1 Introduction

In the study of discourse analysis, there are innumerable theories and methodologies for the representation of discourse structure. Two of the more influential of these to arise in the past two decades are Rhetorical Structure Theory (RST) (Mann & Thompson 88) and the approach described in Grosz & Sidner (1986), informally known as Grosz and Sidner Theory (GST). In this paper, we briefly examine the main features of each theory, and the discussion in the literature motivating a combination of the two. After examining a previous attempt at merging the two, we put forth a new proposal for the combination of both theories, illustrating with one example text, and a brief discussion of possible applications for dialogue.

2 Two Approaches to Discourse

2.1 Rhetorical Structure Theory

Rhetorical Structure Theory, as formulated in Mann & Thompson (1988), is a relatively fine grained approach, in which text is segmented at the clause level, and clauses are combined into larger text spans by discourse relations. As more and more complex spans are combined, the final result is a tree-structure which encompasses the entire text. In the initial presentation, there are twenty-four discourse relations with which the rhetorical structure can be constructed, however, this is an open set, and later implementations have exploited expanded sets of relations. Crucial to the construction of the structure is the notion of nuclearity; relations are defined in terms of a nucleus and a satellite, and conditions which must be met for the relation to be said to hold between two spans. In general terms, the satellite contributes to the nucleus, and the nature of this contribution is given by the rhetorical relation binding
the two. Simple connectives, such as and or but are realized as multi-nuclear relations, along with a relation for simple temporal sequencing. There are two major classes of relations: intentional and informational. For the informational relations, the reader/hearer is merely expected to recognize the relation. Informational relations, on the other hand, encode more complex relationships, such as justification or motivation, whereby some inclination in the reader/hearer is being modified.

Restrictions on compositionality are introduced in a first-order formalization of RST (Marcu 1996), intended as a first step for automated parsing of natural language texts; these restrictions come into play with the notion of the promotion set. As two clauses are merged to form a larger text span, one will be a nucleus, and the other a satellite; the nucleus of this span will be designated the span's promotion set. As spans combine, the resulting larger span will always adopt the promotion set of its nucleus, and the promotion set of the root node, encompassing the whole text, will be the most important part of the text. Furthermore, as complex spans are merged, the stipulation is added that for a relation to hold between two spans, it must independently hold between the two promotion sets. By adding this stipulation, spurious analyses are avoided. However, it must be noted that it is a widely discussed fact that for any given text, there are multiple possible analyses. One guideline for getting the “right” analyses given in Mann & Thompson (1988) is that when a situation is encountered where either an intentional relation or an informational relation could hold, the intentional one should be chosen.

2.2 Grosz and Sidner Theory

In contrast to RST, Grosz and Sidner Theory (GST) employs a coarser and, on the whole, less-constrained discourse segmentation. Aside from the whole discourse, labeled Discourse Segment (DS) 0, there are no fixed guidelines for setting the size or boundaries of smaller segments. Further distinguishing this approach from RST, GST uses only two discourse relations, and each relation ties directly to the structural description of the discourse. Additionally, each discourse segment is linked to an intention, and the interplay between segments is determined by the interplay between their respective intentions. Formally, these discourse segment intentions are given as
an ordered list, expressed using first-order predicate logic notation. Dominance (DOM) is an embedding relation, wherein the satisfaction of the dominated segment's intention contributes to the satisfaction of the intention of the segment in which it is embedded. Satisfaction-Precedence (SP) is an ordering relation, whereby the satisfaction of one segment's intention must precede the satisfaction of the next segment's intention. Structurally, this appears as a non-relation, with the two related segments existing as sisters at the same level of embedding. This is ambiguous however, as not every pair of sisters at the same level of representation is necessarily related by SP; it could simply be that two unrelated segments are both dominated by one common super-ordinate segment.

