Signalling of coherence relations in discourse, beyond discourse markers*

Debopam Das and Maite Taboada

University of Potsdam, Simon Fraser University

debdas@uni-potsdam.de, mtaboada@sfu.ca

Abstract

We argue that coherence relations (relations between propositions, such as *Concession* or *Purpose*) are signalled more frequently and by more means than is generally believed. We examine how coherence relations in text are indicated by all possible textual signals, and whether every relation is signalled. To that end, we conducted a corpus study on the RST Discourse Treebank (Carlson et al., 2002), a corpus of newspaper articles annotated for rhetorical (or coherence) relations. Results from our corpus study show that the majority of relations in text (over 90%) are signalled, and also that the majority of signalled relations (over 80%) are indicated not only by discourse markers (*and, but, if, since*), but also by a wide variety of signals other than discourse markers, such as *reference, lexical, semantic, syntactic* and *graphical* features. These findings suggest that signalling of coherence relations is much more sophisticated than previously thought.

Keywords: coherence relations, Rhetorical Structure Theory, signalling, discourse markers, RST Discourse Treebank, RST Signalling Corpus

* Das, D. and M. Taboada (to appear) Signalling of coherence relations in discourse. *Discourse Processes*. Current version: September 2017.

1. Introduction

One of the ways to achieve coherence in discourse is through establishing meaningful links between discourse components. Coherence relations define and characterize the nature of relationships between discourse components, and thus contribute to creating and interpreting the discourse structure of a text. Consider Example (1)[1], which consists of two units of discourse, the two sentences. These units are connected to each other by an *Evidence* relation: The claim that consumers change their brand loyalty as a result of a greater number of choices available to them is evidenced by the majority of car-buyers' tendency to switch brand, as reported by the Wall Street Journal's "American Way of Buying" survey.

(1) When consumers have so many choices, brand loyalty is much harder to maintain. The Wall Street Journal's "American Way of Buying" survey found that 53% of today's car buyers tend to switch brand. (wsj_1377)

One of the most important questions in discourse analysis is how readers or hearers identify the presence and type of coherence relations. Coherence relations are often signalled by discourse markers or DMs, such as *because* indicating a causal coherence relation, or *if* a condition. In many instances, however, as with Example (1), no discourse marker is present. We are interested in the general signalling of relations, by discourse markers or by other means. We explore signals beyond discourse markers for two reasons: (1) The majority of relations in a text do not contain a DM; and (2) signalling by certain DMs can be underspecified, since the same DM can be used to indicate different types of coherence relations (e.g., the DM *and* as a signal for *Elaboration, List* and *Consequence* relations).

In this study, we investigate how coherence relations are signalled in discourse and what signals are used to indicate them. A secondary goal is to examine whether coherence relations are more frequently

[1] Most of the examples in the paper are from the RST Discourse Treebank (Carlson et al., 2002). The text in parentheses at the end refers to the file number in the RST Discourse Treebank from which the example has been taken. If no file number is mentioned, then the example is invented.

explicit or implicit in terms of the type of signalling involved. By *signalling* we mean the cues that indicate that a coherence relation is present, such as the conjunction *because* as a signal for a causal relation. We use the term *signalling* rather than *marking* because the latter has been associated with discourse markers or DMs, which we believe are only one type of many possible signalling devices.

We undertake a large scale annotation project in which we select an existing corpus of coherence relations called the RST Discourse Treebank (Carlson et al., 2002), and add to those relations in the corpus relevant signalling information. The final product of this annotation project is a newly-annotated discourse corpus, known as the RST Signalling Corpus (Das et al., 2015), which provides annotation not only for DMs, but also for many other textual signals such as syntactic, semantic, lexical or graphical features. More information about the annotation project can be found in Das and Taboada (2017).

The paper is organized as follows: In Section 1, we provide an introduction to the concept of coherence relations and explain how coherence relations are treated in Rhetorical Structure Theory, chosen as the theoretical framework of the study. Section 2 presents a short account of the existing research on signalling in discourse, focusing on the psychological processing of coherence relations in the presence as well as absence of DMs. In Section 3, we describe the corpus study, the annotation scheme and annotation procedure. In Section 4, we present the results, including the statistical distributions of relations and signals in the corpus. Finally, Section 5 discusses the significance of those results, summarizes the study and provides the conclusion.

1 Coherence relations and RST

A discourse is characterized by the connectedness among its different parts. This connectedness is often explained by linguists in terms of two concepts: *cohesion* and *coherence* (Halliday & Hasan, 1976; Hasan, 1985; Hobbs, 1979; Kintsch & van Dijk, 1978; Poesio et al., 2004). Cohesion refers to the grammatical and lexical connections that link one element (typically, an entity) of a discourse to another. Coherence, on the other hand, is defined as a semantic or pragmatic relationship that links one informational unit in a discourse to another unit or to a group of units. For example, consider the following text.

(2) Chris is a fan of Steven Spielberg. She has seen all his movies.

In this example, *she* refers to *Chris* while *his* refers to *Steven Spielberg*, and hence these expressions are associated by cohesion. On the other hand, the interpretation that Chris' fondness for Steven Spielberg's movies is evidenced by the fact that she has seen all of Spielberg's movies is an example of coherence. Building on the notion of coherence, coherence relations are defined in terms of how two (or more) discourse segments are connected to each other in a meaningful way. They specify the semantic or pragmatic types of relationships that hold between two or more discourse components.

Coherence relations are known by different names such as discourse relations or rhetorical relations, and have been extensively studied in discourse theories such as Rhetorical Structure Theory or RST (Mann & Thompson, 1988), Segmented Discourse Representation Theory or SDRT (Asher & Lascarides, 2003; Lascarides & Asher, 2007), the cognitive approach to coherence relations (Sanders et al., 1992), the Unified Linguistic Discourse Model (Polanyi et al., 2004), or Hobbs' theory (Hobbs, 1985), further expanded by Kehler (2002). Despite the apparent dissimilarities involving these labels and among these different discourse frameworks, we believe that all theories refer to fundamentally the same phenomenon: relations among propositions, which are the building blocks of discourse and which help explain coherence. Although we have worked within RST (Mann & Thompson, 1988), and will use some of its constructs here, the discussion that follows likely applies to most views of coherence relations.

Text organization in Rhetorical Structure Theory (RST henceforth)[2] is primarily described in terms of relations that hold between two (or sometimes more) non-overlapping text spans. 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. Relation

[2] For more information on RST, see Mann and Thompson (1988), Taboada and Mann (2006), and the RST website: http://www.sfu.ca/rst/

inventories are open, but the most common ones include names such as *Cause, Concession, Condition, Elaboration, Result* or *Summary*.

Relations in RST are defined in terms of four fields: (1) constraints on the nucleus; (2) constraints on the satellite; (3) constraints on the combination of nucleus and satellite; and (4) effect (on the reader). The *locus* of the effect, derived from the *effect* field, is identified as either the nucleus alone or the nucleus-satellite combination. An analyst builds the RST structure of a text based on the particular judgements that are specified by these four fields.

Texts, according to RST, are built out of basic clausal units that enter into rhetorical (or discourse, or coherence) relations with each other in a recursive manner. Mann and Thompson (1988) 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.

For illustration purposes, we provide the annotation of a short text taken from the RST Discourse Treebank (Carlson et al., 2002).

(3) 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. (wsj_650)

The graphical representation of the RST analysis of this text using the RSTTool (O'Donnell, 1997) is provided in Figure 1.

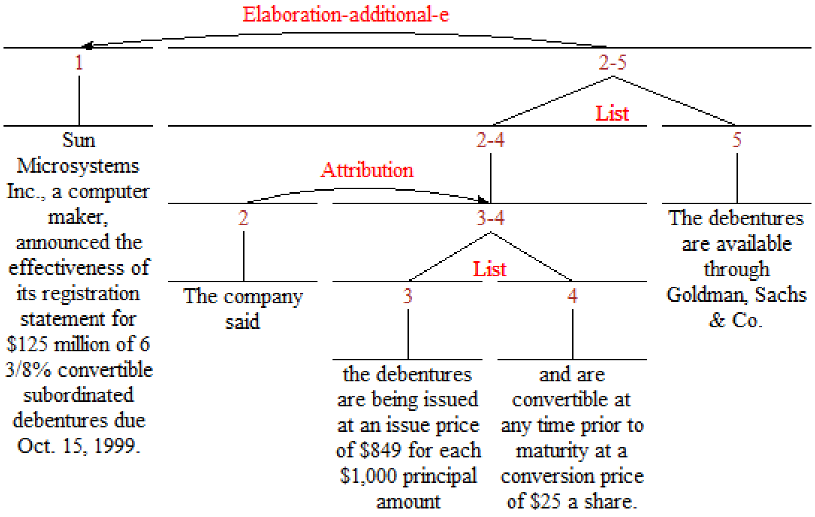
Elaboration-additional-e
1
2-5
List
2-4
5
Attribution
2
3-4
List
3
4
Sun Microsystems Inc., a computer maker, announced the effectiveness of its registration statement for $125 million of 6 3/8% convertible subordinated debentures due Oct. 15, 1999.
The company said
the debentures are being issued at 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.

Figure 1: Graphical representation of an RST analysis

The RST analysis shows that the text can be segmented into five elementary units (spans) which are represented in the diagram by the numbers, 1, 2, 3, 4 and 5, respectively, with horizontal lines above each unit. Elementary units may combine to form spans of more than one unit. Straight vertical lines above a span (whether elementary or complex) mean that it is a nucleus. Lines with arrowheads are used to indicate how a satellite connects to its nucleus, with the arrowhead pointing away from the satellite to the nucleus.

In the diagram, we can see that Span 3 (as a nucleus) and span 4 (another nucleus) are connected to each other by a multinuclear *List* relation, and together they make the combined span 3-4. Span 2 (satellite) is connected to span 3-4 (a nucleus) by an *Attribution* relation, and together they make the combined span 2-4. Then, a *List* relation holds between spans 2-4 (nucleus) and 5 (nucleus), and together they make the combined span 2-5. This relation has two straight lines joining 2-4 and 5, indicating that they are both nuclei. This is a type of coordinating relation, as opposed to a nucleus-satellite relation, which is subordinating. Finally, span 2-5 (as a satellite) is connected to span 1 (a nucleus) by an *Elaboration* relation (more specifically, *Elaboration-addition-e*).

One of the most active and lively debates in RST and other discourse theories has centered around how coherence relations are recognized and interpreted, that is, their cognitive status: Are relations present in the minds of speakers and hearers[3], or are they analysis constructs? The former postulates that coherence relations are part of the process of constructing a coherent text representation. In RST, the relations are presented as being recognizable to an analyst, and in general to a reader. The process is one of uncovering the author's intention in presenting pieces of text in a particular order and combination. In carrying out an RST analysis of a text, "the analyst effectively provides plausible reasons for why the writer might have included each part of the entire text" (Mann & Thompson, 1988: 246). But further cognitive claims have not been strong within RST.

