1985، مع بعض الزملاء أمس

Yesterday colleagues some with, 1985 year built which Victoria museum we-visit

*We visited a Victorian museum, which was built in 1985, with our colleagues yesterday.*

c) Both the arguments and the discourse connective are parts of a relative clause. The relative pronoun التي/الذي/الذين**/...** */who/that/which* should not be included within the argument spans (see Example 40).

(40)

تمثل قضية مستقبل القدس الشرقية التي *احتلتها اسرائيل عام 1967* و **تشرد ابنائها الأصليين** مستقبل الفلسطينين

The future of East Jerusalem, which *Israel occupied in 1967* **and her native people were vagabond**, represents the future of all Palestinians

### 4.3.3 Attribution

The proposed discourse annotation does not consider attribution relations. However, some connectives are ambiguous; they can be used as discourse connectives in some instances, and signal attribution in other instances, such as كما ذكر الدكتور جاك *as Dr. Jack said*. Thus, distinguishing between them is essential.

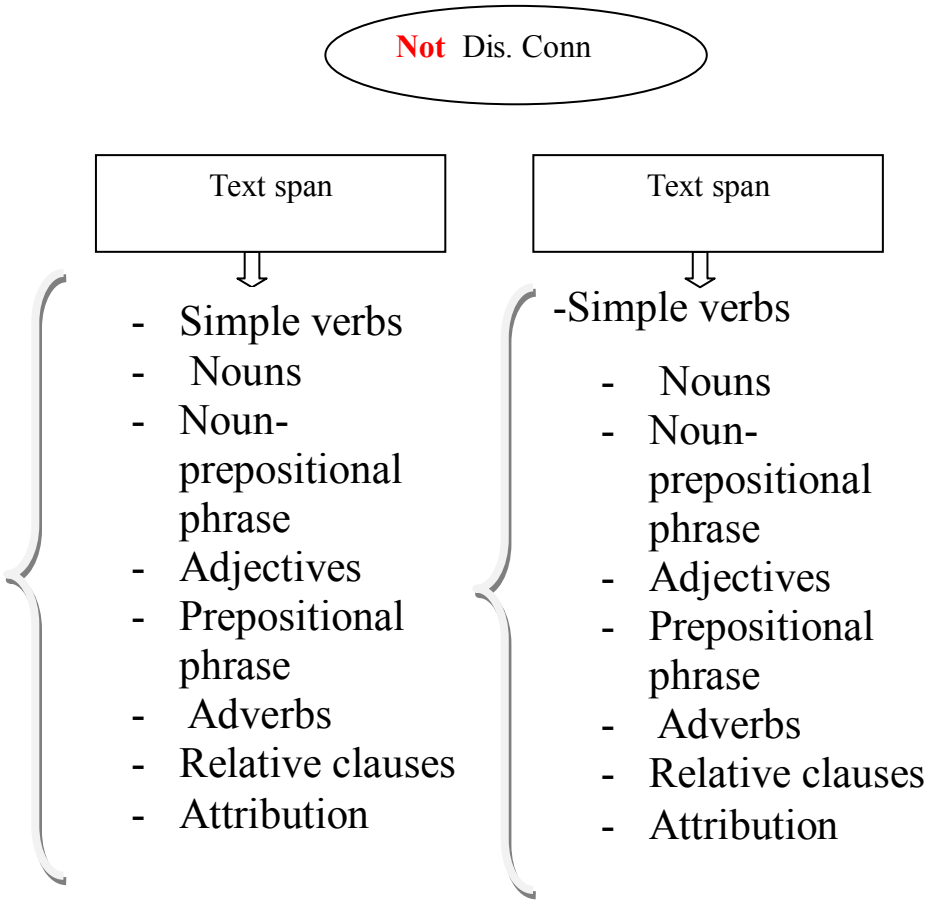



**Figure 4.1 : A summary of text spans that cannot be arguments linked by a discourse relation**

## 4.4 The minimality principle

Each argument should be coherent, that is to say *include all critical parts that play a role in expressing the complete abstract object but not any additional information.* This is called the *mimimality principle* in the PDTB annotation guidelines and we adopt it for Arabic. We should consider only the minimal interpretation of a relation when annotating its arguments including complements such as temporal adverbs, relative clauses, prepositional phrases Example 41 shows that Arg1 is not only (three people were injured), but should include two complements ( *ليلة البارحة - ممن كانوا يقفون قرب الحادث/ who were standing near the accident - last night*).

(41)

**تم اصابة ثلاث أشخاص ممن كانوا يقفون قرب الحادث ليلة البارحة** مما **أدى إلى تسبب فوضى شاملة في المنطقة**

region in massive mess causing to led which last night accident near standing persons **]**
**[**three injured

*Three people were injured who were standing near the accident last night*. Thus **a** }
{**massive mess was caused in the region**

# 5 Discourse relations

One of the main concerns in discourse annotation is identifying the discourse relations between arguments that are connected explicitly by discourse connectives. These discourse relations can be indicated by more than one explicit connective. Similarly, a discourse connective might indicate more than one discourse relation. Thus, we have a many-to-many relationship!

## 5.1 Hierarchy of discourse relations

The relation hierarchy in the PDTB for English (Prasad, Dinesh et al. 2008) and all related schemes for other languages have advantages over a flat list of discourse relations. The hierarchical structure allows for more flexible annotation as the annotator has the right to choose one or more discourse relations for a DC at any level in the hierarchy. For example, if the discourse relation of the connective is hard to be recognized at the type or subtype levels, the annotator can just choose the equivalent discourse relation from the class level. This can also increase reliability of annotation as it allows backoff to a higher level. The hierarchy also makes it easy to insert/delete a discourse relation at any level or to compress/merge relations.

Therefore, we preferred using a hierarchy of discourse relations to represent our relations taxonomy for Arabic. We have built the taxonomy in two steps: first, our discourse analysis of more than 60 Arabic articles resulted in a list of discourse relations and examples using our own terminology and definitions. Second, we then mapped this list onto the PDTB relation hierarchy. We kept only the relations that have been recognized for Arabic, modifying definitions slightly as required. In addition, we do not annotate some of the very fine-grained relations in the PDTB in this annotation exercise. We also added two new discourse relations.

We use the same top level, *class level,* as the PDTB,which consists of the relations TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. Each class has several *types* and further *subtypes* expressing more fine-grained relations. Figure 5.4 shows our discourse relations hierarchy.

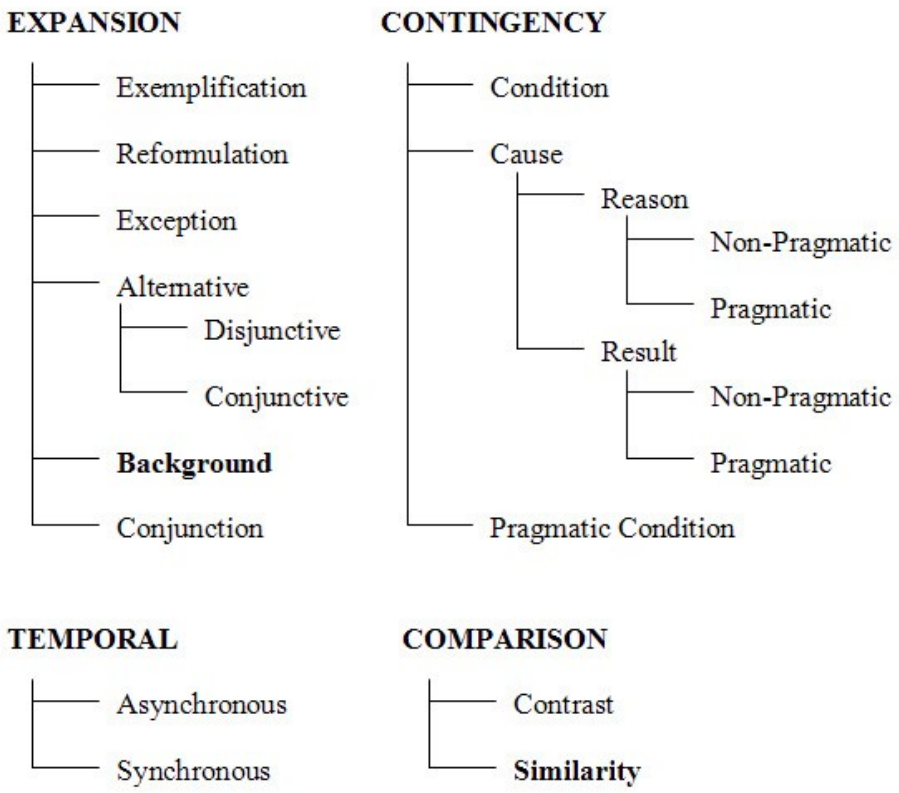



**Figure 5.1: The hierarchy of discourse relations for Arabic**

## 5.2 Discourse relations descriptions

We will specify for each relation whether the relation description is exactly the same as the corresponding PDTB relation (SAME_as_PDTB), has been slightly changed (ADAPTED_from_PDTB) or is completely new/different (NEW).

### 5.2.1 Class: "TEMPORAL"

The tag TEMPORAL is used when the connective indicates that the abstract objects described in the arguments are related temporally. There are two types of TEMPORAL relations (SAME_as_PDTB).

#### 5.2.1.1 Type: "Asynchronous"

The tag *Asynchronous* is used when the situations described in the two arguments are temporally ordered. One of the events happened before/after the other. Typical connectives are قبل/*before* and بعد/*after*.

(42)

تم افتتاح المعرض الثقافي العالمي يوم الأحد بعد **وصول جميع ممثلين اللجان الثقافية في البلدان المشاركة**

[representatives all arrival after Sunday day international cultural exhibition opened participating countries in cultural committees]

*The international cultural exhibition was opened on Sunday* after **the arrival of all {cultural committees representatives from participating countries}**

#### 5.2.1.2 Type: "Synchronous"

The tag *Synchronous* applies when the situations described in Arg1 and Arg2 overlap temporally.

(43)

وصلنا المسجد لصلاة الجمعة اذ **بالمصلين يخرجون من المسجد**

[the-mosque from leaving prayers when Friday for-praying the-mosque arrive]

*{We arrived at the mosque for Friday prayers* when **prayers were leaving the mosque}**

#### 5.2.1.3 Synchronous or Asynchronous:

The length of the event plays a role in distinguishing between the two temporal relations. In Example 44, the start of the clashes is an event that happened at a specific point in the time line. We focus here on the start of the clashes and not the clashes themselves. Thus, the connective ( **منذ**/*since)* indicates an Asynchronous relation.

(44)

اعربت عن قلقها للاستخدام المفرط للقوة من قبل اسرائيل منذ **بدء المواجهات في 28 ايلول/سبتمبر**

She expressed concern at *the excessive use of force by Israel* since **the start of the clashes on September 28**

### 5.2.2 Class: "CONTINGENCY"

The class level tag "CONTINGENCY" is used when one of the Aos described in Arg1 and Arg2 causally influences the other.

#### 5.2.2.1 Type: "Cause"

The type *Cause* is used when one of the situations described in Arg1 and Arg2 causally influences the other and the two **are not in a conditional relation.** The directionality of causality is not specified at this level: when "Cause" is used in the annotation, it means that

the annotators could not uniquely specify its directionality. The directionality is specified depending on the situation in Arg2 and the temporal order. The two subtypes might be pragmatic relations as well (ADAPTED_from_PDTB).

#### 5.2.2.2 Subtype: Reason

The subtype *Reason* is used when the situation described in Arg2 is the cause and the situation described in Arg1 is the effect. Example 45.

(45)

*بلغت استراليا الدور النهائي لمسابقة كأس ديفيس لكرة المضرب* **بتقدمها على البرازيل 3-صفر**

*Australia reached the final round of the Davis Cup Tennis Tournament* because of **her progress against Brazil 3 – zero**

The situation in Arg2 might be a direct reason (CONTINGENCY.Cause.Reason. NonPragmatic) or an indirect reason that provides a *justification or evidence* for the claim in Arg1 (CONTINGENCY.Cause.Reason.Pragmatic). For example, the speed cameras in Example 46 do not cause the withdrawal of driving licenses but are used to detect speed violations, which cause the withdrawal. Similarly, in Example 47, Arg2 ( لقد شاهده العاملون يزور في الأرقام/*the workers saw him updating the accounting figures*) justifies the sentence of the project accountant in Arg1.

