# Modelling Discourse Relations for Arabic

**Amal Alsaif**
University of Leeds
Leeds, UK
LS2 9JT
amalalsaif@yahoo.co.uk

**Katja Markert**
University of Leeds
Leeds, UK
LS2 9JT
markert@comp.leeds.ac.uk

## Abstract

We present the first algorithms to automatically identify explicit discourse connectives and the relations they signal for Arabic text. First we show that, for Arabic news, most adjacent sentences are connected via explicit connectives in contrast to English, making the treatment of explicit discourse connectives for Arabic highly important. We also show that explicit Arabic discourse connectives are far more ambiguous than English ones, making their treatment challenging. In the second part of the paper, we present supervised algorithms to address automatic discourse connective identification and discourse relation recognition. Our connective identifier based on gold standard syntactic features achieves almost human performance. In addition, an identifier based solely on simple lexical and automatically derived morphological and POS features performs with high reliability, essential for languages that do not have high-quality parsers yet. 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.

## 1 Introduction

The automatic detection of discourse relations, such as causal, contrast or temporal relations, is useful for many applications such as automatic summarization (Marcu, 2000), question answering (Girju, 2003), sentiment analysis (Somasundaran et al., 2008) and readability assessment (Pitler and Nenkova, 2008). This task has recently seen renewed interest due to the growing availability of large-scale corpora annotated for discourse relations, such as the Penn Discourse Treebank (Prasad et al., 2008a).

In the Penn Discourse Treebank (PDTB), local discourse relations (also called *senses*) such as CAUSAL or CONTRAST are annotated. They hold between two text segments (so-called *arguments*) that express abstract entities such as events, facts and propositions. Annotated discourse relations can be signalled explicitly by so-called *discourse connectives* (Marcu, 2000; Webber et al., 1999; Prasad et al., 2008a) or hold implicitly between adjacent sentences in the same paragraph, i.e. are not signalled by a specific surface string. In Ex. 1, the connective *while* indicates an explicit CONTRAST between the attitudes of John and Richard. In Ex. 2, the connective *while* indicates an explicit TEMPORAL relation. In Ex. 3, an implicit CAUSAL relation between the first and second sentence holds. We indicate discourse connectives and the two arguments they relate via annotated square brackets.

(1) [John liked adventure,]$_{Arg2}$ [ while]$_{DC}$[Richard was cautious]$_{Arg2}$

(2) [The children were crying loudly]$_{Arg1}$[while]$_{DC}$,[their mother was cooking]$_{Arg2}$

(3) [I cannot eat any dessert.]$_{Arg1}$ [I have eaten far too much already.]$_{Arg2}$

Although similar corpora for other languages are being developed such as for Hindi (Prasad et al., 2008b), Turkish (Zeyrek and Webber, 2008), Chinese (Xue, 2005) and, by ourselves, for Arabic (Al-

*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 736–747,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

Saif and Markert, 2010), efforts in the automated recognition of discourse connectives, arguments and relations have so far almost exclusively centered on English.

In contrast we present the first models for discourse relations for Arabic, focusing on explicit connectives. This focus is partially justified by the fact that this first study for a new language should center on the superficially more straightforward case and that no annotations for implicit relations are yet available for Arabic. More importantly, however, we make two essential claims (Section 4). Firstly, Arabic discourse connectives are more ambiguous than their English counterparts, i.e cases such as *while* which can signal different relations dependent on context (see Example 1 and 2) are far more frequent. This makes their treatment more *challenging*. Secondly, discourse relations between adjacent sentences in Arabic tend to be expressed via an explicit connective, at least for the news genre, i.e. cases such as Example 3 are rarer. This makes the treatment of explicit connectives *essential.*

We tackle two tasks for explicit Arabic connectives in this paper, which are further discussed in Section 2. Discourse connective recognition needs to distinguish between discourse usage of potential connectives and non-discourse usage (such as the use of *while* as a noun). We show in Section 5 that we can distinguish discourse- and non-discourse usage for potential connectives in Arabic with very high reliability, even without parsed data, a fact that is important for languages with fewer high quality NLP tools available. We then present an algorithm for relation identification in Section 6 that shows small but significant gains over assigning the most frequent relation for each connective. We discuss future work and conclude in Section 7.

