The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic

Amal Al-Saif, Katja Markert

School of Computing
University of Leeds, Leeds, UK, LS2 9JT
assaif@comp.leeds.ac.uk, markert@comp.leeds.ac.uk

Abstract
We present the first effort towards producing an Arabic Discourse Treebank, a news corpus where all discourse connectives are identified and annotated with the discourse relations they convey as well as with the two arguments they relate. We discuss our collection of Arabic discourse connectives as well as principles for identifying and annotating them in context, taking into account properties specific to Arabic. In particular, we deal with the fact that Arabic has a rich morphology: we therefore include clitics as connectives as well as a wide range of nominalizations as potential arguments. We present a dedicated discourse annotation tool for Arabic and a large-scale annotation study. We show that both the human identification of discourse connectives and the determination of the discourse relations they convey is reliable. Our current annotated corpus encompasses a final 5651 annotated discourse connectives in 537 news texts. In future, we will release the annotated corpus to other researchers and use it for training and testing automated methods for discourse connective and relation recognition.

1. Introduction
Discourse relations such as CAUSAL or CONTRAST relations between textual units play an important role in producing a coherent discourse. They are widely studied in theoretical linguistics (Halliday and Hasan, 1976; Hobbs, 1985), where also different relation taxonomies have been derived (Hobbs, 1985; Knott and Sanders, 1998; Mann and Thompson, 1988; Marcu, 2000). Discourse relations can be signalled by explicit lexical indicators, so-called discourse connectives (Marcu, 2000; Webber et al., 1999; Prasad et al., 2008a). We follow (Prasad et al., 2008a) in defining discourse connectives as lexical expressions that relate two text segments that express abstract entities such as events, belief, facts or propositions. These text segments are called the arguments of the discourse connective. In Ex. 1 the connective because indicates a CAUSAL relation between Jack not getting a high mark and his fatigue at exam time. In Ex. 2, the connective however indicates a CONTRAST relation. We indicate discourse connectives and the two arguments via annotated square brackets.

1 [Because]_
   [he was very tired during the exam.]_
   [Arg2]
   [Jack did not achieve a high mark.]_
   [Arg1]

2 [The TV was broken.]_
   [However]_
   [I was able to fix it.]_

Discourse connectives are often used as an important feature in the automatic recognition of discourse relations, a task useful for many applications such as automatic summarization, question answering and text generation (Hovy, 1993; Marcu, 2000). Recently, to enable corpus studies and automatic discourse relation recognition algorithms, the Penn Discourse Treebank (PDTB) has been developed (Prasad et al., 2008a) – an English corpus which is annotated for discourse connectives, the relations they convey, which they call senses, and their arguments. One of its main attractions is that its annotation is theory-neutral (for example, it does not subscribe to any restrictions on the distance of the two arguments of a connective). It has also been shown to be extensible to other languages such as Hindi (Prasad et al., 2008b), Turkish (Zeyrek and Webber, 2008) and Chinese (Xue, 2005).

We extend these efforts to Modern Standard Arabic (MSA) by producing the Leeds Arabic Discourse Treebank (LADTB). The remainder of this paper is organized as follows: Section 2, describes related work. Section 3, describes our methodology for collecting the potential discourse connectives in MSA. A brief description of the annotation scheme follows in Section 4. The corpus, annotation tool, and the annotation methodology are discussed in Section 5. The results of our agreement studies and the gold standard corpus

---

1 The PDTB project has been extended to also annotate implicit discourse relations, i.e. discourse relations which are not indicated via discourse connectives. In this first study for Arabic, we focus on discourse relations signaled explicitly via connectives.
2. Related Work

Several textual corpora of Arabic exist. Some of them are available with Part-of-Speech and syntactic annotation such as the Arabic Treebank (ATB) (Maamouri and Bies, 2004). The Prague Arabic Dependency Treebank (PADT), which is smaller in scale than the ATB, contains multilevel annotations, including morphological and analytical level of linguistic representation (Hajic et al., 2004). Moreover, a recent effort by Dukes and Habash (2010) has produced The Quranic Arabic Corpus, a free annotated linguistic resource which provides morphological annotation and syntactic analysis (using dependency grammar) of the Holy Quran.

