# RST Signalling Corpus: A corpus of signals of coherence relations*

Debopam Das
Simon Fraser University
8888 University Drive
Burnaby, BC V5A 1S6 Canada
Email: ddas@sfu.ca
Phone: +1-778-628-8440

Maite Taboada
Simon Fraser University
8888 University Drive
Burnaby, BC V5A 1S6 Canada
Email: mtaboada@sfu.ca
Phone: +1-778-782-5585

## Abstract

We present the RST Signalling Corpus (Das et al., 2015), a corpus annotated for signals of coherence relations. The corpus is developed over the RST Discourse Treebank (Carlson et al., 2002) which is annotated for coherence relations. In the RST Signalling Corpus, these relations are further annotated with signalling information. The corpus includes annotation not only for discourse markers which are considered to be the most typical (or sometimes the only type of) signals in discourse, but also for a wide array of other signals such as reference, lexical, semantic, syntactic, graphical and genre features as potential indicators of coherence relations. We describe the research underlying the development of the corpus and the annotation process, and provide details of the corpus. We also present the results of an inter-annotator agreement study, illustrating the validity and reproducibility of the annotation. The corpus is available through the Linguistic Data Consortium (LDC), and can be used to investigate the psycholinguistic mechanisms behind the interpretation of relations through signalling, and also to develop discourse-specific computational systems such as discourse parsing applications.

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

## Acknowledgements

We are greatly indebted to the late Dr. Paul McFetridge for his invaluable contribution to this work. Dr. McFetridge was the Senior Supervisor of Debopam Das' PhD dissertation (Das, 2014). Sadly, he passed away on March 14, 2014, only a few months before the completion of the final version of the RST Signalling Corpus. He was a major driving force and a source of constant support for our work.

Funding for this research was provided by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant 261104-2008).

* Das, D. and M. Taboada (to appear) RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*. Current version: December 2016.

# 1 Introduction

One of the most intriguing issues in discourse analysis is how coherence relations (relations between propositions) are interpreted by readers or hearers, particularly when accompanied by underspecified signals and, most importantly, in the absence of any signal. Coherence relations are often signalled by discourse markers or DMs, which are generally considered to be the most typical (or sometimes the only type of) signals in discourse. DMs constitute a broad class of items used to indicate a link between propositions, and include conjunctions, prepositional phrases and lexicalized expressions (such as *all in all*). They are also referred to as connectives or cue phrases. Interpreting relations solely based on DMs seems to be problematic for two reasons: (1) DMs are used to indicate only a small number of relations in discourse, thereby leaving the majority of relations supposedly unsignalled; 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). If we believe that the successful interpretation of relations is essentially dependent on signalling, it is then possible that the phenomenon of signalling is not confined to the use of DMs alone, and relations can be indicated by signals which may extend beyond the category of DMs.

In this paper, we present the product of a signalling annotation effort, called the RST Signalling Corpus (Das et al., 2015). The corpus is annotated for signals of coherence relations, and is built over the RST Discourse Treebank (Carlson et al., 2002) which includes annotation of coherence relations. The RST Signalling Corpus uses the existing relations in the RST Discourse Treebank as its source data to which it adds relevant signalling information.

The RST Signalling Corpus has three significant features. First, the corpus includes annotation not only for DMs, but also for diverse types of other signals (signals other than DMs), such as reference, lexical, semantic, syntactic, graphical and genre features as potential indicators of coherence relations. Second, the corpus shows that the phenomenon of signalling is widespread in discourse as the overwhelming majority of relations (over 90%) in the corpus are found to be signalled, sometimes by multiple signals. Third, the significant majority of signalled relations (over 80%) in the corpus are indicated not by DMs, but by other signals.

The corpus has two main potential applications. It can be used in psycholinguistic studies to determine how readers or hearers use signals to identify relations, particularly those without DMs. It can also be used to develop discourse-specific computational systems, such as discourse parsing applications to automatically categorize coherence relations.

The paper is organized as follows: In Section 2, we outline the theoretical background underlying the development of the RST Signalling Corpus, including an introduction to the concept of coherence relations, Rhetorical Structure Theory and the signalling of coherence relations. In Section 3, a description of the research project (including relevant details on research motivation, hypotheses and methodology) is provided. Section 4 presents the annotation process, describing the annotation scheme, annotation tool and procedure, and providing an example of the signalling annotation from the corpus. Reliability of annotation is discussed in Section 5, illustrating the validity and reproducibility of the annotation. In Section 6, we provide the details of the corpus, with relevant statistics on extracted signals. Section 7 provides a brief account of a few recent corpus-based projects on signalling annotation. Finally, Section 8 summarizes the paper, and discusses potential applications of the corpus.

## 2 Coherence relations, Rhetorical Structure Theory and signalling

The perception of coherence in discourse is largely dependent on how well the text components are linked together. Coherence relations (also known as discourse or rhetorical relations) refer to the types of semantic or pragmatic connections that bind one discourse component to another. For example, consider the following text[1].

(1) New York City bonds have been beaten down for three straight weeks. On Friday, some issues fell nearly one point, or close to $10 for each $1,000 face amount. [wsj_671]

In Example (1) there are two components which are represented by the two sentences. These components are connected to each other by an *Evidence* relation, that is, the fact that New York City bonds have been beaten down is evidenced by the fall of some issues by significant values.

Coherence relations have been extensively investigated in the framework of Rhetorical Structure Theory or RST (Mann & Thompson, 1988) which is adopted as the theoretical framework of the present study. This is because it is a framework that we have worked with for years, and we believe best captures coherence relations. We chose it mostly, however, because the existing resource that we were using, the RST Discourse Treebank (Carlson et al., 2002), is annotated following the general RST principles.

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

Texts, according to RST, are built out of basic clausal units which are known as elementary discourse units or EDUs. EDUs constitute text spans that enter into rhetorical (or coherence) relations with each other, in a recursive manner. The authors of RST proposed that most texts can be analyzed in their entirety as recursive applications of different types of relations (Mann & Thompson, 1988). In effect, this means that an entire text can be analyzed as a tree structure, with clausal units being the branches and relations the nodes.

We provide below the RST annotation of a short text taken from an RST file in the RST Discourse Treebank (Carlson et al., 2002).

(2) President Bush insists it would be a great tool for curbing the budget deficit and slicing the lard out of government programs. He wants it now. [wsj_609]

The graphical representation of the RST analysis of this text in Example (2) is provided in Figure 1.

[1] Most of the examples in the paper were extracted from the RST Discourse Treebank (Carlson et al., 2002). The content inside the square brackets following an example refers to the file number in the RST Discourse Treebank from which the example has been taken.

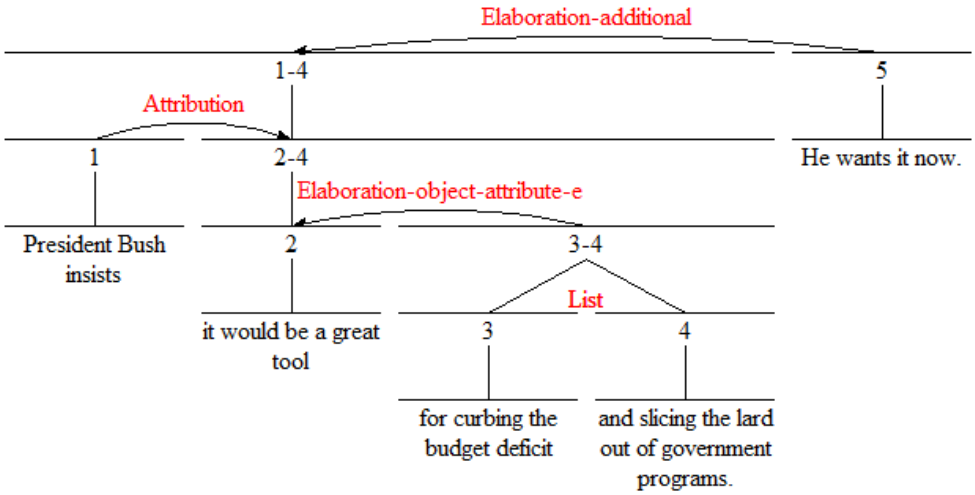


**Fig.1 RST analysis of the text in Example (**2**)**

The RST analysis shows that the text comprises five spans which are represented in the diagram in Figure 1 by the cardinal numbers, 1, 2, 3, 4, and 5, respectively. In the diagram, the arrow points to a span called the nucleus, and away from another span called the satellite. For a multinuclear relation, the nuclei are connected to each other by two or more straight lines (resembling the top part of a pyramid). Span 3 (nucleus) and Span 4 (nucleus) are connected to each other by a multinuclear *List* relation, and together they make the combined span 3-4[2]. Span 3-4 (satellite) is connected to span 2 (nucleus) by an *Elaboration* (more specifically, *Elaboration-object-attribute-e*) relation, and together they make the combined span 2-4. Then, span 1 (satellite) is linked to span 2-4 by an *Attribution* relation, and together they make a combined span 1-4. Finally, span 5 (satellite) is connected to span 1-4 (nucleus) by an *Elaboration-additional* relation.

