Chapter #

Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory

Lynn Carlson,[1] Daniel Marcu,[2] and Mary Ellen Okurowski[1]
[1]U.S. Department of Defense; [2]Information Sciences Institute, University of Southern California

Abstract: We describe our experience in developing a discourse-annotated corpus for community-wide use. Working in the framework of Rhetorical Structure Theory, we were able to create a large annotated resource with very high consistency, using a well-defined methodology and protocol. This resource is made publicly available through the Linguistic Data Consortium to enable researchers to develop empirically grounded, discourse-specific applications.

Key words: discourse, corpus, annotation, rhetorical structure

1. INTRODUCTION

The advent of large-scale collections of annotated data has marked a paradigm shift in the research community for natural language processing. These corpora, now also common in many languages, have accelerated development efforts and energized the community. Annotation ranges from broad characterization of document-level information, such as topic or relevance judgments (Voorhees and Harman, 1999; Wayne, 2000) to discrete analysis of a wide range of linguistic phenomena. However, rich theoretical approaches to discourse/text analysis (Van Dijk and Kintsch, 1983; Meyer, 1985; Grosz and Sidner, 1986; Mann and Thompson, 1988) have yet to be applied on a large scale. So far, the annotation of discourse structure of documents has been applied primarily to identifying topical segments (Hearst, 1997), inter-sentential relations (Nomoto and Matsumoto, 1999; Ts'ou *et al.* 2000), and hierarchical analyses of small corpora (Moser and Moore, 1995; Marcu *et al.* 1999).

In this paper, we recount our experience in developing a large resource with discourse-level annotation for NLP research. Our main goal in undertaking this effort was to create a reference corpus for community-wide use. Two essential considerations from the outset were that the corpus needed to be consistently annotated, and that it would be made publicly available through the Linguistic Data Consortium for a nominal fee to cover distribution costs. The paper describes the challenges we faced in building a corpus of this level of complexity and scope – including selection of theoretical approach, annotation methodology, training, and quality assurance. The resulting corpus contains 385 documents of American English selected from the Penn Treebank (Marcus *et al.* 1993), hierarchically annotated in the framework of Rhetorical Structure Theory (Mann and Thompson, 1988). In the paper, we also show how the corpus can be mined in order to study a variety of linguistic phenomena that range from the role of cue phrases in signaling discourse relations to issues pertaining to high-level writing strategies. Our preliminary analysis illustrates the potential of this corpus as a rich new source of multi-layered discourse information to support multiple lines of research for language understanding applications.

2. FRAMEWORK

Two principle goals underpin the creation of this discourse-tagged corpus: 1) The corpus should be grounded in a particular theoretical approach, and 2) it should be sufficiently large to offer potential for wide-scale use – including linguistic analysis, training of statistical models of discourse, and other computational linguistic applications. These goals necessitated a number of constraints to our approach. We focused on annotating a large corpus of textual material, and did not address the applicability of our approach to spoken language corpora. The theoretical framework had to be practical and repeatable over a large set of documents in a reasonable amount of time, with a significant level of consistency across annotators. Thus, our approach contributes to the community quite differently from detailed analyses of specific discourse phenomena in depth, such as anaphoric relations (Garside *et al.* 1997) or style types (Leech *et al*. 1997); analysis of a single text from multiple perspectives (Mann and Thompson, 1992); or illustrations of a theoretical model on a single representative text (Britton and Black, 1985; Van Dijk and Kintsch, 1983).

Our annotation work is grounded in the Rhetorical Structure Theory (RST) framework (Mann and Thompson, 1988). We decided to use RST for three reasons:

- It is a framework that yields rich annotations that uniformly capture intentional, semantic, and textual features that are specific to a given text.
- Previous research on annotating texts with rhetorical structure trees (Marcu *et al.* 1999) has shown that texts can be annotated by multiple judges at relatively high levels of agreement. We aimed to produce annotation protocols that would yield even higher agreement figures.
- Previous research has shown that RST trees can play a crucial role in building natural language generation systems (Hovy, 1993; Moore and Paris, 1993; Moore, 1995) and text summarization systems (Marcu, 2000); can be used to increase the naturalness of machine translation outputs (Marcu *et al*. 2000); and can be used to build essay-scoring systems that provide students with discourse-based feedback (Burstein *et al*. 2001). We suspect that RST trees can be exploited successfully in the context of other applications as well.

In the RST framework, the discourse structure of a text can be represented as a tree defined in terms of four aspects:

- The leaves of the tree correspond to text fragments that represent the minimal units of the discourse, called *elementary discourse units*
- The internal nodes of the tree correspond to contiguous text *spans*
- Each node is characterized by its *nuclearity* – a nucleus indicates a more essential unit of information, while a satellite indicates a supporting or background unit of information.
- Each node is characterized by a *rhetorical relation* that holds between two or more non-overlapping, adjacent text spans. Relations can be intentional, semantic, or textual in nature.

Below, we describe the protocol that we used to build consistent RST annotations.

2.1 Segmenting Texts into Units

The first step in characterizing the discourse structure of a text in our protocol is to determine the elementary discourse units (EDUs), which are the minimal building blocks of a discourse tree. Mann and Thompson (1988, p. 244) state that "RST provides a general way to describe the relations among clauses in a text, whether or not they are grammatically or lexically signalled." Yet, applying this intuitive notion to the task of producing a large, consistently annotated corpus is extremely difficult, because the boundary between discourse and syntax can be very blurry. The examples below, which range from two distinct sentences to a single clause, all convey essentially the same meaning, packaged in different ways:

1. [Xerox Corp.'s third-quarter net income grew 6.2% on 7.3% higher revenue.] [This earned mixed reviews from Wall Street analysts.]
2. [Xerox Corp's third-quarter net income grew 6.2% on 7.3% higher revenue,] [which earned mixed reviews from Wall Street analysts.]
3. [Xerox Corp's third-quarter net income grew 6.2% on 7.3% higher revenue,] [earning mixed reviews from Wall Street analysts.]
4. [The 6.2% growth of Xerox Corp.'s third-quarter net income on 7.3% higher revenue earned mixed reviews from Wall Street analysts.]

In Example 1, there is a consequential relation between the first and second sentences. Ideally, we would like to capture that kind of rhetorical information regardless of the syntactic form in which it is conveyed. However, as examples 2-4 illustrate, separating rhetorical from syntactic analysis is not always easy. It is inevitable that any decision on how to bracket elementary discourse units necessarily involves some compromises.

Reseachers in the field have proposed a number of competing hypotheses about what constitutes an elementary discourse unit. While some take the elementary units to be clauses (Grimes, 1975; Givon, 1983; Longacre, 1983), others take them to be prosodic units (Hirschberg and Litman, 1993), turns of talk (Sacks, 1974), sentences (Polanyi, 1988), intentionally defined discourse segments (Grosz and Sidner, 1986), or the "contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse world," (Polanyi, 1996, p.5). Regardless of their theoretical stance, all agree that the elementary discourse units are non-overlapping spans of text.

Our goal was to find a balance between granularity of tagging and ability to identify units consistently on a large scale. In the end, we chose the clause as the elementary unit of discourse, using lexical and syntactic clues to help determine boundaries:

5. [**Although** Mr. Freeman is retiring,] [he will continue to work as a consultant for American Express on a project basis.]$_{\text{wsj\_1317}}$
6. [Bond Corp., a brewing, property, media and resources company, is selling many of its assets] [**to reduce** its debts.]$_{\text{wsj\_0630}}$

However, clauses that are subjects, objects, or complements of a main verb are not treated as EDUs:

7. [**Making computers smaller** often means **sacrificing memory**.]$_{\text{wsj\_2387}}$
8. [The company's current management found itself **locked into this**," he said.]$_{\text{wsj\_1103}}$

Relative clauses, nominal postmodifiers, or clauses that break up other legitimate EDUs, are treated as embedded discourse units:

9. [The results underscore Sears's difficulties] [**in implementing the "everyday low pricing" strategy…**]$_{\text{wsj\_1105}}$
10. [The Bush Administration,][1] [**trying to blunt growing demands from Western Europe for a relaxation of controls on exports to the Soviet bloc**,] [is questioning…]$_{\text{wsj\_2326}}$

[1] In this example, *The Bush Administration is questioning* is actually a single EDU, interrupted by the embedded discourse unit, *trying to blunt…* Using the annotation tool, the SAME-UNIT relation is selected to group the two parts of the unit back together.