(46)

*تم سحب رخص القيادة من 34 سائقاً الاسبوع الماضي* حيث **تم استخدام كامرات مراقبة السرعة للتعرف على مستوى السرعة الغير قانونية**

[monitoring cameras used was since last week drivers 34 from driving licences withdraw legal non speed level on identify to speed]

*Driving licences were withdrawn from 34 drivers last week*, as **speed cameras were used** } {**to identify the level of illegal speed**

(47)

*حكم على محاسب المشروع بالسجن 3 سنوات بتهمة التلاعب* فقد **شاهده العاملون يزور في الأرقام**

[workers saw as cheating on-charges years 3 prison in project accountant sentenced figures in updating]

*A project accountant was sentenced to 3 years in prison on charges of cheating* as **the** *}* **{workers saw him updating the accounting figures**

#### 5.2.2.3 Subtype: Result

The subtype *Result* applies when the situation in Arg2 is the effect brought about by the situation described in Arg1.

(48)

*في كرة القدم لكل مباراة ظروفها وحساباتها الخاصة* وبالتالي **يصعب التكهن بما سيحصل غدا**

[consequently and special calculations and its-circumstance match for-each football in tomorrow happen will what predicting difficult]

I*n football, each match has its own circumstance and calculations* so **it is difficult to** } **{predict what will happen tomorrow**

(49)

*ارتدت الرصاصات على الحصى* **و اصيب ابو غيدا بشظاياها**

*The Bullets ricocheted on the gravel*. (and) **Abu Ghida was injured with fragments**

(50)

*اصيب ابو غيدا بشظاياها* **مما استدعى نقله الى المستشفى**

*Abu Ghida was injured with fragments*. As a result, **he was rushed to a hospital**

Similar to reason relations, the situation in Arg2 might be a direct result (CONTINGENCY.Cause.Result.NonPragmatic) or indirect result (CONTINGENCY.Cause.Result.Pragmatic) for a *justification* in Arg1. For example, in Example 51, Arg1 ('there are no diplomatic relations with Israel') is a justification for the result in Arg2. Also, confirming the break team in Example 52 is not a direct result of *the violence in the team* in Arg1.

(51)

*لا تقيم اندونيسيا علاقات دبلوماسية مع اسرائيل* و **يبدو ان التكتم الذي احاط بزيارة بيريز الى اندونيسيا كان يهدف الى تفادي اثارة ردود فعل معادية في البلاد**

*Indonesia does not maintain diplomatic relations with Israel* and **it seems that the secrecy surrounding Peres visit to Indonesia was aimed at avoiding negative reactions in the country**

(52)

*جاء اعتداء هشام حنفي على زميله شادي على مراي و مسمع الجميع اثناءالمباراة* **ليؤكد تفكك الفريق**

*The violence attack by Hesham on his colleague Shadi in front of all audience during the match, happened to* **confirm the team breaking**

**Note:**

Cause relations (Reason/Result) implicitly indicate a temporal relation. Generally, the cause happens before the result. There is no need to specify this temporal relation explicitly unless the discourse connective is a temporal connective in the first place, such as the connective (after/بعد) in Example 53. In this case, the relation to be annotated is a combination of CONTINGENCY.Cause.Reason.NonPragmatic and TEMPORAL.Asynchronous.

(53)

*شعرت ليلى بسعادة غامرة* بعد **ان سمعت خبر عودة ابيها من الحج**

[Hajj from her-father back news heard after except relax taste Laila Feel]

*{Laila felt extremely happy* after **she heard the news that her was father back from the Hajj}**

### 5.2.2.4 Type: Condition

The tag *Condition* is used when the situation in Arg2 is taken to be the condition and the situation described in Arg1 is taken to be the consequence. (ADAPTED_from_PDTB). Examples 54, 55 and 56.

(54)

*سوف تمنح جائزة أفضل مشروع* إذا **تم استكمال تقييم جميع المشاريع المقدمة**

the-proposed the-projects evaluation completing finish when project best prize awarded ] [will

*A prize will be awarded for the best project* once **the evaluation is completed for all** } {**proposed projects**

(55)

إذا **كانت الحديقة نظيفة** *نستطيع إقامة حفلة شواء هذه الليلة*

[night this barbeque party establish can the-garden cleaned be If]

{If **the garden is cleaned** ***we can make the barbeque party this night***}

(56)

*قد سمح للجنود الاسرائيليين باطلاق رصاص حي* اذا **شعروا انهم معرضون للخطر**

*Israeli soldiers are permitted to fire real bullets* if **they feel they are in danger**

### 5.2.2.5 Type: Pragmatic Condition

The tag *pragmatic condition* is used for instances of conditional constructions whose interpretation deviates from that of the semantics of *Condition*, specifically, when a condition-indicating connective such as اذا/if is used but Arg1 and Arg2 are not causally

related (SAME_as_PDTB). In these cases, Arg1 holds true independently of Arg2. The box of biscuit in Example 57 is on the kitchen table whether the second speaker enters the kitchen or not.

(57)

إذا **دخلت المطبخ** *هناك علبة بسكويت على الطاولة*

table on biscuit box there the kitchen enter If

If *you get in the kitchen*, **there is a box of biscuits on the table**

(58)

إذا **أردت ان تتحدث عن قظية السلام في الشرق الأوسط فـ***الحرب مع اسرائيل من أبرز النقاط الشائكة*

[is Israel with the-war then the-Middle the-East in peace issue about talk want If

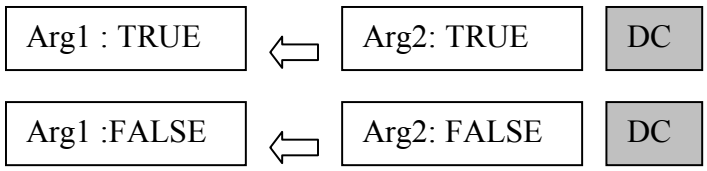
shocking points obvious one ]

If **you want to talk about impure peace in the Middle East**, *the war with Israel is one of* }

{*the most obvious issues*

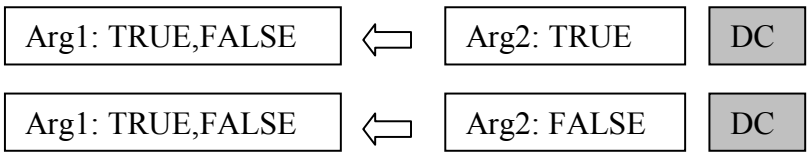
### 5.2.2.6 Condition v. Pragmatic condition

We distinguish among conditional and pragmatic conditional relations using the truth values of both arguments. A Condition relation is considered when the truth of Arg2 affects the truth of Arg1, see the diagram Fig 5.2 (a). In contrast, a pragmatic condition relation is indicated by explicit conditional connectives but there is no clear direct causal relation between Arg1 and Arg2. For example, Arg2 can be true while Arg1 is not, see Fig 5.2 (b).




(A) *ان اجلاء الطاقم سيبدأ بعد الظهر* إذا **تأكد تحسن الاحوال الجوية**

*The evacuation of the crew will happen this afternoon* if **weather conditions improve**




(B) إذا **دخلت المطبخ** *هناك علبة بسكويت على الطاولة*

If *you get in the kitchen*, **there is a box of biscuit on the table**

**Figure 5.2: NonPragmatic (A) and Pragmatic (B) Condition relations**

In general, discourse relations are *pragmatic* when there is no clear direct relationship between Arg1 and Arg2. However, the reader can infer an indirect relation between the arguments such as indirect Cause or Condition. They are signaled either by:

- Explicit connectives which are typically used to indicate a clear discourse relation

- Flexible connectives which can indicate any relations in context such as و/wa (rarely).

### 5.2.3 Class: COMPARISON

The class tag COMPARISON applies when a discourse relation is established between Arg1 and Arg2 in order to highlight prominent differences or similarities between the two AOs. There are two relations here Contrast and Similarity.

#### 5.2.3.1 Type: Contrast

The relation *Contrast* applies when Arg1 and Arg2 share a predicate or property but one or more differences are highlighted in the text. Such differences can be, for example, with respect to an *expectation* as in Example 59 or values assigned to *a shared property* as in Example 60. (SAME_as_PDTB)

(59)

*نجح أحمد في الإمتحانات* بينما **توقع المدرس فشله**

*Ahmad succeeded in the exam* while **his teacher expected him to fail**

(60)

*ارتفعت اسعار النفط في الربع الأخير الى 146 دولارا للبرميل* لكن **ايرادات شركات النفط حققت تراجعا 12% خلال نفس الفترة**

oil companies revenues but for barrel $ 146 to last the-quarter in Oil prices rose]

[the-period same during 12% declining have

*{Oil prices rose in the fourth quarter to $ 146 a barrel*, but **oil companies revenues declined 12% over the same period}**

(61)

يعد الفلم ناجحا حتى و **ان كانت مبيعاته العالمية قليلة**

*The film is successful* even though **the global sales are few**

(62)

واضاف الدبلوماسي الغربي *ستبقى قوات الطوارىء الدولية كما عهدناها حتى مطلع صيف العام* **2001**. مع ان **اي قرار بهذا الشان لم يتخذ حتى الان** بانتظار انتهاء قمة الامم المتحدة

The western diplomat added "*the international peacekeeping forces will stay as usual until the beginning of summer 2001*. **However, any decision in this regard has not been taken so far,** waiting for the end of the United Nations summit

Contrast relation applies also when the situation in Arg2 is not directly influenced by the situation in Arg1 but a typical contrast connective such as (الا ان/ لكن/غير ان/but/however). Is present (see Example 63). In the PDTB, a type *pragmatic contrast* is used for such cases, but we do not distinguish between pragmatic and other contrasts.

(63)

*اتمنى ان ارى المشروع ناجحا* و لكن **يجب عليكم ان تجذبوا السياح الأجانب**

[foreign tourists attract to you must but and succeed project see to hope]

{*I hope to see the project successful*, but **you must attract foreign tourists**}

### COMPARISON.Similarity

The type *Similarity* applies when the connective indicates that the two arguments express similar abstract objects. It is therefore a complement to the contrast relation (NEW). The two arguments in Example 64 are presenting a similar action in the way of giving a present to others.

(64)

*انك تتألم من فراق الوطن* **كما تتألم الأم على فقد رضيعها**

[her-child losing from mum suffer as home- country leaving from suffer You]

*You are suffering from leaving your home country* **as a mother suffers from losing her child**

## 5.2.4 Class: EXPANSION

The class tag “EXPANSION” applies when Arg2 expands or gives more details about the situation in Arg1. The extra information can be classified according to the following types.

### 5.2.4.1 Type: Exemplification

The tag *Exemplification* is used when Arg1 evokes a set and Arg2 exemplifies Arg1 and describes it in further detail (SAME_as_PDTB). For example (*الاحتياطات الأمنية/ safety regulations*) in Example 65 is a set of behaviours and (ربط الحزام/fasten the belt) is one instance of following safety regulations.