The introduction to RST presented here is quite brief. More detail is available in the original RST publication (Mann & Thompson, 1988), and in a later review of developments since the original publication (Taboada & Mann, 2006a, 2006b).

The most typical signals of coherence relations are discourse markers or DMs. Schiffrin defines DMs as "sequentially dependent elements which bracket units of talk" (Schiffrin, 1987: 31), and notes that DMs "add to discourse coherence" (Schiffrin, 1987: 326). DMs, According to Schiffrin, include verbal expressions (such as *and, because, but, I mean, oh, so, then, well* and *y'know*) as well as paralinguistic features and non-verbal gestures. Our treatment of DMs, however, is primarily motivated by the notion of discourse coherence, as outlined in Halliday and Hasan (1976). We are interested in how DMs contribute to discourse coherence by signalling relations in text. That is why our definition of DMs is mainly based on the definition provided by Fraser (1999, 2006, 2009), that is, DMs are lexical expressions (*and, if, since, thus*, etc.) which belong to different syntactic classes, such as conjunctions, adverbials and prepositional phrases. Furthermore, DMs are used to connect discourse components, and they signal the coherence relations that hold between those components. Consider the following example.

(3) A country is considered financially healthy **if** its reserves cover three months of its imports. [wsj_1391]

[2] A combined span comprises two or more spans (Elementary Discourse Units, or EDUs), and is represented by the starting span and the ending span, with a hyphen between them.

In the short text in Example (3), the two discourse components are the two sentences which are connected to each other by the DM *if* which signals a *Condition* relation between them.

Traditionally, coherence relations are believed to be signalled only by DMs, and accordingly, relations, based on the presence or absence of DMs, are classified into two groups, explicit and implicit relations, respectively (Knott & Dale, 1994; Martin, 1992; Meyer & Webber, 2013; Renkema, 2004; Taboada, 2009; Taboada & Mann, 2006b; Versley, 2013). Explicit relations contain DMs or are signalled by DMs. For instance, the relation in example (3) above will be considered to be explicit since it is signalled by the DM *if*. Implicit relations, in contrast, are not signalled by DMs, and thereby, they remain (supposedly) unsignalled. For instance, consider the following (invented) example.

(4) John is tall. Mary is short.

In Example (4), the discourse components are the two sentences which are connected to each other by a *Contrast* relation. Traditionally, this relation will be considered to be an implicit relation since it does not contain a DM, or it is not signalled by a DM. In more recent research, coherence relations are also considered to be signalled by lexical means (other than DMs) such as indicative phrases (*quite the contrary*, *what's more*, *that would follow*, etc.) known as Alternative Lexicalization (AltLex) in the Penn Discourse Treebank (Prasad et al., 2008). Research into the relationship between discourse markers, alternative lexicalizations and coherence relations is flourishing, and an important pan-European network is currently devoted to studying that relationship[3].

# 3 The RST Signalling Corpus project

## 3.1 Motivation and hypothesis

Although DMs are often used to signal coherence relations, they actually account for only a small number (usually between 20% and 50%) of relations present in discourse, as documented in a number of corpus studies (Prasad et al., 2007; Renkema, 2009; Taboada, 2006). The fact that relations without DMs are omnipresent in discourse raises an important question: How are coherence relations recognized in the absence of DMs? Psycholinguistic research shows that coherence relations are recognized in the process of text comprehension, and that they contribute to differences in reading time and comprehension effects (Knott & Sanders, 1998; Mak & Sanders, 2013; Sanders & Spooren, 2007, 2009; Sanders et al., 1992, 1993). Furthermore, relations are recognized even when no DMs are present (Kamalski, 2007; Mulder, 2008; Mulder & Sanders, 2012; Murray, 1995; Sanders & Noordman, 2000). 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 the relatively infrequent DMs.

There is also a considerable number of studies in computational linguistics which have investigated the signalling of coherence relations by signals other than DMs. In these studies, various linguistic and textual features have been used to identify the presence and type of relations. Some of these features include tense or mood (Scott & de Souza, 1990), anaphora and deixis (Corston-Oliver, 1998), lexical chains (Marcu, 2000), punctuation and graphical markers

[3] TextLink, Structuring Discourse in Multilingual Europe, COST Action IS1312, http://textlink.ii.metu.edu.tr/

(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).

Building on the psychological assumption, and considering the success of the above-mentioned computational studies, we hypothesize that the signalling of coherence relations is not confined to the use of DMs alone. For this reason, we use the more general term *signalling*, rather than *marking*, to indicate that signalling can be carried out by means other than discourse markers alone. There exist a wide variety of textual signals other than DMs, such as *lexical*, *semantic*, *syntactic*, *graphical* and *genre* features, which are frequently used to convey coherence relations. We argue that the *Contrast* relation in Example (4) in Section 2 is actually signalled, even though it is not signalled by a DM. The relation is indicated by two types of other signals. One can notice that the two discourse components (or two sentences) in the text share a parallel syntactic construction (subject–copula–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.

Furthermore, we also hypothesize that every relation in 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. This study is unique in that it includes a wide range of signals, which have not previously been annotated for the same corpus.

### 3.2 Selection and description of source corpus

One of the research objectives of our study is to discover as many signals of coherence relations as possible. Accordingly, our research design required a large database from which a considerably large number of tokens (representing relational signals) can be extracted. Thus, we specifically looked for a large corpus in which coherence relations are already annotated, and from which additional information about signalling of relations can be extracted.

We choose to use the RST Discourse Treebank or RST-DT (Carlson et al., 2002) as our source of data, first of all because it is already annotated for coherence relations, based on Rhetorical Structure Theory. Additionally, the RST-DT annotations include all levels of discourse. The RST-DT, unlike other resources like the Penn Discourse Treebank (Prasad et al., 2008), 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.

The RST-DT contains a collection of 385 Wall Street Journal articles (representing over 176,000 words of text) selected from the Penn Treebank (Marcus et al., 1993). The corpus is distributed by the Linguistic Data Consortium (LDC)[4], from which the corpus can be downloaded (for a fee). The articles chosen for annotation in the RST-DT cover a variety of topics, such as financial reports, general interest stories, business-related news, cultural reviews, editorials and letters to the editor. The texts in these articles are annotated manually by a group of annotators. 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 tree diagrams.

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 includes 20,123 relations in total, employing a large set of 78 relation types which are divided into 16 major relation groups. The (concise) taxonomy of RST relations in the RST-DT is provided 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* and *Same-Unit*, were used in the annotation of the RST-DT in order to impose certain structure-specific requirements on the discourse trees. More information on the detailed taxonomy of relations and relation definitions can be found in the RST-DT annotation manual (Carlson & Marcu, 2001).

# 4 Annotation process

## 4.1 Annotation scheme

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

The most important aspect of the signalling annotation task is to select and classify the types of signals to annotate. There are mainly three ways to build resources of discourse signalling. First, one can manually build a repository of signals. For instance, Stede and Umbach (1998) developed a dictionary of German and English DMs called DIscourse Marker LEXicon (DiMLex), compiling entries from available sources such as standard dictionaries and grammars; Alonso et al. (2002) gathered a set of 577 DMs in Spanish from previous work in addition to a corpus study, and provided a data-driven classification of those DMs using clustering techniques; Roze et al. (2012) constructed LEXCONN, a French lexicon of DMs manually extracted from a corpus. Second, one can automatically infer a list of markers from existing (manually-annotated) discourse corpora: Al-Saif and Markert (2010) collected a set of 107 Arabic DMs using machine learning algorithms to identify DMs along with the relations they convey. Third, one can use available corpora in a source language to automatically build a signalling lexicon in a target language, as in Meyer and Webber (2013).