Surprisingly, the annotation level of existing Arabic corpora has not yet included the discourse layer. Al-Sanie et al. (2005) and Seif et al. (2005) discuss a limited set of rhetorical relations and discourse connectives. However, they did not distinguish between discourse connectives such as لأن /tan/because and other syntactic connectors such as prepositions like في /lyin or مع /mt/with, where the latter signal a semantic relation between two concrete objects instead of a discourse relation between abstract entities such as clauses or sentences. Moreover, the studies had a small empirical basis using only a small number of Arabic texts and no agreement studies on identification of discourse connectives and relations in context have been carried out. Therefore, our work is the first principled discourse annotation effort for Arabic.

We work on the syntactically annotated Arabic Penn Treebank v.2 (Maamouri and Bies, 2004), which we extend to a discourse-level resource by identifying its explicit discourse connectives and annotating them with the discourse relations they convey as well as their arguments. We based our annotation guidelines on the same principles as the PDTB but adapt and expand the annotation to take into account properties specific to Arabic.

3. Collecting Arabic Connectives

Although several references in the Arabic literature (Al-Warraki and Hasanayn, 1994; Ryding, 2005; Alansari, 1985; Alfarabi, 1990) point out the discourse usage of connectives such as لأن /tan/because and لكن /kn/ but, no single exhaustive list of Arabic discourse connectives exist.

We collected a large set of Arabic discourse connectives using text analysis and corpus-based techniques. We enhanced the ones mentioned in the literature with manual extraction of all connectives from 50 randomly selected texts from the Penn Arabic Treebank and from 10 different web sites. In addition, we extracted all lexical items with connective-typical POS tags (such as conjunctions) automatically from the Penn Arabic Treebank (Al-Saif et al., 2009). The resulting list was manually verified by two Arabic native speakers.

Our final list contains 91 basic Arabic discourse connectives, enhanced with 16 modified forms of basic connectives (such as إذا /hā /then/ if as a modified form of إذا /hā /then/ if), yielding 107 discourse connectives overall. This number is comparable to the number of 100 distinct English connectives in the PDTB. Table 4 shows the most frequent connectives in the LADTB.

4. Annotation Scheme

We followed the annotation principles in the PDTB as far as possible. Necessary adaptations were made to take into account properties specific to Arabic. PDTB annotation is based on lexicalized grammar theory. The anchor of the annotation is the lexical item - a discourse connective (DC). The argument labels of the signalled relation are partially syntactically driven, in that the Arg2 label is assigned to the argument with which the connective was syntactically associated. The Arg1 label, however, can refer to an abstract object at any distance from the connective.

4.1. Types of Discourse Connectives

Discourse connectives in the PDTB are coordinating or subordinating conjunctions such as and, but and or, adverbials such as then, later and otherwise, and prepositional phrases such as in contrast and as a result. All these are also used for MSA (see Examples 3, 4 and 5).
However, our analysis shows that, in addition, many typical discourse relations are expressed in Arabic via prepositions where normally only one argument of the connective is a nominalization (Al-Mazdar).\textsuperscript{3} Thus, in Ex. 6  

\textbf{Ex. 6}\textsuperscript{1} 

\textit{Al-Mazdar}\textsuperscript{3} \textit{in Arabic, these might be one or more, tensed or untensed, verbal sentences or clauses \textit{gmlh fylh}, noun sentences \textit{gmlh smyh}, anaphoric expressions (if they refer to an abstract object such as many demonstrative pronouns) or verb ellipses.

\textsuperscript{3}Al-Mazdar is a well defined noun category in the Arabic literature with 58 noun forms.

\textbf{Figure 1: Discourse relations in the LADTB.}

The main difference to English is the inclusion of certain noun sentences. The Arabic noun sentence is equivalent to one of two English sentences/clauses: (i) a verbal phrase of the form (x verb-to-be y) (such as \textit{the university is famous}) or (ii) a noun phrase (such as \textit{famous university}). The latter is normally not an abstract object, except if it is a nominalization. We allow the first type and nominalizations (Al-Mazdar) from the second type as arguments of a discourse connective.

\textbf{4.3. Types of Relations}

We use the same 4 main relation classes as the PDTB does for English: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. However, we reduce the number of subclass relations we use to 18. We especially do not currently annotate further fine-grained distinctions, such as whether a conditional is counterfactual, as done in the PDTB. Future versions of the LADTB might include finer-grained distinctions. Figure 1 shows the hierarchy of discourse relations for Arabic.

We introduce two new relations at subclass level as we found them necessary in our pilot annotation for Arabic. These are EXPANSION.Background and COMPARISON.Similarity.

\textbf{EXPANSION.Background}\textsuperscript{2} applies when the argument that is syntactically associated with the connective describes prior eventualities which are background information of the other argument. This was frequent in news reports (see Ex. 7).

