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.

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 14,040 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 LST relation in Fig. 1 demonstrates: although the mean complexity of LST is 8.2 EDUs, its median complexity is only 2 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 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 = c2/n2) of c2 is greater than c1 or c2/n1, if c1 is greater than c2, where c1 is the rhetorical complexity of the satellite and c2 is that of the nucleus. Thus, b varies between 0 and 1, with a value of 0.5 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 the new satellite complexities and satellite proportion i.e., (c1/c2 + c1) 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).

APPLICATION

When the input to an NLP 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-E has a higher complexity than CONDITION in the RST-DT corpus, and is thus more likely to occur higher up the tree.