Using Discourse Analysis and Automatic Text Generation to Study Discourse Cue Usage

Megan Moser
University of Pittsburgh
Learning Research & Development Center
and Department of Linguistics
Pittsburgh, PA 15260
moser@isp.pitt.edu

Johanna D. Moore
University of Pittsburgh
Department of Computer Science and
Learning Research & Development Center
Pittsburgh, PA 15260
jmoore@cs.pitt.edu

Abstract

Our two stage methodology for the study of cue usage coordinates an exhaustive corpus analysis with a system for text generation. Coding of the corpus uses Relational Discourse Analysis, a synthesis of two previous accounts of discourse structure. In the first stage of our study, hypotheses about cue usage are evaluated and refined using the corpus analysis. Several initial results concerning how cues mark segment structure are presented here. In the second stage of our study, the results of the corpus analysis are used to determine a set of heuristics to be implemented in a system for text generation. The automatic generation of texts is then used to exercise and further evaluate the heuristics for cue placement.

Discourse cues play a crucial role in many discourse processing tasks, including plan recognition (Litman & Allen 1987), anaphora resolution (Grosz & Sidner 1986), and generation of coherent multisentential texts (Elhadad & McKeown 1990; Rosner & Stede 1992; Scott & de Souza 1990; Zukerman 1990). Cues are words or phrases such as BECAUSE, FIRST, ALTHOUGH and ALSO that mark structural and semantic relationships between discourse entities. Research in reading comprehension presents a mixed picture (Goldman & Murray 1992; Lorch 1989) suggesting that felicitous use of cues improves comprehension and recall, but that indiscriminate use of cues may have detrimental effects on recall (Millis, Graesser, & Haberlandt 1993) and that the benefit of cues may depend on the subjects' reading skill and level of domain knowledge. However, interpreting the research is problematic because the manipulation of cues both within and across studies has been very unsystematic (Lorch 1989).

In this paper, we describe a methodology for identifying the factors that influence effective cue selection and placement. Our methodology coordinates linguistic analysis of "good texts" with a system for automatic generation of texts conforming to the rules identified by the corpus analysis. Our linguistic analysis, which we call Relational Discourse Analysis (RDA), is a synthesis of two accounts of discourse structure (Grosz & Sidner 1986; Mann & Thompson 1988) previously thought incompatible. We demonstrate the use of the analysis by presenting some hypotheses concerning how cues mark segment structure and the initial results from for hypotheses from our corpus study. The generation system we are implementing will provide a means for evaluation and further refinement of our strategies for cue selection and placement. Our ultimate goal is to provide a text generation component that can be used in a variety of application systems. In addition, the text generator will provide a tool for the systematic construction of materials for reading comprehension experiments.

The study is part of a project to improve the explanation component of a computer system that trains avionics technicians to troubleshoot complex electronic circuitry. The tutoring system gives the student a troubleshooting problem to solve, allows the student to solve the problem with minimal tutor interaction, and then provides a critique of the student's solution in a post-problem session. During this session, the system replays the solution step by step pointing out good aspects of the student's solution as well as ways in which the student's solution could be improved. To determine how to build an automated explanation component, we collected protocols of a human expert tutor providing explanations during the critiquing session. Because the explanation component we are building interacts with users via text and menus, the student and human tutor were required to communicate in written form. In addition, in order to study effective explanation, we chose experts who were rated as excellent tutors by their peers, students, and superiors.

Methodology: Integration of corpus analysis and automatic text generation

Because the recognition of discourse coherence and structure is complex and dependent on many types of non-linguistic knowledge, determining the way in which cues and other linguistic markers aid that recognition is a difficult problem. Our approach to the study of cues proceeds in two complementary stages. The first stage is the corpus study reported on here. The design of our corpus study is coordinated with a sys-

tem for automatic text generation. In the second stage, we generate texts that reflect our conclusions from the corpus study. The evaluation of these texts suggests further exploration of the corpus analysis, the results of which can in turn can be incorporated into the text generation system.