Finally, a small number of phrasal EDUs are allowed, provided that the phrase begins with a strong discourse marker, such as *because, in spite of, as a result of, according to*. We opted for consistency in segmenting, sacrificing some potentially discourse-relevant phrases in the process.

2.2 Building up the Discourse Structure

Once the elementary units of discourse have been determined, adjacent spans are linked together via rhetorical relations, creating a hierarchical structure. Relations may be mononuclear or multinuclear. Mononuclear relations hold between two spans and reflect the situation in which one span, the *nucleus,* is more salient to the discourse structure, while the other span, the *satellite,* represents supporting information. Multinuclear relations hold among two or more spans, each of which has equal weight in the discourse structure. A total of 53 mononuclear and 25 multinuclear relations were used for the tagging of the RST Corpus. The final inventory of rhetorical relations is data driven, and is based on extensive analysis of the corpus. Although this inventory is highly detailed, annotators strongly preferred keeping a higher level of granularity in the selections available to them during the tagging process. More extensive analysis of the final tagged corpus will demonstrate the extent to which individual relations that are similar in semantic content were distinguished consistently during the tagging process. Although our corpus contained a number of different genres (e.g., editorials, letters and informative articles), applicability of this relation set to a broader range of genres may give rise to the need for addition rhetorical relations.

The 78 relations used in annotating the corpus can be partitioned into 16 classes that share some type of rhetorical meaning[2]:

- *Attribution*: attribution, attribution-negative
- *Background:* background, circumstance
- *Cause:* cause, result, consequence
- *Comparison:* comparison, preference, analogy, proportion
- *Condition:* condition, hypothetical, contingency, otherwise
- *Contrast:* contrast, concession, antithesis
- *Elaboration:* elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
- *Enablement:* purpose, enablement
- *Evaluation:* evaluation, interpretation, conclusion, comment
- *Explanation:* evidence, explanation-argumentative, reason
- *Joint:* list, disjunction
- *Manner-Means:* manner, means
- *Topic-Comment:* problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
- *Summary:* summary, restatement
- *Temporal:* temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence
- *Topic Change:* topic-shift, topic-drift

In addition, three relations are used to impose structure on the tree: textual-organization, span, and same-unit (used to link parts of units separated by an embedded unit or span).

[2] Many relations include variants based on nuclearity assignment, which are not included here. The complete list of relations can be viewed in the tagging guidelines (Carlson and Marcu, 2001).

3. DISCOURSE ANNOTATIONTASK

Our methodology for annotating the RST Corpus builds on prior corpus work in the Rhetorical Structure Theory framework by Marcu *et al.* (1999). Because the goal of this effort was to build a high-quality, consistently annotated reference corpus, the task required that we employ people as annotators whose primary professional experience was in the area of language analysis and reporting, provide extensive annotator training, and specify a rigorous set of annotation guidelines.

3.1 Annotator Profile and Training

The annotators hired to build the corpus were all professional language analysts with prior experience in other types of data annotation. They underwent extensive hands-on training, which took place roughly in three phases. During the orientation phase, the annotators were introduced to the principles of Rhetorical Structure Theory and the discourse-tagging tool used for the project (Marcu et al., 1999). The tool enables an annotator to segment a text into units, and then build up a hierarchical structure of the discourse. In this stage of the training, the focus was on segmenting hard copy texts into EDUs, and learning the mechanics of the tool.

In the second phase, annotators began to explore interpretations of discourse structure, by independently tagging a short document, based on an initial set of tagging guidelines, and then meeting as a group to compare results. The initial focus was on resolving segmentation differences, but over time this shifted to addressing issues of relations and nuclearity. These exploratory sessions led to enhancements in the tagging guidelines. To reinforce new rules, annotators re-tagged the document. During this process, we regularly tracked inter-annotator agreement (see Section 4.2). In the final phase, the annotation team concentrated on ways to reduce differences by adopting some heuristics for handling higher levels of the discourse structure.

Wiebe *et al.* (1999) presents a methodology for improving inter-coder reliability using automatically generated, bias-corrected tags. It is likely that Wiebe's method could be used to improve the reliability of our annotations as well. However, applying this method in our context would have been by no means a trivial process. The annotation task considered by Wiebe *et al.* (1999) consisted in labelling as “objective” or “subjective” a sample of independently generated sentences. In contrast, in our task, the examples were not independent. Decisions made by an annotator at a given step affected the decisions made at subsequent steps of the annotation. Our methodology for determining the “best” guidelines was much more of a consensus-building process, taking into consideration multiple factors at each step. The final tagging manual, about 80 pages in length, contains extensive examples from the corpus to illustrate text segmentation, nuclearity, selection of relations, and discourse cues. The manual can be downloaded from the following web site: *http://www.isi.edu/~marcu/discourse.*

The actual tagging of the corpus progressed in three developmental phases. During the initial phase of about four months, the team created a preliminary corpus of 100 tagged documents. This was followed by a one-month reassessment phase, during which we measured consistency across the group on a select set of documents, and refined the annotation rules. At this point, we decided to proceed by pre-segmenting all of the texts on hard copy, to ensure a higher overall quality to the final corpus. Each text was pre-segmented by two annotators; discrepancies were resolved by the author of the tagging guidelines. In the final phase (about six months) all 100 documents were re-tagged with the new approach and guidelines. The remainder of the corpus was tagged in this manner.

3.2 Tagging Strategies

Annotators developed different strategies for analyzing a document and building up the corresponding discourse tree. There were two basic orientations for document analysis – hard copy or graphical visualization with the tool. Hard copy analysis ranged from jotting of notes in the margins to marking up the document into discourse segments. Those who preferred a graphical orientation performed their analysis simultaneously with building the discourse structure, and were more likely to build the discourse tree in chunks, rather than incrementally.

We observed a variety of annotation styles for the actual building of a discourse tree. Two of the more representative styles are illustrated below.

1. *The annotator segments the text one unit at a time, then incrementally builds up the discourse tree by immediately attaching the current node to a previous node.* When building the tree in this fashion, the annotator must anticipate the upcoming discourse structure, possibly for a large span. Yet, often an appropriate choice of relation for an unseen segment may not be obvious from the current (rightmost) unit that needs to be attached. That is why annotators typically used this approach on short documents, but resorted to other strategies for longer documents.

2. *The annotator segments multiple units at a time, then builds discourse sub-trees for each sentence. Adjacent sentences are then linked, and larger sub-trees begin to emerge. The final tree is produced by linking major chunks of the discourse structure.* This strategy allows the annotator to see the emerging discourse structure more globally; thus, it was the preferred approach for longer documents.

Consider the text fragment below, consisting of four sentences, and 11 EDUs:

[Still, analysts don't expect the buy-back to significantly affect per-share earnings in the short term.]^{16} ["The impact won't be that great,"]^{17} [said Graeme Lidgerwood of First Boston Corp.]^{18} [This is in part because of the effect]^{19} [of having to average the number of shares outstanding,]^{20} [she said.]^{21} [In addition,]^{22} [Mrs. Lidgerwood said,]^{23} [Norfolk is likely to draw down its cash initially]^{24} [to finance the purchases]^{25} [and thus forfeit some interest income.]^{26} $_{\text{wsj\_1111}}$

The discourse sub-tree for this text fragment is given in Figure 1. Using Style 1 the annotator, upon segmenting unit [17], must anticipate the upcoming *example* relation, which spans units [17-26]. However, even if the annotator selects an incorrect relation at that point, the tool allows great flexibility in changing the structure of the tree later on.

