Experiments in Constructing a Corpus of Discourse Trees

Daniel Marcu
Information Sciences Institute and
Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu

Estibaliz Amorrortu
Department of Linguistics
University of Southern California
Los Angeles, CA 90089
amorrort@usc.edu

Magdalena Romera
Department of Linguistics
University of Southern California
Los Angeles, CA 90089
romera@usc.edu

Abstract

We discuss a tagging schema and a tagging tool for labeling the rhetorical structure of texts. We also propose a statistical method for measuring agreement of hierarchical structure annotations and we discuss its strengths and weaknesses. The statistical measure we use suggests that annotators can achieve good levels of agreement on the task of determining the high-level, rhetorical structure of texts. Our empirical experiments also suggest that building discourse parsers that incrementally derive correct rhetorical structures of unrestricted texts without applying any form of backtracking is unfeasible.

1 Introduction

Empirical studies of discourse structure have primarily focused on identifying discourse segment boundaries and their linguistic correlates. Very little attention has been paid so far to the high-level, *rhetorical relations* that hold between discourse segments. In some cases, the role of these relations was considered to fall outside the scope of a study (Flammia and Zue, 1995); in other cases, judgements were made with respect to a taxonomy of very few intention-based relations (usually *dominance* and *satisfaction-precedence*) (Grosz and Hirschberg, 1992; Nakatani et al., 1995; Hirschberg and Litman, 1987; Passonneau and Litman, 1997; Carletta et al., 1997). And in the only case in which a rich taxonomy of 29 relations was used (Moser and Moore, 1997), the corpus was small and specific to a very restricted genre: written interactions between a student and tutor on the subject of fault location and repair in electronic circuitry.

In spite of many influential proposals in the linguistic of discourse structures and relations (Ballard et al., 1971; Grimes, 1975; Halliday and Hasan, 1976; Martin, 1992; Mann and Thompson, 1988; Sanders et al., 1992; Sanders et al., 1993; Asher, 1993; Lascarides and Asher, 1993; Knott, 1995; Hovy and Maier, 1993), a number of empirical questions remain to be answered.

- Can human judges construct *rich* discourse structures in a manner that ensures inter-judge agreement that is statistically significant?
- How can one measure the agreement?
- How should judges (and programs) construct the discourse structure of texts; should they follow a top-down, bottom-up, or an incremental procedure?
- How does the genre of a text influence the degree to which judges achieve agreement on the task of rhetorical tagging?

In this paper, we describe an experiment designed to answer these questions.

2 The experiment

2.1 Tools

We used as starting point O'Donnell's discourse annotation tool (1997), which we improved significantly. The original tool constrains human judges to construct rhetorical structures in a bottom-up fashion: as a first step, judges determine the elementary discourse units *(edu)* of a text; subsequently, they recursively assemble the units into discourse trees, in a bottom-up fashion. As texts get larger, the annotation process becomes impractical.

We modified O'Donnell's tool in order to enable annotators to construct discourse structures in an incremental fashion as well. At any time t during the annotation process, annotators have access to two panels (see figure 1 for an example):

- The upper panel displays in the style of Mann and Thompson (1988) the discourse structure built up to time t. The discourse structure is

