Identifying the Linguistic Correlates of Rhetorical Relations

Simon H. Corston-Oliver

Microsoft Research
One Microsoft Way
Redmond WA 98052-6399, USA
simonco@microsoft.com

Abstract

RASTA (Rhetorical Structure Theory Analyzer), a system for automatic discourse analysis, reliably identifies rhetorical relations present in written discourse by examining information available in syntactic and logical form analyses. Since there is a many-to-many relationship between rhetorical relations and elements of linguistic form, RASTA identifies relations by the convergence of a number of pieces of evidence, many of which would be insufficient in isolation to reliably identify a relation.

1. Introduction

Within Rhetorical Structure Theory (RST) (Mann and Thompson 1986, 1988), the discourse structure of a text is represented by means of a hierarchical tree diagram in which contiguous text spans are related by labeled relations. Hierarchical structure results from the fact that each text span in a labeled relation may itself have a complex internal discourse structure.

Traditionally, human analysts have constructed RST analyses for texts by employing tacit, subjective, intuitive judgments. RASTA (Corston-Oliver 1998a, 1998b), a discourse analysis component within the Microsoft English Grammar, automatically produces RST analyses of texts. To do so, it proceeds in three stages. In the first stage, RASTA identifies the clauses that function as terminal nodes in an RST analysis. In the second stage, RASTA examines all possible pairs of terminal nodes to determine which discourse relation, if any, might hold between the two nodes. In the third stage, RASTA combines the terminal nodes according to the discourse relations that it hypothesized to form RST analyses of a complete text.

This paper discusses the second stage of processing, during which RASTA identifies discourse relations. Whereas introspection is a viable strategy for human analysts, a computational discourse analysis system like RASTA requires explicit methods for identifying discourse relations. This paper therefore describes (section 2) the kinds of linguistic evidence that RASTA considers in positing discourse structure. Intuitively, cues to discourse relations are not all equally compelling. This intuition is reflected in the use of heuristic scores (section 3) to measure the plausibility of a relation. Section 5 describes in detail the cues used to identify the SEQUENCE relation and gives a worked example. For a more complete description of the workings of RASTA, the reader is referred to Corston-Oliver (1998b).

The Microsoft English Grammar (MEG) is a broad-coverage grammar of English that performs a morphological analysis, a conventional syntactic constituent analysis and a logical form analysis (involving the normalization of syntactic alternations to yield a representation with the flavor of a predicate representation). Functional roles such as subject and object are identified and anaphoric references are resolved during linguistic analysis.

To date, I have focused on the text of *Encarta 96* (Microsoft Corporation 1995, henceforth *Encarta*), a general purpose electronic encyclopedia whose articles exhibit a variety of complex discourse structures. All examples in this paper are taken from *Encarta*. References given are to the titles of articles.

2. Identifying rhetorical relations

In the computational discourse analysis literature, there are three strands concerning the identification of rhetorical relations. The first strand (Knott and Dale 1995; Kurohashi and Nagao 1994; Marcu 1997; Ono et al. 1994; Sanders 1992; Sanders et al. 1992, 1993; Sanders and van Wijk 1996; Sumita et al. 1992) concerns the identification of rhetorical relations by fairly superficial means. Typically simple pattern matching is used to identify cue phrases. These cue phrases are then assumed to be in a one-to-

one relationship to rhetorical relations.

The second strand (Fukumoto and Tsujii 1994; Hobbs 1979), in contrast to the first strand, eschews an examination of the form of a text in favor of reasoning with more abstract representations such as predicate representations of linguistic content and axiomatic representations of world knowledge.

The third strand concerns programmatic descriptions of how a computational discourse analysis might proceed (Polanyi 1988; Wu and Lytinen 1990), with no specific details about how discourse relations might be identified.

RASTA identifies rhetorical relations by directly examining a text, and is therefore most closely aligned with the first of these three strands. Like previous work in this vein, RASTA considers cue phrases to be a useful indicator of rhetorical relations (section 2.3). However, RASTA goes beyond a simple examination of cue phrases and considers such linguistic evidence as clausal status (section 2.1), anaphora, deixis and referential continuity (section 2.2) and tense, aspect, and polarity (section 5).

