

# Why Can't Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity

Yang Janet Liu and Amir Zeldes

Department of Linguistics

Georgetown University

{y1879, amir.zeldes}@georgetown.edu

## Abstract

Recent advances in discourse parsing performance create the impression that, as in other NLP tasks, performance for high-resource languages such as English is finally becoming reliable. In this paper we demonstrate that this is not the case, and thoroughly investigate the impact of data diversity on RST parsing stability. We show that state-of-the-art architectures trained on the standard English newswire benchmark do not generalize well, even within the news domain. Using the two largest RST corpora of English with text from multiple genres, we quantify the impact of genre diversity in training data for achieving generalization to text types unseen during training. Our results show that a heterogeneous training regime is critical for stable and generalizable models, across parser architectures. We also provide error analyses of model outputs and out-of-domain performance. To our knowledge, this study is the first to fully evaluate cross-corpus RST parsing generalizability on complete trees, examine between-genre degradation within an RST corpus, and investigate the impact of genre diversity in training data composition.

## 1 Introduction

Discourse parsing is the task of identifying and classifying the coherence relations that hold between different parts of a text, such as recognizing that one sentence specifies the cause of events in another, or that a subordinate clause indicates the purpose of a main clause. In hierarchical frameworks, such as the Rhetorical Structure Theory (RST, Mann and Thompson 1988), parsers construct a labeled constituent tree of discourse units as shown in Figure 1. Such trees have numerous applications: retrieving specific relations (e.g. all CONCESSIONS made in any speech by some politician, e.g. unit 26-27 in Figure 1), or hierarchically finding the most central unit in an arbitrary span of text (unit 24 in Figure 1). They can also contribute to downstream tasks such as text generation

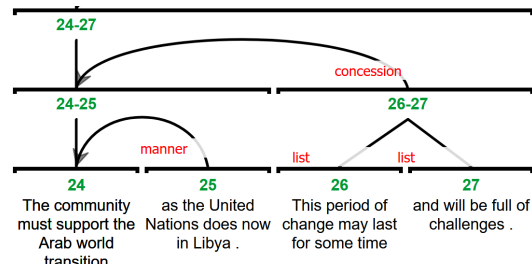


Figure 1: RST Fragment from GUM (Zeldes, 2017).

(Maskharashvili et al., 2021) and summarization (e.g. Louis et al. 2010; Li et al. 2016; Xu et al. 2020; Hewett and Stede 2022) and to qualitative analysis and comparison of texts (Wan et al., 2019).<sup>1</sup>

Recent advances in NLP have resulted in increasingly accurate systems for tasks such as part-of-speech tagging (Heinzerling and Strube, 2019) and dependency parsing (Mrini et al., 2020), which are now commonly applied with confidence to novel data, at least for high-resource languages such as English. At the same time, rising scores on the main standard English RST benchmark, the RST Discourse Treebank (RST-DT, Carlson et al. 2003), with data from the 1989 Wall Street Journal (WSJ), create the impression that discourse parsing too is becoming reliable, and by proxy, applicable to arbitrary text. This has resulted in use of automatic RST parsing to automatically generate large-scale datasets, such as MEGA-DT (Huber and Carenini, 2020) or AMALGUM (Gessler et al., 2020). However, there are some indications that this picture is too optimistic, with some studies revealing problems in RST-DT-based parsing once the target domain changes (Ferracane et al. 2019; Wang et al. 2019; Linscheid et al. 2021; Atwell et al. 2021, 2022; Nishida and Matsumoto 2022; Yu et al. 2022), and qualitative inspection of parser outputs casting doubt on their reliability.

<sup>1</sup>This applies to other discourse frameworks too, such as PDTB (Prasad et al., 2019) or SDRT (Asher and Lascarides, 2003), but we limit the scope of this paper to RST parsing.

A problem in evaluating the stability of RST parsing for English is the scarcity of out-of-domain (OOD) data, as well as potential differences between the annotation schemes of different corpora. Recently, the availability of increasing amounts of non-newswire data, as well as non-WSJ news data, in the English GUM corpus (Zeldes, 2017) has made some detailed comparisons possible, which we explore below. Beyond evaluating OOD degradation, the existence of these large datasets for English RST allows us to test different joint training strategies to increase both the accuracy and generalizability of RST parsing for English.<sup>2</sup>

Our main goals here are: 1) to demonstrate the generalizability limitations of English RST parsing based on RST-DT and quantify them for users; 2) to explore reasons for generalizability issues, with a focus on the genre composition of training sets, pointing the way to the kind of data that robust discourse parsing requires; and 3) to promote multi-genre benchmarks for RST parsing based on our experimental results. Overall we find that diverse training data leads to better generalization on unseen genres regardless of model architecture.

## 2 Related Work

### 2.1 English RST Corpora

The RST Discourse Treebank (RST-DT, Carlson et al. 2003) is the standard English RST benchmark, with data from the WSJ section of the Penn Treebank (PTB, Marcus et al. 1993). Another human-annotated English RST corpus is the Georgetown University Multilayer (GUM) corpus, which is freely available online and covers 12 written and spoken genres (Zeldes, 2017). GUM is continuously growing, with new data added in each version. For this paper, we used Version 8 of the corpus, which is described in Table 1 next to information about RST-DT (the current version of GUM, V9, was not yet available at submission time, and now contains 213 documents, 203K tokens and 26K EDUs, and the next version 10 is set to add 4 additional new genres).

As Table 1 shows, although GUM V8 is smaller than RST-DT in number of tokens and documents,

<sup>2</sup>Though we focus on English in this paper, understanding the size and nature of data needed in English will hopefully shed light on what other languages may require.

<sup>3</sup>There is no established dev partition for RST-DT.

<sup>4</sup>Previous work has identified 4 sub-genres in the WSJ data (Webber, 2009), but these are very unevenly distributed, and 29 RST-DT documents were not included in that analysis.

	RST-DT	GUM V8
<b>documents</b>	385	193
<i>train/dev/test</i>	347/— <sup>3</sup> /38	145/24/24
<b>tokens</b>	203,352	180,851
<b>EDUs</b>	21,789	23,107
<b>relation instances</b>	20,163	21,903
<b>relation labels</b>	78 / 17 classes	32 / 15 classes
<b>genres</b>	1 <sup>4</sup>	12

Table 1: Overview of the English RST Corpora.

it contains more instances of elementary discourse units (EDUs) and relations, since newswire units from RST-DT are longer on average. EDU segmentation guidelines are identical for the two corpora, and the label set is very similar, with both corpora using the SAME-UNIT pseudo-relation for discontinuous EDUs. GUM’s 12 genres make it possible to conduct experiments to investigate parsing generalizability (see Table 12 in Appendix B for their exact size breakdowns). Virtually all work on RST constituent parsing uses only the collapsed coarse relation classes (i.e. 17 labels for RST-DT); we will follow this practice for the evaluation below, but will refer to some issues relating to fine-grained labels in our analysis as well. In addition, we conducted an analysis of satellite-nuclearity patterns in the two corpora, confirming that they are quite similar: NS dominates with 77.7% in RST-DT, similar to 74.8% in GUM news; interestingly, proportions in GUM overall are somewhat lower (70.1%).

There are also other corpora annotated in RST dependencies, such as SciDTB (Yang and Li, 2018) and full RST constituent treebanks in other languages (see Zeldes et al. 2021 for an overview); since we focus on hierarchical English RST constituent parsing, we will use the corpora in Table 1 for the experiments below: they differ in content, vocabulary, domains, and more importantly, underlying communicative intents mirroring discourse relations, all of which are the backbone of our experiments to investigate English RST parsing generalizability as a function of training data.

### 2.2 RST Discourse Parsers

Several approaches have been proposed for RST parsing, primarily distinguished by a BOTTOM-UP vs. TOP-DOWN approach, and the algorithm used (often a neural shift-reduce architecture). Table 2 compares recent high scoring parsers on three metrics: **Span** (whether subtrees span the right EDUs), **Nuclearity** (whether edges point the right way), and **Relation** (whether labels are correct). Following

	S	N	R	Neural (Y/N)	Model Architectures	Additional Features / Resources	Pre-trained LM
BOTTOM-UP							
Ji and Eisenstein (2014)*	64.1	54.2	46.8	N	transition-based + SVMs	lexical, dependency, graphical features	–
Yu et al. (2018)◇	71.4	60.3	49.2	Y	transition-based + Bi-LSTM; encoder-decoder + dynamic oracle	implicit syntactic features	–
Guz and Carenini (2020)♣	<b>76.5</b>	65.9	<b>54.8</b>	Y	transition-based + MLPs	organizational features (Wang et al., 2017)	SpanBERT-base
Yu et al. (2022)♣	76.4	<b>66.1</b>	54.5	Y	Yu et al. (2018) + 2 pre-training tasks: Next EDU Prediction; Discourse Marker Prediction	–	XLNet
TOP-DOWN							
Koto et al. (2021)♣	73.1	62.3	51.5	Y	LSTM + dynamic oracle	sentence and paragraph boundaries	–
Nguyen et al. (2021)♣	74.3	64.3	51.6	Y	Bi-LSTM encoder + unidirectional LSTM decoder	–	XLNet-base
Zhang et al. (2021)♣	76.3	<b>65.5</b>	<b>55.6</b>	Y	split-point encoding and ranking + adversarial learning	–	XLNet-base
Liu et al. (2021)♣	<b>76.5</b>	65.2	54.2	Y	pointer-network decoder + depth-first span splitting + biaffine classifiers	–	XLNet-base

Table 2: Micro-averaged Parser Scores on RST-DT with Gold Segmentation. \* = scores from Morey et al. (2017). ◇ = converted Parseval scores for Yu et al. (2018)’s parser reported by Koto et al. (2021). ♣ = 5 run average Original Parseval replication of respective work. ♣ = scores from the original paper (not necessarily averaged scores).