In our project, we started with the first strategy, i.e., manually building the repository of relational signals, and then followed the second strategy, i.e., extracting more signals from the corpus. First, we built our taxonomy of signals based on the different classes of relational signals that have been mentioned in previous studies on signalling in discourse (Blakemore, 1987, 1992, 2002; Fraser, 1990, 1999, 2006, 2009; Halliday & Hasan, 1976; Knott, 1996; Le Thanh, 2007; Lin et al., 2009; Marcu, 1999; Polanyi et al., 2004; Prasad et al., 2010; Sanders et al., 1992, 1993; Schiffrin, 1987, 2001). Second, we added to the taxonomy more signals identified in our preliminary corpus work (Das, 2012; Das & Taboada, 2013; Taboada & Das, 2013). Signals found by using these strategies include not only DMs but also other textual signals such as reference, semantic and syntactic features.

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 are only illustrative, not exhaustive. The complete taxonomy is provided in Table 6 in Appendix. More details on the taxonomy can be found in the annotation manual for the corpus (Das et al., 2015).

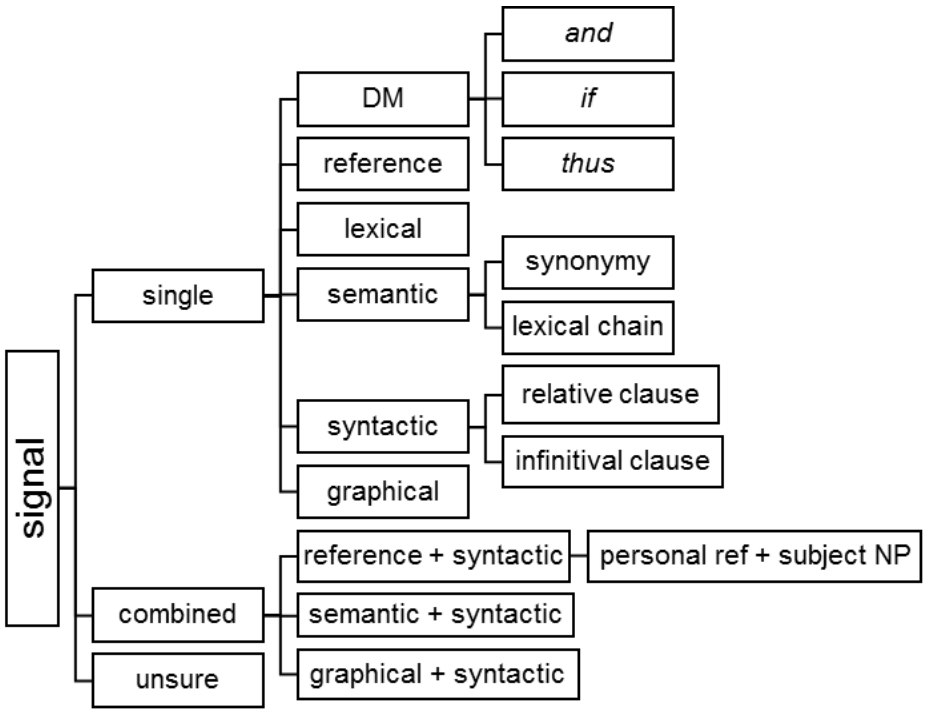


**Fig.2 Hierarchical taxonomy of signals (fragment)**

A *single* signal is made of one (and only one) feature used to indicate a particular relation. Consider the following examples from the RST-DT[5]. In Example (5) below, the DM *although* is a single signal, and is used to indicate the *Antithesis* relation.

(5) [Although Larsen & Toubro hadn't raised money from the public in 38 years,]S [its new owners frequently raise funds on the local market.]N – Antithesis [wsj_629: 142-143]

In Example (6), the *Contingency* relation[6] is indicated by a lexical signal, the indicative word *contingent*, which represents a single signalling feature.

(6) [Iran's President Rafsanjani offered to help gain freedom for Western hostages in Lebanon,]N [but said the assistance was contingent on U.S. aid in resolving the cases of three Iranians kidnapped in Lebanon in 1982 or the release of frozen Iranian assets.]S – Contingency [wsj_1353: 77-82]

The *Purpose* relation in Example (7) is signalled by a syntactic feature, the *infinitival clause* (underlined), which is also a single signal[7].

[5] Conventions for interpreting examples from the RST-DT: The text within square brackets denotes a span. Each pair of square brackets is followed by either the uppercase character N, referring to the nucleus span, or the uppercase character S, referring to the satellite span. A pair of two spans (N and S, or N and N) is respectively followed by a dash and the name of the relation that holds between the spans. The square brackets at the end contain the file number of the source document, and the location of the relation in the document. The signal under discussion is underlined.

[6] The Contingency relation is also signalled by the DM *but* here.

[7] We chose to use the entire infinitival clause as the relevant signal rather than the infinitive particle *to*, as it can be confused with the preposition *to*.

(7) [To encourage more competition among exporting countries,]S [the U.S. is proposing that export subsidies, including tax incentives for exporters, be phased out in five years.]N – Purpose [wsj_1135: 54-58]

A c*ombined* signal[8], on the other hand, comprises two single signals or features (other than DMs) which work in combination with each other to indicate a particular relation. Consider the following example from the RST-DT. In Example (8), two types of single signals, a *reference* feature and a *syntactic* feature, are operative together in signalling the *Elaboration-additional* relation. The reference feature indicates that the word *He* in the satellite span is a personal pronoun because it refers back to Gerald C. Beddall, an entity mentioned (or introduced) in the nucleus span. Syntactically, the personal pronoun *He* is also in the subject position of the sentence the satellite span starts with, representing the topic of the *Elaboration-additional* relation. Therefore, the combined signal, comprising the *reference* and *syntactic* features – in the form of a *personal reference* plus a *subject NP*, represented as (*personal reference + subject NP*) – functions here as a signal for the *Elaboration-additional* relation.

(8) [Gerald C. Beddall, 47 years old, was named president of the Clairol division of this pharmaceuticals and health-care company.]N [He succeeds C. Benjamin Brooks Jr.,…]S – Elaboration-additional [wsj_1341: 3-8]

Finally, *unsure* refers to those cases in which no potential signals are found or specified, as represented in Examples (9) and (10).

(9) [This hasn't been Kellogg Co.'s year.]S [The oat-bran craze has cost the world's largest cereal maker market share.]N – Cause [wsj_610: 1-2]

(10) ["This is a democratic process]N [-- you can't slam-dunk anything around here."]N – Consequence [wsj_1963: 33-34]

Relations can also be indicated by multiple signals (by more than one signal), as can be seen in example (4) in Section 2. 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*), as in Example (8) above, 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, every signal functions independently and separately from each other, but they all contribute to signalling the relation. For example, an *Elaboration(-additional)* relation with multiple signals (like the one in Table 2, Section 4.4) can be indicated by multiple signals, such as a genre feature (e.g., *inverted pyramid scheme*) and a semantic feature (e.g., *lexical chain*). The signals do not have any connection, but they separately signal the relation.

[8] A combined signal is represented within parentheses, including two features conjoined by the '+' symbol. For example, a combined signal, containing feature 1 and feature 2, is represented in the following form: (feature 1 + feature 2).

### 4.2 Annotation tool

We use UAM CorpusTool (O'Donnell, 2008), a software for text annotation, for performing our signalling annotation task. When choosing UAM CorpusTool as the means of annotation, we considered the following requirements, based on Dipper et al. (2004).

(1) **Importability:** In our annotation task, we need to select individual relations from the RST-DT, and tag them with appropriate labels of signalling information. This requirement specifies that the text to be annotated must be imported into the annotation tool along with the relevant relational information, in the original LISP, in XML, or in similar format that allows for visualization of the relations.

(2) **Annotation scheme:** The signal tags in our taxonomy are organized hierarchically in three levels (see Figure 2). To annotate each relation, we need to record the signalling information for all these three levels. This requirement necessitates that the annotation tool supports the structure of the taxonomy.

(3) **Customizability:** This requirement implies a convenient access to edit or modify the annotation tags while continuing with the annotation.

(4) **Multiple annotations:** In our annotation, a single relation can be indicated by more than one signal, and the annotation of that relation accordingly requires the attachment of two or more sets of signalling tags to a single relation.

