HITS at DISRPT 2025: Discourse Segmentation, Connective Detection, and Relation Classification

Yi Fan* **and Banerjee Souvik*** **and Michael Strube**
Heidelberg Institute for Theoretical Studies
{yi.fan, souvik.banerjee, michael.strube}@h-its.org

Abstract

This paper describes the submission of the HITS team to the DISRPT 2025 shared task. The shared task includes three sub-tasks: (1) discourse unit segmentation across formalisms, (2) cross-lingual discourse connective identification, and (3) cross-formalism discourse relation classification. For task (1), we use the google/mt5-xl model as our base model. Additionally, we combine the weighted cross-entropy loss function and adversarial training techniques. For task (2), we propose an ensemble of three encoder models whose embeddings are fused together with multi-head attention. We also integrate linguistic features and employ a CRF layer with label smoothing and focal loss to further improve performance. Finally for task (3), we introduce a two-stage curriculum learning framework with knowledge distillation. A smaller "student" model internalizes a larger "teacher" model's reasoning by first learning simple label prediction and then learning to analyze Chain-of-Thought explanations before the label prediction for more difficult samples.

The source code for our models is publicly available at: https://github.com/HereticFy/disrpt2025

1 Shared Task and Related Work

The shared task of Discourse Relation Parsing and Treebanking (DISRPT), since 2019, has been aiming to broaden the scope of discourse studies by including datasets and inviting researchers from different discourse theories, to facilitate cross-framework studies (Zeldes et al., 2019, Zeldes et al., 2021, Braud et al., 2023). The 2025 shared task proposes a unified typology of 17 discourse relations and contains three sub-tasks across sixteen different languages. It also adds a unique constraint of submitting only one multilingual model per sub-task and the model also has a size constraint of less than or equal to 4 billion parameters (for the closed track). Task 1 of the shared task addresses discourse unit segmentation, the foundational step of partitioning a text into discourse segments. The primary challenge lies in the significant diversity of segmentation guidelines across different annotation formalisms, such as Rhetorical Structure Theory (RST, MANN and Thompson, 1988), Segmented Discourse Representation Theory (SDRT, Lascarides and Asher, 2007) and languages. Therefore, the task aims to promote the development of a single, flexible model capable of handling this cross-formalism and cross-lingual variation.

Task 2 of the shared task is focused on discourse connective identification. The goal is to automatically locate and extract the explicit words or phrases (e.g., but, because, on the other hand) that signal a relationship between two spans of text. The provided datasets span multiple languages and are annotated using two different formalisms: the Penn Discourse Treebank (PDTB, Miltsakaki et al., 2004) and the International Organization for Standardization's framework for discourse relations (ISO, Pustejovsky et al., 2008). The primary challenge lies in the linguistic diversity of connectives and the structural differences between the two annotation schemes, requiring systems to handle both forms of variation. Both segmentation and connective identification remains an easy task in English owing to the large availability of English based corpora. However, it remains a bit of a challenge to train more resource constrained languages (for example, Farsi).

Task 3 concentrates on discourse relation classification between two discourse units. This is a challenging task even in a monolingual setting, as evidenced by the existence of implicit connectives. Implicit connective classification is a well studied work in discourse parsing literature (Liu and Strube,

*Equal contribution. Yi works on discourse segmentation while Souvik is responsible for connective detection and relation classification.

Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 63–78

2023, Liu et al., 2024a, Zhou et al., 2010, Shi et al., 2017). The task is fundamentally ambiguity heavy and more so in low resource corpora. Consequently, building a successful multilingual model requires a well-designed architecture capable of modeling the complex relationships between discourse units across all the diverse formalisms. The datasets' use of all the formalisms also means that systems must contend with potential differences in the sense inventories and annotation criteria between all the standards.

Most recent work relies on fine-tuning pretrained language models to achieve the best performance (Bakshi and Sharma, 2021, Lu et al., 2023). This is further demonstrated by the winning teams in the previous edition of the shared task. In 2023, the best performance in the discourse segmentation and connective identification task was achieved by the MELODI team (Metheniti et al., 2023). They fine-tuned a multilingual RoBERTa model for each language separately. For the relation classification task, the best performance was achieved by our previous team (Liu et al., 2023). They fine tune multilingual RoBERTa model for large datasets separately. But for others, they group datasets by their frameworks and jointly train model on framework groups.

Now with the advent of LLM, it remains to be seen how generative approaches would benefit such tasks. (Eichin et al., 2025) probes large language models (LLMs) to see whether they capture discourse knowledge that generalizes across languages and frameworks. This work provides wonderful insight into what model would be best suitable for the shared task objectives.