Morey et al. (2017), we use the more stringent original Parseval procedure on binary trees.

As Table 2 shows, BOTTOM-UP and TOP-DOWN approaches are currently very close, and all state-of-the-art (SOTA) systems rely on transformer word embeddings as the primary input. To rule out an impact of parser architecture, we test both the BOTTOM-UP from Guz and Carenini (2020), using their best SpanBERT–NoCoref setting, and the TOP-DOWN from Liu et al. (2021) as their code is public and both have near-SOTA S and N metrics, which are more robust to differences between label inventories across corpora (relevant hyperparameters and dev performance are provided in Appendix C).

### 2.3 Cross-Genre Variation in RST

Previous work suggests RST trees vary widely across text types,<sup>5</sup> in relation distributions (Taboada and Lavid, 2003; Zeldes, 2018), ways they are signaled (e.g. Eggins and Martin 1997; Liu 2019; Demberg et al. 2019), and variation across languages (Cao et al. 2018; Stede and Neumann 2014; Iruskieta et al. 2013; Redeker et al. 2012; da Cunha et al. 2011; Pisarevskaya et al. 2017).

These differences suggest OOD parsing will be challenging: For example, Taboada and Lavid (2003) showed that in appointment-scheduling conversations, CONCESSION, CONDITION, CAUSE, and RESULT were disproportionately frequent due to the nature of the data, correlating with stages of a conversation (i.e. Opening, Task Performance, Closing): e.g. more RESTATEMENT, EVALUATION, and SUMMARY in the Closing stage. Similarly, Zeldes (2018) showed MOTIVATION often encodes how-to guides influencing one’s willingness to act,

<sup>5</sup>We use “text type”, “genre”, and “domain” interchangeably here to mean different kinds of data, though they are not exactly the same (e.g. see Biber 1988 and Lee 2001).

which almost never occurs in the objective style of news. We therefore expect parsers trained on a single domain to generalize poorly to novel ones.

### 2.4 Generalizability in RST Parsing

Although some first studies are beginning to appear, generalizability in full RST parsing remains understudied. Three recent papers include some cross-dataset numbers, but not a systematic analysis of full RST constituent parser generalizability. Atwell et al. (2021) examined the subtask of relation classification between PDTB (Prasad et al., 2019) and RST annotated data, and included F1 scores for identifying relations in GUM V5 using the RST-DT training data for relations that appear in both datasets (however, GUM V5’s EDU segmentation differed strongly from RST-DT and the parser in the paper was an older non-neural one).

Yu et al. (2022) evaluate a neural parser on RST-DT and report scores on 11/12 genres in GUM V7 (same segmentation as RST-DT, but only 25 relations and less data than V8); however because their focus is not on cross-genre generalization, they do not report micro-averaged scores on the whole GUM corpus, do not evaluate training on GUM or joint training, and do not test held-out genres within GUM. Finally, Atwell et al. (2022) used GUM V6 (with RST-DT’s segmentation and 8 genres) to identify relation types between pairs of provided argument spans, similar to PDTB-style shallow discourse parsing in the CoNLL’16 shared task (Xue et al., 2016) and recent DISRPT shared task on relation classification (Zeldes et al., 2021).

This study is therefore the first to fully evaluate cross-genre RST parsing generalizability on complete trees in datasets with the same EDU segmentation. To our knowledge, no previous work has tested an OOD-trained parser on RST-DT, at-

tempted joint training on both corpora, evaluated between-genre degradation within an RST corpus, or the impact of training data genre composition.

### 3 Experiments

In this section we conduct four experiments. First, we want to find out how compatible the two English RST corpora are, and how well they generalize to each other’s domains (§3.1). Second, we want to test whether we can incorporate training data from GUM in order to exceed SOTA performance on RST-DT, with one of three approaches: joint training, model stacking, and pretraining (§3.2). Third, we would like to know whether generalizability issues are mainly due to differences between corpora (details of annotation guidelines or common practices) or the nature of the texts’ genres and domains themselves (§3.3). We test this by leaving out different subsets of GUM genres from training, and quantify OOD degradation for each genre, which also informs users of SOTA parsers of what degradation they can expect on genres ‘in the wild’. The final experiment (§3.4) tests whether, given a fixed training set size, having relatively many ‘small’ genres leads to better OOD stability than having few genres with more material each, when the target is a totally disjoint set of genres. The code, trained and finetuned models, and predicted GUM parses are available at <https://github.com/janetlaueung/crossGENRE4RST>.

#### 3.1 Cross-Corpus Generalization

In this section we hypothesize that, since GUM contains many genres, models trained on it will degrade less when testing on RST-DT than in the opposite scenario. This is non-trivial, since the corpora are very close in total size, and it is possible that diverse training genres with little data each will prevent effective learning for any genre. We train the two parsers identified in §2.2 from [Guz and Carenini \(2020\)](#) (BOTTOM-UP) and [Liu et al. \(2021\)](#) (TOP-DOWN) on the train partition of each dataset, and report scores on the test set (since RST-DT has no dev partition, we follow previous work in using 10% of training data stratified by the number of EDUs in each document as a dev set, which remains the same in the training).

A possible limitation of our experiment is annotation differences between the corpora, especially for their relation inventories, which are similar, but not identical. We therefore omit R scores in Ta-

bles 3 and 4, though we will return to relations in more detail below. For S and N, however, we expect differences to be small: both corpora use the same segmentation guidelines, and the same guideline of subordinating less prominent satellite units to more prominent nuclei using tests such as deletability and subjective prominence. Although EDU segmentation is not our focus, we can confirm the compatibility of segmentation guidelines by cross-testing the current SOTA EDU segmenter from [Gessler et al. \(2021\)](#). The results show fairly modest degradation training on RST-DT and testing on GUM, from 94.9 to 89.9, while the opposite direction shows a 92.9 to 91 drop (see Table 11 in Appendix A for full numbers). We speculate that segmentation generalizes well due to the syntactic, and hence less sparse and domain-specific nature of the task, which resembles clause boundary detection.

For cross-corpus S and N scores, both parsers show a very significant degradation when training on RST-DT to parse OOD data from GUM, as shown in Tables 3 and 4. For instance, there is a degradation of  $\sim -11$  points for S and  $\sim -16$  for N using the the BOTTOM-UP architecture. We were also curious whether this applies to GUM’s news genre too: indeed, testing only on GUM *news* reduces degradation by about 50% using the BOTTOM-UP parser, but still shows a substantial performance hit, even if the model is trained exclusively on news (albeit from much older, WSJ news data). By contrast, the GUM-trained model actually scores better on RST-DT than on GUM, with minor improvements of +2.8 on S and +0.4 on N using the BOTTOM-UP parser. The same degradation pattern is observed in the TOP-DOWN parser performance: substantial degradation overall, and worse when training on news and testing other genres.<sup>6</sup>

train	test	S	N	R
RST-DT	RST-DT	76.5	65.9	54.8
	GUM	65.3 (-11.2)	49.5 (-16.4)	–
	GUM <i>news</i>	71.0 (-5.5)	57.5 (-8.4)	–
GUM	GUM	69.9	57.0	48.5
	RST-DT	72.7 (+2.8)	57.4 (+0.4)	–
	GUM <i>news</i>	71.6	58.5	49.5

Table 3: Cross-Corpus Results (5 run average) of the BOTTOM-UP Parser from [Guz and Carenini \(2020\)](#).

<sup>6</sup>We further analyzed left vs. right-branching trees during testing. In both corpora, NS outperformed SN: RST-DT in-domain scores are F1=82.6% vs. 73.4% for NS vs. SN; GUM V8 has 81.6% vs. 72.4%, meaning SN transitions are ‘easier’.

train	test	S	N	R
RST-DT	RST-DT	76.5	65.2	54.2
	GUM	66.2 (-10.3)	50.8 (-14.4)	-
	GUM <i>news</i>	67.9 (-8.6)	55.8 (-9.4)	-
GUM	GUM	68.6	54.9	46.1
	RST-DT	71.1 (+2.5)	55.9 (+1.0)	-
	GUM <i>news</i>	73.4	63.3	57.2

Table 4: Cross-Corpus Results (5 run average) of the TOP-DOWN Parser from Liu et al. (2021).

We interpret this result to mean that unlike in EDU segmentation, where results are close in both directions, genre composition of the train and test data plays a crucial role in the generalizability of RST constituent parsing, regardless of parser architecture. Thus, we opt for the BOTTOM-UP parser for the experiments presented in the rest of the paper as it outperformed the TOP-DOWN parser on both corpora.