Traditionally, RST analysts have been averse to tying their analyses of discourse structure to specific elements of linguistic form. The descriptions of rhetorical relations in Mann and Thompson (1986, 1988), for example, studiously avoid all mention of the correlates of discourse structure. This aversion is apparently intended to avoid "naïve mono-functionalism", i.e. the overly-simplistic assumption of a one-to-one mapping between linguistic form and rhetorical structure. This laudable concern is accompanied by a general pessimism. For example, Mann and Thompson (1986:71-72) note that "we do not believe that there are undiscovered signal forms, and we do not believe that text form can ever provide a definitive basis for describing how relational propositions can be discerned." Instead of looking for simple one-to-one mappings between linguistic form and discourse structure, RASTA considers a number of small cues that stand in many-to-many relations to rhetorical relations. By allowing these minor cues to converge in identifying discourse relations, the prospects for identifying rhetorical relations appear rosy, as this paper demonstrates.

2.1. Clausal status

Each RST relation can be classified as a member of one of two structural types: symmetric and asymmetric. Symmetric relations (CONTRAST, JOINT, LIST, SEQUENCE, etc.) consist of two or more *co-nuclei*, each of which is equally important in realizing the writer's communicative goals. Asymmetric relations (CAUSE, ELABORATION, CONCESSION, etc.) have two constituents: a *nucleus*, the more central element in realizing the writer's goals, and a *satellite*, a less important element that is in a dependency relation to the nucleus.

Matthiessen and Thompson (1988) suggest that the grammatical distinction between paratactic clause combining (including coordination, apposition and quoting) and hypotactic clause combining (including various kinds of clausal subordination) represents the grammaticization of the two different kinds of RST relation. This proposal motivates the most important discriminator of rhetorical relations employed by RASTA. Hypotactic clause combining, identified by the syntactic analysis performed by MEG, always suggests an asymmetric RST relation in which the matrix clause is posited to be the nucleus and the subordinate clause to be the satellite. Paratactic clause combining, however, may correspond to either a symmetric or an asymmetric RST relation.

In rare cases, this correlation between clausal status and rhetorical status is the only clue to discourse structure that RASTA is able to identify, i.e. having correctly identified a hypotactic relationship, RASTA is unable to identify a specific corresponding asymmetric rhetorical relation. In such cases, RASTA proposes an asymmetric relationship which it then labels with a question mark, as illustrated in Figure 1. Clause$_2$ is clearly a satellite of Clause$_1$. However, it is not quite clear exactly what RST relation holds. The PURPOSE or RESULT relations are weak candidates, but certainly not inviting enough to warrant a commitment to either.

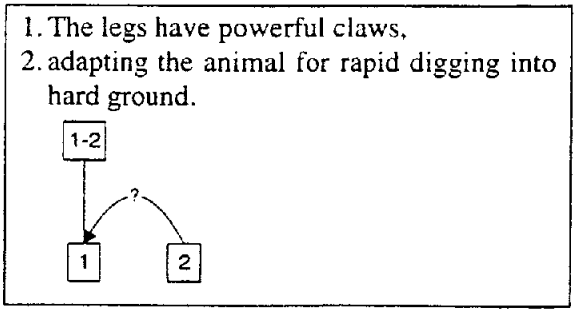

Figure 1 Echidna

2.2. Anaphora, deixis and referential continuity

Anaphora, deixis and referential continuity are strongly cohesive devices (Halliday and Hasan 1976). Often, RASTA need do no more than identify the form of a referring device, without

actually resolving the referent. Pronouns and demonstratives, for example, are frequently positively correlated with the satellite of an asymmetric relation, especially when they occur as syntactic subjects or as modifiers of subjects, and negatively correlated with the co-nucleus of a symmetric relation (see for example criterion 4 for the SEQUENCE relation in Figure 3, section 5).

In other cases, the form of a referring expression is insufficient, and RASTA must consider referential continuity. The MEG system resolves pronominal anaphoric references during the construction of the logical form. Although MEG is sometimes able to identify a single antecedent for a pronoun, it often proposes a list of plausible antecedents. In determining subject continuity, the most important kind of referential continuity for identifying discourse relations, RASTA considers whether the subject of one clause is one of the possible antecedents of the subject of another clause. For a pronominal subject, RASTA examines the list of proposed antecedents. For a subject modified by a possessive pronoun, RASTA considers the proposed antecedents of the possessive pronoun. For lexical subjects, RASTA considers simply whether the head of the subject noun phrase of one clause is identical to the head of the subject noun phrase of the other clause. (MEG does not currently perform anaphora resolution for lexical noun phrases.)

