EXPLICIT AND IMPLICIT COHERENCE RELATIONS: A CORPUS STUDY

Debopam Das and Maite Taboada
Simon Fraser University

1. 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 (also known as discourse or rhetorical 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 components: (i) *John could not go to the party* and (ii) *He was busy with his work*. These components 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 holding between those components. 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 on signalling (the linguistic marking of a relation), DMs are considered to be the only signals of coherence relations (Taboada and Mann 2006). Consequently, coherence relations, based on the presence or absence of DMs, are divided into two groups: explicit (also called signalled) relations and implicit (also called unsignalled) relations (Martin 1992; Renkema 2004; Taboada 2009). 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. Consider the following text.

(3) John is tall. Mary is short.

Actes du congrès annuel de l'Association canadienne de linguistique 2013.
Proceedings of the 2013 annual conference of the Canadian Linguistic Association.

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, we question the traditional notion about the relation signalling, and evaluate the validity of the classification of explicit and implicit relations. We hypothesize that the signalling of coherence relations is not confined to the use of DMs alone. In other words, the absence of DMs does not imply that there is no signal, as a signal must be necessary for correct interpretation. We argue that the Contrast relation in example (3) is not unsignalled; rather it is indicated by two signals other than DMs. First, the two components (or sentences) in the text share a parallel syntactic construction (subject-copula-adjective) which is a strong signal for Contrast relation. Second, the relation is also signalled by the antonymy relationship between the words *tall* and *short* in the respective sentences.

In order to test this hypothesis, we examine what signals are used to convey coherence relations and how they are used. We also examine whether coherence relations are more frequently explicit or implicit in terms of the type of signalling involved. For this purpose, we first select a corpus already annotated for coherence relations, then examine the relations in the corpus, and finally add information on how those relations are signalled, including a variety of possible signals.

In this paper, we begin with discussing the notion of coherence relations in Rhetorical Structure Theory (Mann and Thompson 1988), also adopted as the theoretical framework of this study. Then we provide a brief account of the previous studies on the signalling of coherence relations, and propose a classification of signalling devices, which we use to annotate our corpus. We also describe the corpus and annotation procedure, followed by the discussion of the results. The paper concludes with the future development of this study, and the applications that the corpus will have.

2. Coherence Relations and Rhetorical Structure Theory

Coherence relations have been extensively investigated in the framework of Rhetorical Structure Theory or RST (Mann and Thompson 1988). In RST, relations are defined through different fields, the most important of which is the Effect, the intention of the writer (or speaker) in presenting their discourse. Relation inventories are open, and the most common ones include names such as Cause, Concession, Condition, Elaboration, Result or Summary. Relations can be multinuclear, reflecting a paratactic relationship, or nucleus-satellite, a hypotactic type of relation. The names nucleus and satellite refer to the relative importance of each of the relation components.

Texts, according to RST, are built out of basic clausal units that enter into rhetorical relations with each other, in a recursive manner. Mann and Thompson proposed that most texts can be analyzed in their entirety as recursive applications of different types of relations. In effect, this means that an entire text can be analyzed as a tree structure, with clausal units being the branches and relations the nodes.

We provide the RST annotation of a text taken from the RST Discourse Treebank (Carlson *et al.* 2002). The file contains the following text.

(4) Sun Microsystems Inc., a computer maker, announced the effectiveness of its registration statement for $125 million of 6 3/8% convertible subordinated debentures due Oct. 15, 1999.

The company said the debentures are being issued at an issue price of $849 for each $1,000 principal amount and are convertible at any time prior to maturity at a conversion price of $25 a share.

The debentures are available through Goldman, Sachs & Co.

The graphical representation of the RST analysis of this text is provided in Figure 1.