(65)

*احرص على أخذ جميع الاحتياطات الأمنية أثناء السفر بالطائرة* كأن **تربط حزام الأمان طيلة الرحلة**

[fastening for-example by-plane travelling during safety protection all taking on aware the-flight during safety seatbelt]

*Make sure that you follow all necessary safety regulations when you travel by plane* **for example** *}* **{ fasten your belt during the flight**

### 5.2.4.2 Type: Reformulation

A connective is marked as *Reformulation* when Arg2 mainly restates the content of Arg1. It could be that (i) Arg2 specifies and describes the situation in Arg1 in more details as in Example 66 (ii) Arg2 summarizes Arg1, such as in Example 67. (iii) Arg2 describes the same situation as Arg1 from a different perspective, such as in Example 68. In all cases, the situations described in Arg1 and Arg2 are both true or false. (ADAPTED_from_PDTB)**.**

(66)

*ضرب القرية زلزال مدمر و خلف وارءه دمارا في المنازل و الطرقات*

*A devastating earthquake hit the village* and **it left a massive destruction in houses and roads**

(67)

*ازدادت حالة الشعب الفلسطيني سوءاً فبعد التضيق على الفلسطسنين بالعبور عبر المعابر إلى انقطاع الكهرباء و الماء لعدة ساعات يومياً .* و بصورة عامة **بات الشعب الفلسطيني على مهب الريح**

*The Situation of the Palestinian people has got worse; they dont have the right for passing the crossing points, electricity and water are interrupted for several hours a day*. (And) In general **the Palestinian people are caught in a storm**

(68)

*ستكون هذه الاتفاقية فيما بعد بمثابة أساس للاتفاقيات و المعاهدات الأخرى التي ستجري حول العلاقات المتبادلة في شتى المجالات .* بعبارة أخرى **إن هذه الاتفاقية تفسح المجال لإمكانية حدوث اتفاقيات أخرى في المستقبل**

around done will that other deals and for-conventions basic as Later-on convention this will field will-allow the-agreement this In other words. fields various in exchanged relations [future in other conventions happening possibility-for

*This convention will be later as the basis for other conventions and deals that will take place on mutual relations in various fields*. In other words, **this agreement will allow for more cooperation in the future.**

#### 5.2.4.3 Type: Alternative

The type *Alternative* applies when the two arguments denote alternative situations. (SAME_as_PDTB). Example 69.

(69)

إما *أن تقول الحقيقة* أو **تعاقب على فعلتك**

[act for punish or the-truth say either]

{Either *you tell the truth* or you **will be punished for your act**}

#### 5.2.4.4 Subtype: conjunctive

The *conjunctive* subtype is used when the connective indicates that both alternatives hold or are possible. Example 70.

(70)

من المقرر ان يختار اعضاء "أوبك" أيضا ضخ *جزء من هذه الفوائض في مشاريع انمائية* او **التخفيف قليلا من ديونهم**، وفقا للاسبوعية

It is scheduled that *OPEC members choose to pump a part of their profits into developing new projects* or **into reducing their debts slightly**, according to the weekly press

#### 5.2.4.5 Subtype: disjunctive

The *disjunctive* subtype is used when two situations are evoked in the discourse but only one of them can hold. Example 71.

(71)

إما **أن تقول الحقيقة** أو *تعاقب على فعلتك*

{Either *you tell the truth* or you **will be punished for your act**}

#### 5.2.4.6 Type: Exception

The type *Exception* applies when Arg2 specifies an exception to the generalization specified by Arg1. The generalization in Arg1 can be a negative situation and the exception is the positive situation in Arg2 as in Example 72, or the other way around. Alternatively, both generalization and exception situations have positive impacts but the situation in Arg2 is an exception from the situation in Arg1 (SAME_as_PDTB).

(72)

*سوف لن يبقى في ذاكرة الناس*الا **عمل الخير و السعي في مصالح الناس**

[people needs in seeking and charity work except the-people memory in remained not will]

*Nothing will remain in the people memory* except **doing charity and looking after people** } {**needs**

### 5.2.4.7 Type: Background

The type *Background* applies when Arg2 describes a new situation related to the situation in Arg1 by giving more details as *background information* in order to give the reader a wider view about the situation in Arg1 to improve understanding. The new situation (information) in Arg2 happens before the situation in Arg1 (NEW). Examples: 73 and 74.

(73)

*غادر الرئيس جورج بوش العراق بخيبة أمل من إيجاد حل سياسي للحرب على الإرهاب في العراق.* وقد **بدأت الحرب في العراق عام 2005 اثر مزاعم امريكية بنية العراق امتلاك سلاح نووي**

*President George W. Bush, left Iraq, disappointed about the failure to find a political solution to the terrorism war in Iraq*. (and) **The war in Iraq began in 2005 after U.S. allegations that Iraq has intention of acquiring nuclear weapons**

(74)

*ان الطائرة التي تقل الوفد اللبناني الرسمي وصلت اليوم الثلاثاء الى طرابلس.* **وقد اتى الوفد لاصطحاب الرهينة اللبنانية ماري ميشال معربس المحتجزة في الفيليبين**

*The plane, which was carrying the official Lebanese delegation, arrived in Tripoli on Tuesday*. (and)**The delegation came to accompany the Lebanese hostage Marie Michel Maarbes, who held in the Philippines.**

### 5.2.4.8 Type: Conjunction

The type *Conjunction* is used when the situation described in Arg2 provides additional, new information to the situation described in Arg1, but the relation does not fit any of the relations described above **(**ADAPTED_from_PDTB). Examples: 75 and 76.

(75)

*وكان رودريغيز اعلن امس الجمعة ان المنظمة النفطية لن تعقد اجتماعا استثنائيا الثلاثاء في فيينا* **كما لم يرتب لمناقشة تأثير ارتفاع الأسعار على مخزون النفط**

*Rodriguez was announced yesterday, Friday, that the oil organization will hold an extraordinary meeting in Vienna Tuesday* **also it was not arranged to discuss the impact of the prices rising of oil stocks**

(76)

قال الإتحاد الأوربي انه[ *سوف يدرس موضوع تخفيض الرسوم الجمركية]* و [**سيتبنى مشروع السلام في الشرق الأوسط]**

The EU said he *will study the issue of tariff cuts* and **will agree to the peace project in the Middle East**

## 5.3 Entity relations

In this first discourse annotation effort for Arabic, we annotate these relations as Conjunction relations if they are introduced by an explicit discourse connective (NEW).

(77)

نالت محاضرة الدكتور الحبيب استحسان جميع الحضور **خصوصا الجزء الذي ناقش فيه الروبوتات و علاقتها بإبداع الأطفال**

within discussed that the-part especially attendance *all welcoming Dr. Habib lecture [received*

children creativity relationship and robots]

*Dr. Habibs lecture received a strong welcoming from* attendants **especially the part that discussed robots and their relationship to children creativity**

## 5.4 Multiple discourse relations (combined relations)

Annotators are allowed to assign more than one relation to a DC. For example, the connective (بعدما/after that) indicates two discourse relations (Temporal.Asynchronous/Contingency.Cause.Reason.NonPragmatic) in Example 78.

(78)

*اتخذ الشقيقان قرارهما* **بعدما وجدا تجاهلا تاما من قبل ادارة الكرة في الاهلي**

*The brothers made their decision* **after they were disregarded completely by the } {department of the football in the Alahli Club**

The connective (الا بعد/*except after*) in Example 79 indicates the relations Temporal.Synchronous/ Expansion.Exception. In contrast, the same connective ( الا بعد/*only after*) indicates in addition a relation Comparison.Condition in Example 80.

(79)

*لم تشعر ليلى بطعم الراحة* الا بعد **ان سمعت خبر عودة ابيها من السفر**

[travel from her-father back news heard after except relax taste Laila Feel not]

*{Laila did not feel relaxed* except after **she heard news about return back her father from a way}**

(80)

*سوف لن تذهب معنا للتسوق* إلا بعد **أن تكمل أداء جميع واجباتك المدرسية**

*You are not allowed to go shopping with us* except after **you finished doing all your homework**

# 6 Discourse Annotation Procedure

Follow the subsequent procedure for each raw text file.

1) Read the article fully to get a comprehensive view about what knowledge the writer intended to pass to the readers.

2) Go through each highlighted potential connective (listed in the suggested connectives list in the READSD tool) in order and make the following decision according to our guidelines:

- The highlighted connective is a discourse connective. If so, go to Step3.
- The highlighted connective is **not** a discourse connective; remove it from the list (into the Non-discourse connective list in the tool using the arrows). Jump to the beginning of step 2with the next highlighted potential connective.

3) Mark the first argument (Arg1) and the second Argument (Arg2).

4) Select suitable discourse relations from our relations taxonomy.

5) If the connective is paired, you should mark the second part of the connective as well.

6) Write down any comment or suggestion about this annotation in the comment box.

7) Save the annotation and go to Step 2 for the next highlighted potential connective.

At the end, there should be no suggested connectives left without a decision. Section 7 describes the annotation procedure using the newly developed annotation tool in detail.

**Table 6-1: Hints for discourse annotation**

**Hints!**

- The highlighted potential discourse connective is not a discourse connective unless it relates two abstract objects Arg1 and Arg2.
- The connective string should not include attached pronoun clitics. The pronoun is a part of the argument.
- Arguments should not include irrelevant connectives such as a connective of a different annotation.
- Remember that the connective always introduces Arg2
- Function words such as ‘ ’قد’,’وان and ’ ’كانare parts of arguments.
- The Annotator must indicate that the current connective is a paired connective by clicking a check box ‘Paired Conn?. ’The paired connective should be annotated as:
    - The first part is the highlighted connective.
    - The second part could be any token/clitic.
- The Annotator is not allowed to add new connectives. However, he can record his comments in a comment box.
- Annotators should look for a relation between the two AOs (Arg1 and Arg2) following the sequence:
    a) The DC expresses a TEMPORAL, CONTINGENCY, COMPARISON relation. If not:
    b) It expresses an EXPANSION relation other than Background and Conjunction. If not:
    c) It expresses the Background relation? If not:
    d) It expresses the Conjunction relation.

# 7 The Discourse Annotation Tool for Arabic and English

(This section is almost similar to Chapter 6 in the main thesis)

# 8 References


AL-SANIE, W., A. TOUIR and H. MATHKOUR. 2005. Towards a Rhetorical Parsing of Arabic Text. *In: The International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC'05)*: IEEE Computer Society.

FRASER, B. 1999. What are discourse markers? *Journal of Pragmatics,* **31**(7), pp.931-952.

HOBBS, J. R. 1985. 85-37. *On the Coherence and Structure of Discourse*. Center for the Study of Language and Information (CSLI), Stanford University.

HOVY, E. H. and E. MAIER. 1993. *Parsimonious and Profligate: How Many and Which Discourse Structure Relations? Discourse Processes* University of Southern Claifornia.

M. ABDL AL LATIF, A. U., M. ZAHRAN and D. A. AL-ARABI. 1997. *Alnhw AlAsAsi.* CSLI.

MAAMOURI, M., A. BIES, T. BUCKWALTER and W. MEKKI. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus *In: NEMLAR Conference on Arabic Language Resources and Tools,* Cairo, Egypt.

MANN, W. C. and S. A. THOMPSON. 1987. *Rhetorical Structure Theory: a theory of text organization*. Technical Report ISI/RS- Information Sciences Institute

MARCUS, M. P., B. SANTORINI and M. A. MARCINKIEWICZ. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics,* **19**, pp.313--330.

MILTSAKAKI, E., R. PRASAD, A. JOSHI and B. WEBBER. 2006. The Penn Discourse Treebank.

OZA, U., R. PRASAD, S. KOLACHINA, D. M. SHARMA and A. JOSHI. 2009. The hindi discourse relation bank. *In: Proceedings of the Third Linguistic Annotation Workshop,* Suntec, Singapore.

PITLER, E., M. RAGHUPATHY, H. MEHTA, A. NENKOVA, A. LEE and A. JOSHI. 2008. Easily identifiable discourse relations. *In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008),* Manchester, UK.

PRASAD, R., S. HUSAIN, D. M. SHARMA and A. JOSHI. 2008. Towards an Annotated Corpus of Discourse Relations in Hindi. *In: In The Third International Joint Conference on Natural Language Processing*, *January 2008, India*.

PRASAD, R., E. MILTSAKAKI, N. DINESH, A. LEE, A. JOSHI, L. ROBALDO and B. WEBBER. 2007. The penn discourse treebank 2.0 annotation manual. *The PDTB Research Group (2007)*.

RYDING, K. C. 2005. *A reference grammar of modern standard Arabic.* Cambridge: Cambridge University Press.

SANDERS, T. J. M. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes,* **15**(1), pp.1-35.

SEIF, A., H. MATHKOUR and A. TOUIR. 2005. An RST Computational Tool for the Arabic Language. *In: iiWAS, Malaysia, Kuala Lumpur*.

WOLF, F., E. GIBSON, A. FISHER and M. KNIGHT. 2003. A procedure for collecting a database of texts annotated with coherence relations. *Documentation accompanying the Discourse GraphBank, LDC2005T08*.