For more details on the statistics of the shared task dataset, we kindly invite the reader to refer to `https://github.com/disrpt/sharedtask2025`.

2 Discourse Unit Segmentation across Formalisms

2.1 Method

Following the shared task requirements for a single multilingual model under 4 billion parameters, we select ***google/mt5-xl*** (3.7B parameters) as our base model for Task 1. We employ the Low-Rank Adaptation (LoRA) technique (Hu et al., 2021) for parameter-efficient fine-tuning. Building upon findings that demonstrate the effectiveness of multilingual training in fields such as machine translation (Johnson et al., 2017; Dong et al., 2015; Aharoni et al., 2019), we adopt a multilingual joint fine-tuning strategy. This approach has been shown to outperform monolingual fine-tuning for each language, corroborated by Chen et al. (2024).

To investigate the impact of data composition on model performance, we designed and compared three distinct experimental configurations. Our primary setup involves fine-tuning a single fully multilingual model on the combined training data from all available languages, which is subsequently evaluated across all corresponding test sets. For comparison, we established a monolingual baseline, in which a separate model is trained and evaluated exclusively on the data for each language. We found that the group-specific configuration explores an intermediate approach by partitioning the corpora into two macro-groups: one for Chinese and another for all other languages, which can achieve the best performance for Task 1. For this setup, a specialized model was trained for each group and evaluated only within its respective language set.

Besides, due to time constraints, we are unable to investigate the role of linguistic typology in cross-lingual transfer for our task. Although our current experiments examine broad data compositions, a more fine-grained analysis could involve partitioning the training data based on language families. For example, we could train specialized models on families such as Romance or Germanic languages. This approach would enable a systematic evaluation of how typological proximity influences knowledge sharing and performance in discourse segmentation. A more interesting setup in these experiments would be to include language isolates like Basque. Such a setup could offer valuable insights into the boundaries and mechanisms of cross-lingual transfer. We aim to address this in future work.

Besides, the discourse segmentation task shows a significant class imbalance, with the *Seg=O* (non-boundary) tag being overwhelmingly dominant. We employ a weighted cross-entropy loss function during training to address this issue and encourage the model to focus on the rare but essential *Seg=B-seg* (boundary) tags. The weight for each class c, represented as w_c, is calculated using the inverse of the class frequency, a standard method for managing imbalance. The formula is defined as Equation 1, where N is the total number of tokens in the training set, C is the total number of unique classes, and N_c is the count of occurrences

Corpus	F_1	Corpus	F_1
nld.rst.nldt (Redeker et al., 2012a)	97.47	rus.rst.rrt (Pisarevskaya et al., 2017; Toldova et al., 2017a)	92.75
eng.rst.rstdt (Lynn Carlson, 2002; Carlson et al., 2003)	97.40	eng.erst.gentle (Aoyama et al., 2023a)	92.00
eng.sdrt.msdc (Thompson et al., 2024a)	95.64	eus.rst.ert (Iruskieta et al., 2013a; Aranzabe et al., 2015)	90.97
por.rst.cstn (Cardoso et al., 2011a)	95.64	zho.rst.gcdt (Peng et al., 2022c,a)	90.90
eng.dep.scidtb (Yang and Li, 2018a)	95.08	eng.rst.umuc (Zaczynska and Stede, 2024a)	88.21
deu.rst.pcc (Stede and Neumann, 2014a)	94.52	fra.sdrt.annodis (Afantenos et al., 2012a)	88.06
fas.rst.prstc (Shahmohammadi et al., 2021a)	94.03	zho.dep.scidtb (Cheng and Li, 2019a; Yi et al., 2021a)	87.83
ces.rst.crdt (Poláková et al., 2023a)	93.71	spa.rst.sctb (Cao et al., 2018a, 2017c,a, 2016a)	86.80
eng.erst.gum (Zeldes et al., 2025)	93.56	eng.rst.oll (Potter, 2008a)	86.66
eng.dep.covdtb (Nishida and Matsumoto, 2022a)	93.36	eng.rst.sts (Potter, 2023)	82.90
eng.sdrt.stac (Asher et al., 2016a)	93.34	zho.rst.sctb (Cao et al., 2018b, 2017d,b, 2016b)	73.24
spa.rst.rststb (da Cunha et al., 2011a)	93.05	fra.sdrt.summre (Hunter et al., 2024b)	65.04
Mean			**90.09**

Table 1: Discourse Segmentation: Results per datasets on the Treebanked data, on test set

of class c. This approach assigns a higher penalty to misclassifications of minority classes, thereby improving the model's F1 score on these critical tags.