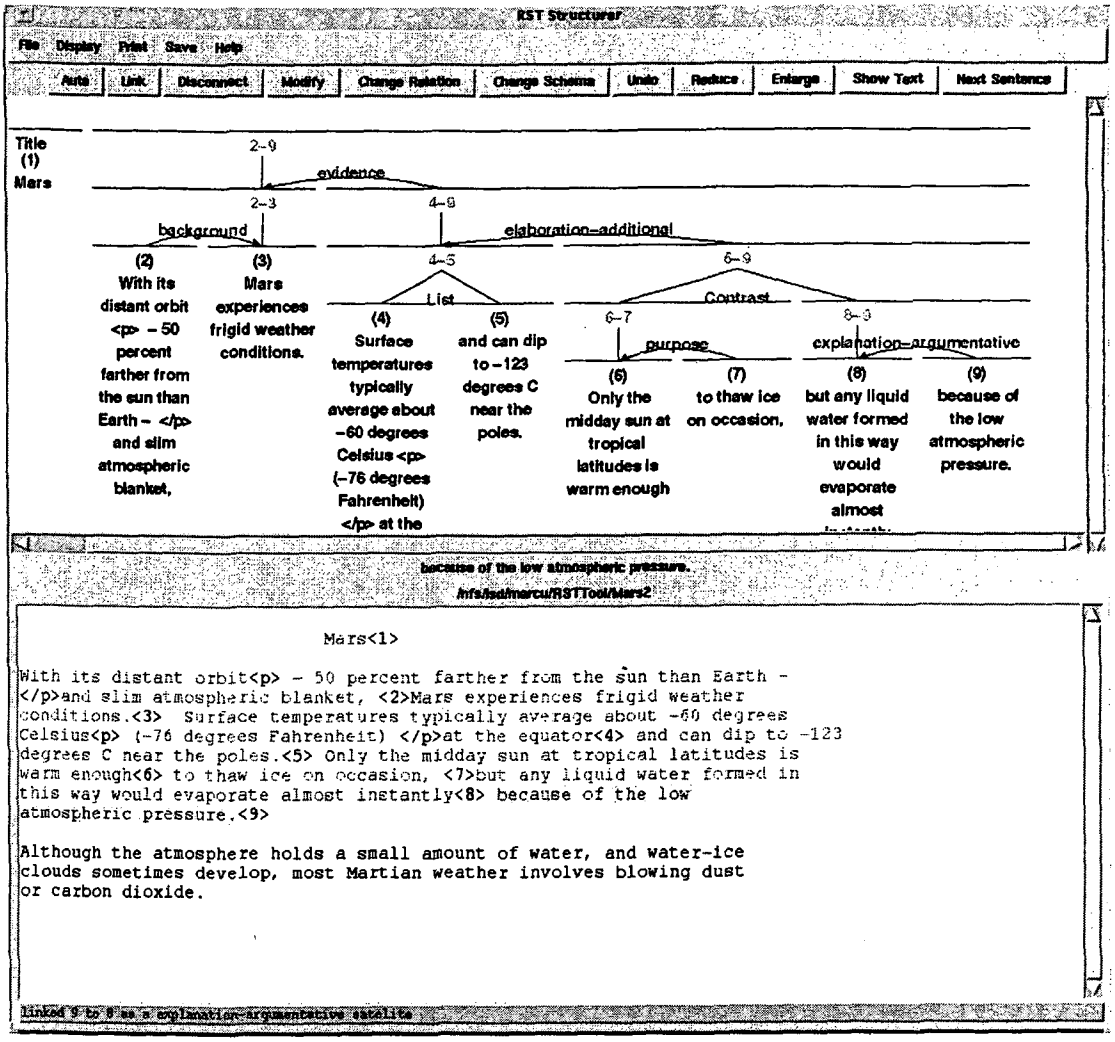


Figure 1: A snapshot of our annotation tool

a tree whose leaves correspond to *edus* and whose internal nodes correspond to contiguous text spans. Each internal node is characterized by a *rhetorical relation*, which is a relation that holds between two non-overlapping text spans called NUCLEUS and SATELLITE. (There are a few exceptions to this rule: some relations, such as the LIST relation that holds between units 4 and 5 and the CONTRAST relation that holds between spans [6,7] and [8,9] in figure 1, are multinuclear.) The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's purpose/intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. Some *edus* may contain *parenthetical units*, i.e., embedded units whose deletion does not affect the understanding of the *edu* to which they belong. For example, the unit shown in italics in (1) is parenthetic.

> This book, *which I have received from John,* is the best book that I have read in a while. (1)

- The lower panel displays the text read by the annotator up to time t and *only the first sentence* that immediately follows the labeled *edus*.

Annotators can create elementary and parenthetical units by clicking on their boundaries; immediately

add a newly created unit to a partial discourse structure using operations specific to tree-adjoining and bottom-up parsers; postpone the construction of a partial discourse structure until the understanding of the text enables them to do so; take discourse structures apart and re-connect them; change relation names and nuclearity assignments; undo any number of steps; etc. In other words, annotators have complete control over the discourse construction strategy that they employ.

All actions taken by annotators are automatically logged.

2.2 Annotation protocol

One of us initially prepared a manual that contained instructions pertaining to the functionality of the tool, definitions of *edu*s and rhetorical relations, and a protocol that was supposed to be followed during the annotation process (Marcu, 1998).

*Edu*s were defined functionally as clauses or clause-like units that are unequivocally the NUCLEUS or SATELLITE of a rhetorical relation that adds some significant information to the text. For example, *because of the low atmospheric pressure* in text (2) is not a fully fleshed clause. However, since it is the SATELLITE of an EXPLANATION relation, it should be treated as elementary.

