Abstract: In this study, we examine how subject matter and presentational coherence relations in Rhetorical Structure Theory (Mann and Thompson 1988) are signalled in written discourse, and whether they differ quantitatively or qualitatively in terms of the signalling devices involved. By signalling we mean textual signals (discourse markers such as although, because and thus, and also signals such as tense, lexical chains or punctuation) that indicate a relation is present. We hypothesize that, because of their different nature, subject matter and presentational relations may be indicated by different signals, or at different frequencies. We conducted a corpus study examining 40 articles, comprising 1,306 relations, from the RST Discourse Treebank (Carlson, Marcu, and Okurowski 2002). We identified the signals for those relations, and added a new layer of annotation to them, to include signalling information. Results from our corpus analysis show that subject matter and presentational relations do not differ quantitatively or qualitatively in terms of signalling. These findings, although negative, point to a number of theoretical possibilities about the validity and accuracy of subject matter and presentational relation classification.

Keywords: Rhetorical Structure Theory, subject matter relations, presentational relations, signals of relations, corpus study, RST Discourse Treebank

Languages: English

Interpretation of a text requires not only the understanding of the meanings of individual discourse components (clauses, sentences, etc.), but also the understanding of how the meanings of those components are connected to each other. Coherence relations (also known as discourse or rhetorical relations) refer to the types of semantic or pragmatic connections that bind one discourse component to another. For example, in the following text,

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

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

1. COHERENCE RELATIONS AND RHETORICAL STRUCTURE THEORY. Coherence relations have been extensively investigated in the framework of Rhetorical Structure Theory or RST (Mann
and Thompson 1988). RST is a functional theory of text organization. It describes what parts a text is made of, what kinds of relationships exist between these parts, and how these parts are organized with respect to each other to constitute a coherent piece of discourse. RST originally emerged as computational theory for analyzing text structure, but later the use of RST has been extended to various other applications, including the analysis of spoken discourse.

In RST, relations are defined through different fields, the most important of which is the *Effect* (achieved on the text receiver), referring to the intention of the writer (or speaker) in presenting their discourse. Relation inventories are open, and the most common relation taxonomies include names such as *Cause, Concession, Condition, Elaboration, Result* and *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 clausal units (smallest discourse components) that enter into rhetorical (coherence) relations with each other, in a recursive manner. Mann and Thompson proposed that most texts can be analyzed in their entirety as recursive applications of different types of relations. In effect, this means that an entire text can be analyzed as a tree structure, with clausal units being the branches and relations the nodes.

In **Figure 1**, we provide an RST analysis of a text taken from the RST Discourse Treebank (Carlson, Marcu, and Okurowski 2002).

![Figure 1: Graphical representation of an RST analysis](image)

The RST analysis in **Figure 1** shows that the text comprises five spans which are represented by the cardinal numbers, 1, 2, 3, 4, and 5, respectively. In the diagram, the arrowhead points to a span refer to the nuclei, and the arrow points away from another span refer to the satellites. Span 1 (satellite) is connected to Span 2 (nucleus) by an *Attribution* relation. Span 3 (nucleus) and 4 (nucleus) are in a multinuclear *List* relation, and together they make the combined span 3-4. Then, 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 1-4. Finally, span 5 (satellite) is connected to span 1-4 (nucleus) by an Elaboration (more specifically, Elaboration-additional) relation. For more information about the relational definitions, see Carlson and Marcu (2001).

2. SUBJECT MATTER AND PRESENTATIONAL RELATIONS IN RST. Relations in RST, in terms of their intended effects, are divided into two groups: subject matter relations (e.g., Elaboration, Circumstance, Solutionhood, Cause, Restatement) and presentational relations (e.g., Motivation, Background, Justify, Concession). In subject matter relations, the text producer intends for the reader to understand the relation. In presentational relations, on the other hand, the intended effect is to increase some inclination in the reader (positive regard, belief, or acceptance of the nucleus). The distinction is related to the semantic/pragmatic divide proposed by van Dijk to classify discourse connectives according to what type of relation they signal: “[p]ragmatic connectives express relations between speech acts, whereas semantic connectives express relations between denoted facts” (van Dijk 1979: 449). The distinction is also represented by the following labels (corresponding to subject matter and presentational relations, respectively):

