Investigating the Role of Discourse Markers in Signalling Coherence Relations: A Corpus Study

Debopam Das

Simon Fraser University

E-mail: ddas@sfu.ca

Abstract

In this study, I investigate how coherence relations are signalled by discourse markers, and examine the relationship between coherence relations and discourse markers. In the traditional discourse literature, discourse markers (DMs) are considered to be the only type of signals in discourse. Consequently, coherence relations, based on the presence or absence of DMs, are divided into two groups: explicit (also called signalled) relations: relations signalled by a DM, and implicit (also called unsignalled) relations: relations not signalled by a DM (Martin, 1992; Renkema, 2004; Taboada, 2009; Taboada & Mann, 2006). In order to test the validity of such classification, I conduct a corpus study examining part of the RST Discourse Treebank (Carlson et al., 2002) which includes a collection of 385 Wall Street Journal articles annotated for rhetorical relations (or coherence relations). I select 40 articles from the corpus, examine each and every relation in those articles, and add a new layer of annotation of signalling information to them. In addition, I identify, extract and analyze those relations which are specifically signalled by DMs. Results show that the majority (86%) of the relations examined are signalled, with a breakdown into 22% of the relations signalled by a DM and the remaining 78% of the relations indicated by other types of signals. Furthermore, relations such as Conditions, Contrast and List are most frequently signalled by a DM, while relations such as Background, Elaboration and Summary are rarely or never signalled by a DM. DM-wise, *and, but* and *if* are the most frequently used DMs, while DMs such as *as a result, despite* and *thus* are rarely or never used for signalling a coherence relation.

Keywords: coherence relations, discourse markers, explicit relations, implicit relations, RST, corpus study

Introduction

A discourse is not merely a collection of random utterances. Rather, the components of a discourse are connected to each other in a meaningful way. Coherence relations refer to the types of semantic or pragmatic connections that bind one discourse component to another. For example, in the following text,

(1) John could not go to the party. He was busy with his work.

there are two parts: (i) *John could not go to the party* and (ii) *He was busy with his work.* These parts are connected to each other by a causal relation: John's inability to go to the party is caused by the fact that he was doing his work.

Coherence relations are often signalled by discourse markers (DMs). DMs are lexical expressions (such as *although, because, since* and *thus*) which belong to different syntactic classes, such as conjunctions, adverbials, and prepositional phrases. DMs are used to connect discourse components, and they signal the coherence relations that hold between them. For example, in the following text,

(2) The coach will drop the player from the team **if** he fails the fitness test.

the discourse components are: (i) *The coach will drop the player from the team* and (ii*) he fails the fitness test.* These components are connected to each other by the DM *if*, and this DM signals a Condition relation that holds between these components.

In the traditional discourse literature, DMs are considered to be the only type of signals in discourse. Consequently, coherence relations, based on the presence or absence of the DMs, are divided into two groups: explicit (also called signalled) and implicit (also called unsignalled) relations (Martin, 1992; Renkema, 2004; Taboada, 2009; Taboada & Mann, 2006). Explicit relations are those which are signalled by a DM. For instance, the relation in example (2) will be considered to be explicit since it is signalled by the DM *if*. Implicit relations, in contrast, are not signalled by DMs, and thereby, they remain unsignalled. For instance, consider the following text.

(3) John is tall. Mary is short.

In this text, the discourse components are two sentences, *John is tall* and *Mary is short*, respectively. These components are connected to each other by a Contrast relation. Traditionally, this relation will be considered to be an implicit relation since it does not contain a DM, or, it is not signalled by a DM.

In this study, I question the validity of the traditional notion about the signalling of coherence relations, and test whether DMs are the only type of signals in discourse. Furthermore, I also evaluate the validity of the traditional classification of explicit and implicit relations which is based on the presence and absence of DMs. For this purpose, I investigate the role of DMs in signalling coherence relations, and examine how coherence relations are signalled by DMs. In particular, I examine issues such as exactly what types of relationship hold between relations and DMs, what relations are most (and also least) frequently signalled by DMs, what DMs are most (also least) frequently used in discourse, and finally, whether there are any relations which are signalled by other means in the absence of DMs.

The paper is organized as follows: following the Introduction, I give a brief account of the previous studies on the signalling of coherence relations by DMs. Then, I describe the methodology of the present study. Next, I provide the experimental results followed by the discussion. Finally, I summarize the paper, and provide the conclusion.

Previous Studies

DMs are generally considered to be the most important type of signals in discourse, and accordingly, DMs, among the various types of signals, have been the centre of research on the signalling of relations for a long time (Taboada & Mann, 2006).