> [Only the midday sun at tropical latitudes is warm enough] [to thaw ice on occasion,] [but any liquid water formed in this way would evaporate almost instantly] [because of the low atmospheric pressure.] (2)

A total of 70 rhetorical relations were partitioned into clusters, each cluster containing a subset of relations that shared some rhetorical meaning. For example, one cluster contained the contrast-like rhetorical relations of ANTITHESIS, CONTRAST, and CONCESSION. Another cluster contained REASON, EVIDENCE, and EXPLANATION. Each relation was paired with an informal definition given in the style of Mann and Thompson (1988) and Moser and Moore (1997) and one or more examples. No explicit distinction was made between intentional and informational relations. In addition, we also marked two constituency relations that were ubiquitous in our corpora and that often subsumed complex rhetorical constituents, and one textual relation. The constituency relations were ATTRIBUTION, which was used to label the relation between a reporting and a reported clause, and APPOSITION. The textual relation was TEXTUALORGANIZATION; it was used to connect in an RST-like manner the textual spans that corresponded to the title, author, and textual body of each document in the corpus. We also enabled the annotators to use the label OTHER-RELATION whenever they felt that no relation in the manual captured sufficiently well the meaning of a rhetorical relation that held between two text spans.

In an attempt to manage the inherent rhetorical ambiguity of texts, we also devised a protocol that listed the clusters of relations in decreasing order of specificity. Hence, the relations at the beginning of the protocol were more specific than the relations at the end of the protocol. The protocol specified that in assigning rhetorical relations judges should choose the first relation in the protocol whose definition was consistent with the case under consideration. For example, it is often the case that when an EVIDENCE relation holds between two segments, an ELABORATION relation holds as well. Because EVIDENCE is more specific than ELABORATION, it comes first in the protocol, and hence, whenever both of these relations hold, only EVIDENCE is supposed to be used for tagging.

The protocol specified that the rhetorical tagging should be performed incrementally. That is, if an annotator created an *edu* at time t and if she knew how to attach that *edu* to the discourse structure, she was supposed to do so at time $t+1$. If the text read up to time t did not warrant such a decision, the annotator was supposed to determine the *edu*s of the subsequent text and complete the construction of the discourse structure as soon as sufficient information became available.

2.3 Materials and method

Since we were aware of no previous study that investigated thoroughly the coverage of any set of rhetorical relations or any protocol, we felt necessary to divide the experiment into a training and an annotation stage. During the training stage, each of us built the discourse structures of 10 texts that varied in size from 162 to 1924 words. The texts belonged to the news story, editorial, and scientific genres. We had extensive discussions after the tagging of each text. During these discussions, we refined the definition of *edu*, the definitions and number of rhetorical relations that we used, and the order of the relations in the protocol. Eventually, our protocol comprised 50 mononuclear relations and 23 multinuclear relations. All relations were divided into 23 clusters of rhetorical similarity

MUC Corpus		WSJ Corpus		Brown Corpus	
Relation	Percent	Relation	Percent	Relation	Percent
ELABORATION-ADDITIONAL	13.80	ELABORATION-ADDITIONAL	17.41	ELABORATION-ADDITIONAL	21.64
ATTRIBUTION	12.07	ATTRIBUTION	14.78	LIST	16.29
LIST	9.99	LIST	11.25	JOINT	6.58
TEXTUALORGANIZATION	6.23	CONTRAST	6.84	CONTRAST	5.60
APPOSITION	5.02	JOINT	4.35	TEXTUALORGANIZATION	3.22
TOPICSHIFT	4.76	EVIDENCE	3.82	PURPOSE	2.88
JOINT	4.19	APPOSITION	3.31	EXPLANATION-ARGUMENTATIVE	2.68
CONTRAST	3.99	TOPIC-SHIFT	2.96	SEQUENCE	2.57
ELABORATION-OBJECT-ATTIBUTE	2.88	BACKGROUND	2.41	ELABORATION-GENERAL-SPECIFIC	2.23
EVIDENCE	2.54	ELABORATION-OBJECT-ATTIBUTE	2.37	TOPIC-SHIFT	2.12
BACKGROUND	2.42	PURPOSE	2.21	BACKGROUND	1.96
PURPOSE	2.26	ELABORATION-GENERAL-SPECIFIC	2.19	CONCESSION	1.84
ELABORATION-GENERAL-SPECIFIC	2.21	TOPIC-DRIFT	1.88	ELABORATION-OBJECT-ATTIBUTE	1.76
TOPIC-DRIFT	1.85	CONDITION	1.77	CONDITION	1.76
SEQUENCE	1.59	SEQUENCE	1.31	EVIDENCE	1.62
...		...		...	
OTHER-RELATION	0.38	OTHER-RELATION	0.32	OTHER-RELATION	0.19

Table 1: Distribution of the most frequent rhetorical relations in each of the three corpora.