2.3. Cue phrases

Many clauses contain cue phrases that provide evidence of rhetorical structure. Like other approaches to identifying rhetorical structure (Ono et al. 1994; Knott and Dale 1995; Marcu 1997), RASTA recognizes cue phrases as a valuable source of evidence. RASTA, however, attempts to overcome two problems related to cue phrases: compositionality, i.e. some cue phrases are amenable to different compositional analyses, and coverage, i.e. not all clauses contain cue phrases.

Some phrases ought to be treated as lexicalized units in some contexts and as phrases with internal constituency in other contexts. The *Encarta* article *Quasar*, for example, contains the phrase *as long as* in sentence medial position: "...their observed light would have been traveling practically as long as the age of the universe." Such instances of the phrase *as long as* are amenable to a compositional analysis. In other cases, the same phrase in sentence-medial position ought to be treated as a lexicalized unit, analogous to the subordinating conjunction *provided*, as illustrated in Figure 2.

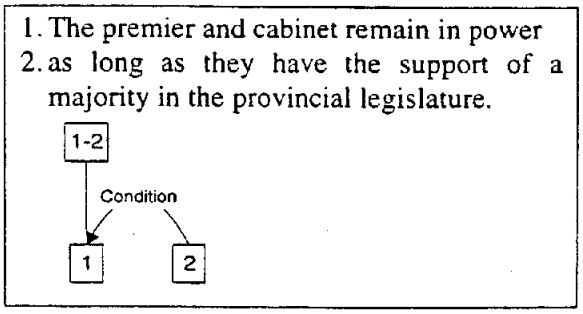

Figure 2 Prince Edward Island

RASTA examines cue phrases as a form of evidence for rhetorical structure, distinguishing ambiguous readings of phrases like *as long as* on the basis of the syntactic analysis performed by MEG.

Unfortunately, it is not the case that all clauses contain useful cue phrases. Cue phrases are therefore insufficient for the task of constructing discourse representations that cover an entire text. To overcome this deficiency, RASTA augments cue phrases with additional evidence available in a text.

3. Necessary criteria and cues

The process of hypothesizing discourse relations involves tension between two competing concerns. On the one hand, it is desirable to postulate all possible discourse relations that might hold between two terminal nodes, in order to ensure that the preferred RST analysis for a text is always in the set of analyses produced by RASTA. On the other hand, considerations of computational efficiency lead us to desire a small set of relations, since as the number of possible discourse relations increases, the number of possible discourse trees to be considered increases exponentially; the smaller the set of hypothesized relations, the more quickly the algorithm for constructing RST trees (Corston-Oliver 1998a, 1998b) can test all possibilities.

RASTA resolves this tension by distinguishing two kinds of evidence. The first kind of evidence is the set of necessary criteria—the conditions that simply must be met before RASTA is even willing to "consider" a given discourse relation. The second kind of evidence is the set of cues that are only applied if the necessary criteria are satisfied. Coordination by means of the conjunction *and*, for example, correlates with the SEQUENCE conjunction (Figure 6, section 5), but only weakly. If we were to posit a SEQUENCE relation every time we observed the conjunction *and*, we would posit a great many spurious relations. However, RASTA only tests this cue if

an extensive set of necessary criteria for the SEQUENCE relation have been satisfied (Figure 3, section 5).

The cues that RASTA uses to identify rhetorical relations by no means constitute an exhaustive list of the correlates of each relation. Rather, the cues that RASTA employs are sufficient to enable it to distinguish reliably among the thirteen relations (ASYMMETRICCONTRAST, CAUSE, CIRCUMSTANCE, CONCESSION, CONDITION, CONTRAST, ELABORATION, JOINT, LIST, MEANS, PURPOSE, RESULT, SEQUENCE) necessary for an adequate discourse analysis of the text of the articles in *Encarta*. The extent to which the cues used by RASTA correspond to the evidence that human readers use when attempting to understand the discourse structure of a text is a matter for independent experimental investigation.

4. Heuristic scores

RASTA examines many cues in identifying rhetorical relations. Intuitively, these different cues are not of equal weight. To reflect this intuition, RASTA associates a heuristic score with each cue. Each cue is thus able to "vote" for a relation. Each relation receives a score, equal to the sum of the heuristic scores of the cues that voted in favor of that relation.