$$w_c = \frac{N}{C \times N_c} \tag{1}$$

To enhance the model's robustness and generalization capabilities, particularly on subtle discourse cues, we incorporate adversarial training into our fine-tuning process. Specifically, we use the Fast Gradient Method (FGM) inspired by Goodfellow et al. (2015) to create adversarial perturbations on the word embedding layer. During each training step, after the standard backpropagation, FGM determines a perturbation, r_{adv}, for the embedding parameters θ_{emb} based on the gradient of the loss L, as shown in Equation 2, where ϵ is a hyperparameter controlling the size of the perturbation. This perturbation is then added to the original embeddings, the model then performs a second forward and backward pass to compute and gather the adversarial loss. This approach helps the model learn a smoother and more resilient decision boundary in the embedding space.

$$r_{\text{adv}} = \epsilon \frac{\nabla_{\theta_{\text{emb}}} L(\theta)}{\|\nabla_{\theta_{\text{emb}}} L(\theta)\|_2} \tag{2}$$

2.2 Results

Table 1 shows our experiment results for Task 1. The results in Table 1 show that our model performs strongly across most English-language datasets. This aligns with previous findings (Liu et al., 2023). However, we notice considerably lower performance on two specific corpora, fra.sdrt.summre and zho.rst.sctb, which warrants a closer qualitative analysis.

Our model performs the worst on the fra.sdrt.summre corpus. Our detailed investigation shows that its content, which comes from multi-party meeting dialogues, exhibits frequent linguistic disfluencies (e.g., "euh"), repetitions (e.g., "ok, voilà, donc"), and non-standard punctuation. This spoken, spontaneous style contrasts sharply with the formal news articles or blog articles prevalent in other datasets. We hypothesize that the leading cause of performance decline is the lack of sentence-ending periods and proper capitalization, along with differences between

spoken and written language. This is supported by the fact that several different models tested on this dataset also produced poor results. This points out two major limitations of current models: the input text needs to be properly formatted with correct punctuation and capitalization, and while they do well with formal written text, they struggle to identify segmentation cues in noisy, conversational dialogue.

Another dataset where our model underperforms is zho.rst.sctb. We attribute this to a potential data imbalance. Compared to the other two Chinese corpora in Task 1, zho.rst.sctb includes a broader range of genres and topics. However, this variety is paired with a smaller amount of training data, which likely hampers the model's ability to generalize effectively across its diverse content.

These findings highlight the significant challenges of out-of-domain generalization for discourse segmentation. Bridging the performance gap between written and spoken language, as well as between well-structured and disorganized texts, remains an important area for future research. We leave this as a direction for future work.

3 Discourse Connective Identification across Languages

3.1 Methodology: A Linguistically-Aware Ensemble with Multi-Feature Fusion

For the task of identifying discourse connectives across languages, we found encoder-only models to be significantly more effective and efficient than decoder-based generative architectures. The inherent bidirectionality of encoders is crucial for this task, and their smaller size enabled us to construct a powerful ensemble of multilingual models. This ensemble approach allows the strengths of each encoder to complement one another, leading to more robust performance. Recognizing that connective detection is a fundamentally linguistic challenge, we also enhanced our models by explicitly injecting linguistic information.

3.1.1 Model Architecture

Our proposed system for multilingual discourse connective identification is centered around a powerful ensemble of pretrained transformer-based encoders. They are further enhanced with explicit linguistic features: Part of Speech tags and dependency relations. It employs a fusion mechanism to fuse the hidden representations of the different encoders. A structured output layer that consists of a classification layer and a CRF layer. This section details the core components of our model architecture and training strategy.

Multi-Encoder Ensemble Backbone Our approach uses an ensemble of three heterogeneous multilingual models to create a robust feature representation that mitigates model-specific biases. We selected ***RemBERT*** for its strong cross-lingual transfer (Chung et al., 2021), ***XLM-RoBERTa (Large)*** for its proven performance on multilingual tasks (Conneau et al., 2020), and ***mDeBERTa-v3 (Base)*** for its improved disentangled attention mechanism (He et al., 2023)

For a given input sequence, each encoder independently generates contextualized hidden state representations, $H_i \in \mathbb{R}^{L \times D_i}$, where L is the sequence length and D_i is the hidden dimension of encoder i. This is our system to leverage the complementary strengths of each architecture.