The role of DMs in discourse organization has been investigated in many studies in computational linguistics, and more specifically, the knowledge of DMs is most widely used in discourse parsing. In discourse parsing, the discourse structure of a given text is determined by identifying the relationships that hold between the components of that text. Since DMs are the most prominent signals of coherence relations, they are frequently used by many discourse parsing applications to identify the relations as well as to determine the structure of discourse (da Cunha et al., 2012; Forbes et al., 2001; Hernault et al., 2011; Hernault et al., 2010; Le Thanh, 2007; Marcu, 2000; Mithun & Kosseim, 2011; Pardo & Nunes, 2008; Schilder, 2002; Subba & Eugenio, 2009).

It is also important to note that while coherence relations are often signalled by a DM, the correlation between DMs and the relations they signal is not one-to-one; but in the majority of cases, it is one-to-many. For instance, while a single relation can well be signalled by a number of DMs (e.g., Condition is signalled by a number of DMs, such as *if, unless, given* and *since*), a single DM can also indicate a number of relations (e.g., the DM *'but'* is used to indicate a number of relations, such as Circumstance, Concession and Elaboration). As a result, the information solely based on the list of DMs indicative of particular relations often proves to be insufficient for determining relations in unseen texts. As a solution to this problem, discourse parsing systems often rely on different criteria or heuristics in order to specify the accurate relationship between relations and DMs in a more straightforward way. For instance, Marcu (2000) imposes specific conditions upon the usage of DMs, and Pardo and Nunes (2008) employs different templates based on the information of the distribution of DMs.

In the present study, I also adopt a data-driven approach as I examine the relationship between coherence relations and DMs in the context of a corpus analysis.

Theoretical Framework: RST

I used Rhetorical Structure Theory or RST as the theoretical framework of this study. RST is a functional theory of text organization. It describes what parts a text is made of, what kind of relationships exists between these parts, and how parts in a text are organized with respect to each other in order to constitute a coherent piece of discourse. RST was developed by William Mann, Christian Matthiessen and Sandra Thompson in the 1980s. The theory was originally developed as part of research on computational text generation; however, the use of RST has not been restricted to this particular area alone, and it has been used later in various other applications as well.

Text organization in RST is primarily described in terms of relations that hold between two (or sometimes more) non-overlapping text spans. These spans include nucleus, referring to the central or the most important span, and satellite, referring to the peripheral or the secondary span. In cases where each span is equally important, they are considered to be the nuclei, and the relation is multinuclear. Clauses are considered to be basic discourse units, and they enter into coherence relations with each other in a recursive manner.

Relations are mainly defined in terms of Effect, the intention of the writer or speaker in presenting their discourse. The relation inventory in RST is an open set, and the most common relations include Cause, Concession, Condition, Elaboration, Result and Summary.

In Figure 1, I provide an RST analysis of a text (file no. 650) taken from the RST Discourse Treebank.

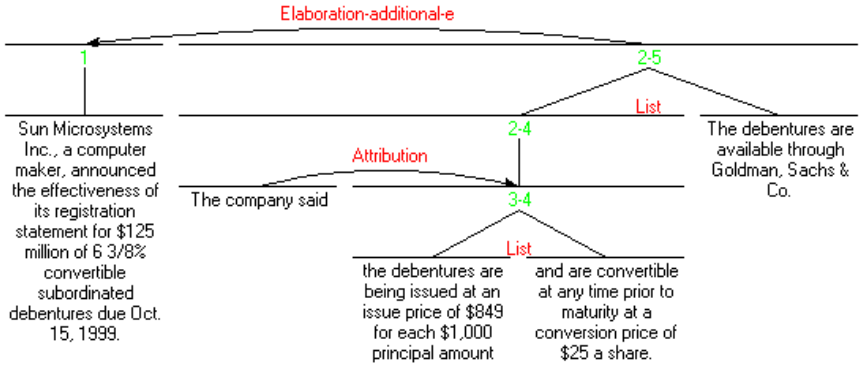

Figure 1: Graphical Representation of an RST Analysis

The RST analysis shows that the text comprises five spans which are represented in the diagram (in Figure 1) by the numbers, 1, 2, 3, 4, and 5, respectively. In the diagram, the arrowheads point to the nuclei and the spans with the tail of an arrow are called the satellites. Span 3 (nucleus) and span 4 (nucleus) are in a multinuclear List relation, and together they make the combined span 3-4. Span 2 (satellite) is connected to span 3-4 (nucleus) by an Attribution relation, and together they make the combined span 2-4. A multinuclear List relation holds between spans 2-4 (nucleus) and 5 (nucleus), and together they make the combined span 2-5. Finally, span 2-5 (satellite) is connected to span 1 (nucleus) by an Elaboration (more specifically, Elaboration-addition) relation.