\textsuperscript{2}As noted in Section 4.2, we reduce the number of subclass relations we use to 18.

\textsuperscript{3}These are \textit{explication}, \textit{reformulation}, \textit{exception}, \textit{alternative}, \textit{disjunctive} and \textit{conjunctive}.

\textsuperscript{4}Non-pragmatic reasons are contextually based (e.g. \textit{He was tired}. \textit{[In contrast]}). Pragmatic reasons are based on the choice of the speaker (e.g. \textit{He was tired}. \textit{[Although]}).
The comparison similarity applies when the connective indicates that the two arguments express similar abstract objects. It is therefore a complement to the contrast relation.

**Comparison. Similarity** appears when the connective indicates that the two arguments express similar abstract objects. It is therefore a complement to the contrast relation.

\[\text{إن العسكريان قتلا برصاصة في الرأس ثم سوهت} \quad (8) \quad \text{جثموه [كما}][\text{يحصل غالبًا في عمليات}} \quad \text{الخطف على يد القوات المسلحة}}\]

[an al\text{s}k\text{ryyn qtl\text{a} br\text{s}ash fy al\text{r}as tm \text{š}wht \text{ált\text{h}}nâ\text{a-mı́}]
\quad \text{Arg1} \quad [\text{kmâ}]\text{PC } \text{ýyl\text{a} g\text{alb\text{ä} fy \text{mly}t\text{ä} \text{altf} \text{lä yd \text{alq}w\text{ä}t \text{alm}lšh}t}] \quad \text{Arg2}

[The military were killed by a bullet in the head and their bodies disfigured[\text{Arg1}[\text{as]PC [often happens in abductions by armed groups}][\text{Arg2}}

5. Agreement Studies

5.1. The Corpus

We base our study on the Penn Arabic Treebank (Part 1 v. 2.0) as part of the largest syntactically annotated corpus for Arabic. It consists of 734 files containing roughly 166K words of written Modern Standard Arabic newswire from the Agence France Press.

5.2. Arabic Discourse Annotation Tool (ADA) and Annotation Process

We developed a dedicated discourse annotation tool to deal with requirements specific to Arabic discourse annotation such as the annotation of clitics and right-to-left script order. The tool allows selection of Arabic or English annotation (see Figure 2).

It also highlights all potential discourse connectives from our connectives list (see Section 3.), including potential clitics. This is also shown in Figure 2. The annotator reads the text to get an overall understanding, and then makes a series of decisions for each potential connective in context.

1. If the potential connective is in this particular context not a connective (for example, because it does not relate two abstract entities), highlighting is removed and all annotation ceases.

2. If the potential discourse connective is indeed used as a connective, annotators will mark the text segments that express the two arguments of
the connective as well the discourse relation it conveys from a drop-down list of relations. Similar to the English annotation, annotators are allowed to use more than one relation, if a connective is deemed to express two relations at the same time. The screenshot in Figure 3 shows an example annotation.

3. Annotators are allowed to add comments into a comment box.

5.3. Annotation Methodology

Annotation was conducted by two independent native speakers of Arabic who were not involved in tool or scheme development. Agreement is measured on two tasks. The first task TASK I measures whether annotators agree on the binary decision on whether an item constitutes a discourse connective in context (Step 1 in the annotation procedure described above). Due to clitics and Arabic’s complex morphology this task is potentially harder than in English. Agreement is measured by the kappa statistic (Siegel and Castellan, 1956). The second task TASK 2 measures whether annotators agree on which discourse relation an identified connective expresses. As annotators can use sets of relations for a connective, we use kappa as well a variant of kappa called alpha, which allows us to measure partial agreement on sets while keeping kappa’s advantage of factoring out random agreement (Artstein and Poesio, 2008). A pilot annotation on 121 texts was used to train the annotators and to clarify the annotation guidelines, if necessary. The actual annotation after training has been conducted on 537 texts.

6. Results

Agreement on TASK I is highly reliable (N= 23331, percentage agreement of 0.95, kappa of 0.88). Full agreement is shown in Table 1. Due to proliferation of ambiguous clitics, most potential connective tokens are actually not connectives so that only 5586 of a potential 23331 connectives are actually really discourse connectives.