- External and Internal (Halliday and Hasan 1976a, Martin 1992).
- Ideational and Interpersonal. This is a distinction proposed in Systemic Functional Linguistics (Halliday 1985), and applied to coherence relations by others (Maier and Hovy 1993, Redeker 2000). The SFL distinction includes a third category, Textual, which does not fit clearly against the other dual classifications.
- Experiential and Rhetorical (Benwell 1999).
- Causal and Diagnostic, restricted to relations linked by connectives such as because, e.g., The streets are wet because it is raining versus It is raining, because the streets are wet (Traxler et al. 1997).

As it happens with most attempts at classification, the different classifications above represent distinctions that are not fully equivalent. The semantic/pragmatic distinction proposed by van Dijk (1979) separates relations between propositions (semantic) from those between speech acts (pragmatic). This distinction seems to be orthogonal to the subject matter/presentational divide in RST, since both propositions and speech acts can be in the same types of relations. Events denoted by propositions may be related through Cause (subject matter type) or Background (presentational type) relations; speech acts can be equally related through either type of relation. RST has concentrated on intended effects, more than on how they are achieved, whether it is through presenting relations containing events, facts, propositions or speech acts.

3. RESEARCH GOALS. In this study, we examine how subject matter and presentational relations are signalled, and whether they differ quantitatively and qualitatively in terms of the types of signalling involved. By signalling we mean textual signals that indicate a relation is present. These include discourse markers or DMs (e.g., although, because, since, thus), but also signals such as tense, lexical chains or punctuation. We hypothesize that, because of their different nature, subject matter and presentational relations may be indicated by different types of signals, or at different frequencies. Because subject matter relations are meant to be recognized as such,
they may have explicit signals, whereas in presentational relations the connection is more abstract, and the effect probably stems more out of the content than the form.

Research on coherence relations has often focused on cues that indicate the presence of a relation, or the lack of such cues, as many relations seem to be un signalled or implicit. Whereas it is true that many coherence relations are not signalled by a DM, it is also often the case that other signals have been understudied (Taboada and Mann 2006, Taboada 2009). We explore how many and what types of cues can be found if we study signalling beyond DMs, and how they are used to signal subject matter and presentational relations. For this purpose, we undertake a large-scale signalling annotation project, in which we first select a corpus already annotated for coherence relations, then examine the relations in the corpus, and finally add information on how those relations are signalled, including a variety of possible signals.


The taxonomy of signals is organized hierarchically in three levels: signal class, signal type and specific signal. The top level, signal class, has three tags representing three major classes of signals: single, combined and unsure. For each class, a second level of types is defined; for example, the class single is divided into nine types (DM, reference, lexical, semantic, morphological, syntactic, graphical, genre and numerical features). Finally, the third level in the hierarchy refers to the specific signals; for example, reference type has four specific signals: personal, demonstrative, comparative and propositional reference. The taxonomy of signals is illustrated in Figure 2. Note that subcategories are only illustrative, not exhaustive. More detail on the taxonomy can be found in Taboada and Das (2013)\(^1\).

In addition, we find that many relations are indicated by combined signals. Combined signals are made of two or more single signals which work together to indicate a particular relation. We have identified 10 broad types of combined signals\(^2\): (i) entity + positional, (ii) entity + syntactic + lexical, (iii) entity + syntactic, (iv) graphical + syntactic, (v) lexical + positional, (vi) lexical + syntactic + positional, (vii) lexical + syntactic, (viii) syntactic + lexical, (ix) syntactic + positional, and (x) semantic + syntactic.

---


\(^2\) For more detail on combined signals, see Taboada and Das (2013).
5. ANNOTATION PROCESS. For our corpus, we have selected the RST Discourse Treebank or RST-DT (Carlson, Marcu, and Okurowski 2002) which comprises a collection of 385 Wall Street Journal articles (financial reports, general interest stories, editorials, etc.) annotated for coherence relations. We chose to use the RST-DT because it is the only available large-scale corpus of RST relations for written texts. The annotated texts in the RST-DT are stored as LISP files which can be opened with RSTTool (O'Donnell 1997) for visual representation.