The heuristic scores assigned accord well with human linguistic intuitions. However, the primary role of the heuristic scores is to guide RASTA in subsequent stages of processing. When constructing RST trees, RASTA applies the relations with the highest scores first. This causes RASTA to converge on better analyses of a text before producing less plausible analyses (Corston-Oliver 1998a, 1998b).

5. The SEQUENCE relation

To illustrate the kinds of evidence that RASTA considers, let us consider how RASTA identifies the SEQUENCE relation. The SEQUENCE relation is a symmetric relation in which two or more clauses report events that are in a relationship of temporal succession. Figure 3 gives the necessary criteria for the SEQUENCE relation. If the necessary criteria are satisfied, then it is reasonable to posit a SEQUENCE relation between two clauses. The criteria are sufficiently stringent that an initial heuristic score of 20 is associated with this hypothesized relation.

1. $Clause_1$ precedes $Clause_2$.
2. $Clause_1$ is not syntactically subordinate to $Clause_2$.
3. $Clause_2$ is not syntactically subordinate to $Clause_1$.
4. The subject of $Clause_2$ is not a demonstrative pronoun, nor is it modified by a demonstrative.
5. Neither $Clause_1$ nor $Clause_2$ has progressive aspect (marked by the *-ing* verbal suffix).
6. If either $Clause_1$ nor $Clause_2$ has negative polarity, then it must also have an explicit indication of time.
7. Neither $Clause_1$ nor $Clause_2$ is a Wh question.
8. Neither $Clause_1$ nor $Clause_2$ has an attributive predicate.
9. The event expressed in $Clause_2$ does not temporally precede the event in $Clause_1$.
10. $Clause_1$ and $Clause_2$ match in tense and aspect.
11. $Clause_2$ must not be immediately governed by a contrast conjunction.

Figure 3 Necessary criteria for the SEQUENCE relation

A few of the necessary criteria merit special discussion. Criteria 2 and 3 are intended to exclude situations in which one clause is syntactically subordinate to another, since a relationship of grammatical subordination always corresponds to an asymmetric relation (section 2.1), whereas the SEQUENCE relation is a symmetric relation.

Criterion 4, "The subject of $Clause_2$ is not a demonstrative pronoun, nor is it modified by a demonstrative", is intended to block cases in which the correlations between deixis and discourse structure (section 2.2) would make an asymmetric relation more likely than the symmetric SEQUENCE relation. For example, in the following excerpt, a SEQUENCE relation is dispreferred in the face of a more plausible RESULT relation because the subject of the second main clause, *this study*, contains a demonstrative:

> "He made a study of the famous Adams family of Massachusetts, to which he was not related; **this study** resulted in "The Adams Family"..." (*Adams, James Truslow*).

Since the SEQUENCE relation involves a narrative sequence of events, criterion 5, "Neither Clause$_1$ nor Clause$_2$ has progressive aspect...", excludes clauses which are not eventive, as in the following example:

> "Abbott was willing to admit a number of manufactured goods from the United States duty-free" (*Abbot, Sir John Joseph Caldwell*).

For the most part, clauses with negative polarity do not express events and therefore cannot enter into the SEQUENCE relation. One notable exception to this generalization is clauses with negative polarity which also contain an explicit indication of time (criterion 6), as illustrated in Figure 4. The clause with negative polarity and an explicit indication of time is given in bold type. This clause entails an event which is in a SEQUENCE relation with other events. RASTA does not require a sophisticated reasoning module to detect this entailment. Rather, the mere presence of an explicit indication of time within a negative clause appears to be sufficient to identify the entailment in this instance and in other similar instances. Prepositional phrases and subordinate clauses introduced by *before* or *until* are the most common means of explicitly indicating time for clauses with negative polarity in *Encarta*.

1. Although AIDS has been tracked since 1981,
2. **the identification of HIV as the causative agent was not made until 1983.**
3. In 1985 the first blood test for HIV, developed by the research group led by Robert Gallo, was approved for use in blood banks.