Support for the cognitive status of coherence relations comes from experimental work on the effect of particular types of relations on text comprehension. Sanders and colleagues have best articulated this view. Knott and Sanders (1998) argue that text processing consists of building a representation of the information contained in the text. Part of the process of building involves integrating individual propositions in the text into a whole. Coherence relations model the ways in which propositions are integrated. The evidence presented comes from studies on the recognition of different types of relations, whether as a binary classification, causal versus non-causal (Keenan et al., 1984; Myers et al., 1987; Sanders & Noordman, 2000; Trabasso & Sperry, 1985), or as a more specific type of distinction, such as the difference between *Problem-Solution* and *List* (Sanders & Noordman, 2000). It seems clear that coherence relations are different in nature among themselves. The second source of evidence on the cognitive status of relations is from studies on how the presence of DMs or connectives tends to facilitate text processing (Gaddy et al., 2001; Haberlandt, 1982; Sanders et al., 2007; Sanders & Noordman, 2000; Sanders et al., 1992). If coherence relations were not cognitive entities, then there should not be any effect in indicating their presence. The conclusion is, then, that processing coherence relations is part of understanding text.

This line of research has explored the identification and classification of coherence relations through DMs (or connectives). The problem with such an approach is that it does not address the issue of relations

[3] We will use speakers/hearers and writers/readers interchangeably. It is arguably the case that most of what can be said about coherence relations applies equally to spoken and written discourse. Indeed, if we postulate the psychological validity for coherence relations, both forms of discourse must be accounted for.

which appear to be unsignalled, because no DM is present. It is clear to most researchers that one can postulate relations (and presumably, readers understand them) even when they are not signalled by a DM. If all relations are of the same type, that is, if all relations are cognitive entities, then signalling through DMs only facilitates their comprehension. Lack of signalling does not mean that no relation is present.

2 Signalling of coherence relations

From the viewpoint of signalling, coherence relations are divided into two groups: signalled and unsignalled relations. The distinction is also represented by other labels such as explicit and implicit relations or marked and unmarked relations, and has widely been discussed in the discourse literature (Knott & Dale, 1994; Martin, 1992; Meyer & Webber, 2013; Renkema, 2004; Taboada, 2009; Taboada & Mann, 2006; van der Vliet & Redeker, 2014; Versley, 2013). Traditionally, the distinguishing criterion for such a classification has always been the presence or absence of DMs which are considered to be the most typical (sometimes the only type of) signals of coherence relations. DMs are lexical expressions (*and, because, since, thus*, etc.) which belong to different syntactic classes, such as conjunctions, conjunctive adverbs, adverbial and prepositional phrases (see Redeker (1990), Fischer (2006) and Fraser (2009) for definitions and classifications). They have received a variety of names, including connectives, discourse cues or discourse relational devices, but we will use the very general 'discourse marker'. DMs are used to connect discourse components, and they help readers understand the coherence relations that hold between those components[4]. Consider the following examples:

(4) Pat quit his job because he was tired of the long hours.

(5) Pat quit his job. He was tired of the long hours.

In Example (4), the discourse components (two propositions represented by the two clauses) are connected by a *Reason* relation. Since the relation is specified by the DM *because*, the relation is signalled

[4] In spoken discourse, DMs (such as *so* and *well*) also have a topic-organizing function, and can be used indicate a change of topic or a new discourse move (Schiffrin, 1987). Sometimes, DMs in conversation (such as *y'know*) signals the speaker's attitude to the content of interaction, and primarily serves an interpersonal function rather than an ideational one (Georgakopoulou & Goutsos, 2004).

(or explicit, or marked). On the other hand, the *Reason* relation in Example (5) does not contain a DM, and hence, is considered to be unsignalled (or implicit, or unmarked). One interesting aspect of signalling is that for unsignalled relation the implicature (the meaning inferred from or suggested by an utterance) can be cancelled with the insertion of an appropriate DM, as shown in Example (6).

(6) Pat quit his job. He was tired of long hours, anyway.

Although DMs are considered to be the most useful signals of coherence relations, studies on signalling show that the majority of relations occur in a text without DMs (Das, 2014). Taboada (2009) notes that over 50% of the relations in different types of text are not signalled by DMs. For instance, in the largest available discourse-annotated corpus, the Penn Discourse Treebank (Prasad et al., 2008), 54.37% of the relations are not signalled by DMs (Prasad et al., 2007).

The issue of unsignalled relations or the fact that relations without DMs are omnipresent in discourse can be approached from a Gricean point of view using the Cooperative Principle, particularly the Quantity maxim (Grice, 1975). Grice formulates the Cooperative Principle as: “Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.” (Grice, 1975: 45). In the Quantity maxim, he specifies that speakers (or writers) should make their contribution as informative as is required, and not more. If we believe, as Spooren (1997) suggests, that underspecified or unsignalled relations obey the Cooperative Principle and the Quantity maxim (“say no more than necessary”)[5], then unsignalled relations are such because no signal is necessary. The task of a writer or speaker, then, is one of determining how much signalling is enough. A writer may decide that no connective is necessary because signals or other cues that suffice to identify the relation are present, thus obeying the Quantity maxim. Unsignalled relations may be more difficult to process, but, under the Cooperative Principle, not impossible.

Research on text processing shows that connectedness in discourse is a mental phenomenon, and language users, when interpreting a text, make a coherent representation of the information from that text

[5] Spooren actually makes reference to Horn’s (1984) take on the Cooperative Principle, which can be summarized as “say no more than necessary”.

(Sanders & Spooren, 2007). This representation is aided by connecting different parts of the text with appropriate coherence relations. Cognitive linguists often hypothesize that if coherence relations play a significant role in establishing the mental representation of a coherent discourse, then the linguistic signals of coherence relations must have some influence on the reading process, and also on the mental representation that a reader achieves after reading. DMs, in psycholinguistic studies, are considered to be the processing instructions which guide the readers to recognize the coherence relations that hold between text segments. 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 of 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 (for references, see Das (2014)). Haberlandt (1982), for instance, found that sentences which include causal or concessive connectives are processed faster than sentences without connectives. Sanders et al. (2007) showed that explicitly marked relations led to better performance in text comprehension questions, both in laboratory and realistic situations.

The effects of signalling on recall and some aspects of comprehension have been more mixed. Meyer et al. (1980) found no positive effect on recalling content[6]. They did, however, find that subjects recalled the structure of the original text more faithfully when it was signalled. Millis and Just (1994) saw an increase in processing time, but observed more accurate answers to comprehension questions when a connective was present. Sanders and Noordman (2000) found that connectives had a positive effect on processing, but no noticeable effect on recall. Sanders and Noordman's conclusion about the recall effect is that the effect of the marker decreases over time, just as the surface representation of the text is lost, but the semantic content is preserved longer. Degand and Sanders (2002) report better answers on comprehension questions if the texts include a relational marker.

[6] Meyer et al.'s (1980) signalling included explicit statements of the structure of the text and connectives. As noted later on in this section, the results were different for different types of students (poor vs. good readers).

Studies have indeed shown that the effect of signalling is different for different types of readers. Meyer et al. (1980) discovered that explicit connectives helped only underachieving students, those readers that need signalling to identify the top-level structure of a text. Kamalski et al. (2008)[7] examined the impact of DMs on the understanding of informative and persuasive texts by high knowledge readers (with prior knowledge) and low knowledge readers (without prior knowledge). Results showed that, while the low knowledge readers had a better understanding of the explicit text (with DMs), the high knowledge readers had a better understanding of the implicit text (without DMs). On the other hand, the presence of the DMs in the persuasive genre proved to be beneficial for comprehension for both types of readers.

A significant issue in the psycholinguistic research involving the manipulation of DMs concerns the naturalness of the texts to be used as test materials. In most psycholinguistic studies, a set of two alternative text versions is used, the first being characterized by the presence of naturally occurring DMs, and the second being created from the first version by removing the DMs. The question is how well a relation holds after a DM which occurs naturally in a text is removed. Sporleder and Lascarides (2008) suggest (in a computational experiment context) that marked and unmarked texts might be linguistically very dissimilar, and removing unambiguous markers might result in a change of meaning in the original text. In other words, the contexts containing a DM could be very different from the contexts without a DM. This can be shown by the following examples (also used previously).

(7) Pat quit his job because he was tired of the long hours.

(8) Pat quit his job. He was tired of long hours, anyway.

While removing the DM *because* in Example (7) does not affect the reason relation between the discourse segments, the removal of *anyway* in Example (8) results in a strong causal connection that was previously not available.

[7] Kamalski et al.'s study was a replication of McNamara and Kintsch's (1996) study which investigated the effects of prior knowledge on learning of high-and-low-coherence history texts. Results showed that readers with prior knowledge were more successful in answering the open-ended questions after reading the low-coherence text. Also, the reading time experiments showed that the low-coherence text required more inference processes for all readers.

If we restrict the scope of signalling exclusively to the use of DMs, then the most vital question is whether relations are correctly interpreted in the absence of signalling. Theoretically, there can be two possible answers to this question. First, if it is only DMs which entail or justify the presence of coherence relations, then the lack of signalling (by DMs) results in the absence of relations. In other words, if there are no signals, then there are no relations. Second, signalling of relations can be achieved through the use of signals other than DMs. Thus, 'no signalling' means the absence of DMs, but most importantly it implies signalling by other signals which may actively facilitate the understanding of coherence relations and hence, the comprehension process, as well.

The issue of signalling of coherence relations has been dealt by large more successfully in computational linguistics. With the common goal of automatically identifying and characterizing coherence relations in unseen texts, most computational studies used DMs and similar cue phrases as the primary signals of coherence relations (Feng & Hirst, 2012; Forbes et al., 2001; Hernault et al., 2010; Le Thanh, 2007; Marcu, 2000; Schilder, 2002; Subba & Eugenio, 2009). However, most importantly, a lot of those studies also investigated the signalling of coherence relations beyond DMs by looking at other linguistic or textual features. Some of these features exploited in these studies include tense or mood (Scott & de Souza, 1990), anaphora and deixis (Corston-Oliver, 1998), lexical chains (Marcu, 2000), punctuation and graphical markers (Dale, 1991a, 1991b), textual layout (Bateman et al., 2001), NP and VP cues (Le Thanh, 2007), reference and discourse features (Theijssen, 2007; Theijssen et al., 2008), specific genre-related features (Maziero et al., 2011; Pardo & Nunes, 2008), collocations (Berzlánovich & Redeker, 2012), polarity, modality and word-pairs (Pitler et al., 2009), coreference, givenness and lexical features (Louis et al., 2010), word co-occurrences (Marcu & Echihabi, 2002), noun and verb identity/class, argument structure (Lapata & Lascarides, 2004), or positional features, length features and part-of-speech features (Sporleder & Lascarides, 2005, 2008). For a summary of these, see Das (2014).

In our previous studies (Das, 2012; Das & Taboada, 2013a; Taboada & Das, 2013), we have shown that coherence relations can indeed be indicated by a wide variety of signals other than DMs. For example, a morphological marker such as *tense* is a good predictor of *Background* or *Temporal* relations; a syntactic marker such as a *parallel syntactic construction* can indicate a *Contrast* or *List* relation; a semantic relationships between words such as *synonymy* may signal *Elaboration* relations; a semantic feature such as *lexical overlap* in two discourse components can serve as a signal for *Summary* relations; and a graphical marker such as an *enumerated* or *itemized list* is present in some *List* relations. In the present study, we want to push that line of research further, as we attempt to explore every possible signal of coherence relations, and investigate their role in discourse organization.