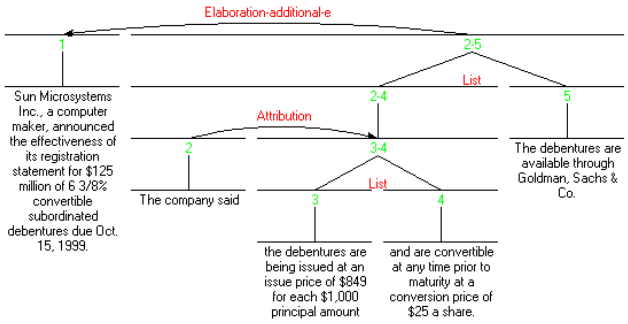

Figure 1: Graphical representation of an RST analysis

The RST analysis in Figure 1 shows that the text comprises five spans, represented in the diagram by the numbers, 1, 2, 3, 4, and 5, respectively. The arrowheads point to the nuclei and the spans with the tail of an arrow refer to 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), together making the combined span 2-5. Finally, span 2-5 (satellite) is connected to span 1 (nucleus) by an Elaboration (more specifically, Elaboration-addition) relation.

2. Signalling of Coherence Relations

Research on coherence relations has often focused on cues that indicate the presence of a relation, or the lack of such cues, as many relations seem to be unsignalled. Whereas it is true that many coherence relations are not signalled by a DM, that is, they are *implicit*, it is also often the case that other markers have been understudied (Taboada 2009; Taboada and Mann 2006). Our goal in this paper is to push that line of research further. We explore how many, and what types of cues can be found if we study signalling beyond DMs. A secondary goal aims at discovering whether unsignalled or implicit relations can be said to exist at all.

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 relation signalling for a long time (Taboada and Mann 2006). The knowledge of DMs is investigated in various NLP applications, such as discourse parsing. In discourse parsing, the discourse structure of a given text is determined by identifying the relationships that hold between the text components. Since DMs are the most prominent signals of coherence relations, they are frequently used by many discourse parsing applications to identify the relations in a text as well as to determine the structure of a discourse (da Cunha *et al.* 2012; Forbes *et al.* 2001; Hernault *et al.* 2011; Hernault *et al.* 2010; Le Thanh 2007; Marcu 2000; Mithun and Kosseim 2011; Pardo and Nunes 2008; Schilder 2002; Subba and Eugenio 2009).

In psycholinguistic research, DMs are considered to be the processing instructions which guide the readers to recognize coherence relations. Subsequently, it is assumed that DMs must have a positive influence on the readers' understanding of a discourse and on the readers' recall performance in retrieving the textual information. Most studies on text processing suggest that DMs accelerate text processing, i.e., the presence of DMs, during reading tasks, leads to a faster processing of the immediately following text segment (Britton *et al.* 1982; Haberlandt 1982; Sanders and Noordman 2000; Sanders *et al.* 1992). However, the effects of signalling by DMs on recall show a somewhat mixed pattern. For instance, some studies suggest that the presence of DMs have a positive effect on the mental representation of a discourse, i.e., subjects perform better while a DM is present (Loman and Mayer 1983; Lorch and Lorch 1986; Meyer *et al.* 1980; Millis and Just 1994). In contrast, studies such as Meyer (1975), Sanders and Noordman (2000), Sanders et al. (1992), or Spyridakis & Standal (1987) show that DMs do not contribute to cognitive representation, and they do not have any significant effect on it. Furthermore, some scholars even claim that DMs have a negative effect on the readers' recall performances, and they hinder the process of cognitive representations (Millis *et al.* 1993).

The problem of considering DMs to be the only type of signals is that DMs account for only a small fraction of relations present in a discourse, thereby leaving the majority of relations without DMs. This raises an obvious question: how are coherence relations signalled in the absence of DMs? If we postulate the psychological validity for coherence relations, that is, if we assume that coherence relations are present in discourse and that they are recognized by speakers, then there must be signals through which speakers identify relations when parsing discourse. Unfortunately, research on the signalling of relations by signals other than DMs is not abundant. There are only a few computational studies which employ the knowledge of other signals for identifying the presence or nature of coherence relations in the absence of DMs. The signals used in those studies include features such as tense and mood (Scott and de Souza 1990); clausal status, anaphora and polarity (Corston-Oliver 1998); lexical chains and cohesive devices (Marcu 1999, 2000); punctuation and graphical features (Dale 1991a, 1991b); textual layout (Bateman *et al.* 2001); synonym/antonym, parallel syntactic structure and topic/focus (Polanyi *et al.* 2004); NP/VP cues and Reiterative devices (Le Thanh 2007); morphosyntactic and genre-related information (Pardo and Nunes 2008); and syntactic similarity, word overlap, proper nouns and definite articles (Theijssen 2007).

