

CLaC at DISRPT 2025: Hierarchical Adapters for Cross-Framework Multi-lingual Discourse Relation Classification

Nawar Turk, Daniele Comitogianni, Leila Kosseim

Computational Linguistics at Concordia (CLaC) Lab

Dept. of Computer Science and Software Engineering

Concordia University, Montréal, Québec, Canada

nawar.turk@mail.concordia.ca, danielecomitogianni@outlook.com, leila.kosseim@concordia.ca

Abstract

We present our submission to Task 3 (Discourse Relation Classification) of the DISRPT 2025 shared task. Task 3 introduces a unified set of 17 discourse relation labels across 39 corpora in 16 languages and six discourse frameworks, posing significant multilingual and cross-formalism challenges. We first benchmark the task by fine-tuning multilingual BERT-based models (*mBERT*, *XLM-RoBERTa-Base*, and *XLM-RoBERTa-Large*) with two argument-ordering strategies and progressive unfreezing ratios to establish strong baselines. We then evaluate prompt-based large language models (namely Claude Opus 4.0) in zero-shot and few-shot settings to understand how LLMs respond to the newly proposed unified labels. Finally, we introduce HiDAC, a Hierarchical Dual-Adapter Contrastive learning model. Results show that while larger transformer models achieve higher accuracy, the improvements are modest, and that unfreezing the top 75% of encoder layers yields performance comparable to full fine-tuning while training far fewer parameters. Prompt-based models lag significantly behind fine-tuned transformers, and HiDAC achieves the highest overall accuracy (67.5%) while remaining more parameter-efficient than full fine-tuning.

1 Introduction

The 2025 DISRPT shared task addresses three challenges in discourse parsing¹: discourse unit segmentation (Task 1), connective detection (Task 2), and discourse relation classification (Task 3) which aims at identifying logical and rhetorical relationships between text segments. This paper presents our approach to Task 3 which expands the task from previous years to 39 corpora across 16 languages and six discourse frameworks.

¹<https://sites.google.com/view/d isrpt2025>

This year, the task proposed a unified set of 17 labels to classify the relations across multiple languages and frameworks, making the classification problem more challenging due to the diversity of data and the need to generalize across different frameworks and languages. This work contributes to the shared effort of building a multilingual and cross-framework discourse parser.

For our participation, we evaluated three methods: transformer-based baselines, prompt engineering with LLMs, and a custom made model HiDAC Hierarchical Dual-Adapter Contrastive learning, a novel parameter-efficient framework which employs a hierarchical adapter backbone and is trained with a dual-loss objective. Results show that HiDAC achieved slightly higher overall accuracy than the transformer baselines with considerably fewer training parameters, while the prompting approaches did not perform well.

2 Related Work

Recent work has explored cross-framework and multilingual discourse relation classification. Costa et al. (2023) proposed mappings between RST-DT and PDTB 3.0 frameworks, while Costa and Kosseim (2025a,b) developed multilabel hierarchical models for multilingual implicit discourse relation recognition. Recently, Eichin et al. (2025) conducted a comprehensive analysis evaluating open-weight LLMs, in which they defined a unified set of discourse relation labels to better understand how these models generalize across languages and annotation frameworks. The study demonstrated that LLMs, especially those with multilingual training corpora, can generalize discourse information across languages and frameworks. Their error analysis highlighted overlapping relations, such as frequent confusions between *Elaboration*, *Framing*, and *Explanation*.

The DISRPT Task 3 itself was first intro-

duced in 2021, where two teams submitted systems: DisCoDisCo (Gessler et al., 2021) achieved 61.82%² accuracy using language-specific BERT-base models (varying per language) fine-tuned per corpus with additional hand-crafted features, while DiscRel (Varachkina and Pannach, 2021) reached 54.23%² accuracy using multilingual Sentence-BERT to embed discourse units, then fed the Euclidean distance between the unit embeddings and directionality as features into a two-level stacked Random Forest, first predicting five coarse classes, then fine-grained labels within each class.

The task was proposed again in 2023 where three teams submitted systems: MELODI (Metheniti et al., 2023) used three mBERT-based models: a baseline fine-tuned model and two adapter-enhanced variants with layer freezing, achieving 54.44%³ accuracy. HITS (Liu et al., 2023) used BERT-based encoders with adversarial training, training separate models for large corpora and a joint multilingual model for smaller corpora and achieved an accuracy of 62.36%³. On the other hand, DiscoFlan (Anuranjana, 2023) used multilingual Flan-T5 based seq2seq model with instruction fine-tuning to generate discourse relation labels from input prompts, achieving 31.21%³ accuracy.

3 Methodology

To address Task 3 (Discourse Relation Classification), we experimented with three families of models: (1) BERT-based models as baselines (see Section 3.1.1); (2) prompt-based generative models to use as a baselines as well (see Section 3.1.2), and (3) a custom model, HiDAC, based on adapters and contrastive learning (see Section 3.2).

3.1 Baseline Models

3.1.1 BERT-Based Models

Our first baseline models are based on multilingual BERT transformers obtained from Hugging Face⁴: *bert-base-multilingual-cased*, *xlm-roberta-base*, and *xlm-roberta-large*. We conducted two sets of experiments: (1) we tested both natural argument ordering as well as relation-directed ordering with no freezing of the BERT models; the relation ordering reorders the argument so as to respect the direction of the relation. For example, given the

²<https://sites.google.com/georgetown.edu/disrpt2021/results>

³<https://sites.google.com/view/disrpt2023/results>

⁴<https://huggingface.co/>

following instance with label *purpose* and direction annotated as $1 < 2$:

Arg1: We propose a neural network approach

Arg2: to benefit from the non-linearity of corpus-wide statistics for part-of-speech (POS) tagging.

