A Cross-Genre Analysis of Discourse Relation Signaling in the GUM Corpus

Lauren Levine
lel76@georgetown.edu
GEORGETOWN UNIVERSITY

Research Questions

- Are discourse relations signaled differently in different genres?
- Which discourse relations display the most inter-genre variation in how they are signaled?

Study Overview

- **In this study we:**
  - Investigate the **cross-genre variation** in how **discourse relations** are **signaled** in the **RST** annotations of the **GUM** corpus
  - Provide a methodology for ranking the inter-genre variability of the signaling of individual discourse relations
  - Analyze which discourse relations display the most inter-genre variation in how they are signaled
- **We find:**
  - Discourse relations are signaled in a **relatively stable** manner across genres in GUM
  - We produce **stable rankings of inter-genre variability** for the signaling of discourse relations, finding that ***organization***, ***restatement***, and ***explanation*** relations display the most inter-genre variation.
  - However, we find that genre specific graphical norms can account for a large portion of the observed variation.

Background

- **Discourse Relations**: the meaning that arises from the combination of multiple linguistic units in a discourse
- **Discourse Signals**: linguistic units that mark (explicitly or implicitly) the occurrence of a discourse relation
- Rhetorical Structure Theory (**RST;** Mann and Thompson, 1988): Pragmatic formalism for creating discourse relation tree structures of a text (example in Figure 1)
- The RST Signalling Corpus (RST-SC; Das and Taboada, 2018): established a **taxonomy of signal types for RST data:**
- The **eRST** project (Zeldes, 2024): added the Das and Taboada signal taxonomy to the existing GUM corpus
- In this study, we leverage the **signaling annotations** added to the **GUM RST treebank** from the eRST project.

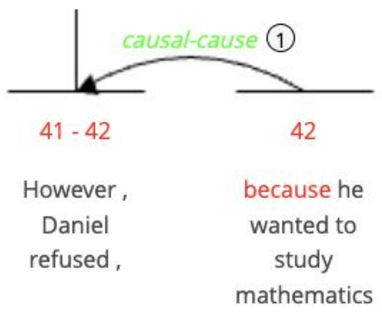

Figure 1: Example from the GUM corpus of a **causal** discourse relation, overtly signaled by the explicit discourse marker (dm) *because*.

Data

- Georgetown Multilayer Corpus GUM V10 (Zeldes, 2017):
  - 228k token corpus of English
  - 235 documents across 16 genres:
    - academic, biographies, courtroom, conversation, essay, fiction, interview, letters, news, podcasts, speeches, textbooks, travel, vlogs, how-to and Reddit forum discussions
- 30,774 discourse relation annotations in GUM RST
  - 69.35% (**21,343 instances**) occur with one or more signaling annotation
- Two-tiered relation inventory; 15 coarse relation and 32 fine-grained relations
  - **Coarse relations**: *adversative, attribution, causal, context, contingency, elaboration, explanation, evaluation, joint, mode, organization, purpose, restatement, topic, same-unit*
- Annotated with Das and Taboada signal types:
  - discourse markers (dm)
  - graphical (grf)
  - lexical (lex)
  - morphological (mrf)
  - numerical (num)
  - reference (ref)
  - semantic (sem)
  - syntactic (syn)
- We analyze the distribution and cross-genre variation of each signal type co-occurring with each relation type

Methods

- We use the Jensen-Shannon Distance (JSD) to quantify how different the distributions of relation signals are between a pair of genres for a particular relation.
- We calculate the JSD scores between all possible pairs of genres, and average these scores to create a an inter-genre variability score for the individual relation.
- The **inter-genre variability score** for a discourse relation *R* using **Avg. Pairwise JSD** is defined as:

$$Avg.PairwiseJSD(R) = \frac{\sum_{i,j \in G} JSD(SD_i(R), SD_j(R))}{\binom{|G|}{2}}$$

where ***G*** is the set of genres, **JSD** is the JS Distance, and ***SDx(R)*** is the frequency distribution of relation signal types for relation ***R*** in genre ***x***.

- Sorting relations by **Avg. Pairwise JSD** creates a **relative ranking** of inter-genre variability of signaling for individual relations.
- To **establish reliability**, we sample documents across genres, calculating 50 independent rankings. We report avg. correlation metrics of the rankings in Table 1.
- The strength of these metrics shows that **Avg. Pairwise JSD** is a reliable method for constructing a relative ranking of inter-genre variation.

Rank Correlation Metric	Relation Type	
	Coarse	Fine-grained
Avg. Kendall's Tau	0.82	0.76
Avg. Spearman Rank	0.93	0.90
Avg. Pearson Correlation	0.95	0.92

Table 1: Avg. of correlation metrics comparing rankings of inter-genre variation for the signaling of RST relations, computed from randomly sampled subsets of the GUM corpus.