(5) **Convertibility:** The output of the annotation should be stored in XML-based format which would facilitate the reuse of the existing annotation, standardization of annotation format and application of other tools to the same data.

(6) **Simplicity:** This requirement specifies that the users of the tool should be able to use it even without having advanced computational knowledge about the inner workings of the tool. In addition, the tool should also have a graphical user interface which can provide adequate visualization of both the source data and annotated data.

With respect to the annotation scheme requirements, UAM CorpusTool allows us to create a hierarchically-organized tagging scheme (including all three categories of signals: signal class, signal type and specific signal). It also provides multiple annotations for a single element. In terms of simplicity, UAM CorpusTool is primarily aimed at those users with little or no prior computational knowledge. It also 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 LISP format to a simple text file format. This allows 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. For all these reasons, we choose UAM CorpusTool for our annotation task.

We used the 2.8.12 version of UAM CorpusTool to perform our signalling annotation in the RST Signalling Corpus. The annotations are also accessible using later versions of UAM CorpusTool, usually by importing the CorpusTool file into the new version.

### 4.3 Procedure

In our signalling annotation, we performed 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. The annotation was carried out by the first author, after a reliability study showed good inter-annotator reliability (see Section 5). We used the taxonomy of signals presented in Figure 2 in Section 4.1 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 practice, we annotated 21,400 relations in total. This number is higher than the number of relations in the corpus (20,123). This is because in the RST-DT some multinuclear relations, mostly *List*, *Contrast* and *Sequence*, often occur with more than two nuclei, and these relations (with three or more nuclei) are considered to be single relations. For example, the following *List* relation in the RST-DT (file number: wsj_1369) Figure 3 represents such a situation.

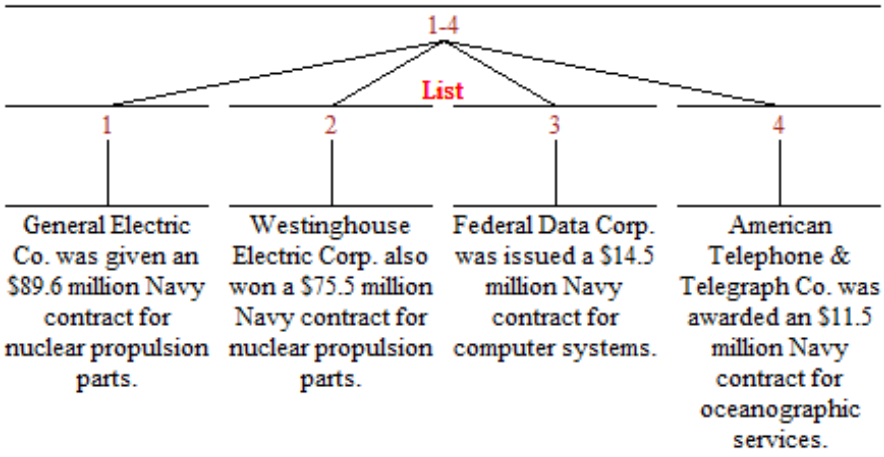


**Fig.3 Example of a multinuclear relation with more than two nuclei**

In this *List* relation, there are four nuclei, connected to each other by a single relation. However, in our signalling annotation we consider such a relation to be a number of binuclear relations (relation with two nuclei). In other words, we divide a multinuclear relation containing three or more nuclei into a series of two or more binuclear relations in which every linear pair of adjacent spans are considered to be linked by a distinct binuclear relation. For example, we divide the *List* relation in Figure 3 into *three* single *List* relations, respectively holding between span 1 and span 2, then between span 2 and span 3, and finally, between span 3 and span 4. We take this approach so that we can focus on signals. It is more difficult to reliably identify signals when looking at a relation holistically, as the entire unit will typically contain multiple links across spans. We encountered numerous instances of such relations in the corpus, and this means that our annotation includes a higher number of relations (21,400 relations) than that (20,123 relations) originally in the RST-DT.

In the annotation process, we imported the RST files (in a text file format, converted from the LISP format) into UAM CorpusTool. The visualization window of UAM CorpusTool shows the existing relational annotations, including the RST-segmented texts and the names of the relations holding between text spans. To tag a particular relation instance, we selected the name of the relation, and then chose from the annotation scheme (the taxonomy of signals is incorporated into the tool before the annotation starts) the appropriate set of signalling tags (organized into three levels: *signal class*, *signal type* and *specific signal*) in order to assign signalling information to that relation. If the relation contained more signals, we selected the relation again (and again, if necessary) and re-did the above-mentioned steps. A snapshot of the annotation window in UAM CorpusTool is provided in Figure 4, where the unit highlighted, the Purpose relation has been annotated as having a single signal, of the *syntactic* type, and, more specifically, *an infinitival clause* (the clause *to produce 12 low-budget movies a year*).