In this case, the natural ordering is the way the arguments appear in the text, whereas the relation-directed ordering is obtained by reversing the arguments to respect the direction of the relation $1 < 2$:

Arg1: to benefit from the non-linearity of corpus-wide statistics for part-of-speech (POS) tagging.

Arg2: We propose a neural network approach

(2) using the original order only, we explored gradual unfreezing strategies, where we initially froze all layers and then progressively unfroze 25%, 50%, and 75% of the encoder top layers.

This multilingual encoder approach was tested because previous BERT-based work mainly fine-tuned single-language BERT-based models or combined small languages, with, to our knowledge, across both the 2021 and 2023 versions of the shared task (Zeldes et al., 2021; Braud et al., 2023), only one team, MELODI (Metheniti et al., 2023), attempted a fully multilingual model. With DISRPT 2025’s expanded data and languages, reevaluating a unified multilingual BERT-based approach is now feasible and promising for cross-lingual knowledge transfer.

3.1.2 Prompt Engineering

As a second family of baselines, we evaluated both zero-shot and few-shot prompting strategies using Claude Opus 4.0. We chose the Claude model because in a previous work (Turk et al., 2025), this LLM achieved the highest average macroF1 compared to GPT and Gemini models that were evaluated to identify PDTB 3.0 Level 2 discourse relations.

Due to cost constraints of prompting the full development set, we used stratified sampling to ensure a representative evaluation. We first divided the validation set into 27 equal stratified subsets using StratifiedKFold from scikit-learn⁵ which preserves label distribution, we then randomly selected 4 groups ($\approx 1k$ samples each). Each selected subset was evaluated under both zero-shot and few-shot prompting settings.

In zero-shot prompting, we provided a structured prompt template (available in our repository⁶) without examples and tested both natural-order

⁵<https://scikit-learn.org>

⁶<https://github.com/CLaC-Lab/DISRPT-2025>

and relation-directed order argument arrangements. The template includes DISRPT 2025’s unified set of 17 discourse relation labels in the label list.

For few-shot prompting, we used the same structured template but included examples. We developed a balanced pool of $\approx 1\text{k}$ examples, stratified across framework, language, and label (3 examples for each of the 327 unique framework-language-label combinations). We then ran three few-shot prompting experiments: (1) Exp 1 used 4 examples in total (not per label), sampled randomly from the balanced example pool in the same language as the input instance (note that the task instructions remained in English). (2) Exp 2 used 4 English only examples from the same pool regardless of the instance language; and (3) Exp 3 analyzed Exp 1 & 2 results to identify six labels with the lowest F1 scores, then used 6 English examples from these low-performing labels plus 2 random English examples from the remaining pool, doubling the number of examples.

3.2 The HiDAC Model

Our third method which we call Hierarchical Dual-Adapter Contrastive (HiDAC) is based on a parameter-efficient fine-tuning approach. HiDAC is based on a pre-trained language model (PLM) with two main additions: (1) a hierarchical adapter backbone varying adapter types based on layer depth, and (2) a dual-loss training objective that applies two different losses at two different points in the network. An intermediate-layer contrastive loss is used to train the lower layers to build better representations, while a final-layer cross-entropy loss is used to train the upper layers for the classification task.

3.2.1 Model Architecture

The overall architecture of HiDAC is shown in Figure 1. The model processes each discourse argument independently through a dual-encoder PLM backbone (*XLM-RoBERTa-Large*). The architecture is named ‘hierarchical’ because it partitions the encoder layers into two distinct functional levels, each with a different adaptation method and training objective. The lower layers (1-8) are trained using a Contrastive loss, which operates on the output representations from layer 8. The goal of this objective is to learn foundational representations by explicitly pulling the embedding of an instance closer to its correct class prototype, while simultaneously pushing it away from the prototypes

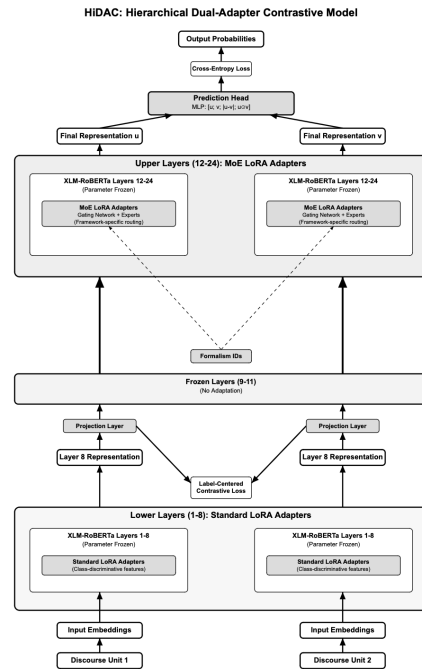


Figure 1: Overview of the HiDAC architecture. The contrastive loss is applied to an intermediate layer (8) to structure the embedding space, while the cross-entropy loss is applied to the representation at the last layer for the final classification.

of all incorrect classes. The upper layers (12-24) then receive these structured representations and are trained by a separate Cross-Entropy loss on the final-layer outputs to perform the classification task. The entire model is trained by minimizing a weighted sum of these two complementary losses.

3.2.2 Discourse Unit Representation and Enhanced Prediction Head

As shown in Figure 1, HiDAC uses a dual-encoder framework. From the final layer’s hidden states, we extract the [CLS] token representation of each discourse unit, Arg1 and Arg2, resulting in vectors $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$, respectively, where $d = 1024$, the model size of *XLM-RoBERTa-Large*.