3. Signals for Reliable Annotation

The most important aspect of the annotation was to select and classify the types of signals to annotate. We built our taxonomy of signals based on the different classes of relational markers that we identified in our preliminary corpus work, or that have been mentioned in previous studies (Bateman *et al.* 2001; Blakemore 1987, 1992, 2002; Corston-Oliver 1998; Dale 1991a, 1991b; Fraser 1990, 1999, 2006, 2009; Halliday and Hasan 1976; Knott 1996; Knott and Dale 1994; Lapata and Lascarides 2004; Le Thanh 2007; Lin *et al.* 2009; Louis *et al.* 2010; Marcu 1999, 2000; Pardo and Nunes 2008; Pitler *et al.* 2009; Polanyi *et al.* 2004; Prasad *et al.* 2010; Prasad *et al.* 2007; Sanders *et al.* 1992, 1993; Schiffrin 1987, 2001; Scott and de Souza 1990; Sporleder and Lascarides 2005, 2008; Theijssen 2007). The taxonomy is organized hierarchically in three levels: *signal class*, *signal type* and *specific signal*. The top level, *signal class*, has three tags representing three major classes of signals: *single*, *combined* and *unsure*. For each class, a second level of types is defined; for example, the class *single* is divided into nine types (*DM*, *reference*, *lexical*, *semantic*, *morphological*, *syntactic*, *graphical*, *genre* and *numerical features*). Finally, the third level in the hierarchy refers to the specific signals; for example, *reference type* has four specific signals: *personal*, *demonstrative*, *comparative* and *propositional reference*. A glimpse of the taxonomy is provided in Figure 2. Note that subcategories are only illustrative, not exhaustive. More detail on the taxonomy can be found in Taboada and Das (2013).

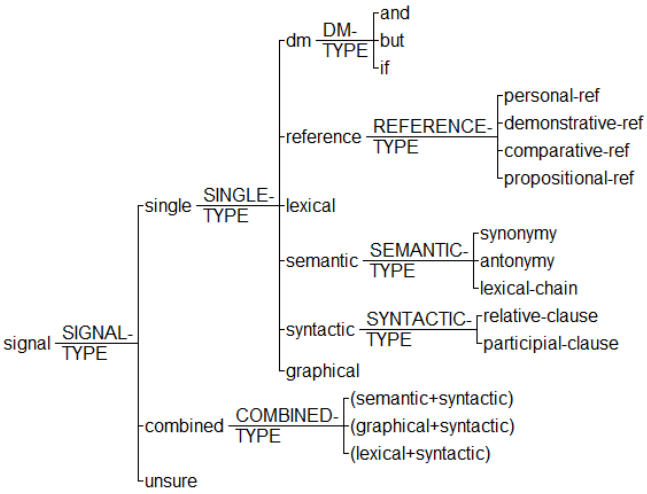

Figure 2: Hierarchical taxonomy of signals

In addition, we find that many relations are indicated by combined signals. Combined signals are made of two or more single signals which work together to indicate a particular relation. We have identified 10 broad types of combined signals[1]: (i) *entity + positional*, (ii) *entity + syntactic + lexical*, (iii)

[1] For more detail on combined signals, see Taboada and Das (2013).

entity + syntactic, (iv) *graphical + syntactic*, (v) *lexical + positional*, (vi) *lexical + syntactic + positional*, (vii) *lexical + syntactic*, (viii) *syntactic + lexical*, (ix) *syntactic + positional*, and (x) *semantic + syntactic*.