## 2 The Tasks

The handling of explicit connectives can be split into three tasks (Pitler and Nenkova, 2009). The first task of *discourse connective recognition* distinguishes between the *discourse usage and non-discourse usage* of potential connectives. Whereas some potential connectives such as the Arabic connective لكن */lkn/but* almost always have discourse usage, this is not true for all potential connectives.[1] Thus, the discourse usage of Arabic رغبة */rġbh/desire* needs to be distinguished from its use as a noun. Conjunctions such as و */w/and,* او */āw/or* can have discourse usage or just conjoin two non-abstract entities as in عمر و ساره */ʿmr w sārh/Omar and Sarah.*

The second task is *discourse connective interpretation* where a discourse connective in context is assigned a discourse relation. Again, some connectives are largely unambiguous in this respect. For example, لكن */lkn/but* signals almost always a CONTRAST relation. However, there are connectives where this is not the case, such as منذ */mnḏ/since* which has a CAUSAL and a TEMPORAL sense.

The third task is argument identification which identifies the arguments' position and extent. In this paper we tackle Task 1 and Task 2 for Arabic in a supervised machine learning framework.

## 3 Related work

**Annotated Discourse Corpora and Linguistic Background.** Discourse relations 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). Different inventories have been used in English corpora annotated for discourse relations (Hobbs, 1985; Prasad et al., 2008a; Carlson et al., 2002) which also differ in other respects (such as whether they prescribe a tree structure for discourse annotation). However, the annotation level of existing Arabic corpora has not yet included the discourse layer, making our work the first to address this problem for Arabic on a larger scale.

**Automatic discourse parsing: explicit relations.** There is no work on discourse connective recognition, interpretation and argument assignment for Arabic, so that we break entirely new ground here. However, the two tasks we explore (discourse connective recognition and discourse connective disambiguation) have been tackled for English.[2] (Pitler

[1] Arabic examples contain in order: the Arabic right-to-left script, the transliteration (standards ISO/R 233 and DIN 31635) and the English translation (if possible).

[2] There is also substantial work on argument identification (Wellner and Pustejovski, 2007; Elwell and Baldridge, 2008)

and Nenkova, 2009) use gold standard syntactic features as well as the connective surface string in a supervised model for discourse connective recognition. They achieve very high results with this approach. 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.

With regard to discourse connective interpretation, (Miltsakaki et al., 2005) concentrate on disambiguating the three connectives *since, while, when* only, 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. A more comprehensive study on all discourse connectives in the PDTB (Pitler et al., 2008; Pitler and Nenkova, 2009) reveals that most connectives are not ambiguous in English. Using syntactic features of the connective, they achieve only a very small improvement over a "most frequent relation per connective baseline" for which significance tests are not given. We will show that for Arabic, discourse connectives are more highly ambiguous with regard to the relations they convey. We will present a supervised learning model that uses a wider feature set and that achieves small but significant improvements over the most frequent relation per connective baseline.

**Automatic discourse parsing: implicit relations.** Implicit relations have excited substantial interest for English. This includes work in the framework of RST (Soricut and Marcu, 2003; duVerle and Prendinger, 2009; Marcu and Echihabi, 2002), SDRT (Baldridge and Lascarides, 2005), GraphBank (Wellner et al., 2006), the PDTB (Blair-Goldensohn et al., 2007; Pitler et al., 2009; Lin et al., 2009; Wang et al., 2010; Zhou et al., 2010; Louis and Nenkova, 2010) or framework-independent (Sporleder and Lascarides, 2008).[3] The task is challenging as implicits behave substantially differently from explicits (Sporleder and Lascarides, 2008) and often need world knowledge (Lin et al., 2009). However, features/approaches that have shown improvement over a baseline are word pairs (Sporleder and Lascarides, 2008), production rules and syntactic trees (Wang et al., 2010; Lin et al., 2009) as well as language modelling (Zhou et al., 2010). As we only deal with explicit connectives this work is not directly comparable to ours, although we do explore some of the suggested features for improving explicit connective disambiguation.

## 4 An Arabic Discourse Corpus

We annotate news articles from the Arabic Penn Treebank (Part 1 v2.0) (Maamouri and Bies, 2004) for explicitly marked discourse relations. This is the first discourse-annotated corpus for Arabic, whose initial development stages we have described in (Al-Saif and Markert, 2010). We summarize this previous work and extend it by including agreement studies for arguments in Sections 4.1 and 4.2. In Sections 4.3, 4.4 and 4.5. we then present a corpus study on the corpus which shows our major claim as to the importance and high levels of ambiguity of Arabic discourse connectives.

### 4.1 Annotation Principles

We overall follow the annotation principles in the Penn Discourse Treebank for explicit connectives (for example, arguments can occur at any distance from the connectives). The relation set we use is a more coarse-grained version of the PDTB relations with two relations added — BACKGROUND and SIMILARITY — that we found in our Arabic news texts. The final, hierarchically organized, relation set of 17 discourse relations is shown in Fig 1.