To capture relational information for classification, we create an enhanced feature vector, \mathbf{z} , by concatenating \mathbf{u} and \mathbf{v} along with their element-wise difference and product (Conneau et al., 2017). We opted for this richer feature representation after an ablation study showed it outperformed the simpler concatenation of $[\mathbf{u}; \mathbf{v}]$, which achieved a development set accuracy of 67.18% and an F1-score of 64.64%. As this was lower than the 67.46% accuracy and 65.14% F1-score of the enhanced method (see Section 5), we chose the latter for our

final model. The enhanced vector \mathbf{z} serves as the input to the multi-layer perceptron (MLP) classifier and is defined as $\mathbf{z} = [\mathbf{u}, \mathbf{v}, \mathbf{u} - \mathbf{v}, \mathbf{u} * \mathbf{v}]$.

3.2.3 Hierarchical Adapter Architecture

HiDAC uses a hierarchical Parameter-Efficient Fine-Tuning (PEFT) strategy, segmenting the backbone PLM into lower layers (1 to 8) and upper layers (12 to 24), each employing distinct adapter mechanisms.

Lower Layers: Layers (1-8) aim to learn foundational representations using standard LoRA adapters (Hu et al., 2022). The purpose of these adapters is to learn basic patterns that help separate the different discourse relation classes. They achieve this by adjusting the model’s early representations, grouping instances with the same discourse relation together in the embedding space. These adapters, optimized exclusively by the contrastive loss, add minimal computational overhead and encourage early formation of a well-organized embedding space.

Upper Layers: Layers (12-24) aim to learn task-specific representations. The upper layers use Mixture-of-Experts LoRA adapters (MoE-LoRA), enabling dynamic, formalism-specific specialization. Each MoE-LoRA layer is made of multiple LoRA adapters acting as "experts." For each input token, a gating network computes a softmax distribution over these experts. The final output is a weighted sum of all expert outputs, allowing the model to learn a soft combination of specialized adaptations for each token. These adapters are optimized by the cross-entropy loss.

3.2.4 Dual-Loss Objective

HiDAC uses a dual-loss framework integrating Cross-Entropy loss (L_{ce}) and Contrastive loss (L_{lcl}) to optimize model representations and decision boundaries concurrently. The total loss (L_{total}) is a weighted sum $L_{total} = \lambda_{ce} \cdot L_{ce} + \lambda_{cl} \cdot L_{lcl}$, where λ_{ce} and λ_{cl} are hyperparameters.

Cross-Entropy Loss (L_{ce}) This loss is applied at the final classification layer as the primary objective for the prediction task.

Contrastive Loss (L_{lcl}) This loss provides the training signal for the foundational adapters in the lower layers (1-8) of the encoder. We experimented with two contrastive objectives: a Label-Centered

loss and a more traditional instance-vs-instance loss.

As detailed in our analysis (see Section 6), the Label-Centered method yielded superior performance and training stability, and was therefore selected for our final model. The chosen Label-Centered SCL, inspired by (Wu et al., 2024), simplifies the contrastive objective by introducing stable, learnable embeddings corresponding directly to each class label. Instead of comparing an instance to other random instances, the model’s task is now much clearer: it learns to make the representation of an instance (from layer 8) more similar to the embedding of its correct label, while making it dissimilar to the embedding of all incorrect labels. The loss is computed as:

$$L_{lcl} = -\log \frac{\exp(\text{sim}(\mathbf{h}_{\text{contrastive}}, \mathbf{e}_y)/\tau)}{\sum_{j \in C} \exp(\text{sim}(\mathbf{h}_{\text{contrastive}}, \mathbf{e}_j)/\tau)}$$

Here, $\mathbf{h}_{\text{contrastive}}$ is formed by taking the [CLS] token representation for each discourse unit from the intermediate layer’s output (layer 8), and then averaging these two vectors. The terms \mathbf{e}_y and \mathbf{e}_j refer to the learnable class prototypes; where \mathbf{e}_y is the embedding of the ground-truth class y , while the sum in the denominator is over the embeddings for all 17 labels in the set C . Finally, sim is the cosine similarity, and τ is the temperature hyperparameter.

On the other hand, the instance-vs-instance contrastive loss works by creating two slightly different vector representations for the same input text. The model is then trained to solve a simple matching task: it learns to pull these two views of the same text together, while simultaneously pushing them away from the representations of all other different texts in the batch. To provide a larger and more consistent set of negative examples, this method is augmented with a momentum-updated encoder and a negative queue (He et al., 2020).

4 Experimental Setup

4.1 Datasets

We trained and evaluated our models using the datasets provided by the shared task organizers.⁷ In total, the benchmark is composed of 39 datasets, covering 16 languages and 6 frameworks. These datasets were obtained from the following corpora: the Czech RST Discourse Treebank 1.0 (Poláková

⁷<https://github.com/disrpt/sharedtask2025>

et al., 2023), the Potsdam Commentary Corpus (Stede and Neumann, 2014; Bourgonje and Stede, 2020), the COVID-19 Discourse Dependency Treebank (Nishida and Matsumoto, 2022), the Discourse Dependency TreeBank for Scientific Abstracts (Yang and Li, 2018; Yi et al., 2021; Cheng and Li, 2019), the Genre Tests for Linguistic Evaluation corpus (Aoyama et al., 2023), the Georgetown University Multilayer corpus (Zeldes, 2017), the RST Discourse Treebank (Carlson et al., 2001), the Science, Technology, and Society corpus (Potter, 2008), the University of Potsdam Multilayer UNSC Corpus (Zaczynska and Stede, 2024), the Minecraft Structured Dialogue Corpus (Thompson et al., 2024), the Strategic Conversations corpus (Asher et al., 2016), the Basque RST Treebank (Iruskieta et al., 2013), the Persian RST Corpus (Shahmohammadi et al., 2021), the ANNOtation DIScursive corpus (Afantenos et al., 2012), the SUMM-RE corpus (Hunter et al., 2024; Prévot et al., 2025), the Dutch Discourse Treebank (Redeker et al., 2012), the Polish Discourse Corpus (Ogrodniczuk et al., 2024; Calzolari et al., 2024), the Cross-document Structure Theory News Corpus (Cardoso et al., 2011), the Russian RST Treebank (Toldova et al., 2017), the RST Spanish Treebank (da Cunha et al., 2011), the RST Spanish-Chinese Treebank (Cao et al., 2018), the Georgetown Chinese Discourse Treebank (Peng et al., 2022b,a), the DiscoNaija corpus (Scholman et al., 2025), the Penn Discourse Treebank (Prasad et al., 2014; Webber et al., 2019), the TED-Multilingual Discourse Bank (English) (Zeyrek et al., 2018, 2019), the LUNA Corpus Discourse Data Set (Tonelli et al., 2010; Riccardi et al., 2016), the Portuguese Discourse Bank (Mendes and Lejeune, 2022; Génereux et al., 2012), the Thai Discourse Treebank (Prasertsom et al., 2024), the Turkish Discourse Bank (Zeyrek and Webber, 2008; Zeyrek and Kurfalı, 2017), and the Chinese Discourse Treebank (Zhou et al., 2014).

