Rhetorical Parallelism Detection: Revised Annotation Guidelines

Stephen Bothwell

July 26th, 2023*

1 Introduction

In this document, we provide our annotation guidelines for parallelism data. We do so in order to allow for data collection to be done effectively and with minimal ambiguity and inconsistency.

2 Guidelines

2.1 Terminology

Before we begin, we describe some terminology.

- Given a document, a **span** is a sequence of words in that document.
- A **branch** is any span which is involved in a parallelism. In other words, every branch is a span, but not every span is necessarily a branch.
- A **parallelism** is a group of spans, called branches, which relate to each other in two respects: proximity and linguistic similarity. Note that all spans contained in a parallelism must be distinct; a span cannot be parallel with itself.
  - *Proximity* has to do with the idea that the two spans of text occur close together—they are close enough that, in conversation, one might reasonably remember an earlier span which relates to a later span.
  - *Linguistic similarity* concerns the way in which two texts relate in terms of their phonology, morphology, syntax, and semantics. Two spans are linguistically similar if they share a sufficient amount of such features.
- A parallelism is *synchystic* if the relations of words in branches are in-order. Meanwhile, a parallelism is *chiastic* if the relations of words in branches are reversed.

2.2 Guidelines

In general, the detection of a branch should be on the basis of exhibiting at least *two* of these criteria:

- **Length**: They contain a nearly identical (*e.g.*, ±2) number of words.
- **Distance**: They are a short distance from one another (*i.e.*, they had few [roughly ±2] intervening words).

*The date presented here is the "last updated" date.

- **Phonology**: They have at least two words that are phonetically similar (*e.g.*, that rhyme).
- **Morphology**: They have two or more pairs of words that are lexically identical.
- **Syntax**: They have identical syntactic structure, or they consist of two or more pairs of words of an identical grammatical form in identical order.
- **Semantics**: They have two or more pairs of words that are semantically related (*e.g.*, are synonyms, cognates, or antonyms).

A branch cannot be considered in isolation; entire parallelisms come together on the basis of multiple branches. Because the smallest unit of any span is a single word, there are many ways in which one could determine how to divide up and connect spans into branches of a parallel structure. Because of this, we provide further principles on how to select branches from multiple possibilities.

- **Maximality**: In general, the largest possible branches should be selected. Parallelisms should be kept separate only if the pertinent spans concern distinct linguistic relationships.
  - *Interlocking*: Branches should **not** interlock—that is, maintain a synchystic (A-B-A-B) structure on their own if each "A" and "B" can be merged to form a larger branch containing each of the smaller branches and pairing those branches in effectively the same manner.
  - *Nesting*: Branches are permitted to *nest* in one another; that is, a span can be labeled more than once so long as one of the branches entirely contains the other. Each branch should also be a part of a distinct parallelism.
- **Intent**: One critical fixture of *rhetorical* as opposed to *syntactic* parallelism is its intentionality. Syntactic parallelism can be required by the grammatical constructions of a given language; meanwhile, rhetorical parallelism is an effort by a speaker or writer to create an effect. We elaborate upon a few items to aid in determining what is intentional.
  - *Conjunctions*: In many cases, it is unclear whether conjunctions are simply fulfilling a syntactic role or are involved in a wider juxtaposition. We rule that conjunctions should be included in parallelisms for a series of branches only if polysyndeton is maximal; in other words, *every branch* must contain a conjunct. Otherwise, conjuncts should be left out.
  - *Laundry Lists*: As is implied by the description above, syntactic parallelism is required but is not necessarily intentional, whereas rhetorical parallelism is not necessarily required but is intended. By this logic, a series of items piled up in a "laundry list" is not innately a rhetorical parallelism unless some other feature-based connection ties the items together.

2.3 Software Notes

We use the BRAT annotation tool [1] for our annotations. Our annotation scheme uses BRAT's entities and relationships. In BRAT, an entity is used to label a span. We have three types of entities: `ParallelArm`, `ChiasmA`, and `ChiasmB`. Relationships are used to connect spans to form parallelisms. We have two types of relationships: `Parallel` and `Chiasm`. These are used as follows:

- To label a synchystic parallelism, each of the (two or more) branches of the parallelism are labeled as `ParallelArm` entities. Then, they are connected with `Parallel` relationships.
  - Note that, in a parallelism, all spans in a parallelism are each parallel with one another. As such, every branch only needs to be connected with one other branch in a parallelism to be connected with all other branches. For the sake of simplicity, we recommend connecting branches sequentially in the order that they appear.

- To label a chiastic parallelism, the first branch in sequential orer should be labeled as `ChiasmA`. Meanwhile, the second branch should be labeled as `ChiasmB`. Finally, these branches should be connected with the `Chiasm` relationship.

References

[1] Pontus Stenetorp et al. "brat: A Web-Based Tool for NLP-assisted Text Annotation". In: *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France: Association for Computational Linguistics, Apr. 2012.