**Appendix S1. Performance of the Explicit Discourse Connectives Tagger on different types of connectives**

Table S1.1 summarizes the overall performance of the EDCT as well its performance on each score band. Table S1.2 summarizes the precision, recall, and F-score of the EDCT’s tagging of different types of connectives in the 30 texts randomly sampled from our dataset. Overall, the EDCT performed well on all types of connectives, with the exception that the recall for connectives indicating Comparison was only 67.2%. Specifically, it missed 19 of the 58 connectives of this type identified by the two annotators. A close look at the missed instances showed that the EDCT tagged a number of cases of *but* and *unless* indicating Comparison as non-discourse connectives. This fine-grained evaluation provides insight into areas for improvement of the tagger in future research.

**Table S1.1.** Performance of the EDCT on 30 texts randomly sampled from the dataset

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Learner Use | | | Tool Label | | |
| Score | Nessays | Nsents | Nconnectives | Nproper | Nimproper | Accuracy | Ncorrect | Nincorrect | Accuracy |
| Low | 10 | 118 | 255 | 178 | 77 | 69.8% | 163 | 15 | 91.6% |
| Mid | 10 | 137 | 419 | 402 | 17 | 95.9% | 350 | 52 | 87.1% |
| High | 10 | 234 | 331 | 319 | 12 | 96.4% | 300 | 19 | 94.0% |
| All | 30 | 489 | 1005 | 899 | 106 | 89.5% | 813 | 86 | 90.4% |

Nessays = number of essays, Nsents = number sentences, Nconnectives = number of connective instances, Nproper = number of connectives determined to be clear/proper, Nimproper = number of connectives determined to be unclear/improper, Ncorrect = number of clear/proper connectives correctly tagged by the EDCT, Nincorrect = number of clear/proper connectives incorrectly tagged by the EDCT.

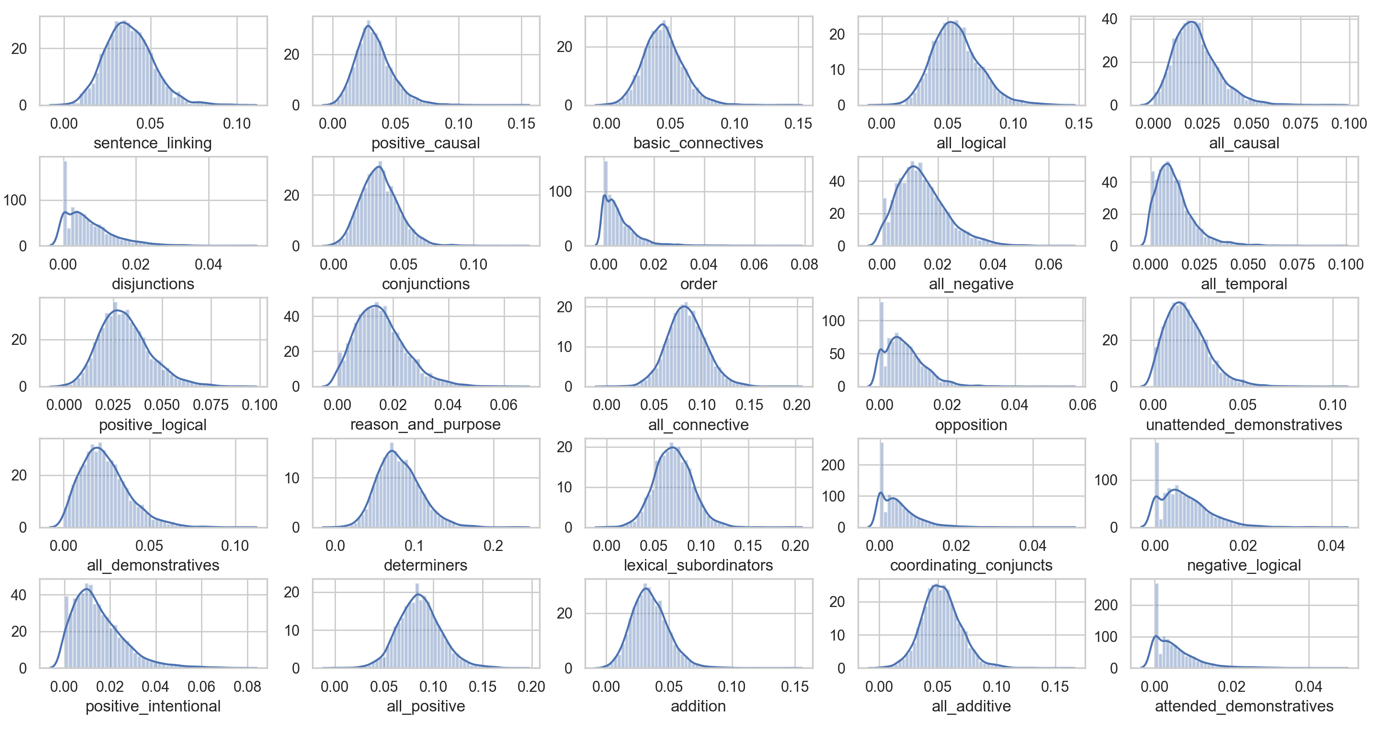
**Table S1.2.** Performance of the EDCT on different types of connectives