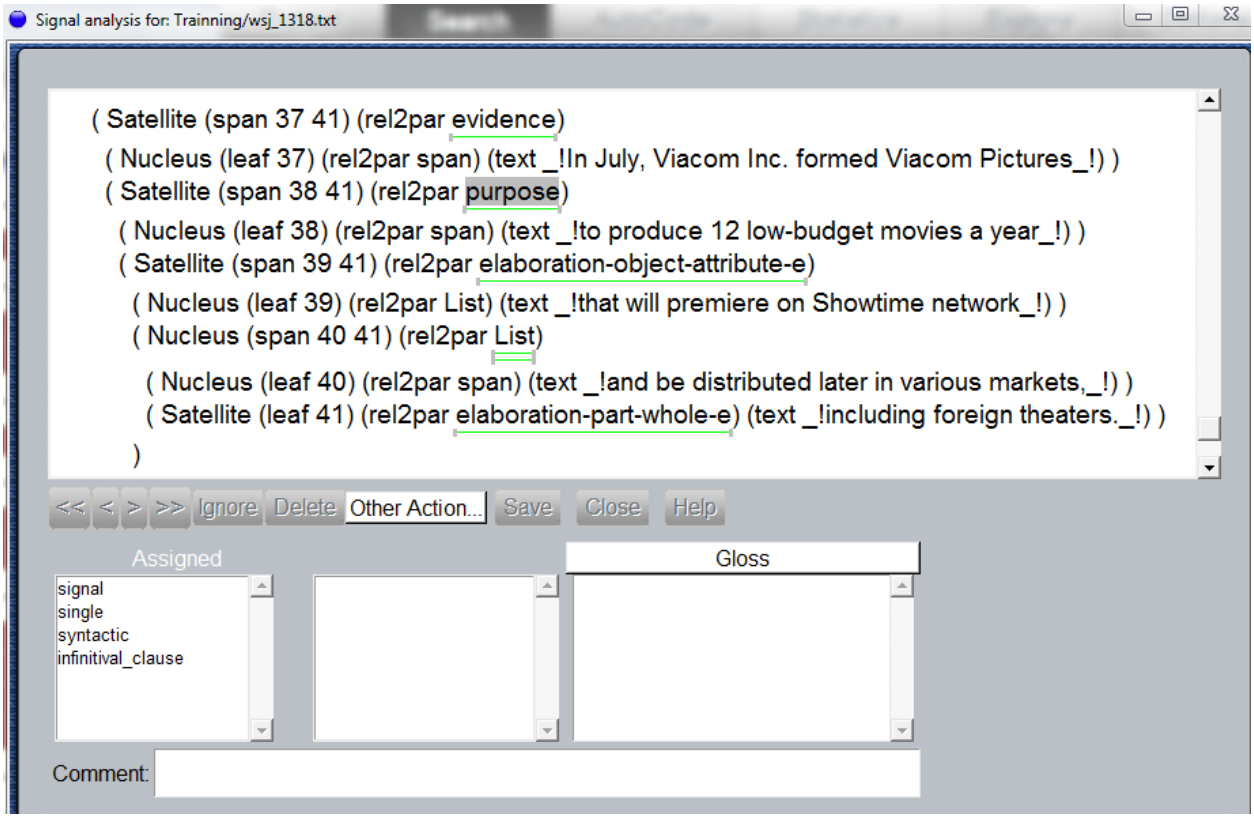


**Fig.4 Signalling annotation in UAM CorpusTool**

The text box on top in the annotation window in Figure 4 shows the original RST annotation (imported from the RST-DT) in a text format roughly following a LISP-style bracketing structure. It represents the relevant relational information, including the text spans (text segments functioning as the discourse units), span numbers, span status (nucleus or satellite) and relation names (e.g., *evidence*, *purpose*, *elaboration-object-attribute-e*). The indented text structures (along with the nested parenthetical structures) represent the hierarchical discourse structures which are typical of an RST annotation (see Section 2). The relation names are underlined as they are (manually) selected and tagged for signalling annotation. The number of underlines beneath a relation name corresponds to the number of signals annotated for that relation. For example, in the top text box the *evidence* or *purpose* relation with a single underline has only one signal while the *list* relation has two underlines which signify that the relation is indicated by two signals and has been annotated twice. The bottom-left box called 'Assigned' in the annotation window shows the signalling information for individual relations across four levels: signal (the generic category), signal class, signal type and specific signals. When the snapshot was taken, the underline beneath the *purpose* relation (highlighted in grey color) in the top text box was selected. Accordingly, the 'Assigned' box represented the signalling annotation for the relation. The labels in the box show that the relation is indicated by a *single* signal which is of the *syntactic* type, and the specific signal is an *infinitival clause*. For the *list* relation with two underlines, clicking on each underline will show in the 'Assigned' box one of the two signals annotated for that relation. The same kind of signalling information for other relations can similarly be accessed by clicking on the corresponding underlines.

## 4.4 An example of signalling annotation

We provide the annotation of a short RST file from the RST-DT with signalling information. The file contains the following text.

(11) 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 RST analysis of the text in Example (11) using RSTTool (O'Donnell, 1997) is provided in Figure 5.

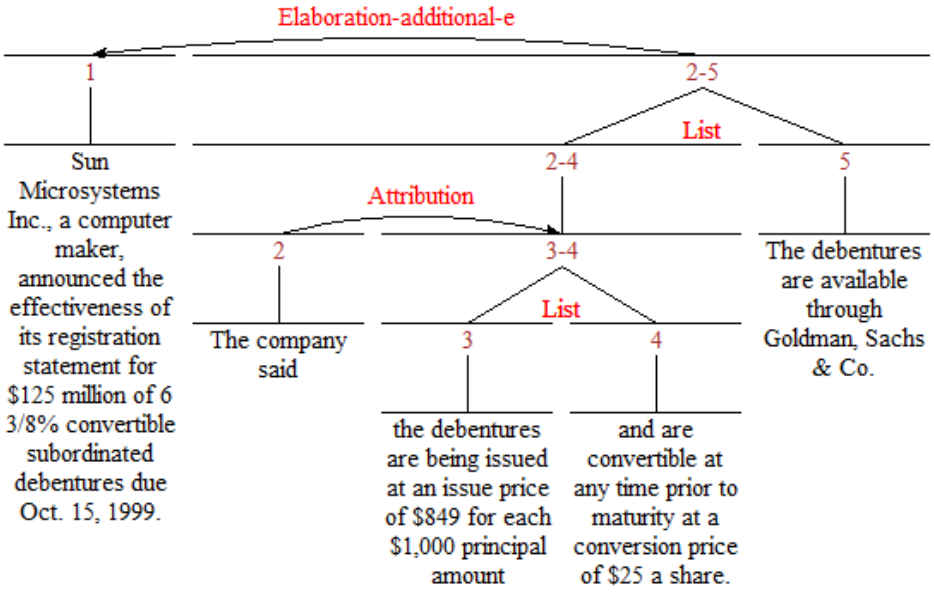


**Fig 5. Graphical representation of an RST analysis, text in Example (11)**

The RST analysis shows that the text comprises five spans which are represented in the diagram (in Figure 5) by the numbers, 1, 2, 3, 4 and 5, respectively. In the diagram, span 3 (nucleus) and span 4 (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 (nucleus) by an *Attribution* relation, and together they make the combined span 2-4. Then, a multinuclear *List* relation holds between spans 2-4 (nucleus) and 5 (nucleus), and together they make the combined span 2-5. Finally, span 2-5 (satellite) is connected to span 1 (nucleus) by an *Elaboration* (more specifically, *Elaboration-addition-e*) relation.

We annotate the relations in the text with appropriate signalling information, as presented in Table 2. A detailed description is provided in Table 6 in the Appendix.

| File | N | S | Relation | Signal type | Specific signal | Explanation: How signalling works |
|---|---|---|---|---|---|---|
| wsj_650 | 1 | 2-5 | Elaboration-additional | 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. |
| | | | | semantic | 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. |
| | 2-5 | | 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 (in Table 2), the *Elaboration* 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 he/she is reading a newspaper text, 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 being its 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). Our definition of genre is very informal here, but recent research explores the connection between genre/register and rhetorical relations (Matthiessen, 2015; Matthiessen & Teruya, 2015). 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*. The words such as *issued* and *convertible* (in the first nucleus) and words *debentures* and *available* (in the second nucleus) are semantically related, indicating (perhaps loosely) a *List* relation between the spans.

## 5 Reliability of annotation

Our list of signals and the annotation procedure were agreed upon after several iterations of the taxonomy, and after adding more signals when our initial analysis revealed more than what we had originally listed.

In order to check the validity and reproducibility of our initial annotation and original taxonomy, we conducted a reliability study. We selected 130 relations from two files in the RST-DT, after ensuring that the relation sample is representative of the overall distribution of the entire collection of the relations in the corpus. The two authors annotated the 130 relations independently, and then compared the annotations. 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, as provided in Figure 2 in Section 4.1. We concentrated on whether we agreed on the type of signal, not necessarily on where it is conveyed in the text (e.g., for a lexical chain, we annotated *semantic*, but not what words or phrases are involved in the chain).

Our original taxonomy of signals evaluated in the reliability study included nine types of single signals[9]. The description of these signals is provided in Table 3. A more detailed description of these signals can be found in Taboada and Das (2013).

[9] The original taxonomy also included ten types of combined signals. See Taboada and Das (2013) for more information on our pilot study.

| # | Signal type | Description |
|---|---|---|
| 1 | DM | DMs are lexical expressions (e.g., *and, if, since, then*) which are primarily drawn from different syntactic categories, such as conjunctions, adverbials and prepositional phrases. DMs connect discourse segments, and they signal a coherence relation between those segments. |
| 2 | Entity | Entity features include links where entities, similar or dissimilar, help interpret the relation. Entities are of different types, such as *given entity*, *different entities* and *mutually exclusive entities*. |
| 3 | Genre | Genre features guide the interpretation of relations when a particular genre is well known to the reader. In the case of the newspaper genre (which all the texts in the corpus belong to), it is common to start the text with the most important information, and to continue with additional details. This results in *Elaboration* relations, with the nucleus being the first sentence or paragraph, and the rest of the article acting as a satellite that expands on the nucleus or the beginning part of the text. |
| 4 | Graphical | Graphical and other punctuation features, such as lists and headings, and other forms of layout are sometimes indicators of a relation. |
| 5 | Lexical | Lexical features include the use of indicative words and phrases, such as individual words that indicate a relation, for example, the verbs *concede* and *cause* for *Concession* and *Cause* respectively. |
| 6 | Morphological | Among morphological features, *tense* is the most prominent one, indicating *Temporal* relations or *Circumstance* relations, as is the case of some instances of non-finite verbs. |
| 7 | Numerical | Numerical elements are present in *List* relations, but also in more subtle ways, when an *Elaboration* consists of providing a general word (in this case, digit(s) or number(s)) and then listing the contents of that word. |
| 8 | Semantic | A semantic feature has two components, each belonging to one of the spans. The components are in a semantic relationship with each other, such as *synonymy*, *antonymy* and *lexical chain*. |
| 9 | Syntactic | At the syntactic level, there are a host of constructions that help identify a relation. From word order, such as subject-verb inversion for *Condition* (*Had he known…*) to sentence mood, such as the use of interrogatives to signal *Solutionhood*. |

**Table 3. Description of signals in the original taxonomy**

We used Cohen's Kappa (Siegel & Castellan, 1988) to calculate the agreement value, with nominal data representing the aforementioned nine categories in our classification, plus an additional category *unsure* (used to indicate the situations in which the annotators did not find any identifiable signal). The confusion matrix representing the agreements and disagreements between the two annotators for 130 instances of signalling annotation is provided in Table 4.

| | | A2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DM | ent | genre | graph | lex | morph | numer | sem | syn | nosig | Total |
| A1 | DM | 25 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 26 |
| | ent | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| | genre | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | graph | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| | lex | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 1 | 5 |
| | morph | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | numer | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | sem | 1 | 14 | 0 | 0 | 0 | 2 | 0 | 7 | 0 | 6 | 30 |
| | syn | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 53 | 3 | 56 |
| | no sig | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 4 |
| | Total | 28 | 20 | 0 | 4 | 4 | 2 | 0 | 7 | 54 | 11 | 130 |
| Note: A1 = Annotator 1; A2 = Annotator 2 | | | | | | | | | | | | |

**Table 4. Confusion matrix of agreements and disagreements between two annotators**

The unweighted and weighted kappa values for our reliability study are 0.67 and 0.71, respectively, which indicate 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.

