# Enhancing Discourse Parsing for Local Structures from Social Media with LLM-Generated Data

**Martial Pastor**[1] **Nelleke Oostdijk**[1] **Patricia Martín-Rodilla**[2] **Javier Parapar**[3]
[1]Centre for Language Studies, Radboud University, The Netherlands
[2]IEGPS CSIC, Santiago de Compostela, Spain
[3]IRLab, CITIC Research Centre, Universidade da Coruña, La Coruña, Spain
martial.pastor@ru.nl nelleke.oostdijk@ru.nl
p.m.rodilla@iegps.csic.es javier.parapar@udc.es

## Abstract

We explore the use of discourse parsers for extracting a particular discourse structure in a real-world social media scenario. Specifically, we focus on enhancing parser performance through the integration of synthetic data generated by large language models (LLMs). We conduct experiments using a newly developed dataset of 1,170 local RST discourse structures, including 900 synthetic and 270 gold examples, covering three social media platforms: online news comments sections, a discussion forum (Reddit), and a social media messaging platform (Twitter). Our primary goal is to assess the impact of LLM-generated synthetic training data on parser performance in a raw text setting without pre-identified discourse units. While both top-down and bottom-up RST architectures greatly benefit from synthetic data, challenges remain in classifying evaluative discourse structures.

## 1 Introduction

Recent advancements in discourse parsing have sparked a range of applications in discourse analysis. Indeed, automatically extracting discourse structures from text has been shown to be useful for the analysis of political discourse (Allen et al., 2014; Pastor et al., 2024; Wang et al., 2023) or of good scientific writing (Gonçalves et al., 2020; Kiepura et al., 2024). In particular, RST (Rhetorical Structure Theory; Mann and Thompson, 1988) has served as an innovative tool for studying rhetorical relations. However, the task of RST discourse parsing has not achieved the same level of success as other NLP tasks at the sentence level. Notably, current RST datasets have been developed for a limited set of genres and text types, and it has been shown that existing RST parsers do not generalize well across different genres (Liu and Zeldes, 2023).

With recent interests in investigating rhetoric in social media and micro-blogging using RST (Chenlo et al., 2013; Liu and Liu, 2021), we explore the use of Large Language Models for generating synthetic training data to compensate for the lack of resources developed for these platforms. Specifically, we examine the effect of using synthetic data on parser performance in a discourse parsing 'in the wild' setting[1], with a focus on recognizing a local discourse structure prevalent in polarized discussions on social media. This particular structure (see Figure 1) consists of the arrangement of JOINT JOINT EVALUATION (JJE) coherence relations, which mimics a logical-argumentative flow by presenting a series of implicitly related statements[2], leading to a climactic statement (Pastor et al., 2024).

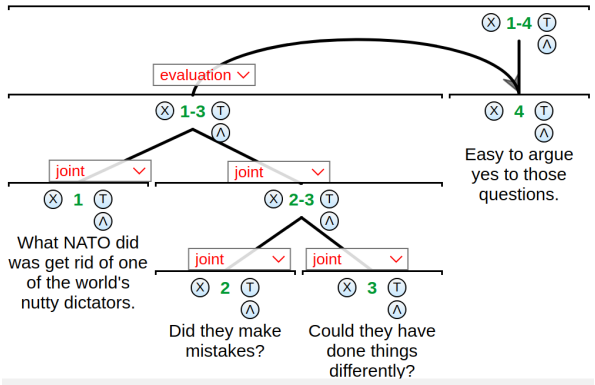


Figure 1: An RST analysis of a comment exhibiting the JOINT JOINT EVALUATION (JJE) discourse structure. The arrow denotes the relation between nucleus and satellite and points to the nucleus.

The structure of the remainder of the paper is as follows. We first provide a brief overview of related work. Then, we describe our process for annotation and the generation of synthetic training data. We explain how we initially extracted our original set of JJEs, which we later annotated using refined guidelines, also detailed here. We high-

[1]Raw text is parsed, no gold EDU breaks are given.
[2]Here, implicitness pertains to the absence of explicit connectives or syntactic elements that link independent clauses.



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 8739–8748
January 19–24, 2025. 

light key considerations in annotating RST structures for social media, particularly for the specific RST constellation we focus on. This resulted in a dataset of 1,170 JJE instances (900 synthetic and 270 gold) across three social media platforms: online news comment sections, Reddit, and Twitter. Next we go on to discuss the results obtained by two different parsers—a bottom-up parser (Guz and Carenini, 2020) and a top-down parser (Liu et al., 2021)—that have been trained on this new synthetic data, the quality of which is further assessed. While we demonstrate that both fundamental RST architectures benefit from synthetic data across the three different social media platforms, we find that RST structures involving evaluative elements remain difficult to classify.

