

Deriving rhetorical complexity data from the RST-DT Corpus

Sandra Williams and Richard Power

The Open University, Walton Hall, Milton Keynes, U.K.

s.h.williams@open.ac.uk r.power@open.ac.uk

RHETORICAL COMPLEXITY

Rhetorical complexity can be defined as the number of Elementary Discourse Units (EDUs) that a relation subsumes.

In Fig. 1,

RESULT subsumes 5 EDUs,
PURPOSE subsumes 4 EDUs,
LIST subsumes 3 EDUs, and
ELABORATION-OBJECT-ATTRIBUTE-E subsumes 2 EDUs.

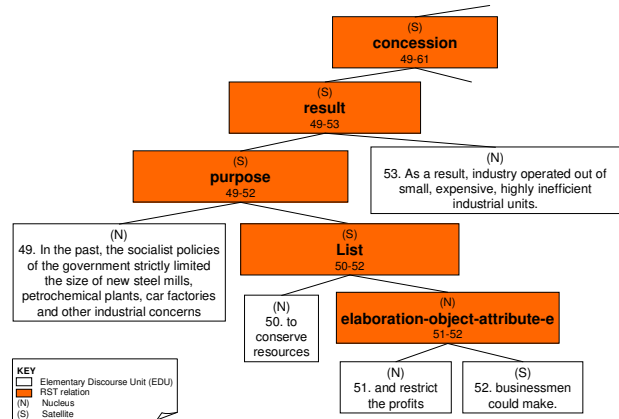


Figure 1 Extract from the Rhetorical Structure Theory Discourse Treebank (RST-DT) Corpus, Carlson et al., (2003)

RESULTS

1. Rhetorical complexity varies for different relations

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 Fig. 1 demonstrates: although the mean complexity of LIST is 8.2 EDUs, its median complexity is only 3 EDUs.

2. Subject-matter relations have lower complexity than presentational relations

Subject-matter relations (concerning the semantic domain) tend to have lower complexity values, suggesting that they are concentrated in the lower and middle levels of RST trees, whereas presentational relations (concerning the author's intentions) have higher complexities, suggesting a greater concentration towards the upper levels.

3. Content is more unbalanced between nucleus and satellite in presentational relations than in subject-matter relations.

Balance (b) = c_N/c_S if c_S is greater than c_N , or c_S/c_N if c_N is greater than c_S , where c_S is the rhetorical complexity of the satellite and c_N is that of the nucleus. Thus, b varies between 0 and 1, with a value of 1 if the satellite and nucleus are of equal complexity, and values tending to 0 as the complexities of satellite and nucleus 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 ($p < 0.0001$). 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 ($p < 0.0001$).

4. Low-complexity relations occur more often in the nucleus role, high-complexity relations in the satellite role

See Figure 3, where the nucleus role is more frequent for relations with complexities of 2-3 EDUs, while the satellite role is more frequent for relations with complexities over 4 EDUs. An independent samples t-test shows that the difference is significant ($p < 0.0001$), with mean complexities of 7.20(nucleus role) and 8.21 (satellite role).

Relation	N	Type	Mean Rhetorical Complexity in EDUs	Median	Standard Deviation from the Mean
evaluation-s	140	SM	19.3	10	24.7
background	186	P	17.6	9	21.4
interpretation-s	158	SM	15.5	9	16
elaboration-additional	2,820	SM	15.3	8	20.3
comment	135	P	14.4	8	19.4
explanation-argumentative	484	P	12.5	7	16.8
example	223	P	12.4	9	11.8
evidence	146	P	12.2	8	12.6
Contrast	337	MN	11.6	5	19.2
antithesis	323	P	8.6	5	13.5
elaboration-general-specific	326	SM	8.6	5	13.4
List	1,153	MN	8.2	3	14.9
Sequence	130	MN	7.9	4	12.6
consequence-s	202	SM	7.2	4	10.2
concession	212	P	6.8	4	9
circumstance	511	SM	6.2	3	10.4
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
result	109	SM	5.2	3	5.4
reason	150	SM	4.3	3	4.9
condition	166	SM	3.3	3	3.6
attribution	2,367	SM	3.3	3	3.7
purpose	422	SM	2.6	2	2.1
elaboration-additional-e	690	SM	2.5	2	1.1
elaboration-object-attribute-e	2,218	SM	2.4	2	1.3
attribution-e	102	SM	2.1	2	0.4

Table 1: RST relations in order of increasing Rhetorical Complexity. SM= subject-matter, P= presentational, MN= multi-nuclear

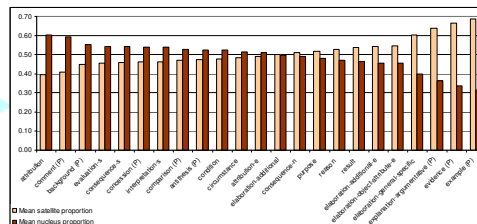


Figure 2. Mean proportions of satellite and nucleus for each Nucleus-Satellite relation (P=presentational)

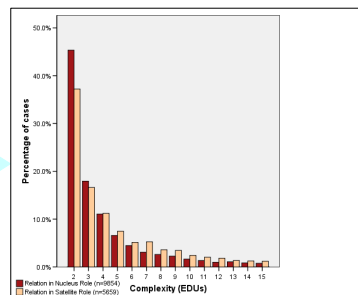


Figure 3. Frequencies of relations in nucleus and satellite role for different complexities

APPLICATION

When the input to an NLG system allows a range of different rhetorical structure trees, complexity data can provide an empirical basis for preferring one option to another.

In the example in Figure 4, we can prefer option 1 to option 2 on the grounds that EXAMPLE has a higher complexity than CONDITION in the RST-DT corpus, and is thus more likely to occur higher up the tree.

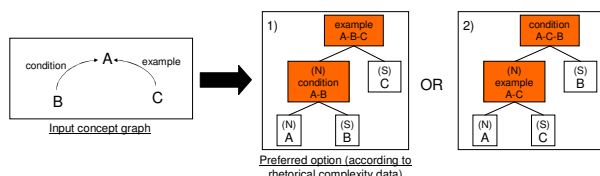


Figure 4. Choosing between different RST trees expressing the same rhetorical input