Methodology

In order to investigate the signalling mechanisms of coherence relations, I conducted a corpus study which comprises a sequence of different tasks: (i) I selected an existing corpus already annotated for coherence relations, (ii) examined each and every relation in that corpus, (iii) identified the signals involved to indicate those relations, and (iv) added to those relations a new layer of annotation of signalling information. In addition, the relations signalled specifically by a DM were identified, extracted and analyzed.

Corpus

I selected the RST Discourse Treebank (Carlson et al., 2002) as my source corpus. The RST Discourse Treebank contains a collection of 385 Wall Street Journal articles (representing over 176,000 words of text) selected from the Penn Treebank (Marcus et al., 1993). The texts of these articles are annotated for rhetorical (or coherence) relations. The corpus is distributed by the Linguistic Data Consortium (LDC), from which the corpus can be downloaded (for a fee).

Clauses are generally considered to be the elementary discourse units in the RST corpus, although there are a few exceptions. The theoretical foundation of the RST corpus is Rhetorical Structure Theory, and the annotations in the corpus are based on a large set of 16 major relation groups which are further divided into 78 RST relations.

The motivation for selecting the RST corpus is two-fold. First, the choice of the RST corpus is at par with the theoretical framework of the present study. I chose to use RST as the theoretical framework for this study, and the RST corpus, as it is already mentioned, is also annotated (for coherence relations) based on RST. Second, in this study I attempt to examine the signalling of relations by DMs at different levels of discourse, and the RST corpus, unlike many other available corpora such as the Penn Discourse Treebank (Prasad et al., 2008), provides annotations not only for the local level relations (between elementary discourse units) but also for the global level relations (between units larger than elementary discourse units).

Procedure

I examined 40 articles from the RST Discourse Treebank which is approximately one-tenth of the 385 articles present in the corpus. The texts in these articles contain 1,304 coherence relations. I examined manually each of those relations, and identified the signals used to mark those relations.

For visualizing the discourse structure of a text, I used an existing tool, called the RST Tool (O'Donnell, 1997), to open the corresponding RST file in the corpus. The RST Tool provides a graphical representation of the annotated RST files in the form of tree-diagrams. Next, I identified the signal(s) that potentially indicate(s) the corresponding relation. For identifying the relations signalled by DMs, I mainly followed Fraser's (1999; 2006; 2009) definition of DMs, that is, DMs constitute a functional class of lexical expressions drawn from different syntactic classes, such as conjunctions, adverbials and prepositional phrases. In addition, they connect discourse segments, and signal a semantic or pragmatic relationship between them.

Beyond DMs, I also found other classes of cues that have been mentioned in previous studies, or that I identified in the present corpus work. The classification has a top-level breakdown into morphological, syntactic, semantic, entity, lexical, genre and graphical features, plus heuristics specific to each relation. I started my annotation by consulting previous studies for indication of what signalling devices other than DMs have been found in corpora (Bateman et al., 2001; Corston-Oliver, 1998; Dale, 1991a, 1991b; Halliday & Hasan, 1976; Knott, 1996; Knott & Dale, 1994; Lapata & Lascarides, 2004; Le Thanh, 2007; Lin et al., 2009; Louis et al., 2010; Marcu, 1999, 2000; Pardo & Nunes, 2008; Pitler et al., 2009; Polanyi et al., 2004; Prasad et al., 2010; Scott & de Souza, 1990; Sporleder & Lascarides, 2005, 2008; Theijssen, 2007; Theijssen et al., 2008). When confronted with a new instance of a particular type of relation, I consulted the signals found from previous studies, and tried to search from there appropriate signal(s) that could best function as the indicator for that relation instance. If my search led me to assigning an appropriate signal (or more than one signal) to the relation, I considered myself successful in identifying the signal(s) for that relation. If there

was no appropriate match, then I examined the context (comprising the spans) of the actual relation instance, and intuitively tried to discover a new signal that was actually present in the context spans to signal that relation.

Coding

In the coding task, I provided annotations for signals of coherence relations, or in other words, I added signalling information to the existing relations from the RST corpus. For this purpose, I extracted the signals identified, and documented them along with the relevant information about the relation in question, the document number (to which the relation belongs), the status of the spans (i.e., nucleus or satellite), and the span numbers (i.e., the location of the spans in the text). I annotated the signalling information in a separate Excel file, since the RST Tool does not allow multiple levels of annotation. The Excel file represents two different levels of annotations: (i) the annotation of the relations signalled by a DM, and (ii) the annotations of the relations indicated by other signals in the absence of a DM.