## 2 Related Work

### 2.1 Computational Approaches to the Analysis of Local Discourse Structures

Local discourse structures allow for the formal definition of textual-level entities or relations within targeted discourse segments. These relations or entities encompass coherence relations, discourse categories, argumentative relations, and more.

Automatic identification of local discourse structures has facilitated numerous applications. For example, it helps in discerning stances aligned with persuasive argumentation (Chakrabarty et al., 2019; Stab and Gurevych, 2017) or in generating questions about information in texts using precise arrangements of rhetorical relations (Desai et al., 2018). While these structures have been used to represent logical-argumentative components of discourse segments and informative structures in reading comprehension, they have gained significant traction in analyzing the rhetoric at play in various parts of texts.

A notable application area is the study of discourse categories in scientific writing (Dernoncourt and Lee, 2017; Gonçalves et al., 2020; Kiepura et al., 2024), which, although it does not focus on political discourse, is relevant to our topic because it explores the factors that contribute to local-level text coherence. Here this approach has served to examine and define established conventions of effective scientific communication, whether for critical discourse analysis or for automatically extracting the parts of a paper which are most relevant. For example, Kiepura et al. (2024) explored the ability of Pre-trained Language Models (PLMs) to accurately detect Topic Sentences in scientific writing. They found that these models struggle to differentiate between Concluding, Transition, and Supporting Sentences. Surprisingly, their study showed that paragraphs with fully developed discourse structures are actually harder to read, challenging common beliefs about the readability of well-constructed scientific paragraphs. This last work provides an example of analyses conducted through the extraction of local discourse structures, illustrating how such extraction can be relevant.

### 2.2 Local RST Discourse Structures in Political Discourse

Delving deeper into rhetoric and the analysis of conventions of coherence, we note that the RST (Rhetorical Structure Theory) framework has been extensively used for analyzing political discourse. It has not only been proven to provide an effective set of features in NLP deep-learning models for the recognition of persuasive strategies (Chernyavskiy et al., 2024; Li and Xiao, 2021) but has also served as an innovative tool in discourse analysis for studying patterns of rhetorical relations in political discourse.

Specifically, identifying relations of interest when parsing speeches of a public figure or political party as to extract relevant relations (such as concession, contrast, etc.) for inspection has proven to be a viable method for discourse analysis (Zeldes et al., 2024).

Analyzing patterns of coherence relations on social media has yielded interesting insights into the formation of coherent textual elements in online persuasive texts. For instance, in Pastor et al. (2024) a targeted parser evaluation was proposed for extracting JOINT JOINT EVALUATION structures, which have been found to be highly prevalent in social media discussions about immigration. The ability to massively extract these discourse structures from datasets built around specific topics is valuable for understanding the discursive phenomena responsible for controversial content such as polarized debates or misinformation (de Rijk, 2020).

## 3 RST and the JJE pattern

### 3.1 Framework: RST

Rhetorical Structure Theory (Mann and Thompson, 1988) is a text-analysis model used to describe coherence within texts. With RST, the text is seg-

mented into 'elementary discourse units' (EDUs), which are contiguous spans of tokens roughly equivalent to clauses. The EDUs are then classified or annotated with various types of coherence relations, like ELABORATION, CONTRAST, CAUSAL, TEMPORAL, and others, which are categories provided by the RST framework. Relations are established not just between individual EDUs but also among groups of EDUs, forming a hierarchical tree that represents the entire text (e.g., a book, chapter, article, or social media comment).

RST differentiates between two types of EDU: "Nucleus" and "Satellite". The Nucleus contains essential information, while the Satellite provides additional details. Some relations, like JOINT and SAME-UNIT can be multi-nuclear. In Figure 1, the arrow is used to indicate that the EVALUATION refers to the nucleus.

In this article, we focus specifically on the EVALUATION relation, so it is crucial to define it clearly here. The EVALUATION relation involves directing the reader's attention towards a central point (the nucleus). This can serve either an evidential purpose, aiming to increase the reader's belief in the central material, or a justificatory purpose, seeking to enhance "the reader's willingness to accept the writer's authority in presenting the central material" (Mann and Thompson, 1988). Unlike the multinuclear nature of the JOINT relation, the EVALUATION relation is mononuclear, encompassing a single segment that presents a primary claim. In our context, this involves a subjective assessment (positive/negative, desirable/undesirable) from the perspective of the writer (Stede et al., 2017).