To be clear, the results do not suggest that training on GUM is a way of achieving top performance on RST-DT: the GUM model’s scores are still almost 4 and 8 points below the RST-DT trained model for S and N. However, it seems that RST-DT news data is less surprising for the GUM model which has already seen some news, and in sum, RST-DT data appears to be a comparatively ‘easy’ target given the broad genre inventory that the GUM model is trained to tackle.

These results also confirm that training on multiple genres, each with comparatively fewer documents, can lead to good performance with only minor degradation on the very narrow WSJ domain from RST-DT. While we think this result by itself is important and suggests that RST parsing work should devote more attention to multi-genre corpora as benchmarks, it leaves a question open: Can we combine data from both corpora to boost English RST parsing performance on RST-DT?

### 3.2 Joint Training

If RST parsing generalizes across domains, it may be possible to see gains through joint training. For this we compare three approaches: naive concatenation, model stacking, and pretraining, which we evaluate on the RST-DT benchmark.

For concatenation (CONCAT) we must map relations across corpora. As in previous work, we target coarse relation classes, most of which are identical in both corpora. Exceptions include GUM’s PHATIC relation used for dysfluencies and back-channeling (due to conversational

data), GUM’s PREPARATION, and the mapping between RST-DT’s TEMPORAL and COMPARISON classes, which map differently onto GUM’s CIRCUMSTANCE and SEQUENCE, or GUM’s ANTITHESIS, depending on the precise senses. Although the number of relation instances affected by these mismatches is modest, it is not negligible (13.3%), and we do not expect this approach to outperform in-domain training, and mainly report on it for completeness. Our mapping is given in Appendix D.

For model stacking, we test three variants: 1) FLAIR-LABEL: train an LSTM using FLAIR (Akbiik et al., 2019) to predict EDU dependency labels: the LSTM receives text for three-EDU chunks, set apart by separators, and predicts the middle EDU’s label in the GUM corpus scheme (using the RST dependency conversion from Li et al. 2014). We then train the parser on RST-DT with predicted GUM labels for each EDU as an additional feature encoded as a dense embedding, requiring no relation mapping. 2) SR-LABEL: train a full shift-reduce parser on GUM, generate predictions for RST-DT in the GUM scheme, and collapse these into dependencies to create the same kind of features; this approach gives the GUM classifier more access to global context than the previous LSTM’s EDU triples, but may be more vulnerable to sparseness. 3) SR-GRAPH: because it is possible that incompatibility of RST-DT and GUM labels may cause confusion, we also attempted using the same parser as in the previous approach, but featurizing each EDU’s predicted dependency attachment direction, and EDU distance to the parent EDU, instead of the label itself. Finally, we test a simple pretraining approach, in which we finetune the underlying SpanBERT model (Joshi et al., 2020) on full parsing of the GUM corpus, then load the fine-tuned SpanBERT, and train again on RST-DT (SR-FT).

Table 5 shows that SR-FT achieved the best performance compared to the other approaches and is on par with the pure in-domain training on RST-DT for S and R compared to previous work. There is a minor but stable gain on average (65.9→66.2) for N in SR-FT, which was verified by rerunning the experiment, as well as the selected SOTA system from Guz and Carenini (2020), 5 times. The remaining scenarios are virtually equivalent to training on RST-DT alone, suggesting that added features are more distracting than helpful.

This result is somewhat surprising given that scores are not very high, and there should still be

	S	N	R	architecture
Zhang et al. (2021)*	76.3	65.5	<b>55.6</b>	TOP-DOWN
Liu et al. (2021) <sup>◇</sup>	<b>76.5</b>	65.2	54.2	TOP-DOWN
Guz and Carenini (2020) <sup>◇</sup>	<b>76.5</b>	<b>65.9</b>	54.8	BOTTOM-UP
<i>this paper</i> (CONCAT) <sup>♣</sup>	75.9	64.8	54.1	BOTTOM-UP
<i>this paper</i> (FLAIR-LABEL) <sup>♣</sup>	75.8	65.6	55.3	
<i>this paper</i> (SR-LABEL) <sup>♣</sup>	76.2	66.0	55.3	
<i>this paper</i> (SR-GRAPH) <sup>♣</sup>	75.8	65.5	54.7	
<i>this paper</i> (SR-FT) <sup>◇</sup>	<b>76.3</b>	<b>66.2</b>	<b>55.5</b>	
Human (Morey et al., 2017)	78.7	66.8	57.1	—

Table 5: Joint Training Performance on RST-DT. \* = original paper score. <sup>◇</sup> = 5 run avg.; <sup>♣</sup> = 3 run avg.

headroom for improvement; however, we suspect some of the missing information responsible for errors may relate to global structure and pragmatic understanding which cannot easily be compensated for by adding more genres with potentially disjoint vocabulary. A further surprising result is that CONCAT is not much worse than the base system, suggesting that most of the score comes from obvious cases (e.g. relative and infinitival clauses) which do not differ substantially across corpora or genres, and were already learned without the added data. Similar results have been found for other NLP tasks where adding a second dataset for joint training creates a ‘break-even’ effect: the benefit of more data helps about as much as the disparate domains harm within-corpus performance, e.g. for dependency parsing in Hebrew (Zeldes et al., 2022) and English (Zeldes and Schneider, 2023), English coreference resolution (Zhu et al., 2021), and very recent similar results for RST parsing for Chinese (Peng et al., 2022).

### 3.3 OOD Multi-Genre Degradation

In this section we explore our next question: How badly will a multi-genre trained model degrade on unseen genres, when the annotation scheme remains identical? This question is important for applied, methodological, and ethical reasons. From the practical perspective, while we already know that training only on news leads to severe degradation, we want to inform users of discourse parsing about expected performance on unseen domains if training data is already domain-rich. Methodologically, we want to test the benefit of adding multiple genres and weigh the differences between a few-genre corpus design and a many-genres design, given a fixed total data size capacity for dataset creation. Finally from an ethical perspective, OOD degradation has real life implications for less common types of data, whether they come from un-

derrepresented genres, communicative situations, or speaker demographics (Mengesha et al., 2021). While we cannot fix these problems without more data, we can point them out and increase awareness of skewed data biases at the discourse level.

To explore OOD degradation, we conducted 10 experiments, comparing the normal genre-balanced scenario (GUM test) with testing on each genre when it is not in ‘train’ (ONE-VS-ALL or OVA). For consistency, we test OVA for the 8 roughly equal sized non-growing genres in GUM. Since data for the smaller 4 growing genres may be less reliable and non-comparable, we separately report scores for training on all 8 large genres (ALL-LARGE), tested on each of the four growing genres.

In the last scenario in particular, we would like to see whether small test genres, for which we cannot obtain enough training data, perform better if a training genre exists which may offer a near substitute. For instance, the small *textbook* and *speech* genres have structural organization and formal language similar to *academic* and *news* respectively, and there is some overlap between *speech* vocabulary and *interviews* with politicians. By contrast, *vlog* and *conversation* are highly informal and colloquial, perhaps closest to the *reddit* or *interview* genres, but still likely more challenging. For reproducibility, exact training data compositions are given in Table 17 in Appendix E.

Table 6 shows the results, with the same 4 documents from each held-out genre used to evaluate each OVA model (these are not included in any training set for consistency). The *degradation* column shows that the parser suffers when a genre is removed from training across the board, except for *news* and the Span level of *reddit*; *bio* and *how-to* genres suffer the most: qualitative inspection reveals that this is due to frequent errors on ORGANIZATION and EXPLANATION, which tend to lack explicit cues and require understanding of macrostructure and global topic or pragmatic reasoning to interpret. It is likely that lack of a similar substitute to these genres contributes to the degradation.

By contrast, degradation on *interview* is low (perhaps covered by *conversation* and *news* in training), and the inverse result for *news* suggests that collecting more news data may not be a priority. The minimal degradation for *reddit* is surprising, and is likely due to more instances of explicitly signaled and ‘easy’ relations such as PURPOSE (e.g. *in order to*) and CONTINGENCY (e.g. *if*), but we note

that ordinary scores on *reddit* are low to begin with (second worst after *fiction*, which degrades more). The bottom of the table shows smaller genres suffering, except for a minor gain in *speech* at the Span level. It is unsurprising that *conversation* degrades the most, especially at the R level, since it contains challenging phenomena absent in other genres, such as back-channeling, dysfluencies, and abrupt topic changes. This is mirrored in bad performance on the ORGANIZATION and TOPIC relation classes (see more detailed analysis in §4).

