Deriving rhetorical complexity data from the RST-DT Corpus

Sandra Williams, Richard Power

The Open University
Milton Keynes, MK7 6AA, U.K.
s.h.williams@open.ac.uk, r.power@open.ac.uk

Abstract

This paper describes a study of the levels at which different rhetorical relations occur in rhetorical structure trees. In a previous empirical study (Williams and Reiter, 2003) of the RST-DT (Rhetorical Structure Theory Discourse Treebank) Corpus (Carlson et al., 2003), we noticed that certain rhetorical relations tended to occur more frequently at higher levels in a rhetorical structure tree, whereas others seemed to occur more often at lower levels. The present study takes a closer look at the data, partly to test this observation, and partly to investigate related issues such as the relative complexity of satellite and nucleus for each type of relation. One practical application of this investigation would be to guide discourse planning in Natural Language Generation (NLG), so that it reflects more accurately the structures found in documents written by human authors. We present our preliminary findings and discuss their relevance for discourse planning.

1. Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) has been adopted in many NLG systems as a convenient theory of discourse structure (Power et al., 2003; O'Donnell et al., 2001). Briefly, the theory claims that texts express a hierarchical rhetorical structure (see Figure 1), formally represented by a tree in which non-terminal nodes are labelled with rhetorical relations, and terminal nodes are Elementary Discourse Units (EDUs); the latter are typically realised by clauses, and are assumed to have no internal rhetorical structure. Near the bottom of the tree, rhetorical relationships will typically be expressed within sentences, but at higher levels the nodes can embrace large spans of text such as paragraphs or even sections. The original theory defines some 30 rhetorical relations, subclassified into Nucleus-Satellite (two arguments of unequal importance), and Multi-nuclear (two or more arguments of equal importance); the Nucleus-Satellite relations are further subdivided into Subject-Matter (concerning the semantic domain) and Presentational (concerning the author's intentions). The RST-DT Corpus is a corpus of Wall Street Journal articles annotated with RST relations, using a similar but enlarged relation set, with the tree structure fully specified all the way from the root node (governing a whole document) to the EDUs.

An interesting consequence of RST is that we can associate each node in the tree (and its associated text span) with an indicator of its *rhetorical complexity*. The simplest and most direct measure of rhetorical complexity is the number of EDUs beneath the node. In Figure 1, the rhetorical complexity of the RESULT relation is 5, since there are five EDUs, numbered 49 to 53, beneath it; the complexity of the LIST relation is 3; and so on. The minimal rhetorical complexity for a non-terminal node is then 2, with no upper limit on the maximum. Using this concept, we can investigate quantitatively the observations mentioned earlier — for instance, that some relations tend to occur higher up the RST tree than others. We know of no previous study that has looked directly at this issue, even though the relevant data are easily recoverable from existing corpora. The RST-DT corpus has been used in discourse parsing (Marcu and Echihabi, 2002), but mostly with the aim of detecting low-level structure using discourse connectives. Studies of higher-level structure (Sporleder and Lascarides, 2004) have looked at the rhetorical roles played by paragraphs and sections, but without exploiting corpus markup of the RST span structure using a notion of rhetorical complexity.

2. Issues

In an initial exploration of the data, we look at five issues.

1. *Does rhetorical complexity vary for different relations?* This issue is fundamental to all the others: if complexity is randomly distributed, then we have nothing further to investigate.

2. *Does rhetorical complexity vary for different classes of relations?* In particular, we are interested in whether there are differences among the three main groups identified in RST: subject-matter, presentational, and multi-nuclear.

3. *Is there a typical complexity for each relation, or do complexities spread across a wide spectrum with no tendency to focus?* The answer to this question might vary from one relation to another: some might have a typical complexity, some not.

4. W*ithin nucleus-satellite relations, is content distributed equally between nucleus and satellite, or does one typically have more complexity than the other?* Again this is likely to vary from one relation to another.

5. I*s the complexity of a node influenced more by its role (i.e., as nucleus or satellite), or by the relation it expresses?* For instance, in Figure 1, where the PURPOSE node serves as the satellite of the RESULT relationship, which of these features most strongly influences its complexity?