The dataset is divided into official train/dev/test splits; however, some corpora are included only as surprise or out-of-domain (OOD) evaluation sets and are not represented in the training data. Additionally, four corpora are fully masked (i.e., argument texts are replaced with underscores), and two are partially masked due to LDC subscription requirements.

For our experiments, we used the official train and dev splits to train and evaluate our models, excluding the instances requiring an LDC subscription. As Table 1 shows, the class distribution of

	Label	Train (\approx170k)	Dev (\approx28k)
1	elaboration	19.9%	23.3%
2	conjunction	17.5%	16.5%
3	causal	11.7%	10.6%
4	temporal	9.4%	8.1%
5	query	6.5%	5.1%
6	contrast	5.2%	4.6%
7	concession	4.5%	4.7%
8	comment	4.0%	3.3%
9	explanation	3.7%	3.9%
10	purpose	3.1%	3.7%
11	condition	3.0%	2.7%
12	attribution	3.0%	3.2%
13	organization	2.9%	3.5%
14	frame	2.3%	2.9%
15	mode	1.4%	2.0%
16	reformulation	1.2%	1.4%
17	alternation	0.7%	0.7%

Table 1: Class distribution of discourse relation labels in the training and development sets.

the training dataset (\approx 170K instances) and development dataset (\approx 28K instances) is not balanced with *elaboration* and *conjunction* being overrepresented while *alternation* and *reformulation* are severely underrepresented.

4.2 Training Details

For the BERT-based models (see Section 3.1.1), we used standard fixed hyperparameters: We use the AdamW optimizer with a learning rate of $3e-5$, batch size of 32, weight decay of 0.01, and warm-up ratio of 0.1, we set a maximum of 20 epochs and applied early stopping with a patience of 3 epochs based on the validation loss and we used fixed random seeds (seed=42) for reproducibility across all experiments. We used the default `AutoModelForSequenceClassification` implementation from Hugging Face Transformers⁸ with 17 output classes corresponding to DISRPT 2025’s unified label set.

For the prompt-based experiments (see Section 3.1.2), we set the temperature to 0 to ensure fully deterministic outputs. We also fixed a random seed (seed=42) when selecting examples for the few-shot prompting setting for reproducibility.

The implementation of HiDAC uses *XLM-RoBERTa-Large* as the foundational model. Adapter modules are applied using forward hooks, preserving the pre-trained weights while updating only the small adapter matrices. During inference,