3 Large-scale corpus study

We question the validity of the signalled/unsignalled classification based on the presence or absence of DMs, and re-examine the scope of signalling in discourse from a broader viewpoint. We illustrate how signalling works in the absence of DMs through the analysis of the following text.

(9) Chris is tall. Pat is short.

In this mini-text, the discourse components (two sentences) are connected to each other by a *Contrast* relation. Traditionally, this relation will be considered to be unsignalled (or implicit, or unmarked) since it does not contain a DM. However, we argue the relation is signalled by two types of other signals. One can notice that the two discourse components, the two sentences in the text, share a parallel syntactic construction (Subject – Copular Verb – Adjective). This syntactic feature is often used to indicate a *Contrast* relation. Furthermore, the relation is also signalled by the words *tall* and *short* in the respective sentences. These words are antonyms, and this particular meaning relationship is also a good indicator for *Contrast* relations.

The omnipresence of coherence relations without DMs in a discourse and their successful interpretation by readers or hearers raises one important question: How are coherence relations recognized in the absence of DMs? As discussed in the previous section, psycholinguistic research has shown that coherence relations are recognized (Kamalski, 2007; Knott & Sanders, 1998; Mak & Sanders, 2012; Mulder, 2008; Sanders & Noordman, 2000; Sanders & Spooren, 2007, 2009; Sanders et al., 1992, 1993).

This leads one to assume that if readers or hearers can understand a variety of relations, then there must be indicators which guide the interpretation process, beyond DMs.

Building on this assumption, we hypothesize that the signalling of coherence relations is achieved not only by DMs, but also through the use of a wide variety of textual signals beyond DMs. We refer to these signals as *other signals* in this paper and classify them into major types such as *lexical, semantic, syntactic, graphical* and *genre* features. In addition, we also hypothesize that every relation in a discourse is signalled (hence explicit), as a signal must be necessary for correct interpretation. In order to test these hypotheses, we conducted a corpus study.

3.1 Corpus

One of the research objectives in our study is to discover as many signals of coherence relations as possible. We chose to use the RST Discourse Treebank or RST-DT (Carlson et al., 2002) as our source of data, for two reasons. First, we wanted to work on a discourse annotated corpus whose theoretical foundation is similar to the theoretical framework that we have worked with in previous research. The RST-DT, as its name implies, is annotated for coherence relations based on RST. Second, we are interested in examining the signalling of relations at different levels of discourse. The RST-DT provides annotations not only for relations between elementary discourse units (usually clauses), but also for relations between larger chunks of texts (between sentences, groups of sentences, or even paragraphs). This is because RST follows a hierarchy principle in which a discourse sequence (the combined span comprising the nucleus and the satellite of a relation) can often function as a larger discourse segment, and can combine as a nucleus or a satellite with another discourse segment in order to form a global level relation (see Section 1 for the hierarchy principle in RST).

The RST-DT contains a collection of 385 Wall Street Journal articles (about 176,000 words of text) selected from the Penn Treebank (Marcus et al., 1993). The corpus is distributed by the Linguistic Data Consortium (LDC)[8], from which it can be downloaded (for a fee). The articles chosen for annotation in the

[8] https://www.ldc.upenn.edu/

RST-DT come from a variety of topics, such as financial reports, general interest stories, business-related news, cultural reviews, editorials and letters to the editor. The annotation process is aided by a modified version of RSTTool (O'Donnell, 1997) which provides a graphical representation of the RST analysis of a text in the form of a tree-diagram. For a description of the original annotation, see Das (2014) and Das and Taboada (2017).

The elementary discourse units in the RST-DT are considered to be clauses, with a few exceptions, as documented in the RST-DT annotation manual (Carlson & Marcu, 2001). The RST-DT employs a large set of 78 relations which are divided into 16 major relation groups. For example, the corpus includes a relation group called *Contrast* which comprises three individual relations: *Contrast*, *Concession* and *Antithesis*. The (concise) taxonomy of RST relations in the RST-DT can be found in Table 1.

#	**Relation Group**	**Relation**
1.	Attribution	Attribution, Attribution-negative
2.	Background	Background, Circumstance
3.	Cause	Cause, Result, Consequence
4.	Comparison	Comparison, Preference, Analogy, Proportion
5.	Condition	Condition, Hypothetical, Contingency, Otherwise
6.	Contrast	Contrast, Concession, Antithesis
7.	Elaboration	Elaboration-additional, Elaboration-general-specific, Elaboration-part-whole, Elaboration-process-step, Elaboration-object-attribute, Elaboration-set-member, Example, Definition
8.	Enablement	Purpose, Enablement
9.	Evaluation	Evaluation, Interpretation, Conclusion, Comment
10.	Explanation	Evidence, Explanation-argumentative, Reason
11.	Joint	List, Disjunction
12.	Manner-Means	Manner, Means
13.	Topic-Comment	Problem-solution, Question-answer, Statement-response, Topic-comment, Comment-topic, Rhetorical-question
14.	Summary	Summary, Restatement
15.	Temporal	Temporal-before, Temporal-after, Temporal-same-time, Sequence, Inverted-sequence
16.	Topic Change	Topic-shift, Topic-drift

Table 1: Taxonomy of RST relations in the RST-DT

Furthermore, three additional relations: *Textual-Organization*, *Span*[9] and *Same-Unit* were used in the annotation of the RST-DT in order to impose certain structure-specific requirements on the discourse trees.

[9] Among these three additional relations, *Span* was exclusively used for structural reasons, and not as a coherence relation proper, which connects two discourse segments. For this reason, *Span* was excluded from our signalling analyses.

More information on the taxonomy of relations and relation definitions can be found in the RST-DT annotation manual (Carlson & Marcu, 2001). The annotation was performed by a group of trained annotators, and the inter-annotator reliability reported by the corpus creators was quite reasonable. We do not, however, agree with every annotation decision, and such is the nature of annotation and corpus work. We chose to make use of an existing resource to build upon, as we believe we can provide better added value this way (see Taboada and Das (2013) for further discussion).

3.2 Taxonomy of signals

The first step in a signalling annotation task involves selection and classification of the types of signals which are to be annotated. We built our taxonomy of signals following two strategies. First, we manually built the repository of relational signals based on different classes of relational markers that have been mentioned in previous studies on the signalling in discourse (for references, see Das (2014)). Second, we extracted more markers by adding to the taxonomy signals identified in our preliminary corpus work (Das, 2012; Das & Taboada, 2013a, 2013b; Taboada & Das, 2013).

The signals in our taxonomy are 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 is defined; for example, the class *single* is divided into nine types (*DMs*, *reference*, *lexical*, *semantic*, *morphological*, *syntactic*, *graphical*, *genre* and *numerical* features). Finally, the third level in the hierarchy refers to specific signals; for example, *reference type* has four specific signals: *personal*, *demonstrative*, *comparative* and *propositional reference*. The hierarchical organization of the signalling taxonomy is provided in Figure 2. Note that subcategories in the figure are only illustrative, not exhaustive. For the detailed taxonomy and more information about the definitions of signals, see Das (2014), Das and Taboada (2017) and the RST Signalling Corpus (Das et al., 2015), together with the annotation manual (Das & Taboada, 2014), available online[10].

[10] http://www.sfu.ca/~mtaboada/docs/RST_Signalling_Corpus_Annotation_Manual.pdf

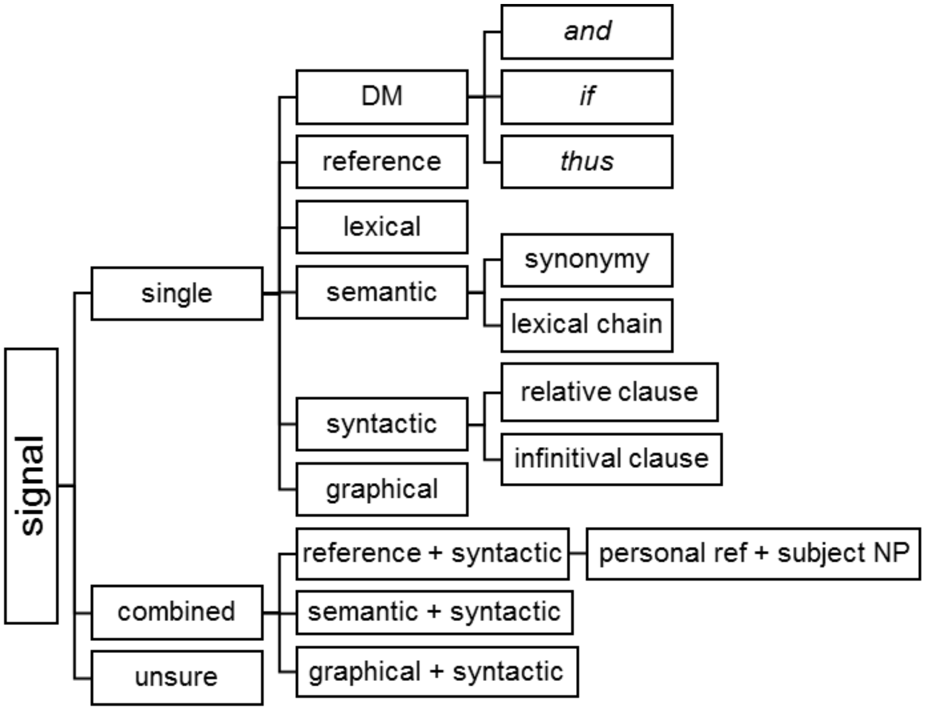

Figure 2: Hierarchical taxonomy of signals

A *single* signal is made of one (and only one) feature used to indicate a particular relation. In Example (10) below[11], the DM *because*, which is a single signal, is used to signal the Explanation-argumentative relation.

(10) [The Christmas quarter is important to retailers]N [because it represents roughly a third of their sales and nearly half of their profits.]S – Explanation-argumentative (wsj_640: 22/23)

In Example (11), the *Interpretation* relation is indicated by a lexical signal, the alternate expression *That means*, a single signalling feature.

[11] Conventions for annotated examples: The text within square brackets denotes a span. Each pair of square brackets is followed by either N, referring to the nucleus span, or S, referring to the satellite span. A pair of two spans (N and S) is followed by a dash and the name of the relation that holds between the spans. The parentheses at the end contain the file number of the source document, and the span numbers (the location of the relation in the document). In addition, the file number and the span numbers within the parentheses are separated by a colon, and each span number is separated from the other span number by a forward slash. The particular signal being discussed is underlined.

(11) [Production of full-sized vans will be consolidated into a single plant in Flint, Mich.]N [That means two plants -- one in Scarborough, Ontario, and the other in Lordstown, Ohio -- probably will be shut down after the end of 1991...]S – Interpretation (wsj_2338: 45/46-53)

We would like to point out that DMs and the lexical type are very closely-related categories, and can be argued to belong to a single broad type, such as ‘cue phrases’ as in Knott (1996). This is particularly true for alternate expressions (short tensed clauses) such as *that means* in Example (11) which could potentially function as a linking element between two discourse segments, and indicate a relation such as correction, repetition or restatement. From a relational point of view, these expressions could be considered as belonging to the category of DMs. However, in our study we use a fairly strict definition of DMs which include words or phrases (conjunctions, conjunctive adverbs, adverbial and prepositional phrases) but exclude clauses. For this reason, we assign clausal expressions (such as *that means*) under the lexical category which include words, phrases as well as clauses. Another important difference between DMs and the lexical type is that while DMs primarily (if not always) function as linking elements and indicators of relations, words/phrases/clauses constituting the lexical type (as indicative words and alternate expressions) mainly have other functions (conceptual or grammatical or both) in a text. This is not only true for the lexical type, but also for all the other types of signals, and this is precisely what distinguishes DMs from all other signals: Signalling a relation is the primary function of DMs, while the signalling function is secondary for other types of signals.