The study of cues must begin with descriptive work using intuition and observation to identify the factors affecting cue usage. Previous research (Hobbs 1985; Grosz & Sidner 1986; Schiffrin 1987; Mann & Thompson 1988; Elhadad & McKeown 1990) suggests that these factors include structural features of the discourse, intentional and informational relations in that structure, givenness of information in the discourse, and syntactic form of discourse constituents. In order to devise an algorithm for cue selection and placement, we must determine how cue usage is affected by combinations of these factors. The corpus study is intended to enable us to gather this information, and is therefore conducted directly in terms of the factors thought responsible for cue placement. Because it is important to detect the contrast between occurrence and nonoccurrence of cues, the corpus study must be be exhaustive, i.e., it must include all of the factors thought to contribute to cue usage and all of the text must be analyzed. From this study, we are deriving a system of hypotheses about cues.

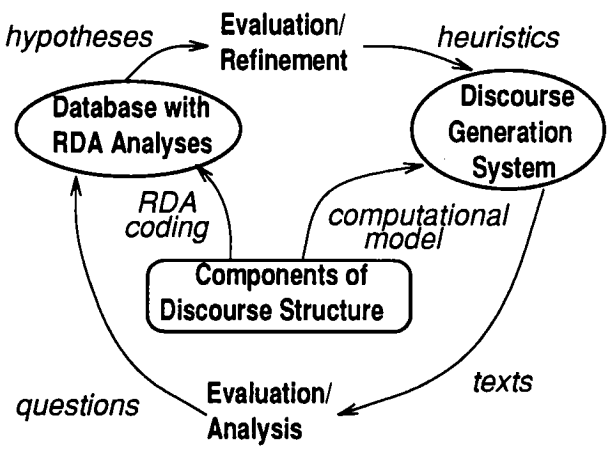

Figure 1: Interleaving of corpus study and evaluation of generated texts

For example, based on observations about a portion of our data, we formulated the hypothesis (discussed further in Section), that cues are more frequent in "wide" segments, where the core, an expression of the idea or action the hearer is intended to adopt, is supported by more than one additional segment constituent that contributes to this adoption. One hypothesis is that cues are placed with every contributor in these wide segments. Note, however, that other factors must determine the particular choice of cue because this hypothesis leaves that choice entirely unspecified. To imagine a simple case, perhaps selecting a cue that is not semantically anomalous (e.g., suggesting causality where it is impossible) is the only additional constraint. If the conclusion of our study is that these two constraints provide the best account of cue placement in our corpus, then our algorithm will place semantically appropriate cues to accompany contributors in wide segments.

Because our corpus analysis is done directly in terms of factors that are modeled by our text generation system, the resulting algorithm can be exercised by automatic generation. This feature allows the corpus analysis and the text generation to interact productively, as suggested by Figure 1. We anticipate that, initially, the texts generated may be obviously flawed. Inspection of these texts may suggest new hypotheses or refinements which can be tested by returning to our tool of corpus analysis and subsequently incorporated into the algorithm. Through this interleaving of corpus analysis and text generation, our algorithm becomes more refined. When the generated texts pass our inspection, they can then be evaluated more thoroughly using panels of human judges and reading comprehension experiments.

Relational Discourse Analysis

In this section we describe our approach to the analysis of a single speaker's discourse, which we call **Relational Discourse Analysis (RDA)**. Applying RDA to a tutor's explanation is exhaustive, i.e., every word in the explanation belongs to exactly one element in the analysis. All elements of the analysis, from the largest constituents of an explanation to the minimal units, are determined by their function in the discourse. A tutor may offer an explanation in multiple **segments**, the topmost constituents of the explanation. Multiple segments arise when a tutor's explanation has several steps, e.g., he may enumerate several reasons why the student's action was inefficient, or he may point out the flaws in the student's step and then describe a better alternative. Each segment originates with an intention of the speaker; it is determined by being a set of clauses that taken together serve a purpose. Segments are internally structured and consist of a **core**, i.e., that element that most directly expresses the segment purpose, and any number of **contributors**, the remaining constituents in the segment each of which plays a role in serving the purpose expressed by the core. For each contributor in a segment, we analyze its relation to the core from an intentional perspective, i.e., how it is intended to support the core, and from an informational perspective, i.e., how its content relates to that of the core. Each segment constituent, both core and contributors, may itself be a segment with a *core:contributor* structure or may be a simpler functional element. There are three types of simpler functional elements: (1) **units**, which are descriptions

of domain states and actions, (2) **matrix** elements, which express a mental attitude, a prescription or an evaluation by embedding another element, and (3) relation **clusters**, which are otherwise like segments except that they have no *core:contributor* structure.