ZEYREK, D. and B. WEBBER. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. *In: Proceedings of IJCNLP-2008. ,* Hyderabad, India.

## Appendix A: A List of Potential Discourse Connectives for Arabic

(The content of this appendix is similar to the finial deposit of potential discourse connectives for Arabic, Table 4-2, in the main thesis)

## Appendix B: Al-maSdar Morphological Patterns

(The content of this appendix is similar to Appendix A in the main thesis)

# Appendix C

# Distribution of Arabic discourse connectives

This appendix provides the distribution of the types of explicit connectives in the LADTB v.1, and the discourse relation types they signal. The full distribution is presented in the following tables. There are 80 distinct types of explicit connectives including modified connectives. The total number of Explicit discourse connective tokens annotated is 6,328 (the total for the third column). Each connective type is described by how often it has discourse function (the second and third columns), how often it has not discourse function in context (the fourth and fifth columns), its total (the sixth column), the last two columns present the discourse relations of the discourse connective signal in the LADTB. Each relation signal the connective is presented with a frequency and a percentage. The number of how many relations are labelled for the connective is presented in the last column. The multiple relations are separated by a slash sign. The association between discourse relations and the full forms of connectives is shown in Appendix D. Note, there might be more than one possible translation of the Arabic connective which varies depending on the context. Only one approximate translation is attached to the connective type in the table.

**Table C: Distribution of discourse connectives in the LADTB.**

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| و/w/and | 3999 | 54% | 3376 | 46% | 7375 | {76%:EXPANSION.Conjunction (3070)}; {7%:EXPANSION.Reformulation (287)}; {4%:EXPANSION.Background (184)}; {3%:CONTINGENCY.Cause.Result.NonPragmatic (134)}; {2%:TEMPORAL.Asynchronous (109)}; {1%:COMPARISON.Contrast (55)}; {0%:CONTINGENCY.Cause.Reason.NonPragmatic (31)}; {0%:TEMPORAL.Synchronous (29)}; {0%:EXPANSION.Exemplification (24)}; {0%:CONTINGENCY.Cause.Result.Pragmatic (23)}; {0%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (12)}; {0%:CONTINGENCY.Cause.Reason.Pragmatic (11)}; {0%:COMPARISON.Similarity (5)}; {0%:EXPANSION.Exemplification/EXPANSION.Reformulation (3)}; {0%:COMPARISON.Contrast/TEMPORAL.Synchronous (3)}; {0%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (2)}; {0%:COMPARISON.Contrast/TEMPORAL.Asynchronous (2)}; {0%:COMPARISON.Contrast/CONTINGENCY.Cause.Result.NonPragmatic (2)}; {0%:EXPANSION.Reformulation/TEMPORAL.Asynchronous (1)}; {0%:EXPANSION.Exemplification/TEMPORAL.Asynchronous (1)}; {0%:EXPANSION.Exception (1)}; {0%:CONTINGENCY.Condition (1)}; {0%:CONTINGENCY.Cause.Result.Pragmatic/TEMPORAL.Asynchronous (1)}; {0%:CONTINGENCY.Cause.Result.Pragmatic/EXPANSION.Conjunction (1)}; {0%:CONTINGENCY.Cause.Result.Pragmatic/CONTINGENCY.Condition (1)}; {0%:CONTINGENCY.Cause.Result.NonPragmatic/EXPANSION.Background (1)}; {0%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Asynchronous (1)}; {0%:CONTINGENCY.Cause.Reason.NonPragmatic/EXPANSION.Background (1)}; {0%:COMPARISON.Similarity/EXPANSION.Exemplification (1)}; {0%:COMPARISON.Contrast/EXPANSION.Background (1)}; {0%:COMPARISON.Contrast/CONTINGENCY.PragmaticCondition (1)} | 31 |
| ل/l/for | 468 | 11% | 3838 | 89% | 4306 | {93%:CONTINGENCY.Cause.Reason.NonPragmatic (437)}; {5%:CONTINGENCY.Cause.Result.NonPragmatic (25)}; {0%:CONTINGENCY.Cause.Result.Pragmatic (3)}; {0%:CONTINGENCY.Cause.Reason.Pragmatic (3)} | 4 |

**Distribution of discourse connectives in the LADTB (cont.)**

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| لكن/lkn/however | 204 | 99% | 3 | 1% | 207 | {97%:COMPARISON.Contrast (198)}; {0%:EXPANSION.Conjunction (2)}; {0%:COMPARISON.Contrast/CONTINGENCY.Cause.Reason.NonPragmatic (2)}; {0%:EXPANSION.Reformulation (1)}; {0%:COMPARISON.Contrast/EXPANSION.Exception (1)} | 5 |
| بعد/bEd/after | 194 | 62% | 121 | 38% | 315 | {51%:TEMPORAL.Asynchronous (100)}; {39%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (76)}; {4%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Asynchronous (9)}; {2%:COMPARISON.Contrast/TEMPORAL.Asynchronous (4)}; {1%:CONTINGENCY.Cause.Reason.NonPragmatic (3)}; {0%:CONTINGENCY.Cause.Result.Pragmatic/TEMPORAL.Asynchronous (1)}; {0%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Synchronous (1)} | 7 |
| خلال/xlAl/during | 102 | 81% | 24 | 19% | 126 | {100%:TEMPORAL.Synchronous (102)} | 1 |
| ف/f/then | 99 | 6% | 1426 | 94% | 1525 | {29%:CONTINGENCY.Cause.Result.NonPragmatic (29)}; {20%:CONTINGENCY.Cause.Reason.NonPragmatic (20)}; {18%:EXPANSION.Reformulation (18)}; {12%:EXPANSION.Exemplification (12)}; {6%:TEMPORAL.Asynchronous (6)}; {4%:CONTINGENCY.Cause.Reason.Pragmatic (4)}; {3%:TEMPORAL.Synchronous (3)}; {2%:COMPARISON.Contrast (2)}; {1%:EXPANSION.Background (1)}; {1%:CONTINGENCY.Cause.Result.Pragmatic (1)}; {1%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (1)}; {1%:CONTINGENCY.Cause.Reason.Pragmatic/EXPANSION.Background (1)}; {1%:CONTINGENCY.Cause.Reason.NonPragmatic/EXPANSION.Exemplification (1)} | 13 |
| ب/b/by | 96 | 2% | 4072 | 98% | 4168 | {89%:CONTINGENCY.Cause.Reason.NonPragmatic (86)}; {5%:TEMPORAL.Synchronous (5)}; {4%:CONTINGENCY.Cause.Reason.Pragmatic (4)}; {1%:COMPARISON.Contrast (1)} | 4 |
| قبل/qbl/before | 84 | 52% | 77 | 48% | 161 | {98%:TEMPORAL.Asynchronous (83)}; {1%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (1)} | 2 |
| لان/lAn/because | 77 | 73% | 29 | 27% | 106 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (77)} | 1 |
| منذ/mn*/since | 69 | 31% | 151 | 69% | 220 | {69%:TEMPORAL.Asynchronous (48)}; {15%:TEMPORAL.Synchronous (11)}; {11%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (8)}; {1%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Synchronous (1)}; {1%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Asynchronous (1)} | 5 |

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| كما/kmA/as | 69 | 66% | 36 | 34% | 105 | {57%:EXPANSION.Conjunction (40)}; {13%:TEMPORAL.Synchronous (9)}; {13%:COMPARISON.Similarity (9)}; {4%:TEMPORAL.Asynchronous (3)}; {2%:COMPARISON.Contrast (2)}; {1%:EXPANSION.Reformulation/TEMPORAL.Synchronous (1)}; {1%:EXPANSION.Reformulation (1)}; {1%:EXPANSION.Exemplification (1)}; {1%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Synchronous (1)}; {1%:COMPARISON.Similarity/EXPANSION.Exemplification (1)}; {1%:COMPARISON.Contrast/COMPARISON.Similarity (1)} | 11 |
| اثر/Avr/after | 67 | 97% | 2 | 3% | 69 | {73%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (49)}; {13%:TEMPORAL.Asynchronous (9)}; {13%:CONTINGENCY.Cause.Reason.NonPragmatic (9)} | 3 |
| عندما/EndmA/ when | 54 | 98% | 1 | 2% | 55 | | |
| بسبب/bsbb/ because of | 49 | 94% | 3 | 6% | 52 | {51%:TEMPORAL.Synchronous (28)}; {16%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Synchronous (9)}; {7%:TEMPORAL.Asynchronous (4)}; {7%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (4)}; {3%:CONTINGENCY.Condition/TEMPORAL.Synchronous (2)}; {3%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (2)}; {3%:CONTINGENCY.Cause.Reason.NonPragmatic (2)}; {1%:CONTINGENCY.Condition (1)}; {1%:COMPARISON.Similarity/EXPANSION.Reformulation (1)}; {1%:COMPARISON.Contrast/TEMPORAL.Asynchronous (1)} | 10 |
| الا ان/AlA An/however | 41 | 100% | 0 | 0% | 41 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (49)} | 1 |
| فيما/fymA/ while | 36 | 88% | 5 | 12% | 41 | {92%:COMPARISON.Contrast (38)}; {2%:EXPANSION.Exception (1)}; {2%:EXPANSION.Conjunction (1)}; {2%:COMPARISON.Contrast/EXPANSION.Reformulation (1)} | 4 |

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| عندما/EndmA/ when | 54 | 98% | 1 | 2% | 55 | | |
| بسبب/bsbb/ because of | 49 | 94% | 3 | 6% | 52 | {51%:TEMPORAL.Synchronous (28)}; {16%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Synchronous (9)}; {7%:TEMPORAL.Asynchronous (4)}; {7%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (4)}; {3%:CONTINGENCY.Condition/TEMPORAL.Synchronous (2)}; {3%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (2)}; {3%:CONTINGENCY.Cause.Reason.NonPragmatic (2)}; {1%:CONTINGENCY.Condition (1)}; {1%:COMPARISON.Similarity/EXPANSION.Reformulation (1)}; {1%:COMPARISON.Contrast/TEMPORAL.Asynchronous (1)} | 10 |
| الا ان/AlA An/however | 41 | 100% | 0 | 0% | 41 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (49)} | 1 |
| فيما/fymA/ while | 36 | 88% | 5 | 12% | 41 | {92%:COMPARISON.Contrast (38)}; {2%:EXPANSION.Exception (1)}; {2%:EXPANSION.Conjunction (1)}; {2%:COMPARISON.Contrast/EXPANSION.Reformulation (1)} | 4 |

| Connective | Discourse Conn | | NonDis. Conn | Total | | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| ثم/vm/then | 36 | 75% | 12 | 25% | 48 | {91%:TEMPORAL.Asynchronous (33)}; {2%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (1)}; {2%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (1)}; {2%:COMPARISON.Contrast/TEMPORAL.Asynchronous (1)} | 4 |
| او/Aw/or | 35 | 38% | 58 | 62% | 93 | {80%:EXPANSION.Alternative.Conjunctive (28)}; {20%:EXPANSION.Alternative.Disjunctive (7)} | 2 |
| في حال/fy HAl/in case | 35 | 83% | 7 | 17% | 42 | {100%:CONTINGENCY.Condition (35)} | 1 |
| اذا/A*A/if | 34 | 69% | 15 | 31% | 49 | {94%:CONTINGENCY.Condition (32)}; {2%:CONTINGENCY.PragmaticCondition (1)}; {2%:CONTINGENCY.Condition/EXPANSION.Exception (1)} | 3 |
| حيث/Hyv/ where-since | 32 | 33% | 64 | 67% | 96 | {40%:CONTINGENCY.Cause.Reason.NonPragmatic (13)}; {21%:EXPANSION.Reformulation (7)}; {9%:TEMPORAL.Synchronous (3)}; {6%:EXPANSION.Conjunction (2)}; {6%:CONTINGENCY.Cause.Reason.Pragmatic (2)}; {3%:EXPANSION.Exemplification (1)}; {3%:EXPANSION.Background (1)}; {3%:CONTINGENCY.Cause.Result.NonPragmatic (1)}; {3%:CONTINGENCY.Cause.Reason.Pragmatic/EXPANSION.Exemplification (1)}; {3%:CONTINGENCY.Cause.Reason.NonPragmatic/EXPANSION.Exemplification (1)} | 10 |
| رغم/rgm/ although | 31 | 82% | 7 | 18% | 38 | {100%:COMPARISON.Contrast (31)} | 1 |
| حتى/HtY/until | 29 | 39% | 46 | 61% | 75 | {20%:CONTINGENCY.Cause.Reason.NonPragmatic (6)}; {20%:COMPARISON.Contrast (6)}; {13%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (4)}; {10%:CONTINGENCY.Cause.Result.NonPragmatic (3)}; {6%:CONTINGENCY.PragmaticCondition (2)}; {6%:CONTINGENCY.Cause.Result.Pragmatic (2)}; {3%:TEMPORAL.Synchronous (1)}; {3%:EXPANSION.Conjunction (1)}; {3%:CONTINGENCY.Condition (1)}; {3%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Synchronous (1)}; {3%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (1)}; {3%:COMPARISON.Contrast/CONTINGENCY.Condition (1)} | 12 |
| في حين/fy Hyn/while | 27 | 96% | 1 | 4% | 28 | {44%:COMPARISON.Contrast (12)}; {25%:TEMPORAL.Synchronous (7)}; {25%:COMPARISON.Contrast/TEMPORAL.Synchronous (7)}; {3%:EXPANSION.Conjunction (1)} | 4 |