Further adaptations necessary for Arabic are the inclusion of clitics as connectives such as ل */l/for,* ب */b/by,with* and ف */f/then* . In addition, differently to English, prepositions were included as connectives as these are frequently used to express discourse relations in Arabic. In these cases, normally argument 2 is the so-called *Al-Masdar*.[4] Typical examples are وصول */wṣwl/arrival* from the verb وصل */wṣ/to arrive* and محاولة */mḥāwlh/attempt* from the verb حاول

but we do not discuss this work in depth here.

[3] Some work does not make the distinction between implicit and explicit and/or treats them in a joint framework (Soricut and Marcu, 2003; Wellner et al., 2006; Wang et al., 2010).

[4] The medieval Arabic grammar schools, the Basra and Kufa, debated whether the noun (almasdar) or the verb is the most basic element of language (Ryding, 2005).

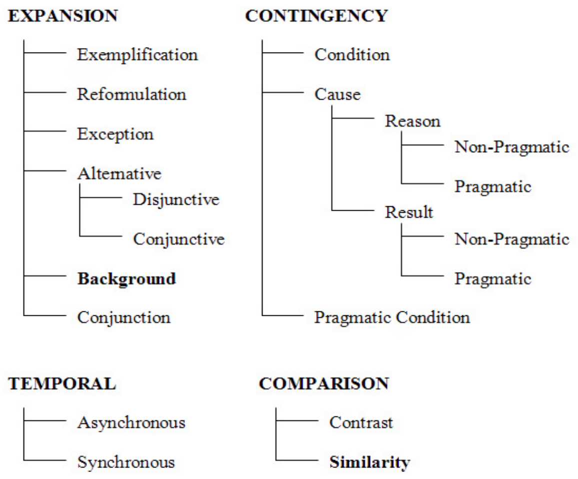

Figure 1: Discourse relations for Arabic

*/ḥāwl/to try*. Al-Masdar is formed using morphological patterns well-known in the Arabic grammatical tradition: major Arabic grammars list around 60 patterns although some other references also claim that the patterns are many more as well as more unpredictable (Abdl al latif et al., 1997; Wright, 2008; Ryding, 2005). Al-Masdar forms do not fit into one grammatical or morphological category in English: they might correspond to a gerund, a nominalization or a noun which is not a nominalization. Some examples are listed in Table 1.

Table 1: A list of Al-MaSdar patterns, examples and their English correspondence

| Root | Pattern | MaSdar | Translation |
|---|---|---|---|
| سبح */sbḥ* | فعالة */fʿalh* | سباحة */sbaḥh* | swimming |
| نفذ */nfḏ* | تفعيل */tfʿyl* | تنفيذ */tnfyḏ* | execution |
| دفع */dfʿ* | فعال */fʿāl* | دفاع */dfāʿ* | defence |
| زرع */zrʿ* | فعالة */fʿālh* | زراعة */zrāʿh* | agriculture |
| حرب */ḥrb* | فعل */fʿl* | حرب */ḥrb* | war |

An example of Al-MaSdar as argument of a discourse relation is Ex. 4, where تبليغ*/tblyġ/informing* is the *Al-MaSdar* form of بلغ*/blġ/inform*.

(4) [ذهبنا الى مركز الشرطة]$_{Arg1}$ [ل]$_{DC}$ [لّتبليغ عن فقدان وثائق الشركة الرسمية]$_{Arg2}$

[*ḏhbnā ’lā mrkz al-šrṭt.*]$_{Arg1}$[*l*]$_{DC}$[*ltblyġ ʿn fqdān wṯāʾiq alšrkh alrsmyh*]$_{Arg2}$

[We went to the police station]$_{Arg1}$ [for]$_{DC}$ [informing about the loss of the company's official documents.]$_{Arg2}$

### 4.2 Agreement Studies

The occurrences of a precompiled list of 107 potential discourse connectives were annotated independently by 2 native Arabic speakers on 537 news texts. Agreement was measured for the distinction of discourse vs. non-discourse usage, relation assignment and argument assignment.