Results and Discussion

Among the 1,304 relations examined, the distribution of signalled relations (indicated either by a DM or by some other signal) and unsignalled relations (not indicated by any signal) is provided in Table 1.

Relation Type	Tokens	Percentage	Total
Relations signalled (by DMs or other signals)	1,127	86.43%	1,304
Relations not signalled (by any signals)	177	13.57%	

Table 1: Distribution of signalled and unsignalled relations

The results show that 1,127 relations (86.43%), out of all the 1,304 relations, are indicated, either by a DM or with the help of some other signal. On the other hand, no significant signals are found for the remaining 177 relations (13.57%).

Among the 1,127 signalled relations, the basic distribution of the relations signalled by a DM and the relations indicated by other signal(s) is provided in Table 2.

Relation Type	Tokens	Tokens of Overlapping Relations	Percentage	Total
Relations signalled by a DM	251	2 (2 instances of List relation)	22.27%	1,127
Relations indicated by other signals	878		77.91%	

Table 2: Distribution of relations signalled by a DM and of relations indicated by other signals

Among the 1,127 signalled relations, DMs are used to signal 251 relations (22.27% of the signalled relations), while 878 relations (77.91% of the signalled relations) are marked with the help of some other signal in the absence of a DM. In addition, there are two instances of List relation which are indicated by both a DM and some other signal. This is because these relations are multinuclear, consisting of three or four nuclei, and it is found that while a nucleus is connected to another nucleus by a DM, a third nucleus is related to any of the two former nuclei (in case of a tri-nuclei relation), or to a fourth nuclei (in case of a tetra-nuclei relation) by means of some other signal.

In the corpus analysis, 22.27% of the signalled relations, and 19.25% of all the relations (251 out of 1,304 relations) are indicated by DMs. This measure is much lower than the results documented in many previous studies on the signalling of relations by DMs. For instance, Taboada (2006) shows that, in smaller corpora, 31% of the relations in conversation and 43% of the relations in newspaper articles are signalled by DMs. In the PDTB, 45.47% relations are shown to be explicit, or are signalled by a DM (Prasad et al., 2007). Renkema (2009) states that in the standard RST corpus (Carlson et al., 2002), 28% of the relations contain DMs. One possible reason for the lower score in the present experiment could be that I used a fairly strict definition of DMs, and the criteria I used for considering an expression to be a DM excludes many expressions which are treated as DMs elsewhere. For instance, I did not consider expressions such as *at the same time, for the simple reason,* and *in other respects* to be the examples of DMs, but considered them to be *indicative phrases* of *lexical* type of signal; however, these expressions are included within the class of DMs in other studies such as in Knott (1996).

The relative distribution of relations with respect to whether they are indicated by a DM, by some other signal, or whether they are unsignalled is provided in Table 3.

No.	Relation Group	Relation	# relations signalled by DMs	# relation indicated by other signals	# relations not signalled	Total
1.	Attribution	Attribution	0	228	3	231
		Attribution-negative	0	0	0	0
2.	Background	Background	2	8	6	16
		Circumstance	21	9	9	39
3.	Cause	Cause	2	1	1	4
		Result	3	0	0	3
		Consequence	14	1	12	27
4.	Comparison	Comparison	5	9	4	18
		Preference	0	0	0	0
		Analogy	0	0	0	0
		Proportion	0	0	0	0
5.	Condition	Condition	15	1	1	17
		Hypothetical	1	1	0	2
		Contingency	0	0	0	0
		Otherwise	0	0	0	0
6.	Contrast	Contrast	19	2	2	23
		Concession	13	0	1	14
		Antithesis	25	1	4	30
7.	Elaboration	Elaboration-additional	23	238	41	302
		Elaboration-general-specific	1	16	4	21
		Elaboration-part-whole	0	0	0	0
		Elaboration-process-step	0	0	0	0
		Elaboration-object-attribute	4	179	3	186
		Elaboration-set-member	0	6	1	7
		Example	3	6	8	17
		Definition	0	2	0	2
8.	Enablement	Purpose	0	39	0	39
		Enablement	0	0	0	0
9.	Evaluation	Evaluation	1	3	1	5
		Interpretation	1	0	9	10
		Conclusion	0	0	0	0
		Comment	0	0	9	9