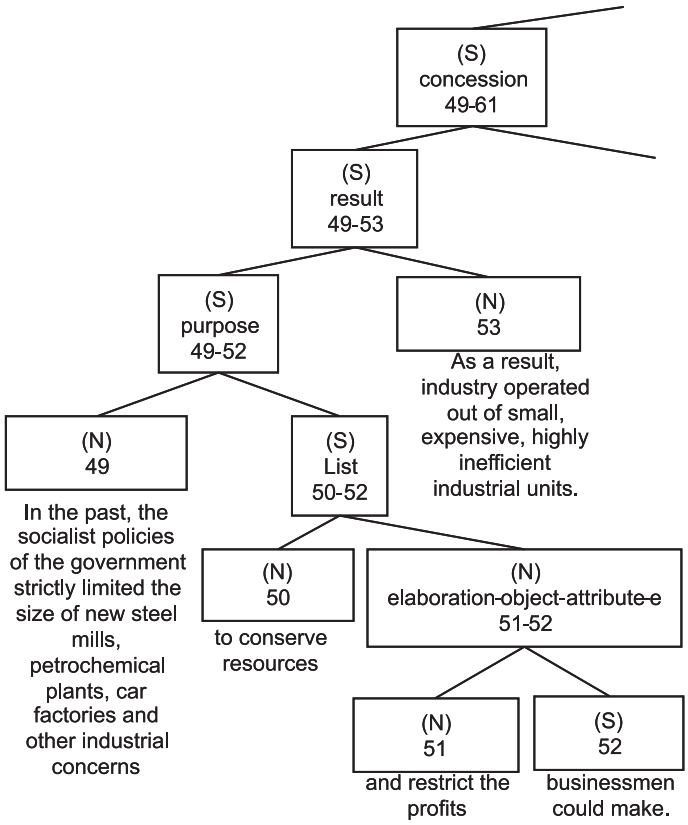

Figure 1: Fragment of an RST tree from the RST-DT Corpus

3. Method

For each non-terminal node in the RST-DT corpus, we measured the number of EDUs that it subsumed. As explained, this serves as a measure of the rhetorical complexity of the relation, as well as an indication of the level at which it occurs in an RST tree. We used the Training part of the corpus, which contains 332 texts in trees of almost 35,000 nodes, noting the complexity of each span in EDUs; then for each relation we computed frequency of occurrence for each complexity value, the mean and median complexities, and the standard deviation. A zero or small standard deviation from the mean indicates that a relation occurs almost without exception at a particular level in an RST tree, whereas a large standard deviation means that a relation occurs at a wide range of levels. Finally, we noted the role (nucleus, satellite, or root) that a node takes in an RST tree, and computed role frequencies for each relation.

4. Results

Table 1 gives aggregate data (mean, median, standard deviation) for all relations in the RST-DT corpus for which we found at least 100 instances (i.e., N$\geq$100). It addresses issues 1-3 on the above list. All relations in the RST-DT Corpus are defined and explained in the RST-DT Corpus Tagging Manual (Carlson and Marcu, 2001). For example, names ending in "-e" denote embedded relations; those ending in "-s" mean that the relation is focussed semantically in the satellite rather than the nucleus (so in EVALUATION-S, the evaluation is presented in the satellite); whilst those ending in "-n" are focussed in the nucleus (therefore CONSEQUENCE-N presents a consequence in the nucleus).

Figure 1 shows the distribution of complexity between nucleus and satellite for each Nucleus-Satellite relation, ordered by increasing complexity of satellite. Since the bars represent proportions, they add up to 1.0 in every case. These results address issue 4.

Figure 2 shows the distribution of roles for each relation[1] – that is, whether the node in question was a satellite or nucleus of the parent node, or the root of the tree. Presentational relations are shown by the label P, and multi-nuclear relations begin with a capital letter. The relations are ordered by increasing frequency of the satellite role. These results are used for assessing issue 5.

In brief, our analysis for each issue is as follows.