Agreement on TASK II (relation assignment) is relatively low (N = 5586, percentage agreement of 0.66, kappa of 0.57, and alpha of 0.58). It turns out that one of the major sources of disagreement is due to a convention in Arabic newswire writing: each sentence (if not introduced by an alternative connective) is introduced by /w/and, mostly without a specific discourse relation conveyed. This caused a high level of confusion. We therefore report agreement on three different datasets (see Table 2): the set of all identified connectives, the set of identified connectives excluding /w/and and the set of identified connectives excluding /w/and at the beginning of a paragraph (BOP). We see that reliability for connectives excluding rhetorical use of /w/and is good.

Connectives are mostly unambiguous in English (Pitler et al., 2008). However, for Arabic we encountered higher levels of ambiguity. The most ambiguous connectives at class level are in order ف/then, ك/ymaâl/while, ك/kmâlas and ك/ة في حي/ن in the same time. The most ambiguous connectives at sub-class level are again ك/w/and, then in order ك/ة b/due to/because, ك/ff/then and ك/ll/for/due to. This also highlights the value of this study of connectives in context as we discovered several context-dependent usages of discourse connectives that were not discussed in previous work on discourse connectives for Arabic (Al-Warraki and Hasanayn, 1994; Ryding, 2005; Alansari, 1985; Alfarabi, 1990). For example, ك/w/and is normally just associated with Conjunction in the literature but we discovered various other relations it expresses.

Table 1: Inter-annotator reliability for discourse connective identification (TASK I)

| All potential connectives (23331) | Observed agreement | 0.95 |
|----------------------------------|--------------------|------|
| Kappa                            | 0.88               |
| Potential connectives w/o ك/w/and (15602) | Observed agreement | 0.95 |
| Kappa                            | 0.82               |
| Potential connectives w/o ك/w/and at BOP (21200) | Observed agreement | 0.95 |
| Kappa                            | 0.84               |

Table 2: Inter-annotator reliability for discourse relations (TASK II)

| Connectives excluding ك/w/and (1886) | Observed agreement | 0.80 |
|--------------------------------------|--------------------|------|
| Kappa                                | 0.77               |
| Alpha                                | 0.80               |
| Connectives excluding ك/w/and at BOP (3500) | Observed agreement | 0.74 |
| Kappa                                | 0.69               |
| Alpha                                | 0.71               |

7. Gold Standard

We are now in the process of reconciling the annotations into a gold standard. First, we realized that
of discourse connectives as well as to Bonnie Webber and other members of the PDTB team for useful discussions.

9. References

Amal Al-Saif, Katja Markert, and Hussein Abdul-Raof. 2009. Corpus-Based Study: Extensive Collection of Discourse Connectives For Arabic. In *Proceedings of The Saudi International Conference 2009 (SIC09)*, Surrey, UK.

W. Al-Sanie, 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 (IAWTIC05)*.

N.N. Al-Warraki and A.T. Hasanyn. 1994. *The connectors in modern standard Arabic*. American University in Cairo Press.

I.H. Alansari. 1985. *Mogny Allabib*. Dar Alfekur, Beirut.

H. Alfarabi. 1990. *Ketab AlHorof*. Dar AlMashreg, Lebanon.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics (survey article). *Computational Linguistics*, 34(4):555–596.

K. Dukes and N. Habas. 2010. Morphological annotation of quranic arabic. In *International Conference on Language Resources and Evaluation (LREC 2010)*.

J. Hajic, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117. Citeseer.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman London.

J.R. Hobbs. 1985. *On the coherence and structure of discourse*. Center for the Study of Language and Information, Stanford, Calif.

E.H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1-2):341–385.

A. Knott and T. Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135–175.

M. Maamouri 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 (COLING)*, Geneva.

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

---

N. Al-Warraki and A. Touir. 2005. *Towards a rhetorical parsing of Arabic text*. In *The International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC05)*.

Amal Al-Saif is supported by a PhD scholarship from the Imam Muhammad Ibn Saud University, Saudi Arabia. We thank the British Academy for additional funding for the annotation study via Grant SG51944. We would also like to acknowledge the contributions of the annotators Latifa Alsulaiti and Abdul-baqi Sharif in the actual annotation and Basmah Al-Soli, Boshra Al-shyban and Maryam Al-Gawi in the pilot annotation. A special thank you goes to Dr. Hussein Abdul-Raof for linguistic advice on the collection

---

Note that other instances of *lwland* at BOP are not treated this way.
| Metric                              | Value       |
|------------------------------------|-------------|
| Number of files                    | 537         |
| Connective types                   | 68          |
| Discourse relation types           | 18 plus 22 combinations |
| Potential connective tokens        | 23147       |
| Real discourse connectives         | 5651        |
| Most frequent connective           | /w/and (3826) |
| Least frequent connectives         |             |
|                                   |             |
|                                   |             |
| Most frequent relations            |             |
|                                   |             |
| Rare relations                     |             |

Table 3: Statistics of the LADTB gold standard

D. Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.