⁸<https://huggingface.co/docs/transformers>

Category	Size (%)	BERT-based Models																	
		mBERT						XLM-R-Base						XLM-R-Large					
		NO	RO	0	0.25	0.5	0.75	NO	RO	0	0.25	0.5	0.75	NO	RO	0	0.25	0.5	0.75
Overall	-	64.4%	63.5%	53.4%	64.0%	64.1%	64.3%	65.4%	64.6%	33.3%	63.8%	65.3%	65.1%	66.4%	65.8%	32.2%	65.9%	66.5%	66.8%
Framework																			
pdftb	27.0	65.6%	65.1%	51.6%	65.0%	65.5%	65.3%	67.3%	66.6%	25.5%	65.5%	67.1%	67.1%	68.8%	67.7%	22.9%	67.8%	68.8%	68.5%
rst	26.5	57.8%	57.4%	47.8%	58.4%	58.4%	58.5%	59.2%	58.3%	28.7%	57.2%	59.0%	59.0%	60.2%	60.5%	27.6%	60.2%	59.8%	60.2%
dep	16.7	72.2%	70.3%	62.5%	70.6%	71.2%	71.7%	72.0%	69.8%	46.4%	71.7%	70.1%	71.2%	72.2%	70.2%	46.1%	70.6%	72.4%	72.1%
sdrt	14.5	75.3%	74.1%	64.4%	74.7%	73.9%	75.0%	75.6%	75.4%	49.5%	74.4%	75.8%	75.0%	76.2%	75.6%	51.2%	76.4%	76.7%	76.7%
erst	12.5	56.0%	55.6%	46.8%	56.2%	55.7%	56.3%	56.7%	56.9%	24.2%	54.4%	57.6%	57.2%	58.2%	58.9%	21.6%	58.4%	59.5%	60.5%
iso	2.8	47.2%	45.7%	40.7%	47.2%	48.2%	46.3%	52.1%	50.7%	29.2%	49.2%	53.4%	50.8%	55.3%	51.4%	31.4%	53.0%	52.6%	56.7%
Language																			
eng	51.0	67.0%	66.0%	57.7%	66.5%	66.3%	66.7%	67.6%	66.6%	38.0%	66.1%	67.2%	67.6%	68.1%	67.2%	37.1%	67.7%	68.6%	68.9%
zho	9.8	60.3%	62.0%	47.6%	59.9%	60.4%	60.3%	60.7%	61.8%	26.9%	60.1%	62.2%	60.6%	64.2%	64.0%	24.9%	61.0%	63.9%	63.1%
rus	8.2	62.5%	61.2%	52.2%	62.0%	62.8%	62.9%	63.8%	62.6%	30.4%	60.5%	63.5%	63.4%	63.5%	63.8%	28.9%	63.7%	63.7%	64.2%
por	7.4	64.7%	64.2%	55.6%	65.8%	65.8%	65.2%	67.0%	66.9%	33.3%	65.6%	66.9%	66.3%	68.3%	68.0%	30.0%	68.9%	67.8%	68.0%
tha	4.5	93.5%	94.2%	66.3%	94.4%	94.0%	94.4%	96.0%	95.9%	26.7%	95.0%	96.1%	95.7%	95.7%	95.9%	26.5%	95.5%	95.9%	96.0%
pcm	3.8	51.0%	49.6%	33.3%	49.6%	48.8%	50.0%	52.9%	51.0%	22.9%	47.4%	50.5%	51.7%	56.1%	54.2%	21.0%	52.3%	53.5%	52.8%
pol	2.8	47.2%	45.7%	40.7%	47.2%	48.2%	46.3%	52.1%	50.7%	29.2%	49.2%	53.4%	50.8%	55.3%	51.4%	31.4%	53.0%	52.6%	56.7%
eus	2.2	50.3%	48.4%	37.8%	49.3%	50.8%	52.6%	52.4%	51.3%	22.5%	52.3%	54.1%	54.2%	54.6%	57.2%	21.8%	56.7%	53.7%	58.6%
fra	1.9	61.2%	54.5%	51.8%	60.6%	59.5%	59.1%	58.5%	57.6%	38.4%	60.0%	59.3%	54.5%	62.0%	58.1%	37.1%	62.9%	62.5%	62.5%
fas	1.8	55.1%	53.5%	45.9%	52.9%	53.5%	54.5%	54.3%	54.3%	28.1%	54.7%	54.1%	55.5%	57.9%	57.7%	30.1%	56.5%	56.5%	57.7%
spa	1.7	68.1%	65.6%	50.9%	67.7%	67.7%	69.2%	66.2%	66.7%	27.5%	63.5%	68.3%	64.8%	68.3%	68.1%	26.8%	68.3%	70.4%	66.9%
deu	1.6	44.7%	44.2%	37.6%	45.4%	47.6%	45.8%	47.1%	46.2%	18.6%	47.3%	48.2%	45.4%	49.3%	46.9%	19.2%	49.3%	45.6%	44.9%
nld	1.2	53.2%	51.1%	38.4%	54.4%	54.7%	56.5%	56.5%	50.5%	29.9%	56.5%	54.7%	55.9%	55.3%	57.1%	28.4%	58.9%	55.9%	54.7%
tur	0.8	51.7%	44.5%	44.1%	46.4%	50.2%	49.8%	54.0%	45.0%	30.8%	52.1%	52.1%	52.6%	52.6%	50.7%	32.2%	53.6%	54.5%	54.5%
ita	0.7	60.7%	63.1%	54.4%	62.1%	60.2%	59.2%	63.1%	63.6%	22.8%	59.7%	59.2%	60.7%	58.7%	60.7%	21.8%	62.6%	62.1%	65.0%
ces	0.4	40.7%	47.2%	33.3%	45.5%	39.8%	45.5%	48.0%	46.3%	22.0%	41.5%	39.8%	44.7%	45.5%	55.3%	23.6%	46.3%	53.7%	48.0%

Table 2: Accuracy of the BERT-based models (*mBERT*, *XLM-RoBERTa-Base* & *XLM-RoBERTa-Large*) across the two argument ordering and freezing strategies with subset size (%) on the development ($\approx 28k$) dataset. Note: NO = Natural ordering; RO = Relation-based ordering. The percentages represent the portion of the encoder’s top layers that we unfreeze after initially freezing the entire model. All progressive unfreezing experiments were conducted with NO argument ordering only. Light green cells indicate the highest accuracy within each model family for a given framework/language, while dark green cells indicate the overall best accuracy across all families.

Category	Size (%)	Prompt-based Models					Custom Model			
		Zero-Shot		Few-Shot			HiDAC			
		NO	RO	Exp 1	Exp 2	Exp 3	Instance *	Instance	Label Centered	Final HiDAC
Overall	-	40.85%	31.93%	39.90%	39.6%	34.7%	66.88%		66.97%	67.46%
Framework										
pdftb	27.0	43.22%	35.30%	43.50%	41.8%	37.50%	67.89%		67.64%	69.00%
rst	26.5	39.36%	28.80%	36.33%	36.9%	31.60%	60.63%		60.98%	60.92%
dep	16.7	51.78%	35.70%	48.60%	48.0%	38.65%	75.70%		75.96%	75.89%
sdrt	14.5	30.76%	31.30%	31.48%	32.2%	31.13%	76.17%		76.62%	76.67%
erst	12.5	36.90%	26.05%	37.85%	37.9%	33.50%	58.97%		58.68%	59.49%
iso	2.8	37.33%	39.18%	39.35%	40.4%	36.55%	50.79%		50.53%	51.97%
Language										
eng	51.0	40.36%	30.38%	39.90%	40.0%	34.70%	69.67%		69.75%	70.00%
zho	9.8	36.15%	27.68%	37.18%	35.5%	32.83%	64.35%		64.24%	65.06%
rus	8.2	45.12%	34.90%	39.15%	40.3%	37.53%	64.17%		63.64%	64.61%
por	7.4	47.38%	31.38%	41.53%	39.7%	34.78%	68.16%		68.36%	69.92%
tha	4.5	64.30%	64.30%	66.18%	64.4%	57.70%	95.82%		95.58%	95.90%
pcm	3.8	28.23%	23.75%	36.23%	32.3%	24.33%	51.24%		50.76%	53.33%
pol	2.8	37.33%	39.18%	39.35%	40.4%	36.55%	50.79%		50.53%	51.97%
eus	2.2	33.18%	26.35%	34.83%	35.9%	30.78%	55.54%		55.70%	53.75%
fra	1.9	45.63%	45.63%	35.30%	38.7%	37.38%	56.02%		57.36%	57.36%
fas	1.8	35.83%	29.60%	37.28%	38.8%	32.60%	52.30%		55.51%	55.31%
spa	1.7	34.68%	17.68%	38.80%	38.1%	30.33%	66.88%		66.88%	67.71%
deu	1.6	27.25%	25.88%	23.00%	23.0%	23.95%	56.86%		57.52%	56.64%
nld	1.2	34.50%	34.18%	31.65%	25.4%	24.03%	49.85%		53.47%	52.27%
tur	0.8	40.08%	28.98%	40.08%	36.5%	19.05%	48.82%		51.66%	52.61%
ita	0.7	33.75%	20.00%	30.00%	30.0%	37.50%	60.68%		60.19%	59.71%
ces	0.4	16.08%	16.08%	6.25%	13.4%	9.38%	49.59%		51.22%	49.59%