1. *Overall differences in complexity:* Table 1 shows clear differences in mean complexity for different rhetorical relations. An Analysis of Variance (ANOVA) test performed on all 14,042 cases for the 27 relations in Table 1 indicates that these differences in complexity are significant ($p < 0.0001$). The median values are useful for indicating the typical complexity for each relation, as the LIST relation in Figure 1 demonstrates, since although the mean complexity of LIST is 8.2 EDUs, in

[1]Embedded relations were omitted from this analysis, since they are always represented in the corpus as Same-Unit nuclei.

Relation	N	Type	Mean Length in EDUs	Median	Standard Deviation from the Mean
attribution-e	102	SM	2.1	2	0.4
elaboration-object-attribute-e	2,218	SM	2.4	2	1.3
elaboration-additional-e	690	SM	2.5	2	1.1
purpose	422	SM	2.6	2	2.1
condition	166	SM	3.3	3	3.6
attribution	2,367	SM	3.3	3	3.7
reason	150	SM	4.3	3	4.9
result	109	SM	5.2	3	5.4
consequence-n	110	SM	5.3	3	12.2
Comparison	100	MN	5.3	3	9.6
comparison	122	P	5.3	2	7.8
circumstance	511	SM	6.2	3	10.4
concession	212	P	6.8	4	9.0
consequence-s	202	SM	7.2	4	10.2
Sequence	130	MN	7.9	4	12.6
List	1,153	MN	8.2	3	14.9
antithesis	323	P	8.6	5	13.5
elaboration-general-specific	326	SM	8.6	5	13.4
Contrast	337	MN	11.6	5	19.2
evidence	146	P	12.2	8	12.6
example	223	P	12.4	9	11.8
explanation-argumentative	484	P	12.5	7	16.8
comment	135	P	14.4	8	19.4
elaboration-additional	2,820	SM	15.3	8	20.3
interpretation-s	158	SM	15.5	9	16.0
background	186	P	17.6	9	21.4
evaluation-s	140	SM	19.3	10	24.7

Table 1: RST relations in order of increasing size in EDUs (SM= subject-matter, P= presentational, MN= multi-nuclear)

Figure 1 it has a complexity of only three, the median value.

2. *Complexity and category of relation:* Subject-matter relations in Table 1 tend to have lower complexity values, suggesting that they are concentrated in the lower and middle levels of RST trees, whereas presentational relations have higher complexities, suggesting a greater concentration towards the upper levels. This hypothesis was confirmed by an independent samples t-test performed on the 12,322 cases of subject-matter and presentational relations (Table 2, $p < 0.0001$).

3. *Whether complexities are focussed or versatile:* The relations in Table 1 from ATTRIBUTION-E to ATTRIBUTION have low means and standard deviations, indicating that they are sharply foccussed at lower levels of the RST trees; those from CONTRAST to EVALUATION-S tend towards the higher levels, but their large standard deviations suggest that they are also distributed through the middle levels.

4. *Relative complexities of nucleus and satellite:* Figure 2 suggests that complexities tend to be more unbalanced in presentational relationships than in subject-matter relationships (e.g., comments are shorter than the material being commented on, whereas examples tend to be much longer than the items being exemplified). To measure balance we first computed $r_{SN} = c_S/c_N$ for each nucleus-satellite relationship, where c_S is the complexity of the satellite and c_N the complexity of the nucleus, then defined the balance index b as equal to r_{SN} if this was less than or equal to 1, or its reciprocal $r_{NS} = c_N/c_S$ otherwise. The index b thus varies from 0 to 1, with a value of 1 if the complexities of nucleus and satellite are equal, and a value tending to zero as the complexities diverge. An independent samples t-test comparing b values for the two types of nucleus-satellite relationship showed presentational relationships significantly more unbalanced than subject-matter ones, with means of 0.53 and 0.69 respectively (Table 2, $p < 0.0001$).

   Figure 2 also suggests that satellites in presentational relationships tend to be more complex and constitute a greater proportion of total complexity than those of subject-matter relationships. This is confirmed by the t-tests on raw satellite complexities and satellite proportion (i.e., $c_S/(c_S + c_N)$) in Table 2 ($p < 0.0001$).