3. Annotation Process

For our corpus, we have selected the RST Discourse Treebank or RST-DT (Carlson *et al.* 2002), a collection of 385 Wall Street Journal articles (financial reports, general interest stories, editorials, etc.) annotated for RST relations[2]. The annotated texts in the RST-DT are stored as LISP files which can be opened with RSTTool (O'Donnell 1997) for visual representation.

The motivation for selecting the RST-DT is two-fold. First, the choice of the RST-DT is at par with the theoretical framework of the present study. We chose to use RST as the theoretical framework for this study, and the RST-DT, as it is already mentioned, is also annotated (for coherence relations) based on RST. Second, in this study we attempt to examine the signalling of relations at different levels of discourse, and the RST-DT, 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).

In our preliminary corpus study, we annotated 1,306 relations in 40 articles which constitute approximately ten percent of the 385 articles in the RST-DT. The annotation process involves examining each relation and, assuming the relation annotation is correct, searching for cues that indicate that such relation is present. In some cases, more than one cue may be present. When confronted with a new instance of a particular type of relation, we consult our taxonomy of signals, and find appropriate signal(s) that could best function as the indicator(s) for that relation instance. If our search led us to assigning an appropriate signal (or more than one appropriate signal) to that relation, we declared success in identifying the signal(s) for that relation. If our search does not match any of the signals in the taxonomy, then we examine the context (comprising the spans) to discover any potential new signals. If a new signal is identified, we include it in the appropriate category in our existing taxonomy. In this way, we proceed through identifying the signals of the relations in the corpus, and, at the same time, keep on updating our taxonomy with new signalling information, if necessary. We found that after approximately 20 files, or 650 relations, we added very few new signals to the taxonomy.

In the coding task, we provided annotations for signals of coherence relations, or in other words, we added signalling information to the existing relations from the RST-DT. For this purpose, we extracted the signals identified, and documented them along with the relevant information about the relation in question, the document number (the file to which the relation belongs), the status of the spans (nucleus or satellite), and the span numbers (the location of the spans in the text). We annotated the signalling information in a separate Excel file, since RSTTool does not allow for multiple levels of annotation.

---

[2] The taxonomy used in the RST-DT comprises a set of 16 major relation groups which are further divided into 78 RST relations.

As an example of our annotation task, we provide the signalling annotation of a short RST file. The file is taken from the RST-DT, and its relational structure is provided in Figure 1. A detailed description of our annotation for the file is provided in Table 1.

Nuc	Sat	Relation	Signal class	Signal type	Specific signal	Explanation on signalling
1	2-5	Elaboration-additional	single	genre	textual organization	In newspaper reports, the content of the first paragraph (or the first few paragraphs) is elaborated on in the following paragraphs.
1	2-5	Elaboration-additional	combined	entity + syntactic	given entity + subject NP	*Sun Microsystems Inc.*, mentioned in the first span, is the subject of the sentence which the second starts with.
1	2-5	Elaboration-additional	single	lexical	lexical overlap	The words *debentures* and *convertible* are present in both spans.
3/4		List	single	DM	*and*	DM *and* signals the List relation.
3-4	2	Attribution	single	syntactic	reported speech pattern	The reported speech pattern "The company said…" signals the Attribution relation.
2-4/5		List	combined	entity + syntactic	given entity + subject NP	*The debenture* in the first span is the subject of the sentence which the second span starts with.
2-4/5		List	single	semantic	lexical chain	The words *issued* and *available* in the respective spans are semantically related.

Table 1: Annotation of an RST file with relevant signalling information

According to the annotation provided in Table 1, the Elaboration relation between spans 1 and 2-5 is indicated by three types of signals: (i) *genre*; (ii) *entity + syntactic*; and (iii) *lexical features*. First, the text represents the newspaper genre (since it is taken from a Wall Street Journal article), and in newspaper texts the content of the first (or the first few) paragraphs is typically elaborated on in the subsequent paragraphs. In this particular example, the entire first paragraph functions as the nucleus of the Elaboration relation, with the two following paragraphs being its satellite. Thus, we postulate that the Elaboration relation is conveyed by the genre feature (more specifically by a feature which we call *textual organization*). Second, we postulate that a combined signal *entity + syntactic* (specifically, *given entity + subject NP*), made of two individual features, is operative here in signalling the Elaboration relation. One can notice that the entity *Sun Microsystems Inc.*, mentioned in the nucleus, is elaborated on in the satellite. Syntactically, the entity is also used as the subject NP of the sentence the satellite starts with, representing the topic of the Elaboration relation. Finally, the Elaboration relation is also (perhaps rather loosely)