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

In the coding task, we provided annotations for signals of coherence relations, or in other words, we added signalling information to the existing relations from the RST-DT. For this
purpose, we extracted the signals identified, and documented them along with the relevant information about the relation in question, the document number (the file to which the relation belongs), the status of the spans (nucleus or satellite), and the span numbers (the location of the spans in the text). We annotated the signalling information in a separate Excel file, since RSTTool does not allow for multiple levels of annotation. For more information about the annotation process, see Das and Taboada (2013) and Taboada and Das (2013).

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

<table>
<thead>
<tr>
<th>Relation</th>
<th>Tokens</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signalled relation</td>
<td>1,129</td>
<td>86.45%</td>
</tr>
<tr>
<td>Unsignalled relation</td>
<td>177</td>
<td>13.55%</td>
</tr>
<tr>
<td>Total</td>
<td>1,306</td>
<td></td>
</tr>
<tr>
<td>Relations indicated by a DM</td>
<td>251</td>
<td>22.23%</td>
</tr>
<tr>
<td>Relations indicated by other signals</td>
<td>878</td>
<td>77.77%</td>
</tr>
<tr>
<td>Total</td>
<td>1,129</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Distribution of signalled and unsignalled relations

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

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

For the 251 instances of relations signalled by a DM, we have found 58 different DMs. Examples of some of these DMs include after, although, and, as, as a result, because, before, despite, for example, however, if, in addition, moreover, or, since, so, thus, unless, when and yet. For a full list of these extracted markers, see Taboada and Das (2013).

For the 878 signalled relations without DMs, we have found that a wide variety of signals are used to indicate them. These signals include all the eight types of signals other than DMs, mentioned in Section 3, along with numerous specific signals belonging to each type. In our corpus analysis, 81.67% of the signalled relations (922 out of 1,129 signalled relations) are exclusively indicated by a single signal (including DMs), whereas 5.67% of the signalled relations (64 out of 1,129 signalled relations) are indicated by a combined signal. In addition, the distribution also shows that 12.49% of the signalled relations (141 out of 1,129 signalled relations) contain multiple signals.

In terms of relation types, we divided the relations in the RST-DT in three groups: subject matter, presentational and undetermined relations, as shown in Table 2.

4 Multiple signals refer to two or more types of signals (single or combined) which are separately used to indicate a particular relation instance. For more detail, see Taboada and Das (2013).
Table 2: Relation classification on subject matter/presentational/undetermined basis

For subject matter and presentational relations, we mainly followed Mann and Thompson’s (1988) original classification RST relations. For the new relations, which were not part of the original RST taxonomy but used exclusively for annotating the RST-DT, we examined the definitions of those relations from the RST-DT annotation manual (Carlson and Marcu 2001), and assigned them to what we believed was the most suitable category. However, there were also some new relations whose definitions did not fit adequately to any of the two major relation types (i.e., subject matter or presentational), and hence, we assigned them to a new category called undetermined relations.

We also found (as shown in Table 3) that among the 1,306 relations examined there are 762 subject matter and 358 presentational relations, along with the remaining 186 undetermined relations.

Table 3: Distribution of relation types

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Tokens</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject matter</td>
<td>762</td>
<td>58.35%</td>
</tr>
<tr>
<td>Presentational</td>
<td>358</td>
<td>27.41%</td>
</tr>
<tr>
<td>Undetermined</td>
<td>186</td>
<td>14.24%</td>
</tr>
<tr>
<td>Total</td>
<td>1,306</td>
<td></td>
</tr>
</tbody>
</table>

The distribution of relation types with respect to signalling is provided in Table 4.