Following the annotation guidelines provided by Stede et al. (2017), we further detail the standards used for annotating the EVALUATION relation: In an EVALUATION relation, one span judges the situation presented in the other span on a scale from good to bad. This can take the form of an appraisal, estimation, rating, interpretation, or assessment of a situation. The viewpoint can be that of the writer or another agent within the text. The assessment may be located in the satellite (EVALUATION-S) or the nucleus (EVALUATION-N). Additionally, in cases where the spans representing the situation and the assessment carry equal weight, it may occur in a multinuclear relation.

### 3.2 The JOINT JOINT EVALUATION Pattern

Our objective is to assess the ability of RST discourse parsers to identify a specific RST coherence relation pattern that corresponds to coherent textual segments and see what the effects are of augmenting the training data with synthetic data. The pattern we selected was inspired by previous work investigating a recurring discourse strategy that creates stylistic pragmatic effects in immigration-related discourse (Pastor et al., (2024). This pattern is captured by variations of a triadic RST sub-tree containing 4 to 6 elementary discourse units (EDUs) that fit the JOINT JOINT EVALUATION structure, as illustrated in Figure 2 below.

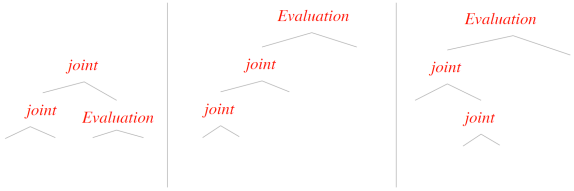


Figure 2: RST trees of extracted patterns of coherence relations. Where either a JOINT has another JOINT as a left child and EVALUATION as right child or an EVALUATION has a JOINT left child which in turn has either a left or right JOINT child.

## 4 Experimental Set-up

In this section, we detail our process for generating synthetic data for training in the context of a binary classification task (JJE or non-JJE). Before prompting the LLM to generate the required training data points, we first outline our method for annotating a gold-standard dataset of 270 cases of JOINT JOINT EVALUATION (JJE). From this annotated dataset, we select a root subset, which serves as the foundation from which the synthetic data stems. The section concludes with a description of the parsers used and our evaluation method.

### 4.1 Data

The 270 cases of JOINT JOINT EVALUATION (JJE) originate from the following three data sets coming from three different platforms. This aims to investigate how variations in social media platform styles influence parser performance.

The data originate from a subset of the SFU Opinion and Comments Corpus (SOCC; Kolhatkar et al., 2020), the 2020 Tensions over Race and Heritage Collection (TRHC; Otero et al., 2021) and the Australian Election 2019 Tweets (AUSPOL[3])

[3]https://www.kaggle.com/datasets/taniaj/australian-election-2019-tweets

The SOCC has been designed to analyze online news commentary. It includes opinion articles from The Globe and Mail, a leading Canadian daily newspaper, and comments from its readers. The curated data preserve reply structures and other metadata. We here focus on a subset of the corpus which includes 15,658 comments associated with 114 articles, all discussing the topic of immigration in Europe. In our dataset only comments were included. The TRCH was created to study antiracist protests and patrimonial attacks associated with the Black Lives Matter movement. It preserves the tree structures of 296 Reddit threads, comprising a total of 260,578 posts. The AUSPOL dataset containing 180,000 tweets, along with metadata on likes and retweets, was collected over two weeks during the Australian 2019 elections, using the Twitter API.

## 4.2 Annotating JJE Patterns

### 4.2.1 Original JJE candidates

In order to arrive at a collection of JJE structures, we used the DMRST parser (Liu et al., 2021) checkpoint[4] provided at the authors' GitHub to identify JOINT JOINT X patterns, where X represents any coherence relation extracted by the parser. We extracted 1,582 data points from the three datasets mentioned above and annotated them, focusing only on the presence of JJE patterns. All remaining JJX structures were classified as 'not JJE'.

### 4.2.2 JJE and JJX

By not limiting our initial search to JJE patterns but including JJX patterns, we were able to establish that this method of collecting JJE cases did not miss out on particular JJE types. However, we did encounter a few compelling examples that were originally labeled as JOINT JOINT JOINT or JOINT JOINT EXPLANATION, though they were scarce.