signalled by a *lexical feature*, or *lexical overlap*. Words such as *debentures* and *convertible* occur in both the nucleus and satellite, indicating the presence of the same topic in both spans, with an elaboration in the second span of some topic introduced in the first span.

The List relation between spans 3 and 4 is conveyed in a straightforward (albeit underspecified) way by the use of the DM *and*.

The Attribution relation between spans 2 and 3-4 is indicated by a *syntactic* signal, a *reported speech pattern* in which the reporting clause (span 2) functions as the satellite and the reported clause (span 3-4) functions as the nucleus. The key is the S+V (Subject + Verb) combination with a reported speech verb (*said*).

Finally, the List relation between spans 2-4 and 5 is indicated by two types of signals: (i) *entity + syntactic* and (ii) *semantic feature*. For the combined feature *entity + syntactic*, the specific signal is *given entity + subject NP*, according to which the entity, *the debentures*, mentioned in the first span is used as the subject NP of the sentence the second span starts with. For the *semantic feature*, the specific signal is a *lexical chain* which means that semantically similar or related words occur in the respective text spans. We notice that words such as *issued* and *available* are semantically related, and they are used in both spans, indicating a List relation holding between them.

4. Results

Among the 1,306 relations examined, the distribution of signalled relations (indicated either by DMs or by some other signal) and unsignalled relations (not indicated by any signal) is provided in Table 2.

Relation type	**Tokens**	**Percentage**
Signalled relations	1,129	86.45%
Unsignalled relations	177	13.55%
Total	1,306	
Relations indicated by a DM	251	22.23%
Relations indicated by other signals	878	77.77%
Total	1,129	

Table 2: Distribution of signalled and unsignalled relations

The results show that 1,129 relations (86.45%) out of all the 1,306 relations are signalled, either by a DM or with the help of some other signalling device. On the other hand, no significant signals are found for the remaining 177 relations (13.55%).

Among the 1,129 signalled relations, we find that DMs are used to signal 251 relations (22.23% of the signalled relations), while 878 relations (77.77% of the signalled relations) are indicated with the help of some other signals.

For the 251 instances of relations signalled by a DM, we 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*. For a full list of these extracted markers, see Taboada and Das (2013).

For the 878 signalled relations without DMs, we have found that a wide variety of signals are used to indicate them. In our corpus analysis, 81.67% of the signalled relations (922 out of 1,129 signalled relations) are exclusively indicated by a single signal (including DMs), whereas 5.67% of the signalled relations (64 out of 1,129 signalled relations) are indicated by a combined signal. In addition, the distribution also shows that 12.49% of the signalled relations (141 out of 1,127 signalled relations) contain multiple signals[3].

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

	Relation group	Relation	# Relations with DMs	# Relations with other markers	# 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

[3] Multiple signals refer to two or more types of signals (single or combined) which are separately used to indicate a particular relation instance, as shown in example (3).