Table 4: Distribution of relation types for signalled and unsignalled relations

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Tokens</th>
<th># relations signalled</th>
<th># relations unsignalled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject matter</td>
<td>762</td>
<td>660 (86.61%)</td>
<td>102 (13.39%)</td>
</tr>
<tr>
<td>Presentational</td>
<td>358</td>
<td>300 (83.80%)</td>
<td>58 (16.20%)</td>
</tr>
<tr>
<td>Undetermined</td>
<td>186</td>
<td>169 (90.86%)</td>
<td>17 (9.14%)</td>
</tr>
<tr>
<td>Total</td>
<td>1,306</td>
<td>1,129 (86.45%)</td>
<td>177 (13.55%)</td>
</tr>
</tbody>
</table>

As Table 4 shows, out of the 762 subject matter relations 660 relations (86.61%) are signalled while the remaining 102 (13.39%) relations are not signalled. For presentational relations, out of 358 relations 300 relations (83.80%) are signalled and 58 (16.20%) relations are unsignalled. Finally, the breakdown for the 186 undetermined relations is into 169 relations (86.45%) signalled and 17 relations (9.14%) unsignalled.
Furthermore, analyzing the distribution of signalled relations of each relation type (as shown in Table 5), we found that among the 660 signalled subject matter relations 135 relations (20.45%) are indicated by a DM while the remaining 525 relations (79.55%) are indicated by signals other than DMs. For presentational relations, the distribution is between 58 (19.33%) relations with DMs and 242 (80.67%) relations with other signals, out of a total of 300 signalled relations.

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Signal type</th>
<th>Tokens</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject matter</td>
<td>Indicated by DMs</td>
<td>135</td>
<td>20.45%</td>
</tr>
<tr>
<td></td>
<td>Indicated by other signals</td>
<td>525</td>
<td>79.55%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>660</td>
<td></td>
</tr>
<tr>
<td>Presentational</td>
<td>Indicated by DMs</td>
<td>58</td>
<td>19.33%</td>
</tr>
<tr>
<td></td>
<td>Indicated by other signals</td>
<td>242</td>
<td>80.67%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Distribution of signalled relation types indicated by a DM and by other signals

The results (see Table 6) also show that the most frequently signals for subject matter relations include DMs such as and, if and but, and other signals such as entity, semantic and syntactic features. For presentational relations, DMs such as but, because and although, and other signals such as syntactic feature are the most common signals.

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Signal type</th>
<th>Most common signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject matter</td>
<td>Indicated by DMs</td>
<td>and, if, but</td>
</tr>
<tr>
<td></td>
<td>Indicated by other signals</td>
<td>entity, semantic, syntactic</td>
</tr>
<tr>
<td>Presentational</td>
<td>Indicated by DMs</td>
<td>but, because, although</td>
</tr>
<tr>
<td></td>
<td>Indicated by other signals</td>
<td>syntactic</td>
</tr>
</tbody>
</table>

Table 6: Most common signals for relation types

Finally, the signals for the most frequently signalled individual relations from each relation type are shown in Table 7.

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Relation</th>
<th>DM</th>
<th>Other signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject matter</td>
<td>Circumstance</td>
<td>Yes</td>
<td>syntactic, lexical</td>
</tr>
<tr>
<td></td>
<td>Cause</td>
<td>Yes</td>
<td>semantic</td>
</tr>
<tr>
<td></td>
<td>Condition</td>
<td>Yes</td>
<td>lexical, syntactic</td>
</tr>
<tr>
<td></td>
<td>Elaboration</td>
<td>Yes</td>
<td>entity, semantic, syntactic, genre</td>
</tr>
<tr>
<td></td>
<td>Purpose</td>
<td>-</td>
<td>syntactic</td>
</tr>
<tr>
<td>Presentational</td>
<td>Attribution</td>
<td>-</td>
<td>syntactic</td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>Yes</td>
<td>lexical, semantic, morphological</td>
</tr>
<tr>
<td></td>
<td>Antithesis</td>
<td>Yes</td>
<td>semantic</td>
</tr>
<tr>
<td></td>
<td>Evidence</td>
<td>-</td>
<td>lexical</td>
</tr>
<tr>
<td></td>
<td>Reason</td>
<td>Yes</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7: Distribution of most frequently signalled individual relations by signalling
We found that almost every subject matter relation (except a few, such as Purpose) is signalled by DMs, while most of the presentational relations are signalled by DMs. In terms of using other signals, almost every subject matter relation is indicated by other signals, while for the presentational type most individual relations are indicated by them, with a few such as Attribution and Evidence exclusively signalled by other signals.