Linguistic Feature Integration To make the model explicitly aware of grammatical context, we integrate two types of syntactic features derived from CoNLL-U file annotations: Part-of-Speech (POS) tags and Dependency Relations (Dep-Rels). These categorical features are converted into dense vectors via separate embedding layers, E_{pos} and E_{dep}. The resulting embeddings are concatenated and passed through a linear projection layer with a ReLU activation, allowing the model to learn complex interactions between these features. (Kiperwasser and Goldberg, 2016)

Feature Fusion Module We explore three strategies to fuse the outputs from the multiple encoders:

1. **Concatenation (`concat`):** The hidden states from all encoders are concatenated along the feature dimension:

$$H_{\text{fused}} = [H_1, H_2, \ldots, H_N] \quad (3)$$

2. **Weighted Fusion (`weighted`):** Each encoder's hidden state H_i is projected to a common dimension and the weights $\mathbf{w}$ are normalized via a softmax function.

$$H_{\text{fused}} = \sum_{i=1}^{N} \text{softmax}(\mathbf{w})_i \cdot \text{Linear}_i(H_i) \quad (4)$$

3. **Attention Fusion (`attention`):** Multi-Head Attention layer processes the concatenated

Corpus	F_1	Corpus	F_1
eng.pdtb.pdtb (Prasad et al., 2008, 2018, 2019)	93.15	deu.pdtb.pcc (Bourgonje and Stede, 2020)	79.37
tur.pdtb.tdb (Zeyrek and Kurfalı, 2017)	93.07	por.pdtb.tedm (Zeyrek et al., 2019, 2018a)	78.38
eng.pdtb.gentle (Aoyama et al., 2023b)	89.20	eng.pdtb.tedm (Zeyrek et al., 2019, 2018a)	78.18
eng.pdtb.gum (Liu et al., 2024b)	87.09	zho.pdtb.ted (Long et al., 2020)	76.04
tha.pdtb.tdtb (Sriwirote et al., in press; Boonkwan et al., 2020)	86.14	pol.iso.pdc (Ogrodniczuk et al., 2024a)	72.18
zho.pdtb.cdtb (Zhou et al., 2014a)	84.01	ita.pdtb.luna (Tonelli et al., 2010; Riccardi et al., 2016)	70.81
por.pdtb.crpc (Mendes and Lejeune, 2022)	80.86	tur.pdtb.tedm (Zeyrek et al., 2018a, 2019)	65.80
pcm.pdtb.disconaija (Scholman et al., 2025)	80.82		
Mean			**81.00**

Table 2: Discourse Connective Identification: Results per datasets on the Treebanked data, on test set

hidden states to dynamically learn token-level combinations of the different representations.

The final fused representation is concatenated with our linguistic feature embeddings. We found that the attention fusion works best empirically. The results provided in Table 2 use the same fusion method.

Classifier Head and CRF Layer The combined representation is passed through a multi-layer classifier head before a final linear layer projects the features into the label space, producing logits. Instead of making independent predictions, we employ a **Conditional Random Field (CRF)** layer. A CRF models dependencies between adjacent labels by learning a matrix of transition scores. The final output is determined by the Viterbi algorithm, which finds the globally optimal sequence of labels, thus ensuring syntactically valid tag sequences (e.g., an 'I-conn' must follow a 'B-conn').

3.1.2 Training and Optimization

The model is trained end-to-end using a strategy designed for robustness and performance on imbalanced data.

Hybrid Loss Function We train the model end-to-end with a hybrid loss function designed for robustness on imbalanced data. The total loss combines the following components:

- **CRF Loss:** The negative log-likelihood of the gold-standard label sequence, calculated by a final Conditional Random Field (CRF) layer (Lafferty et al., 2001).
- **Focal Loss:** To address the severe class imbalance between 'O' (outside) labels and connective labels ('B-conn', 'I-conn'), we incorporate Focal Loss (Lin et al., 2017). Similar to the method in task 1, this loss modifies the standard cross-entropy to focus training on hard-to-classify examples with a weight calculation dependent on the training set itself:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \quad (5)$$

  where γ is a tunable focusing parameter. The loss is computed on the logits before the CRF layer.
- **Label Smoothing:** We also apply Label Smoothing (Szegedy et al., 2016), a regularization technique that discourages overconfidence in the model's predictions to improve calibration and generalization.