8	Enablement	Purpose	0	39	0	39
8	Enablement	Enablement	0	0	0	0
9	Evaluation	Evaluation	1	3	1	5
9	Evaluation	Interpretation	1	0	9	10
9	Evaluation	Conclusion	0	0	0	0
9	Evaluation	Comment	0	0	9	9
10	Explanation	Evidence	0	3	8	11
10	Explanation	Explanation-argumentative	6	1	23	30
10	Explanation	Reason	12	1	4	17
11	Joint	List	50	27	6	83
11	Joint	Disjunction	3	0	0	3
12	Manner-Means	Manner	3	0	0	3
12	Manner-Means	Means	1	4	0	5
13	Topic-Comment	Problem-solution	2	2	2	6
13	Topic-Comment	Question-answer	0	0	0	0
13	Topic-Comment	Statement-response	0	2	0	2
13	Topic-Comment	Topic-comment	1	0	0	1
13	Topic-Comment	Comment-topic	0	0	0	0
13	Topic-Comment	Rhetorical-question	0	0	0	0
14	Summary	Summary	0	0	8	8
14	Summary	Restatement	0	9	0	9
15	Temporal	Temporal-before	3	0	0	3
15	Temporal	Temporal-after	7	1	0	8
15	Temporal	Temporal-same-time	3	1	0	4
15	Temporal	Sequence	5	0	0	5
15	Temporal	Inverted-sequence	0	0	0	0
16	Topic-change	Topic-shift	0	0	4	4
16	Topic-change	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 (19.22%)	878 (67.23%)	177 (13.55%)	1,306

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

The distribution of relations in Table 3 shows that almost every group of relations is more or less signalled. In particular, we find that relation groups such as Attribution, Elaboration, Enablement, and Joint are most frequently signalled, either by DMs or by some other signals. We also found that that there is only one group of relation, Evaluation, which is rarely indicated by any signal.

Among the signalled relations, DMs are most frequently used to signal relations such as Circumstance, Result, Consequence, Condition, Concession, Contrast, Antithesis, Reason and List. In contrast, 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. Our findings are also parallel to the results presented in our earlier work (Taboada 2006), where we found that relations such as Concession, Condition and Purpose are most frequently signalled (by a DM), while Background and Summary are rarely signalled (by a DM).

Relations which are mostly indicated by other signals include Attribution, Elaboration-additional, Elaboration-general-specific, Elaboration-object-attribute, Purpose and Restatement. In contrast, relations which are rarely or never indicated by other signals include Circumstance, Consequence, Condition, Contrast, Antithesis, Explanation-argumentative and Temporal-after.

Finally, the relations for which no signals (neither a DM nor any other signal) are found include Comment, Summary and Topic-change.

5. Discussion

The first goal of our study was to investigate whether signals other than DMs exist for coherence relations. In this respect, we can confidently say that this is, indeed, the case: Out of the 1,129 signalled relations examined, 878 (77.77%) of the relations contain a signal other than a DM. Furthermore, the signals of coherence relations are diverse in nature, and can be broadly classified in major groups, such as *DM*, *reference*, *lexical*, *semantic*, *syntactic*, *graphical* and *genre features*. The individual signal groups also contain different specific signals in themselves. For example, *syntactic feature* includes specific signals such as *relative clause*, *participial clause* and *parallel syntactic construction*.

We would like to point out that what we have found are *positive* signals, that is, indicators that a relation exists. This does not mean that such signals are used exclusively to indicate that relation (as we have seen in the many-to-many correspondences). It also means that the signals, as linguistic devices, are not exclusively used to mark a relation; they may well have other purposes in the text. In a sense, this means that the signals are compatible with a relation, not necessarily indicators of the relation exclusively.

The other objective of our study was to evaluate the validity of the traditional dichotomy of explicit and implicit relations. Traditionally, explicit relations are signalled by DMs while implicit relations are not signalled by them. Our results show that relations can be signalled by DMs as well as by other signals. In addition, signals of relations, regardless of their types, have an explicit presence in discourse since they are all textual in nature. This implies that the category of explicit relations should not comprise only those relations indicated by DMs, but also those indicated by other signals. Implicit relations, on the other hand, can be characterized in terms of the absence of any signal.

Finally, although we discovered signalling evidence for the majority of the relations in the corpus, we also found that some of the relations (13.55% of the total 1,306) are not signalled. As for the 177 relations for which we could not identify a signal, there are three different reasons why we believe that is the

case. First of all, in some cases we found that there were errors in the existing annotation of relations in the RST-DT, and a relation was postulated, whereas we would not have annotated a relation. In those cases, the lack of signalling is perfectly understandable. Secondly, some of the RST-DT relations are not true RST relations. Relations such as Comment or Topic-shift, in our opinion, belong in the realm of discourse organization, not together with relations among propositions. Again, finding no signals in those cases is not surprising, as such phenomena are not likely to be indicated by the same type of signals as coherence relations proper. Finally, in many cases, one or both of the annotators had a sense that the relation was clear, but could not pinpoint the specific signal used. This is the case with tenuous entity relations, or relations that rely on world knowledge.