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Ntrue | Npred | Ncorrect | Precision | Recall | F-score |
| Contingency | 163 | 144 | 134 | 93.1% | 82.2% | 87.3% |
| Expansion | 114 | 107 | 100 | 93.5% | 87.7% | 90.5% |
| Temporal | 61 | 63 | 55 | 87.3% | 90.2% | 88.7% |
| Comparison | 58 | 43 | 39 | 90.7% | 67.2% | 77.2% |
| Non-DC | 503 | 542 | 485 | 89.5% | 96.4% | 92.8% |
| All | 899 | 899 | 813 | 90.4% | 90.4% | 90.4% |

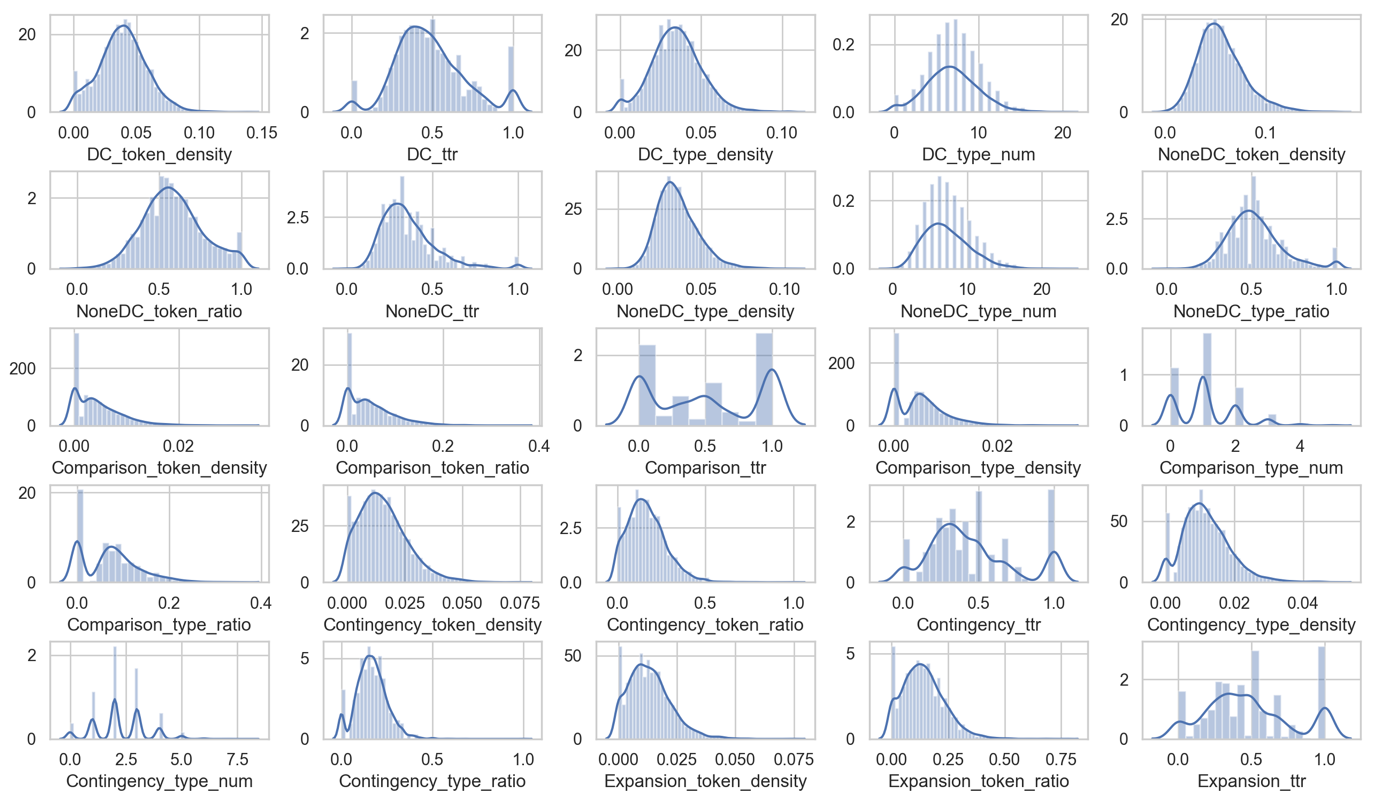
Ntrue = Number of connectives judged to be of this type by the two annotators, Npred = number of connectives tagged by the EDCT as of this type, Ncorrect = number of connectives correctly tagged by the EDCT as of this type.