Initially, our goal was to broaden the scope to include a wider variety of JJE types, but we found only a limited number of precise instances that aligned with the phenomenon's definition.

Additionally, from the false positives identified during this phase, we created a set that was further re-annotated. This set now serves as a reference for non-JJE examples, which we use to build the 'non' class of the testing set. By doing so, we aim to assess the extent to which augmenting the training set with true cases of JJE helps the parser correctly classify closely related instances as non-JJE. This approach was taken to better simulate a real-world scenario, as selecting other triadic constellations of RST as non-classes would have made them too easy to classify as non-JJE.

### 4.2.3 Refined Annotation Guidelines for JJE

The following guidelines were informed by a pilot annotation campaign on a first (smaller) subset of the SOCC that we later extended to arrive at the subset used for our experiment. Prior to reaching consensus on the annotation criteria, the inter-annotator agreement between the two annotators had a Cohen's kappa score of 0.64.

Although we were successful in identifying valid instances of JJE in the pilot annotation phase, we later identified several problematic cases that required reconsideration before being selected as gold data.

Our criteria for selecting gold JJE instances were initially based on the annotation guidelines by Carlson and Marcu (2001) and subsequently aimed at addressing the following gray areas in RST theory for social media content.

Certain types of verbless structures, such as phatic expressions, sometimes function as independent EDUs but cannot be characterized by any relations from the RST label set. As a result, all evaluative instances that resembled phatic expressions, such as 'well said', 'that's that', 'that's all', 'congrats', 'lol', 'haha', etc. have been removed. Many of these cases, especially those used sarcastically, were challenging to classify, as only individual comments were parsed. Additionally we found that the Twitter data contained numerous symbols or series of symbols whose effects were difficult to label. Consequently, we decided to exclude any JJE instances containing links, sequences of hashtags, emojis, or other ostensible displays of punctuation or special symbols that diverged from the original definitions of the JJE discourse structure.

Lastly, we also removed all JJEs involving personal user attacks through various types of hateful assessments, where the focus was on evaluating the traits of a user rather than a particular state of affairs or opinion. Though these personal attacks fit the description of JJE in terms of structure and sequential flow, they show little interest in further analyzing the polarized social attitudes and ideological stances on immigration-related topics. These

[4]This parser is trained on nine different RST datasets spanning six languages, which have been cross-translated to augment the training data. We use this parser in our experiment as it was the one originally tested by the authors of the JJE paper (Pastor et al., 2024).

last instances being particularly prevalent in the TRHC collection.

### 4.3 Generating Synthetic Data

From the 270 collected and annotated JJEs, for each platform we randomly selected 30 from the 90 cases obtained to use for synthetic data generation. Thus the three platforms were equally represented in the prompts for the generated data. More specifically, for each of the source JJEs, we prompted the LLM to generate 10 new JJEs inspired by the example structure, yielding a total synthetic dataset of 900 cases[5]. Initially, to expand the dataset, we generated 20 new JJEs. However, as we observed that in some random instances, the LLM began generating duplicates for the last of the 20 new data points we decided to limit their number to 10.

The LLM used is GPT-4-turbo[6]. It accepts a sequence of messages as input and produces a generated message as output. We utilized the OpenAI Python library[7] to automate our prompts, allowing us to obtain more precise responses and a larger volume of data more efficiently using the GPT-4-turbo model through its API. We should note that we used the LLM in a zero-shot setting.

Figure 3 shows an example of a prompt used for generation.

message : {

*role.system*: "You are an expert discourse analyst in the field of rhetorical structure theory (RST)."

*role.user*: "I have a graph RST representation of a small text of 4 to 5 edus, Those edus are connected using joint, joint and an evaluation relation. You have to generate similar joint joint evaluation structures by changing the edus text, make the subject about a polarized topic. Make sure it's still joint joint evaluation relations: Make sure you output the code of the graph and, no extra text in the answer message, just a JSON storing a 10 examples like this:

{"1" : "//graph code 1", "2" : "//graph code 2", ...}

From the following examples, generate a JSON with 10 new original joint joint evaluation structures from the following graph: <insertgraphcode>"

}

Figure 3: Example of a prompt used for generating JOINT JOINT EVALUATION RST structures.