Coming back to the discussion of single signals, we provide an instance of *Condition* relation in Example (12) which is signalled by a syntactic feature, *subject auxiliary inversion*, which is also a single signal.

(12) [Should the courts uphold the validity of this type of defense,]S [ASKO will then ask the court to overturn such a vote-diluting maneuver recently deployed by Koninklijke Ahold NV.]N – Condition (wsj_2383: 11/12-13)

A combined signal comprises two single signals or features which work in combination with each other to signal a particular relation. In Example (13), two types of single signals, *reference* and *syntactic* feature, operate together to signal the *Elaboration-general-specific* relation. The reference feature indicates that the word *These* in the satellite span is a demonstrative pronoun because it refers back to the object *$100 million of insured senior lien bonds*, mentioned in the nucleus span. Syntactically, the demonstrative pronoun, *These,* is also in the subject position of the sentence the satellite span starts with, providing more detail about the object *$100 million of insured senior lien bonds* in the *Elaboration-general-specific* relation. Therefore, the combined signal, comprising the *reference* and *syntactic* feature — in the form of a *demonstrative reference* plus a *subject NP*—functions here as a signal for the *Elaboration-general-specific* relation.

(13) [The issue includes $100 million of insured senior lien bonds.]N [These consist of current interest bonds due 1990-2002, 2010 and 2015, and capital appreciation bonds due 2003 and 2004,…]S – Elaboration-general-specific (wsj_1161: 69/70-73)

We would like to point out that every single signal in the taxonomy could possibly be used in combination with some other single signal and constitute a combined signal. However, we came up with only a certain set of combined signals because they occurred in the corpus. Those single signals which were not used as part of a combined signal in this study could well be found as such in corpora belonging to different genres or different languages.

Finally, *unsure* refers to those cases in which no signal was found, as represented in Example (14) and (15). We discuss these in Section 4.

(14) ["Mastergate" is subtitled "a play on words," and Mr. Gelbart plays that game as well as anyone.]N [He describes a Mastergate flunky as one who experienced a "meteoric disappearance" and found himself "handling blanket appeals at the Bureau of Indian Affairs."]S – Evidence (wsj_1984: 79-80/81-83)

(15) [First Boston Corp. projects that 10 of the 15 companies it follows will report lower profit.]N Most of the 10 have big commodity-chemical operations.]S – Explanation-argumentative (wsj_2398: 26-27/28)

Relations can also be indicated by multiple signals (by more than one signal), as can be seen in Example (9), at the beginning part of Section 3. The difference between combined signals and multiple signals is one of independence of operability. In a combined signal, there are two signals, one of which is an independent signal, while the other one is dependent on the first signal. For example, in a combined signal such as (*personal reference* + *subject NP*), the feature *personal reference* is the independent signal because it directly (and independently) refers back to the entity introduced in the first span. In contrast, the feature *subject NP* is the dependent signal because it is used to specify additional attributes of the first signal. In this particular case, the syntactic role of the personal reference (i.e., a subject NP) in the second span is specified by the use of the second signal *subject NP*. For multiple signals, on the other hand, each signal functions independently and separately from each other, but they all contribute to signalling the relation. For example, in an Elaboration relation with multiple signals, such as a genre feature (e.g., *inverted pyramid scheme*) and a lexical feature (e.g., *indicative word*), the signals do not have any connection, but they separately signal the relation.

3.3 Procedure

In our signalling annotation, we perform a sequence of three tasks: (i) We examined each relation in the RST-DT; (ii) Assuming that the relational annotation is correct, we searched for signals that indicate that such relation is present; and finally, (iii) We added to those relations a new layer of annotation of signalling information.

We annotated all the 385 documents in the RST-DT (divided into 347 training documents and 38 test documents) containing 20,123 relations in total[12]. We used the taxonomy of signals presented in Figure 2 in Section 3.2 to annotate the signals for those relations in the corpus. In some cases, more than one signal may be present. When confronted with a new instance of a particular type of relation, we consulted our taxonomy, and tried to find the 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 did not match any of the signals in the taxonomy, then we examined the context (comprising the spans) to discover any potential new signals. If a new signal was identified, we included it in the appropriate category in our existing taxonomy. In this way, we proceeded through identifying the signals of the relations in the corpus, and, at the same time, continued to update our taxonomy with new signalling information, if necessary. We found that after approximately 50 files, or 2,000 relations, we added very few new signals to the taxonomy.

In order to facilitate the annotation process, we used UAM CorpusTool (O'Donnell, 2008), a software for text annotation. UAM CorpusTool allowed us to create a hierarchically-organized tagging scheme, including all three levels of signals: signal class, signal type and specific signal. It also provides the option for multiple annotations for a single element. The tool is easy to use, does not require advanced computational knowledge, and provides an adequate visualization of source and annotated data.

UAM CorpusTool can directly import RST files, and show the discourse structure of a text in the form of RST trees, although it does not support layered annotation on top of RST-level structures. We, however, found out that it is possible to import the RST base files (along with all relational information) into UAM CorpusTool after converting them from their original LISP-style format to a simple text file

[12] In practice, we annotated 21,400 relations in total. This number is higher than the number of relations (20,123) stated above. This is because we considered multinuclear relations with more than two nuclei to be a number of individual binuclear relations (sets of relations with two nuclei). For more information, see Das (2014) and Das and Taboada (2017).

format. This allowed us to select individual relations and tag them with relevant signal tags. In addition, the annotated data in UAM CorpusTool is stored in XML.

UAM CorpusTool has two added advantages. First, it provides an excellent tag-specific search option for finding required annotated segments. Second, UAM CorpusTool provides various types of statistical analyses of the corpus, some of which we present here. Additional studies and other types of feature extraction are possible with the combination of the annotated corpus and UAM CorpusTool.

3.4 An example of signalling annotation

For illustration purposes, we provide the annotation of an RST file from the RST-DT (file number: wsj_650) with signalling information. The text is the same as in Example (3) above, and the graphical representation can be found in Figure 1. A detailed description of our annotation is provided in Table 2.

File	N	S	Relation	Signal type	Specific signal	Explanation: How signalling works
				genre	inverted pyramid scheme	In the newspaper genre, the content of the first paragraph (or the first few paragraphs) is elaborated on in the subsequent paragraphs.
					lexical overlap	The word *debenture* occurs both in the nucleus and satellite.
					lexical chain	Words such as *debentures*, *issue price*, *convertible*, *conversion price* and *share* are in a lexical chain.
				(semantic + syntactic)	(lexical chain + subject NP)	The phrases *Sun Microsystems Inc.* and *the company* in the respective spans are in a lexical chain, and the latter is syntactically used as the subject NP of the sentence the satellite starts with.
	3/4		List	DM	*and*	The DM *and* functions as a signal for the *List* relation.
	3-4	2	Attribution	syntactic	reported speech	The reporting clause plus the reported clause construction is a signal for the *Attribution* relation.
	2-4/5		List	semantic	lexical chain	The words, *issued*, *convertible*, *debentures*, *available*, in the respective spans are semantically related.

Table 2: Annotation of an RST file with relevant signalling information

According to our annotation, the *Elaboration (-additional)* relation between span 1 and span 2-5 is indicated by three types of signals, more specifically by two types of *single* signals: *genre* and *semantic* features; and by a *combined* type of signal: (*semantic + syntactic*) feature. First, the text represents the newspaper genre (since it is taken from a Wall Street Journal article). In newspaper texts, the content of the first (or the first few) paragraphs is typically elaborated on in the subsequent paragraphs. A reader, being conscious of the fact that they are reading a newspaper article, expects the presence of an *Elaboration* relation between the first paragraph (or the first few paragraphs) and subsequent paragraphs. It is this prior knowledge about the textual organization of the newspaper genre that guides the reader to interpret an *Elaboration* relation between paragraphs in a news text. In this particular example, the entire first paragraph is the nucleus of the *Elaboration* relation, with the two following paragraphs constituting the satellite. Thus, we postulate that the *Elaboration* relation is conveyed by the *genre* feature more specifically by a feature which we call *inverted pyramid scheme* (Scanlan, 2000). Second, the *Elaboration* relation is also signalled by two *semantic* features: *lexical overlap* and *lexical chain*. The word *debentures* occurs in both the nucleus and satellite spans, 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. Also, words such as *convertible* and *debentures* in the first span and words (or phrases) such as *issue price, convertible, conversion price* and *share* in the second span are semantically related. These words form a lexical chain which is a strong signal for an *Elaboration* relation. Finally, we postulate that a *combined* feature (*semantic* + *syntactic*), made of two individual features is operative in signalling the *Elaboration* relation: The entity *Sun Microsystems Inc.*, mentioned in the nucleus, is elaborated on in the satellite. The phrase *Sun Microsystems Inc.* is semantically related to the phrase *the company* in the satellite, and hence, they are in a lexical chain. Syntactically, the phrase *the company* is used as the subject NP of the sentence the satellite starts with, representing the topic of the *Elaboration* relation.

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

The *Attribution* relation between span 2 and span 3-4 is indicated by a *syntactic* signal, the *reported speech* feature, 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 subject-verb combination with a reported speech verb (*said*).

Finally, the *List* relation between span 2-4 and span 5 is indicated by a *semantic* feature, *lexical chain*. Words such as *issued* and *convertible* (in the first nucleus) and words *debentures* and *available* (in the second nucleus) are semantically related, indicating a *List* relation between the spans.

3.5 Reliability of annotation

In order to check the validity and reproducibility of our initial annotation and original taxonomy, we conducted a reliability study. We selected two files from the corpus, containing 130 relations in total, and both authors annotated them independently. We concentrated on whether we agreed on each of the signals for every single relation. Some relations have multiple signals (more than one signal), and some relations have combined signals. As calculating agreement on those would become very complex quite quickly, we stayed with single signals. Also because of the complexity of the task, we calculated agreement focusing only the *signal types* in the signalling taxonomy, and not involving *specific signals*.

We used Cohen's Kappa (Siegel & Castellan, 1988) for calculating the agreement value, with nominal data representing the nine categories of signals in our classification, plus an additional category *unsure* (used to indicate those situations in which the annotators did not find any identifiable signal). The unweighted and weighted kappa values for our reliability study are 0.67 and 0.71, respectively, which imply moderate agreement. Given that there are 10 different categories to choose from, we feel that this is a good level of agreement, and we do believe that our annotation is reproducible. For more information about the reliability study, see Das (2014) and Das and Taboada (2017).