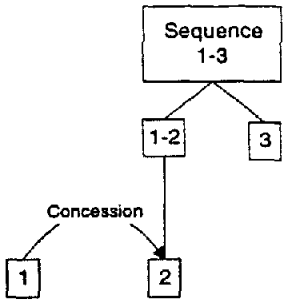

Figure 4 Acquired Immune Deficiency Syndrome

Neither Wh questions (criterion 7) nor attributive predicates (criterion 8) report events. They therefore cannot participate in SEQUENCE relations. Changes in state, unlike attributive predicates, can however participate in SEQUENCE relations. Clause 2 in Figure 5, *and [Abacha] became a captain in the army in 1967*, illustrates a change of state.

1. Abacha graduated from the Nigerian Military Training College in Zaria in 1963,
2. and became a captain in the army in 1967.

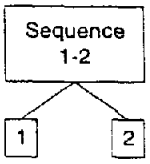

Figure 5 Abacha, Sani

Criteria 1 and 9 together constitute the traditional minimal definition of a narrative (Labov 1972; Reinhart 1984): a narrative sequence is one in which a series of tensed clauses reports a sequence of events, with the linear order of the clauses expressing the events matching the real-world temporal order of those events.

Provided that the necessary criteria for the SEQUENCE relation are satisfied, RASTA tests the cues given in Figure 6.

1. Clause$_2$ contains a sequence conjunction (*and later then...*). Score = 10.
2. Clause$_1$ and Clause$_2$ are coordinated. Score = 5.
3. There is an explicit indication that the event expressed by Clause$_1$ temporally precedes the event expressed by Clause$_2$. Score = 5.

Figure 6 Cues for the SEQUENCE relation

Note that RASTA does identify SEQUENCE conjunctions (cue 1). However, the presence of a SEQUENCE conjunction is not a necessary criterion in identifying the SEQUENCE relation.

Explicit indications of time are of great value in identifying the SEQUENCE relation (criterion 9 and cue 3). In Figure 7, for example, the events described in clauses 2 through 7—conferences being held, agreements being made, and so on—occur during the 1920s, the timeframe described in clause 1. RASTA identifies the timeframe of the expression *the 1920s* by the presence of a definite

article with a numeric year, together with the presence of the plural suffix *-s*. The timeframe thus identified spans the first day of 1920 to the last day of 1929. It is a matter of simple math to determine that the dates 1920 (clause 2), 1921-1922 (clause 4), 1925 (clause 5) and 1928 (clause 6) fall within this interval.

Clause 1 describes a temporal interval within which the events described in clauses 2 through 7 occur, rather than describing any event that precedes the events in clauses 2 through 7. RASTA therefore does not posit a SEQUENCE relation between clause 1 and any of the following clauses. Rather, clause 1, the topic sentence of this paragraph, is in an ELABORATION relation with the SEQUENCE node that spans clauses 2 through 7.

Clauses 2 through 7 satisfy criterion 9, since the temporal order of the events described matches the temporal order of the events in the world and none of the clauses describes a temporal interval within which the events of any other clauses occur. Cue 3 identifies the appropriate sequencing of the temporal expressions in each of the relevant clauses, leading RASTA to posit the SEQUENCE node depicted in Figure 7.

6. Conclusion

RASTA posits plausible rhetorical relations between clauses by identifying the linguistic correlates of rhetorical relations. The evidence that RASTA examines goes beyond cue phrases, including such cues as clausal status, anaphora, deixis and referential continuity.

The form of a text represents the sum of a number of the decisions made by a writer. These decisions include the rhetorical structuring of the text, motivating the choice of linguistic devices such as specific grammatical constructions and tense and aspect sequencing. By examining the linguistic form of a text, we are able to make plausible inferences about rhetorical structure. Even subtle entailments (criterion 6, Figure 3, section 5) can be identified by an examination of linguistic form alone.

RASTA allows for a many-to-many mapping between elements of linguistic form and specific rhetorical relations. Specific relations are identified by the convergence of multiple pieces of evidence. Future research in this vein will seek to mine the wealth of information present in a text for more cues to rhetorical structure.

1. During the 1920s, attempts were made to achieve a stable peace.
2. The first was the establishment (1920) of the League of Nations as a forum in which nations could settle their disputes.
3. The league's powers were limited to persuasion and various levels of moral and economic sanctions that the members were free to carry out as they saw fit.
4. At the Washington Conference of 1921-22, the principal naval powers agreed to limit their navies according to a fixed ratio.
5. The Locarno Conference (1925) produced a treaty guarantee of the German-French boundary and an arbitration agreement between Germany and Poland.
6. In the Paris Peace Pact (1928), 63 countries, including all the great powers except the USSR, renounced war as an instrument of national policy
7. and pledged to resolve all disputes among them "by pacific means."