7. DISCUSSION. In an effort to investigate the signalling of coherence relations beyond DMs, we found out that the majority of the relations present in a written discourse are signalled: in our corpus analysis, 1,129 (86.45%) relations out of the 1,306 relations examined are signalled. We also observed that coherence relations can well be indicated by signals other than DMs: in our study, out of the 1,129 signalled relations, 878 (77.77%) relations contain a signal other than a DM. Furthermore, signals of coherence relations in written discourse are diverse in nature, and can be broadly classified in major groups, such as DM, reference, lexical, semantic, syntactic, graphical and genre features. The individual signal groups also contain different specific signals in themselves. For example, syntactic feature includes specific signals such as relative clause, participial clause and parallel syntactic construction.

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

Examining the signalling of subject matter and presentational relations, we observed that there is not much statistical difference between the two relation types: around 85% of each type of relations (86.61% for subject matter and 83.80% for presentational relations) present in a corpus are indicated by some linguistic signals. Furthermore, among the signalled relations of each type, around 20% of the relations (20.45% for subject matter and 19.33% for presentational relations) are signalled by a DM while around 80% of relations (79.55% for subject matter and 80.67% for presentational relations) are indicated by signals other than DMs.

Qualitatively, we also found that the signalling of two relation types do not show any noticeable difference. For example, both subject matter and presentational relations are signalled by DMs as well as by other signals. Furthermore, although there are a few differences between the two relation types in using certain signals, the majority of signals for both the subject matter and presentational relations overlap as both relation types are conveyed through the same textual signals, be it either a specific DM or certain other signal type.

Finally, for the 177 relations for which we could not identify a signal, there are three different reasons why we believe that is the case. First of all, in some cases we found that there were errors in the existing annotation of relations in the RST-DT, and a relation was postulated, whereas we would not have annotated a relation. In those cases, the lack of signalling is perfectly understandable. Secondly, some of the RST-DT relations are not true RST relations. Relations such as Comment or Topic-shift, in our opinion, belong in the realm of discourse organization, not together with relations among propositions. Again, finding no signals in those cases is not surprising, as such phenomena are not likely to be indicated by the same type of signals as coherence relations proper. Finally, in many cases, one or both of the annotators had a sense that the relation was clear, but could not pinpoint the specific signal used. This is the case with tenuous entity relations, or relations that rely on world knowledge.
In sum, our findings show negative results, i.e., there are no significant quantitative or qualitative differences between subject matter and presentational relations in terms of signalling. However, these null findings point to a number of theoretical possibilities. First, subject matter and presentational relations can differ in their intended effects, but the difference in their intended effects may not actually lead to different kinds of signalling. In other words, signalling of coherence relations may be independent of their intended effects; otherwise, the difference in effects should have been reflected in their use of signals as well. Second, the dichotomy between subject matter and presentational relations may be theoretically invalid, i.e., relations may not differ in their intended effects. Finally, even if the subject matter vs. presentational classification is correct, or even if relations do differ in their intended effects, the assignation of individual relations to each type may not be absolutely correct, and the classification may require further review.

8. CONCLUSION. The purpose of this paper was to examine whether subject matter and presentational relations in RST differ in terms of signalling. Our results showed that quantitatively there is not a significant difference between the two relation types in terms of how often they are signalled and unsignalled in a text, and also as to the proportion of signals that are discourse markers (DMs) versus other linguistic devices. We also found that the qualitative difference between the two types is also minor in terms of the types of signals (both DMs and other signals) used by them. Thus, we conclude that subject matter and presentational relations do not differ quantitatively or qualitatively in their signalling. However, these findings, although negative, point to a number of theoretical possibilities, suggesting a revision of the theoretical validity of such relation classifications and also a re-examination of the accuracy in assigning individual relations to a particular relation type.

The annotation described in this paper is, however, a preliminary pilot study, comprising only 10% of the total corpus. In future work, we will expand to cover the entire corpus, and examine the difference between the signalling of subject matter and presentational relations more extensively. Parallel to this work, we would also like to examine the signalling of different relation types using other relation classifications, such as the three-way distinction between semantic, pragmatic and expansion relations as proposed by Redeker (2012), and to compare them with the existing classification in RST.

References


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


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