This approach synthesizes ideas which were previously thought incompatible from two theories of discourse structure, the theory proposed by Grosz and Sidner 1986 and Rhetorical Structure Theory (RST) proposed by Mann and Thompson 1988. The idea that the hierarchical segment structure of discourse originates with intentions of the speaker, and thus the defining feature of a segment is that there be a recognizable segment purpose, is due to Grosz and Sidner. The idea that discourse is hierarchically structured by pairwise relations in which one relatum is more central to the speaker's purpose is due to Mann and Thompson. Work by Moore and Pollack 1992 modified the RST assumption that these pairwise relations are unique, demonstrating that intentional and informational relations occur simultaneously. Moser and Moore 1993 point out the correspondence between the relation of dominance among intentions in Grosz and Sidner and the nucleus-satellite distinction in RST. Because our analysis realizes this relation/distinction in a form different from both intention dominance and nuclearity, we have chosen the new terms core and contributor.

To illustrate the application of RDA, consider the partial tutor explanation in (1)[1]. The purpose of this segment is to inform the student that she made the strategy error of testing inside part3 too soon. The constituent that makes the purpose obvious, in this case (1-B), is the core of the segment. The other constituents help to serve the segment purpose by contributing to it. How each contributor relates to the core is analyzed from two perspectives, intentional and informational, as illustrated below. Each constituent may in turn be a segment with its own *core:contributor* structure. For example, (1-C) is a subsegment whose purpose is to give a reason for testing part2 first, namely that part2 is more susceptible to damage and therefore a more likely source of the circuit fault. The core of this subsegment is (1-C.2) because it most directly expresses the purpose. The contributor in (1-C.1) provides a reason for this susceptibility, i.e., that part2 is moved frequently.

Due to space limitations, we can provide only a brief description of *core:contributor* relations, and omit altogether the analysis of the example into the minimal RDA units of state and action units, matrix expressions and clusters. (For more details, see (Moser & Moore In preparation).) A contributor is analyzed for both its intentional and informational relations to its core. Intentional relations describe how a contributor may affect the hearer's adoption of the core. For example, (1-A) acknowledges a fact that might have led the student to make the mistake. Such a **concession** contributes to the hearer's adoption of the core in (1-B) by acknowledging something that might otherwise interfere with this intended effect. Another kind of intentional relation is **evidence**, in which the contributors are intended to increase the hearer's belief in the core. For example, (1-C) through (1-E) each stand in the evidence relation to (1-B). The set of intentional relations in RDA is a modification of the presentational relations of RST. In addition to the intentional relation, each *core:contributor* pair is also analyzed for its informational relation. These relations describe how the situations referred to by the core and contributor are related in the domain. The informational relations are similar to the subject matter relations defined for RST.

The RDA analysis of (1) is shown schematically in Figure 2. As a convention, the core appears as the mother of all the relations it participates in. Each relation is labeled with both its intentional and informational relation, with the order of relata in the label indicating the linear order in discourse. Each relation node has up to two daughters: the cue, if any, and the contributor, in the order they appear in the discourse.

Reliability of Relational Discourse Analysis

To assess inter-coder reliability of RDA analyses, we compared two independent analyses of the same data. Because the results reported in this paper depend only on the structural aspects of the analysis, our reliability assessment is confined to these structural aspects. The categorization of *core:contributor* relations will not be assessed here.

The reliability coder coded one quarter of the currently analyzed corpus, consisting of 132 clauses, 51 segments, and 70 relations. Here we report the percentage of relations in this reliability subset for which the reliability coder agreed with the main coder. To do this, we identified 91 possible points of agreement, or judgment sites, in the main coder's analyses. A judgement site was either a *core:contributor* structure or a cluster, a relation without an intentional distinction. The percentages reported are based on this total of 91 possible agreements.

In analyzing a relation, there are several possible types of disagreements. First, the two coders could analyze a contributor as supporting different cores. This occurred 7 times (92% agreement). Second, the coders could disagree on the core of a segment. This occurred 2 times (98% agreement). Third, the coders could disagree on whether a relatum should be further analyzed into an embedded *core:contributor* structure. This occurred 8 times (91% agreement). Fourth, the coders could disagree on which relation a cue was associated with. This occurred 1 times (99% agreement). Overall, there were 18 disagreements, or 80% agreement.

[1] In order to make the example more intelligible to the reader, we replaced references to parts of the circuit with the simple labels *part1*, *part2* and *part3*.