Agreement for the classification tasks of discourse connective recognition and relation assignment was measured using kappa (Siegel and Castellan, 1956). Argument agreement was measured by `agr`, a directional measure (Wiebe et al., 2005). It measures the word overlap between the text spans of two judges (ann1 and ann2). `agr(ann1||ann2)` measures the proportion of words ann1 annotated that were also annotated by ann2.

$$agr(ann1||ann2) = \frac{|ann1\ matching\ ann2|}{|ann1|}$$

Discourse connective recognition proved to be highly reliable with percentage agreement of 0.95 and a kappa of 0.88 on the 23,331 occurrences of the 107 potential discourse connectives. 5586 of the potential connectives were agreed on by both annotators to have discourse usage and agreement for relations and argument assignment was measured on these. As shown in Table 2, kappa on all 17 relations was low with 0.57 — it turned out that this was due to the frequent, almost rhetorical use of the connective و */w/and* at the beginning of paragraphs, which is a genre convention for Arabic news that normally does not convey a specific discourse relation. Disregarding such occurrences of و */w/and*, kappa rises to good agreement: 0.69 for fine-grained relations and 0.75 when measuring agreement between the 4 major relations EXPANSION, CONTINGENCY, COMPARISON and TEMPORAL.

Argument agreement on the 5586 agreed connectives is shown in Table 3. We report high word overlap via `agr` (over 90%) for Arg2, which is the argument syntactically attached to the connective, and lesser but still substantial agreement for Arg1.

Table 2: Inter-annotator reliability for discourse relation assignment

| **All connectives (5586)** | |
|---|---|
| Observed agreement | 0.66 |
| Kappa | 0.57 |
| Class level | |
| Observed agreement | 0.8 |
| Kappa | 0.67 |
| **Connectives excluding و */w/and* at BOP (3500)** | |
| Observed agreement | 0.74 |
| Kappa | 0.69 |
| Class level | |
| Observed agreement | 0.71 |
| Kappa | 0.75 |

| **Agreed disc. conn** | 5586 | |
|---|---|---|
| | Arg1 | Arg2 |
| **a) exact match** | | |
| exact match =1 | 2361 (42%) | 3803 (68%) |
| exact match =0 | 699 (13%) | 18 (0.3%) |
| partial match | 2526 (45%) | 1765 (32%) |
| **b) agr metric** | | |
| agr(ann1\|\|ann2) | 78% | 93% |
| agr(ann2\|\|ann1) | 74% | 93% |
| **Avr (agr)** | 76% | 93% |

Table 3: Inter-annotator reliability for arguments Arg1 and Arg2, using two different measurements (a) exact match (b) agr

### 4.3 Gold standard

We produced a unified gold standard. First, we automatically corrected easily made annotator mistakes. With regard to argument extent, we automatically corrected mistakes such as the erroneous inclusion of punctuation marks at the end of clauses/sentences or not including all obligatory complements in a verb phrase argument. The latter relied on the syntactic annotation in the ATB. Second, with regard to discourse relation assignment, we automatically assigned EXPANSION.CONJUNCTION to all disagreed instances of و */w/and* at BOP.[5] A further disambiguation study is necessary for و */w/and* at BOP, which is beyond the scope of this paper.

Finally, an adjudicator not initially involved in annotation reconciled the remaining disagreements at all levels and included annotations for 5 new potential discourse connective types not in our initial connective list but commented on by the annotators during annotation. 3 news files were removed from the corpus — they contained no actual news reports but just a list of headlines.

The final discourse treebank we use has 6328 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 من*/bālrġm mn,* رغم ان */rġm ān* as modified forms of رغم */rġm/although.*

Most discourse connectives were only annotated with a single relation but 5% were annotated with two or more relations (as also allowed in the PDTB). These statistics are summarised in Table 4.

| | |
|---|---|
| Files | 534 |
| Total tagged tokens | 126,046 (125KB) |
| Sentences | 3607 |
| Paragraphs | 3312 |
| Discourse connectives (tokens) | 6328 |
| Distinct connective (types)<br>including modifed form connectives | 68<br>80 |
| Clitic discourse connectives (tokens) | 4779 (76%) |
| Non-clitic discourse connectives (tokens) | 1549 (24%) |
| Relations types (17 single, 38 combined) | 55 |
| Single relations (tokens)<br>Combined relations (tokens) | 6039 (95%)<br>289 (5%) |

Table 4: Statistics of the final gold standard corpus

### 4.4 Importance of explicitly signalled relations

We compared the number of relations between 2 adjacent sentences that were explicitly signalled in English vs. the ones that were explicitly signalled in Arabic, using the PDTB and our corpus (both containing texts of the news genre). Out of a total 44,470 adjacent sentence pairs in the PDTB, 5355

[5] Other instances of و */w/and* are not treated this way.

(12%) were linked by an explicit connective.[6] In contrast, out of the 3073 adjacent sentence pairs in our corpus, 2140 (70%) were linked by an explicit connective, 948 (30%) were linked via non-*wa* connectives. Thus, for our corpus, modeling of explicit connectives is primary: intrasentential relations tend to be marked by connectives anyway in both English and Arabic, and our corpus shows that this is true for most local intersentential relations as well.