3.2 Results

Table 2 shows the F1-score for the various PDTB corpora and the ISO corpus. The F1-scores span a wide range, from a low of 65.80 to a high of 93.15, with a mean of 81.00. This highlights the varying difficulty and perhaps the maturity of the annotation schemes and resources across different languages and domains. Corpora like Italian (ita.pdtb.luna at 70.81) and Polish (pol.iso.pdc at

72.18) are on the lower end of the performance spectrum. The ita.pdtb.luna corpus is a corpus of conversational spoken dialogues. This is a significant difference from corpora based on written text, like the Penn Discourse Treebank (PDTB), which uses news articles. Spoken language is often less structured and can contain interruptions, overlaps, and less formal grammatical constructions, making discourse relations more ambiguous. As for pol.iso.pdc, we note that the dataset is not very large. Another reason could be Polish is a strongly inflected language, resulting in high type counts and bad generalizability for word-piece based models across different forms of the same lexical item.

The corpora based on TED talks consistently have lower F1-scores compared to other corpora in the same language (tur.pdtb.tedm, zho.pdtb.ted, eng.pdtb.tedm, por.pdtb.tedm). This is most likely because they have no training set and are test-only, which suggests our method can't generalize well from other datasets. Clearly, the choice of corpus (i.e., the type of text) has a massive impact on performance, often more so than the language itself. Further analysis needs to be done to understand the nuances in the score difference. We also tried out adversarial training strategies, but the scores were barely affected by the strategy.

4 Discourse Relation Classification across Formalisms

4.1 Method Introduction: A Two-Stage Curriculum Learning Framework

Our approach to multilingual discourse relation classification is a two-stage fine-tuning framework designed to transfer the nuanced reasoning capabilities of a very large "teacher" model to a compact "student" model ($\leq$ 4B parameters). We call this methodology Rationale-Enhanced Curriculum Learning (RECL). It combines supervised fine-tuning with hard-sample mining and a weighted curriculum, structured in a way that mimics a student's learning process: first, a broad initial study, followed by targeted tutoring on difficult topics.

Given size constraints, the core idea is to use knowledge distillation (Hinton et al., 2015), not by copying the output probabilities, but by transferring the explicit reasoning process of the teacher model to the student through chain-of-thought (CoT) rationales (Wei et al., 2022). This is particularly suited for a relatively complex task like discourse relation classification. Our framework is also explicitly designed to mitigate catastrophic forgetting (Kirkpatrick et al., 2017) by ensuring that the model consolidates its existing knowledge while learning from its mistakes.

4.2 Foundational Model and Task Formulation

Our student model is ***google/gemma-2-2b-it***. A 2B parameter model was chosen because it empirically outperformed the larger 3-4B models we evaluated by a small margin. We formulate the task as a generative problem. The model is prompted with two text units (Argument 1 and Argument 2), the full sentence they are part of (full context), the direction of the relation, and a list of all 17 labels. All this information is parsed from the training files themselves. Finally, the model's task is to generate a single, structured JSON output.

This strict output format simplifies parsing and ensures reliable evaluation. The system prompt explicitly instructs the model on its role and output format. The output format is `{"label": "classification"}`.

```
You are a discourse relation
classifier.  Your task is to
analyze text pairs and classify
their discourse relationship and
label them from the given labels.


IMPORTANT: Your response must be
ONLY a JSON object in the format
{"label": "your_classification"}


Do not include any other text or
explanations outside of the JSON.
```

4.3 Stage 1: Initial Supervised Fine-Tuning (SFT)

Objective To train a competent classifier that learns the general patterns of discourse relations across multiple languages. This stage is analogous to a student attending a general lecture course.

Data The full training set, with samples from all available languages, is loaded and combined into a unified dataset for comprehensive training. It exposes the model to the full diversity of the task.

Corpus	Accuracy	Corpus	Accuracy
tha.pdtb.tdtb	93.97	spa.rst.rststb	66.43
eng.sdrt.msdc	88.90	eng.pdtb.tedm	66.38
eng.dep.scidtb	81.41	por.pdtb.tedm	65.93
eng.pdtb.pdtb	79.32	eng.pdtb.gentle	65.65
por.pdtb.crpc	75.48	nld.rst.nldt	64.92
eng.sdrt.stac	75.00	eng.rst.rstdt	64.64
spa.rst.sctb	74.84	eng.erst.gentle	62.66
rus.rst.rrt	71.87	eng.rst.umuc	61.57
zho.pdtb.cdtb	71.37	zho.rst.sctb	59.75
zho.dep.scidtb	70.23	tur.pdtb.tedm	58.68
eng.dep.covdtb	70.07	fas.rst.prstc	58.45
pol.iso.pdc	69.99	deu.rst.pcc	58.24
por.rst.cstn	69.49	pcm.pdtb.disconaija	57.82
zho.rst.gcdt	68.52	deu.pdtb.pcc	56.70
eng.pdtb.gum	67.46	eng.rst.oll	54.98
eng.erst.gum	67.26	eus.rst.ert	53.20
ita.pdtb.luna	67.20	fra.sdrt.annodis	52.82
zho.pdtb.ted	67.07	eng.rst.sts	52.74
tur.pdtb.tdb	66.75	ces.rst.crdt	52.03
Macro Average			**66.78**
Micro Average			**72.24**