Table 3: Accuracy of Prompt Engineering using Claude Opus 4.0 (Zero-Shot, Few-Shot) and HiDAC on the development ($\approx 28k$) dataset. Note: NO = Natural ordering; RO = Relation-based ordering; Exp 1 = Multilingual examples; Exp 2 = English-only examples; Exp 3 = English-only mainly from weak-label examples. Light green cells indicate the highest accuracy within each model family for a given framework/language, while dark green cells indicate the overall best accuracy across all families.

the trained LoRA adapter weights are loaded on top of the frozen base model. This approach adds negligible computational latency compared to using the base model without any adapters, as the forward pass only involves a few small matrix multiplications (Hu et al., 2022). The hyperparameters were tuned empirically based on performance on the development set. Our final model uses a LoRA rank (r) of 128 with a scaling factor (α) of 256 for all adapters. For the contrastive loss, the temperature (τ) was set to 0.1. The final loss weighting coefficients were set to $\lambda_{ce} = 1.0$ and $\lambda_{cl} = 0.3$. We trained the model using the AdamW optimizer with a learning rate of $2e-5$ and a cosine learning rate scheduler with a warmup period covering the first 2 epochs. The model was trained with a batch size of 32, fixed random seed (seed = 42) and we used early stopping with patience of 2 epochs based on the accuracy of the development set. All experiments were conducted on Google Colab using a single NVIDIA A100 GPU. Training run took approximately 5 hours for the argument-ordering and HiDAC models, while the progressive unfreezing ones were shorter and varied depending on the unfreezing ratio.

5 Results

Tables 2 and 3 summarize the results across all three model families where we report the accuracy values overall, as well as by framework and language. Table 2 shows the results for the two argument-order arrangements and for the progressive unfreezing experiments at 0%, 25%, 50%, and 75% unfreezing ratios. Overall, accuracy increases across unfreezing ratios for all models, with the highest scores observed at the 75% unfreezing ratio, and natural argument ordering consistently outperformed relation-directed ordering across all models. Table 3 shows prompting results, with zero-shot achieving 40.85% and few-shot reaching 39.90%. HiDAC achieved the highest overall performance at 67.46%, outperforming both the best BERT baseline without progressive unfreezing, where the model is fully trainable (*XLM-RoBERTa-Large* at 66.43%) and with progressive unfreezing (66.76% with 75% unfreezing).

6 Analysis

We analyze the results of Tables 2 and 3 to better understand the impact of model choices.

6.1 BERT-Based Models

Effect of Model Size: As shown in Table 2 larger models tend to perform slightly better. *XLM-RoBERTa-Large* (550M parameters) achieves 66.4%, compared to *XLM-RoBERTa-Base* (65.4%, 270M parameters) and *mBERT* (64.4%, 179M parameters). Although the absolute gains are modest, this trend suggests that increased capacity may help the model capture complex semantic relationships across languages and annotation frameworks.

Effect of Progressive Unfreezing: Progressive unfreezing leads to modest performance improvements as more layers are unfrozen; accuracy steadily increases from the 25% to the 75% unfreezing ratio across all models. At 75% unfreezing, performance is comparable to or slightly higher than full fine-tuning while being significantly more parameter-efficient. For example, *XLM-RoBERTa-Large* achieves 66.8% accuracy while training only 41% of parameters (vs. 66.4% fully trained), and *mBERT* reaches 64.3% while training 36% of parameters (as opposed to 64.4% fully trained). This suggests that full fine-tuning may be unnecessary to achieve strong performance, allowing reductions in computational resources.

Cross-Formalism Analysis: The training data is dominated by PDTB (28.3%) and RST (31.8%). In contrast, SDRT and DEP represent only 18.7% and 4% of the training data, respectively. However, on the development set, SDRT (14.5% of dev) and DEP (16.7% of dev) achieve notably higher accuracy than PDTB (27% of dev) and RST (26.5% of dev) across models. For example, on the *XLM-R-Large* with unfreezing ratio of 0.75, we have the following accuracy values: SDRT = 76.7%, DEP = 72.1% vs. PDTB = 68.5%, RST = 60.2%. This raises the question of whether the 17 proposed unified labels may align more with these frameworks or whether other factors, such as the ratio of implicit to explicit relations, contribute to this discrepancy. Further analysis is needed to understand the source of these differences.