### 4.5 Ambiguity for Arabic discourse connectives

We investigate the ambiguity of Arabic connectives with regard to their sense at class level (4 relations) as well as the more fine-grained level (all 17 relations). We restrict our investigation to the connective occurrences that were annotated with a single relation (6039 tokens) and also exclude و */w/and* at the beginning of paragraph, leaving 3813 tokens.[7] Of 80 connective types, 52 were unambiguous at the class level and 47 at the fine-grained level. However, many of the most frequent connectives are highly ambiguous. If we just assign the most frequent reading to each of the 3813 connectives, 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 error margin. 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.[8]

## 5 Discourse Connective Recognition

We distinguished discourse vs. non-discourse usage for all potential connectives in the 534 gold standard files. As headers and footers in the news files never contained true discourse connectives, we disregarded these, leaving 20,312 potential discourse connectives of which 6328 are actual connectives.

### 5.1 Features

Apart from the surface string of the potential connective *Conn*, we use the following features. Features are either extracted from raw files tokenized by white space only (M2) or from raw files tokenized by white space and tagged by the Stanford tagger[9] (Models M3, M4) or from the Arabic Treebank (ATB) gold standard part-of-speech and parse annotation (models M5-M9). The syntactic features (Syn) are inspired by (Pitler and Nenkova, 2009). Lexical/POS patterns of surrounding words, clitic features and Al-Masdar are novel.

**Surface Features (SConn).** These include the position of the potential connective (sentence-initial, medial or final). The *type* of the potential connective is `Simple` when the potential connective is a single token not attached to other tokens, `PotClitic` when it is attached. Potential connectives containing more than one token have `MoreThanToken` type.

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. Models that use ATB annotation also distinguish between potential connectives that correspond to a phrase in the ATB (`MorethanToken_Phrase`) and the ones that do not (`MorethanToken_NonPhrase`).

**Lexical features of surrounding words (Lex).** We encode the surface strings of the three words before and 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 */ān/* it most likely has a discourse function (see Ex. 5).

(5) ان الاطفال يمكن [ان يصابوا بالارهاق]$_{Arg1}$ [ و ]$_{DC}$
[ان يشعروا بالنعاس]$_{Arg2}$ خلال الدراسة اذا لم يناموا
جيدا

[ *ān ālāṭfāl ymkn [ān yṣābwā bālārhāq*]$_{Arg1}$[*w*]$_{DC}$[*ān yšʕrwā bālnʕās*]$_{Arg2}$ *ḫlāl āldrāsh āḏā lm ynāmwā ǧydā*

[Children might be tired]$_{Arg1}$ [and]$_{DC}$ [feel sleepy]$_{Arg2}$ during school time if they did not sleep well

[6] Connections between subclauses or phrases in different, adjacent sentences were included in the count.

[7] We automatically assigned CONJUNCTION to many occurrences of و */w/and* at BOP (Section 4.3) so that it is not sensible to include these occurrences in a study of human-assigned ambiguity.

[8] The second level in the PDTB with its 16 relations corresponds approximately to our fine-grained inventory. This comparison can only be appropriate due to slight differences in the lower-grained relation inventory.

[9] `http://nlp.stanford.edu/software/tagger.shtml`

**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 ḥāl/in case* with its tags (fy PREP)(Hal NOUN)) will receive the pos PREP#NOUN. If a potential connective does not receive a separate POS tag in the ATB/tagger, the value "NONE" is assigned. This allows to distinguish clitics from letters at the start of a word. We also record the POS of the three words before/after the connective (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, 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 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 */ālmdrsh kbyrh w ğmylh/ the school is very large and beautiful*) illustrates this. Unlike English, parents in Arabic often are noun phrases as nominalisations are frequent arguments of prepositional connectives. We also encode the Left sibling category and right sibling category of the connective. For discourse connectives, the right sibling is normally S, SBAR, VP or an NP (if the connective is a preposition).

**Al-Masdar feature.** Potential connectives followed by Al-Masdar are more likely to have discourse usage (see Section 4.1). Especially prepositions with discourse usage are normally attached to Al-masdar such as in لمحادثة */lmḥādṯh/for contacting* or باجراء */bāğrāʔ/by processing*. Al-Masdar information is not included in the ATB so 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 uses a stemmer for Arabic and then determines whether the stem is al-Masdar by a combination of surface-based rules to check whether the stem corresponds to one of the known Al-Masdar patterns.