5. *Complexity and role:* Figure 2 shows that subject-matter relations tend to occupy the nucleus role in RST trees, whereas presentational and multi-nuclear relations tend to be more equally distributed between the satellite and nucleus roles, and that the root role is less

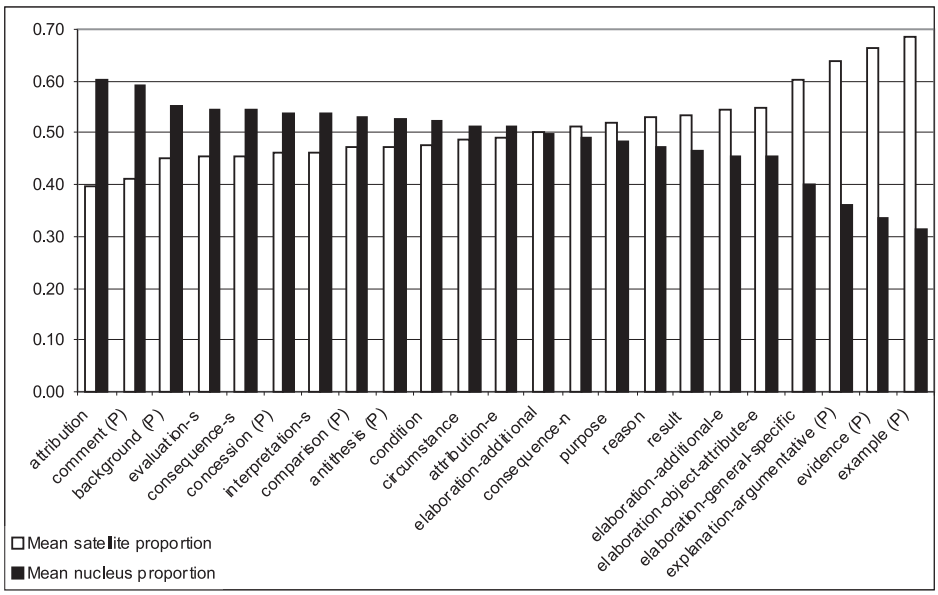

Figure 2: Mean proportions of satellite and nucleus for each Nucleus-Satellite relation (P=presentational)

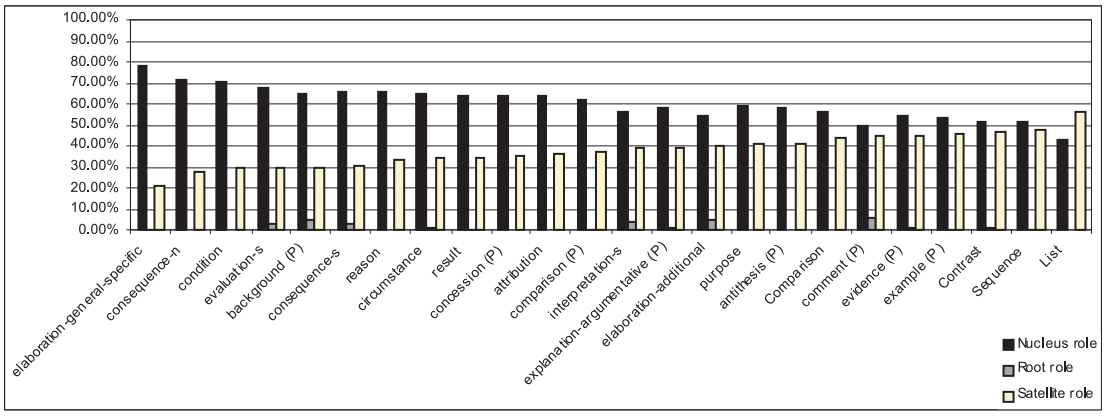

Figure 3: Mean percentages of roles for each relation (P=presentational)

common across all three groups. This was confirmed by a chi-squared test (table 3, $\chi^2 = 253.4, df = 4, p < 0.0001$).