[5]The dataset can be accessed through the following GitHub repository https://github.com/metabolean5/coling2025-jjes

[6]https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[7]https://github.com/openai/openai-python

The format selected to represent the RST trees for the prompts is the *.mermaid* graph code. This format includes the required spans, nuclearity, relations, and hierarchical tree structures to represent JJE patterns. We opted for this format instead of the *.dis* (Carlson et al., 2001) and *.rs3* (Zeldes, 2016) formats because it is widely used for representing graphs, and there are numerous examples available online, which have contributed to the training data of LLMs. Additionally, we discovered that it was easier for the LLM to reproduce the connections between the relations and the EDUs using this format compared to *.dis* or *.rs3*, where the generated data often contained multiple formatting mistakes. Listing 1 is an example of the *.mermaid* graph code that was requested for generation by the LLM.

```
flowchart TD
    2---N2[These immigrants are not like
        us,]
    2(JOINT)---3
    3(JOINT)
    3---3_ltxt[they don't think like we
        do,]
    3---3_rtxt[they don't meld like we
        have.]
    1(EVALUATION)---2
    1---M1[Massive problems ahead.]
```

Listing 1: *.mermaid* graph code example of a JJE from **SOCC_27725714_20_0**

The resulting dataset is summarized in Table 1. Table 1 also shows the data splits used for training and testing.

| Dataset Information | not JJE | JJE |
|---|---|---|
| Original Instances | 180 | 270 |
| Instances for Synthetic Data Generation | - | 90 |
| Generated Synthetic Data for Training | - | 900 |
| Test Set Instances | 180 | 180 |

Table 1: Summary of Dataset and Synthetic Data Generation.

### 4.4 Discourse Parsers and Training

To examine parsing performance across different architectures, we select two SOTA parsers for RST: a BOTTOM-UP model by Guz and Carenini (2020), and a TOP-DOWN model introduced by Liu et al. (2021). Both parsers are configured according to the best practices outlined in their respective papers and GitHub repositories. The bottom-up parser uses the SpanBERT-NoCoref setting and was trained for 20 epochs, while the top-down

parser employs XLM-RoBERTa-base (Conneau et al., 2020) and was trained for 15 epochs. For the present experiment, both parsers were first trained from scratch on the RST-DT + GUM data and then on RST-DT + GUM + synthetic data.[8]

| | Synthetic Data | RST-DT | GUM |
|---|---|---|---|
| # Documents | 900 | 347 | 235 |
| # EDUs | 4,150 | 14,292 | 29,577 |
| # EVALUATION | 920 | 1,014 | 1,187 |

Table 2: Overview of RST datasets used for Training.

### 4.5 Evaluation

To the best of our knowledge, no existing 'parsing in the wild' method has been developed to measure parser performance in extracting local RST discourse structures from larger texts. Discourse parsers are typically evaluated based on their ability to reconstruct tree structures and classify the nuclearity and discourse relations between gold-standard pre-segmented EDUs. However, this approach provides limited insight into how such parsers would perform when applied directly to raw text. In this study, the parser is tasked with parsing raw text without any provided gold EDU breaks. Each parser produces the entire tree[9], and potential JJE candidates are subsequently extracted from the full tree[10] based on the inorder tree patterns shown in Figure 2. For each of the extracted candidates, we then assess whether the structure and content of the EDUs match the gold-standard JJE that is expected to be present in the full text provided for parsing. The method for determining whether a gold JJE is recognized by any candidate is described in detail with the matching function below where *G* is the gold JJE and *J* is the set of JJE candidates.

$$\mathrm{f}(G, J) = \begin{cases} & \exists j \in J_{\text{filtered}}, \\ \text{True}, & \exists c \in C, \\ & \exists e \in E_j : (c \subseteq e \text{ or } e \subseteq c) \\ \text{False}, & \text{otherwise.} \end{cases}$$

First, $J_{\text{filtered}} = \{j \in J : 4 \leq \text{EDUlength}(j) \leq 6\}$ is the set of candidates with EDU length between 4 and 6. Then, for each $j \in J_{\text{filtered}}$, $E_j = \{e_1, e_2\}$ is the set of the last two EDUs extracted from $j$ and $C = \{c_1, c_2\}$ is the set of the last two EDUs from the gold discourse structure $G$. Finally the substring matching condition $c \subseteq e$ denotes that $c$ is a substring of $e$.

Since it is unlikely that the parser will achieve perfect segmentation, we opted to match using the last two EDUs to account for potential parser errors. This approach ensures that even if the segmentation is imperfect, the candidate JJE still has a chance of aligning with the gold-standard JJE.