A general issue about reliability studies is whether they are useful at all, particularly in the context of discourse annotation which is performed by the members of the same research groups who share similar points of view. The even larger question is whether providing values for kappa or for similar measures reveals much about the annotation process and its level of difficulty. In this regard, our stance is that discourse annotation is inherently subjective, because many of the decisions rely of interpreting the text, or re-interpreting what the author meant. We believe what is more required than arriving at an acceptable measure of agreement is an acceptance of the intrinsic difficulty of annotation, together with a reasonable explanation of how the annotation was performed. For more discussion on this issue, see Taboada and Das (2013) and Das and Taboada (2017).

3.6 Final product: RST Signalling Corpus

The final outcome of our study is the RST Signalling Corpus (Das et al., 2015), a discourse-annotated corpus of signals of coherence relations. The corpus is available from the Linguistic Data Consortium or LDC (https://catalog.ldc.upenn.edu/LDC2015T10), for a fee as a single user, or free to LDC members.

The RST Signalling Corpus includes 29,297 signal tokens for 21,400 relation instances, with a breakdown into 24,220 (82.7%) single signals, 3,524 (12.0%) combined signals and 1,553 (5.3%) unsure cases (in which the appropriate signals for relations were not found). The distribution of the signals is provided in Table 5 in the next section. More information about the corpus can be found in Das and Taboada (2017).

4 Results: Relation distribution and signalling

In this section, we provide descriptive statistics of the frequency of relations and how often each of them is signalled. In addition, we carried out statistical tests of significance, to establish whether there are differences across relations in terms of their association with particular signals.

We divided the annotated relations in two broad groups: signalled and unsignalled. Then, the signalled relations are divided in three sub-groups: (1) relations exclusively signalled by DMs, (2) relations exclusively signalled by other signals and (3) relations signalled by both DMs and other signals. The distribution is provided in Table 3, which shows that 19,847 relations (92.74%), out of all the 21,400 annotated relations, are signalled either by DMs or by means of other signals or by both. On the other hand, no significant signalling evidence is found for the remaining 1,553 relations (7.26%). We discuss the apparently unsignalled relations at the end of this section. The distribution also shows that 10.65% of the relations are exclusively signalled by DMs while 74.54% of the relations are exclusively indicated by other signals. In addition, 1,616 relations or 7.55% of the relations in the corpus are indicated by both DMs and other signals. This result suggests that, if we limit the signalling phenomenon only to DMs (as postulated in most previous studies on signalling), then the degree of signalling is indeed very low: Only 18.21% of the relations in the corpus (2,280 + 1,616 = 3,896 relations out of 21,400 relations) are signalled (by DMs).

Relation type	Signalling type	Frequency	Percentage
	Relations exclusively signalled by DMs	2,280	10.65%
	Relations exclusively signalled by other signals	15,951	74.54%
	Relations signalled by both DMs and other signals	1,616	7.55%
	TOTAL	**19,847**	**92.74%**
Unsignalled relations	Relations not signalled by DMs or other signals	1,553	7.26%
TOTAL		**21,400**	**100.00%**

Table 3: Distribution of signalled and unsignalled relations

We would like to note that the proportion of DMs in our corpus (18.21% of all the annotated relations (3,896 relations out of 21,400 relations) and 19.63% of the signalled relations (3,896 relations out of 19,847 signalled relations) is lower than the results documented in many previous studies on the signalling of coherence relations by DMs (see Section 2). We believe that there are two reasons for this. First, we use a fairly strict definition of DMs, and our criteria for considering an expression to be a DM excludes many expressions which are treated as DMs elsewhere. For instance, we do not consider expressions such as *always assuming that*, *for the simple reason* and *in other respects* to be examples of DMs, but consider them to be *indicative phrases* (under the *lexical* feature). However, these expressions are included under the class of DMs in other studies such as Knott (1996). Second, the RST-DT uses a very finely-grained definition of the atomic units, producing a number of relations which are not usually recognized as coherence relations in classical RST or in in most studies on DMs. These relations include *Attribution* (relation between a reporting clause and a reported speech clause), *Same-unit* (relation between discontinuous clauses), *Elaboration-e* (relation between a main clause and non-restrictive relative clause) and *Elaboration-object-attribute-e* (relation between a main clause and a restrictive relative clause). These relations occur in high frequencies in the RST-DT and constitute a significantly large portion in the corpus (as shown in Table 4). However, the fact that they are not signalled by DMs (although they most frequently signalled, particularly by syntactic signals, as shown in Table 4, Table 7 and Table 11) has probably contributed to yield an overall lower proportion of relations with DMs.

For the 3,896 instances of relations signalled by DMs, we found 201 different DMs. Examples of some of these markers 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*. A full list of these DMs can be found in Das (2014) and the RST Signalling Corpus annotation manual (Das & Taboada, 2014).

In Table 4, we provide the detailed distribution of individual relations with respect to whether they are signalled or unsignalled. The table also contains the distribution of relation types in the RST-DT. (Note: The percentage figures in column 6 refer to the proportions of signalled and unsignalled relations for a relation type, and should be interpreted horizontally across the rows, while the percentage figures in column 8 refer to the proportion of relation types against the total number of relations in the corpus and should be interpreted vertically, along column 7 and 8).

#	Relation group	Relation	# signalled	# unsignalled	% signalled	Total relation	% total relation
1	Attribution	Attribution	3061	9	99.71%	3070	14.35%
		Background	185	42	81.50%	227	1.06%
		Circumstance	635	75	89.44%	710	3.32%
		Cause	43	9	82.69%	52	0.24%
		Result	122	37	76.73%	159	0.74%
		Cause-Result	56	9	86.15%	65	0.30%
		Consequence	343	74	82.25%	417	1.95%
		Comparison	242	23	91.32%	265	1.24%
		Preference	15	0	100%	15	0.07%
		Analogy	16	4	80.00%	20	0.09%
		Proportion	3	0	100%	3	0.01%
		Condition	234	5	97.91%	239	1.12%
		Hypothetical	8	38	17.39%	46	0.21%
		Contingency	24	3	88.89%	27	0.13%
		Otherwise	15	1	93.75%	16	0.07%
		Contrast	388	47	89.20%	435	2.03%
		Concession	277	16	94.54%	293	1.37%
		Antithesis	369	33	91.79%	402	1.88%
		Elaboration-additional	4043	101	97.56%	4144	19.36%
		Elaboration-general-specific	452	21	95.56%	473	2.21%
		Elaboration-part-whole	44	0	100%	44	0.21%
		Elaboration-process-step	2	1	66.67%	3	0.01%
		Elaboration-object-attribute	2685	13	99.52%	2698	12.61%
		Elaboration-set-member	126	3	97.67%	129	0.60%

		Example	276	56	83.13%	332	1.55%
		Definition	46	33	58.23%	79	0.37%
		Purpose	526	11	97.95%	537	2.51%
		Enablement	9	22	29.03%	31	0.14%
		Evaluation	183	9	95.31%	192	0.90%
		Interpretation	185	28	86.85%	213	1.00%
		Conclusion	2	3	40.00%	5	0.02%
		Comment	155	35	81.58%	190	0.89%
		Evidence	110	64	63.22%	174	0.81%
		Explanation-argumentative	392	214	64.69%	606	2.83%
		Reason	173	33	83.98%	206	0.96%
		List	1843	112	94.27%	1955	9.14%
		Disjunction	27	0	100%	27	0.13%
		Manner	85	11	88.54%	96	0.45%
		Means	121	9	93.08%	130	0.61%
		Problem-solution	46	19	70.77%	65	0.30%
		Question-answer	8	25	24.24%	33	0.15%
		Statement-response	18	14	56.25%	32	0.15%
		Topic-comment	0	5	0.00%	5	0.02%
		Comment-topic	1	1	50.00%	2	0.01%
		Rhetorical-question	3	16	15.79%	19	0.09%
		Summary	69	14	83.13%	83	0.39%
		Restatement	111	29	79.29%	140	0.65%
		Temporal-before	42	2	95.45%	44	0.21%
		Temporal-after	87	6	93.55%	93	0.43%
		Temporal-same-time	135	25	84.38%	160	0.75%
		Sequence	188	30	86.24%	218	1.02%
		Inverted-sequence	13	2	86.67%	15	0.07%
		Topic-shift	31	87	26.27%	118	0.55%
		Topic-drift	19	68	21.84%	87	0.41%
17	Textual Organization	Textual-organization	156	1	99.36%	157	0.73%
18	Same-Unit	Same-unit	1399	5	99.64%	1404	6.56%
TOTAL			19847	1553	92.74%	21400	100.00%

Table 4: Distribution of relations and relation groups by signalled and unsignalled categories

Table 4 shows that almost every individual relation type (and almost every group of relations) contains signals. Individual relations such as *Attribution, Circumstance, Comparison, Condition, Contrast, Concession, Elaboration-additional*, *Elaboration-set-member*, *List*, *Means*, *Temporal-before* and *Textual-*

organization are most frequently signalled. In fact, relations such as *Elaboration-part-whole* and *Disjunction* are always signalled. On the other hand, there are only a few relations such as *Hypothetical*[13], *Enablement, Question-answer* and *Topic-shift* for which signalling is not very common.

The newly annotated signalling corpus includes 29,297 signal tokens for 21,400 relation instances[14], with a breakdown into 24,220 (82.7%) single signals, 3,524 (12.0%) combined signals and 1,553 (5.3%) unsure cases (in which the appropriate signals for relations were not found). The detailed distribution of signals in the corpus is provided in Table 5.

#	Signal class	Signal type	Specific signal	# of tokens	Total	%
		DM	*and, but, if, since, then,* etc.	3,909	3,909	13.34%
			personal reference	260		
			demonstrative reference	134		
			comparative reference	182		
			propositional reference	10		
			indicative word	1,399		
			alternate expression	41		
			synonymy	38		
			antonymy	37		
			meronymy	34		
			repetition	1,405		
			indicative word pair	19		
			lexical chain	5,700		
			general word	29		
		morphological	tense	313	313	1.07%
			relative clause	1,621		
			infinitival clause	524		
			present participial clause	91		
			past participial clause	12		
			imperative clause	5		
			interrupted matrix clause	1,399		
			parallel syntactic construction	149		
			reported speech	3,023		
			subject auxiliary inversion	7		
			nominal modifier	1,881		
			adjectival modifier	11		
			colon	222		
			semicolon	20		
			dash	273		
			parentheses	247		

[13] One interesting observation is that Hypothetical relations are rarely signalled even though they belong to the broad group of Condition relations. Unlike Condition relations, Hypotheticals do not contain DMs. Sometimes, they include a modal verb, but that could only be considered as a very weak signal for the relation. Hypothetical relations tend to occur between larger chunks of text (as compared to more local Condition relations), thus making it even more difficult for annotators to find a reliable signal for them.

[14] The number of signal tokens is higher than the number of relation instances because many relations contain multiple signals.

			items in sequence	252		
			inverted pyramid scheme	720		
			newspaper layout	189		
			newspaper style attribution	26		
			newspaper style definition	8		
		numerical	same count	26	26	0.09%
			(personal reference + subject NP)	504		
			(demonstrative reference + subject NP)	23		
			(comparative reference + subject NP)	1		
			(propositional reference + subject NP)	15		
			(repetition + subject NP)	972		
			(lexical chain + subject NP)	1,042		
			(synonymy + subject NP)	22		
			(meronymy + subject NP)	84		
			(general word + subject NP)	35		
		(lexical + syntactic)	(indicative word + present participial clause)	120	120	0.41%
		(syntactic + semantic)	(parallel syntactic construction + lexical chain)	410	410	1.40%
			(past participial clause + beginning)	41		
			(present participial clause + beginning)	28		
			(comma + present participial clause)	216		
			(comma + past participial clause)	10		
3	unsure	unsure	unsure	1,553	1,553	5.3%
Total				29,297	29,297	100%

Table 5: Distribution of signals in the RST Signalling Corpus

In order to determine whether certain relations and certain signals are more frequently associated with each other, we computed several measures of association[15]. We describe each in detail in the next subsections.