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| اما/AmA/while | 24 | 92% | 2 | 8% | 26 | {75%:COMPARISON.Contrast (18)}; {20%:EXPANSION.Conjunction (5)}; {4%:TEMPORAL.Asynchronous (1)} | 3 |
| خصوصا/ xSwSA/ especially | 23 | 36% | 41 | 64% | 64 | {39%:EXPANSION.Reformulation (9)}; {21%:EXPANSION.Exemplification (5)}; {13%:CONTINGENCY.Cause.Reason.NonPragmatic/EXPANSION.Reformulation (3)}; {8%:CONTINGENCY.Cause.Reason.Pragmatic (2)}; {8%:CONTINGENCY.Cause.Reason.NonPragmatic (2)}; {4%:CONTINGENCY.Cause.Reason.Pragmatic/EXPANSION.Reformulation (1)}; {4%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (1)} | 7 |
| بعدما/bEdmA/ after that | 23 | 100% | 0 | 0% | 23 | {52%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (12)}; {30%:TEMPORAL.Asynchronous (7)}; {8%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Asynchronous (2)}; {8%:COMPARISON.Contrast/TEMPORAL.Asynchronous (2)} | 4 |
| اذ/A*/as | 22 | 100% | 0 | 0% | 22 | {45%:CONTINGENCY.Cause.Reason.NonPragmatic (10)}; {22%:EXPANSION.Reformulation (5)}; {9%:CONTINGENCY.Cause.Reason.NonPragmatic/EXPANSION.Reformulation (2)}; {4%:EXPANSION.Exemplification (1)}; {4%:EXPANSION.Conjunction (1)}; {4%:CONTINGENCY.Cause.Result.Pragmatic (1)}; {4%:CONTINGENCY.Cause.Reason.Pragmatic/EXPANSION.Reformulation (1)}; {4%:CONTINGENCY.Cause.Reason.Pragmatic (1)} | 8 |
| مما/mmA/ which lead to | 21 | 81% | 5 | 19% | 26 | {100%:CONTINGENCY.Cause.Result.NonPragmatic (21)} | 1 |
| ايضا/AyDA/ also | 17 | 17% | 85 | 83% | 102 | {94%:EXPANSION.Conjunction (16)}; {5%:TEMPORAL.Asynchronous (1)} | 2 |
| بينما/bynmA/ while | 16 | 100% | 0 | 0% | 16 | {50%:TEMPORAL.Synchronous (8)}; {37%:COMPARISON.Contrast (6)}; {12%:COMPARISON.Contrast/TEMPORAL.Synchronous (2)} | 3 |
| بل/bl/but | 15 | 94% | 1 | 6% | 16 | {73%:COMPARISON.Contrast (11)}; {20%:EXPANSION.Conjunction (3)}; {6%:EXPANSION.Reformulation (1)} | 3 |
| بهدف/bhdf/ because of | 15 | 56% | 12 | 44% | 27 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (15)} | 1 |

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| بالتالي/bAltAly/ consequently | 14 | 93% | 1 | 7% | 15 | {85%:CONTINGENCY.Cause.Result.NonPragmatic (12)}; {14%:CONTINGENCY.Cause.Result.Pragmatic (2)} | 2 |
| جراء/jrA'/ because | 10 | 100% | 0 | 0% | 10 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (10)} | 1 |
| على الرغم/ ElY Alrgm | 9 | 100% | 0 | 0% | 9 | {100%:COMPARISON.Contrast (9)} | 1 |
| نظرا ل/nZrA l/ because of | 9 | 100% | 0 | 0% | 9 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (9)} | 1 |
| انما/AnmA/but | 7 | 70% | 3 | 30% | 10 | {57%:COMPARISON.Contrast (4)}; {14%:EXPANSION.Conjunction (1)}; {14%:CONTINGENCY.Cause.Result.NonPragmatic (1)}; {14%:CONTINGENCY.Cause.Reason.NonPragmatic (1)} | 4 |
| لو/lw/if (in the past) | 6 | 43% | 8 | 57% | 14 | {33%:CONTINGENCY.Condition (2)}; {33%:COMPARISON.Contrast (2)}; {16%:CONTINGENCY.PragmaticCondition (1)}; {16%:CONTINGENCY.Cause.Result.Pragmatic (1)} | 4 |
| في ظل/fy Zl/ under | 6 | 100% | 0 | 0% | 6 | {50%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Synchronous (3)}; {33%:TEMPORAL.Synchronous (2)}; {16%:CONTINGENCY.Cause.Reason.NonPragmatic (1)} | 3 |
| بيد ان/byd An/ but | 6 | 100% | 0 | 0% | 6 | {100%:COMPARISON.Contrast (6)} | 1 |
| رغم ان /rgm An/although | 6 | 100% | 0 | 0% | 6 | {100%:COMPARISON.Contrast (6)} | 1 |
| غير ان/gyr An/but | 6 | 100% | 0 | 0% | 6 | {100%:COMPARISON.Contrast (6)} | 1 |
| فضلا عن/fDlA En/ as well as | 6 | 43% | 8 | 57% | 14 | {100%:EXPANSION.Conjunction (6)} | 1 |
| كذلك/k*lk/and that | 6 | 30% | 14 | 70% | 20 | {100%:EXPANSION.Conjunction (6)} | 1 |
| عقب/Eqb/shortly after | 5 | 100% | 0 | 0% | 5 | {40%:TEMPORAL.Asynchronous (2)}; {40%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Asynchronous (2)}; {20%:CONTINGENCY.Cause.Reason.Pragmatic/TEMPORAL.Asynchronous (1)} | 3 |
| لاسيما/lA symA/ Particularly | 5 | 28% | 13 | 72% | 18 | {40%:EXPANSION.Exemplification (2)}; {40%:CONTINGENCY.Cause.Reason.NonPragmatic (2)}; {20%:EXPANSION.Reformulation (1)} | 3 |
| الا بعد/AlA bEd/ except after | 5 | 83% | 1 | 17% | 6 | {80%:EXPANSION.Exception/TEMPORAL.Asynchronous (4)}; {20%:CONTINGENCY.Condition/TEMPORAL.Asynchronous (1)} | 2 |

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| بفضل/bfDl/ thanks to | 5 | 100% | 0 | 0% | 5 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (5)} | 1 |
| قبيل/qbyl/ shortly before | 5 | 100% | 0 | 0% | 5 | {100%:TEMPORAL.Asynchronous (5)} | 1 |
| في المقابل/fy AlmqAbl/in contrast | 5 | 100% | 0 | 0% | 5 | {100%:COMPARISON.Contrast (5)} | 1 |
| بالرغم من/bAlrgm mn/although | 5 | 100% | 0 | 0% | 5 | {100%:COMPARISON.Contrast (5)} | 1 |
| لكي/lky/for | 5 | 83% | 1 | 17% | 6 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (5)} | 1 |
| بغية/bgyp/ desire to | 5 | 100% | 0 | 0% | 5 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (5)} | 1 |
| طالما/TAlmA/ as long as | 4 | 100% | 0 | 0% | 4 | {50%:CONTINGENCY.Condition (2)}; {25%:CONTINGENCY.Cause.Reason.NonPragmatic/TEMPORAL.Synchronous (1)}; {25%:CONTINGENCY.Cause.Reason.NonPragmatic/CONTINGENCY.Condition (1)} | 3 |
| من ثم/mn vm/ then after | 4 | 100% | 0 | 0% | 4 | {50%:TEMPORAL.Asynchronous (2)}; {50%:CONTINGENCY.Cause.Result.NonPragmatic/TEMPORAL.Asynchronous (2)} | 2 |
| بالمقابل/bAlmqAbl/in contrast | 3 | 100% | 0 | 0% | 3 | {33%:EXPANSION.Conjunction (1)}; {33%:COMPARISON.Contrast/TEMPORAL.Synchronous (1)}; {33%:COMPARISON.Contrast (1)} | 3 |
| الا/AlA/except | 3 | 38% | 5 | 63% | 8 | {66%:EXPANSION.Exception (2)}; {33%:EXPANSION.Exception/TEMPORAL.Asynchronous (1)} | 2 |
| نتيجة/ntyjp/ a result of | 3 | 75% | 1 | 25% | 4 | {66%:CONTINGENCY.Cause.Reason.NonPragmatic (2)}; {33%:CONTINGENCY.Cause.Reason.Pragmatic (1)} | 2 |
| بالاضافة الى/bAlADAfp AlY/in addition to | 3 | 30% | 7 | 70% | 10 | {100%:EXPANSION.Conjunction (3)} | 1 |
| لأن/lAn/ because | 3 | 100% | 0 | 0% | 3 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (3)} | 1 |
| قبل ان /qbl An/before that | 3 | 100% | 0 | 0% | 3 | {100%:TEMPORAL.Asynchronous (3)} | 1 |

| Connective | Discourse Conn | | NonDis. Conn | | Total | Discourse Relations and frequency | #Rel |
|---|---|---|---|---|---|---|---|
| حال/HAl/when | 2 | 100% | 0 | 0% | 2 | {50%:EXPANSION.Conjunction (1)}; {50%:CONTINGENCY.Condition (1)} | 2 |
| الا اذا/AlA A*A/except if | 2 | 100% | 0 | 0% | 2 | {50%:EXPANSION.Exception (1)}; {50%:CONTINGENCY.Condition (1)} | 2 |
| حتى لو/HtY lw/even if | 2 | 100% | 0 | 0% | 2 | {100%:CONTINGENCY.PragmaticCondition (2)} | 1 |
| حينها /HynhA/when | 2 | 40% | 3 | 60% | 5 | {100%:TEMPORAL.Synchronous (2)} | 1 |
| كي /ky/for | 2 | 67% | 1 | 33% | 3 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (2)} | 1 |
| وقبل/wqbl/and before | 1 | 100% | 0 | 0% | 1 | {100%:TEMPORAL.Asynchronous (1)} | 1 |
| بيد /byd/but | 1 | 100% | 0 | 0% | 1 | {100%:COMPARISON.Contrast (1)} | 1 |
| اضافة الى/ ADAfp AlY/ additionally | 1 | 5% | 18 | 95% | 19 | {100%:EXPANSION.Conjunction (1)} | 1 |
| كأن /k<n/like | 1 | 100% | 0 | 0% | 1 | {100%:EXPANSION.Exemplification (1)} | 1 |
| بحيث/bHyv/ where/since | 1 | 100% | 0 | 0% | 1 | {100%:CONTINGENCY.Cause.Result.NonPragmatic (1)} | 1 |
| حين/Hyn/when | 1 | 3% | 30 | 97% | 31 | {100%:TEMPORAL.Synchronous (1)} | 1 |
| خلافا ل/xlAfA l /in conflict to | 1 | 100% | 0 | 0% | 1 | {100%:COMPARISON.Contrast (1)} | 1 |
| برغم/brgm/ although | 1 | 100% | 0 | 0% | 1 | {100%:COMPARISON.Contrast (1)} | 1 |
| كلما/klmA/if | 1 | 100% | 0 | 0% | 1 | {100%:CONTINGENCY.Condition/TEMPORAL.Synchronous (1)} | 1 |
| لذا /l*A/for this | 1 | 50% | 1 | 50% | 2 | {100%:CONTINGENCY.Cause.Reason.NonPragmatic (1)} | 1 |
| لذلك/l*lk/for that | 1 | 17% | 5 | 83% | 6 | {100%:CONTINGENCY.Cause.Result.NonPragmatic (1)} | 1 |
| بمعنى آخر/ bmEnY xr/in other words | 1 | 100% | 0 | 0% | 1 | {100%:EXPANSION.Reformulation (1)} | 1 |
| لولا/lwlA/if not | 1 | 100% | 0 | 0% | 1 | {100%:CONTINGENCY.Condition (1)} | 1 |