(1)

ALTHO	A.	you know that part1 is good,
	B.	you should eliminate part2 before troubleshooting inside part3.
THIS IS BECAUSE	C.	
		1. part2 is moved frequently
	AND THUS	2. is more susceptible to damage than part3.
ALSO,	D.	it is more work to open up part3 for testing
AND	E.	the process of opening drawers and extending cards in part3 may induce problems which did not already exist.

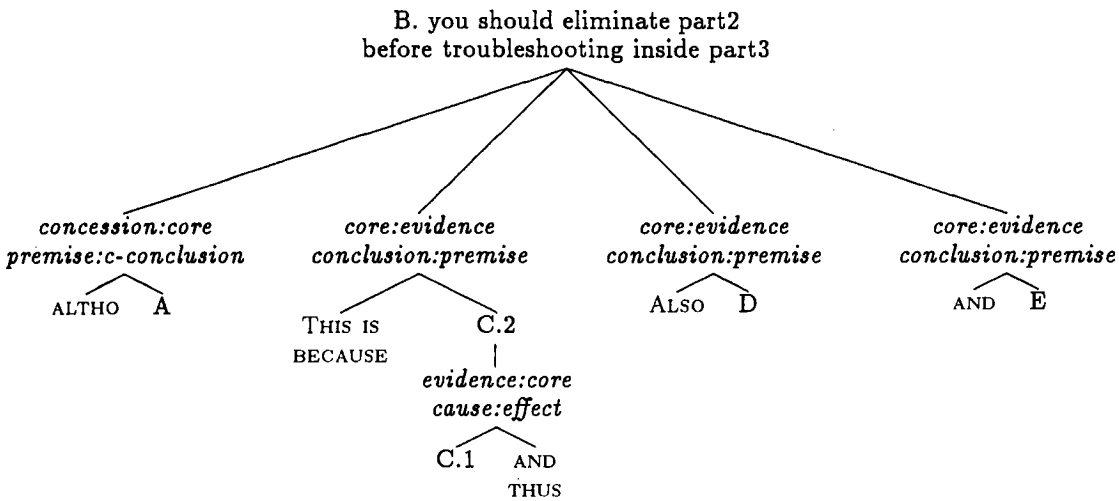

Figure 2: The RDA analysis of (1)

These rates of agreement are similar to those found in studies of (nonembedded) segmentation agreement (Grosz & Hirschberg 1992; Passonneau & Litman 1993; Hearst 1993). However, our assessment of RDA reliability differs from this work in several key ways. For one thing, our subjects/coders are not naive about their task and the data is not spoken. Further, the task is more complex than identifying locations of segment boundaries.

Example hypotheses and initial results

For each tutor explanation in our corpus, a coder analyzes the text as described above, and then enters this analysis into a database. The technique of representing an analysis in a database and then using database queries to test hypotheses is similar to work using RST analyses to investigate the form of purpose clauses (Vander Linden, Cumming, & Martin 1992). Because our analysis is exhaustive, information about both occurrence and nonoccurrence of cues can be retrieved from the database in order to test and modify hypotheses about cue usage. That is, both cue-based and factor-based retrievals are possible. In cue-based retrievals, we use an occurrence of the cue under investigation as the criterion for retrieving the value of its hypothesized descriptive factors. Factor-based retrievals provide information about cues that is unique to this study. In factor-based retrieval, the occurrence of a combination of descriptive factor values is the criteria for retrieving the accompanying cues. Here we illustrate the use of the RDA paradigm with an example hypothesis and the initial results of its query. These results are based on the portion of our corpus that is analyzed and entered into the database, approximately 528 clauses. These clauses comprise 216 segments in which 287 relations were analyzed. Accompanying these relations were 165 cue occurrences, resulting from 39 distinct cues.

The contrast between usage and nonusage of cues is only possible with an exhaustive analysis of the data, and has thus been overlooked by other studies of cues. Yet, given the frequency of cue nonoccurrence, heuristics for cue usage must be able to select no cue as well as a particular cue. Grosz and Sidner 1986 suggest that cues, along with other linguistic features, may function as markers of purely structural aspects of the discourse, such as segment embeddedness. However,

cues are used in only 57% of cases in our corpus where one segment is embedded in another. That is, independent of any particular cue type, nearly as many *core:contributor* relations are without any cue as those that have some cue. Most probably other structural features described by Grosz and Sidner also have a significant number of occurrences without any cue at all. To formulate a testable hypothesis about the choice between usage and nonusage of a cue, we examined a small number of explanations in the corpus. Two factors appeared likely to affect the choice between cue and no cue.

