# 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

GEORGETOWN UNIVERSITY

The Graduate School of Arts & Sciences

Corpling@GU

# Overview

- **Goal #1**: to demonstrate the generalizability limitations of English RST parsing based on RST-DT and quantify the degradation
- **Goal #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 robust discourse parsing requires

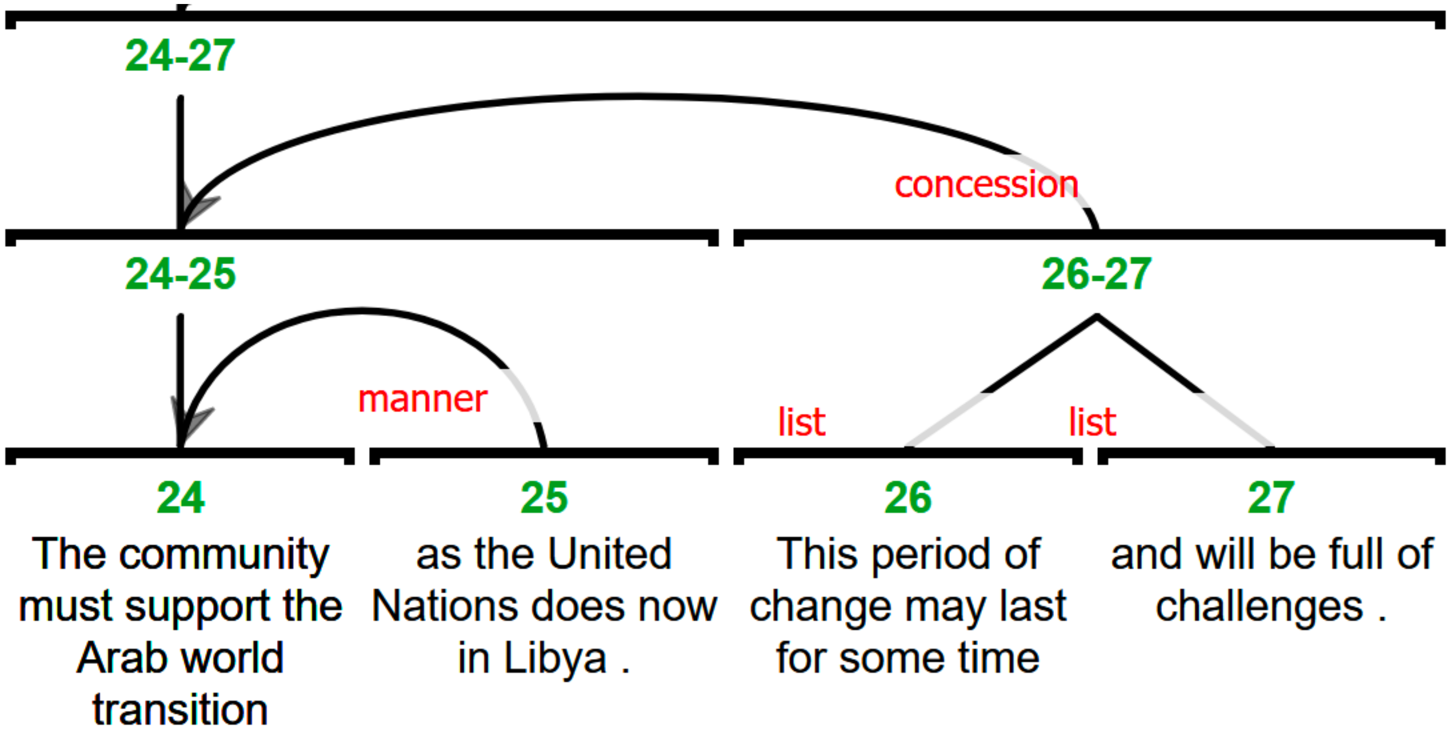
# Overview

- **Takeaway #1**: Diverse training data leads to better generalization on unseen genres regardless of model architecture

- **Takeaway #2**: RST parsing work should devote more attention to multi-genre corpora as benchmarks

# Rhetorical Structure Theory (RST)

- Mann and Thompson (1989)