Using Style 2, the annotator segments each sentence, and builds up corresponding sub-trees for spans [16], [17-18], [19-21] and [22-26]. The second and third sub-trees are then linked via an *explanation-argumentative* relation, after which, the fourth sub-tree is linked via an *elaboration-additional* relation. The resulting span [17-26] is finally attached to node [16] as an *example* satellite.

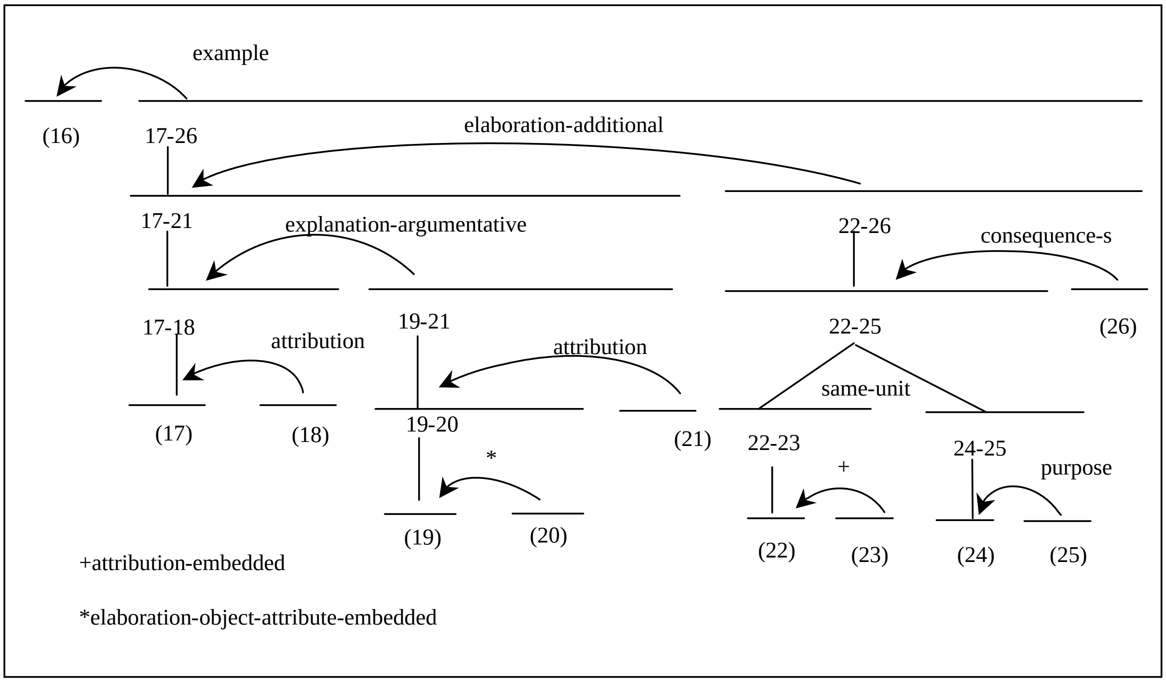

Figure 1. Discourse sub-tree for multiple sentences

4. QUALITY ASSURANCE

A number of steps were taken to ensure the quality of the final discourse corpus. These involved two types of tasks: checking the validity of the trees and tracking inter-annotator consistency.

4.1 Tree Validation Procedures

Annotators reviewed each tree for syntactic and semantic validity. Syntactic checking involved ensuring that the tree had a single root node and comparing the tree to the document to check for missing sentences or fragments from the end of the text. Semantic checking involved reviewing nuclearity assignments, as well as choice of relation and level of attachment in the tree. All trees were checked with a discourse parser and tree traversal program which often identified errors undetected by the manual validation process. In the end, all of the trees worked successfully with these programs.

4.2 Measuring Consistency

We tracked inter-annotator agreement during each phase of the project, using a method developed by Marcu *et al.* (1999) for computing kappa statistics over hierarchical structures. The kappa coefficient (Siegel and Castellan, 1988) has been used extensively in previous empirical studies of discourse (Carletta *et al*. 1997; Flammia and Zue, 1995; Passonneau and Litman, 1997). It measures pairwise agreement among a set of coders who make category judgments, correcting for chance expected agreement. The method described in Marcu *et al.* (1999) maps hierarchical structures into sets of units that are labeled with categorial judgments. The strengths and shortcomings of the approach are also discussed in detail there. Researchers in content analysis (Krippendorff, 1980) suggest that values of kappa > 0.8 reflect very high agreement, while values between 0.6 and 0.8 reflect good agreement.

Table 1 shows average kappa statistics reflecting the agreement of three annotators at various stages of the tasks on selected documents. Different sets of documents were chosen for each stage, with no overlap in documents. The statistics measure annotation reliability at four levels: elementary discourse units, hierarchical spans, hierarchical nuclearity and hierarchical relation assignments.

The results of Table 1 show significant improvement over time at all levels of annotation. At the unit level, the initial (April 00) scores and final (January 01) scores represent agreement on blind segmentation, and are shown in boldface. The interim June and November scores represent agreement based on hard copy pre-segmented texts. In these cases, two annotators independently segmented a hard copy of each document into EDUs. Discrepancies were resolved by the annotation team leader, and a "gold standard" annotated hard copy was produced. This version was given to the annotator (or annotators, for double-tagged documents) responsible for building the discourse tree for that document. Notice that even for these pre-segmented documents, agreement at the unit level is not 100% perfect, ranging from .95 to 1.00, because of human errors that were introduced when transferring the segmentation from hard copy into the annotation tool. A typical example of such an error would be inserting a unit boundary before an end-of-sentence quotation mark or period. As Table 1 shows, all levels demonstrate a marked improvement from April to November (when the final corpus was completed), ranging from about 0.77 to 0.92 at the span level, from 0.70 to 0.88 at the nuclearity level, and from 0.60 to 0.79 at the relation level. In particular, when relations are combined into the 16 rhetorically related classes discussed in Section 2.2, the November results of the annotation process are extremely good. The Fewer-Relations Column shows the improvement in scores on assigning relations when they are grouped in this manner, with November results ranging from 0.78 to 0.82 over the three pairs of annotators. In order to see how much of the improvement had to do with pre-segmenting, we asked the same three annotators to annotate five previously unseen documents in January, without reference to a pre-segmented document. The results of this experiment are given in the last row of Table 1, and they reflect only a small overall decline in performance from the November results. These scores reflect very strong agreement and represent a significant improvement over previously reported results on annotating multiple texts in the RST framework (Marcu *et al.* 1999).

Table 1. Inter-annotator agreement -- periodic results for three taggers

Taggers	*Units*	*Spans*	*Nuclearity*	*Relations*	*Fewer-Relations*	*No. of Docs*	*Avg. No. EDUs*
A, B, E (Apr 00)	**0.874407**	**0.772147**	**0.705330**	**0.601673**	**0.644851**	**4**	**128.750000**
A, B, E (Jun 00)	0.952721	0.844141	0.782589	0.708932	0.739616	5	38.400002
A, E (Nov 00)	0.984471	0.904707	0.835040	0.755486	0.784435	6	57.666668
B, E (Nov 00)	0.960384	0.890481	0.848976	0.782327	0.806389	7	88.285713
A, B (Nov 00)	1.000000	0.929157	0.882437	0.792134	0.822910	5	58.200001
A, B, E (Jan 01)	**0.971613**	**0.899971**	**0.855867**	**0.755539**	**0.782312**	**5**	**68.599998**

Table 2 reports final results for all pairs of taggers who double-annotated four or more documents, representing 30 out of the 53 documents that were double-tagged. Results are based on pre-segmented documents.

Our team was able to reach a significant level of consistency, even though they faced a number of challenges which reflect differences in the agreement scores at the various levels. While operating under the constraints typical of any theoretical approach in an applied environment, the annotators faced a task in which the complexity increased as support from the guidelines tended to decrease. Thus, while rules for segmenting were fairly precise, annotators relied on heuristics requiring more human judgment to assign relations and nuclearity. Another factor is that the cognitive challenge of the task increases as the tree takes shape. It is relatively straightforward for the annotator to make a decision on assignment of nuclearity and relation at the inter-clausal level, but this becomes more complex at the inter-sentential level, and extremely difficult when linking large segments.