(see (Marcu, 1998) for the complete list of rhetorical relations and protocol).

During the annotation stage, we independently built the discourse structures of 90 texts by following the instructions in the protocol; 30 texts were taken from the MUC7 co-reference corpus, 30 texts from the Brown-Learned corpus, and 30 texts from the Wall Street Journal (WSJ) corpus. The MUC corpus contained news stories about changes in corporate executive management personnel; the Brown corpus contained long, highly elaborate scientific articles; and the WSJ corpus contained editorials. The average number of words for each text was 405 in the MUC corpus, 2029 in the Brown corpus, and 878 in the WSJ corpus. The average number of *edu*s in each text was 52 in the MUC corpus, 170 in the Brown corpus, and 95 in the WSJ corpus. Each of the MUC texts was tagged by all three of us; each of the Brown and WSJ texts was tagged by only two of us. Table 1 shows the 15 relations that were used most frequently by annotators in each of the three corpora; the associated percentages reflect averages computed over all annotators. The table also shows the percentage of cases in which the annotators used the label OTHER-RELATION.

Problems with the method. It has been argued that the reliability of a coding schema can be assessed only on the basis of judgments made by naive coders (Carletta, 1996). Although we agree with this, we believe that more experiments of the kind reported here will have to be carried out before we can produce a tagging manual that is usable by naive coders. In our experiment, it is not clear how much of the agreement came from the manual and how much from the common understanding that we reached during the training session. For our annotation task, we felt that it was more important to arrive at a common understanding instead of tightly controlling how this understanding was reached. This position was taken by other computational linguists as well (Carletta et al., 1997, p. 25).

3 Computing agreement among judges

We computed agreement figures with respect to the way we set up *edu* boundaries and the way we built hierarchical discourse structures of texts.

3.1 Reliability of tagging the *edu* and parenthetical unit boundaries

In order to compute how well we agreed on determining the *edu* and parenthetical unit boundaries, we used the kappa coefficient k (Siegel and Castellan, 1988), a statistic used extensively in previous empirical studies of discourse. The kappa coefficient measures pairwise agreement among a set of coders who make category judgements, correcting for chance expected agreement (see equation (3) below, where $P(A)$ is the proportion of times a set of coders agree and $P(E)$ is the proportion of times a set of coders are expected to agree by chance).

$$k = \frac{P(A) - P(E)}{1 - P(E)} \quad (3)$$

Carletta (1996) suggests that the units over which the kappa statistic is computed affects the outcome. To account for this, we computed the kappa statistics in two ways:

1. The first statistic, k_w, reflects inter-annotator agreement under the assumption that *edu* and parenthetical unit boundaries can be inserted after any word in a text. Because many of the words occur within units and not at their boundaries, the chance agreement is very high, and therefore, k_w tends to be higher than the statistic discussed below.

2. The second statistic, k_u, reflects inter-annotator agreement under the assumption that *edu* and parenthetical unit boundaries can occur only at locations judged to be boundaries by at least one annotator. This statistic offers the most conservative measure of agreement.

3.2 Reliability of tagging the discourse structure of texts

3.2.1 Previous work

We are aware of only one proposal for computing agreement with respect to the way human judges construct hierarchical structures, that of Flammia and Zue (1995). This proposal appears to be adequate for computing the observed agreement, but it provides only a lower bound on the chance agreement, and hence, only an upper bound on the kappa coefficient. With the exception of Flammia and Zue, other researchers relied primarily on cascaded schemata for computing agreement among hierarchical structures. For example, Carletta et al. (1997) computed agreement on a coarse segmentation level that was constructed on the top of finer segments, by determining how well coders agreed on where the coarse segments started, and, for agreed starts, by computing how coders agreed on where coarse segments ended. Moser and Moore (1997) determined first the kappa coefficient with respect to the way judges assigned boundaries at the highest level of segmentation. Then judges met and agreed on a particular segmentation. Each high-level segment was then independently broken into smaller segments and the process was repeated recursively until the elementary unit level was reached. Although Moser and Moore's approach accommodates readily the traditional computation of kappa, it is impractical for large texts. In addition, since judges meet and agree on every level, it is likely that the agreement at finer levels of detail is influenced by judges' interaction.