5. Discussion

It is worth bearing in mind that our results are specific to the newspaper genre of the RST-DT corpus. However, they exhibit some clear trends. To give just one example, Figure 2 shows an interesting partition of the presentational relations into ones with a relatively complex nucleus, and ones with a complex satellite; such differences may provide clues allowing a better classification of the relations.

To take this a little further, let us consider a group of eight nucleus-satellite relations with high average complexities. In Table 1, the relations EVIDENCE, EXAMPLE, EXPLANATION-ARGUMENT and ELABORATION-GENERAL-SPECIFIC all have high complexities (respectively 12.2, 12.4, 12.5, 8.6); in Figure 2, all four come at the satellite-heavy end (satellite proportions respectively 0.67, 0.68, 0.65, 0.60). By contrast, the relations COMMENT, INTERPRETATION-S, BACKGROUND and EVALUATION-S, while also having high complexities (14.4, 15.5, 17.6, 19.3), come instead at the nucleus-heavy end (satellite proportions 0.41, 0.46, 0.45, 0.45). With more analysis it might be possible to interpret this dichotomy: it could be, for instance, that heavy satellites tend to aim at increasing the reader's level of understanding or confidence in the nucleus, while light satellites tend to add some kind of interpretive comment to material that is already understood and accepted. We cannot develop this point here, but such examples do suggest that complexities are worth analysing as a source of theoretical insights.

On a more applied level, the complexity data reported here could be used directly in discourse planning algorithms for NLG. To give just one simple example, the semantic graph shown on the LHS of Figure 4 could be realised by two

	Subcategory	N	Mean	Std. Devn.	Std. Error Mean	Significance (2-tailed)
Rhetorical	SM	10491	7.09	13.18	.129	$p < 0.0001$
complexity	P	1831	11.30	15.32	.358	
Satellite	SM	10491	3.56	7.90	.077	$p < 0.0001$
complexity	P	1831	6.49	10.71	.250	
Satellite	SM	10491	.49	.17	.002	$p < 0.0001$
proportion	P	1831	.55	.22	.005	
Balance	SM	10491	.69	.32	.003	$p < 0.0001$
measure	P	1831	.53	.31	.007	

Table 2: Independent Samples t-tests (SM= subject-matter, P= presentational)

Subcategory	Nucleus Role	Root Role	Satellite Role	Total
MN	797	7	916	1720
P	1070	28	733	1831
SM	6772	169	3550	10491
Total	8639	204	5199	14042

Table 3: Cross Tabulation for role of relations (SM= subject-matter, P= presentational, MN= multi-nuclear)

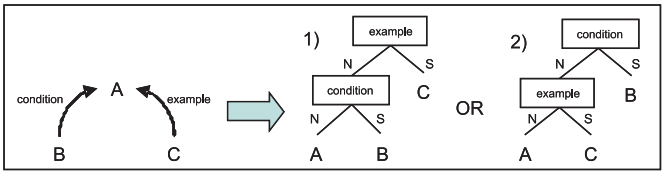

Figure 4: Planning RST trees from a semantic graph

different RST trees: (1) with the EXAMPLE relation at the root, and (2) with CONDITION at the root. If a rhetorical planning algorithm were weighted according to our rhetorical complexity data, it would clearly favour tree (1) over tree (2), since CONDITION relations tend to be less complex than EXAMPLE relations.

6. References

L. Carlson and D. Marcu. 2001. Discourse tagging manual. Technical Report ISI Tech Report ISI-TR-545, Information Sciences Institute (ISI) Reprint Series, University of Southern California, U.S.A.

L. Carlson, D. Marcu, and M. E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt and R. Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375.

M. O'Donnell, C. Mellish, J. Oberlander, and A. Knott. 2001. Ilex: An architecture for a dynamic hypertext generation system. *Journal of Natural Language Engineering*, 7:225–250.

R. Power, D. Scott, and N. Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29:211–260.

C. Sporleder and A. Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 43–49.

S. Williams and E. Reiter. 2003. A corpus analysis of discourse relations for natural language generation. In *Proceedings of Corpus Linguistics*, pages 899–908.