A complete departure from RST comes in the introduction of GST's attentional structure. The attentional structural is conceived as a stack, wherein each discourse segment occupies a distinct focus space. As a segment begins, its focus space is pushed onto the stack, and will remain there until that segment's intention has been fulfilled. Thus, in the case of dominance, the focus space of the dominating segment can not be popped from the focus stack until all of its embedded segments' focus spaces have been popped first. With all of this machinery in place, GST allows for the simultaneous expression of discourse structure, intentions, and the focus attention of the discourse participants.

3 Merging Approaches

3.1 Common Ground
As indicated by the above discussion, both RST and GST have a good deal of exploitable common ground. Both break discourse down into a constituent structure, with this structure being somewhat related to the intentions of the speaker. However, in RST the nature of the relations between text spans is of key importance to the analysis, while in GST it is the intentions underlying these relations which receives greater expression. This contrast is what has led to claims that a merger of these two approaches would be advantageous.
3.2 The Intentional/Informational Divide

In Moore & Pollack (1992), the underspecification of the intentional structure of discourse is noted as a problem for RST. This is complicated by their claim that a given discourse can have an intentional structure and a distinct, non-isomorphic, informational structure. A clear example of this claim is illustrated in the following two line discourse from Moser & Moore (1996):

(1) a. Come home by 5:00
   b. Then we can go to the store before it closes.

Here, Moser & Moore claim that there are two possible analyses, one with an informational relation \textit{MOTIVATION}, where $a$ is the nucleus and $b$ the satellite, and another where $b$ is the nucleus of a \textit{CONDITION} relation with $a$ as the satellite. Unlike the original formulation of RST, both Moore & Pollack and Moser & Moore claim that the informational relation is necessary to a complete representation of the discourse structure, while the guidelines given in Mann & Thompson (1988) state that the intentional relation should be chosen over the informational one.

In the end, after a more thorough discussion of the common ground between the two theories, Moser and Moore reach the conclusion that it would be desirable to synthesize a new approach which combines GST's rich intentional structure with RST's informational structure representations. The exact nature of this synthesis, and the answer to the important question of what and how many relations it should employ is left for further research.

3.3 One Attempt

In response to this “challenge” by Moser & Moore, there has been one attempt at unifying the two theories. This approach exploits another piece of common ground between the two theories which has so far not been discussed. In Moser & Moore (1996), the dominance relation of GST is said to be captured by the nucleus-satellite
relationship of RST. In essence, every RST satellite is, in a GST sense, dominated by the larger text span containing that satellite and its related nucleus.

In Marcu (1999), this structural correspondence forms the basis of a synthesized approach. The synthesis is purely algorithmic, building a GST relation from a previously established RST structure for a given discourse. Given an RST structure, well-formed according to the standards of Marcu (1996), for each text span formed by a nucleus-satellite relation, a GST dominance relation is stated in which the whole span dominates the satellite. Satisfaction-Precedence is identified as a relation holding between paratactic (multi-nuclear) relations, but the formalization of this is expressed only such that all incidences of SP are between nuclei; there is no way to determine precisely which paratactic relations are in fact SP. To capture the intentional structure, an additional two constraints are added to the formal definition of RST. One simply claiming that the primary intention of any discourse segment is unique, and the other that this intention is either \texttt{NONE}, or a function of the promotion sets of the dominating and subordinate segments, and the rhetorical relation holding between them.

3.4 \textit{Wish List for a New Attempt}

While it does have merit, it can be argued that the approach advocated by Marcu (1999) “over-generates.” One of the key distinctions between GST and RST is the granularity of the analysis, so to speak. RST chops up a discourse clause by clause, whereas GST’s structure highlights only the most intentionally salient discourse segments. By piggy-backing GST onto an RST structure, Marcu generates a GST structure which is most likely of a much finer grain than a traditional GST analysis of the same discourse. Furthermore, these two structures will be essentially isomorphic, ignoring the possibility that different structures may arise depending on the choice of either an informational or an intentional focus.