	GUM test			ova			degradation		
	S	N	R	S	N	R	S	N	R
<b>non-growing</b>									
<i>academic</i>	77.0	68.5	59.8	75.2	66.2	55.7	1.7	2.3	4.1
<i>bio</i>	70.4	58.2	51.2	68.8	53.9	43.2	1.6	4.3	<b>8.0</b>
<i>fiction</i>	66.3	53.1	43.7	64.5	50.1	42.1	1.8	3.0	1.7
<i>interview</i>	73.3	59.0	50.9	73.0	56.7	49.7	0.3	2.2	1.2
<i>news</i>	71.7	58.4	49.1	72.2	59.2	51.3	-0.5	-0.8	-2.2
<i>reddit</i>	66.0	52.3	44.2	66.6	51.9	43.3	0.6	0.4	0.8
<i>voyage</i>	78.3	62.1	51.8	77.4	59.7	49.3	0.9	2.4	2.4
<i>how-to</i>	76.5	63.6	54.6	67.1	54.3	44.8	<b>9.3</b>	<b>9.3</b>	<b>9.9</b>
<b>growing</b>	GUM test			ALL-LARGE			degradation		
	S	N	R	S	N	R	S	N	R
<i>conversation</i>	45.4	34.5	26.7	42.7	31.4	21.8	2.7	3.1	<b>4.9</b>
<i>speech</i>	76.0	64.4	55.2	76.4	62.9	54.8	-0.4	1.5	0.4
<i>textbook</i>	77.4	66.8	57.3	76.2	64.3	54.5	1.2	2.6	2.9
<i>vlog</i>	64.8	49.0	42.8	63.3	49.0	40.4	1.5	0.0	2.5

Table 6: Per Genre Scores for GUM test vs. the OVA or ALL-LARGE Experiments (3 run average).

### 3.4 Genre Variety in a Fixed-Size Sample

While §3.3 shows how genres differ in degradation, it falls short of proving that genre diversity promotes generalization when all other things are equal, since train sets for each genre are not identical in size. Our final experiment addresses this: ideally, we want to compare scores on a fixed OOD test set for equal-sized training corpora, divided into fewer or more genres. Although we might expect more genres to be helpful for generalization, this is not trivial or inevitable: If there are not enough recurring examples of infrequent phenomena, because data is so diverse, learning might fail due to sparseness; that is, more genres could be distracting rather than helpful in a meaningful way, which could hurt performance.

Because genres in GUM are small (~2K EDUs) and we want a critical mass of 5K+ EDUs for reasonable parser performance, our combinatory options are limited to 3+ training genres. The parser can only be trained on complete documents (we cannot select  $n$  EDUs from each genre – genres must comprise coherent texts), i.e. we must find permutations of the data which total roughly the same number of training instances, but with dif-

ferent genres. Table 7 gives details on the 3 best training cohorts based on these criteria.

ID	genres	docs	EDUs	ID	genres	docs	EDUs
C1	<i>academic</i>	18	1,970	C3	<i>academic</i>	9	1,004
	<i>bio</i>	19	1,981		<i>bio</i>	9	930
	<i>news</i>	23	1,760		<i>news</i>	10	635
	<b>total</b>	60	5,711				
C2	<i>fiction</i>	15	1,941		<i>fiction</i>	8	1,027
	<i>interview</i>	15	1,931		<i>interview</i>	8	1,199
	<i>how-to</i>	15	1,840		<i>how-to</i>	8	917
	<b>total</b>	45	5,712		<b>total</b>	52	5,712

Table 7: Composition of 3 Fixed-Size Training Cohorts with Different Genre Contents.

As Table 7 shows, the greatest care was taken to ensure that the cohorts sum up to almost exactly as many training instances (~5,712 EDUs), at the price of somewhat diverse amounts of EDUs and documents per genre. If we cannot control all factors, we prefer to constrain the amount of shift-reduce operations learned, in order to prevent any alternative explanation in which total size effects outweigh genre diversity. Additionally, we assume that EDU count per genre will vary in any multi-genre corpus due to complete document constraints; corpus developers are more likely to target a total fixed size budget. If having too many small genres is harmful, we expect cohort 3 (C3) to perform worst; by contrast, if diversity is helpful, C3 should perform best. In either case, we compare how the S/N/R metrics behave in each training regime.

Table 8 shows 5-run averages on each OOD genre, as well as total micro- and across-genres macro-averaged performance. The 3-genre cohorts, C1 and C2, perform similarly overall, though C2 outperforms C1 on test data (which is held constant for all cohorts). R scores are very close on macro-average, but C1 is particularly bad on *conversation*, where C2 is tied with C3 on R (possibly because both have *interview* data). Conversely, C1 is better on textbooks and travel guides (*voyage*) than C2 (possibly because C1 has *academic*, which predicts textbooks reasonably well, and bios – a descriptive informational genre, something missing from C2).

Overall best performance is obtained by C3, with 6 training genres, but the least data in each. Although C3 loses to C2 in the dialogue-oriented *conversation* and *reddit* genres, degradation is very modest and does not affect all metrics (e.g. C3 has a better N score on *reddit*). C3 also outperforms C1 on *academic* and *voyage*, giving the best performance on both. The average gain for C3 across all

test	C1			C2			C3			C3-C1			C3-C2			mean_C3_gain		
	S	N	R	S	N	R	S	N	R	S	N	R	S	N	R	S	N	R
<i>conversation</i>	34.8	23.4	13.9	40.3	27.9	18.0	37.9	26.4	18.0	3.0	3.0	4.1	-2.5	-1.5	0.0	0.3	0.7	2.0
<i>reddit</i>	60.3	45.3	36.0	63.5	46.9	37.6	61.8	47.6	37.3	1.5	2.3	1.4	-1.7	0.7	-0.3	-0.1	1.5	0.6
<i>speech</i>	72.5	58.2	46.9	72.6	59.3	47.7	71.6	57.1	48.0	-0.9	-1.1	1.1	-1.0	-2.1	0.3	-0.9	-1.6	0.7
<i>textbook</i>	73.6	59.0	48.9	70.9	55.0	45.6	74.0	60.5	51.4	0.5	1.5	2.5	3.1	5.5	5.9	1.8	3.5	4.2
<i>vlog</i>	57.8	41.3	35.0	58.8	44.5	35.3	57.7	43.4	34.8	-0.1	2.1	-0.2	-1.1	-1.1	-0.5	-0.6	0.5	-0.3
<i>voyage</i>	76.6	58.1	47.5	76.5	57.4	46.4	78.0	59.1	50.2	1.5	1.0	2.7	1.6	1.7	3.8	1.5	1.4	3.3
macro_avg	62.6	47.6	38.0	<b>63.8</b>	48.5	38.4	63.5	<b>49.0</b>	<b>40.0</b>	0.9	1.5	1.9	-0.3	0.5	1.5	0.3	1.0	1.7
micro_avg	58.7	44.2	34.8	<b>60.5</b>	45.7	35.7	59.8	<b>45.9</b>	<b>36.9</b>	1.1	1.7	2.1	-0.6	0.2	1.2	0.2	1.0	1.6

Table 8: Performance of 3 Fixed-Size Train Cohorts with Different Genre Contents (5 run average).

metrics (3-metric average) is around +1%—taking individual scores from each of five runs, and each genre test set as a data-point, this improvement is significant at  $p < 0.05$ . For individual metrics, we see  $\sim +1.7\%$  on R, +1% on N, and just under +0.3% for S (significant at  $p < 0.05$  except for S).

Although all scores are rather low due to the small corpus sizes (about  $\frac{1}{4}$  of GUM), they suggest that more training genres with smaller portions each promotes OOD generalization, though not by a lot. It is an open question whether this gap would increase or decrease with corpus size: on the one hand, more data would allow for more lexical diversity even with few genres. On the other, it is likely that scores in small data are driven by easy to learn cases (e.g. relative clauses as ELABORATION, or PURPOSE infinitives), which stand to gain less from diversity. If more data means models will tackle more sparse phenomena, then genre diversity should matter *more* for OOD material as the training set grows. To an extent, results in §3.1 showing worse generalization from the large but homogeneous RST-DT to GUM seem to support this hypothesis. We interpret the present results to mean that development of more diverse multi-genre data should take priority over building up material in existing genres to promote generalizable parsing.

#### 4 Error Analysis

Due to space, we limit our error analysis to examining dependency conversions of gold vs. predicted trees, which allow us to break down OOD errors by coarse relation class in Table 9 from all test sets in §3.3. We select models with scores closest to average run scores. It is clear that even in the more modest degradation within GUM, over half of relation classes have  $< 50\%$  accuracy, with the document ROOT being the hardest to identify—i.e. the Central Discourse Unit (CDU)—followed by RESTATEMENT and EVALUATION, which require reasoning over many EDUs and lack consistent overt signals, as opposed to relations such as ATTRIBUTION (marked by speech verbs), PUR-

POSE (marked by purpose infinitives or *in order to*), and CONTINGENCY (usually *if*); interrupted clauses (SAME-UNIT) are easier as well. For CDU identification (Irskieta et al., 2016), which can benefit summarization or long-form QA systems, half of the genres (*academic, fiction, interview, voyage, how-to, vlog*) score 0%; the highest accuracy is only 50% (*bio, news, reddit* and *speech*). More alarmingly, in the cross-corpus setting (§3.1), an RST-DT trained model captures only a single GUM CDU correctly (ACC=0.042 vs. 0.375 for a GUM-trained model); scores on RST-DT are much higher: ACC=0.842 for SR-FT trained on RST-DT vs. 0.553 for a GUM-trained model.

gold coarse relation class	acc	gold coarse relation class	acc
ATTRIBUTION	0.875	CONTEXT	0.471
PURPOSE	0.861	ADVERSATIVE	0.467
SAME-UNIT	0.814	ORGANIZATION	0.463
CONTINGENCY	0.794	EXPLANATION	0.431
ELABORATION	0.666	CAUSAL	0.384
JOINT	0.654	EVALUATION	0.362
TOPIC	0.574	RESTATEMENT	0.308
MODE	0.504	ROOT	0.208

Table 9: OOD Accuracy by Relation Class in GUM.

genre	gold coarse relation class	abs(resid)
<i>textbook</i>	CONTEXT	3.64
<i>speech</i>	EXPLANATION	3.14
<i>reddit</i>	EXPLANATION	3.02
<i>fiction</i>	EVALUATION	2.59
<i>bio</i>	CAUSAL	2.26
<i>vlog</i>	CAUSAL	2.23
<i>conversation</i>	ORGANIZATION	2.14
<i>voyage</i>	CONTEXT	2.13
<i>academic</i>	ORGANIZATION	1.84
<i>how-to</i>	ORGANIZATION	1.62
<i>news</i>	EXPLANATION	1.38
<i>interview</i>	EVALUATION	0.89

Table 10: Max Absolute Error Residuals by Genre.