6. Conclusion

The purpose of the study was to determine to what extent coherence relations carry signals that may help readers and hearers identify the relation. Research so far has focused mainly on one type of signals, DMs, and has thus concluded that the majority of relations are implicit, that is, they contain no overt signal. However, in our study we found out that DMs are not the only type of signals of coherence relations as relations in discourse can well be indicated by a variety of signals other than DMs. We also found that, although there may still exist some implicit relations, the majority of the relations present in a discourse are signalled or explicit. These finding reinforce the psycholinguistic claim that there exist signals for every interpretable relation (Taboada 2009). They also suggest that relation signalling is much more sophisticated than previously thought as relations are conveyed through different types of single, combined and multiple signals.

The annotation described in this paper is a preliminary pilot study, comprising only 10% of the total corpus. In future work, we will expand to cover the entire corpus. The most important qualitative change for the rest of the annotation involves finding a method to layer annotations on top of the existing LISP-style notation for the RST-DT. Although we have not formalized a plan for that, the most likely avenue will be to covert the LISP format to XML, and encode the signalling information as XML.

The finished corpus has two clear applications. From a psycholinguistic point of view, we hope to be able to use it to determine how hearers and readers use signals to identify relations. Most of the psycholinguistic studies to date have manipulated relations by adding or deleting DMs. It would be very useful to extend that work by changing other types of signals, to see what effects that has on comprehension.

The other main application of such an annotated corpus is in discourse parsing. A great deal of recent work (da Cunha *et al.* 2012; Hernault *et al.* 2011; Hernault *et al.* 2010; Mithun and Kosseim 2011) and also earlier approaches (Corston-Oliver 1998; Marcu 2000; Schilder 2002) have used DMs as the main signals to automatically parse relations, and almost exclusively at the sentence level. Our extended set of signals, and the fact that they work at all levels of discourse, will probably facilitate this task.

References

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

Blakemore, Diane. 1987. *Semantic Constraints on Relevance*. Oxford: Blackwell.

Blakemore, Diane. 1992. *Understanding Utterances: An Introduction to Pragmatics*. Oxford: Blackwell.

Blakemore, Diane. 2002. *Relevance and Linguistic Meaning : Th e Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

Britton, B. K., S. M. Glynn, B. J. F. Meyer & M. J. Penland. 1982. Effects of text structure on the use of cognitive capacity during reading. *Journal of Educational Psychology* 74. 51-61.

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski. 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.* Proceedings of AAAI 1998 Spring Symposium Series, Intelligent Text Summarization (pp. 9-15). Madison, Wisconsin.

da Cunha, Iria, Eric San Juan, Juan Manuel Torres-Moreno, María Teresa Cabré & Gerardo Sierra. 2012. *A symbolic approach for automatic detection of nuclearity and rhetorical relations among intra-sentence discourse segments in Spanish.* Proceedings of CICLing (pp. 462-474). New Delhi, India.

Dale, Robert. 1991a. *Exploring the Role of Punctuation in the Signalling of Discourse Structure.* Proceedings of Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI (pp. 110-120). Technical University of Berlin.

Dale, Robert. 1991b. *The role of punctuation in discourse structure.* Proceedings of AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation (pp. 13-14). Asilomar, CA.

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

Fraser, Bruce. 1990. An approach to discourse markers. *Journal of Pragmatics* 14. 383-395.

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

Fraser, Bruce. 2006. Towards a theory of discourse markers. In Kerstin Fischer (ed.), *Approaches to Discourse Particles* (pp. 189-204). Amsterdam: Elsevier.

Fraser, Bruce. 2009. An account of discourse markers. *International Review of Pragmatics* 1. 293-320.