10.	Explanation	Evidence	0	3	8	11
		Explanation-argumentative	6	1	23	30
		Reason	12	1	4	17
11.	Joint	List	50	27	6	83
		Disjunction	3	0	0	3
12.	Manner-Means	Manner	3	0	0	3
		Means	1	4	0	5
13.	Topic-Comment	Problem-solution	2	2	2	6
		Question-answer	0	0	0	0
		Statement-response	0	2	0	2
		Topic-comment	1	0	0	1
		Comment-topic	0	0	0	0
		Rhetorical-question	0	0	0	0
14.	Summary	Summary	0	0	8	8
		Restatement	0	9	0	9
15.	Temporal	Temporal-before	3	0	0	3
		Temporal-after	7	1	0	8
		Temporal-same-time	3	1	0	4
		Sequence	5	0	0	5
		Inverted-sequence	0	0	0	0
16.	Topic-Change	Topic-shift	0	0	4	4
		Topic-drift	0	0	0	0
17.	Same-Unit	Same-unit	2	76	3	81
18.	Span	Span	0	0	0	0
19.	Textual Organization	Textual organization	0	1	0	1
	Total		251 (19.25%)	878 (67.33%)	177 (13.57%)	1,306[1]

Table 3: Distribution of relations indicated by a DM, of relations indicated by some other signal, and of unsignalled relations

The distribution in Table 3 shows that almost every group of relations is more or less marked either by a DM or by some other signal. In particular, it is found that relation groups such as Attribution, Elaboration, Enablement and Joint are most frequently marked, either by a DM or by some other signal. On the other hand, there is only one group, Evaluation, which is rarely marked by any signal.

Among the signalled relations, it is found that DMs are most frequently used to signal individual relations such as Circumstance, Result, Consequence, Condition, Concession, Contrast, Antithesis, Reason and List. In contrast, individual relations such as Attribution, Background, Comparison, Elaboration-additional, Elaboration-general-specific, Elaboration-object-attribute, Example and Purpose are rarely or never signalled by a DM. These findings are also parallel to the results presented by Taboada (2006) who shows that relations such as Concession, Circumstance and Result are most frequently signalled (by a DM), while Background, Elaboration and Summary are rarely signalled (by a DM).

For the 251 instances of relations signalled by a DM, I have found 58 different DMs. Examples of some of these DMs include *after, although, and, as, as a result, because, before, despite, for example, however, if, in addition, moreover, or, since, so, thus, unless, when* and *yet*. I have provided an exhaustive list of the extracted DMs in Table 4.

[1] The total is actually 1,304, but two relations are counted twice, since there are two instances of List relation which are indicated by DMs and other signals at the same time.

No.	DM	No.	DM	No.	DM	No.	DM
1.	(largely) because	16.	Despite	31.	More provocatively	46.	Though
2.	(simply) because	17.	Even before	32.	Moreover	47.	Thus
3.	After	18.	Even though	33.	Now	48.	Unless
4.	Although	19.	Even when	34.	Once	49.	Until
5.	And	20.	Except when	35.	Or	50.	Until recently
6.	And after	21.	For example	36.	Previously	51.	When
7.	As	22.	Further	37.	Rather	52.	Where
8.	As a result	23.	However	38.	Recently	53.	Whereby
9.	As a result of	24.	If	39.	Since	54.	Whether
10.	As well as	25.	In addition	40.	Since… Now	55.	While
11.	Because	26.	In addition to	41.	So	56.	With
12.	Because of	27.	In case	42.	So far	57.	Without
13.	Before	28.	Indeed	43.	So that	58.	Yet
14.	But	29.	Instead	44.	Still		
15.	But now	30.	Meanwhile	45.	Thereafter		

Table 4: List of DMs extracted in the corpus study

The relation-wise distribution of different DMs shows that a significant number of relation groups are frequently signalled by a wide variety of DMs. These groups include Background, Cause, Condition, Contrast, Explanation, Joint, and Temporal relations. In Table 5[2], I have provided the distribution of the most frequently occurring DMs with respect to each of these relation groups (also significant specific relations) which are signalled by them.

Common Relation Group	Common Relation	Most Frequently Occurring DMs
Background (23)	Circumstance (21)	when (5), as (4), with (3),
Cause (19)	Consequence (14)	and (6)
Condition (16)	Condition (15)	if (11), unless (2)
Contrast (57)	Contrast (19)	but (11), however (3)
	Concession (13)	while (3), but (2), though (2)
	Antithesis (25)	but (11), although (3), however (3)
Explanation (18)	Reason (12)	and (4), because (4), because of (3)
Joint (53)	Disjunction (3)	or (3)
	List (50)	and (44), in addition (2), moreover (2)
Temporal (18)	Sequence (5)	and (4)
	Temporal-after (7)	since (3), after (2)
	Temporal-before (3)	before (3)

Table 5: Distribution of most frequently occurring DMs with respect to most common relations signalled by them

The relation-wise distribution of different DMs (shown in Table 5) shows what DMs are most frequently used to convey a particular relation, and how frequently they are used for signalling that relation. For instance, List relations are most frequently signalled by *and*, *in addition*, and *moreover*. In addition, out of the 50 instances of List relation, the DMs *and, in addition* and *moreover* are used 44 (88%), 2 (4%), and 2 (4%) times, respectively.