### 5.2 Results and Discussion

We used the implementation JRip of the rule-based classifier Ripper in the machine learning tool WEKA with its default settings. We used 10-fold cross-validation throughout. Significance tests are reported using the McNemar test at the significance level of 1%. A most frequent category baseline would assign all potential connectives as *not connective*, achieving an accuracy of 68.9% as only 6328 of our potential 20,312 connectives actually have discourse usage. We built several models using different features. The results are shown in Table 5.

A simple model M1 that only uses the connective string improves significantly over the baseline with 75.7% accuracy but a kappa of only 0.48, showing that this is not a reliable strategy. 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-space tokenized raw files in addition to the connective string (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.

With ATB gold standard tokenisation, tagging and parsing, our models (not surprisingly) improve further showing the same pattern of (M1 < M5 < M6 < M7) with all differences being significant. The final best model achieves highly reliable results (accuracy 92.4%, kappa 0.82). We also conclude that syntactic features are more useful than lexical patterns as model M8 (syntax with no lexical patterns) achieves equally good results as M6. Our models also manage to generalise well over individual connectives. If we leave out the connective string (M9), we still achieve a highly reliable result.

## 6 Discourse Relation Recognition

When disambiguating the relation that discourse connectives signal, we assume that the arguments of

|  | Features | Acurr | K |
|---|---|---|---|
|  | Baseline (not conn) | 68.9 | 0 |
| M1 | Conn only | 75.7 | 0.48 |
| Tokenization by white space + auto tagger |  |  |  |
| M2 | Conn+ SConn+Lex | 85.6 | 0.62 |
| M3 | Conn+ SConn+Lex+POS | 87.6 | 0.69 |
| M4 | Conn+SConn+Lex+POS+Masdar | 88.5 | 0.70 |
| ATB-based features |  |  |  |
| M5 | Conn+SConn+Lex | 86.2 | 0.65 |
| M6 | Conn+SConn+Lex+Syn/POS | 91.2 | 0.79 |
| M7 | Conn+SConn+Lex+Syn/POS+Masdar | 92.4 | 0.82 |
| M8 | Conn+SConn+Syn | 91.2 | 0.79 |
| M9 | SConn+Lex+Syn+Masdar | 91.2 | 0.79 |

Table 5: Performance of different models for identifying discourse connectives.

the connective are known. This is well-established for PDTB relation recognition (Wang et al., 2010; Lin et al., 2009; Miltsakaki et al., 2005). Our models predict single relations on two datasets: (i) all instances of connectives signalling single relations (Set `All`, 6039 instances) (2) all instances apart from the connective و */w/and* at beginning of paragraph as they are affected by the auto-correction process (Set `no-wa-atBOP`, 3813 instances). We use 10-fold cross-validation and JRip as well as a McNemar test at the 5% level for significance tests.

## 6.1 Features

Whereas some of the features we use have been used for English implicit relation recognition (Lin et al., 2009; Wang et al., 2010; Pitler et al., 2009) , they are new for Arabic and not widely used for explicit connectives. All features are extracted from the ATB gold standard parses.

**Connective features.** This includes the connective string *Conn*. In addition, we also use the surface connective features and POS of connective described in Section 5. We also use the syntactic path to the connective as a novel feature.

**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. For example, when the first word of Arg2 is قد */qd/might/may* or كان */kān/had*, the relation is likely to be EXPANSION.BACKGROUND or EXPANSION.CONJUNCTION. We also measure word overlap between the arguments, hoping to catch relations such as COMPARISON.SIMILARITY.

**Masdar.** This feature states whether the first or second word in Arg 2 is an Al-Masdar. Many prepositional connectives followed by an Al-Masdar indicate a CONTINGENCY.CAUSE relation (see Ex. 4)

**Tense and Negation.** Each argument is assigned its tense as one of *perfect, imperfect, future or none*. We also indicate whether the tense of Arg1 or 2 are the same and whether a negation is part of Arg 1 or 2. Inspired by (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 are likely to have the same tense. In contrast, arg1_tense is more likely to be prior to arg2_tense for TEMPORAL.ASYNCHRONOUS and CAUSE relations.

**Length, Distance and Order Features.** We use the length of arguments (in words), word distance between a connective and its arguments (-1; for arguments in order Arg1_Conn_Arg2_Arg1), tree distance of connective and arguments (0 if connective and an argument are in the same tree) and a binary feature of whether Arg1 and Arg2 are in different sentences. A 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.ASYNCHRONOUS relations.

**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.

**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 et al., 2009) who use production rules to good effect for implicit relations in English.