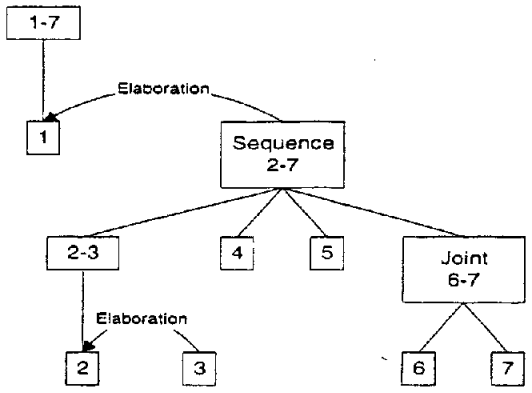

Figure 7 World War II

References

Corston-Oliver, Simon. 1998a. Beyond String Matching and Cue Phrases: Improving Efficiency and Coverage in Discourse Analysis. In Proceedings of the AAAI 1998 Spring Symposium Series, Intelligent Text Summarization. March 23-25, 1998.

Corston-Oliver, Simon. 1998b. Computing Representations of the Structure of Written

Discourse. Ph.D. dissertation. University of California, Santa Barbara, U.S.A.

Fukumoto, J. and Tsujii, J. 1994. Breaking down rhetorical relations for the purpose of analyzing discourse structures. In COLING 94: The Proceedings of the 15th International Conference on Computational Linguistics, vol. 2:1177-1183.

Hobbs, J. R. 1979. Coherence and coreference. Cognitive Science 3:67-90.

Kurohashi, S. and Nagao, M. 1994. Automatic detection of discourse structure by checking surface information in sentences. In Proceedings of COLING 94: The 15th International Conference on Computational Linguistics, vol. 2:1123-1127.

Knott, Alistair and Robert Dale. 1995. Using linguistic phenomena to motivate a set of coherence relations. Discourse Processes 18:35-62.

Labov, W. 1972. Language in the Inner City: Studies in the Black English Vernacular—Conduct and Communication. Philadelphia: University of Pennsylvania Press.

Mann, W. C. and Thompson, S. A. 1986. Relational Propositions in Discourse. Discourse Processes 9:57-90.

Mann, W. C. and Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8:243-281.

Marcu, D. 1997. The Rhetorical Parsing of Natural Language Texts. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL-97), 96-103.

Matthiessen, C. and Thompson, S. A. 1988. The structure of discourse and 'subordination'. In Haiman, J. and Thompson, S. A., (eds.). 1988. Clause Combining in Grammar and Discourse. John Benjamins: Amsterdam and Philadelphia. 275-329.

Microsoft Corporation. 1995. Encarta® 96 Encyclopedia. Redmond: Microsoft.

Ono, K., Sumita, K. and Miike, S. 1994. Abstract generation based on rhetorical structure extraction. In Proceedings of COLING 94: The 15th International Conference on Computational Linguistics, vol. 2:344-348.

Polanyi, L. 1988. A formal model of the structure of discourse. Journal of Pragmatics 12:601-638.

Sanders, Ted J.M. 1992. Discourse Structure and Coherence: Aspects of a Cognitive Theory of Discourse Representation. Lundegem: Nevelland.

Sanders, Ted J.M., W.P.M Spooren and L.G.M. Noordman. 1992. Toward a taxonomy of coherence relations. Discourse Processes 15:1-35

Sanders, Ted J.M., W.P.M Spooren and L.G.M. Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. Cognitive Linguistics 4:93-133.

Sanders, T. J. M. and Wijk, C. van. 1996. PISA: A procedure for analyzing the structure of explanatory texts. Text 16:91-132.

Sumita, K., Ono, K., Chino, T., Ukita, T. and Amano, S. 1992. A discourse structure analyzer for Japanese text. In Proceedings of the International Conference of Fifth Generation Computer Systems, 1133-1140.

Wu, H. J. P. and Lytinen, S. L. 1990. Coherence relation reasoning in persuasive discourse. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, 503-510.