Hearkening back to the title of the original Moore & Pollack paper which spurred this discussion, perhaps a multi-level representation is required: one level to capture the intentional structure of a discourse, and another for a possibly non-isomorphic informational structure. This ability to capture non-isomorphic structures
would be the first major requirement of a new merged approach. Secondly, at least at the informational level, it would be ideal to retain RST's level of granularity. Thirdly, the new approach should retain GST's tracking of attentional states. Finally, the merged approach should be equally well-suited to both text and conversation analysis. In their original formations, RST is generally more well-suited to the former, and GST the latter.

4 A New Approach

4.1 Basis of the New Merger

The new approach to merging RST and GST takes advantage of one additional aspect of GST, as established by Grosz & Sidner (1986), which has not yet been explored. There, while the main thrust of the discussion is on the coarser level of analysis already described, there is mention of possible relations between utterances, even utterance level intentions, separate from the higher level discourse segment intentions. We theorize that these intra-segmental relations could be characterized by RST's informational relations, and that the dominance relations in GST will correspond to intentional relations in RST. Should this be the case, then there is now room for two levels of representation, split along the intentional/informational axis in the set of RST relations. The intentional relations could be captured in a manner similar to GST, being recorded as a list of intentions expressed in first order logical form, derived from a structure which feeds the attentional component of the analysis. In practice, the implementation of this analysis would be similar to that of Marcu's approach. The intentional structure is built up from an RST analysis, but only in a limited fashion. The merged system would begin with an RST analysis following the original guidelines of Mann & Thompson, selecting intentional relations over informational ones. However, there would most likely be cases where only an informational relation would hold; we predict this would be at a lower level in the analysis, leaving a coarse-grained intentional structure. Once this is recorded a la GST, the text could be submitted to re-analysis, using informational relations only, potentially yielding a non-isomorphic structure. This model directly captures the first
three items on the wish list above: non-isomorphism between intentional structures and informational, an utterance-level informational analysis, and mechanisms for retaining GST's attentional structure. To test the performance of this merged system, and address the last remaining issue on the wish list, case studies will need to be conducted.

4.2 Testing the System

The first test of this merged system will be upon a piece of text known as “The Movies Text,” drawn from Grosz & Sidner (1986). This text was selected for its relatively short size, and because there is an “official” GST analysis to serve as a basis of comparison. If it emerges that an RST analysis heavy on intentional relations identifies the same discourse segments as the existing GST analysis, then we will take this as evidence that our system is genuinely capturing the intentional structure as well as GST, although with an enhanced set of relations, as all of RST's intentional relations will replace DOMINANCE. The original GST analysis of this particular text contains no instances of SP, however it will still be worth looking for any correspondences between paratactic relations in the RST analysis and any discourse segments in the GST analysis which are sisters under a common dominating segment. The text is presented below, along with Grosz and Sidner's original segmentation into utterances and larger discourse segments:
Of immediate interest here is the identification of 16 as the least embedded segment. In our system, this should correspond to an RST structure wherein 16 is the nucleus of the least-embedded intentional relation. To check if this is indeed the case, and to examine other correspondences, the first step of our system, the intentional RST analysis is presented below:
Indeed, the RST analysis does indicate that the first 15 utterances serve as evidence for the claim made in 16. However, while the GST analysis treats spans 1-3 and 4-15 as discrete, unrelated entities, RST identifies a more complex internal structure. An exact match between structures would be impossible, as the span 4-15 in the GST analysis has two unembedded utterances, 4 and 15, which surround subordinated material. A structure with essentially two discontinuous nuclei and one satellite is not permitted under RST compositionality rules, however RST does identify both 4 and 15 as nuclei of their respective spans. For the more deeply embedded segments, the structures given by the RST analysis correspond exactly to the original GST: unrelated sister pairs 5-6~7-14 and 9-10~11-14 are picked out as paratactic informational relations, and the internal structure of 11-14 identifies 11 as the unembedded nucleus, with 12-14 as a subordinated satellite. Given this degree of correspondence, the analysis seems to be on the right track. The need for two levels of non-isomorphic representation emerges when examining the informational structure of the text.