This tension between task complexity and guideline under-specification resulted from the practical application of a theoretical model on a broad scale. While other discourse theoretical approaches posit distinctly different treatments for various levels of the discourse (Van Dijk and Kintsch, 1983; Meyer, 1985), RST relies on a standard methodology to analyze the document at all levels. The RST relation set is rich and the concept of nuclearity, somewhat interpretive. This gave our annotators more leeway in interpreting the higher levels of the discourse structure, thus introducing some stylistic differences, which may prove an interesting avenue of future research.

Table 2. Inter-annotator agreement -- final results for six taggers

Taggers	*Units*	*Spans*	*Nuclearity*	*Relations*	*Fewer-Relations*	*No. of Docs*	*Avg. No. EDUs*
B, E	0.960384	0.890481	0.848976	0.782327	0.806389	7	88.285713
A, E	0.984471	0.904707	0.835040	0.755486	0.784435	6	57.666668
A, B	1.000000	0.929157	0.882437	0.792134	0.822910	5	58.200001
A, C	0.950962	0.840187	0.782688	0.676564	0.711109	4	116.500000
A, F	0.952342	0.777553	0.694634	0.597302	0.624908	4	26.500000
A, D	1.000000	0.868280	0.801544	0.720692	0.769894	4	23.250000

5. CORPUS OVERVIEW

The RST Corpus consists of 385 Wall Street Journal articles from the Penn Treebank, representing over 176,000 words of text. In order to measure inter-annotator consistency, 53 of the documents (13.8%) were double-tagged. Various other characteristics of the corpus are reported below:

- The documents range in size from 31 to 2124 words, with an average of 458.14 words per document.
- The final tagged corpus contains 21,789 EDUs (excluding the double-tagged documents).
- The average number of EDUs per document is 56.59. The shortest discourse tree contains two EDUs, while the longest has 304 EDUs.
- The average number of words per EDU is 8.1.

The articles range over a variety of topics, including financial reports, general interest stories, business-related news, cultural reviews, editorials, and letters to the editor. In selecting these documents, we partnered with the Linguistic Data Consortium to select Penn Treebank texts for which the syntactic bracketing was known to be of high caliber. Thus, the RST Corpus provides an additional level of linguistic annotation to supplement existing annotated resources.

For details on obtaining the corpus, annotation software, tagging guidelines, and related documentation and resources, see: *http://www.isi.edu/~marcu/discourse.*

6. MINING THE RST CORPUS

A growing number of groups have developed or are developing discourse-annotated corpora for text. These can be characterized both in terms of the kinds of features annotated as well as by the scope of the annotation. Features may include specific discourse cues or markers, coreference links, identification of rhetorical relations, etc. The scope of the annotation refers to the levels of analysis within the document, and can be characterized as follows:

- *sentential:* annotation of features at the intra-sentential or inter-sentential level, at a single level of depth (Sundheim, 1995; Tsou *et al.* 2000; Nomoto and Matsumoto, 1999; Rebeyrolle, 2000).
- *hierarchical:* annotation of features at multiple levels, building upon lower levels of analysis at the clause or sentence level (Moser and Moore, 1995; Marcu *et al.* 1999)
- *document-level:* broad characterization of document structure such as identification of topical segments (Hearst, 1997), linking of large text segments via specific relations (Ferrari, 1998; Rebeyrolle, 2000), or defining text objects with a text architecture (Pery-Woodley and Rebeyrolle, 1998).

As a *hierarchical* type, the RST Corpus is a rich resource that records an extensive and intricate human interpretation of each text, governed by detailed annotation guidelines. This interpretation lends itself to analysis at many different levels. Below we illustrate our own preliminary mining of the corpus at the leaf-level, text-level and mid-level. These sample analyses show how the RST Corpus can accelerate and enrich computational analysis of discourse structure, because the researcher can extract and exploit the meta-language of the RST theory.

6.1 Leaf-Level Analysis: Comparison of Discourse Markers

Discourse markers have been the subject of a wide range of research both in theoretical (Halliday and Hasan, 1976; Schriffrin, 1987; Martin 1992) and computational linguistics (Hirschberg and Litman, 1987; Litman, 1996; Knott, 1995; Di Eugenio, Moore, and Paolucci, 1997; Marcu 2000).

Though on-line corpora have facilitated empirical investigations of the role of discourse markers in text analysis and generation, none of the previous empirical work could take advantage of a corpus as rich as the one we built – many empirical analyses were carried out with no access to hierarchical annotations of underlying texts (Knott, 1995; Marcu 2000) or with access to a relatively small corpus of hierarchically annotated structures (Di Eugenio, Moore, and Paolucci, 1997).

Having the RST Corpus already annotated and interpreted by human analysts allows the computational linguist to perform a meta-level analysis of discourse cues in multiple contexts. We examined two discourse markers, *since* and *as*, to explore their distribution in the RST-annotated corpus. The meta-language of the corpus gave us ready access to information about frequency, rhetorical relation, nuclearity, and other aspects of these cues. Table 3 summarizes the distribution of these cues in the training corpus (347 documents; 157,930 words).

Table 3. Comparison of discourse markers *since* and *as*

	since	*as*
# of occurrences of word in training corpus	128	730
# of occurrences that trigger discourse relation	42	240
# of different relations selected	10	25
Nuclearity of discourse marker:		
# in nucleus (mononuclear case)	11	9
# in satellite	35	205
# in nucleus (multinuclear case)	6	26
Position of discourse marker:		
# in first span of relation	14	48
# in second span of relation	28	192

	since	as
Scope of relation:		
# of inter-clausal	36	225
# of inter-sentential	3	10
# of multi-sentential	3	5

The information extracted reveals a number of properties of these markers. The first observation is that relative to the frequency of these words in the corpus, they only triggered a discourse relation in about one third of the cases. This is because *as* and *since* were not always factors in relating two EDUs, as defined in our annotation guidelines. Instead, the terms frequently performed a function at the propositional level, rather than the discourse level. For example:

- The term since often appeared in a temporal phrase: "…, says Scott C. Newquist, Kidder's head of investment banking since June." (wsj_0604)
- The term *as* often occurred as a phrasal complement of a main verb: "to serve, among other things, as the court of last resort for most patent disputes." (wsj_0601)

A second observation is that these markers trigger a large number of different rhetorical relations, with *as* triggering a much broader range of relations than *since*.[3] A breakdown of the distribution of these markers across the inventory of rhetorical relations is shown in Table 4. Thirdly, Table 3 shows that when these terms did appear as discourse markers, they occurred a majority of the time in the satellite (83% for *since*; 85% for *as*), and in the second span of the relation (67% for *since*; 80% for *as*). Most significantly, we documented that these discourse markers almost always trigger *local* relations: for *since,* 86% of the cases are inter-clausal, and, for *as*, 94%.

In short, our initial analysis of a number of factors – frequency of occurrence, triggers for discourse relation, types of relation selected, scope of relation across multiple EDUs, nuclearity and span position – demonstrated a limited, and rather local, discourse function for *since* and *as*. In the following two sections, we illustrate how we exploit the meta-language of the annotated discourse trees to examine the higher and middle levels of the discourse structure.

Table 4. Comparison of relations triggered by discourse markers *since* and *as*

Relation Name	Frequency for *since*	Frequency for *as*
analogy		8
antithesis	1	1
attribution		18
background		1
cause-result		2
circumstance	14	69
comment		5
comparison		24
concession		1
conclusion	1	
condition		5
consequence-n	1	9
consequence-s		3
contingency		2
elaboration-additional	1	15
evaluation-s		1
evidence		1
example		1
explanation-argumentative	3	10
interpretation-s		1
list		4
manner		10
means		1

[3] For this preliminary analysis, all cases of *as* that triggered discourse relations were counted, including the following phrasals: *as long as, as soon as, as if, as a result*.