In an alternate combination, I have provided in Table 6 the distribution of the most common relations with respect to the most frequently occurring DMs which are used to signal them.

[2] **Conventions for Table 5 (and Table 6):** The numerical value within parentheses followed by a relation/relation group refers to the number of instances the relation/relation group is signalled by a DM. On the other hand, the numerical value within parentheses followed by a DM refers to the number it is used to signal the corresponding relation (in the same row).

Frequently Occurring DM	Common Relation Group	Common Relation(s)
although (5)	Contrast (5)	Antithesis (3)
and (70)	Cause (7)	Consequence (6)
	Elaboration (8)	Elaboration-additional (8)
	Joint (44)	List (44)
	Explanation (4)	Reason (4)
	Temporal (4)	Sequence (4)
as (8)	Background (4)	Circumstance (4)
	Elaboration (2)	Elaboration-additional (2)
because (8)	Cause (2)	Consequence (2)
	Explanation (6)	Explanation-argumentative (2)
		Reason (4)
because of (6)	Explanation (4)	Reason (3)
before (4)	Temporal	Temporal-before (3)
but (35)	Contrast (25)	Antithesis (11)
		Concession (3)
		Contrast (11)
	Elaboration (6)	Elaboration-additional (6)
however (9)	Contrast (6)	Antithesis (3)
		Contrast (3)
if (13)	Condition (11)	Condition (11)
since (5)	Temporal (3)	Temporal-after (3)
when (10)	Background (5)	Circumstance (5)
while (8)	Comparison (3)	Comparison (3)
	Contrast (4)	Concession (3)
with (4)	Background (3)	Circumstance (3)
without (6)	Manner-Means	Manner (3)

Table 6: Distribution of most common relations with respect to most frequently occurring DMs

The DM-wise distribution of different relations (in Table 6) shows what relations are most frequently signalled by a particular DM, and how frequently they are signalled by that DM. For instance, the DM *but* is mostly frequently used to signal Contrast and Elaboration relation groups. In addition, out of the 35 instances of *but*, Contrast and Elaboration groups are signalled 25 (71%) and 6 (17%) times, respectively.

The distribution of DMs in Table 5 implies that a particular set of DMs can be associated with a single relation which is most frequently signalled by them. Similarly, the distribution of relations in Table 6 implies that a particular set of relations can be associated with a DM which is most frequently used to signal those relations. However, I, at the same time, feel that these small-scale statistics produced by this corpus analysis are not sufficient to make any generalization about the relationship between a relation and a set of DMs, or between a DM and a set of relations. Rather, a statistical analysis which can potentially represent the aforementioned associations requires a large body of data. In fact, one important future development of this study can be re-doing the same experiment with a larger amount of the same type of data, for example, the entire RST corpus. The increase of the data size will definitely provide an opportunity to extract much larger number as well as wider variety of representative samples, and in this way, more conclusive generalizations about the association between relations and their signalling DMs can be drawn.

One of the objectives of this study was to examine whether DMs are the only signals of coherence relations in discourse. The findings from the present study suggest that this is, in fact, not the case. The results show that only 19.25% of the

relations in the corpus are signalled by a DM while the rest of the relations do not contain any DMs. This suggests that although DMs may constitute the most straightforward type of signals, they are used to indicate only a small fraction of relations in discourse. In contrast, results show that 67.33% of the relations in the corpus are indicated by other signals. This has two important implications. First, coherence relations in a discourse, in general, are signalled (either by DMs or by other signals). Second, relations, in the absence of DMs, are indicated by other signals in the majority of cases. For instance, consider the text provided earlier in example (3) in the Introduction once again.

(4) John is tall. Mary is short.

There is a Contrast relation holding between the text components, but the relations is not signaled by a DM. Instead, the relation is indicated by two types of other signals. One can notice that there is a parallel syntactic construction (i.e., Subject – Copular Verb – Adjective) which is shared by the two sentences (functioning as the discourse components). This syntactic signal is often used to indicate a Contrast relation. Furthermore, the relation is also indicated by the words *tall* and *short* in the respective sentences. These words are antonyms, and this particular meaning relationship is also a good marker of Contrast relations. In this way, we can see that the knowledge of signals other than DMs can be beneficial for identifying coherence relations when no DM is present. In fact, another important future development of the present study can be a thorough investigation of the role of other signals in indicating coherence relations.