6.2 Prompt-based Models

Prompt engineering was significantly affected by relation-directed ordering, suggesting that LLMs may prefer the natural flow of arguments regardless of discourse relation source or direction. Surprisingly, performance did not improve when using few-shot prompts with four examples, whether

English-only or instance-language-specific, showing the difficulty of designing effective prompts for the proposed labels; particularly, for the following labels, which achieved F1 scores below 15%: *concession*, *explanation*, *frame*, *mode*, *organization*, and *reformulation*. When doubling the examples in Exp 3, accuracy dropped to 34.7%, which could be because the model became overwhelmed by conflicting patterns and noise. Too many examples could have created confusion rather than clarity. The instructions in all cases were in English only; it is worth investigating the use of language-specific instructions. Also, we only investigated with the Claude model; it may be beneficial to evaluate other LLMs on the unified discourse labels for a broader comparison.

6.3 HiDAC Model

Effect of Contrastive Learning Objective: We assess two distinct supervised contrastive learning (SCL) objectives. As shown in Table 3, the instance-based supervised contrastive learning objective, augmented with a momentum encoder and negative queue (He et al., 2020) achieved strong performance, comparable to our other methods. However, a closer analysis of the training dynamics revealed a significant issue: the contrastive loss value remained stagnant after the initial warm-up period, indicating that the model was failing to optimize this objective. We hypothesize this was caused by two factors. First, random in-batch negatives often included semantically similar pairs (e.g., two *elaboration* instances), which may have provided a contradictory training signal. Second, the constantly changing nature of the negative examples prevented the model from learning against a fixed target. Although the loss stagnated, it did lead to a strong performance.

The Label-Centered SCL improved the learning dynamics. Unlike the stagnant loss observed with the instance-vs-instance method, the Label-Centered objective resulted in a consistent decrease in the contrastive loss throughout training. This indicates that the model was able to effectively learn from the clearer, more stable training signal provided by the fixed label embeddings. While the final performance scores were comparable to the instance-based method, we selected the Label-Centered approach for our final model due to its demonstrably superior training stability.

HiDAC Final Tuning: Our final model incorporates few tuning enhancements to the Label-Centered architecture.

As a regularization technique, we used label smoothing within the cross-entropy loss function and applied gradient clipping during training. These methods mitigate overfitting, stabilize training dynamics, and facilitate smoother convergence.

In addition, we replaced the linear learning rate scheduler with a cosine annealing scheduler. This modification ensures a more gradual and stable reduction in learning rate, improving convergence behavior during the final training stages.

As shown in Table 3, these refinements delivered a final incremental performance boost, ultimately yielding our best overall results.

7 Conclusion

This paper presented three approaches for Task 3 of DISRPT 2025: transformer-based baselines, prompt-based models, and HiDAC, a hierarchical adapter-based model with a dual-loss objective. Our experiments showed that natural argument ordering and progressive unfreezing generally improved performance, while prompt-based approaches underperformed compared to fine-tuned transformers. HiDAC achieved the best overall accuracy while training on fewer parameters, showing that adapter-based methods can reduce training cost without sacrificing accuracy. Future work could focus on improving prompt-based models and exploring multi-task training across Tasks 1-3. Also we plan to explore a multi-layer contrastive loss, applying the SCL objective across several intermediate layers to build a more robust representation. Additionally, we will investigate using Focal Loss with class-aware weighting to better address the severe class imbalance in the dataset and improve performance on underrepresented labels.

8 Limitations

For the BERT-based models, we used fixed hyperparameters across all models and datasets without additional fine-tuning on the validation sets. Our prompt engineering experiments, although one setup included language-based examples, used instructions only in English. While we explored both zero-shot and few-shot prompting, we did not experiment with chain-of-thought reasoning and only used the Claude model.

With respect to HiDAC, the hierarchical dual-adaptor design introduces additional computational overhead during training; specifically, calculating two separate losses and extracting representations from intermediate layers makes the training process slower than a standard approach. Furthermore, experiments revealed that the model’s performance is sensitive to the balance between the cross-entropy and contrastive loss weights (λ_{ce} and λ_{cl}). This suggests that applying this framework to new datasets would require careful hyperparameter tuning.

The architectural choices of HiDAC could also be explored. The division of the PLM into lower (1-8) and upper (12-24) layers was based on prior work (Wu et al., 2024) and was not optimized; other partitioning schemes might yield different outcomes. Similarly, the gating mechanism in the upper-layer MoE adapters uses a simple soft mixture, and more advanced sparse routing strategies were not explored. Finally, the performance of the model is fundamentally dependent on the capabilities of the underlying XLM-RoBERTa backbone, and its effectiveness may vary when applied to different pre-trained encoders.

Finally, due to time constraints, we did not perform tests of statistical significance on the performance difference of the various experiments.

Acknowledgments

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Québec (FRQ), and the Pierre Arbour Foundation.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. [An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

Kaveri Anuranjana. 2023. [DiscoFlan: Instruction fine-tuning and refined text generation for discourse relation label classification](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28, Toronto, Canada.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. [GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation](#). In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. [Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Peter Bourgonje and Manfred Stede. 2020. The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Paris, France. European Language Resources Association (ELRA).

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. [The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, Toronto, Canada.

Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors. 2024. [Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation \(LREC-COLING 2024\)](#). ELRA and ICCL, Torino, Italy.

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.

Paula Christina Figueira Cardoso, Erick Galani Maziero, Maria Lucía del Rosario Castro Jorge, M. Eloize, R. Kibar Aji Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. CST-News - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá, Brazil.

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*, Aalborg, Denmark.