Relation Name	Frequency for *since*	Frequency for *as*
reason	5	1
result	1	18
sequence	4	1
temporal-after	7	
temporal-before		1
temporal-same-time		26
topic-drift	1	
spurious	3	1

6.2 Text-Level Analysis: Comparison of Trees for Different Styles of News Reports

There are few community resources available that capture text-level characteristics aside from genre or register, as in the Lancaster-Oslo/Bergen Corpus (Garside *et al.* 1987; Biber *et al.* 1998), or topic identification, as in the TDT Corpus (Wayne, 2000). The RST Corpus, in contrast, presents a multi-level discourse representation of the entire rhetorical structure of each document. The highest level of a discourse tree depicts the organization of the text and serves as a rhetorical outline. This view of the document is constrained by a finite set of rhetorical relations, the use of which is governed by the annotation guidelines. The tree may be viewed graphically at varying levels of depth. At the most abstract level, the tree provides a non-lexical or rhetorical summary of the content defined by the relations and hierarchical branching structure. Zooming in on any non-leaf node reveals concrete details of the rhetorical structure.

We selected three representative styles of news texts for comparison of text-level characteristics. In each figure, the upper pane of the RST Structurer tool displays the highest level of a text and the lower pane presents a portion of the text (with EDU boundaries marked) that corresponds to text spans in the tree.

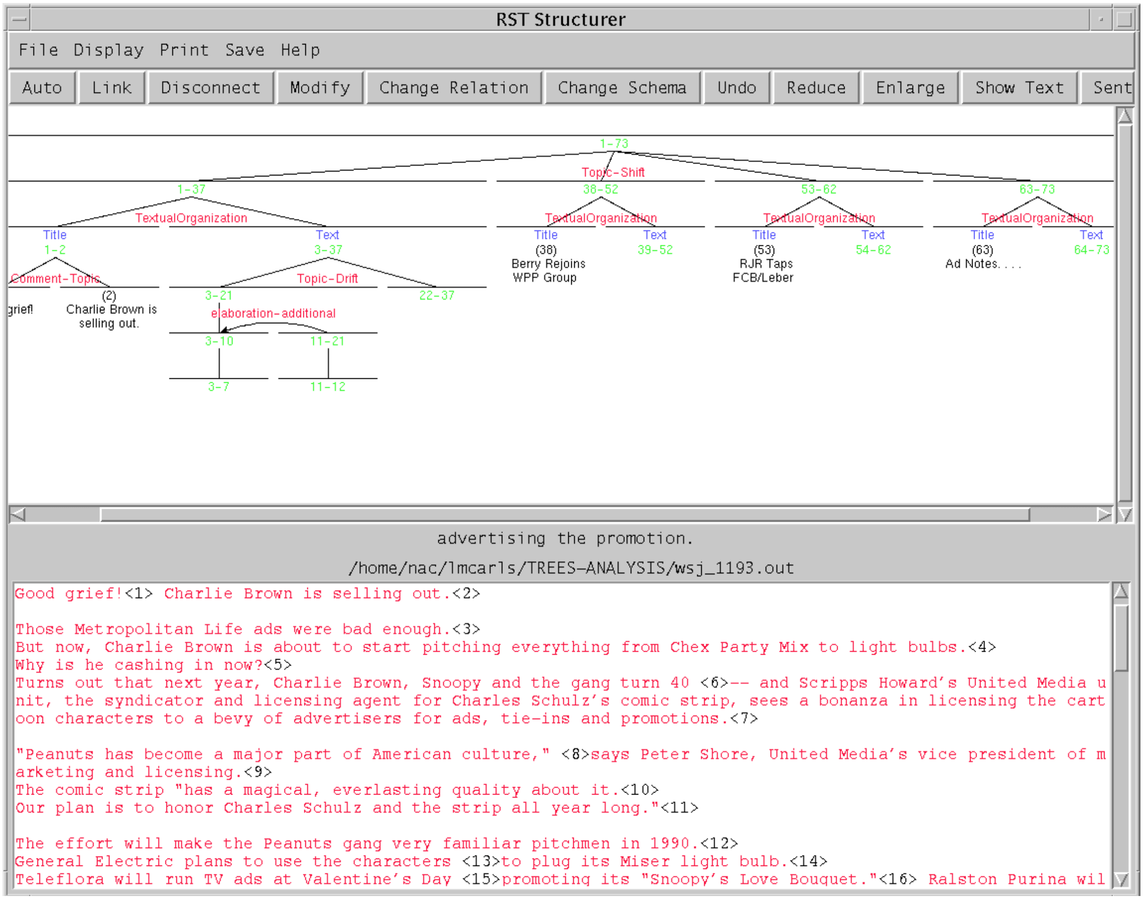

Figure 2. Top Level Discourse Tree for Document #1193: highly structured with topic shifts

Document #1193 (Figure 2) is a good example of a highly structured document, composed of a series of news briefs with clearly marked headings and sections. The rhetorical relation TOPIC-SHIFT links the sub-sections of the document, each of which is actually a separate news story. At the subsequent level, each story contains a TITLE and TEXT section – these labels are schemata, which refer to structural elements of the organization of the text. Schemata are associated with individual nodes in the discourse structure of a text, and represent an annotation level that is independent of the rhetorical relations. Schemata do not reflect relations between text spans, but rather characterize a functional role of an individual text span. The TEXTUAL-ORGANIZATION relation is used to link two or more spans labelled as schemata.

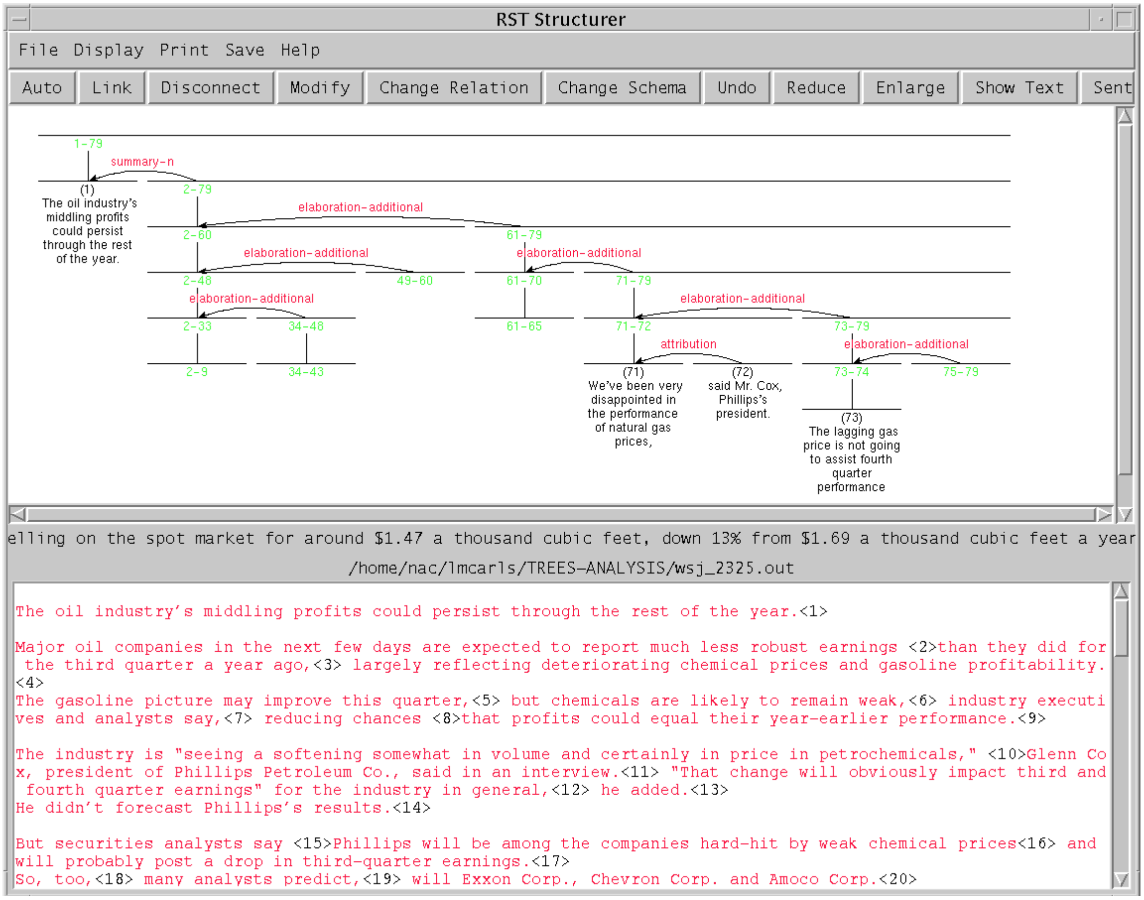

Figure 3. Top Level Discourse Tree for Document #2325: Summary lead followed by supporting details

Document #2325 (Figure 3) typifies the journalistic practice of a business news article with an initial summary lead followed by supporting details. This style is representative of a larger portion of the corpus than Document #1193. Here the text is formed by a series of successive text spans that elaborate on the summary sentence. The high level snapshot shows how the text is chunked into text segments that expand upon the content at different levels.