### 6.2 Results

Table 6 shows the results for fine-grained (17 relations) classification. The baseline of assigning the most frequent relation EXPANSION.CONJUNCTION to every connective performs with an accuracy of 52.5% on Set `All` and 35% on set `no-wa-atBOP`. If we use a model that relies on the discourse connective alone (M1) we achieve results of 77.2%/74.3%, respectively. As noted in Section 4.5 this is substantially lower than what the same model can achieve for English. Including connective and argument features (apart from production rules) in M2, leads to a small but significant improvement.[10] Further incorporation of production rules does not improve the results (M3). In Table 7, we show the results at the class-level (4 relations). Here using additional features over the connective string does not lead to significant improvements.

| Ref | Features | Acc | K |
|---|---|---|---|
| *All connectives (6039)* | | | |
| | Baseline (CONJUNCTION) | 52.5 | 0 |
| M1 | Conn only (1) | 77.2 | 0.60 |
| M2 | Conn+Conn_f+ Arg_f (37) | 78.8 | 0.66 |
| M3 | Conn+Conn_f+ Arg_f+ Production rules (1237) | 78.3 | 0.65 |
| *excluding wa at BOP (3813)* | | | |
| | Baseline (CONJUNCTION) | 35 | 0 |
| M1 | Conn only (1) | 74.3 | 0.65 |
| M2 | Conn+Conn_f+ Arg_f (37) | 77 | 0.69 |
| M3 | Conn+Conn_f+ Arg_f+ Production rules (1237) | 76.7 | 0.69 |

Table 6: Performance of different models for identifying fine-grained discourse relations on two datasets.

| Ref | Features | Acc | K |
|---|---|---|---|
| *All connectives (6039)* | | | |
| | Baseline (EXPANSION) | 62.4 | 0 |
| M1 | Conn only (1) | 88.7 | 0.78 |
| M2 | Conn+Conn_f+ Arg_f (37) | 88.7 | 0.78 |
| *excluding wa at BOP (3813)* | | | |
| | Baseline (EXPANSION) | 41.8 | 0 |
| M1 | Conn only (1) | 82.7 | 0.74 |
| M2 | Conn+Conn_f+ Arg_f (37) | 83.5 | 0.75 |

Table 7: Performance of different models for identifying class-level discourse relations on two datasets.

### 6.3 Discussion and Error Analysis

We concentrate our discussion on fine-grained classification excluding wa at BOP.

Our improvements in M2 over the connective-only classifier (M1) 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 M2 over M1 for *u*nambiguous 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 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. Our data includes 47 unambiguous connective types, accounting for 574 of the 3813 tokens. 30 of these 47 types are so rare that we found mistakes in the connective-only classification, including الا /*ālā*/except (2), عقب /*ʿqb*(2),طالما /*ṭālmā*(2), برغم/*brġm*(1). For 14 of these 30 connectives, model M2 was able to use generalised rules to improve relation assignment.[11] These rules involve mainly connective surface and POS features. Thus, sentence-start adverbials consisting of more than one token such as بيد ان /*byd ān*(6) and غير ان /*ġyr ān*(6) were correctly classified as CONTRAST.

[10] Our corpus includes some texts on similar topics where some sentences are (almost) repeated in different texts. To investigate whether our improvements are due to this repetition, we also performed an experiment excluding all repeated instances of feature vectors from the corpus. The results are almost the same and, most importantly, M2 again improves significantly over M1.

[11] 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 does on truly ambiguous connectives (33 connective types accounting for 3239 tokens of 3813 overall tokens). We conducted a separate significance test on ambiguous connectives only and found that M2 improves over M1 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 (لو /*lw*, انما /*ānmā*, حال /*ḥāl*/) both M1 and M2 do perform badly on them. If connectives are frequent (10 or more occurrences) and have relatively low ambiguity (majority reading accounts for more than 70% of instances), the overall performance of M1 and M2 with regard to accuracy is also similar, often both using just the connective string. On the other hand, if connectives are frequent and have high ambiguity (i.e. no such clear majority reading), then M2 normally improves (often substantially) on M1. Examples of such connectives are كما /*kmā*, فيما /*fymā*and اثر /*āṯr*. 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 a successful recognition of all 9 instances of `Similarity` in the connective kmA (whose majority relation is `Expansion.Conjunction` in 40 out of 65 occurrences). The connective ف /*f/then* is 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, nontensed arguments are most often nominalisations leading to a reason reading, whereas a verb phrase as argument 2 and a sentence as argument 1 often is a result reading. However, it is worth reporting that in cases of very high ambiguity, M2 is still far from perfect such as for connectives f ف /*f*and اثر /*āṯr*.