The other objective of the present study was to evaluate the validity of the traditional distinction of explicit and implicit relations. According to the traditional discourse literature, the explicit relations are those which are signalled by DMs, and implicit relations are those which are not signalled by DMs. The present study shows that relations can be indicated by DMs as well as by other signals. One of the important characteristics of the signals (including DMs) of coherence relations is that they are all textual in nature. They are intrinsic to the texts which contain the relations, and one can infer or determine their identity with the help of linguistics knowledge. This implies that the signals of coherence relations are actually present in the discourse, and since they are textual or linguistic in nature, they must have some explicit manifestations. So, one reasonable suggestion in this case can be that the category explicit relations should not comprise only DMs but also other signals which have explicit presence in discourse. On the other hand, the category implicit can be defined in terms of the absence of any signal.

Conclusion

In this study, I investigated the signalling of coherence relations by DMs, and examined the relationship between coherence relations and DMs. I found out that around 20% of the relations present in a corpus are signalled by DMs, while the remaining 80% of the relations are not signalled by them. Relation-wise, DMs are most frequently used to signal relations such as Condition, Contrast and List, while relations such as Background, Elaboration and Summary are rarely or never signalled by DMs. With respect to DMs, *and*, *but* and *if* are the most frequently

used DMs, while DMs such as *as a result*, *despite* and *thus* are rarely or never used for signalling a coherence relation.

I also found out that the majority of the relations (86%) present in a discourse are signalled. Most importantly, among the signalled relations, only a very small proportion of relations (22%) are signalled by DMs, while the majority of signalled relations (78%) are indicated by various other signals. This crucially points to the fact that the signalling of coherence relations is not confined to the use of DMs alone, and relations can well be indicated by other signals in the absence of DMs. Unfortunately, while research on the signalling of coherence relations has been restricted mainly to the study of DMs alone, the other signalling devices remain relatively understudied (Taboada, 2009). The present study suggests that investigation of other signals can itself be an avenue for future research on the signalling of relations.

Reference

Bateman, John, Kamps, Thomas, Kleinz, Jörg, & Reichenberger, Klaus. (2001). Towards constructive text, diagram, and layout generation for information presentation. Computational Linguistics, 27(3), 409-449.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). RST Discourse Treebank, LDC2002T07 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Corston-Oliver, Simon. (1998). Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. Paper presented at the AAAI 1998 Spring Symposium Series, Intelligent Text Summarization, Madison, Wisconsin.

da Cunha, Iria, Juan, Eric San, Torres-Moreno, Juan Manuel, Cabré, María Teresa, & Sierra, Gerardo. (2012). A symbolic approach for automatic detection of nuclearity and rhetorical relations among intra-sentence discourse segments in Spanish. Paper presented at the CICLing, New Delhi, India.

Dale, Robert. (1991a). Exploring the Role of Punctuation in the Signalling of Discourse Structure. Paper presented at the Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI, Technical University of Berlin.

Dale, Robert. (1991b). The role of punctuation in discourse structure. Paper presented at the AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation, Asilomar, CA.

Forbes, Katherine, Miltsakaki, Eleni, Prasad, Rashmi, Sarkar, Anoop, Joshi, Aravind K., & Webber, Bonnie. (2001). D-LTAG system - Discourse parsing with a lexicalised Tree Adjoining Grammar. Paper presented at the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics, Helsinki, Finland.

Fraser, Bruce. (1999). What are discourse markers? Journal of Pragmatics, 31, 931 - 953.

Fraser, Bruce. (2006). Towards a theory of discourse markers. In K. Fischer (Ed.), Approaches to Discourse Particles (pp. 189 - 204): Elsevier Press.

Fraser, Bruce. (2009). An account of Discourse Markers. International Review of Pragmatics, 1, 293–320.

Halliday, Michael A.K., & Hasan, Ruqaiya. (1976). Cohesion in English. London: Longman.

Hernault, Hugo, Bollegala, Danushka, & Ishizuka, Mitsuru. (2011). Semi-supervised discourse relation classification with structural learning. Paper presented at the 12th international conference on Computational linguistics and intelligent text processing (CICLing '11), Tokyo, Japan.

Hernault, Hugo, Prendinger, Helmut, duVerle, David A., & Ishizuka, Mitsuru. (2010). HILDA: A discourse parser using Support Vector Machine classification. Dialogue and Discourse, 1(3).

Knott, Alistair. (1996). A data-driven methodology for motivating a set of coherence relations. Unpublished Ph.D. dissertation, University of Edinburgh, Edinburgh, UK.