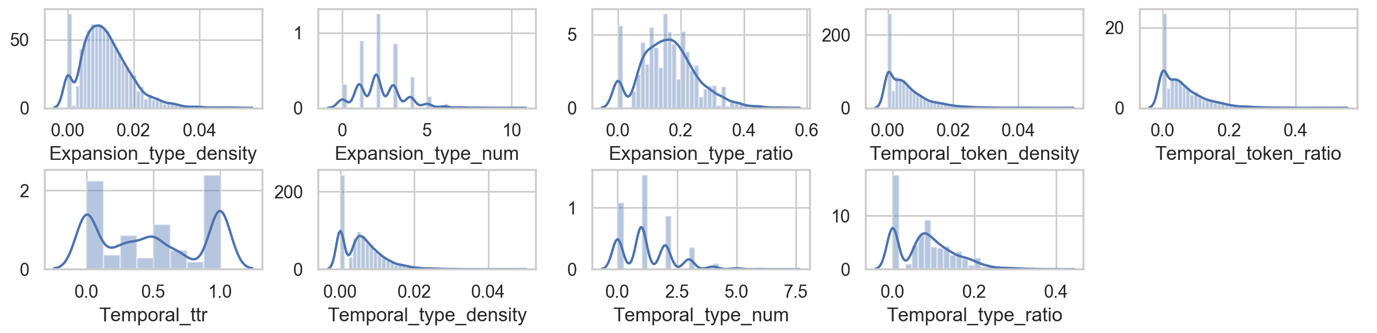
**Appendix S2. Distribution of the indices**

Figures S2.1 and S2.2 provide the distribution plots for each cohesion index generated using the *distplot* function from the Python seaborn library. Many indices exhibited a largely normal distribution. We note that our large sample size (N = 3,911) also helps mitigate the violation of the normality assumption according to the central limit theorem. While some indices (e.g., the type\_num indices) had small values, there were retained in the analysis as they still showed good distinction among different essays.



**Figure S2.1.** Distribution plots of TAACO indices





**Figure S2.2.** Distribution plots of the sense-aware cohesion indices