Haberlandt, K. 1982. Reader expectations in text comprehension. In J.-F. Le Ny & W. Kintsch (eds.), *Language and Comprehension* (pp. 239-249). Amsterdam: North-Holland.

Halliday, Michael & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

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

Hernault, Hugo, Helmut Prendinger, David A. duVerle & Mitsuru Ishizuka. 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.* Edinburgh, UK: University of Edinburgh. Ph.D. dissertation.

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

Lapata, Mirella & Alex Lascarides. 2004. *Inferring sentence-internal temporal relations.* Proceedings of NAACL-04 (pp. 153–160).

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, Min-Yen Kan & Hwee Tou Ng. 2009. *Recognizing implicit discourse relations in the Penn Discourse Treebank.* Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing. Singapore.

Loman, N. L. & R. E. Mayer. 1983. Signaling techniques that increase the understandability of expository prose. *Journal of Educational Psychology* 75. 402-412.

Lorch, R.F. & E.P. Lorch. 1986. On-line processing of summary and importance signals in reading. *Discourse Processes* 9. 489–496.

Louis, Annie, Aravind Joshi, Rashmi Prasad & Ani Nenkova. 2010. *Using Entity Features to Classify Implicit Discourse Relations.* Proceedings of SIGDIAL 2010 (pp. 59–62).

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3). 243-281.

Marcu, Daniel. 1999. *A decision-based approach to rhetorical parsing.* Proceedings of 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 365-372). College Park, Maryland.

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

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

Meyer, B. J. F. 1975. *The organization of prose and its effects on memory*. Amsterdam: North-Holland.

Meyer, B. J. F., D. M. Brandt & G. J. Bluth. 1980. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly* 16. 72-103.

Millis, K. K., A.C. Graesser & K. Haberlandt. 1993. The impact of connectives on the memory for expository texts. *Applied Cognitive Psychology* 7. 317–339.

Millis, K. K. & M. A. Just. 1994. The influence of connectives on sentence comprehension. *Journal of Memory and Language* 33. 128-147.

Mithun, Shamima & Leila Kosseim. 2011. *Comparing approaches to tag discourse relations.* Proceedings of 12th international conference on Computational linguistics and intelligent text processing (CICLing '11) (pp. 328-339). Tokyo, Japan.

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

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

Pitler, Emily, Annie Louis & Ani Nenkova. 2009. *Automatic sense prediction for implicit discourse relations in text.* Proceedings of 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, Chris Culy, Martin van den Berg, Gian Lorenzo Thione & David Ahn. 2004. *A rule based approach to discourse parsing.* Proceedings of SigDIAL 2004. Cambridge, MA.

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

Prasad, Rashmi, Aravind Joshi & Bonnie Webber. 2010. *Realization of Discourse Relations by Other Means: Alternative Lexicalizations.* Proceedings of COLING 2010 (pp. 1023-1031). Beijing.

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

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

Sanders, Ted & Leo Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes* 29(1). 37-60.

Sanders, Ted, Wilbert Spooren & Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15. 1–35.

Sanders, Ted, Wilbert Spooren & Leo Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics* 4(2). 93-133.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schiffrin, Deborah. 2001. Discourse markers: Language, meaning and context. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The Handbook of Discourse Analysis* (pp. 54-75). Malden, MA: Blackwell.

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

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

Sporleder, Caroline & Alex Lascarides. 2005. *Exploiting linguistic cues to classify rhetorical relations.* Proceedings of Recent Advances in Natural Language Processing (RANLP-05).

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

Spyridakis, J. H. & T. C. Standal. 1987. Signals in expository prose: Effects on reading comprehension. *Reading Research Quarterly* 12. 285-298.

Subba, Rajen & Barbara Di Eugenio. 2009. *An effective discourse parser that uses rich linguistic information.* Proceedings of HLT-ACL 2009 (pp. 566-574). 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 & Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse* 4(2). 249-281.

Taboada, Maite & William C. Mann. 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.* Radboud University Nijmegen, The Netherlands. MA.