Yi Cheng and Sujian Li. 2019. [Zero-shot Chinese discourse dependency parsing via cross-lingual mapping](#). In *Proceedings of the 1st Workshop on Dis-*

- course Structure in *Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. [Supervised learning of universal sentence representations from natural language inference data](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 670–680, Copenhagen, Denmark.
- Nelson Filipe Costa and Leila Kosseim. 2025a. [A Multi-Task and Multi-Label Classification Model for Implicit Discourse Relation Recognition](#). In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2025)*, Avignon, France.
- Nelson Filipe Costa and Leila Kosseim. 2025b. [Multi-Lingual Implicit Discourse Relation Recognition with Multi-Label Hierarchical Learning](#). In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2025)*, Avignon, France.
- Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. [Mapping Explicit and Implicit Discourse Relations between the RST-DT and the PDTB 3.0](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP'23)*, pages 344–352, Varna, Bulgaria.
- 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.
- Florian Eichin, Yang Janet Liu, Barbara Plank, and Michael A. Hedderich. 2025. [Probing LLMs for multilingual discourse generalization through a unified label set](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2025)*, pages 18665–18684, Vienna, Austria.
- Michel Génèreux, Iris Hendrickx, and Amália Mendes. 2012. [Introducing the reference corpus of contemporary Portuguese online](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).
- 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.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 9729–9738, Seattle, Washington.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *International Conference on Learning Representations (ICLR 2022)*.
- Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, and Laurent Prévot. 2024. [MEETING: A corpus of French meeting-style conversations](#). In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 508–529, Toulouse, France. ATALA and AFPC.
- 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 *4th Workshop on RST and Discourse Studies*, pages 40–49, Fortaleza, Brasil.
- Wei Liu, Yi Fan, and Michael Strube. 2023. [HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.
- Amália Mendes and Pierre Lejeune. 2022. [CRPC-DB a Discourse Bank for Portuguese](#). In *Proceedings of the 15th International Conference on Computational Processing of Portuguese (PROPOR 2022)*, pages 79–89, Berlin, Heidelberg. Springer-Verlag.
- Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. [DisCut and DiscReT: MELODI at DISRPT 2023](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The 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.
- Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. 2024. [Polish Discourse Corpus \(PDC\): Corpus design, ISO-compliant annotation, data highlights, and parser development](#). In (Calzolari et al., 2024), pages 12829–12835.
- Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022a. [Chinese Discourse Annotation Reference Manual](#). Research Report, Georgetown University (Washington, D.C.).

- Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022b. [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*, pages 382–391, Online only. Association for Computational Linguistics.
- Lucie Poláková, Šárka Zikánová, Jiří Mírovský, and Eva Hajičová. 2023. [Czech RST discourse treebank 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Andrew Potter. 2008. Interactional Coherence in Asynchronous Learning Networks: A Rhetorical Approach. *Internet and Higher Education*, 11(2):87–97.
- Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. [Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation](#). *Computational Linguistics*, 40(4):921–950.
- Ponrawee Prasertsom, Apiwat Jaroonpol, and Attapol T. Rutherford. 2024. [The Thai Discourse Treebank: Annotating and Classifying Thai Discourse Connectives](#). *Transactions of the Association for Computational Linguistics*, 12:613–629.
- Laurent Prévot, Roxane Bertrand, and Julie Hunter. 2025. Segmenting a large french meeting corpus into elementary discourse units. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2025)*, Avignon, France.
- 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 2012)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).
- Giuseppe Riccardi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. 2016. [Discourse connective detection in spoken conversations](#). In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099.
- Merel CJ Scholman, Marian Marchal, AriaRay Brown, and Vera Demberg. 2025. DiscoNaija: a discourse-annotated parallel Nigerian Pidgin-English corpus. *Language Resources and Evaluation*, pages 1–37.
- Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. [Persian Rhetorical Structure Theory](#). *arXiv preprint arXiv:2106.13833*.
- Manfred Stede and Arne Neumann. 2014. [Potsdam Commentary Corpus 2.0: Annotation for discourse research](#). In *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*, pages 925–929, Reykjavik, Iceland.
- Kate Thompson, Julie Hunter, and Nicholas Asher. 2024. [Discourse structure for the Minecraft corpus](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4957–4967, Torino, Italia. ELRA and ICCL.
- Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. [Rhetorical relations markers in Russian RST treebank](#). In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. [Annotation of discourse relations for conversational spoken dialogs](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).
- Nawar Turk, Sevag Kaspar, and Leila Kosseim. 2025. [On the influence of discourse relations in persuasive texts](#). In *Proceedings of the 38th Canadian Conference on Artificial Intelligence (CanAI 2025)*, Calgary, Canada.
- Hanna Varachkina and Franziska Pannach. 2021. [A unified approach to discourse relation classification in nine languages](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, Punta Cana, Dominican Republic.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse TreeBank 3.0 annotation manual. Technical report, University of Edinburgh, Interactions, LLC, University of Pennsylvania.
- Yiheng Wu, Junhui Li, and Muhua Zhu. 2024. [Constrained Multi-Layer Contrastive Learning for Implicit Discourse Relationship Recognition](#). *arXiv preprint arXiv:2409.13716*.
- 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.
- Cheng Yi, Li Sujian, and Li Yueyuan. 2021. [Unifying discourse resources with dependency framework](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1058–1065, Huhhot, China. Chinese Information Processing Society of China.
- Karolina Zaczynska and Manfred Stede. 2024. [Rhetorical strategies in the UN security council: Rhetorical Structure Theory and conflicts](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 15–28, Kyoto, Japan. 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, 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.
- Deniz Zeyrek and Murathan Kurfalı. 2017. [TDB 1.1: Extensions on Turkish discourse bank](#). In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.
- Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogródniczuk. 2019. [TED Multilingual Discourse Bank \(TED-MDB\): A parallel corpus annotated in the PDTB style](#). *Language Resources and Evaluation*, pages 1–27.
- Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. [Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Deniz Zeyrek and Bonnie Webber. 2008. [A discourse resource for Turkish: Annotating discourse connectives in the METU corpus](#). In *Proceedings of the 6th Workshop on Asian Language Resources*.
- Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. [Chinese Discourse Treebank 0.5 LDC2014T21](#).