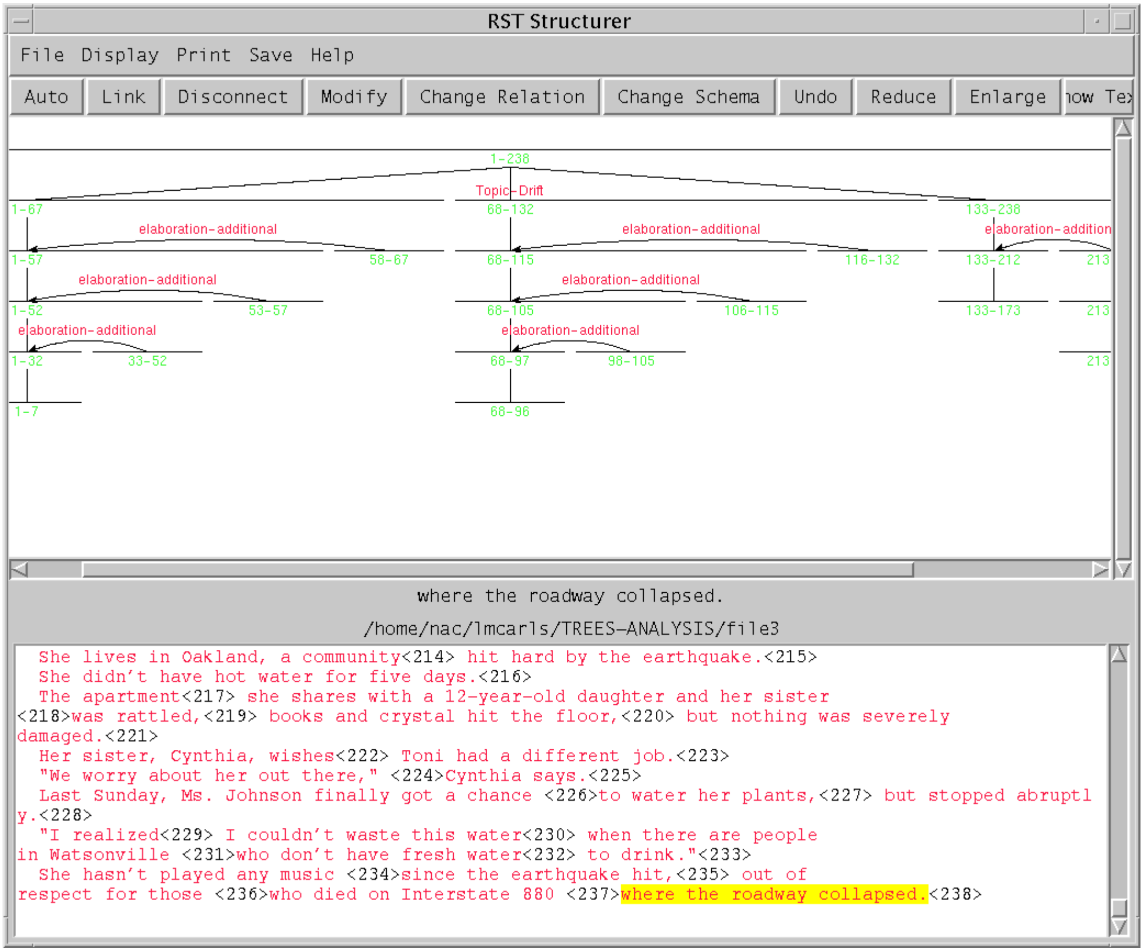

Figure 4. Top Level Discourse Tree for Document #file3: less structured, with topic drifts

The high-level discourse structure of Document #file3 (Figure 4) outlines a document that illustrates a TOPIC-DRIFT style: A human-interest story is woven around an expository text detailing the impact of the San Francisco earthquake on the insurance trade. The text begins with an on-the-scene description of an insurance claims adjuster assessing damages to a house destroyed by the San Francisco earthquake, switches to a general description of the impact on the insurance trade, and then returns to the personal experience of the adjuster. The document is a mixture of expository and narrative styles, which taxes the annotator because of its lack of overt structure.

We believe that the RST Corpus can support the discovery of new lines of inquiry at the text level. The RST Corpus enables the researcher to inventory the mechanisms used to organize the text at the highest level within the context of a single theory. It is available for exploring rhetorical strategies for generating text, discovering the degree of variation in text organization, and comparing RST to other theoretical approaches for representing the high-level rhetorical structure of documents.

6.3 Mid-Level Analysis: Examination of Relations

We define the mid-level discourse structure to be any multi-sentential segment of the text that is captured by a particular rhetorical relation, but is not dominated directly by the root node or a TEXTUALORGANIZATION relation.

Taking a second look at the documents characterized at the text level in Section 6.2, the reader will notice the frequent use of ELABORATION-ADDITIONAL as a rhetorical device for organizing large segments of the document at the upper middle level of the discourse. This occurs in all three of the text styles described. Although the total inventory of relations is quite extensive, one of the most frequently occurring relation sets in the RST Corpus is *elaboration* (see Section 2.2), because a typical rhetorical strategy is for the writer to expand on the previous context. Thus, the relation ELABORATION-ADDITIONAL became a de facto default whenever a more semantically marked relation did not fit the context.

There are, however, numerous occurrences of other multi-sentential rhetorical relations that are characteristic of the mid-level of the discourse. A brief investigation of two very different types of rhetorical relations – LIST and INTERPRETATION – will illustrate to the reader the potential utility of the corpus for analysis at the mid-level of the discourse.[4]

In the corpus annotation guidelines, a LIST relation is defined as "a multinuclear relation whose elements can be listed, but which are not in a comparison, contrast or other, stronger type of multinuclear relation. A LIST relation usually exhibits some sort of parallel structure between the units involved in the relation." Automatic identification of a LIST structure is trivial when the elements are enumerated or signalled by some other overt formatting characteristic such as indentation. However, very often in the corpus, a LIST relation was apparent to the annotator because of some sort of parallel syntactic or semantic structure between the units of the text, as in Figure 5. Here, the LIST structure was selected because of several features which, taken together, create a complex parallel *text* structure:

1. Each element of the LIST presents one example of contrast between two executives, e.g., *comes across as a low-key executive* vs. *has a flashier personality*.
2. Each CONTRAST relation in the LIST is structured as a compound sentence with the elements separated by a semi-colon; both conjoins contain a single clause in the active voice.
3. The names *Mr. Roman* and *Mr.Phillips* occur in a parallel manner in the two listed elements, each as the subject of one of the two conjoins, and in the same order for both items in the LIST.

Either branch of this sub-tree illustrates how parallel structures form a cohesive device (Halliday and Hasan, 1976). Together, they create a parallel text substructure. We have observed that this phenomenon occurs not only with the LIST relation, but also with other multinuclear relations, such as PROBLEM-SOLUTION, QUESTION-ANSWER, CONTRAST, and so on. The rich and varying set of such examples explicitly annotated in the corpus creates an opportunity to explore the phenomenon of textual parallelism, with potential application to various language processing applications, such as text generation or machine translation.