The distribution in Table 4 also shows that agreements are higher for signals such as *DM* and *syntactic* type. This is expected, since both these signal types include the most categorical types of signals. *DMs* are very prominent signals since they are more or less fixed lexical expressions, and occur mostly at the beginning of a text span. The s*yntactic* type also includes a variety of unambiguous types of constructions such as relative clause, participial clause or reported speech which are easy to distinguish.

On the other hand, signals for which the annotators mostly disagree include *entity* and *semantic* types. It is also seen that one of the types is often chosen for the other, that is, while one annotator selects *entity* as the relevant signal for a certain relation, the other annotator annotates it as being *semantic*. This leads us to review the definitions of these signals, and upon closer inspection we observe that many of the attributes of *entity* and *semantic* features actually overlap. Initially, we had reserved the category *entity* for those signals that involved reference to the same referent. The category *semantic* was reserved for semantic relations that do not necessarily involve same reference, such as synonymy. This distinction works along the lines of Halliday and Hasan's (1976) grammatical versus lexical cohesion, with *entity* signals being close to the reference system in Halliday and Hasan's grammatical cohesion. Our *semantic* group of signals contains lexical cohesion relations, such as synonyms, antonyms and hypernyms. The problem, however, is that lexical cohesion also includes repetition of the same item which is, strictly speaking, reference to the same referent, and thus *entity* in our system. As a solution to this problem, in our final annotation we substituted the *entity* type with a new type, *reference*, drawn from Halliday and Hasan (1976), with the latter exclusively represented by pronouns and other referential expressions. The *semantic* type, on the other hand, is kept apart for identifying semantic relationships which are represented by devices of lexical cohesion, and not by pronouns and referential expressions.

In closing this section, we would like to point out that reliability studies are not necessarily reliable (Taboada and Das, 2013). We call into question whether conducting an inter-annotator agreement study, like the one we have just described, is sufficient to provide a stamp of approval for the project. Discourse annotation is inherently subjective, because many of the decisions rely of interpreting the text, or re-interpreting what the author meant. Additionally, agreement studies are carried out by members of the project, or by people trained by members of the project, because it is hardly ever feasible to find an outside expert willing or able to perform annotations. We believe that our annotations are reliable, and that, given enough time and resources, we could train somebody else to perform annotations that are comparable to ours.

## 6 Details of corpus

The RST Signalling Corpus was released on June 15, 2015 through the Linguistic Data Consortium or LDC (https://www.ldc.upenn.edu/) under the authorship of Debopam Das, Maite

Taboada, and Paul McFetridge, and is downloadable from the following URL: https://catalog.ldc.upenn.edu/LDC2015T10 (for a fee as a single user, or free to LDC members). The downloadable version includes the corpus, the annotation manual and the relevant publications.

The 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 detailed distribution of signals in the corpus is provided in Table 5.

| # | Signal class | Signal type | Specific signal | # of tokens | Total | % |
|---|---|---|---|---|---|---|
| 1 | single | DM | *and*, *but*, *if*, *since*, *then*, etc. | 3,909 | 3,909 | 13.34% |
| | | reference | personal reference | 260 | 586 | 2.00% |
| | | | demonstrative reference | 134 | | |
| | | | comparative reference | 182 | | |
| | | | propositional reference | 10 | | |
| | | lexical | indicative word | 1,399 | 1,440 | 3.89% |
| | | | alternate expression | 41 | | |
| | | semantic | synonymy | 38 | 7,265 | 24.80% |
| | | | antonymy | 37 | | |
| | | | meronymy | 34 | | |
| | | | repetition | 1,405 | | |
| | | | indicative word pair | 19 | | |
| | | | lexical chain | 5,700 | | |
| | | | general word | 29 | | |
| | | morphological | tense | 313 | 313 | 1.07% |
| | | syntactic | relative clause | 1,621 | 8,723 | 29.77% |
| | | | 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 | | |
| | | graphical | colon | 222 | 1,014 | 3.46% |
| | | | semicolon | 20 | | |
| | | | dash | 273 | | |
| | | | parentheses | 247 | | |
| | | | items in sequence | 252 | | |
| | | genre | inverted pyramid scheme | 720 | 943 | 3.22% |
| | | | newspaper layout | 189 | | |
| | | | newspaper style attribution | 26 | | |
| | | | newspaper style definition | 8 | | |
| | | numerical | same count | 26 | 26 | 0.09% |
| 2 | combined | (reference + syntactic) | (personal reference + subject NP) | 504 | 544 | 1.86% |
| | | | (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 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | (semantic + syntactic) | (synonymy + subject NP) | 22 | 2,155 | 7.36% |
| | | | (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% |
| | | (syntactic + positional) | (past participial clause + beginning) | 41 | 69 | 0.23% |
| | | | (present participial clause + beginning) | 28 | | |
| | | (graphical + syntactic) | (comma + present participial clause) | 216 | 226 | 0.77% |
| | | | (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**

Although the majority of relations present in the corpus are signalled or contain at least one signal, there are 1,553 relations (7.26% of the 21,400 relations) for which no signals were found. There are four different reasons why we believe no signals could be found for these relations. First, in some cases we observed that there were errors in the original relational annotation in the RST-DT, many of which emerge from the incorrect assignation of relation labels. In a number of cases, we found that a relation was postulated by the annotators of the original corpus, whereas we would not have annotated a relation, or we would have proposed a different one. For example, *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 where the relation holds. We noticed, however, that the interpretation of a relation does not always depend on the recognition of signals from the corresponding relation spans, but 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 some cases, we had a sense that the relation was clear, but it was very difficult to 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.

Thus, our claim that the vast majority of relations are signalled *by some means* still holds. Even in the relatively small percentage of cases (7.26%) where we could not annotate a signal, our intuition is that some signalling is present, but such that reliable annotation is not possible.

More information about the definitions of signals and the statistical distribution of relations and their signals in the RST Signalling Corpus can be found in Debopam Das' PhD dissertation "Signalling of Coherence Relations in Discourse" (Das, 2014). The dissertation is available through the Simon Fraser University library, and is downloadable from the following URL: http://summit.sfu.ca/item/14446. Supplementary material for the dissertation includes the complete distribution of relations and their signals, and is available from the following URL: http://www.sfu.ca/~mtaboada/research/signalling.html. The material contains two types of distributions in the RST Signalling Corpus: (1) the statistical distribution of coherence relations with respect to the signals (both signal types and specific signals) used to indicate those relations, and (2) the statistical distribution of signals with respect to the relations indicated by those signals.

## 7 Related work

Research on the development of discourse annotated corpora with textual signals has recently received considerable attention in discourse communities. We outline some well-known corpus annotation projects.

Studies involving annotating signals in discourse have mainly been restricted to annotating DMs. The largest available discourse annotated corpus, the Penn Discourse Treebank or PDTB (Prasad et al., 2008), presents annotations of 18,459 explicit and 16,053 implicit discourse connectives (DMs) and their corresponding discourse relations in English newspaper texts (Prasad et al., 2007). Following the PDTB, a number of annotated corpora have been developed in other languages, such as Arabic (Al-Saif & Markert, 2010), Czech (Mladová et al., 2008), Hindi (Kolachina et al., 2012), Italian (Tonelli et al., 2010), Turkish (Zeyrek et al., 2010) and German (Versley & Gastel, 2013).

Sometimes, annotation projects focus on annotating a particular relation type and its corresponding DMs. For example, Derczynski and Gaizauskas (2013) develop a corpus called TB-sig annotated for temporal relations and temporal DMs (such as *before* and *as soon as*).