# English RST Corpora

**RST Discourse Treebank (RST-DT, Carlson et al. 2003)**

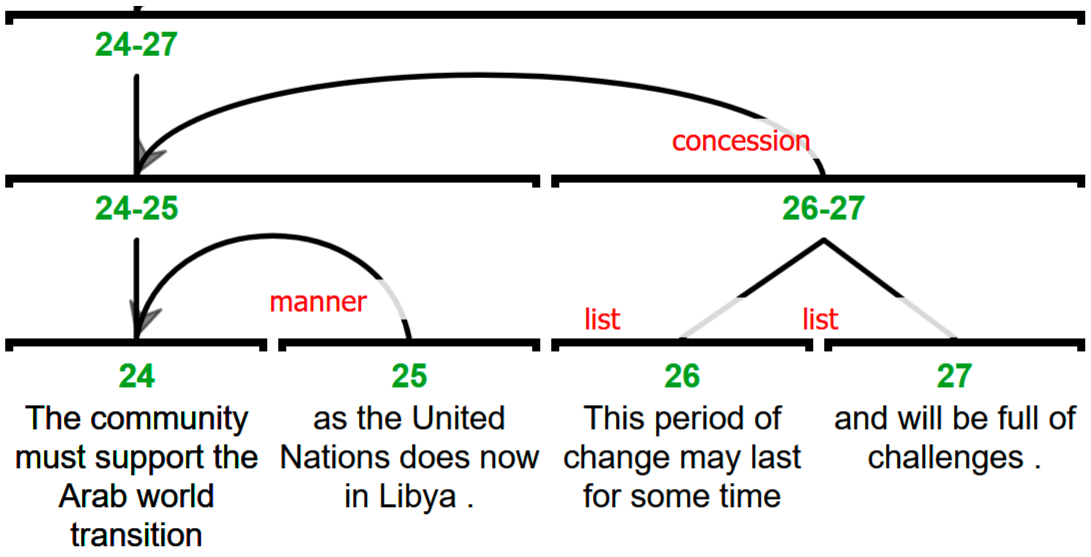
- the standard English RST benchmark, with data from the 1989 Wall Street Journal (WSJ) section of the Penn Treebank (PTB, Marcus et al. 1993)

**GUM (Zeldes, 2017)**

- a multi-genre corpus covering 12 written and spoken genres
- continuously growing, with new data added in each version; for this paper: GUM v8

# Evaluation Metrics

- **S**pan: whether subtrees span the right EDUs
- **N**uclearity: whether edges point the right way
- **R**elation: whether labels are correct



# Experiments

1. Cross-Corpus Generalization (RST-DT & GUM v8)
2. Joint Training (RST-DT)
3. OOD Multi-Genre Degradation (GUM v8)
4. Genre Variety in a Fixed-Size Sample (GUM v8)

# Cross-Corpus Generalization

- Hypothesis: since GUM contains many genres, models trained on it will degrade less when testing on RST-DT than in the opposite scenario

- Parser 1: Guz and Carenini (2020, BOTTOM-UP)
- Parser 2: Liu et al. (2021, TOP-DOWN)

- Setup: train the parsers on the TRAIN partition of each dataset and report scores on the TEST set

# Results

| 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 *news* | 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 *news* | 71.6 | 58.5 | 49.5 |

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

| 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 *news* | 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 *news* | 73.4 | 63.3 | 57.2 |

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

# Discussion

- We interpret this result to mean that **genre composition of the train and test data plays a crucial role in the generalizability of RST constituent parsing, *regardless of parser architecture***.

- 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.

# Experiments

1. Cross-Corpus Generalization (RST-DT & GUM v8)
2. Joint Training (RST-DT)
3. OOD Multi-Genre Degradation (GUM v8)
4. Genre Variety in a Fixed-Size Sample (GUM v8)

# Joint Training

- approach 1: naïve concatenation
- approach 2: model stacking (3 variants)
- approach 3: pretraining

- evaluate on the RST-DT benchmark

# Results

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

Table 5: Joint Training Performance on RST-DT. * = original paper score. ◇ = 5 run avg.; ♠ = 3 run avg.

# Discussion

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

# Experiments

1. Cross-Corpus Generalization (RST-DT & GUM v8)
2. Joint Training (RST-DT)
3. OOD Multi-Genre Degradation (GUM v8)
4. Genre Variety in a Fixed-Size Sample (GUM v8)

# OOD Multi-Genre Degradation

- 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, OVA)

- 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
  - conversation, speech, textbook, vlog

# Results

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

# Experiments

1. Cross-Corpus Generalization (RST-DT & GUM v8)
2. Joint Training (RST-DT)
3. OOD Multi-Genre Degradation (GUM v8)
4. Genre Variety in a Fixed-Size Sample (GUM v8)

# Genre Variety in a Fixed-Size Sample

- Hypothesis: 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.

# Data Composition

- Hypotheses:
  - 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.