When presented with the text, people asked to identify the most salient utterance in the text give two distinct answers. The first is simply to say “movies are bad,” identifying utterance 16 as above. However, utterance 4, the central question underlying the discussion, is also identified. What is striking is that in an RST
analysis using only the informational relations, it is this utterance which emerges as the least embedded nucleus:

Thus, by having a system which allows for two separate representations, this important generalization can be captured. The status of the answer, utterance 16, as the intentional nucleus is recorded at a higher level of representation where the initial RST intentional relations are recorded *a la* GST, using a simple notation of the form `RELATION(satellite)(nucleus):`
The informational relation analysis can stand as is, serving as the second level of representation.

That this system can incorporate GST's representation of focus can be shown using a simple, more embedded relation. Specifically, in the internal structure of utterances 11-14. In the GST analysis, this is illustrated as one discourse segment dominating another, 12-14. In the first GST analysis, the span 11-14 is formed by an evidence relation, with 11 as the nucleus and 12-14 as the satellite. After the first parse, this would be represented as shown in (6):

\[(6)\text{EVIDENCE}(12-14)(11)\]

In effect, this identifies three distinct pieces of structure: the spans 11, 12-14, and the resulting superordinate 11-14. The first two fall right out of the notation, but the third must be calculated from determining the longest span that can be formed out of the two components. As RST requires members of the relations to be adjacent, this should be an easy task. In terms of the attentional structure, this represents two focus spaces: one formed by the superordinate span, and one formed by the satellite. As utterance 11 is encountered, the superordinate focus space is pushed onto the focus stack, with another representing 12-14 being pushed directly on top. Thus, as 14 is reached, 12-14 can be popped off, immediately followed by the superordinate focus space. With the intentional relations identified by the first RST analysis functioning in this way, a full account of the attentional structure of a given text can be given, as in a traditional GST analysis.

5 Dialogue and the Road Ahead

It is in this ability to capture the attentional structure of discourse that the merged system can meet the last item on our earlier-posed wish list. By virtue of its roots in RST, the merged system can easily be demonstrated to be good for handling text.
However, by this same token, it should have problems with spoken dialogue, especially where more than one speaker is involved. As a brief illustration of how this merged system would be superior to conventional RST, we discuss a conversational phenomenon which RST can not cope with: interruptions. In everyday speech, a given conversation may not always stay on track, and there can be multi-utterance digressions on a separate topic before the original conversational thread is picked up again. These interruptions would confound RST, as they could quite possible have no relation whatsoever with adjacent segments, resulting in an ill-formed tree. Conversely, GST handles this phenomenon quite elegantly by means of the attentional structure. Non-permeable focus spaces related to the digression are pushed onto the stack, and once popped, the original focus continues as normal. In effect, the attentional structure allows for the original discourse focus to be saved.

In the merged system proposed here, the same mechanism can be used along with modified rules for discourse-tree well-formedness. If, on the initial RST parse, an unrelated span of utterances is discovered within the larger discourse, then it can be treated essentially as an island, possibly with its own complex internal structure. By employing the notion of the attentional structure as “saving” the thread of discourse through an interruption, our modified RST would now allow for discontinuous spans, greatly expanding the range of genres which can be represented. Of course, the exact conditions for how these islands could be formalized and function will still need to be determined, as the attentional structure described in 4.2 still relies on discourse relations, whereas the phenomenon being discussed here is characterized by a complete lack of relation.

Tailoring this approach to dialogue, while already seeming more possible than simple RST, will require a good deal of future work. Indeed, test analyses with various genres of discourse will be needed to establish the approach given here as a truly viable means of representing discourse. However, given the results found here, it seems that this will be a fruitful line of research, as the correspondences between GST and RST are numerous enough to be worth noting, and the fact that in the one test analysis done so far, the system correctly identifies two different nuclei for the
discourse is a sure sign that there is merit to adopting a multi-level approach to discourse representation.

References