Attempts to annotate signals of coherence relations other than DMs have also been made in a few corpus-based projects. Afantenos et al. (2012) present ANNODIS, a corpus in French, which provides annotation of a wide range of textual signals in discourse such as punctuation, lexico-semantic patterns, layout and syntactic parallelism. Redeker et al. (2012) compile a corpus of Dutch texts annotated with discourse structure and lexical cohesion. The cohesive devices representing lexical cohesion in the corpus include features such as lexical expressions indicative of certain relations, anaphoric chains and ellipsis. Duque (2014) develops a small of corpus of 84 texts in Spanish, annotating only two relations, *Cause* and *Result*, from the RST Spanish Treebank (da Cunha et al., 2011) with signalling information using numerous linguistic features, such as DMs, anaphors, non-finite verbs and genre structure.

# 8 Conclusions

We have presented the RST Signalling Corpus, a corpus annotated for signals of coherence relations, and described how this large-scale annotation project was conceived, designed and executed. The corpus uses the existing relations in the RST Discourse Treebank as its source data to which it adds relevant signalling information. The RST Signalling Corpus contains annotation of different types of signals organized hierarchically at multiple levels, namely signal class, signal type and specific signal. It also includes annotation of relations indicated by multiple signals.

We have shown in our signalling research project that the signalling of coherence relations is not limited to discourse markers (DMs), generally considered to be the most typical (and sometimes the only type of) signals of coherence relations. Rather, relations can be indicated by diverse types of textual signals other than DMs. Consequently, unlike most other contemporary similar corpora, the RST Signalling Corpus provides annotation not only for DMs, but also for a wide range of signals other than DMs such as *lexical, semantic, syntactic, graphical* and *genre* features, which are frequently used to convey coherence relations. Most importantly, analyses of the corpus show that the majority of the relations in discourse are signalled, and furthermore, the majority of signalled relations are indicated not by DMs, but by signals other than DMs.

The corpus with signalling information has two clear applications. From a psycholinguistic point of view, we hope to be able to use it to determine how hearers and readers use signals to identify relations. Most of the psycholinguistic studies to date have investigated the role of DMs (or only a few signals) in the understanding of coherence relations (Cain & Nash, 2011; Cevasco, 2009; Degand & Sanders, 2002; Haberlandt, 1982; Kamalski, 2007; Meyer, 1975; Millis & Just, 1994; Mulder, 2008; Sanders et al., 2007; Sanders & Noordman, 2000; Sanders et al., 1992; Spyridakis & Standal, 1987). It would be very useful to extend such works by examining other types of signals, as found in the RST Signalling Corpus, to see what effects they have on comprehension.

The other main application of such an annotated corpus is in discourse parsing. A great deal of recent work (da Cunha et al., 2012; Feng & Hirst, 2012, 2014; Hernault et al., 2011; Hernault et al., 2010; Maziero et al., 2011; Mithun & Kosseim, 2011) and also earlier approaches (Corston-Oliver, 1998; Marcu, 2000; Schilder, 2002) have used DMs as the main signals to automatically parse relations, and almost exclusively at the sentence level. Our extended set of signals, and the fact that they work at all levels of discourse, will probably facilitate this task.

The RST Signalling Corpus is publicly available through LDC. We believe that the corpus has the potential to be used as a database for future research on signalling in discourse.

## References


Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., . . . Vieu, L. (2012). *An empirical resource for discovering cognitive principles of discourse organization: the ANNODIS corpus.* Paper presented at the the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Al-Saif, A., & Markert, K. (2010). *The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic.* Paper presented at the the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.

Alonso, L., Castellón, I., Gibert, K., & Padró, L. (2002). An Empirical Approach to Discourse Markers by Clustering. In M. T. Escrig, F. Toledo & E. Golobardes (Eds.), *Topics in Artificial Intelligence* (Vol. 2504, pp. 173-183). Berlin: Springer.

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.

Blakemore, D. (1987). *Semantic Constraints on Relevance*. Oxford: Blackwell.

Blakemore, D. (1992). *Understanding Utterances: An Introduction to Pragmatics*. Oxford: Blackwell.

Blakemore, D. (2002). *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

Cain, K., & Nash, H. M. (2011). The Influence of Connectives on Young Readers' Processing and Comprehension of Text. *Journal of Educational Psychology, 103*(2), 429-441.

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

Cevasco, J. (2009). The Role of Connectives in the Comprehension of Spontaneous Spoken Discourse. *The Spanish Journal of Psychology, 12*(1), 56-65.

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.

da Cunha, I., Torres-Moreno, J.-M., & Sierra, G. (2011). *On the development of the RST Spanish Treebank.* Paper presented at the the 5th Linguistic Annotation Workshop, 49th Annual Meeting of the Association for Computational Linguistics (ACL), Portland, OR.

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. (2013). *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., & 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 omprehension in L1 and L2. *Reading and Writing, 15*(7-8), 739-758.

Derczynski, L., & Gaizauskas, R. (2013). *Temporal Signals Help Label Temporal Relations.* Paper presented at the the annual meeting of the Association for Computational Linguistics, ACL, Sofia, Bulgaria.

Dipper, S., Götze, M., & Stede, M. (2004). *Simple Annotation Tools for Complex Annotation Tasks: an Evaluation.* Paper presented at the the LREC Workshop on XML-based Richly Annotated Corpora, Lisbon, Portugal.

Duque, E. (2014). Signaling causal coherence relations. *Discourse Studies, 16*(1), 25-46.

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.

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

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

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

Fraser, B. (2009). An Account of Discourse Markers. *International Review of Pragmatics, 1*, 293-320.

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

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

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).

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

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.

Kolachina, S., Prasad, R., Misra Sharma, D., & Joshi, A. (2012). *Evaluation of discourse relation annotation in the Hindi Discourse Treebank.* Paper presented at the the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

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

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

Lin, Z., Kan, M.-Y., & Ng, H. T. (2009). *Recognizing implicit discourse relations in the Penn Discourse Treebank.* Paper presented at the the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore.

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. (2013). 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. (1999). *A decision-based approach to rhetorical parsing.* Paper presented at the the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland.

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.

Matthiessen, C. M. I. M. (2015). Register in the round: Registerial cartography. *Functional Linguistics, 2*(9), 1-48.

Matthiessen, C. M. I. M., & Teruya, K. (2015). Grammatical realizations of rhetorical relations in different registers. *Word, 61*(3), 232-281.

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.

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

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.

Mladová, L., Zikánová, Š., & Hajičova, E. (2008). *From sentence to discourse: Building an annotation scheme for discourse based on Prague Dependency Treebank.* Paper presented at the the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marakéš, Maroko.

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

Mulder, G., & Sanders, T. J. M. (2012). Causal Coherence Relations and Levels of Discourse Representation. *Discourse Processes, 49*(6), 501-522.

Murray, J. D. (1995). Logical connectives and local coherence. In J. R. F. Lorch & E. J. O'Brien (Eds.), *Sources of Coherence in Reading* (pp. 107-125). Hillsdale, NJ: Lawrence Erlbaum.

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.

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.

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., Joshi, A., & Webber, B. (2010). *Realization of Discourse Relations by Other Means: Alternative Lexicalizations.* Paper presented at the the 23rd International Conference on Computational Linguistics, Beijing.

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., Berzlánovich, I., van der Vliet, N., Bouma, G., & Egg, M. (2012). *Multi-Layer Discourse Annotation of a Dutch Text Corpus.* Paper presented at the the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

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

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

Roze, C., Danlos, L., & Muller, P. (2012). EXCONN: A French Lexicon of Discourse Connectives. *Discours, 10*.

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.

Schiffrin, D. (2001). Discourse markers: language, meaning and context. In D. Schiffrin, D. Tannen & H. E. Hamilton (Eds.), *The Handbook of Discourse Analysis* (pp. 54-75). Malden, MA: Blackwell.

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.

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.

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

Stede, M., & Umbach, C. (1998). *DiMLex: A lexicon of discourse markers for text generation and understanding.* Paper presented at the the COLING-ACL '98 Conference, Montreal.