Some improvements again come from generalised rules: there are some very high-coverage and high precision generalised rules that reduce dependency on the connective string. For example, clitic prepositions (such as ل /*l*/for) can without any further information be clearly classified as `Contingency.Cause.Reason.NonPragmatic` covering 494 occurrences with only 26 mistakes. These are cases where the following argument is normally Al-Masdar.

Our analysis leads us to the following strategy for follow-on work. First of all, a larger corpus is necessary to get more examples for low frequency connectives. Secondly, experiments with different classifiers are worthwhile to conduct to see how our improvements generalise. Third, the most mileage is in further improvements on frequent, ambiguous connectives such as ف /*f*, منذ /*mnḏ*and او /*āw*. This can be achieved with, on the one hand, training connective-specific classifiers on larger data sets but will, on the other hand, also need a wider feature base. From our corpus study, we think that lexico-semantic features such as word pairs and semantic classes of verbal/nominalised arguments are the most promising.

## 7 Conclusions and Future Work

We have presented the first study on the automatic detection and disambiguation of Arabic discourse connectives. A corpus study showed that these are highly frequent and more ambiguous than their English counterparts. Our automatic algorithms achieve very good results on discourse connective identification, using Arabic morphological properties to good effect. It is particular promising that we do not need parsed data to identify discourse usage of potential connectives reliably. Our algorithm for discourse connective interpretation beats the challenging baseline of assigning the most frequent relation per connective. In future, we will explore further features for connective disambiguation as well as connective-specific classification, combined with semi-supervised algorithms to alleviate data sparseness. We will also develop algorithms for argument identification.

**Acknowledgments**

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. Also thank you to the 3 anonymous reviewers for their comments.

## References

M. Abdl al latif, A. Umar, and M. Zahran. 1997. *Alnhw AlAsAsi*. Dar Alfker Al-Arabi, Cairo, Egypt.

A. AlSaif and K. Markert. 2010. The leeds arabic discourse treebank: Annotating discourse connectives for arabic. In *Language Resources and Evaluation Conference (LREC)*.

J. Baldridge and A. Lascarides. 2005. Probabilistic head-driven parsing for discourse structure. In *Proc. Of Conll 2005*.

S. Blair-Goldensohn, K McKeown, and O. Rambow. 2007. Building and refining rhetorical-semantic relation models. In *Proc. of HLT-NAACL 2007*.

L. Carlson, D. Marcu, and M. Okurewski. 2002. Rst discourse treebank. Linguistic Data Consortium.

D. duVerle and H. Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proc. of ACL 2009*.

R. Elwell and J. Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proc. of the International Conference on Semantic Computing*.

R. Girju. 2003. Automatic detection of causal relations for questions answering. In *Proc. of the ACL 2003 Workshop on Multilingual Summarisation and Question Answering*, pages 76–83.

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.

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.

Z. Lin, M. Kan, and H.T. Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proc. of EMNLP 2009*, pages 343–351.

A. Louis and A. Nenkova. 2010. Creating local coherence: An empirical assessment. In *Proc. of NAACL 2010*.

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.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proc. of ACL 2002*.

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

E. Miltsakaki, N. Dinesh, R. Prasad, A. Joshi, and B. Webber. 2005. Experiments on sense annotation and sense disambiguation of discourse connectives. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proc. of EMNLP 2008*, pages 186–195.

E. Pitler and A. Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives. In *Proc of ACL-IJCNLP 2009 (Short Papers)*, pages 13–16.

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*.

E. Pitler, A. Louis, and A. Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proc. of ACL-IJCNLP 2009*, pages 683–691.

R. Prasad, 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)*.

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. Citeseer.

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

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

S. Somasundaran, J. Wiebe, and J. Ruppenhofer. 2008. Discourse-level opinion interpretation. In *Proc. of Coling 2008*.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proc of HLT-NAACL 2003*.

C. Sporleder and A. Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14:369–416.

W. Wang, J. Su, and C. Tan. 2010. Kernel-based discourse relation recognition with temporal ordering information. In *Proc. of ACL 2010*, pages 710–719.

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. Association for Computational Linguistics.

B. Wellner and J. Pustejovski. 2007. Automatically identifying the arguments of discourse connectives. In *Proc. of EMNLP 2007*, pages 92–101.

B. Wellner, J. Pustejovski, A. Havasi, A. Rumshisky, and R. Suair. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proc. of SIGDIAL2006*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*.

W. Wright. 2008. *A grammar of the Arabic language*. Bibliobazaar.

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*.

Z. Zhou, Y. Xu, Z. Niu, M. Lan, . Su, and Tan. C. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proc. of Coling 2010*, pages 1507–1514.