# Appendix D

# Distribution of Arabic Discourse Relations

This appendix provides a distribution of all the distinct discourse relations in the LADTB: 17 distinct single relations plus 38 multiple relations (separated by a slash) were labelled for explicit connectives in the LADTB. The table below shows the full distribution. The second column presents, for each discourse relation (in the first column), a list of all explicit connectives that signal the relation. The list is ordered via frequency of the connectives. Each connective type comes with a percentage and a count of how often it is annotated with the relation. Similar to the distribution in Appendix C, connectives listed in the table also include the modified forms with no distinction between them. The total of counted tokens of a relation is presented in the third column. The last column presents the number of connective types that indicate the relation. Some relations are indicated in the LADTB by only one connective such as EXPANSION.Alternative, while 26 connectives indicate CONTINGENCY.Cause.Reason.NonPragmatic (the relation with the largest number of signalling connectives).

**Table D: Distribution of discourse relations in the LADTB**

| Discourse Relation | Discourse connective | Total (6,328) | #Dis. Conn |
|---|---|---|---|
| EXPANSION.Conjunction | {97%: و/w (3070)}; {1.3%: كما/kmA (40)}; {0.5%: ايضا/AyDA (16)}; {0.2%: كذلك/k*lk (6)}; {0.2%: فيما/fymA (6)}; {0.2%: فضلا عن/fDlA En (6)}; {0.2%: اما/AmA (5)}; {0.1%: بل/bl (3)}; {0.1%: بالاضافة الى/bAlADAfp AlY (3)}; {0.1%: لكن/lkn (2)}; {0.1%: حيث/Hyv (2)}; {0.03%: في حين/fy Hyn (1)}; {0.03%: حتى/HtY (1)}; {0.03%: حال/HAl (1)}; {0.03%: بالمقابل/bAlmqAbl (1)}; {0.03%: انما/AnmA (1)}; {0.03%: الا ان/AlA An (1)}; {0.03%: اضافة الى/ADAfp AlY (1)}; {0.03%: اذ/A* (1)} | 3167 | 19 |
| CONTINGENCY.Cause.Reason. NonPragmatic | {54.2%: ل/l (437)}; {10.7%: ب/b (86)}; {9.6%: لان/lAn (77)}; {6.1%: بسبب/bsbb (49)}; {3.9%: و/w (31)}; {2.5%: ف/f (20)}; {1.9%: بهدف/bhdf (15)}; {1.6%: حيث/Hyv (13)}; {1.2%: جراء/jrA' (10)}; {1.2%: اذ/A* (10)}; {1.117%: نظرا ل/nZrA l (9)}; {1.1%: اثر/Avr (9)}; {0.7%: حتى/HtY (6)}; {0.6%: لكي/lky (5)}; {0.620%: يفضل/bfDl (5)}; {0%: بغية/bgyp (5)}; {0.4%: لأن/l>n (3)}; {0.372%: بعد/bEd (3)}; {0.3%: نتيجة/ntyjp (2)}; {0.248%: لا سيما/lA symA (2)}; {0.3%: كي/ky (2)}; {0.3%: عندما/EndmA (2)}; {0.248%: خصوصا/xSwSA (2)}; {0.1%: لذا/l*A (1)}; {0.1%: في ظل/fy Zl (1)}; {0.1%: انما/AnmA (1)} | 806 | 26 |
| COMPARISON.Contrast | {45%: لكن/lkn (198)}; {12.5%: و/w (55)}; {8.6%: الا ان/AlA An (38)}; {7.1%: رغم/rgm (31)}; {4.1%: اما/AmA (18)}; {3%: فيما/fymA (13)}; {2.7%: في حين/fy Hyn (12)}; {2.5%: بل/bl (11)}; {2.1%: على الرغم/ElY Alrgm (9)}; {1.4%: غير ان/gyr An (6)}; {1.4%: رغم ان/rgm An (6)}; {1.4%: حتى/HtY (6)}; {1.4%: بينما/bynmA (6)}; {1.4%: بيد ان/byd An (6)}; {1.2%: في المقابل/fy AlmqAbl (5)}; {1.2%: بالرغم من/bAlrgm mn (5)}; {0.9%: انما/AnmA (4)}; {0.5%: لو/lw (2)}; {0.5%: كما/kmA (2)}; {0.5%: ف/f (2)}; {0.227%: بيد/byd (1)}; {0.2%: خلافا ل/xlAfA l (1)}; {0.2%: بالمقابل/bAlmqAbl (1)}; {0.2%: برغم/brgm (1)}; {0.2%: ب/b (1)} | 440 | 25 |
| TEMPORAL.Asynchronous | {26.2%: و/w (109)}; {24%: بعد/bEd (100)}; {20%: قبل/qbl (83)}; {11.5%: منذ/mn* (48)}; {8%: ثم/vm (33)}; {2.2%: اثر/Avr (9)}; {1.679%: بعدما/bEdmA (7)}; {1.4%: ف/f (6)}; {1.2%: قبيل/qbyl (5)}; {1%: عندما/EndmA (4)}; {0.7%: كما/kmA (3)}; {0.7%: قبل ان/qbl An (3)}; {0.5%: من ثم/mn vm (2)}; {0.5%: عقب/Eqb (2)}; {0.24%: وقبل/wqbl (1)}; {0.2%: ايضا/AyDA (1)}; {0.2%: اما/AmA (1)} | 417 | 17 |
| EXPANSION.Reformulation | {86.7%: و/w (287)}; {0.3%: بل/bl (1)} {5.4%: ف/f (18)}; {2.7%: خصوصا/xSwSA (9)}; {2.1%: حيث/Hyv (7)}; {1.5%: اذ/A* (5)}; {0.3%: لكن/lkn (1)}; {0.3%: لا سيما/lA symA (1)}; {0.3%: كما/kmA (1)}; {0.3%: بمعنى آخر/bmEnY Axr (1)}; | 331 | 10 |

| Discourse Relation | Discourse connective | Total (6,328) | #Dis. Conn |
|---|---|---|---|
| CONTINGENCY.Cause.Result. NonPragmatic | {58.8%: و/w (134)}; {12.8%: ف/f (29)}; {11%: ل/l (25)}; {9.2%: مما/mmA (21)}; {5.3%: بالتالي/bAltAly (12)}; {1.3%: حتى/HtY (3)}; {0.4%: لذلك/l*lk (1)}; {0.4%: حيث/Hyv (1)}; {0.4%: بحيث/bHyv (1)}; {0.4%: انما/AnmA (1)} | 228 | 10 |
| TEMPORAL.Synchronous | {46.6%: خلال/xlAl (102)}; {13.2%: و/w (29)}; {12.8%: عندما/EndmA (28)}; {5%: منذ/mn* (11)}; {4.1%: كما/kmA (9)}; {3.7%: فيما/fymA (8)}; {3.7%: بينما/bynmA (8)}; {3.2%: في حين/fy Hyn (7)}; {2.3%: ب/b (5)}; {1.4%: ف/f (3)}; {1.4%: حيث/Hyv (3)}; {0.9%: في ظل/fy Zl (2)}; {0.9%: حينها/HynhA (2)}; {0.5%: حين/Hyn (1)}; {0.457%: حتى/HtY (1)} | 219 | 15 |
| EXPANSION.Background | {99%: و/w (184)}; {0.6%: ف/f (1)}; {0.6%: حيث/Hyv (1)} | 186 | 3 |
| CONTINGENCY.Cause.Reason. NonPragmatic/ TEMPORAL.Asynchronous | {48.4%: بعد/bEd (76)}; {31.2%: اثر/Avr (49)}; {7.6%: بعدما/bEdmA (12)}; {5.1%: منذ/mn* (8)}; {2.6%: عندما/EndmA (4)}; {1.3%: و/w (2)}; {1.3%: عقب/Eqb (2)}; {0.6%: قبل/qbl (1)}; {0.6%: خصوصا/xSwSA (1)}; {0.6%: حتى/HtY (1)}; {0.6%: ثم/vm (1)} | 157 | 11 |
| CONTINGENCY.Condition | {45.5%: في حال/fy HAl (35)}; {41.6%: اذا/A*A (32)}; {2.6%: لو/lw (2)}; {2.6%: طالما/TAlmA (2)}; {1.3%: و/w (1)}; {1.3%: لولا/lwlA (1)}; {1.3%: عندما/EndmA (1)}; {1.3%: حتى/HtY (1)}; {1.3%: حال/HAl (1)}; {1.3%: الا اذا/AlA A*A (1)} | 77 | 10 |
| EXPANSION.Exemplification | {51.1%: و/w (24)}; {25.5%: ف/f (12)}; {10.7%: خصوصا/xSwSA (5)}; {4.3%: لا سيما/lA symA (2)}; {2.1%: كما/kmA (1)}; {2.1%: كأن/k>n (1)}; {2.1%: حيث/Hyv (1)}; {2.1%: اذ/A* (1)} | 47 | 8 |
| CONTINGENCY.Cause.Result. Pragmatic | {69.7%: و/w (23)}; {9.1%: ل/l (3)}; {6.1%: حتى/HtY (2)}; {6.1%: بالتالي/bAltAly (2)}; {3.1%: لو/lw (1)}; {3.1%: ف/f (1)}; {3.1%: اذ/A* (1)} | 33 | 7 |
| CONTINGENCY.Cause.Reason. Pragmatic | {39.286%: و/w (11)}; {14.%: ف/f (4)}; {14.3%: ب/b (4)}; {10.7%: ل/l (3)}; {7.1%: خصوصا/xSwSA (2)}; {7.1%: حيث/Hyv (2)}; {3.6%: نتيجة/ntyjp (1)}; {3.571%: اذ/A* (1)} | 28 | 8 |
| EXPANSION.Alternative.Conjunctive | {100%: او/Aw (28)} | 28 | 1 |
| CONTINGENCY.Cause.Result. NonPragmatic/ TEMPORAL.Asynchronous | {54.5%: و/w (12)}; {18.2%: حتى/HtY (4)}; {9.1%: من ثم/mn vm (2)}; {9.1%: عندما/EndmA (2)}; {4.5%: ف/f (1)}; {4.5%: ثم/vm (1)} | 22 | 6 |
| COMPARISON.Contrast/ TEMPORAL.Synchronous | {36.8%: في حين/fy Hyn (7)}; {31.6%: فيما/fymA (6)}; {15.8%: و/w (3)}; {10.5%: بينما/bynmA (2)}; {5.3%: بالمقابل/bAlmqAbl (1)} | 19 | 5 |
| CONTINGENCY.Cause.Reason. Pragmatic/ TEMPORAL.Asynchronous | {64.3%: بعد/bEd (9)}; {14.3%: بعدما/bEdmA (2)}; {7.1%: و/w (1)}; {7.1%: منذ/mn* (1)}; {7.1%: عقب/Eqb (1)} | 14 | 5 |
| CONTINGENCY.Cause.Reason. NonPragmatic/ | {64.3%: عندما/EndmA (9)}; {21.4%: في ظل/fy Zl (3)}; {7.1%: فيما/fymA (1)}; {7.1%: طالما/TAlmA (1)} | 14 | 4 |