3.2.2 Our approach

In order to compute the kappa statistics we devised a new method whose core idea is to map hierarchical structures into sets of units that are labeled with categorial judgments (see (Marcu and Hovy, 1999) for details). Consider, for example, the two hierarchical structures shown in figure 2.a, in which for simplicity, we focus only on the nuclear status of each segment (Nucleus or Satellite). In order to enable the computation of the kappa agreement we take as elementary all textual units found between two consecutive textual boundaries, independent of whether one or multiple judges chose those boundaries. Hence, for the segmentations in figure 2.a we consider that the text is made of 7 units; judge 1 took as elementary segments [0,1], [2,2], [3,3], [4,5] and [6,6], while judge 2 took as elementary segments [0,0], [1,1], [2,2], [3,3], [4,4] and [5,6].

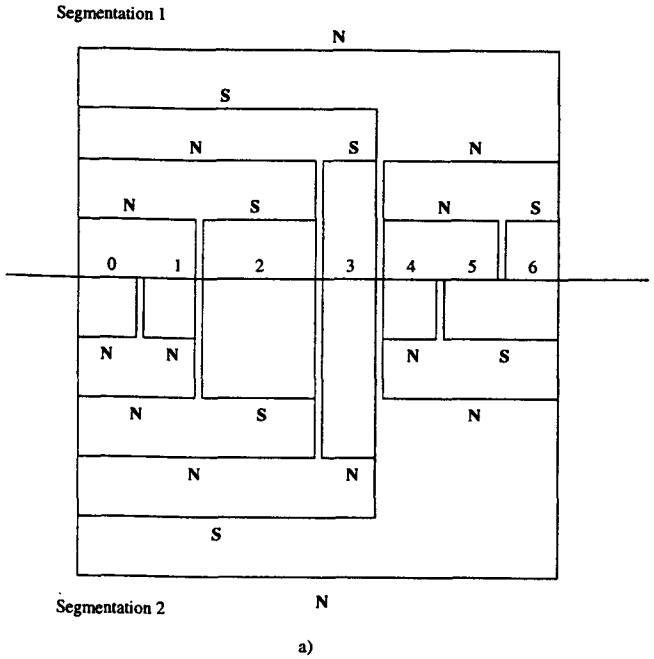
The mapping between the hierarchical structure and a set of units labeled with categorial judgments is straightforward if we consider not only the segments that play an active role in the structure (the nuclei and the satellites) but also the segments that are not explicitly represented. For example, for segmentation 1, there is no active segment across units [2,4], [2,5], and [2,6]. Similarly, for segmentation 2, there is no active segment across units [4,5] and [6,6]. By associating the label NONE to the textual units that do not play an active role in a hierarchical representation, each discourse structure can be mapped into a set that explicitly enumerates all possible spans that range over the units in the text. For a text of n units there are n spans of length 1, $n-1$ spans of length 2, ..., and 1 span of length n. Hence, each hierarchical structure of n units can be mapped into a set of $n+(n-1)+\ldots+1 = n(n+1)/2$ units, each labeled with a categorial judgment. And computing the kappa statistic for such sets is a problem with a textbook solution (Siegel and Castellan, 1988).

In the example in figure 2, we therefore compute the kappa statistics between the two hierarchies by computing the kappa statistics between the two sets that are represented in figure 2.b.

The hierarchical structures in figure 2 correspond to nuclearity judgments. However, the schema we use here is general, since it can accommodate the computation of kappa statistic for judgments at the segmentation and rhetorical levels as well. In fact, the schema can be applied to any discourse, syntactic, or semantic hierarchical labeling.

In our experiment, we computed the kappa statistic with respect to four categorial judgments:

1. k_s reflects the agreement with respect to the hi-



Segment	Segmentation 1	Segmentation 2
[0,0]	none	N
[0,1]	N	N
[0,2]	N	N
[0,3]	S	S
[0,4]	none	none
[0,5]	none	none
[0,6]	N	N
[1,1]	none	N
[1,2]	none	none
[1,3]	none	none
[1,4]	none	none
...		
[4,4]	none	N
[4,5]	N	none
[4,6]	N	N
[5,5]	none	none
[5,6]	none	S
[6,6]	S	none

b)

Figure 2: Computing the kappa statistic among hierarchical structures

erarchical assignment of discourse segments;

2. k_n reflects the agreement with respect to the hierarchical assignment of nuclearity;
3. k_r reflects the agreement with respect to the hierarchical assignment of rhetorical relations;
4. k_{rr} reflects the agreement with respect to the hierarchical assignment of rhetorical relations under the assumption that rhetorical relations are chosen from a reduced set of only 18 relations, each relation representing one or more clusters of rhetorical similarity.[1]

We chose to compute k_{rr} in order to estimate whether the confusion in assigning rhetorical relations lay within the clusters or across them.