## 5 Results and Discussion

In this section, we present and discuss the results obtained by the different parsers, accompanied by an assessment of the generated synthetic data.

### 5.1 Parser Performance in JJE Identification

| System | Label | Precision | Recall | F1 |
|---|---|---|---|---|
| | JJE | 0.87 | 0.11 | 0.20 |
| | not JJE | 0.53 | 0.98 | 0.68 |
| Top Down | avg. | 0.70 | 0.55 | 0.44 |
| | JJE | 0.64 | 0.48 | 0.55 |
| . | not JJE | 0.58 | 0.72 | 0.65 |
| Top Down + Synth. | avg. | 0.61 | 0.60 | 0.60 |
| | JJE | 0.50 | 0.06 | 0.11 |
| | not JJE | 0.50 | 0.94 | 0.65 |
| Bottom-Up | avg. | 0.50 | 0.50 | 0.38 |
| | JJE | 0.68 | 0.41 | 0.51 |
| | not JJE | 0.58 | 0.81 | 0.67 |
| Bottom-Up + Synth. | avg | 0.63 | 0.61 | 0.59 |

Table 3: Comparison of parsing performance across different systems and labels.

As shown in Table 3, upon the addition of the synthetic data both parsers demonstrate impressive gains in overall F1, with increases of 16 points for the top-down approach and 21 points for the bottom-up approach. Of note, the top-down + synthetic approach achieves the best results with an F1 score of 0.60, outperforming the bottom-up parser by just 1 point.

Where the baseline bottom-up performed very poorly on extracting instances of JJE, we observe that the baseline top-down parser exhibited good precision in the cases it identified, but its recall was notably low. The introduction of diverse synthetic data did lead to an improvement in recall; however, this came at the cost of a decline in precision. This is most likely due to the selected non-JJE instances, as observed in the pilot, which have structures very

[8] For a description of the RST Discourse Treebank (RST-DT), see Carlson et al. (2001). The GUM dataset is documented in Zeldes (2017).

[9] Since the bottom-up parser does not include an EDU segmenter, we provided it with the EDU segments generated by the top-down parser.

[10] The trees are reconstructed from the constituency format outputs of both the bottom-up and top-down parsers, and are then processed using a tree evaluation code.

similar to JJE and are thus more prone to confusion, particularly when the final relation in the inorder traversal of the tree is a JOINT.

| System | SOCC | TRHC | AUSPOL |
|---|---|---|---|
| Top-down | 0.39 | 0.42 | 0.49 |
| Top-down + Synth. | 0.64 | 0.64 | 0.56 |
| Bottom-up | 0.42 | 0.41 | 0.32 |
| Bottom-up+ Synth. | 0.57 | 0.57 | 0.62 |

Table 4: F1 Scores for SOCC, TRHC, and AUSPOL by System.

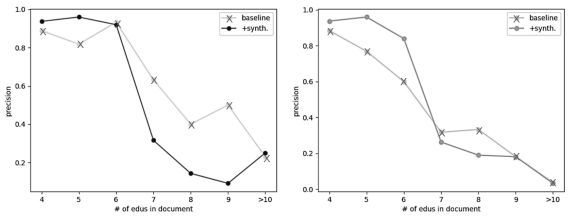


Figure 4: Precision scores by the number of EDUs present in parsed documents: left represents the TOP-DOWN parser, and right represents the BOTTOM-UP parser.

With reference to Table 4, it is important to note that the performance of the augmented parsers remains stable across the three different datasets. However, it is worth mentioning that the augmented top-down parser underperforms compared to the bottom-up parser on the AUSPOL dataset, resulting in its worst performance on the Twitter data. This outcome is surprising, given that the average number of EDUs per document for this genre is significantly lower than for the others (AUSPOL: 7.8 compared to 12.6 and 12.4 for TRHC and SOCC). As will be discussed below, documents with fewer EDUs tend to be parsed more successfully. The underperformance is attributed to the top-down parser's tendency to consistently classify the final hashtag in a tweet as an EVALUATION, leading to a confusion with the scope of the evaluative statement which we elaborate on below.

The graphs in Figure 4 show that the precision of the parser drops significantly when the document to parse contains more than 6 EDUs. For most cases this is because the parser struggles to determine which group of EDUs the evaluative statement should reason over. EVALUATIONs are among the most difficult relations to predict in RST (Liu and Zeldes, 2023). Unlike relations such as ATTRIBUTION, which benefit from clear signals like quotation marks or speech attribution verbs to delineate the set of EDUs attributed to someone, EVALUATION relations often lack consistent overt signals. This results in a tree-shift issue, where, although the relation labels are correctly predicted, entire sub-trees are *shifted* into the satellite material of the EVALUATION.