[4] Analysis at this level is analogous to research conducted on a range of discourse phenomena, such as anaphoric relations (Garside *et al*. 1997), or speech and dialog acts (Levin *et al*. 1998).

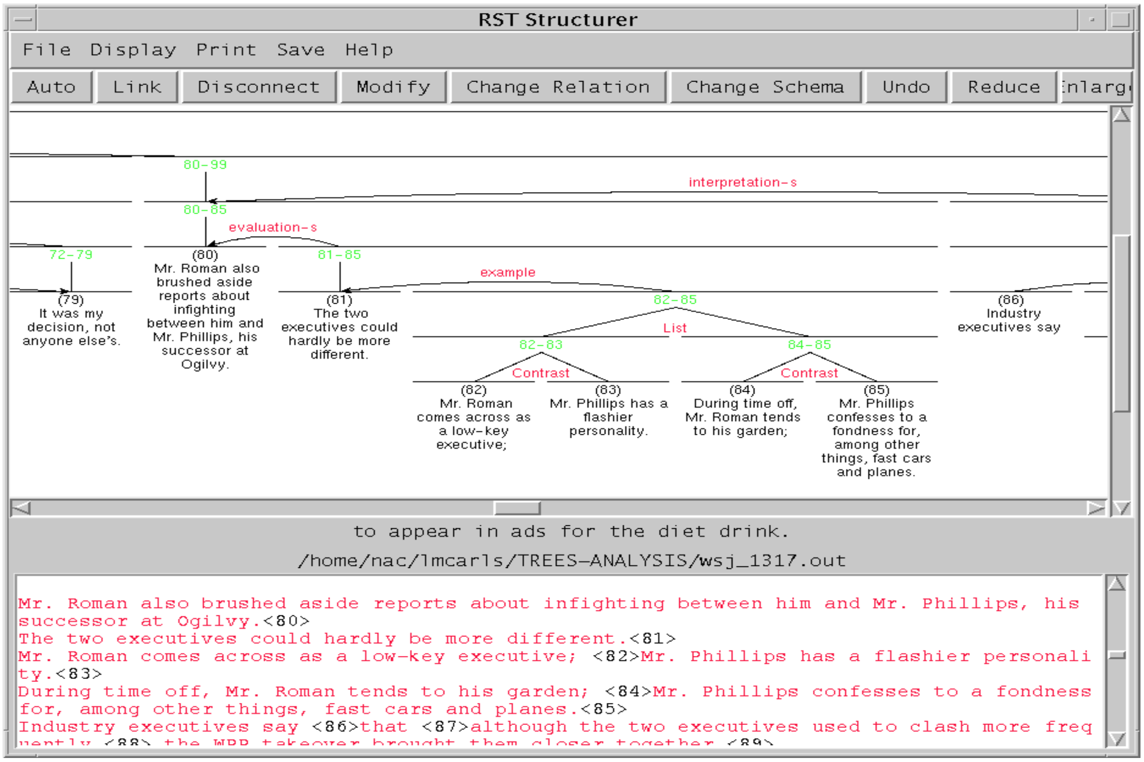

Figure 5. LIST relation with parallel syntactic and semantic structure

Another insight into analysis of the mid-level discourse structure comes from the discovery that subjective relations such as INTERPRETATION are interspersed throughout the RST Corpus, which consists primarily of expository news texts. Surprisingly, even when annotators disagree on the specific nature of the subjective relation[5], they consistently and easily identify the segments of the text that are subjective in nature. Below is a sample text fragment from Document #0628, which contains numerous subjective passages, two of which are shown here in italics:

> *Machine tool executives are hopeful, however, that recent developments in Eastern Europe will expand markets for U.S.-made machine tools in that region.*
>
> There is demand for state-of-the-art machine tools in the Soviet Union and in other Eastern European countries as those nations strive to improve the efficiency of their ailing factories as well as the quality of their goods.
>
> However, there's a continuing dispute between machine tool makers and the Defense Department over whether sophisticated U.S. machine tools would increase the Soviet Union's military might. "The Commerce Department says go, and the Defense Department says stop," complains one machine tool producer.

[5] Other examples of subjective relations found in the corpus are EVALUATION, COMMENT, and CONCLUSION.

If that controversy continues, U.S. machine tool makers say, West German and other foreign producers are likely to grab most of the sales in Eastern Europe.

The discourse tree for this portion of Document #0628 is shown in Figure 6. Note how the subjective passages correspond to the selection of an INTERPRETATION relation by the annotator. In the first case, the annotator has chosen to mark span [38-39] as the satellite of an INTERPRETATION relation. The labelled arc INTERPRETATION-S points to the nucleus of the relation, span [40-41], and indicates that the interpretation occurs in the satellite. In the second case, the annotator decides that the interpretative text, span [48-50], is more salient than the satellite, span [42-47], and selects INTERPRETATION-N as the relation. Access to our data will raise the bar for analysis of more complex linguistic phenomena related to subjectivity in writing.

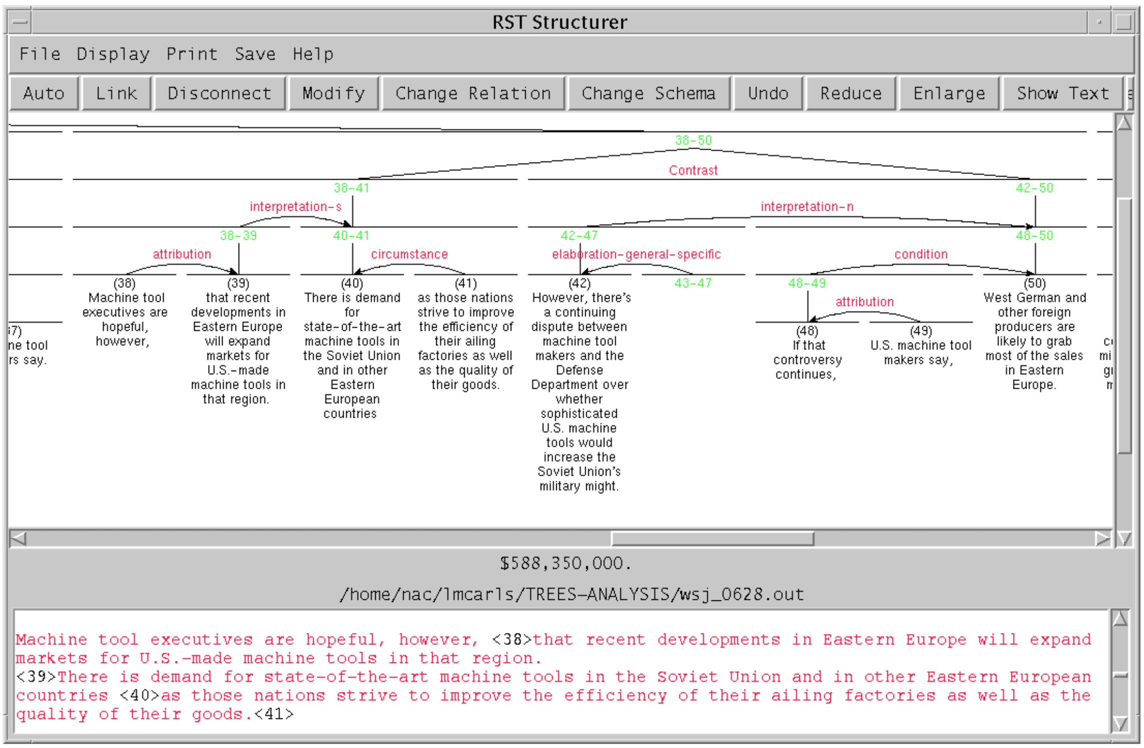

Figure 6. Subjectivity in text marked by INTERPRETATION relation

7. CONCLUSIONS AND FUTURE WORK

Developing corpora with these kinds of rich annotation is a labor-intensive effort. Building the RST Corpus involved more than a dozen people on a full or part time basis over a one-year time frame (Jan-Dec 2000). Annotation of a single document could take anywhere from 30 minutes to several hours, depending on the length and topic. Re-tagging a large number of documents after major enhancements to the annotation guidelines was also time consuming. Notwithstanding our effort to ensure the quality of the final discourse corpus and demonstration of relatively high inter-annotator agreement, we expect that researchers will identify anomalies in the RST Corpus, as typical of all annotation efforts. We believe that some subset of these can be tracked to simple errors. For example, an annotator accidentally highlights the wrong relation in a list or mis-assigns nuclearity.