Problems with the proposed method. In interpreting the statistical significance of the results reported in this paper, the readers should bear in mind that the method we propose and use here is not perfect. The biggest problem we perceive concerns the violation of the independence assumption between the categorial judgments that are usually associated with the computation of the kappa statistics. Obviously, in our proposal, since the decisions taken at one level in the tree affect decisions taken at other levels in the tree, the data points over which the kappa coefficient is computed are not independent. However, it is not clear what the effect on kappa this interdependence has: if two judges agree, for example, on the labels associated with two large spans, they automatically agree on many other spans that do not play an active role in the representation. When this happens, it is likely that the value of kappa increases. However, if two judges disagree on two high-level spans, they automatically disagree on other spans that play an active role in the representation. When this happens, it is likely that the value of kappa decreases. Therefore, it is not very clear how much the final value of kappa is skewed in one direction or another. Most likely, if two judges agree significantly, the kappa coefficient will be skewed to higher values; if two judges disagree significantly, the kappa coefficient will be skewed to smaller values.

Another problem concerns the effect of NONE-agreements on the computation of kappa. Although the kappa statistics makes corrections for chance agreement, it is likely that the kappa coefficient is "artificially" high, because of the large numbers of non-active spans that are labelled NONE. Typically, a hierarchical structure with n leaves will be mapped into $n(n+1)/2$ categorial judgments, of

[1] Some relations were used so rarely in our corpus that we decided to cluster them into only one group in spite of not belonging to the same cluster of rhetorical similarity.

which only $2n-1$ have values different than NONE. Hence, it is possible the kappa coefficient to be "artificially" high because of many agreements on non-active spans. However, the interdependence effect discussed above may equally well "artificially" decrease the value of the kappa coefficient. One may imagine variants of our method in which all NONE-NONE agreements are eliminated, or in which only $2n-1$ are preserved. The first variant may be infelicitous because its adoption may artificially prevent judges to agree on NONE labels. Adopting the second variant is problematic because we don't know exactly how many NONE labels to keep in the mapped representation.

Another potential problem stems from assigning the same importance to agreements at all levels in the hierarchy. For some classes of problems, one may argue that achieving agreement at higher levels in the hierarchy should be more important than achieving agreement at lower levels. Obviously, the method we described here does not enable such an intuition to be properly accounted for. However, for the discourse annotation task, we are quite ambivalent about this intuition. It is not clear to us whether we should consider the annotations that have high agreements with respect to large textual segments and low agreements with respect to small segments better than the annotations that have low agreements with respect to large textual segments and high agreements with respect to small segments. The first group of annotations would correspond to an ability to deal properly with global discourse phenomena, but no ability to deal with local discourse phenomena. The second group of annotations would correspond to an ability to deal properly with local discourse phenomena, but no ability to deal with global discourse phenomena. Which one is "better"? The method we propose treats all spans equally. It is similar in this respect to the labeled recall and precision methods used to evaluate parsers, for example, which also do not consider that it is more important to agree on high level constituents than low level constituents.

The method we propose does not enable one to assess agreement at different levels of granularity; it produces one number, which cannot be used to diagnose where the disagreements are coming from. Although we believe that cascade techniques that were used to measure agreement between hierarchies (Moser and Moore, 1997; Carletta et al., 1997) are more adequate for diagnosing problems in the annotation, we found these techniques difficult to apply on our data. Some of our trees have more than 200 elementary units; and carrying out and interpreting a cascade analysis at potentially 200 levels of granularity is not straightforward either.

Another choice for computing agreement of hierarchical annotations would be to devise a method similar to that used in the Kendall's τ statistic, in which one computes the minimal number of operations that can map one annotation into another. Since the problem of finding the minimal number of operations that rewrite a tree into another tree is NP-complete, devising an operational method for computing agreement does not seem computationally feasible. After all, the number of possible trees that can be built for a text with 200 units is a number larger than 1 followed by 110 zeroes.

3.3 Tagging consistency

For each corpus, table 2 displays the numbers of coders that annotated each text in the corpus (#c) and the average numbers of data points (N_w and N_u) over which the kappas were computed for each text in the corpus. In the first three rows, the table also shows the average kappa statistics computed for each text in the corpus with respect to judges' ability to agree on elementary discourse boundaries (k_w and k_u) and the average value of the corresponding z statistics (z_w and z_u) that were computed to test the significance of kappa (Siegel and Castellan, 1988, p. 289). The last three rows show the same statistics computed over all data points in each corpus.