Table 3: Discourse Relation Classification: Results per datasets on test set

Training We use Parameter-Efficient Fine-Tuning (PEFT) with the LoRA (Low-Rank Adaptation) strategy (Hu et al., 2021). This efficiently adapts the model by training only a small number of parameters in the attention mechanism's projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the feed-forward network layers (`gate_proj`, `up_proj`, `down_proj`).

4.4 Stage 2: Rationale-Enhanced Curriculum Learning

This stage refines the model by focusing on its specific weaknesses, guided by the principle that explicit reasoning can help solve complex problems. It unfolds in three phases.

4.4.1 Identifying the Student's Weaknesses (Hard-Sample Mining)

First, we identify the samples that the Stage 1 model struggles with. We run inference on the entire training set using the model fine-tuned from stage 1. The samples for which the model predicts incorrectly are classified as "hard samples". These samples represent the gaps in the model's initial understanding and form the basis for our targeted curriculum. One should also note that the validation set and test set remain untouched throughout the whole process. We deliberately use the training set for this identification, rather than the development set, to ensure the development set remains a true proxy for unseen test data. Using it to inform the training curriculum would mean it no longer simulates genuine test conditions, which would compromise its ability to provide an unbiased evaluation of the model.

4.4.2 Generating Expert Explanations (Knowledge Distillation)

To provide the necessary "tutoring" for these hard samples, we distill knowledge from a vastly more powerful teacher model, `Qwen/Qwen2.5-72B-Instruct`. We prompt this teacher model to act as a "distinguished computational linguist" and generate a detailed Chain-of-Thought rationale for each hard sample. This rationale explains why a specific label is

correct, citing linguistic evidence and comparing it against other plausible labels. We have also added handwritten Chain-of-Thought rationales for 4 samples from the training dataset. Those handwritten rationales serve as few-shot examples for the model to aid in rationale generation. This process generates high-quality, explanatory data. This large-scale generation task was made feasible by using the `vLLM` (Kwon et al., 2023) library for high-throughput inference on a multi-GPU cluster. For the shared task, we submit the file containing the rationales to avoid the need for loading such a huge model.

4.4.3 Targeted Tutoring with Memory Consolidation (Weighted Fine-Tuning)

Curriculum learning is a training strategy inspired by human education where a model is not shown training samples in a random order, but rather in a meaningful sequence that progresses from easy to more complex examples. This approach helps guide the model towards a better solution and can improve generalization by allowing it to first learn simple concepts before tackling more difficult ones (Bengio et al., 2009). Thus, the final step is to retrain the model, but with a curriculum designed to fix its mistakes while retaining its existing knowledge. We start with the weights of the Stage 1 model, not the original pre-trained model.

The training data for this stage is a strategic mix:

Hard Samples These are the previously misclassified samples. They are now presented to the model with a new prompt that includes the expert-generated CoT rationale under the heading "Expert Analysis." This explicitly guides the model through the reasoning process it failed to grasp initially.

Easy Samples To prevent catastrophic forgetting, the samples that the model classified correctly in Stage 1 are also included. These are presented with the original, simpler prompt from Stage 1, reinforcing the model's existing strengths.

To force the model to prioritize learning from its mistakes, we apply a weighted loss function during training. The hard samples with rationales are assigned a loss weight of 1.5, while the easy samples retain a weight of 1.0. This ensures the training gradient is more significantly influenced by the need to correct prior errors. The learning rate was also much lower compared to the first stage, and the epoch was kept at 1.

4.5 Results

For evaluation, we first merge the Stage 1 LoRA adapter into the base model's weights and then apply a new LoRA adapter for Stage 2. This sequential adaptation approach is a common practice for multi-stage fine-tuning, allowing the model to first acquire broad knowledge before specializing in a subsequent task, a methodology employed in developing specialized models (Wu et al., 2024). This process is efficiently managed using standard libraries designed for parameter-efficient fine-tuning (Mangrulkar et al., 2022)

The ultimate goal of our two-stage process is to produce a more capable standalone classifier. The evaluation protocol measures this outcome directly by tasking the model with classifying unseen samples from the test set using only the standard prompt from stage 1. This approach rigorously tests whether the knowledge distilled from the teacher model's rationales has been successfully integrated into the student model's own parameters, leading to a genuine enhancement of its intrinsic reasoning abilities.