| Discourse Relation | Discourse connective | Total (6,328) | #Dis. Conn |
|---|---|---|---|
| TEMPORAL.Synchronous | | | |
| COMPARISON.Similarity | {64.3%: كما/kmA (9)}; {35.8%: و/w (5)} | 14 | 2 |
| COMPARISON.Contrast/TEMPORAL.Asynchronous | {36.4%: بعد/bEd (4)}; {18.2%: و/w (2)}; {18.2%: بعدما/bEdmA (2)}; {9.1%: فيما/fymA (1)}; {9.1%: عندما/EndmA (1)}; {9.1%: ثم/vm (1)} | 11 | 6 |
| EXPANSION.Alternative.Disjunctive | {100%: او/Aw (7)} | 7 | 1 |
| CONTINGENCY.PragmaticCondition | {33.3%: حتى/HtY (2)}; {33.3%: حتى لو/HtY lw (2)}; {16.7%: لو/lw (1)}; {16.7%: اذا/A*A (1)} | 6 | 4 |
| EXPANSION.Exception | {40%: الا/AlA (2)}; {20%: و/w (1)}; {20%: الا ان/AlA An (1)}; {20%: الا اذا/AlA A*A (1)} | 5 | 4 |
| CONTINGENCY.Cause.Reason.NonPragmatic/ EXPANSION.Reformulation | {60%: خصوصا/xSwSA (3)}; {40%: اذ/A* (2)} | 5 | 2 |
| EXPANSION.Exception/ TEMPORAL.Asynchronous | {80%: الا بعد/AlA bEd (4)}; {20%: الا/AlA (1)} | 5 | 2 |
| CONTINGENCY.Cause.Reason.Pragmatic/ TEMPORAL.Synchronous | {33.3%: منذ/mn* (1)}; {33.3%: كما/kmA (1)}; {33.3%: بعد/bEd (1)} | 3 | 3 |
| CONTINGENCY.Condition/ TEMPORAL.Synchronous | {66.7%: عندما/EndmA (2)}; {33.3%: كلما/klmA (1)} | 3 | 2 |
| EXPANSION.Exemplification/ EXPANSION.Reformulation | {100%: و/w (3)} | 3 | 1 |
| CONTINGENCY.Cause.Reason.NonPragmatic/ EXPANSION.Exemplification | {50%: ف/f (1)}; {50%: حيث/Hyv (1)} | 2 | 2 |
| CONTINGENCY.Cause.Reason.Pragmatic/ EXPANSION.Reformulation | {50%: خصوصا/xSwSA (1)}; {50%: اذ/A* (1)} | 2 | 2 |
| CONTINGENCY.Cause.Result.Pragmatic/ TEMPORAL.Asynchronous | {50%: و/w (1)}; {50%: بعد/bEd (1)} | 2 | 2 |
| COMPARISON.Similarity/ EXPANSION.Exemplification | {50%: و/w (1)}; {50%: كما/kmA (1)} | 2 | 2 |
| COMPARISON.Contrast/ CONTINGENCY.Cause.Reason.NonPragmatic | {100%: لكن/lkn (2)} | 2 | 1 |
| COMPARISON.Contrast/ CONTINGENCY.Cause.Result.NonPragmatic | {100%: و/w (2)} | 2 | 1 |

| Discourse Relation | Discourse connective | Total (6,328) | #Dis. Conn |
|---|---|---|---|
| COMPARISON.Contrast/ COMPARISON.Similarity | {100%: كما/kmA (1)} | 1 | 1 |
| COMPARISON.Contrast/ CONTINGENCY.Condition | {100%: حتى/HtY (1)} | 1 | 1 |
| COMPARISON.Contrast/ CONTINGENCY.PragmaticCondition | {100%: و/w (1)} | 1 | 1 |
| COMPARISON.Contrast/ EXPANSION.Background | {100%: و/w (1)}; | 1 | 1 |
| COMPARISON.Contrast/ EXPANSION.Exception | {100%: لكن/lkn (1)} | 1 | 1 |
| COMPARISON.Contrast/ EXPANSION.Reformulation | {100%: الا ان/AlA An (1)} | 1 | 1 |
| CONTINGENCY.Cause.Reason. NonPragmatic/ CONTINGENCY.Condition | {100%: طالما/TAlmA (1)} | 1 | 1 |
| CONTINGENCY.Cause.Reason. NonPragmatic/ EXPANSION.Background | {100%: و/w (1)} | 1 | 1 |
| CONTINGENCY.Cause.Reason. Pragmatic/ EXPANSION.Background | {100%: ف/f (1)} | 1 | 1 |
| CONTINGENCY.Cause.Reason. Pragmatic/ EXPANSION.Exemplification | {100%: حيث/Hyv (1)} | 1 | 1 |
| CONTINGENCY.Cause.Result. NonPragmatic/ EXPANSION.Background | {100%: و/w (1)} | 1 | 1 |
| CONTINGENCY.Cause.Result. NonPragmatic/TEMPORAL.Synchronous | {100%: حتى/HtY (1)} | 1 | 1 |
| CONTINGENCY.Cause.Result. Pragmatic/ CONTINGENCY.Condition | {100%: و/w (1)} | 1 | 1 |
| CONTINGENCY.Cause.Result. Pragmatic/ EXPANSION.Conjunction | {100%: و/w (1)} | 1 | 1 |
| CONTINGENCY.Condition/ EXPANSION.Exception | {100%: اذا/A*A (1)} | 1 | 1 |
| CONTINGENCY.Condition/ TEMPORAL.Asynchronous | {100%: الا بعد/AlA bEd (1)} | 1 | 1 |

| Discourse Relation | Discourse connective | Total (6,328) | #Dis. Conn |
|---|---|---|---|
| EXPANSION.Exemplification/ TEMPORAL.Asynchronous | {100%: و/w (1)} | 1 | 1 |
| COMPARISON.Similarity/ EXPANSION.Reformulation | {100%: عندما/EndmA (1)} | 1 | 1 |
| EXPANSION.Reformulation/ TEMPORAL.Asynchronous | {100%: و/w (1)} | 1 | 1 |
| EXPANSION.Reformulation/ TEMPORAL.Synchronous | {100%: كما/kmA (1)} | 1 | 1 |
| COMPARISON.Similarity/ TEMPORAL.Asynchronous | {100%: فيما/fymA (1)} | 1 | 1 |

# Appendix E

<publicationStmt>

<distributor>

The discourse annotation tool (READ: Relation annotation for English and Arabic Discourse) was developed by Amal Al-Saif (University of Leeds, UK) and Basmah Asoli. (King Abdul Aziz City for Science and Technology (KACST), Saudi Arabia).

Amal Al-Saif was funded by Imam Muhammad Ibn Saud Islamic University and Basmah Asoli by KACST.

Tool development was partially conducted within the project "A Digital Resource for Discourse Relations in Modern Standard Arabic" (British Academy, SG51944) led by Dr Katja Markert, University of Leeds, Great Britain.

</distributor>

<address> Amal Al-Saif, School of Computing, University of Leeds,

Woodhouse Lane, Leeds LS2 9JT. Email: assaif@comp.leeds.ac.uk or amalalsaif@yahoo.co.uk

Katja Markert, School of Computing, University of Leeds,

Woodhouse Lane, Leeds LS2 9JT. Email: markert@comp.leeds.ac.uk</address>

<availability> You are free to use the annotation tool without charge for all non-commercial purposes under the following conditions: a) Any published material or results using in any way this annotation tool must sufficiently acknowledge the developers with the following wording "We used the annotation tool READ, developed by Amal Al-Saif and Katja Markert, University of Leeds, UK." as well as by citing "Amal Al-Saif, Katja Markert: The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. LREC 2010".

<date> 20.7.2011 </date>

# Appendix F

# The Representation Format of the LADTB Annotation

## 1 Introduction

We describe in this section a representation format of the annotation in the LADTB and the structure of sub-directories in the distribution and how to be linked to the syntactic annotation in the ATB. In general, we followed a similar format of the PDTB annotation for more consistency of the two corpora. However, some useful information was added in our annotation such as POS of the connective, the sequence of trees and words of the connective and the arguments as in the ATB.

## 2 Directory structure

The package has three main directories:

1) *data* directory, which has two subdirectories:
    a. *Text* - refers to the raw text of the LADTB. There are two types of raw text in two folders (i) *ATB_P1_Sgm* contains 537 raw (sgm) files of the Arabic Treebank Part1 without any modifications; they are only the raw files without ATB annotation. (ii) *Raw_without_HTML_tags* folder contains the same raw files but after removing all HTML tags using the attached python program Removing_HTML_tags in *tools* directory.
    *b.* *LADTB_annotation* - refers to the annotated files of the LADTB. The files have similar reference number of the ATB in *Text* directory but with an extension (.ladtb).
2) *doc* directory, which contains:
    a. A text file *list_of_annotated_files.txt* – contains a list of annotated files of this release of the LADTB.
    b. A text file *Files_without_discourseAnnotation.txt* which contains a list of files that do not have any discourse annotations from the 537 ATB files that

we annotated in the LADTB. These files are completely empty in *LADTB_annotation* folder.

c. *Annotation manual.pdf* contains our guidelines for discourse annotation in the LADTB.

d. The published paper *LADTB_LREC2010.pdf* in LREC2010 which describes in brief this first discourse annotation for Arabic connectives.

3) *tools* directory which contains:

a. *READ_Tool* contains the new developed discourse annotation tool for Arabic and English in an executable JAR file *AnnotationTool.jar*. there are also two essential text files: *conn.txt* (contains a list of all potential discourse connectives for Arabic), *conn_clitic.txt (*contains connectives could be clitics in the text), and *conn_eng.txt* (contains English potential discourse connectives of the PDTB. The tool uses those files to highlight the potential connectives in the text. A *Copyright-tool* licence is included too in the directory

b. *Removing_HTML_tags.py:* a python program to remove html tags from the raw files of the ATB. The program should read a list of files in */docs/list_of_annotated_files.txt* and generate new files with an extension (.raw) in the subdirectory */data/Raw_without_HTML_tags*. The indices in the LADTB annotation files and the tool lie on raw files without html tags.

# 3 Linking mechanism of the LADTB and the ATB

The annotated files in the LADTB do have only the discourse annotation of the connectives and associated relations and arguments, using similar reference of the files in the ATB. The two annotations and the raw files are linked via different ways:

1. The *indices* of starting and ending characters of connectives and the two arguments Arg1 and Arg2 in the raw file, after removing HTML tags.
2. The *Gorn address* of each token of connectives and arguments in the ATB. Section 0 illustrates the method of generating these indicators.
3. The token sequence in the ATB Part1 v.2 of tokens of connectives and the two arguments Arg1 and Arg2. The sequence starts with 1 to represent the first tree of the first sentence in the file, excluding trees starting with (X.. ). A sequence

of tokens starts also with 1, to represent the first token the tree and the sentence.

# 4 General outline of the annotation

The explicit connectives are annotated in order of their appearing in the raw file. As shown in Figure . Each annotation is following a format of four parts:

**Part1 (Explicit Conn)** presents the annotation of a connective using information from a raw text, and the syntactic annotation of ATB. The Arabic_Connective_String, the indices Raw_start_index..Raw_end_index are extracted from the raw file, and the Connective_String_Buckwalter_form, and token sequence HostingTree_Sequence_ATB, Word_Sequence_ATB, and Gorn_address_list are extracted from the ATB file.

**Part2 (Features)** presents features belong to the connective. It includes:

- A syntactic feature (POS, extracted from the ATB)
- Surface features (connective type {Simple, Clitic and MoreThanToken} and arguments order{Arg1_Conn_Arg2, Conn_Arg2_Arg1 and Arg1_Conn_Arg2_Arg1 }
- and the discourse function of the connective, single or multiple discourse relations from our the LADTB relations taxonomy.