4.1 Relation groups and signalling

We first computed the mean proportions of relations signalled by each signal. We have a large dataset comprising 19 relation groups (and 78 individual relations) and 16 signal types (including single, combined and unsure types) along with over 50 specific signals. In order to reduce the degree of statistical complexity generated from such a large dataset, we decided to stay only with relation groups (and not individual

[15] The statistical analyses were carried out using the SAS® statistical package, version 9.4.

relations) and only signal types (nine single signal types and the unsure type, thus excluding the combined type and also specific signals). Furthermore, the distribution of relation groups with respect to signal types is extremely diverse, with counts ranging from over 4,000 tokens (e.g., the *Elaboration* group signalled by the semantic type) to zero tokens (e.g., *Enablement* group by the reference type). We also decided to consider only those counts equating 10 or more for improved model fitting.

The predicted mean proportions (least squares means) of relations with respect to the total number of relations in a relation group for DMs are provided in Table 6. A binary logistic regression model was used to calculate these predicted mean proportions.

Relation Least Squares Means						
		Standard				Standard
Relation	**Estimate**	**Error**	**z Value**	**Pr > |z|**	**Mean**	**Error of Mean**
Background	-0.2597	0.06589	-3.94	<.0001	0.4354	0.01620
Cause	-0.2115	0.07640	-2.77	0.0056	0.4473	0.01889
Comparison	-0.7837	0.1238	-6.33	<.0001	0.3135	0.02665
Condition	1.4171	0.1393	10.17	<.0001	0.8049	0.02188
Contrast	1.3808	0.07425	18.60	<.0001	0.7991	0.01192
Elaboration	-3.0657	0.05453	-56.22	<.0001	0.04455	0.002321
Enablement	-3.3636	0.2334	-14.41	<.0001	0.03345	0.007545
Evaluation	-2.0083	0.1264	-15.89	<.0001	0.1183	0.01319
Explanation	-1.2711	0.07700	-16.51	<.0001	0.2191	0.01317
Joint	-0.2947	0.04541	-6.49	<.0001	0.4268	0.01111
Manner-Means	-1.8390	0.1934	-9.51	<.0001	0.1372	0.02288
Temporal	0.6398	0.09136	7.00	<.0001	0.6547	0.02065
Topic-Change	-2.1704	0.2303	-9.42	<.0001	0.1024	0.02118
Topic-Comment	-2.3168	0.2801	-8.27	<.0001	0.08974	0.02288

Table 6: Mean proportions of relation groups for signalling by DMs

Table 6 shows what relation groups are most commonly signalled by DMs and how frequently they are signalled by DMs. The values in the 'Mean' column represent the ratio of number of relations from a particular group signalled by DMs to total number of relations belonging to that group (e.g., how many relations from the *Contrast* relation group are signalled by DMs with respect to the total number of relations in the *Contrast* group). The figures in the 'Mean' column also represent the predicted mean proportion values (i.e., the probabilities of finding those relations which are signalled by DMs from a particular relation group, given the total population of the relations in the group)

Table 6 shows that *Condition* (mean = 0.8049) and *Contrast* (mean = 0.7991) are the two relation groups which are most frequently signalled by DMs. They are followed by groups such as *Background*, *Joint* and *Temporal*, which are moderately signalled by DMs. On the other hand, relation groups such as *Elaboration*, *Enablement*, *Evaluation* and *Topic-Comment* are infrequently signalled by DMs. This broad distinction between causal, concessive and contrastive relations on the one hand, and relations of elaboration on the other, has been well documented (e.g., Taboada (2006)). There are two relation groups, *Attribution* and *Same-unit*, which are not present in Table 6, because these relations are not signalled by DMs in our corpus at all.

We present in Table 7 similar distribution for the syntactic type of signals which is the most frequently occurring group of signals among all types (see Table 5).

relation Least Squares Means						
relation	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**	**Mean**	**Standard Error of Mean**
Attribution	4.2080	0.1502	28.02	<.0001	0.9853	0.002169
Background	-2.4274	0.1197	-20.28	<.0001	0.08111	0.008919
Cause	-2.9042	0.1712	-16.97	<.0001	0.05195	0.008430
Elaboration	-0.2432	0.02267	-10.73	<.0001	0.4395	0.005583
Enablement	1.9141	0.1254	15.27	<.0001	0.8715	0.01404
Joint	-2.5170	0.08545	-29.46	<.0001	0.07467	0.005904
Manner-Means	-2.1777	0.2200	-9.90	<.0001	0.1018	0.02011
Same-Unit	5.6341	0.4480	12.58	<.0001	0.9964	0.001590
Temporal	-3.2387	0.2279	-14.21	<.0001	0.03774	0.008277

Table 7: Mean proportions of relation groups for signalling by syntactic type

We also computed the mean proportions of relation groups for the other seven single signal types (and the unsure type) following the same methodology, and found what relations are most (or least) frequently signalled with respect to any of those signal types. Providing all these distributions would take a considerable amount of space, so we restrict ourselves to providing such distributions only for DMs (in Table 6) and the syntactic type (Table 7). Conclusions on which signals are more frequently found with which relations are summarized in Table 11 in Section 4.5, but all the information will be made available online as supplementary material upon publication.

4.2 Pairwise comparison of relation groups and signals

In order to determine whether a specific relation group stood out in terms of its association with each signal type, we compared the mean proportions between relation groups for every single signal type using binomial logistic regression. Post hoc tests using the Tukey-Kramer adjustment are used to compare mean proportions for all pairs of relation groups with respect to a signal type. We provide the distribution of a few relation group pairs for DMs in Table 8.

Differences of relation Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer						
			Standard			
relation	**_relation**	**Estimate**	**Error**	**z Value**	**Pr > |z|**	**Adj P**
Background	Cause	-0.04826	0.1009	-0.48	0.6324	1.0000
Background	Comparison	0.5239	0.1403	3.74	0.0002	0.0136
Background	Condition	-1.6768	0.1541	-10.88	<.0001	<.0001
Background	Contrast	-1.6405	0.09927	-16.53	<.0001	<.0001
Background	Elaboration	2.8059	0.08553	32.81	<.0001	<.0001
Background	Enablement	3.1039	0.2425	12.80	<.0001	<.0001
Background	Evaluation	1.7486	0.1425	12.27	<.0001	<.0001
Background	Explanation	1.0114	0.1013	9.98	<.0001	<.0001
Background	Joint	0.03503	0.08002	0.44	0.6616	1.0000
Background	Manner-Means	1.5793	0.2043	7.73	<.0001	<.0001
Background	Temporal	-0.8996	0.1126	-7.99	<.0001	<.0001
Background	Topic-Change	1.9107	0.2396	7.98	<.0001	<.0001
Background	Topic-Comment	2.0570	0.2878	7.15	<.0001	<.0001
Cause	Comparison	0.5722	0.1455	3.93	<.0001	0.0064
Cause	Condition	-1.6285	0.1589	-10.25	<.0001	<.0001
Cause	Contrast	-1.5922	0.1065	-14.95	<.0001	<.0001
Cause	Elaboration	2.8542	0.09386	30.41	<.0001	<.0001
Cause	Enablement	3.1522	0.2455	12.84	<.0001	<.0001
Cause	Evaluation	1.7968	0.1477	12.17	<.0001	<.0001
Cause	Explanation	1.0596	0.1085	9.77	<.0001	<.0001
Cause	Joint	0.08329	0.08888	0.94	0.3487	0.9997
Cause	Manner-Means	1.6275	0.2079	7.83	<.0001	<.0001
Cause	Temporal	-0.8513	0.1191	-7.15	<.0001	<.0001
Cause	Topic-Change	1.9590	0.2427	8.07	<.0001	<.0001
Cause	Topic-Comment	2.1053	0.2904	7.25	<.0001	<.0001
Comparison	Condition	-2.2007	0.1864	-11.81	<.0001	<.0001
Comparison	Contrast	-2.1644	0.1444	-14.99	<.0001	<.0001
…	…	…	…	…	…	…

Table 8: Comparison of mean proportions between relation groups for DMs

The complete analysis indicates that the differences in mean proportions between relation groups with respect to DMs are statistically significant (at $p < 0.05$) in the vast majority of cases. There are only a few pairs for which the mean differences between groups are not statistically significant. Examples of such cases are observed for pairs such as *Background ~ Cause*, *Background ~ Joint*, and *Cause ~ Joint* (as presented in Table 8) and also for a few other pairs for DMs (not shown in Table 8). Similar analyses were also conducted for the other eight single signal types and the unsure type, and they also show that the mean differences between relation groups are mostly statistically significant, again with only a few instances not being statistically significant. This means that, for any given pair of relations, and any given signal type, statistically significant mean differences in their distribution would enable us to distinguish them. We also examined whether the mean differences for the relation group pairs are systematically distributed across all signal types. However, we observed no such significant patterns for those pairs, and the mean differences between relation groups seem to be randomly distributed across signal types.

4.3 Signal distribution with respect to relations

We computed the mean proportions of relation groups signalled by a particular signal type with respect to the total number of instances for that signal type. The distribution of relation groups for DMs is provided in Table 9.

Relation Least Squares Means						
		Standard				Standard
Relation	**Estimate**	**Error**	**z Value**	**Pr > |z|**	**Mean**	**Error of Mean**
Background	-2.1495	0.05231	-41.09	<.0001	0.1044	0.004890
Cause	-2.4518	0.05919	-41.42	<.0001	0.07930	0.004322
Comparison	-3.6926	0.1039	-35.55	<.0001	0.02430	0.002463
Condition	-2.6252	0.06374	-41.19	<.0001	0.06754	0.004014
Contrast	-1.2026	0.03795	-31.69	<.0001	0.2310	0.006741
Elaboration	-2.3130	0.05588	-41.40	<.0001	0.09005	0.004578
Enablement	-5.3217	0.2300	-23.14	<.0001	0.004861	0.001112
Evaluation	-3.9900	0.1198	-33.31	<.0001	0.01816	0.002136
Explanation	-2.8389	0.07000	-40.55	<.0001	0.05526	0.003654
Joint	-1.2866	0.03884	-33.13	<.0001	0.2164	0.006587
Manner-Means	-4.8291	0.1803	-26.78	<.0001	0.007930	0.001419
Temporal	-2.3288	0.05624	-41.41	<.0001	0.08877	0.004549
Topic-Change	-5.2211	0.2188	-23.86	<.0001	0.005372	0.001169
Topic-Comment	-5.6284	0.2677	-21.02	<.0001	0.003581	0.000955

Table 9: Mean proportions of relations signalled by DMs across relation groups

Table 9 shows how the population of different relations signalled by DMs is distributed across the total population of DMs in the corpus, and in what proportions. The values in the 'Mean' column represent the ratio of the number of relations from a particular group signalled by DMs to the total number of relations signalled by DMs. The mean values also represent the predicted mean probabilities of finding those relations signalled by DMs from a particular relation group, given the total population of relations signalled by DMs.