## 5.2 Assessing the Quality of Synthetic Discourse Structures

Our assessment of the quality of the synthetic data was motivated by the wish to understand the extent to which the synthetic data adds diversity to the original gold data used for generation. To this end, we annotated a subset of 280 instances derived from 14 gold JJEs. This particular subset comes from our initial experimentation discussed in 4.3, where 20 new JJEs were originally generated. We investigated how the utterances are realized in the synthetic data versus the human (gold) data by examining the differences and similarities.

### 5.2.1 Similarities

The synthetic data closely resembles the examples provided in the prompts regarding text length and topic. However, a potential issue arises when the synthetic data mimics the superficial structure of the gold JJE, like a recipe, offering minimal variation:

(1) [Canada has a higher net immigration rate than Sweden.][Canada's population is 25% foreign born.][Sweden's population is 19% foreign born.][Canada is way ahead of Sweden when it comes to immigration.] **SOCC_26338254_98_2_1_0**

(2) [The US accepts more immigrants annually than Germany.][The US's immigrant population is 15% of its total population.][Germany's immigrant population is 12% of its total population.][The US leads Germany in terms of immigrant acceptance.] **synth_SOCC_26338254_98_2_1_0_1**

### 5.2.2 Differences at the rhetorical level

Though we see that the coherence relations are correct and still aim to create a special emphasis on the evaluative statement, the LLM does fail to generate or replicate sarcasm. The synthetic JJEs use an impersonal style, as evidenced for example by the lack of first-person pronouns and vocatives. Additionally, the stance in the synthetic data is mostly positive (or neutral), while the human data is not.

(3) [Most immigrants are not even interviewed by visa officers anymore.][Their consultants produce "perfect" applications for them,][and they sail through on paper.][The system is out of control.] **SOCC_26338254_0_4_0_0_2_1_0**

(4) [Detention centers for immigrants have been criticized.][Conditions are often poor,][and human rights are at risk.][Immediate reforms are needed.] **synth_SOCC_26338254_0_4_0_0_2_1_0_20**

It is interesting to observe that when the LLM avoids creating sarcasm, it compensates by producing a new form of JJE that provides diversity in terms of syntactic structure.

#### 5.2.3 Differences at the syntactic level

When we compare the syntactic structures in the synthetic data and the human data, we see that the LLM generates variants for various types of sentences, clauses, and phrases. For example, if the provided example contains a *yes-no* interrogative sentence, the LLM might use another type of interrogative (e.g., a *wh*-interrogative) or a declarative sentence.

Altogether, the synthetic data shows some biases in the preference for certain types of sentences and structures that differ from those in the human data. In the synthetic data, it is much more common to use full sentences rather than clauses or phrases. The LLM also tends to generate more declaratives in the JOINT EDUs (75% of the annotated synthetic JJEs are declaratives compared to 62% in the gold data).

While the synthetic and gold data show similar distributions over different complementation types (e.g., intensive complementation[11] being used more frequently in EVALUATIONs than in JOINTs), the use of present and past tense differs. In the gold data, the past tense is used more frequently.

Finally, in terms of the realization of the subject, the synthetic data includes several cases (especially with EVALUATIONs) where the subject is realized by a clause (17%), whereas in the gold data, there are no such cases.

In conclusion, we found that all instances of synthetic JJEs were considered to be intelligible and coherent pieces of text, which stayed on-topic and had correct coherence relations. Though some of the synthetic JJEs almost exactly replicated the structures of the source JJEs at the syntactic level, most of them went beyond simply substituting the lexical items in the provided examples and offered more diversity with respect to the linguistic characteristics described above. Lastly, it is worth mentioning that, apart from using LLMs to increase the amount of data, synthetic data can potentially be used for a qualitative exploration of the data, which may help identify linguistic characteristics that can be ingested as features into an automatic system.

## 6 Conclusion and Future Research

The experiments in this article demonstrate that current discourse parsers, regardless of architecture, benefit from incorporating LLM-generated synthetic data for training. Both the top-down and bottom-up parsers exhibit comparable performance across the three different social media subsets, showing minimal differences in their accuracy with respect to what they classify correctly or incorrectly. In terms of error analysis, this suggests that when dealing with local JJE structures, the RST scope of evaluative statements remains challenging to determine automatically. Nevertheless, the results are promising and highlight the need for further exploration to address the scarcity of specific local structures for different RST constellations across various genres and text types.