Errors are also skewed by genre: Table 10 gives the gold label most surprisingly associated with errors per genre, given global error distributions, based on absolute  $\chi^2$  residuals. While EVALUATION is problematic in *fiction* and *interview*, EX-



PLANATION and ORGANIZATION are surprisingly hard to predict in 3 genres each: the former is used e.g. in speeches and news for supporting evidence or justifications which are difficult to identify using lexical items, while the latter is used in conversations for phatic responses and back-channeling.

To further investigate genre-specific phenomena, we analyzed OOD parsing errors for *how-to guides*, which had the worst OOD performance in Table 6, using a model *not* trained on this genre from §3.3. Figure 2 shows a confusion matrix produced by converting the automatic parses into the RST dependency representation following Li et al. (2014). We focus on relation instances with correct attachment predictions, although we also discuss cases involving attachment errors as well (Figure 5h in Appendix G). We first observe in Figure 2 that there are five gold CONTINGENCY instances predicted to be CONTEXT: four out of five EDUs begin with a discourse marker (DM), ‘once’ or ‘until’, e.g.:

- (1) [Don’t use a joke] humans: CONTINGENCY  
parser: CONTEXT [until  
you’re completely comfortable with it.]

According to The Penn Discourse Treebank (PDTB3, Prasad et al. 2019), these prototypically signal spatio-temporal circumstances: 94% of explicit instances with the DM ‘once’ are annotated as TEMPORAL.ASYNCHRONOUS while less than 5% of the instances (4/85) are annotated as CONTINGENCY.CONDITION. Similarly, 85% of explicit instances with the DM ‘until’ are annotated as TEMPORAL.ASYNCHRONOUS while 10% are CONTINGENCY.CONDITION. However, Liu (2019) found that these two DMs are frequently associated with CONTINGENCY.CONDITION in *how-to guides* because they are essentially never narrative and always part of an instruction which is uncommon in *news*. This suggests that although DMs are usually considered useful devices to identify certain relations, their usage differs across genres and is sometimes too ambiguous to form a reliable signal. We also observe that there are four gold RESTATEMENT confused with ELABORATION: three of these go back to ‘definitional’ RESTATEMENTS, which are likely promoted by the genre’s descriptive and explanatory properties.

Figure 5h shows cases with attachment errors. One major error concerning global structure is seven gold ORGANIZATION instances being confused with JOINT. Two of the errors are unexpected given that there are strong graphical, structural, and

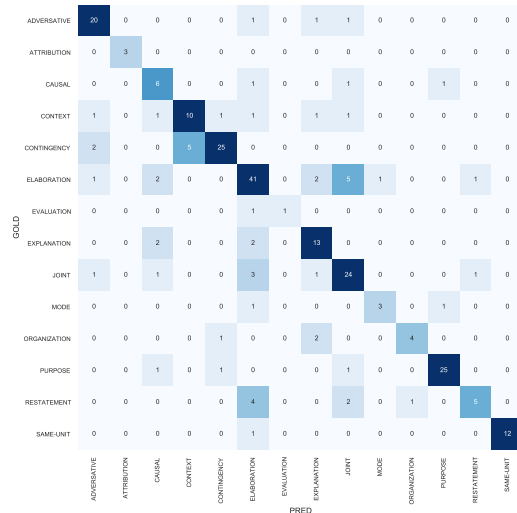


Figure 2: *how-to* confusions (no attachment errors).

semantic cues (see an example in Figure 3 in Appendix F). The remaining cases are harder to pin down: they are all imperatives, which are atypical for headings outside of *how-to guides*, but normal and frequent in instructional texts (see an example in Figure 4). For complete by-genre confusion matrices (with attachment errors) of the multi-genre degradation experiments, see Appendix G.

## 5 Conclusion

The analyses in this paper are meant to inform users of current parsers about what to expect from RST parsing in the wild. Through dozens of experimental runs we have shown a consistent picture: RST parsing has made impressive progress, but OOD degradation is still severe. Our results suggest prioritizing genre diversity in training data is crucial, not only to cover more text types as ‘in domain’, but also to increase performance on unseen text types. Rather than focusing only on the development of better models to beat the single genre RST-DT benchmark, robust RST parsing would be promoted most by creating more annotated data.

We also hope this paper will motivate researchers to prioritize multi-genre benchmarks and OOD settings for RST parsing, and to explore algorithms, representations and features which better capitalize on joint training and foster generalizability, including using data from other sources and theories of discourse analysis (cf. Braud et al. 2016). We see great challenges for tackling errors which implicate complex pragmatic inference or document level reasoning, but we are also optimistic that as more data becomes available, we will be able to learn more and better representations of discourse structure.

## Limitations

One limitation is that the scope of the error analysis presented here is limited, primarily due to space reasons. Although we included full relation confusion matrices (with attachment errors) in Appendix G, the discussion is too limited to describe them in detail, and we encourage interested readers to explore and further compare gold and predicted trees (available in both `.rs3` and `.rsd` formats) on the multi-genre data, which we make available in the repository of this paper.

Another limitation concerns the number of runs of the experiments: some scores were averaged over three rather than five runs. This is due to the fact that SpanBERT-base (Joshi et al., 2020) is fairly large (110M parameters). Training each of our 20 models five times would consume very large amounts of GPU resources, which we feel is hard to justify both financially and environmentally. After completing this study we feel that three-run averaged scores should satisfy the need for reproducibility, though we did use five-run averages for establishing baseline scores and verifying particularly controversial results, such as the improvement on Nuclearity in the fine-tuning condition in §3.2 (i.e. SR-FT). Previous work providing this information reported three-run averaged scores such as Koto et al. (2021) while many papers did not include this information or mention whether the reported scores were averaged over multiple runs.

## Ethics Statement

This work contributes to open source progress in RST discourse parsing, an area which has received less attention than some other NLP tasks, and which, at least in English, is currently suffering from a skewed focus towards the ‘standard’ language of 1990-era Wall Street Journal writing. Previous work has shown that NLP systems retain strong lexical biases mirroring both period and author demographics (Shah et al., 2020), meaning that if language technologies are not pushed to cover diverse data types robustly, they will inevitably perform more poorly on ‘non-standard’ data, with possible discriminatory effects on under-represented populations ranging from the political (think opinion mining social media to guide policy) to financial (e.g. higher/lower search hit rates for YouTube videos by small businesses). By promoting higher quality treatments of diverse language samples in this study covering Reddit, YouTube

vlogs, and demographically diverse conversations, we hope to help level the playing field across text-types, demographics, and domains.

We recognize that NLP research has a computing cost and carbon footprint, which motivates us to release the trained models in this work (preventing the need to retrain similar models), and to avoid extensive hyperparameter optimization which may not generalize to applications in the wild. Specific model configurations such as hyperparameters and validation performance for the reported test results of RST-DT and GUM are detailed in Appendix C.

Finally, we also recognize that NLP tools can be used to do harm. However, we expect that the type of analysis promoted here will do more good than harm by steering tool development away from adhering closely to outdated and narrow-domain data, which this work aims to broaden. Given that discourse parsers already exist, we view the push to reduce topical and authorial bias, as well as the public release of more resources, as net positives.

