muRST: an Analysis of Cross-lingual Variation in Discourse Structure

Junghyun Min* **Emma Thronson*** **Lillian Earhart*** **Kohei Kajikawa** **Lanni Bu**
Georgetown University
{jm3743, et726, le290}@georgetown.edu
https://github.com/aatlantise/multilingual-rst
*Equal contribution

Rhetorical Structure Theory (RST) is a framework of computational discourse analysis, analyzing elementary discourse units (EDUs), discourse markers (DMs), and relations (Mann and Thompson, 1988). While multilingual adaptations exist (Cao et al., 2018; Peng et al., 2022), the field lacks a systematic analysis of a parallel multilingual corpus. We propose to fill the gap with **Multilingual RST (muRST)**: a case study of cross-lingual variation in discourse units, markers, and relations by manually annotating and analyzing two parallel documents in English (en), Spanish (es), Korean (ko), Chinese (zh), and Japanese (jp)[1]. Specifically, we examine parallel texts to understand how typological differences affect discourse structures. We hypothesize that variations in word order (SOV in ko; SVO in en, es), the realization of relative clauses, and pronoun-dropping behaviors will lead to observable differences in RST annotations. For example, we expect more granular EDUs in ko, while relations in en and es may be more strongly driven by discourse markers.

Each author is a native or near-native L2 speaker in a target language, and annotates the document in that language. Our analysis utilizes two open-source parallel news articles sourced from Global Voices[2]. Both articles are originally in English and translated into respective target languages.

A quantitative analysis reveals that the ko annotation is more granular in terms of EDUs and exhibits a higher rate of `same-unit` relations. en and es annotations are similar in the number of discourse units and the most frequent relations, with elaboration and joint-list relations being the most common. We conduct two qualitative case studies to explore specific variation in how relative clauses (RC) and coordinating conjunctions (CC) are annotated.

[1]In this summary, we describe our analysis in en, es, ko.
[2]`mexican-cake`, `iranian-women`

Relative clauses. In en, all satellite relations that include relative pronouns are annotated as elaboration-attribution. In ko, which lacks relative pronouns and often realizes them as adjectival phrases (APs), translation choices frequently convert an RC into a nominal modified by an AP. As a result, ko shows very different tree structures for heavy-RC sentences compared to en and es, yielding frequent `same-unit` relations. However, considering that `same-unit` is not a discourse relation, we conclude that the number of EDUs and the distribution of relation types are fairly stable cross-lingually, after controlling for syntactic differences. Furthermore, translation choices are the primary source of differences between en and es annotations, affecting relation attachment heads.

Coordinating conjunctions. Heavy-CC sentences appear more structurally similar across the three languages and display similar relations, barring translation choices and errors. However, the position of the conjunctions differs. ko CCs that split EDUs are found at the end of segments as verb suffixes, whereas in en and es, they appear at the beginning. For these structures, en and es share the same number of EDUs and are split at equivalent places in the discourse.

Conclusion. Our analysis yields several key takeaways. Overall, the relations and EDUs are fairly stable across languages despite typological syntactic differences, with a similar number of EDUs and the same suite of most frequent relations. The pro-drop characteristics found in es and ko appear to have no effect on discourse relations. Second, heavy-RC sentences result in differing tree structures across languages, while heavy-CC sentences remain largely similar. This may be attributable to how RCs introduce hierarchy by introducing a EDU that modifies one argument in its parent EDU. However, questions remain; subsequent work could focus on the role of DMs and the stability of specific relations across languages.

References

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.