We also outlined the complexities involved in developing RST corpora for social media, emphasizing aspects such as segmentation, relation labeling, and the need for clarity in interpreting certain discourse characteristics like phatic expressions, sarcasm or personal attacks when annotating discourse relations. These considerations were integral to the development of the RST dataset for JJE used in our experiments, resulting in 1,170 annotated RST instances of JJE across three different social media platforms. This dataset can be used for the further qualitative exploration of the linguistic features that characterize EVALUATION relations in polarized social media discussions.

We suggest that future research on enhancing discourse parsers with synthetic data for localized structures should go beyond the JJE structure, exploring its relevance to other discourse patterns. In line with this, the present work only experiments with social media data from polarized discussions and we would be interested in exploring the extent

[11] Found with sentences or clauses with a copular verb and a subject complement, e.g. *Cultural dynamics are unpredictable.*

to which this approach could be applied to other types of text. This would allow us to examine the extent to which parsers benefit from synthetic data, rather than just examples from the same domain-specific origin.

## 7 Limitations

One limitation of this study is that we generated synthetic data using excerpts of comments for training, and training on full tree structures might have yielded different results.

Additionally, the improvements in parser performance are primarily in recall, likely because the current method only augments the parser for detecting a single class (JJE), with no additional non-JJE examples included in the training.

## Acknowledgments

This work was produced as part of the HYBRIDS project, a Marie Skłodowoska-Curie Doctoral Network funded by the European Union under grant no. 101073351 and the UK Research and Innovation (UKRI) Horizon Funding Guarantee.

## References

Kelsey Allen, Giuseppe Carenini, and Raymond Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1169–1180, Doha, Qatar. Association for Computational Linguistics.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. volume 54, page 56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen Mckeown, and Alyssa Hwang. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943.

Jose M Chenlo, Alexander Hogenboom, and David E Losada. 2013. Sentiment-based ranking of blog posts using Rhetorical Structure Theory. In *Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings 18*, pages 13–24. Springer.

Alexander Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2024. Unleashing the power of discourse-enhanced transformers for propaganda detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1452–1462.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Lynn de Rijk. 2020. You said it? How mis-and disinformation tweets surrounding the Corona-5G-conspiracy communicate through implying. In *Working Notes Proceedings of the MediaEval*, Online. CEUR-WS.org.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Takshak Desai, Parag Dakle, and Dan Moldovan. 2018. Generating questions for reading comprehension using coherence relations. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. volume 32, pages 6793–6807. Springer.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for Computational Linguistics.

Anna Kiepura, Yingqiang Gao, Jessica Lam, Nianlong Gu, and Richard H.r. Hahnloser. 2024. SciPara: A new dataset for investigating paragraph discourse structure in scientific papers. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 12–26, St. Julians, Malta. Association for Computational Linguistics.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. volume 4, pages 155–190. Springer.

Jinfen Li and Lu Xiao. 2021. Neural-based RST parsing and analysis in persuasive discourse. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 274–283, Online. Association for Computational Linguistics.

Xingyun Liu and Xiaoqian Liu. 2021. Online suicide identification in the framework of Rhetorical Structure Theory (RST). In *Healthcare*, volume 9, page 847. MDPI.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? A thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. volume 8, pages 243–281. Walter de Gruyter, Berlin/New York Berlin, New York.

David Otero, Patricia Martin-Rodilla, and Javier Parapar. 2021. Building cultural heritage reference collections from social media through pooling strategies: the Case of 2020's tensions over race and heritage. volume 15, pages 1–13. ACM New York, NY.

Martial Pastor, Nelleke Oostdijk, and Martha Larson. 2024. The contribution of coherence relations to understanding paratactic forms of communication in social media comment sections. In *JADT 2024: 17th International Conference on Statistical Analysis of Textual Data*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. volume 43, pages 619–659. MIT Press.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. Annotation guidelines for rhetorical structure. *Manuscript. University of Potsdam and Simon Fraser University*.

Xiaoyu Wang, Hong Zhao, Hongzhi Zhu, and Fang Wang. 2023. Towards intelligent policy analysis: A discourse structure parsing technique for Chinese government document. volume 60, page 103363. Elsevier.

Amir Zeldes. 2016. rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. volume 51, pages 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. *Computational Linguistics*, pages 1–50.