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

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

# Results

| | C1 | | | C2 | | | C3 | | | C3−C1 | | | C3−C2 | | | mean_C3_gain | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **test** | **S** | **N** | **R** | **S** | **N** | **R** | **S** | **N** | **R** | **S** | **N** | **R** | **S** | **N** | **R** | **S** | **N** | **R** |
| *conversation* | 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 |
| *reddit* | 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 |
| *speech* | 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 |
| *textbook* | 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 |
| *vlog* | 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 |
| *voyage* | 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 | **63.8** | 48.5 | 38.4 | 63.5 | **49.0** | **40.0** | 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 | **60.5** | 45.7 | 35.7 | 59.8 | **45.9** | **36.9** | 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).

# Discussion

- Although all scores are rather low due to the small corpus sizes (about ¼ of GUM), they suggest that more training genres with smaller portions each promotes OOD generalization, though not by a lot.

# Discussion

- 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; Purpose infinitives

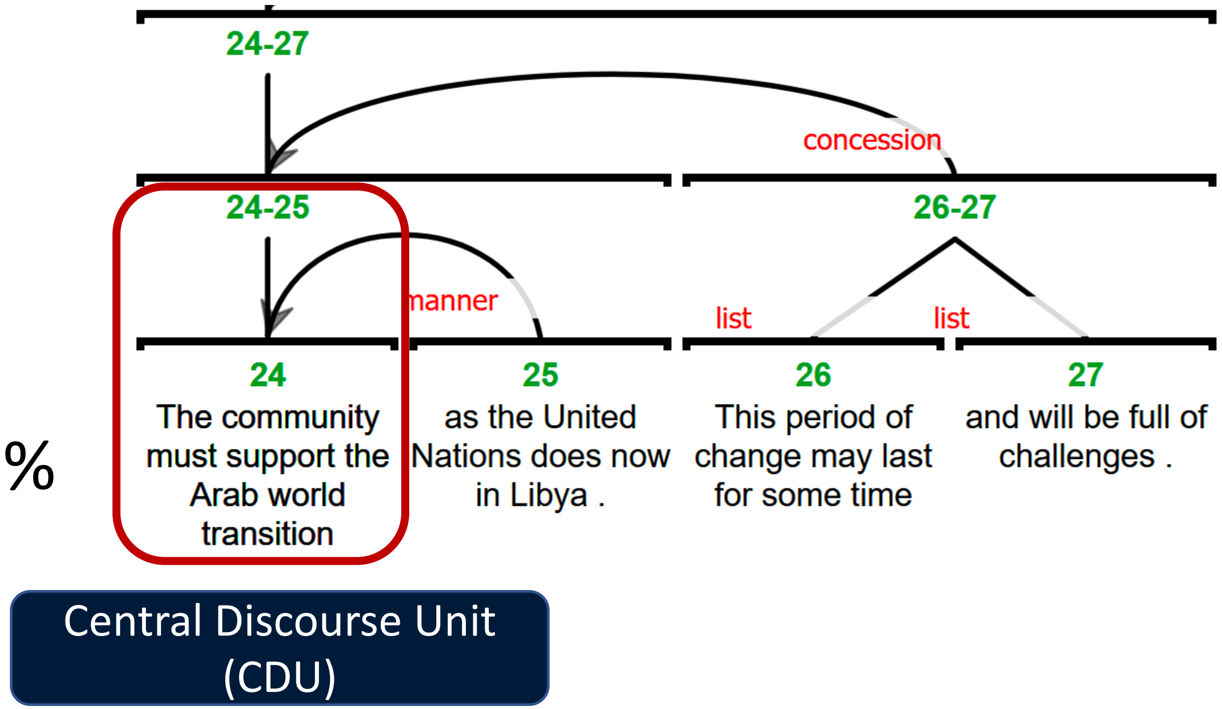
# Discussion

- 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, the results in the cross-corpus generalization experiment showing worse generalization from the large but homogeneous RST-DT to GUM seem to support this hypothesis.

# CDU: OOD Multi-Genre Degradation

- half of the genres **score 0%**
  - *academic, fiction, interview, voyage, how-to, vlog*
- the highest accuracy is only 50%
  - *bio, news, reddit and speech*



# CDU: Cross-Corpus Experiment

- More alarmingly, in the cross-corpus setting, 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

# Takeaways

- Through dozens of experimental runs, we have shown a consistent picture: RST parsing has made impressive progress, but OOD degradation is still severe, regardless of model architecture.

- 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.

# Takeaways

- We want to motivate researchers to **prioritize multi-genre benchmarks and OOD settings for RST parsing**

GEORGETOWN UNIVERSITY

# THANK YOU


yl879@georgetown.edu

amir.zeldes@georgetown.edu


https://github.com/janetlauyeung/crossGENRE4RST