**Part3 (Arg1)** presents annotation of the first argument, from both raw texts and ATB annotation. Starting and ending indices were extracted from raw text. While the rest of the annotation are extracted from ATB annotation: Gorn_address_list of tokens, tree sequence and tokens sequence (HostingTree_Sequence_ATB, Word_Sequence_ATB), tokens as presented in the ATB (ATB_span_of_Arg1_Arabic) and their buckwalter forms (ATB_span_of_Arg1_ Buckwalter_form).

**Part4 (Arg2)** presents the annotation of Arg2 in a similar format of Part3 of Arg1.

In Part3/4, the arguments (Arg1/Arg2) might consist of more than one sentence which are represented by more than one tree in the ATB. The annotation of each line therefore covers all segments of the argument separated by semi-colon (;), except the

line of argument's indices, which has one span. However, for cases of the argument order Arg1_Conn_Arg2_Arg1, there should be two indices sets of the argument Arg1; an indices set for the first part and the other for the second part.

---

##### Explicit Conn #####

Connective_String_Arabic; Connective_String_Buckwalter_form

Raw_start_index..Raw_end_index; HostingTree_Sequence_ATB; Word_Sequence_ATB; Gorn_address_list

##### FEATURES #####

Connective_POS; Connective_Type; Discourse_Relation(s)

Arguments_order

##### ARG1 #####

Raw_start_index .. Raw_end_index

HostingTree_Sequence_ATB ; Word_Sequence_ATB; Gorn_address_list

ATB_span_of_Arg1_Arabic

ATB_span_of_Arg1_ Buckwalter_form

##### ARG2 #####

Raw_start_index .. Raw_end_index

HostingTree_Sequence_ATB ; Word_Sequence_ATB; Gorn_address_list

ATB_span_of_Arg2_Arabic

ATB_span_of_Arg2_ Buckwalter_form

---

**Figure 1: Format of the annotation in the LADTB of one explicit connective**

# 5 Gorn address

"*Gorn address is a method of addressing an interior node within a tree from a phrase structure rule description or parse tree*" *(Gorn, 1967)* [1].

The Gorn address is a series of one or more integers separated by comma, e.g., 0 or 0,0,1. Many programming languages access to nodes in a tree structure using Gorn

[1] http://en.wikipedia.org/wiki/Gorn_address

address technique. Thus the Gorn addresses of connectives and Arg1 and Arg2 in the LADTB are generated automatically using Python modules - NLTK[2]. Figure 2 shows the Gorn address of all internal nodes in a parse tree of a clause (فدخل خالد مسرعاً/ then Kald entered quickly). We consider parent node of only lexical items in GornAddressList. For example the address of (خالد/Kald) is 1,1,0.

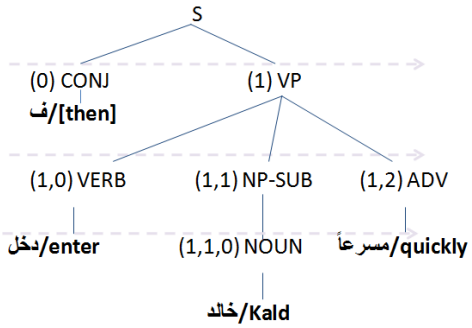



**Figure.2 A parse tree of a clause showing computiation of the Gorn address**

# 6 Sample of LADTB annotation

The following annotation in Ex belongs to a simple subordinating connective, the preposition (بعد/bEd/after) in file *20000715_AFP_ARB_0001.ladtb*. It is annotated with the discourse relation TEMPORAL.Asynchronous and the arguments order is Arg1_Conn_Arg2. Figure 3 shows the equivalent ATB annotation, with gorn address and word_sequence of each token of Arg1 and Arg2.

**Ex 1**

*غادر كنت (45 عاما) مساء الاربعاء المدينة متوجها الى ولاية اوهايو (شمال شرق)* بعد **ان استقل احد باصات شركة غريهاوند الشهيرة التي تجوب كل الولايات الاميركية**

*Kent (45 years) left the city on Wednesday evening on his way to the state of Ohio (North East)* **after he picked up a bus of the famous Greyhound company, which roams all the U.S. states.**

The LADTB annotation:

---

##### Explicit Conn #####

[2] http://www.nltk.org/

بعد ; bEd

486..489 ; Tree No:2 ; Seq:{19}; (1, 5, 0)

##### FEATURES #####

PREP ; Simple ; TEMPORAL.Asynchronous

Arg1_Conn_Arg2

##### ARG1 #####

410..484

Tree No:2 ;Seq:{2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18}; (1, 0), (1, 1, 0, 0), (1, 1, 1, 0), (1, 1, 1, 1, 0), (1, 1, 1, 1, 1), (1, 1, 1, 2), (1, 2, 0), (1, 2, 1, 0), (1, 3, 0), (1, 4, 0, 0), (1, 4, 0, 1, 0), (1, 4, 0, 2, 0), (1, 4, 0, 2, 1, 0, 0), (1, 4, 0, 2, 1, 0, 1, 0), (1, 4, 0, 2, 1, 1, 0), (1, 4, 0, 2, 1, 1, 1, 0, 0), (1, 4, 0, 2, 1, 1, 1, 1, 0), (1,4,0,2,1,1,2)

شمال شرق -LRB- مساء الاربعاء المدينة متوجها الى ولاية اوهايو -RRB- عاما 45 -LRB- غادر كنت

gAdr knt -LRB- 45 EAmA -RRB- msA' AlArbEA' Almdynp mtwjhA AlY wlAyp AwhAyw -LRB- $mAl $rq

##### ARG2 #####

490..562

Tree No:2 ;Seq:{20,21,22,23,24,25,26,27,28,29,30,31}; (1, 5, 1), (1, 5, 2, 0, 0), (1, 5, 2, 0, 1, 0), (1, 5, 2, 0, 2, 0), (1, 5, 2, 0, 2, 1, 0), (1, 5, 2, 0, 2, 1, 1, 0, 0), (1, 5, 2, 0, 2, 1, 1, 1, 0), (1, 5, 2, 0, 2, 1, 1, 2, 0), (1, 5, 2, 0, 2, 1, 1, 3, 0, 0), (1, 5, 2, 0, 2, 1, 1, 3, 1, 0, 0), (1, 5, 2, 0, 2, 1, 1, 3, 1, 0, 1, 0), (1, 5, 2, 0, 2, 1, 1, 3, 1, 0, 2, 0), (1, 5, 2, 0, 2, 1, 1, 3, 1, 0, 2, 1), (1, 5, 2, 0, 2, 1, 1, 3, 1, 0, 2, 2)

ان استقل احد باصات شركة غريهاوند الشهيرة التي تجوب كل الولايات الاميركية

An Astql AHd bASAt $rkp gryhAwnd Al$hyrp Alty tjwb kl AlwlAyAt AlAmyrkyp

---

| | Token | Gorn address | The Penn Arabic Treebank including number of tree, and token |
|---|---|---|---|
| | - | | (S |
| Arg1 | 1 | 0 | (CONJ 2_1_ w_و) |
| | - | 1 | (VP |
| | 2 | 1,0 | (VERB_PERFECT 2_2_ gAdr_غادر) |
| | - | 1,1 | (NP-SBJ-1 |
| | 3 | 1,1,0,0 | (NP (NOUN_PROP 2_3_ knt_كنت)) |
| | - | 1,1,1 | (PRN |
| | 4 | 1,1,1,0 | (PUNC 2_4_ -LRB-_-LRB-) |
| | - | 1,1,1,1 | (NP |
| | 5 | 1,1,1,1,0 | (NUM 2_5_ 45_45) |
| | 6 | 1,1,1,1,1 | (NOUN+NSUFF_MASC_SG_ACC_INDEF 2_6_ EAmA_عاما)) |
| | 7 | 1,1,1,2 | (PUNC 2_7_ -RRB-_-RRB-))) |
| | - | 1,2 | (NP-TMP |
| | 8 | 1,2,1 | (NOUN 2_8_ msA_مساء) |
| | 9 | 1,2,2,0 | (NP (DET+NOUN_PROP 2_9_ AlArbEA'_الاربعاء))) |
| | 10 | 1,3,0 | (NP-OBJ (DET+NOUN+NSUFF_FEM_SG 2_10_ Almdynp_المدينة)) |
| | - | 1,4 | (S-ADV |
| | - | 1,4,0 | (VP |
| | 11 | 1,4,0,0 | (NOUN+NSUFF_MASC_SG_ACC_INDEF 2_11_ mtwjhA_متوجها) |
| | - | 1,4,0,1,0 | (NP-SBJ-1 (-NONE- *)) |
| | - | 1,4,0,2 | (PP-DIR |
| | 12 | 1,4,0,2,0 | (PREP 2_12_ AlY_الى) |
| | - | 1,4,0,2,1 | (NP |

```
        -       1,4,0,2,1,0             (NP
        13      1,4,0,2,1,0,0            (NOUN+NSUFF_FEM_SG 2_13_ wlAyp_ولاية)
        14      1,4,0,2,1,0,1,0          (NP (NO_FUNC 2_14_ AwhAyw_اوهايو)))
        -       1,4,0,2,1,1             (PRN
        15      1,4,0,2,1,1,0            (PUNC 2_15_ -LRB-_-LRB-)
        -       1,4,0,2,1,1,1            (NP
        16      1,4,0,2,1,1,1,0          (NP (NOUN 2_16_ $mAl_شمال))
        17      1,4,0,2,1,1,1,1,0        (NP (NOUN_PROP 2_17_ $rq_شرق)))
        18      1,4,0,2,1,1,2            (PUNC 2_18_ -RRB-_-RRB-))))))

        -       1,5                   (SBAR-TMP
Conn:   19      1,5,0                  (PREP 2_19_ bEd_بعد)
        20      1,5,1                  (FUNC_WORD 2_20_ An_ان)
        -       1,5,2                  (S
        -       1,5,2,0                 (VP
        21      1,5,2,0,0                (VERB_PERFECT 2_21_ Astql_استقل)
        -       1,5,2,0,1,0              (NP-SBJ (-NONE- *))
        -       1,5,2,0,2                (NP-OBJ
        22      1,5,2,0,2,0               (NOUN 2_22_ AHd_احد)
        -       1,5,2,0,2,1               (NP
        23      1,5,2,0,2,1,0              (NOUN+NSUFF_FEM_PL 2_23_ bASAt_باصات)
Arg2    -       1,5,2,0,2,1,1              (NP
        24      1,5,2,0,2,1,1,0,0           (NP (NOUN+NSUFF_FEM_SG 2_24_ $rkp_شركة))
        25      1,5,2,0,2,1,1,1,0           (NP (NO_FUNC 2_25_ gryhAwnd_غريهاوند))
        -       1,5,2,0,2,1,1,2             (ADJP
        26      1,5,2,0,2,1,1,2,0            (DET+ADJ+NSUFF_FEM_SG 2_26_ Al$hyrp_الشهيرة))
        -       1,5,2,0,2,1,1,3             (SBAR
        27      1,5,2,0,2,1,1,3,0            (WHNP-2 (REL_PRON 2_27_ Alty_التي))
        -       1,5,2,0,2,1,1,3,1            (S
        -       1,5,2,0,2,1,1,3,1,0           (VP
        28      1,5,2,0,2,1,1,3,1,0,0          (IV3FS+VERB_IMPERFECT 2_28_ tjwb_تجوب)
        -       1,5,2,0,2,1,1,3,1,0,1,0        (NP-SBJ-2 (-NONE- *T*))
        -       1,5,2,0,2,1,1,3,1,0,2          (NP-OBJ
        29      1,5,2,0,2,1,1,3,1,0,2,0         (NOUN 2_29_ kl_كل)
        30      1,5,2,0,2,1,1,3,1,0,2,1         (DET+NOUN+NSUFF_FEM_PL 2_30_ AlwlAyAt_الولايات)
        31      1,5,2,0,2,1,1,3,1,0,2,2         (DET+ADJ+NSUFF_FEM_SG 2_31_ AlAmyrkyp_الاميركية))))))))))))
        32      1,6                    (PUNC 2_32_ ._.))
```

**Figure 3: A sample of the ATB annotation with corresponding word sequences and Gorn addresses of connective and the two arguments (Arg1 and Arg2) of the example in Ex .**