Knott, Alistair, & Dale, Robert. (1994). Using linguistic phenomena to motivate a set of coherence relations. Discourse Processes, 18(1), 35–62.

Lapata, Mirella, & Lascarides, Alex. (2004). Inferring sentence-internal temporal relations. Paper presented at the NAACL-04.

Le Thanh, Huong. (2007). An approach in automatically generating discourse structure of text. Journal of Computer Science and Cybernetics, 23(3), 212-230.

Lin, Ziheng, Kan, Min-Yen, & Ng, Hwee Tou. (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. Paper presented at the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore.

Louis, Annie, Joshi, Aravind, Prasad, Rashmi, & Nenkova, Ani. (2010). Using Entity Features to Classify Implicit Discourse Relations. Paper presented at the SIGDIAL 2010.

Marcu, Daniel. (1999). A decision-based approach to rhetorical parsing. Paper presented at the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland.

Marcu, Daniel. (2000). The rhetorical parsing of unrestricted texts: A surface based approach. Computational Linguistics, 26(3), 395-448.

Marcus, Mitchell, Santorini, Beatrice, & Marcinkiewicz, Mary Ann. (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19(2), 313-330.

Martin, James R. (1992). English Text: System and Structure. Amsterdam and Philadelphia: John Benjamins.

Mithun, Shamima, & Kosseim, Leila. (2011). Comparing approaches to tag discourse relations. Paper presented at the 12th international conference on Computational linguistics and intelligent text processing (CICLing '11), Tokyo, Japan.

O'Donnell, Michael. (1997). RSTTool, from http://www.wagsoft.com/RSTTool/

Pardo, Thiago Alexandre Salgueiro, & Nunes, Maria das Gracas Volpe. (2008). On the Development and evaluation of a Brazilian Portuguese discourse parser. Journal of Theoretical and Applied Computing, 15(2), 43-64.

Pitler, Emily, Louis, Annie, & Nenkova, Ani. (2009). Automatic sense prediction for implicit discourse relations in text. Paper presented at the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore.

Polanyi, Livia, Culy, Chris, van den Berg, Martin, Thione, Gian Lorenzo, & Ahn, David. (2004). A rule based approach to discourse parsing. Paper presented at the SigDIAL 2004, Cambridge, MA.

Prasad, Rashmi, Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The penn discourse treebank 2.0. Paper presented at the 6th International Conference on Language Resources and Evaluation (LREC).

Prasad, Rashmi, Joshi, Aravind, & Webber, Bonnie. (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. Paper presented at the COLING 2010, Beijing.

Prasad, Rashmi, Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L. , & Webber, B. (2007). The Penn Discourse Treebank 2.0 Annotation Manual.Unpublished manuscript.

Renkema, J. (2004). Introduction to Discourse Studies. Amsterdam: Benjamins.

Renkema, J. (2009). The Texture of Discourse. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Schilder, Frank. (2002). Robust discourse parsing via discourse markers, topicality and position. Natural Language Engineering, 8(2/3), 235-255.

Scott, Donia, & de Souza, Clarisse Sieckenius. (1990). Getting the message across in RST-based text generation. In R. Dale, C. Mellish & M. Zock (Eds.), Current Research in Natural Language Generation (pp. 47-73). London: Academic Press.

Sporleder, Caroline, & Lascarides, Alex. (2005). Exploiting linguistic cues to classify rhetorical relations. Paper presented at the Recent Advances in Natural Language Processing (RANLP-05).

Sporleder, Caroline, & Lascarides, Alex. (2008). Using automatically labelled examples to classify rhetorical relations: An assessment. Natural Language Engineering, 14, 369–416.

Subba, Rajen, & Eugenio, Barbara Di. (2009). An effective discourse parser that uses rich linguistic information. Paper presented at the HLT-ACL 2009, Boulder, CO.

Taboada, Maite. (2006). Discourse markers as signals (or not) of rhetorical relations. Journal of Pragmatics, 38(4), 567-592.

Taboada, Maite. (2009). Implicit and explicit coherence relations. In J. Renkema (Ed.), Discourse, of Course. Amsterdam: John Benjamins.

Taboada, Maite, & Mann, William C. (2006). Rhetorical Structure Theory: Looking Back and Moving Ahead. Discourse Studies, 8(3), 423-459.

Theijssen, Daphne. (2007). Features for automatic discourse analysis of paragraphs. Unpublished MA, Radboud University Nijmegen, The Netherlands.

Theijssen, Daphne, van Halteren, Hans, Verberne, Suzan, & Boves, Lou. (2008). Features for automatic discourse analysis of paragraphs. Paper presented at the 18th meeting of Computational Linguistics in the Netherlands (CLIN 2007).