E. Pitler, 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, August.

R. Prasad, N. Dinesh, A. Lee, E. Mititsakaki, 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)*.

R. Prasad, S. Husain, D.M. Sharma, and A. Joshi. 2008b. Towards an Annotated Corpus of Discourse Relations in Hindi. In *The Third International Joint Conference on Natural Language Processing*, pages 7–12.

K.C. Ryding. 2005. *A reference grammar of modern standard Arabic*. Cambridge Univ Press.

Amal Seif, Hassan Mathkour, and Ameur Touir. 2005. An rst computational tool for the arabic language. In *iiWAS*, pages 527–534.

S. Siegel and N.J. Castellan. 1956. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill New York.

B. Webber, 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*, page 48.

Nianwen Xue. 2005. Annotating discourse connectives in the chinese treebank. In *CorpusAnno ’05: Proceedings of the Workshop on Frontiers in Corpus Annotations II*, pages 84–91, Morristown, NJ, USA. Association for Computational Linguistics.

D. Zeyrek and B. Webber. 2008. A discourse resource for turkish: Annotating discourse connectives in the metu corpus. *Proceedings of IJCNLP-2008. Hyderabad, India.*
| Connective | English equivalent | Syntactic category | Type      | Buckwalter | ATB tag          | Freq |
|-----------|--------------------|--------------------|-----------|------------|------------------|------|
| و          | and                | Coordinating conj  | Simple    | wa         | CONJ             | 3826 |
| ل          | for/of/in order to | Preposition        | Clitic    | li         | PREP             | 261  |
| لكن        | but                | Coordinating conj  | Simple/clitic | lkn       | CONJ             | 208  |
| بعد        | after              | Adverbial          | Simple/clitic | bEd       | PREP             | 167  |
| في         | then               | Coordinating conj  | Clitic    | fa         | CONJ             | 91   |
| لأن         | because            | Subordinating conj | Simple/clitic | lAn       | CONJ             | 82   |
| قبل        | before             | Adverbial          | Simple    | qbl        | PREP             | 79   |
| أثر        | after              | Subordinating conj | Simple    | Avr        | PREP             | 63   |
| ب           | due to/because    | Preposition        | Clitic    | bi         | PREP             | 63   |
| كما         | as/and/similarly   | Coordinating conj  | Simple    | mA         | CONJ             | 60   |
| منذ        | since              | Adverbial          | Simple    | mn*        | PREP             | 59   |
| بسبب       | because of         | Prepositional phrase | Simple/Paired | bsbb | PP_PREP _NOUN/PREP | 45  |
| عندما      | when/due           | Adverbial          | Simple    | EnmA       | CONJ             | 44   |
| إلا أن      | however            | Subordinating conj | Simple    | AAn        | EXCEPT-PART _FUNC-WORD | 42  |
| في حال     | in case/if         | Prepositional phrase | Simple    | fy HAI     | PREP_NOUN        | 36   |
| فيما       | while/as           | Subordinating conj | Simple    | fymA       | PREP _REL-_PRON  | 36   |
| إذا         | if                 | Subordinating conj | Simple/Paired | A*A       | CONJ             | 31   |
| ثم          | then               | Coordinating conj  | Simple    | vm         | ADV              | 30   |
| أو          | or                 | Coordinating conj  | Simple    | Aw         | CONJ             | 29   |
| رغم        | although            | Subordinating conj | Simple/Paired | rgm       | PREP             | 29   |
| في حين    | while/in the same time | Prepositional phrase | Simple/Clitic | fy Hyn | PREP_NOUN        | 26   |
| أما         | while/as           | Subordinating conj | Simple    | AmA        | PREP             | 25   |
| أذ          | because            | Coordinating conj  | Simple    | A*         | CONJ             | 21   |
| مما         | therefore           | Subordinating conj | Simple    | mA         | PP_PREP _REL-_PRON | 21  |
| خصوصا      | especially          | Adverbial          | Simple    | xSwSA      | ADV_SSUFF        | 18   |
| بينما       | while/as           | Subordinating conj | Simple    | bynmA      | CONJ             | 17   |

Table 4: The most frequent connectives in LADTB