The field of content analysis suggests that values of k higher than 0.6 indicate good agreement. Values of z that are higher than 2.32 correspond to significance levels that are higher than $\alpha = 0.01$. The results in table 2 indicate that high, statistically significant agreement was achieved for all three corpora with respect to the task of determining the elementary discourse units.

For each corpus, table 3 displays in its first three rows the number of coders (#c) that annotated the texts and the average number of points (N) over which the agreements were computed for each text in the corpus. In the first three rows, the table displays the average kappa statistics with respect to the judges' ability to agree on each text on discourse segmentation, k_s, nuclearity assignments, k_n, and rhetorical relation assignments, k_r and k_{rr}. In the last three rows, the table displays the kappa statistics computed over all the data points in each corpus. If

Corpus	#c	Word level		Unit level	
		N_w	k_w/z_w	N_u	k_u/z_u
MUC–avg/text	3	408	0.930/11.1	52	0.799/9.2
WSJ–avg/text	2	909	0.906/9.5	95	0.722/70.6
Brown–avg/text	2	2062	0.894/10.6	170	0.685/92.9
MUC–all	3	12242	0.919/66.0	1528	0.769/57.1
WSJ–all	2	27283	0.905/52.4	2836	0.717/419.3
Brown–all	2	61888	0.895/58.1	5100	0.688/841.8

Table 2: Inter-annotator agreement — *edu* boundaries.

Corpus	#c	N	Spans k_s/z_s	Nuclei k_n/z_n	Relations k_r/z_r	Fewer relations k_{rr}/z_{rr}
MUC–avg/text	3	1326	0.792/11.3	0.744/10.6	0.646/8.9	0.689/9.5
WSJ–avg/text	2	3654	0.753/5.9	0.691/5.5	0.588/4.7	0.626/5.0
Brown–avg/text	2	11634	0.733/5.4	0.658/4.9	0.539/4.0	0.586/4.3
MUC–all	3	39807	0.778/61.8	0.722/56.4	0.617/48.3	0.659/51.7
WSJ–all	2	109649	0.751/29.8	0.688/27.6	0.565/27.6	0.623/25.1
Brown–all	2	349039	0.736/29.7	0.661/26.8	0.543/22.1	0.589/23.9

Table 3: Inter-annotator agreement — discourse trees.

the statistical method we proposed does not skew the values of k — a fact that we have not demonstrated — the data in table 3 suggest that reliable agreement is obtained across all three corpora with respect to the assignment of discourse segments and nuclear statuses. Reliable agreement is obtained with respect to the rhetorical labeling only for the MUC corpus. The results in table 3 also show that a significant reduction in the size of the taxonomy of relations may not have a significant impact on agreement (k_{rr} is only about 4% higher than k_r). This suggests that choosing one relation from a set of rhetorically similar relations produces some, but not too much, confusion. However, it may also suggest that it is more difficult to assess *where* to attach an *edu* in a discourse tree than *what* relation to use.

The results in tables 2 and 3 also show that the agreement figures vary significantly from one corpus to another: the news story genre of the MUC texts yields higher agreement figures than the editorial genre of the WSJ texts, which yields higher agreement figures than the scientific genre of the Brown texts. One possible explanation is that some of the Brown texts, which dealt with advanced topics in mathematics, physics, and chemistry, were difficult to understand.

Overall, if our method for computing the kappa statistic is not skewed towards higher values, our experiment suggests that even simple, intuitive definitions of rhetorical relations, textual saliency, and discourse structure can lead to reliable annotation schemata. However, the results do not exclude that better definitions of *edu* and parenthetical units and rhetorical relations can lead to significant improvements in the reliability scores.

4 Tagging style

The vast majority of the computational approaches to discourse parsing rely on models that implicitly or explicitly assume that parsing is incremental (Polanyi, 1988; Lascarides and Asher, 1993; Gardent, 1997; Schilder, 1997; van den Berg, 1996; Cristea and Webber, 1997). That is, as *edu*s are processed, they are *immediately* added to *one* partial discourse structure that subsumes all previous text. However, the logs of our experiment show that, quite often, annotators are unable to decide where to attach a newly created *edu*. The annotation style varies significantly among annotators; but nevertheless, even the most aggressive annotator still needs to postpone 9.2% of the time the decision of where to attach a newly created *edu* (see table 4). Note that this percentage does not reflect UNDO steps, which may also correlate with attachment decisions that are eventually proven to be incorrect.[2]

We noticed that managing multiple partial dis-

[2] UNDO operations may reflect typo-like errors as well.