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. (2006a). Applications of rhetorical structure theory. *Discourse Studies, 8*(4), 567-588.
Taboada, M., & Mann, W. C. (2006b). 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).
Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. (2010). *Annotation of discourse relations for conversational spoken dialogs.* Paper presented at the the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.
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.
Versley, Y., & Gastel, A. (2013). Linguistic Tests for Discourse Relations in the TüBa-D/Z Corpus of Written German. *Dialogue and Discourse, 4*(2), 142-173.
Zeyrek, D., Demirşahin, I., Sevdik-Çalli, A. B., Balaban, H. Ö., Yalçinkaya, I., & Turan, Ü. D. (2010). *The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotation.* Paper presented at the the Fourth Linguistic Annotation Workshop (LAW-IV).

# Appendix

Complete taxonomy of signals used to annotate the RST Signalling Corpus

| Signal class | Signal type | Specific signal | Definition |
|---|---|---|---|
| single | DM | *and, but, if, since,*, etc. | Lexical expressions (conjunctions, adverbials and prepositional phrases, etc.) |
| | reference | personal reference | Pronouns, possessive determiners and possessive pronouns which are present in one span, and refer to an object or entity (or a pronoun) in the other span |
| | | demonstrative reference | Demonstrative determiners, demonstrative pronouns, adverbs (*here, there, now* and *then)*, which are present in one span and refer to an object or entity in the other span |
| | | comparative reference | Reference items (*equal, identical, similar, differently, more, less, better, worse, etc.*) which are present in one span and refer to an object or entity in the other span by means of identity or similarity |
| | | propositional reference | Represented by pronouns: *it, this* and *that,* in one span, refers to a proposition (a process, phenomenon or fact, and NOT an object or entity) in the other span |
| | lexical | indicative word | A word or phrase which signals a relation |
| | | alternate expression | A short tensed clause which functions as the signal of a relation |
| | semantic | synonymy | Words or phrases in respective spans are in a synonymy relationship, or a proper noun or a name in one span is abbreviated or mentioned as an acronym (referring to the same object or entity) in the other span. |
| | | antonymy | Words or phrases in respective spans are in an antonymy relationship. |

| | | | |
|---|---|---|---|
| | | meronymy | A set of objects or entities is introduced in one span, and a member object or entity from that set is mentioned in the other span. |
| | | repetition | An entity is introduced in one span, and the entity (or its name) is repeated in the other span. |
| | | indicative word pair | Words or phrases in the respective spans form a word (or phrasal) pair as they are very closely related by their semantic content. |
| | | lexical chain | Words or phrases in the respective spans are identical or semantically related. |
| | | general word | Words such as *thing*, *matter* and *issue* which are present in one span, and refer to an object, entity, fact or proposition in the other span in a more general way. |
| | morphological | tense | A change of tense, aspect or mood between the relevant clauses or sentences in the respective spans |
| | syntactic | relative clause | One span, functioning as the satellite, is a relative clause modifying an object or entity (or a proposition in a few instances) present in the other span or nucleus. |
| | | infinitival clause | One span, functioning as the satellite, is an infinitival clause embedded under the main clause or nucleus. |
| | | present participial clause | One span, functioning as the satellite, is a present participial clause embedded under the main clause or nucleus. |
| | | past participial clause | One span, functioning as the satellite, is a past participial clause embedded under the main clause or nucleus. |
| | | imperative clause | One span, functioning as the satellite, is an imperative clause. |
| | | interrupted matrix clause | The nucleus span is a sentence which is interrupted by the insertion of a clause or phrase functioning as the satellite span. |
| | | parallel syntactic construction | The spans (clausal segments) or part of the spans (phrasal segments) are parallel to each other in syntactic construction. |
| | | reported speech | The satellite span is the reporting speech and the nucleus span is the reported speech. |
| | | subject auxiliary inversion | The position of the subject and auxiliary verb in a subordinate clause (functioning as the satellite) is interchanged. |
| | | nominal modifier | The satellite span is a reduced relative clause or a non-finite clause functioning as the modifier of an object or entity present in the main clause or nucleus. |
| | | adjectival modifier | The satellite span is a non-finite clause functioning as the modifier of an adjective present in the main clause or nucleus. |
| | graphical | colon | The first span ends with a colon followed by the second span. |
| | | semicolon | The first span ends with a colon followed by the second span. |
| | | dash | The first span ends with a dash followed by the second span, or one of the spans is within dashes. |
| | | parentheses | The satellite span is inside parentheses. |
| | | items in sequence | The nuclei in a multinuclear relation are presented as a numbered list or as items occurring in a sequential order. |
| | genre | inverted pyramid scheme | The content of the first paragraph (or the first few paragraphs) is elaborated on in the subsequent |

| | | | |
|---|---|---|---|
| | | | paragraphs. |
| | | newspaper layout | Visual features that helps understanding of the organization of a newspaper text (e.g., heading, date and place, body of text, information about the author) |
| | | newspaper style attribution | Features characteristic of the newspaper genre, indicative of *Attribution* relations |
| | | newspaper style definition | Features characteristic of the newspaper genre, indicative of *Definition* relations |
| | numerical | same count | The number of certain objects or entities represented by a word (e.g., *five, two*) in one span is equal to the numerical count of those objects or entities present in the other span. |
| combined | (reference + syntactic) | (personal reference + subject NP) | An object or entity (or a pronoun) is mentioned in the first (also nucleus) span, and a personal pronoun (*I, she, they*) referring to the same object or entity (or that previously mentioned pronoun) is used as the subject NP of the sentence the satellite span starts with. |
| | | (demonstrative reference + subject NP) | An object or entity (or demonstrative pronoun) is mentioned in the first (also nucleus) span, and a demonstrative pronoun (*this, that, those*) referring to the same object or entity (or that previously mentioned pronoun) is used as the subject NP of the sentence the satellite span starts with. |
| | | (comparative reference + subject NP) | An object or entity is introduced in the first (also nucleus) span, and a comparative referential item (e.g., *other, another*) referring to the same object or entity is used as the subject NP of the sentence the satellite span starts with. |
| | | (propositional reference + subject NP) | A fact, process or proposition in the first span (also nucleus) is referred to by the pronouns *it, this* or *that*, and the pronoun also occurs as the subject NP of the sentence the satellite span starts with. |
| | (semantic + syntactic) | (repetition + subject NP) | An object or entity in the first (also nucleus) span is repeated, and it occurs as the head of the subject NP of the sentence the satellite span starts with. |
| | | (lexical chain + subject NP) | A word (or phrase) in the satellite (also second) span is either identical or semantically related to a certain word(s) present in the nucleus (also first) span, and that word (or phrase) in the satellite span is used as the head of the subject NP of the sentence the satellite span starts with. |
| | | (synonymy + subject NP) | A word (or phrase) in the satellite (also second) span is synonymous to (or is an acronym of) a certain word(s) present in the nucleus (also first) span, and that word (or phrase) in the satellite span is used as the head of the subject NP of the sentence the satellite span starts with. |
| | | (meronymy + subject NP ) | A set of objects or entities is introduced in the first (also nucleus) span, and a member object or entity from that set is mentioned in the satellite span, and used as the head of the subject NP of the sentence the satellite span starts with. |
| | | (general word + subject NP) | A general word (e.g., *thing, matter* and *issue*), referring to an object, entity, fact or proposition in the first or nucleus span is used as the head of the subject NP of the sentence the satellite span starts with. |
| | (lexical + | (indicative word | The second span which is a present participial clause is |

| | | | |
|---|---|---|---|
| | syntactic) | + present participial clause) | preceded by an indicative word (e.g., *by, in*). |
| | (syntactic + semantic) | (parallel syntactic construction + lexical chain) | The spans (clausal segments) or part of the spans (phrasal segments) are parallel to each other in syntactic construction. The syntactic parallelism is also strengthened by the occurrence of lexical items present in a lexical chain between the spans. |
| | (syntactic + positional) | (past participial clause + beginning) | The satellite span which is a past participial clause is used in the beginning of the sentence containing both spans. |
| | | (present participial clause + beginning) | The satellite span which is a present participial clause is used in the beginning of the sentence containing both spans. |
| | (graphical + syntactic) | (comma + present participial clause) | The first span (usually the nucleus) is respectively followed by a comma and a present participial clause which is the second span (usually the satellite). |
| | | (comma + past participial clause) | The first span (usually the nucleus) is respectively followed by a comma and a past participial clause which is the second span (usually the satellite). |
| unsure | unsure | unsure | No potential signals were found or were specified. |

**Table 6. Complete taxonomy of signals**