First, we hypothesized that relations in wide segments, i.e., those segments with more than one contributor, are more likely to be cued. A narrow segment, with only one contributor, is a simpler structure that would seem to require less processing for both interpretation and generation and thus have less need of a cue. As shown in Figure 3, this hypothesis is not supported by the data. While there are more cues in wide segments, the difference is just as likely to be by chance as to be significant. If we refine the hypothesis to consider a certain type of cue, the results are significant. Specifically, consider the choice between semantically weak cues, defined as conjunctive (e.g., *and, also*) and enumerative (e.g., *first*) cues, and no cue. As shown in Figure 4, the segment width does affect this choice. Based on this result, we would include in our generation algorithm the heuristic to consider a weak cue with relations in wide segments when no other cue is selected. Note, however, that such relations are more likely to have no cue at all than to have a weak cue. That is, additional factors must be considered before electing to actually use the weak cue. One such factor to be explored in future work is the adjacency of contributor to the core.

Segment width:	Cue used? Yes	No
1	65	59
>1	85	62

Figure 3: The width of segments does not affect choice between cue and no cue ($\chi^2 = 0.926, p < 0.50$).

Segment width:	Cue used: Weak cue	No cue
1	16	59
>1	45	62

Figure 4: The width of segments affects choice between weak cue and no cue ($\chi^2 = 8.498, p < 0.01$).

The second factor we hypothesized affected the choice between cue and no cue is the order of core and contributor in the segment. Intuitively, the most natural order is to tell the hearer what belief or action is recommended and then give support for that belief or action. If the support comes first, then the hearer may be confused about the point of what the speaker is saying. According to this reasoning, a cue is more necessary when the core follows the contributor. Figure 5 shows that this hypothesis is strongly supported by the RDA analyses. Based on this result, we would include in our generation algorithm the heuristic to use a cue when the contributor precedes the core. Again, other factors must also be considered in the choice between cue and no cue.

Core 1st?	Cue used? Yes	No
Yes	73	92
No	77	29

Figure 5: Relative order of core and contributor affects the choice between cue and no cue ($\chi^2 = 21.062, p < 0.001$).

Choosing whether or not to use a cue is only one of many questions concerning cue generation. When a cue is used, there is also the question of its placement in the relation. To address this issue, we noted for each cue occurrence whether it was placed with the first or second relatum and whether it was placed with the core or contributor. A general result is that these two factors describing cue placement are not independent. As shown in Figure 6, the relative order of core and contributor affects whether the cue is placed with the core or contributor. Based on this result, a heuristic for generation would be, when the core precedes the contributor, to always place the cue with the contributor.

Core 1st?	Cue placed with: Core	Contributor
Yes	3	70
No	50	40

Figure 6: Relative order of core and contributor affects the placement of cue ($\chi^2 = 48.617, p < 0.001$).

Next, we categorized cues according to their literal meaning, consistent with the taxonomy identified by (Knott & Dale 1994). Figure 7 shows, for each cue class, the number of cue occurrences in the different positions. Using these two factors to describe cue placement, all the cue classes except for causal cues typically appear in a certain position. Based on these results, a possible heuristic to use in cue generation would be to select a cue and place it in the usual position for its cue class. For enumerative and temporal cues, this

Cue class	Cue placed with: 1st relatum Core	1st relatum Contributor	2nd relatum Core	2nd relatum Contributor	Usually placed with	Examples of class
conjunctive	0	2	12	18	final relatum	*and, also*
enumerative	2	0	0	23	final contributor	*first, thirdly*
causal	0	10	15	15	depends on cue	*because, so*
temporal	0	25	2	1	initial contributor	*since, when*
contrastive	0	2	10	6	final relatum	*but, although*
other	1	1	11	7	final relatum	*in this case*
TOTALS	3	40	50	70		

Figure 7: The distribution of cue classes with respect to cue placement.

would determine the relative order of core and contributor. For the other cue classes, this order would be determined by some other consideration. Or, an alternative heuristic would be to determine the relative order of core and contributor and then use this factor to constrain the choice of a cue.

In this section, we have explored various hypotheses of cue usage relating to the general issue of how cues mark segment structure. Implicit in this discussion are several kinds of heuristics about cue usage: (1) whether or not to use any cue, (2) which cue to use and (3) where to place the cue. In order to implement a cue generation algorithm, we must decide how these three aspects should interact with each other and with other features of the text such as the order of core and contributor and the size of a segment. This decision will be based on further exploration and refinement of hypotheses and on evaluation of texts generated using competing sets of heuristics.