## References

- Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. [FLAIR: An easy-to-use framework for state-of-the-art NLP](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.
- Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.
- Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. [Where are we in discourse relation recognition?](#) In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.
- Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. [The change that matters in discourse parsing: Estimating the impact of domain shift on parser error](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845, Dublin, Ireland. Association for Computational Linguistics.
- Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.
- Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. [Multi-view and multi-task training of RST discourse parsers](#). In *Proceedings of COLING 2016, the 26th*

- International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.
- 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.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Text, Speech and Language Technology 22, pages 85–112. Kluwer, Dordrecht.
- 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.
- Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. [On the development of the RST Spanish treebank](#). In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.
- Vera Demberg, Merel C.J. Scholman, and Fatiemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.
- Suzanne Eggins and J. Martin. 1997. [Genres and Registers of Discourse](#). In *Discourse as Structure and Process: Discourse Studies: A Multidisciplinary Introduction*. SAGE Publications Ltd.
- Elisa Ferracane, Titan Page, Junyi Jessy Li, and Katrin Erk. 2019. [From news to medical: Cross-domain discourse segmentation](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 22–29, Minneapolis, MN. Association for Computational Linguistics.
- Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. [DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. [AMALGUM – a free, balanced, multilayer English web corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5267–5275, Marseille, France. European Language Resources Association.
- 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.
- Benjamin Heinzerling and Michael Strube. 2019. [Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy. Association for Computational Linguistics.
- Freya Hewett and Manfred Stede. 2022. [Extractive summarisation for German-language data: A text-level approach with discourse features](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 756–765, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Patrick Huber and Giuseppe Carenini. 2020. [MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7442–7457, Online. Association for Computational Linguistics.
- Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: An Online Search Interface to Check Rhetorical Relations. In *The 4th Workshop RST and Discourse Studies*, Fortaleza, Brasil.
- Mikel Iruskieta, Gorka Labaka, and Juliano Desiderato Antonio. 2016. Detecting the central units in two different genres and languages: a preliminary study of Brazilian Portuguese and Basque texts. *Sociedad Española para el Procesamiento del Lenguaje Natural*, pages 65–72.
- Yangfeng Ji and Jacob Eisenstein. 2014. [Representation learning for text-level discourse parsing](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [SpanBERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. [Top-down discourse parsing via sequence labelling](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational*

- Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.
- David. Y. Lee. 2001. Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, 5:37–72.
- Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.
- Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.
- Pia Linscheid, Peter Bourgonje, and Georg Rehm. 2021. Parsing discourse structures for semantic storytelling: Evaluating an efficient RST parser. In *Proceedings of QURATOR 2021 – Conference on Digital Curation Technologies. (QURATOR-2021)*, February 8-12, Berlin, Germany. CEUR Workshop Proceedings.
- Yang Liu. 2019. Beyond the Wall Street Journal: Anchoring and comparing discourse signals across genres. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 72–81, Minneapolis, MN. Association for Computational Linguistics.
- 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.
- Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. 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.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics*, 19(2):313–330.
- Aleksandre Maskharashvili, Symon Stevens-Guille, Xintong Li, and Michael White. 2021. Neural methodus revisited: Do discourse relations help with pre-trained models too? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 12–23, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. “I don’t think these devices are very culturally sensitive.” — impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence*, 4.
- Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.
- Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.
- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. RST parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.
- Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.
- 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.
- Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, A. Nasedkin, S. Nikiforova, I. Pavlova, and A. Shelepov. 2017. Towards building a discourse-annotated corpus of Russian. In *23rd International Conference on Computational Linguistics and Intellectual Technologies. Dialogue*, pages 194–204, Moscow, Russia.
- Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05.

- Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. [Multi-layer discourse annotation of a Dutch text corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).
- Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. [Predictive biases in natural language processing models: A conceptual framework and overview](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.
- Manfred Stede and Arne Neumann. 2014. [Potsdam commentary corpus 2.0: Annotation for discourse research](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Maitte Taboada and Julia Lavid. 2003. [Rhetorical and thematic patterns in scheduling dialogues: A generic characterization](#). *Functions of Language*, 10(2):147–178.
- Shujun Wan, Tino Kutschbach, Anke Lüdeling, and Manfred Stede. 2019. [RST-tace a tool for automatic comparison and evaluation of RST trees](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 88–96, Minneapolis, MN. Association for Computational Linguistics.
- Xinhao Wang, Binod Gyawali, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2019. [Using Rhetorical Structure Theory to assess discourse coherence for non-native spontaneous speech](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 153–162, Minneapolis, MN. Association for Computational Linguistics.
- Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. [A two-stage parsing method for text-level discourse analysis](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.
- Bonnie Webber. 2009. [Genre distinctions for discourse in the Penn TreeBank](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore. Association for Computational Linguistics.
- Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Discourse-aware neural extractive text summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.
- Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. [CoNLL 2016 shared task on multilingual shallow discourse parsing](#). In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.
- An Yang and Sujian Li. 2018. [SciDTB: Discourse dependency TreeBank for scientific abstracts](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.
- Nan Yu, Meishan Zhang, and Guohong Fu. 2018. [Transition-based neural RST parsing with implicit syntax features](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. [RST discourse parsing with second-stage EDU-level pre-training](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.
- Amir Zeldes. 2017. [The GUM Corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.
- Amir Zeldes. 2018. *Multilayer Corpus Studies*. Routledge Advances in Corpus Linguistics 22. Routledge, London.
- Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. [A second wave of UD Hebrew treebanking and cross-domain parsing](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4331–4344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. [The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Amir Zeldes and Nathan Schneider. 2023. [Are UD Treebanks Getting More Consistent? A Report Card for English UD](#). In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, SyntaxFest 2023)*, Washington, DC. Association for Computational Linguistics.
- Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. [Adversarial learning for discourse rhetorical structure parsing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3946–3957, Online. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. [OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

## A Cross-Corpus Segmentation Scores

The scores in Table 11 (3 run average) show fairly modest degradation on cross-corpus EDU segmentation scores in both directions (a little worse for RST-DT→GUM), using DisCoDisCo (Gessler et al., 2021), the winning EDU segmentation system from the 2021 DISRPT shared task (Zeldes et al., 2021).

train	test	P	R	F1
RST-DT	RST-DT	95.16	94.64	94.90
RST-DT	GUM V8	90.09	89.69	89.89
GUM V8	GUM V8	92.09	93.73	92.90
GUM V8	RST-DT	92.55	89.57	91.03

Table 11: Cross-Corpus Segmentation Performance.

## B Breakdown of the GUM RST Data

The breakdown of genres in GUM V8 is shown in Table 12. Four of the genres in the corpus are still growing, and therefore include less material than other genres at present.

genres	docs	tokens	EDUs	growing
<i>academic</i>	18	17,168	1,969	
<i>bio</i>	20	18,209	2,066	
<i>fiction</i>	19	17,508	2,458	
<i>how-to</i>	19	17,085	2,367	
<i>interview</i>	19	18,189	2,404	
<i>news</i>	23	16,140	1,760	
<i>reddit</i>	18	16,364	2,231	
<i>travel</i>	18	16,513	1,785	
<i>conversation</i>	9	10,451	1,878	✓
<i>speech</i>	10	10,827	1,249	✓
<i>textbook</i>	10	11,190	1,397	✓
<i>vlog</i>	10	11,200	1,543	✓
<b>total</b>	<b>193</b>	<b>180,844</b>	<b>23,107</b>	

Table 12: Overview of RST Data by Genre in GUM V8.

## C Model Configurations and Training Details

To ensure reproducibility and following the reproducibility criteria and checklist provided by EACL 2023, we provide model configurations and training details relevant to the experimental results presented in this paper.

For the BOTTOM-UP parser, the configurations of SpanBERT-NoCoref from Guz and Carenini (2020) was used as our base system, and we followed the hyperparameters and training settings therein across the board except the batch size due to memory limitations and embedding dimensions for organizational features.<sup>7</sup> Specifically, AdamW was used as the optimizer with a learning rate of  $1e^{-5}$  for SpanBERT-base (Joshi et al., 2020) and  $2e^{-4}$  for model-specific components (Guz and Carenini, 2020). Batch size was set to 1 (as opposed to 5 in the original implementation) and there were 20 epochs for each run. The organizational features used in Guz and Carenini (2020) followed Wang et al. (2017) and were represented as binary features in a learnable 5-dimensional embedding (as opposed to 10 in the original implementation). A vector of zeros of the same shape was used when a given feature is unavailable for the aforementioned organizational features as well as the categorical features experimented in §3.2, which all had an embedding dimension of 10.

For the TOP-DOWN parser, overall we followed the original hyperparameters and training settings therein; however, due to memory limitations, we modified the batch size to 3 from 12. The same dev set of respective corpora was used during training as in the training of the BOTTOM-UP parser from Guz and Carenini (2020). Additionally, XLM-RoBERTa-base (Conneau et al., 2020) was used as the language backbone.

All the training sessions were conducted using 1 NVIDIA Tesla T4 GPU on Google Cloud Platform. Table 13 shows the corresponding validation performance on RST-DT test and GUM V8 test reported in Tables 3 and 5 respectively using the BOTTOM-UP parser. Table 14 presents the corresponding validation performance (averaged over 5 runs) on RST-DT test and GUM V8 test reported in Table 4.

<sup>7</sup>Based on our reproducible results of this model (see Table 2) and comparing them to the reported results in the original paper, we believe that the change of these two configurations do not impact model performance significantly.

	S	N	R	# runs
RST-DT BASELINE	76.0	64.9	55.2	5
GUM test	67.0	54.5	45.9	5
RST-DT CONCAT	76.4	65.7	56.0	3
RST-DT FLAIR-LABEL	76.3	65.1	54.9	3
RST-DT SR-LABEL	76.3	64.7	55.2	3
RST-DT SR-GRAPH	76.6	65.5	55.5	3
RS-TDT SR-FT	76.3	64.9	55.4	5

Table 13: Validation Performance on RST-DT test and GUM test Results in Tables 3 and 5 (i.e. BOTTOM-UP).