Table 9 indicates that, given a set of different types of relations signalled by DMs, the most frequently occurring relations would most likely belong to the *Contrast*, *Joint*, *Background* and *Cause* groups. On the other hand, chances of finding relations from groups such as *Enablement*, *Manner-Means* and *Topic-Comment* are low, as they are infrequently signalled by DMs.

In the same way, we computed the predicted mean proportions of relation groups for the other eight single signal types. As with the analyses above, we only provide the distribution for DMs as a representative sample, and the conclusions are summarized in Table 11.

Analyzing the distributions for other signal types, we observed the following tendencies: The reference type is most likely associated with *Elaboration* and *Attribution*; the lexical type with *Background*, *Elaboration* and *Evaluation*; the semantic type with *Elaboration* and *Joint*; the morphological type with *Background*; the syntactic type with *Attribution* and *Elaboration*; the graphical type with *Elaboration* and *Joint*; the genre type with *Elaboration* and *Textual Organization*; and finally, the numerical feature with *Elaboration*. One interesting pattern emerging from these distributions is that *Elaboration* is the typical relation group for the majority types of signals. This implies that, contrary to the popular opinion that *Elaboration* relations are rarely signalled, identification of *Elaboration* relations is achieved using a variety of signals. This would make it challenging to identify automatically, as no particular signal is associated with it, but it is not the case that *Elaboration* is an unsignalled or implicit relation in all its occurrences.

4.4 Multiple signals

As previously mentioned, a single relation instance can be indicated by more than one signal. We found that a considerable number of relations are signalled by multiple signals. In Table 10, we provide the descriptive statistics of the distribution of individual relations with respect to being signalled by multiple signals.

Multiple signals	**# signalled**	**% signalled**	**Common signalled relations**	**# common signalled relations**	**% common signalled relations**
			Elaboration-additional	4,043	97.56%
			Attribution	3,061	99.71%
			Elaboration-object-attribute	2,685	99.52%
			List	1,843	94.27%
			Circumstance	635	89.44%
			Purpose	526	97.95%
			Explanation-argumentative	392	64.69%
			Antithesis	369	91.79%
			Elaboration-additional	2,745	66.24%
			List	861	44.04%
			Elaboration-general-specific	185	39.11%
			Contrast	182	41.84%
			Circumstance	148	20.85%
			Example	142	42.77%
			Antithesis	133	33.08%
			Elaboration-set-member	108	83.72%
			Elaboration-additional	1,561	37.67%
			Elaboration-general-specific	104	21.99%
			Summary	62	74.70%
			List	56	2.86%
			Elaboration-additional	552	13.32%
			Summary	35	42.17%
			Elaboration-general-specific	25	5.29%
			Elaboration-additional	80	1.93%
			Elaboration-general-specific	6	1.27%
6 signals	9	0.04%	Elaboration-additional	9	0.22%

Table 10: Distribution of relations with respect to being signalled by more than one signal

The distribution in Table 10 shows that most relations have at least one signal. This means that relation identification, by humans and machines, relies or can rely on signals as indicators that a relation is present. Given that a large number of relations have only signal, however, bootstrapping is not always available, and, if that signal is ambiguous or underspecified, then accurate identification will be more difficult.

4.5 Summary of results

We provide a summary of the relationship between relation groups and signal types (including the unsure type) in Table 11. The check marks point to high compatibility between a relation group and a signal type (or the unsure type), while the cross marks suggest a weak association or no association between them. 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 the relation (as we have seen in the many-to-many correspondences between relations and their signals). It also means that the signals, as textual devices, are not exclusively used to mark a relation; they may well have other purposes in the text (for instance, a pronoun, an instance of the *reference* signal type, also contributes to cohesion, in addition to signalling a relation). In a sense, this means that the signals are *compatible* with a relation, not necessarily indicators of the relation exclusively. We believe, however, that our results provide evidence that relation signalling is widespread, and has clear potential for interesting applications (see the Conclusions section).

Relation group	**Signal type**									
	DM	Ref	Lex	Sem	Morph	Syn	Graph	Genre	Num	Unsure
Attribution	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Background	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗
Cause	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗
Comparison	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Condition	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Contrast	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
Elaboration	✗	✓	✗	✓	✗	✓	✓	✓	✓	✗
Enablement	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Evaluation	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
Explanation	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓
Joint	✓	✗	✓	✓	✗	✓	✓	✗	✗	✗
Manner-Means	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗
Topic-Comment	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓
Summary	✗	✗	✗	✓	✗	✗	✓	✓	✗	✓
Temporal	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗
Topic-Change	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓
TextualOrganization	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Same-Unit	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗

Table 11: Compatibility between relation groups and signal types

As we close this section, we would like to add a note about unsignalled relations in the corpus. Although we found that the majority of relations in the corpus are signalled, we could not clearly identify a signal for 1,553 relations (7.26% of the 21,400 relations). There are four different reasons why we believe no signals could be found. First of all, in some cases we found that there were errors in the original relational annotation in the RST-DT. The errors usually emerge from an incorrect assignation of relation labels. In many cases, we found that a relation was postulated, whereas we would not have annotated a relation, or we would have proposed a different one. *Summary* and *Elaboration-additional* in the RST-DT seem to be used in very similar contexts, so when a *Summary* was annotated, but we believed the relation was not in fact a summary, it was more difficult to find signals that would identify the relation as *Summary*. Second, some of the relations in the RST-DT are not ‘true’ RST relations. Relations such as *Comment*, *Topic-Comment* or *Topic-shift*, in our opinion, belong in the realm of discourse organization, not together with relations among propositions. 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. Third, in annotating a relation we only considered the immediate spans holding the relation. However, we noticed that the interpretation of a relation does not always depend on the recognition of signals from the surrounding relation spans, but it is sometimes determined by the knowledge extracted from the prior or following parts of the discourse which are outside the immediate relation spans. 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. What may be happening in those cases is that the relation is being evoked, in the same way frames and constructions may be evoked (Dancygier & Sweetser, 2005). Dancygier and Sweetser propose that, in some constructions, only one aspect of the construction is necessary in order to evoke the entire construction. Such is the case with some instances of sentence juxtaposition, which give rise to a conditional relation reading, as in “Steal a bait car. Go to jail” (the slogan for a car-theft prevention campaign by the Vancouver police). No conditional connective is necessary. The juxtaposition of the two sentences, together with the imperative and a certain amount of world knowledge lead to the conditional interpretation.

5 Conclusions

We investigated how coherence relations are signalled in discourse quantitatively and qualitatively. Our first research objective was to establish how frequently coherence relations are signalled. The results (Table 3) show that the majority of the relations in the corpus contain signals. Out of the 21,400 relations annotated, 19,847 (92.74%) relations (Table 3) contain at least one signal (either DMs or other signals), and 7,901 (36.92%) relations (Table 10) contain two or more signals. Although some of the relations (1,553 relations, or 7.26% of the total 21,400 relations) are not signalled (Table 3), the overwhelming majority of them are.

Analyzing the distribution of signalled relations, we found (Table 3) that only 18.21% of the relations (3,896 relations out of 21,400 annotated relations) are signalled by DMs, and furthermore, only 10.65% of the relations (2,280 relations out of 21,400 annotated relations) are exclusively signalled by DMs (i.e., they are not conveyed through other signals, but contain DMs as their only signals). On the other hand, 82.08% of the total relations (17,567 relations out of 21,400 annotated relations) contain a signal other than DMs, and also 74.54% of the relations (15,951 relations out of 21,400 annotated relations) are exclusively indicated by other signals (i.e., they are not signalled by DMs, but contain other signals as their only signals).

Based on these findings, we can draw two main conclusions: First, the majority (over 90%) of coherence relations in discourse contain a signal (either DMs or other signals). Second, only a small proportion of the signalled relations (19.63%, 3,896 out of 19,847 signalled relations) are indicated by DMs while the majority of the signalled relations (88.51%, 17,567 out of 19,847 signalled relations) are indicated by means of different textual features other than DMs. This is a novel result since most studies in coherence relations have shown a low level of signalling (usually below 50%) due to the narrow focus on DMs as signals.

The second objective of the study was to examine what signals other than DMs are used to indicate coherence relations. We observed that relations are indicated by a wide variety of signals. In our corpus analysis, signals which are successfully identified as potential indicators of coherence relations belong to two broad classes: single signals and combined signals. The former includes eight types of signals other than DMs: *reference, lexical, semantic, morphological, syntactic, graphical, genre* and *numerical* features. The latter comprises six types of signals: (*reference* + *syntactic*), (*semantic* + *syntactic*), (*lexical* + *syntactic*), (*syntactic* + *semantic*), (*syntactic* + *positional*) and (*graphical* + *syntactic*). Then, the signal types are divided into specific signals. For instance, the *graphical* type includes specific signals such as *colon, parenthesis, dash* and *items in sequence* feature. Specific signals for a combined type such as (*semantic* + *syntactic*) include specific features such as (*lexical chain* + *subject NP*), (*repetition* + *subject NP*) and (*synonymy* + *subject NP*). Relations are also indicated by multiple signals. For example, an *Elaboration-additional* relation can be signalled by a *DM*, a number of *lexical chains* and a combined signal such as the (*reference* + *subject NP*) feature. Signalling by more than one signal of the same type is also very common. For example, a *Contrast* relation can be indicated by two *semantic* signals, such as *antonymy* and *lexical chain*. Even two different DMs (in a very few instances though) can be used to signal a single relation instance. For example, the *Disjunction* relation in the following example is signalled by two DMs, *or* and *alternatively*.

(16) [that would allow them to acknowledge that Sverdlovsk violated the 1972 agreement]N [or, alternatively, that would give U.S. specialists reasonable confidence that this was a wholly civilian accident.]N – Disjunction (wsj_1143: 78-79/80-81)

Finally, we evaluated the validity of the traditional classification of explicit and implicit relations. As mentioned in Section 2, coherence relations tend to be divided into two groups: explicit and implicit, based on the presence or absence of DMs. This classification hinges on the concept of signalling in discourse which considers DMs to be the main signals for coherence relations. Accordingly, it is postulated that the majority of relations are implicit because they do not contain a DM, depending on the corpus and text type under study. In our study and in our corpus, we address the explicit-implicit classification from a broader point of view. We have shown that the signalling of relations is achieved not only by the use of DMs, but mostly by means of signals other than DMs. These signals, including a wide variety of textual features, are omnipresent in discourse (or at least in our corpus), as they are used to signal the vast majority of relations.

In sum, our study shows that most coherence relations contain signals, sometimes more than one, and the signals are different kinds of textual devices, extending well beyond the category of DMs and including signals such as reference, graphical, genre or syntactic features. The fact that signals other than DMs are profusely distributed throughout a text suggests that readers or listeners while interpreting a relation make extensive use of other signals, in the absence of DMs or in addition to them wherever available.

We also examined the distribution of relations with respect to signal types, and explored the likelihood of the occurrence of certain relations for a given signal type. We computed the probabilities of finding particular relations signalled by a signal type within a relation group. We also computed the probabilities of finding particular relations within the population of all relations indicated by a particular signal type. We believe that this kind of information would be extremely useful in developing applications in computational discourse, and can be successfully deployed to automatically extract and label coherence relations.