A larger issue, though, stems from variation in stylistic interpretation among annotators. The RST theory does not differentiate between different micro- and macro-levels of the discourse structure, and thus, a fairly fine-grained set of relations operates at all levels. This, along with the concept of nuclearity, increased the variation in annotator interpretation. Even though we had very well defined rules for segmenting the text into EDUs, it proved quite difficult to make our already extensive guidelines more explicit in dictating how to assign nuclearity and relations. Other researchers (Ferrari, 1998; Meyer, 1985) have posited a few macro-level relations for text segments, or have conducted studies on a much more limited set of relations (Rebeyrolle, 2000). This approach has the advantage of limiting variability in annotation. However, our goal was to conduct a large-scale implementation within the framework of a single discourse theory in its entirety, with the expectation that this would allow for a better assessment of both its strengths and its limitations. We believe that the annotated corpus itself, along with the subset of documents with double annotations, will lead to refinements in the RST theory.

Based on our hands-on work and initial analysis of this substantial corpus, we anticipate that the RST Corpus will be multifunctional and support a wide range of language engineering applications. The added value of multiple layers of overt linguistic phenomena enhancing the Penn Treebank information can be exploited to advance the study of discourse, to enhance language technologies such as text summarization, text generation, machine translation or information retrieval, or to provide a testbed for new and creative natural language processing techniques.

ACKNOWLEDGEMENTS

We would like to acknowledge the significant contributions of the annotators who participated in the tagging of this corpus and in prior tagging experiments that led directly to this effort. Without their dedication and keen insights, this work would not have been possible: Estibaliz Amorrortu, Jean Hobbs, John Kovarik, Toby Merriken, Norb Rieg, Magdalena Romera, Maki Watanabe, and Markell West.

REFERENCES

Douglas Biber, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bruce Britton and John Black. 1985. *Understanding Expository Text.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Jill Burstein, Daniel Marcu, Slava Andreyev, and Martin Chodorow. 2001. Towards automatic identification of discourse elements in essays. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* Toulouse, France.

Lynn Carlson and Daniel Marcu. 2001. Discourse Tagging Reference Manual. ISI Technical Report. ISI-TR-545. (http://www.isi.edu/~marcu/discourse/).

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics* 23(1): 13-32.

Barbara Di Eugenio, Johanna Moore and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997),* pages 80-87, Madrid, Spain, July 7-12, 1997.

Giacomo Ferrari. 1998. Preliminary steps toward the creation of a discourse and text resource. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998),* Granada, Spain, 999-1001.

Giovanni Flammia and Victor Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Proceedings of the 4th European Conference on Speech Communication and Technology,* Madrid, Spain, vol. 3, 1965-1968.

Roger Garside, Steve Fligelstone and Simon Botley. 1997. Discourse Annotation: Anaphoric Relations in Corpora. In *Corpus annotation: Linguistic information from computer text corpora,* edited by R. Garside, G. Leech, and T. McEnery. London: Longman, 66-84.

Roger Garside, Geoffrey Leech and Geoffrey Sampson, eds. 1987. *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.
Talmy Givon. 1983. Topic continuity in discourse. In *Topic Continuity in Discourse: a Quantitative Cross-Language Study*. Amsterdam/Philadelphia: John Benjamins, 1-41.
Joseph Evans Grimes. 1975. *The Thread of Discourse.* The Hague, Paris: Mouton.
Barbara Grosz and Candice Sidner. 1986. Attentions, intentions, and the structure of discourse. *Computational Linguistics,* 12(3): 175-204.
M. A. K. Halliday and Ruqaiya Hasan, 1976. *Cohesion in English*. London: Longman.
Marti Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1): 33-64.
Julia Hirschberg and Diane Litman. 1987. Now Let's Talk About *Now*: Identifying Cue Phrases Intonationally. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL--87)*, pages 163-171.
Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3): 501-530.
Eduard Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence* 63(1-2): 341-386.
Alistair Knott. 1995. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD Thesis, University of Edinburgh.
Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology.* Beverly Hills, CA: Sage Publications.
Geoffrey Leech, Anthony McEnery, and Martin Wynne. 1997. Further levels of annotation. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, edited by R. Garside, G. Leech, and T. McEnery. London: Longman, 85-101.
Lori Levin, Ann Thyme-Gobbel, Klaus Ries, Alon Lavie, and Monika Woszczyna. 1998. A discourse coding scheme for conversation Spanish. In *Proceedings of the Fifth International Conference on Speech and Language Processing*. Sydney, Australia.
Diane Litman. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53-94.
Robert Longacre. 1983. *The Grammar of Discourse.* New York: Plenum Press.
William Mann and Sandra Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text,* 8(3): 243-281.
William Mann and Sandra Thompson, eds. 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text.* Amsterdam/Philadelphia: John Benjamins.
Daniel Marcu. 2000. The Theory and Practice of Discourse Parsing and Summarization. Cambridge, MA: The MIT Press.
Daniel Marcu, Estibaliz Amorrortu, and Magdelena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging,* College Park, MD, 48-57.
Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics,* Seattle, WA, 9-17.
Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics* 19(2), 313-330.
James R. Martin. 1992. *English Text. System and Structure.* John Benjamin Publishing Company, Philadelphia/Amsterdam.
Bonnie Meyer. 1985. Prose Analysis: Purposes, Procedures, and Problems. In *Understanding Expository Text*, edited by B. Britton and J. Black. Hillsdale, NJ: Lawrence Erlbaum Associates, 11-64.
Johanna Moore. 1995. Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context. Cambridge, MA: MIT Press.
Johanna Moore and Cecile Paris. 1993. Planning text for advisory dialogues: capturing intentional and rhetorical information. *Computational Linguistics* 19(4): 651-694.
Megan Moser and Johanna Moore. 1995. Investigating cue selection and placement in tutorial discourse. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge, MA, 130-135.
Tadashi Nomoto and Yuji Matsumoto. 1999. Learning discourse relations with active data selection. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, 158-167.
Rebecca Passonneau and Diane Litman. 1997. Discourse segmentation by human and automatic means. *Computational Linguistics* 23(1): 103-140.
Marie-Paule Pery-Woodley and Josette Rebeyrolle. 1998. Domain and genre in sublanguage text: definitional microtexts in three corpora. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC-1998)*, Granada, Spain, 987-992.
Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12: 601-638.
Livia Polanyi. 1996. The linguistic structure of discourse. Center for the Study of Language and Information. CSLI-96-200.
Josette Rebeyrolle. 2000. Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. In *Actes Journées Francophones d'Ingénierie de la Connaissance (IC'2000),* Toulouse, IRIT, 105-114.
Harvey Sacks, Emmanuel Schegloff, and Gail Jefferson. 1974. A simple systematics for the organization of turntaking in conversation. *Language* 50: 696-735.

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

Sidney Siegal and N.J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Beth Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6),* Columbia, MD, 13-31.

Benjamin K. T'sou, Tom B.Y. Lai, Samuel W.K. Chan, Weijun Gao, and Xuegang Zhan. 2000. Enhancement of Chinese discourse marker tagger with C.4.5. In *Proceedings of the Second Chinese Language Processing Workshop,* Hong Kong, 38-45.

Teun A. Van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension.* New York: Academic Press.

Ellen Voorhees and Donna Harman. 1999. *The Eighth Text Retrieval Conference (TREC-8).* NIST Special Publication 500-246.

Charles Wayne. 2000. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000),* Athens, Greece, 1487-1493.

Janyce Wiebe, Rebecca Bruce, and Thomas O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* College Park, MD, 246-253.