	S	N	R	# runs
RST-DT BASELINE	75.3	65.0	55.9	5
GUM test	71.3	58.6	50.1	5

Table 14: Validation Performance on RST-DT test and GUM test Results in Table 4 (i.e. TOP-DOWN).

Note that since RST-DT does not have an established dev set, 10% of training data stratified by the number of EDUs in each document is used as the dev set following Guz and Carenini (2020) and is held constant across all the RST-DT-related experiments in this work. The list of the document names used in the dev set is provided in the code repository.<sup>8</sup> Additionally, since the test documents used in all the experiments in §3.3 and §3.4 are OOD data, the corresponding validation performance is not applicable and thus not reported. Information on the average runtime of each epoch for each model is provided in Table 15 below.

RST-DT			
model	avg. runtime	model	avg. runtime
BASELINE	2 hours 53 mins	SR-LABEL	2 hours 53 mins
CONCAT	5 hours 52 mins	SR-GRAPH	2 hours 50 mins
FLAIR-LABEL	2 hours 55 mins	SR-FT	2 hours 56 mins
GUM			
model	avg. runtime	model	avg. runtime
GUM test	3 hours	No Reddit	2 hours 43 mins
ALL-LARGE	2 hours 15 mins	No Voyage	2 hours 47 mins
No Academic	2 hours 43 mins	No How-to	2 hours 45 mins
No Bio	2 hours 45 mins	C1	56 mins
No Fiction	2 hours 44 mins	C2	56 mins
No Interview	2 hours 40 mins	C3	54 mins
No News	2 hours 47 mins		

Table 15: Average Runtime of Every Training Session’s Epoch for Each Model using the BOTTOM-UP Parser.

<sup>8</sup><https://github.com/janetlaueyung/crossGENRE4RST>

## D Relation Mapping

In the interest of reproducibility, Table 16 gives the exact relation mapping used for cross-corpus experiments in which relation labels were targeted.

GUM V8 Relations	GUM V8 Classes	Corresponding RST-DT Classes
adversative-antithesis	Adversative	Contrast
adversative-concession	Adversative	Contrast
adversative-contrast	Adversative	Contrast
attribution-positive	Attribution	Attribution
attribution-negative	Attribution	Attribution
causal-cause	Causal	Cause
causal-result	Causal	Cause
context-background	Context	Background
context-circumstance	Context	Background
contingency-condition	Contingency	Condition
elaboration-attribute	Elaboration	Elaboration
elaboration-additional	Elaboration	Elaboration
explanation-evidence	Explanation	Explanation
explanation-justify	Explanation	Explanation
explanation-motivation	Explanation	Explanation
evaluation-comment	Evaluation	Evaluation
joint-disjunction	Joint	Joint
joint-list	Joint	Joint
joint-sequence	Joint	Temporal
joint-other	Joint	Topic-Change
mode-manner	Mode	Manner-Means
mode-means	Mode	Manner-Means
organization-heading	Organization	Textual-Organization
organization-phatic	Organization	Topic-Comment
organization-preparation	Organization	Textual-Organization
purpose-attribute	Purpose	Elaboration
purpose-goal	Purpose	Enablement
restatement-partial	Restatement	Summary
restatement-repetition	Restatement	Summary
topic-question	Topic	Topic-Comment
topic-solutionhood	Topic	Topic-Comment
same-unit	same-unit	Same-Unit

Table 16: Relation Mapping of GUM V8 to RST-DT.

## E Data Description for the GUM OOD Multi-Genre Experiments

Table 17 details the number of genres, documents, and EDUs used for training the models in the ten experiments in §3.3. For maximum reliability, testing was always conducted on the held-out genres’ standard dev+test partitions specified in the official GUM V8 release (which is also provided in the repository of this paper), since the dev set for the targeted genres cannot be used for early stopping to simulate real OOD data in the wild.

## F An Example OOD Parser Output

Figure 3 provides a comparison of a fragment of parser output versus the corresponding gold annotation to exemplify the distinction between gold ORGANIZATION and predicted JOINT discussed

models	genres	docs	EDUs
No Academic	11	131	16,088
No Bio	11	129	15,901
No Fiction	11	130	15,640
No Interview	11	130	15,599
No News	11	126	16,252
No Reddit	11	131	15,892
No Voyage	11	131	16,133
No How-to	11	130	15,672
ALL-LARGE	8	122	13,703
GUM test	12	145	17,610

Table 17: Training Data Composition used in §3.3.

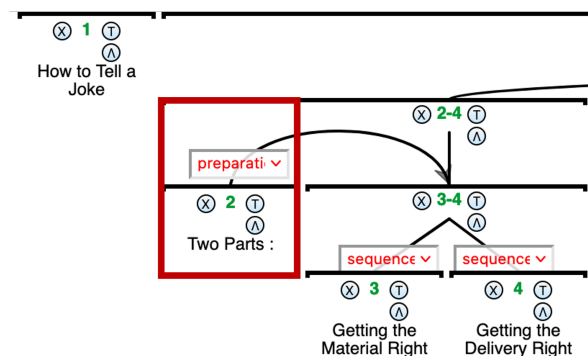
in Section 4. Additionally, Figure 4 provides a fragment of parser output from one of the OVA models, No How-to, which exemplifies some of the observations and discussions brought up in the error analysis such as the most erroneous relation class ORGANIZATION in how-to guides (EDU 107). Moreover, these figures also demonstrate that although discourse markers (DMs) are cues in many cases, they can lead to errors if they are ambiguous, by distracting the parser from other (non-DM) signals such as the misidentified ELABORATION of EDU 110, where the gold ORGANIZATION is signaled by the colon and the numerical matching of ‘Two parts’ to the two nucleus EDUs, 3 and 4 in Figure 3; and the CAUSAL of EDUs 112-116 in Figure 4, where an intensifying ‘so’ in 112 (‘so comfortable’) may have misled the parser into a resultative ‘so’ reading.

## G Confusion Matrices

Figures 5a–5h show confusion matrices for all the non-growing genres from their corresponding OVA models, and Figures 5i–5l are for all the growing genres based on the automatic parses from the ALL-LARGE model where the training data contains only the 8 non-growing genres. Note that all the matrices were produced by converting the automatic parses into the RST dependency representation following Li et al. (2014) and do not reflect attachment errors. The conversion code is available at <https://github.com/amir-zeldes/rst2dep>.



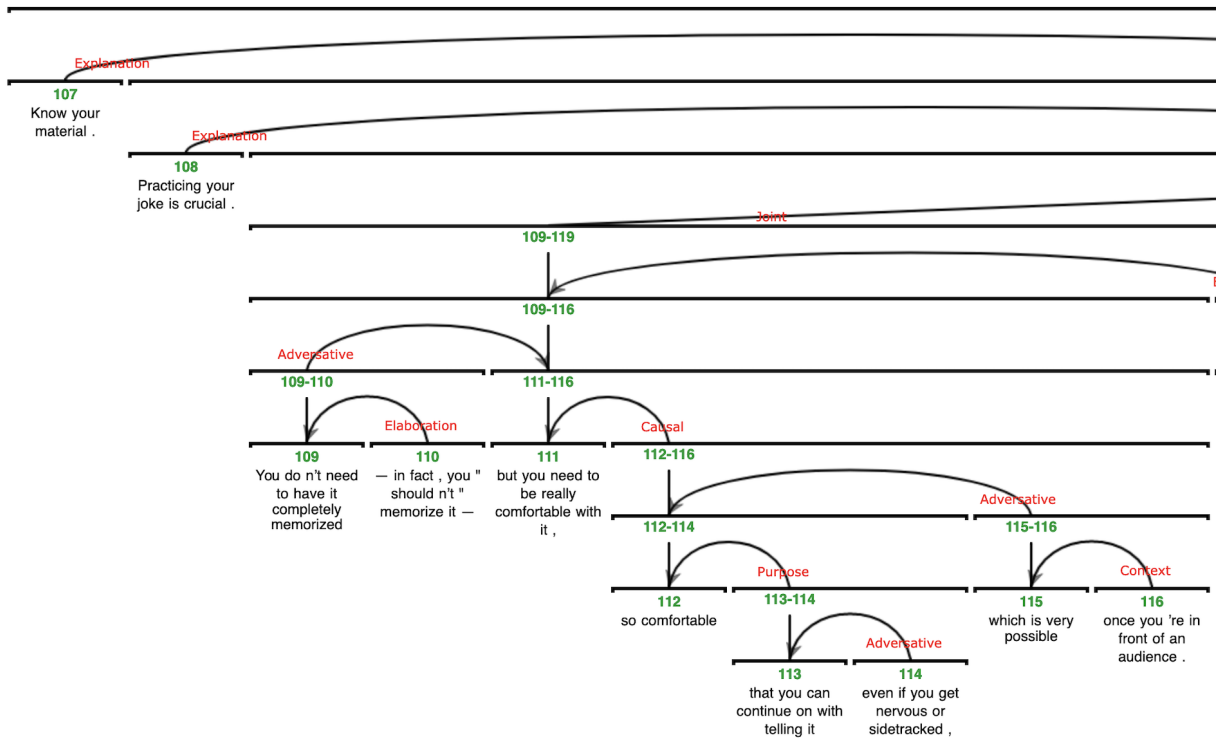
(a) Parser Output Fragment from the *No How-to* Model.