Although a small proportion of relations in our corpus are not signalled, we observed that the lack of signalling often stems from technical issues such as errors in the original relational annotation, the presence of questionable coherence relations or the consideration of only immediate spans. Our findings thus suggest that there is a possibility for all relations in discourse to be signalled, and in this way, the findings also reinforce the psychological claim that there exist signals for every interpretable relation.

We would like to end the paper highlighting two clear applications of the RST Signalling Corpus. 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 investigated the role of DMs (or only a few signals) in the understanding of coherence relations (see Section 2 for references, and Das (2014) for more information). It would be very useful to extend this work by examining other types of signals, as found in the RST Signalling Corpus, to see what effects they have on comprehension. Extending this research is not trivial: As mentioned in Section 2, the manipulation of DMs (presence vs. absence) as practised in many psycholinguistic studies (e.g., Degand & Sanders, 2012) could lead to a change in the relational meaning. The caveat for such manipulation involving other signals is probably stronger. Since other signals are typically integral part of sentences, and primarily contribute to the propositional content or grammar of a sentence, removing or modifying such signals (such as lexical or syntactic) may result in significant changes in the propositional content or grammar of the sentences being compared. Thus, the experimental design would have to be more complex.

The other main application of such an annotated corpus is in discourse parsing. A great deal of recent work (da Cunha et al., 2012; Feng & Hirst, 2012, 2014; Hernault et al., 2011; Hernault et al., 2010; Joty et al., 2015; Maziero et al., 2011; Mithun & Kosseim, 2011; Perret et al., 2016) 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

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

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

Berzlánovich, I., & Redeker, G. (2012). Genre-dependent interaction of coherence and lexical cohesion in written discourse. *Corpus Linguistics and Linguistic Theory, 8*(1), 183-208.

Carlson, L., & Marcu, D. (2001). Discourse Tagging Manual: University of Southern California.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). RST Discourse Treebank, LDC2002T07. from https://catalog.ldc.upenn.edu/LDC2002T07

Corston-Oliver, S. (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, I., Juan, E. S., Torres-Moreno, J. M., Cabré, M. T., & Sierra, G. (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, R. (1991a). *Exploring the Role of Punctuation in the Signalling of Discourse Structure.* Paper presented at the the Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI, Technical University of Berlin.

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

Dancygier, B., & Sweetser, E. (2005). *Mental Spaces in Grammar: Conditional Constructions*: Cambridge University Press.

Das, D. (2012). *Investigating the Role of Discourse Markers in Signalling Coherence Relations: A Corpus Study*. Paper presented at the the Northwest Linguistics Conference, University of Washington, Seattle.

Das, D. (2014). *Signalling of Coherence Relations in Discourse.* (PhD dissertation), Simon Fraser University, Burnaby, Canada.

Das, D., & Taboada, M. (2013a). *Explicit and Implicit Coherence Relations: A Corpus Study.* Paper presented at the the Canadian Linguistic Association (CLA) Conference, University of Victoria, Canada.

Das, D., & Taboada, M. (2013b). *Signalling Subject Matter and Presentational Coherence relations in Discourse: A Corpus Study.* Paper presented at the 2013 LACUS Conference, Brooklyn College, Brooklyn, New York.

Das, D., & Taboada, M. (2014). *RST Signalling Corpus Annotation Manual*. Simon Fraser University. Available from: http://www.sfu.ca/~mtaboada/docs/RST_Signalling_Corpus_Annotation_Manual.pdf.

Das, D., & Taboada, M. (2017). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources & Evaluation*, 1-36.

Das, D., Taboada, M., & McFetridge, P. (2015). RST Signalling Corpus, LDC2015T10. from https://catalog.ldc.upenn.edu/LDC2015T10

Degand, L., & Sanders, T. (2002). The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing, 15*(7-8), 739-758.

Feng, V. W., & Hirst, G. (2012). *Text-level discourse parsing with rich linguistic features.* Paper presented at the the 50th Annual Meeting of the Association for Computational Linguistics.

Feng, V. W., & Hirst, G. (2014). *A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing.* Paper presented at the the 52th Annual Meeting of the Association for Computational Linguistics (ACL-2014), Baltimore, USA.

Fischer, K. (Ed.). (2006). *Approaches to Discourse Particles*. Amsterdam: Elsevier

Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A. K., & Webber, B. (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, B. (2009). An Account of Discourse Markers. *International Review of Pragmatics, 1,* 293-320.

Gaddy, M. L., van den Broek, P., & Sung, Y.-C. (2001). The influence of text cues on the allocation of attention during reading. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 89–110). Amsterdam/Philadelphia: Benjamins.

Georgakopoulou, A., & Goutsos, D. (2004). *Discourse Analysis: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

Grice, P. H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech Acts. Syntax and Semantics* (Vol. 3, pp. 41-58). New York: Academic Press.

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, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hasan, R. (1985). The texture of a text. In M. A. K. Halliday & R. Hasan (Eds.), *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* (pp. 70-96). Oxford: Oxford University Press.

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

Hernault, H., Prendinger, H., duVerle, D. A., & Ishizuka, M. (2010). HILDA: A discourse parser using Support Vector Machine classification. *Dialogue and Discourse, 1*(3).

Hobbs, J. (1979). Coherence and coreference. *Cognitive Science, 6*, 67-90.

Hobbs, J. (1985). On the Coherence and Structure of Discourse. Stanford, CA: CSLI.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form and use in context: Linguistic Implications* (pp. 11-42). Washington, DC: Georgetown University Press.

Joty, S., Carenini, G., & Ng, R. (2015). CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics, 41*(3), 385-435.

Kamalski, J. (2007). *Coherence marking, comprehension and persuasion: On the processing and representation of discourse*. Utrecht: LOT.

Kamalski, J., Lentz, L., Sanders, T., & Zwaan, R. A. (2008). The forewarning effect of coherence markers in persuasive discourse: evidence from persuasion and processing. *Discourse Processes, 45*, 546–579.

Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior, 23*, 115-126.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford, CA: Center for the Study of Language and Information.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of discourse comprehension and production. *Psychological Review, 85*(363-394).

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

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

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

Lapata, M., & Lascarides, A. (2004). *Inferring sentence-internal temporal relations.* Paper presented at the the North American Chapter of the Assocation of Computational Linguistics.

Lascarides, A., & Asher, N. (2007). Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure. In H. Bunt & R. Muskens (Eds.), *Computing Meaning* (Vol. 3, pp. 87–124).

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

Louis, A., Joshi, A., Prasad, R., & Nenkova, A. (2010). *Using Entity Features to Classify Implicit Discourse Relations.* Paper presented at the the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL'10.

Mak, W. M., & Sanders, T. J. M. (2012). The role of causality in discourse processing: effects on expectation and coherence relations. *Language and Cognitive Processes, 28*(9), 1414-1437.

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

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

Marcu, D., & Echihabi, A. (2002). *An unsupervised approach to recognising discourse relations.* Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA,.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

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

Maziero, E. G., Pardo, T. A. S., da Cunha, I., Torres-Moreno, J.-M., & SanJuan, E. (2011). *DiZer 2.0 – An Adaptable On-line Discourse Parser.* Paper presented at the the III RST Meeting (8th Brazilian Symposium in Information and Human Language Technology, Cuiaba, MT, Brazil.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288.

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

Meyer, T., & Webber, B. (2013). *Implicitation of Discourse Connectives in (Machine) Translation.* Paper presented at the the 1st DiscoMT Workshop at ACL 2013 (51th Annual Meeting of the Association for Computational Linguistics), Sofia, Bulgaria.

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

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

Mulder, G. (2008). *Undestanding causal coherence relations.* (PhD Dissertation), Utrecht University, The Netherlands.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language, 26*, 453-465.

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

O'Donnell, M. (2008). *The UAM CorpusTool: Software for corpus annotation and exploration.* Paper presented at the the XXVI Congreso de AESLA, Almeria, Spain.

Pardo, T. A. S., & Nunes, M. d. G. V. (2008). On the development and evaluation of a Brazilian Portuguese discourse parser. *Journal of Theoretical and Applied Computing, 15*(2), 43-64.

Perret, J., Afantenos, S. D., Asher, N., & Morey, M. (2016). *Integer linear programming for discourse parsing.* Paper presented at the NAACL-HLT, San Diego, CA.

Pitler, E., Louis, A., & Nenkova, A. (2009). *Automatic sense prediction for implicit discourse relations in text.* Paper presented at the 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.

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics, 30*(3), 309-363.

Polanyi, L., Culy, C., van den Berg, M., Thione, G. L., & Ahn, D. (2004). *A rule based approach to discourse parsing.* Paper presented at the the 5th SIGdial Workshop on Discourse and Dialogue, ACL, Cambridge, MA.

Prasad, R., 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 2008), Marrackech, Morocco.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. (2007). *The Penn Discourse Treebank 2.0 Annotation Manual*. The PDTB Research Group (University of Pennsylvania).

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics, 14*, 367 - 381.

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

Sanders, T., Land, J., & Mulder, G. (2007). Linguistic markers of coherence improve text comprehension in funtional contexts – on text representation and document design. *Information Design Journal, 15*(3), 219-235.

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

Sanders, T., & Spooren, W. (2007). Discourse and text structure. In D. Geeraerts & J. Cuykens (Eds.), *Handbook of Cognitive Linguistics* (pp. 916-941). Oxford: Oxford University Press.

Sanders, T., & Spooren, W. (2009). The cognition of discourse coherence. In J. Renkema (Ed.), *Discourse, of Course* (pp. 197-212). Amsterdam: Benjamins.

Sanders, T., Spooren, W., & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*, 1-35.

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

Scanlan, C. (2000). *Reporting and Writing: Basics for the 21st Century*. Oxford: Oxford University Press.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

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

Scott, D., & de Souza, C. S. (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.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hil.

Spooren, W. (1997). The processing of underspecified coherence relations. *Discourse Processes*, *24*, 149-168.

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

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

Subba, R., & Eugenio, B. D. (2009). *An effective discourse parser that uses rich linguistic information.* Paper presented at the HLT-ACL 2009, Boulder, CO.

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

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

Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse, 4*(2), 249-281.

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

Theijssen, D. (2007). *Features for automatic discourse analysis of paragraphs.* (MA Dissertation), Radboud University Nijmegen, The Netherlands.

Theijssen, D., van Halteren, H., Verberne, S., & Boves, L. (2008). *Features for automatic discourse analysis of paragraphs.* Paper presented at the 18th meeting of Computational Linguistics in the Netherlands (CLIN 2007).

Trabasso, T., & Sperry, L. L. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language, 24*, 595-611.

van der Vliet, N., & Redeker, G. (2014). Explicit and implicit coherence relations in Dutch texts. In H. Gruber & G. Redeker (Eds.), *The pragmatics of discourse coherence: Theory and Applications* (pp. 23-52). Amsterdam: Benjamins

Versley, Y. (2013). *Subgraph-based Classification of Explicit and Implicit Discourse Relations.* Paper presented at the the 10th International Conference on Computational Semantics (IWCS 2013), Potsdam, Germany.