	Annotator 1		Annotator 2		Annotator 3	
	#	%	#	%	#	%
# incremental operations	2834	79.54	3938	58.41	6034	86.43
# non-incremental ops.	670	18.80	2509	37.21	642	9.20
# change-relation ops.	42	1.18	191	2.83	156	2.23
# change-tree-structure ops.	17	0.48	104	1.54	149	2.13
# operations	3563		6742		6981	
# undo cycles	247		515		466	
# undo operations/cycle	2.38		3.60		2.29	

Table 4: Distribution of tagging operations.

course trees during the annotation process is the norm rather than the exception. In fact it is not that *edus* are attached incrementally to *one* partial discourse structure, although the annotators were asked to do so, but rather that multiple partial discourse structures are created and then assembled using a rich variety of operations, which are specific to tree-adjoining and bottom-up parsers. Moreover, even this strategy proves to be somewhat inadequate, since annotators need from time to time to change rhetorical relation labels (2–3% of the operations) and re-structure completely the discourse (1–2% of the operations).

This data suggests that it is unlikely that we will be able to build perfect discourse parsers that can incrementally derive discourse trees without applying any form of backtracking. If humans are unable to decide incrementally, in 100% of the cases, where to attach the *edus*, it is unlikely we can build computer programs that are.

Note.* Estibaliz Amorrortu and Magdalena Romera cotributed equally to this paper.

Note.** The tool described in this paper can be obtained by emailing the first author or by downloading it from *http://www.isi.edu/~marcu/*.

Acknowledgements. We are grateful to Mick O'Donnell for making publically available his discourse annotation tool and to Benjamin Liberman and Ulrich Germann for contributing to the development of the annotation tool described in this paper. We also thank Eduard Hovy, Kevin Knight, and three anonymous reviewers for extensive comments on a previous version of this paper.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht.

D. Lee Ballard, Robert Conrad, and Robert E. Longacre. 1971. The deep and surface grammar of interclausal relations. *Foundations of language*, 4:70–118.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

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

Dan Cristea and Bonnie L. Webber. 1997. Expectations in incremental discourse processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL–97)*, pages 88–95, Madrid, Spain, July 7–12.

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*, volume 3, pages 1965–1968, Madrid, Spain, September.

Claire Gardent. 1997. Discourse TAG. Technical Report CLAUS-Report Nr. 89, Universität des Saarlandes, Saarbrücken, April.

J.E. Grimes. 1975. *The Thread of Discourse*. Mouton, The Hague, Paris.

Barbara Grosz and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Julia B. Hirschberg and Diane Litman. 1987. Now let's talk about *now*: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual*

Meeting of the Association for Computational Linguistics (ACL–87), pages 163–171.

Eduard H. Hovy and Elisabeth Maier. 1993. Parsimonious or profligate: How many and which discourse structure relations? Unpublished Manuscript.

Alistair Knott. 1995. *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* Ph.D. thesis, University of Edinburgh.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations, and common sense entailment. *Linguistics and Philosophy*, 16(5):437–493.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu, 1998. *Instructions for Manually Annotating the Discourse Structures of Texts.*

Daniel Marcu and Eduard Hovy. 1999. Computing the kappa statistic for hierarchical structures. In preparation.

James R. Martin. 1992. *English Text. System and Structure.* John Benjamin Publishing Company, Philadelphia/Amsterdam.

Megan Moser and Johanna D. Moore. 1997. On the correlation of cues with discourse structure: Results from a corpus study. Forthcoming.

Christine H. Nakatani, Julia Hirschberg, and Barbara J. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 106–112, Stanford, CA, March.

Mick O'Donnell. 1997. RST-Tool: An RST analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, Duisburg, Germany, March 24–26.

Rebbeca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–140, March.

Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

Ted J.M. Sanders, Wilbert P.M. Spooren, and Leo G.M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Ted J.M. Sanders, Wilbert P.M. Spooren, and Leo G.M. Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, 4(2):93–133.

Frank Schilder. 1997. Tree discourse grammar, or how to get attached a discourse. In *Proceedings of the Second International Workshop on Computational Semantics (IWCS-II)*, pages 261–273, Tilburg, The Netherlands, January.

Sidney Siegel and N.J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, second edition.

Martin H. van den Berg. 1996. Discourse grammar and dynamic logic. In P. Dekker and M. Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*, pages 93–112. Department of Philosophy, University of Amsterdam.