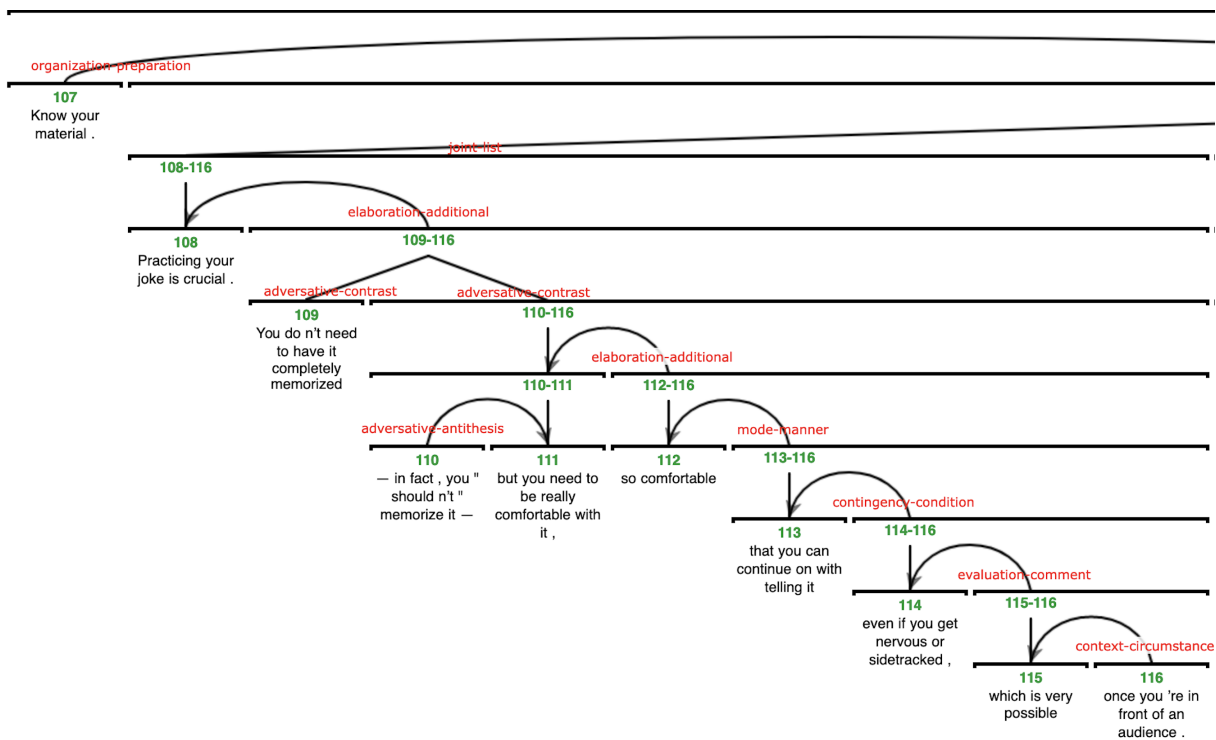
(b) The Corresponding Gold RST Annotation.

Figure 3: An Example of an OOD Parser Output vs. the Corresponding Gold Annotations from *how-to guides*: ORGANIZATION vs. JOINT.



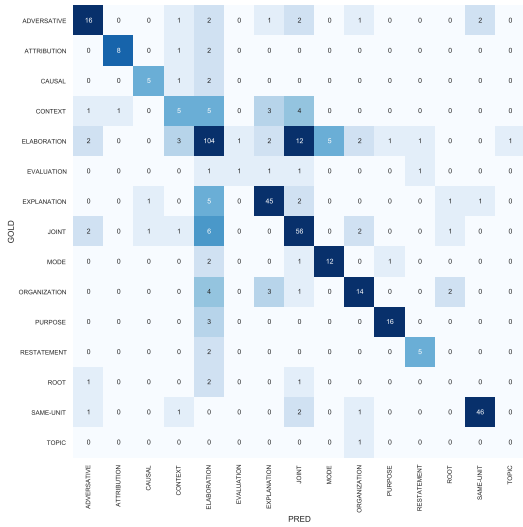


(a) Parser Output Fragment from the OVA (No How-to) Model in §3.3.

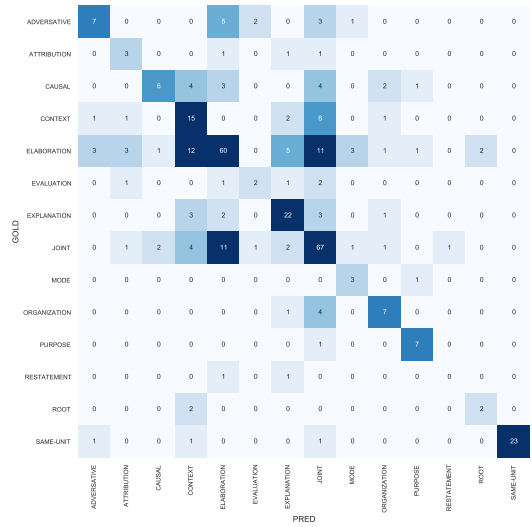


(b) The Corresponding Gold RST Annotation.

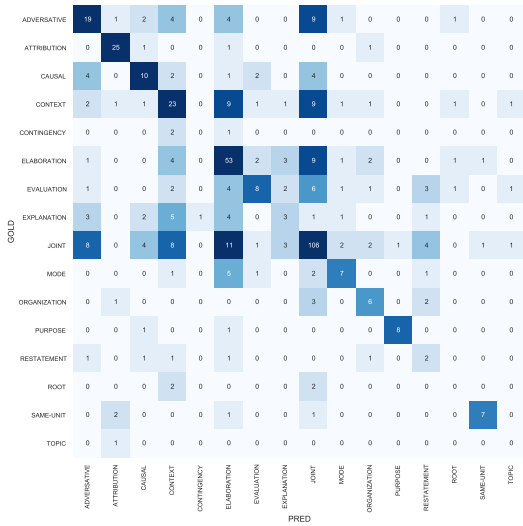
Figure 4: An Example of an OOD Parser Output vs. the Corresponding Gold Annotations from *how-to guides*. This fragment is selected from the document GUM\_whow\_joke, which gives advice on how to tell jokes.



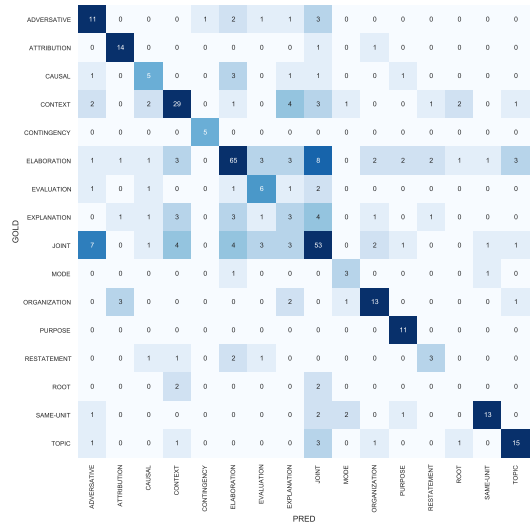
(a) academic



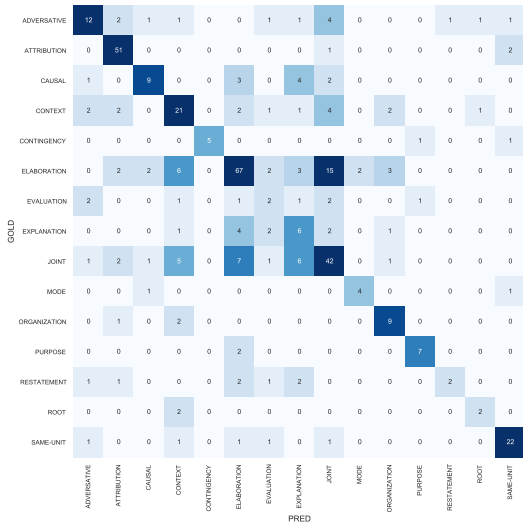
(b) bio



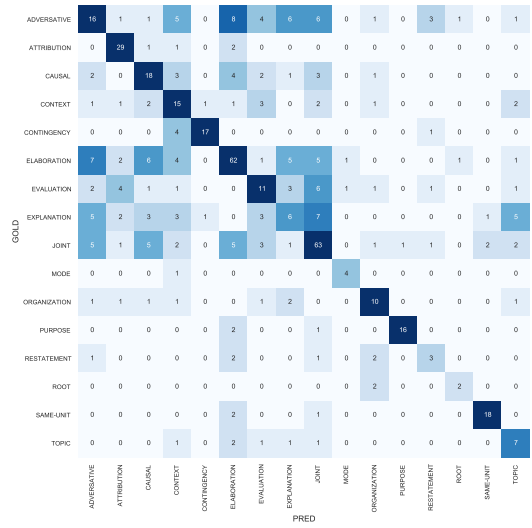
(c) fiction



(d) interview



(e) news



(f) reddit

Figure 5: Confusion Matrices for All Genres from their OVA Models or the ALL-LARGE Model from §3.3.