Results

- There is a considerable amount of inter-relation variation in the proportions of signal types used (Figure 2)
  - *evaluation* relations are signaled exclusively by lexical features
  - *adversative*, *causal*, and *contingency* relations are dominated by overt discourse markers
- Figure 3 shows the proportions of signals present in each genre, adjusted for the relative frequencies of the relations present in that genre.
  - Signal proportions are surprisingly consistent across the various genres of the GUM corpus

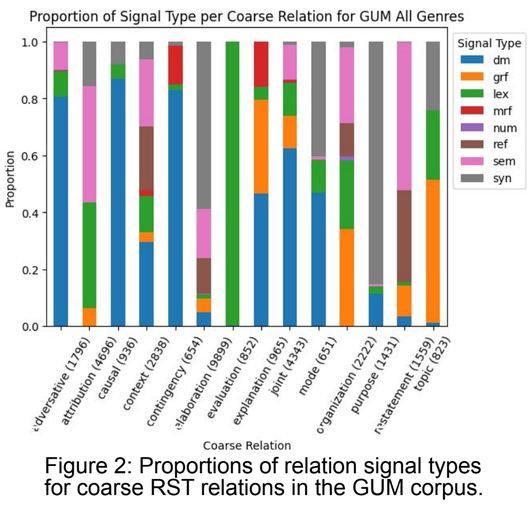

Figure 2: Proportions of relation signal types for coarse RST relations in the GUM corpus.

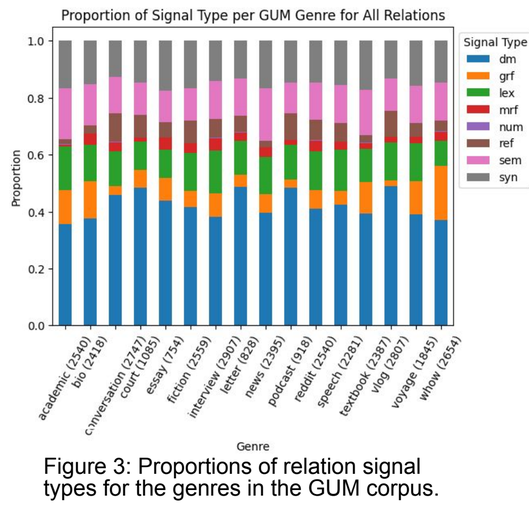

Figure 3: Proportions of relation signal types for the genres in the GUM corpus.

- The inter-genre variation ranking for the coarse RST relations is shown in Figure 4:
  - **Most inter-genre variation:**
    - *organization*, *restatement*, and *explanation*
  - **Least inter-genre variation:**
    - *attribution*, *adversative*, and *evaluation*
- The **signal type distributions across genres** shown for the relations with the **most inter-genre variation**: ***organization***, ***restatement***, and ***explanation***
  - From the visualization, **Avg. Pairwise JSD accurately reflects the relative inter-genre variation of the relations**

Figure 4: (Top) **Ranking of inter-genre variation** of relation signal distributions for coarse RST relations (based on Avg. Pairwise JSD). Proportions of relation signals across genres for: (bottom left) the coarse relation showing the **most variation: organization**, (bottom middle) the coarse relation showing the **median variation: causal**, and (bottom right) the coarse relation showing the **least variation: evaluation**.

Conclusions

- **Inter-genre signaling** of individual discourse relations is **relatively stable**
- In two of the coarse relations which showed the **most inter-genre variation** in their signaling, **organization** and **explanation**, **genre specific graphical norms** seemed to contribute more to the existing variation than the language content.
  - If there is a large variation in the signal types used in two genres that goes beyond graphical norms, it may be because those genres call for different relations to be used, rather than because the genre is signaling the same relations differently.
- As RST is a **pragmatic formalism**, without restrictions on the structural components required to apply a specific discourse relation, it is **surprising** that we see such **limited variation** in the signaling of individual relations across genres.
  - Our results suggests that, **despite being pragmatically defined**, the **discourse relations** in the **RST** relation inventory display some degree of **structural consistency in their manner of signaling**.

References

- Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. RST discourse treebank. Linguistic Data Consortium, University of Pennsylvania.
- Debopam Das. 2019. Nuclearity in RST and signals of coherence relations. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pages 30–37, Minneapolis, MN. Association for Computational Linguistics.
- Debopam Das and Maite Taboada. 2018a. RST signalling corpus: A corpus of signals of coherence relations. Language Resources and Evaluation, 52:149–184.
- Debopam Das and Maite Taboada. 2018b. Signalling of coherence relations in discourse, beyond discourse markers. Discourse Processes, 55(8):743–770.
- Markus Egg and Debopam Das. 2022. Signalling conditional relations. Linguistics Vanguard, 8(s4):383–392.
- Luke Gessler, Yang Janet Liu, and Amir Zeldes. 2019. A discourse signal annotation system for RST trees. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pages 56–61.
- Yang Liu and Amir Zeldes. 2019. Discourse relations and signaling information: Anchoring discourse signals in RST-DT. Society for Computation in Linguistics, 2(1).
- William C Mann and Sandra A Thompson. 1988.Rhetorical structure theory: Toward a functional theory of text organization. Text-interdisciplinary Journal for the Study of Discourse, 8(3):243–281.
- Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. Dialogue & Discourse, 4(2):249–281.
- Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. Language Resources and Evaluation, 51(3):581–612.
- Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2024. eRST: A signaled graph theory of discourse relations and organization. arXiv preprint arXiv:2403.13560.