As can be seen from the results in 3, the accuracy scores for a lot of languages are quite low. This highlights the incredibly difficult nature of the task itself. There does not seem to be any sort of clear trend, but the ted datasets perform poorly here as well. A more thorough investigation is required that involves ablation studies. This would reveal which component of our two-stage fine-tuning process contributes the most or, conversely, least to the accuracy score. We found that there was a 2.12 % increase in micro average score from stage 1 to stage 2. This suggests that the model does use the Chain-Of-Thought rationales to its advantage, but not quite to the extent of warranting the use of such a technique on a wider scale. Future work could look at using task vectors or changing the model's internal, like the representation space, to explicitly make the model "internalise" the rationales for the harder samples.

5 Conclusion

This paper presents our strategies for the DISRPT 2025 Shared Task. In Task 1, our approach involves fine-tuning through multilingual joint training on linguistically motivated language groups. We in-

corporated two key techniques to improve model performance: a weighted loss function to address the task's significant class imbalance and Fast Gradient Method (FGM) adversarial training to boost the model's robustness.

In task 2, our approach involves building an ensemble of three encoder models whose embeddings are smartly fused together with a multi-head attention layer. We also add Part-Of-Speech tags and dependency relations present in the training file as linguistic features. A CRF layer is added after the classification layer to account for dependencies between adjacent labels. To account for label imbalance, we use focal loss and label smoothing. This ensures our model is robust and flexible enough to handle different languages.

In task 3, we use a two-stage fine-tuning framework designed to transfer the nuanced reasoning capabilities of a very large "teacher" model to a compact "student" model so that the smaller model can learn complex discourse relationships. The fine-tuning process follows a curriculum learning framework. In such a framework, the model learns to perform increasingly harder tasks. In our case, the model first learns to look at the discourse units and then predict the label, followed by looking at Chain-Of-Thought reasoning for harder examples. This way, it can learn to internalise such reasoning and increase prediction accuracy on the harder samples. Future work could use this method of knowledge distillation and curriculum learning for more complex discourse-related tasks.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany.

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. 2012a. 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'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

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. 2012b. 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'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023a. 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.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023b. 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.

María Jesús Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Díaz, Deliana Ilarraza, Iakes Goenaga, and Koldo Gojenola. 2015. Automatic conversion of the basque dependency treebank to universal dependencies. In *the fourteenth international workshop on treebanks an linguistic theories (TLT14)*, pages 233–241, Warsaw, Poland.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016a. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016b. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Sahil Bakshi and Dipti Sharma. 2021. A transformer based approach towards identification of discourse

unit segments and connectives. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 13–21, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pages 41–48, New York, NY, USA. ACM.

Prachya Boonkwan, Vorapon Luantangsrisuk, Sitthaa Phaholphinyo, Kanyanat Kriengket, Dhanon Leenoi, Charun Phrombut, Monthika Boriboon, Krit Kosawat, and Thepchai Supnithi. 2020. The annotation guideline of lst20 corpus. *arXiv preprint arXiv:2008.05055*.

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) (to appear)*, 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)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

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. 2016a. A corpus-based approach for Spanish-Chinese language learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 97–106, Osaka, Japan. The COLING 2016 Organizing Committee.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2016b. A corpus-based approach for Spanish-Chinese language learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 97–106, Osaka, Japan. The COLING 2016 Organizing Committee.

Shuyuan Cao, Iria Da-Cunha, and Mikel Iruskieta. 2017a. Toward the elaboration of a spanish-chinese parallel annotated corpus. In *Professional and Academic Discourse: an Interdisciplinary Perspective*, volume 2 of *EPiC Series in Language and Linguistics*, pages 315–324. EasyChair.

Shuyuan Cao, Iria Da-Cunha, and Mikel Iruskieta. 2017b. Toward the elaboration of a spanish-chinese parallel annotated corpus. In *Professional and Academic Discourse: an Interdisciplinary Perspective*, volume 2 of *EPiC Series in Language and Linguistics*, pages 315–324. EasyChair.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018a. 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.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018b. 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.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018c. 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.

Shuyuan Cao, Nianwen Xue, Iria da Cunha, Mikel Iruskieta, and Chuan Wang. 2017c. Discourse segmentation for building a RST Chinese treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 73–81, Santiago de Compostela, Spain. Association for Computational Linguistics.