Conclusions

In conclusion, the first stage of our two-stage approach to the study of discourse cues is an exhaustive corpus study applying our RDA approach. RDA is a synthesis of ideas from two theories of discourse structure (Grosz & Sidner 1986; Mann & Thompson 1988). It provides a system for analyzing discourse and for formulating hypotheses about cue selection and placement. Hypotheses are tested by querying a database containing the results of the RDA application. We presented several hypotheses and initial results. Whether or not to use a cue was affected by the segment width (for a weak cue) and the order of core and contributor. All the cue classes except for causal cues typically appear in a certain position described by the dependent factors of whether the cue is placed with the first or second relatum and with the core or contributor. For each result, a possible heuristic for cue usage was suggested.

As hypotheses are tested and refined, we accumulate heuristics for cue usage. In the second stage of our approach, these hypotheses determine an algorithm for cue usage which we are implementing in an automatic text generator. By automatically generating texts we are able to systematically construct large numbers of texts that reflect the predictions of our hypotheses. Evaluation of these texts forms the basis for further exploration of the corpus and subsequent refinement of the algorithm. In addition, our text generation system can be used to create material for reading comprehension experiments in which cues are systematically manipulated, and thus serves a useful tool for more effective research in reading comprehension.

Acknowledgments.

The research described in this paper was supported by the Office of Naval Research, Cognitive and Neural Sciences Division (Grant Number: N00014-91-J-1694), and a grant from the DoD FY93 Augmentation of Awards for Science and Engineering Research Training (ASSERT) Program (Grant Number: N00014-93-I-0812). We are grateful to Erin Glendening for her patient and careful coding and database entry, and to Maria Gordin for her reliability coding.

References

Elhadad, M., and McKeown, K. 1990. Generating connectives. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 97–101.

Goldman, S. R., and Murray, J. D. 1992. Knowledge of connectors as cohesion devices in text: A compar-

ative study of native-english speakers. *Journal of Educational Psychology* 44(4):504–519.

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

Grosz, B. J., and Sidner, C. L. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Hearst, M. 1993. Texttiling: A quantitative approach to discourse segmentation. Technical Report 93/24, U. of Californa, Berkeley.

Hobbs, J. R. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Leland Stanford Junior University, Stanford, California.

Knott, A., and Dale, R. 1994. Using linguistic pheomena to motivate a set of coherence relations. *Discourse Processes* 18(1):35–62.

Litman, D. J., and Allen, J. F. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science* 11:163–200.

Lorch, R. 1989. Text signaling devices and their effects on reading and memory processes. *Educational Psychology Review* 1:209–234.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *TEXT* 8(3):243–281.

Millis, K.; Graesser, A.; and Haberlandt, K. 1993. The impact of connectives on the memory for expository text. *Applied Cognitive Psychology* 7:317–339.

Moore, J. D., and Pollack, M. E. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics* 18(4):537–544.

Moser, M., and Moore, J. D. 1993. Investigating discourse relations. In *Proceedings of the ACL Workshop on Intentionality and Sturcture in Discourse Relations*, 94–98.

Moser, M., and Moore, J. D. In preparation. Corpus analysis of discourse cues. Technical report, University of Pittsburgh, Learning Research and Development Center.

Passonneau, R., and Litman, D. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.*

Rosner, D., and Stede, M. 1992. Customizing RST for the automatic production of technical manuals. In Dale, R.; Hovy, E.; Rosner, D.; and Stock, O., eds., *Proceedings of the Sixth International Workshop on Natural Language Generation*, 199–215. Berlin: Springer-Verlag.

Schiffrin, D. 1987. *Discourse Markers.* New York: Cambridge University Press.

Scott, D., and de Souza, C. S. 1990. Getting the message across in RST-based text generation. In Dale, R.; Mellish, C.; and Zock, M., eds., *Current Research in Natural Language Generation.* New York: Academic Press. 47–73.

Vander Linden, K.; Cumming, S.; and Martin, J. 1992. Expressing local rhetorical relations in instructional text. Technical Report 92-43, University of Colorado. To appear in *Computational Linguistics.*

Zukerman, I. 1990. A predictive approach for the generation of rhetorical devices. *Computational Intelligence* 6(1):25–40.