Shuyuan Cao, Nianwen Xue, Iria da Cunha, Mikel Iruskieta, and Chuan Wang. 2017d. Discourse segmentation for building a RST Chinese treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 73–81, Santiago de Compostela, Spain. 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. 2011a. 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.

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. 2011b. 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*.

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.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Yi Cheng and Sujian Li. 2019a. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Yi Cheng and Sujian Li. 2019b. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Hyung Won Chung, Carlos Riquelme, Jiahe Vu, and Cagan Anil. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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. *Preprint*, arXiv:1911.02116.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011a. 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.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011b. 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.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. 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)*, pages 18665–18684, Vienna, Austria. Association for Computational Linguistics.

Michel Généreux, 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'12)*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, and Laurent Prévot. 2024a. 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.

Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, and Laurent Prévot. 2024b. SUMM-RE: A corpus of French meeting-style conversations. In *35èmes Journées d'Études sur la Parole (JEP 2024)*, volume 1 : articles longs et prises de position, pages 508–529, Toulouse, France. ATALA & AFPC.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013a. 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.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013b. 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.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In H. Bunt and R. Muskens, editors, *Computing Meaning: Volume 3*, pages 87–124. Kluwer Academic Publishers.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

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.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Stephen Wan, and Michael Strube. 2024a. What causes the failure of explicit to implicit discourse relation recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024b. GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

Wanqiu Long, Bonnie Webber, and Deyi Xiong. 2020. TED-CDB: A large-scale Chinese discourse relation dataset on TED talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803, Online. Association for Computational Linguistics.

Xin Lu, Zhuojun Li, Yanpeng Tong, Yanyan Zhao, and Bing Qin. 2023. HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

Mary Ellen Okurowski Lynn Carlson, Daniel Marcu. 2002. RST Discourse Treebank LDC2002T07.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

WILLIAM MANN and Sandra Thompson. 1988. Rethorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Amália Mendes and Pierre Lejeune. 2022. Crpc-db a discourse bank for portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 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.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*

(LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

Noriki Nishida and Yuji Matsumoto. 2022a. 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.

Noriki Nishida and Yuji Matsumoto. 2022b. 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. 2024a. Polish discourse corpus (PDC): Corpus design, ISO-compliant annotation, data highlights, and parser development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12829–12835, Torino, Italia. ELRA and ICCL.

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. 2024b. 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. Chinese Discourse Annotation Reference Manual. Research Report, Georgetown University (Washington, D.C.).

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022c. 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.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022d. 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.

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 *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 194–204.

Lucie Poláková, Šárka Zikánová, Jiří Mírovský, and Eva Hajičová. 2023a. Czech RST discourse treebank 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Lucie Poláková, Šárka Zikánová, Jiří Mírovský, and Eva Hajičová. 2023b. Czech RST discourse treebank 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Andrew Potter. 2008a. Interactional coherence in asynchronous learning networks: A rhetorical approach. *The Internet and Higher Education*, 11:87–97.

Andrew Potter. 2008b. Interactional coherence in asynchronous learning networks: A rhetorical approach. *Internet and Higher Education*, 11(2):87–97.

Andrew Potter. 2023. STS-Corpus. `https://github.com/anpotter/STS-Corpus`. Retrieved from: `https://github.com/anpotter/STS-Corpus`.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05.

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 SIGDIAL*.

James Pustejovsky, Kiyong Lee, Harry Bunt Harry, Branmir Boguraev, and Nancy Ide. 2008. Language resource management—semantic annotation framework (semaf)—part 1: Time and events. *International Organization*.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012a. Multilayer 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).

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012b. 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).

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. 2021a. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021b. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Panyut Sriwirote, Wei Qi Leong, Charin Polpanumas, Santhawat Thanyawong, William Chandra Tjhi, Wirote Aroonmanakun, and Attapol T. Rutherford. in press. The thai universal dependency treebank. *Transactions of the Association for Computational Linguistics*.

Manfred Stede and Arne Neumann. 2014a. 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).

Manfred Stede and Arne Neumann. 2014b. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Language Resources and Evaluation Conference (LREC '14)*, pages 925–929, Reykjavik.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024a. 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.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024b. 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. 2017a. 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.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017b. 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'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

An Yang and Sujian Li. 2018a. 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.

An Yang and Sujian Li. 2018b. 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. 2021a. 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.

Cheng Yi, Li Sujian, and Li Yueyuan. 2021b. 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. 2024a. 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.

Karolina Zaczynska and Manfred Stede. 2024b. 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, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. 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.

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, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (tedmdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38.

Deniz Zeyrek, Amalia Mendes, and Murathan Kurfali. 2018a. Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. In *LREC*.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018b. 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. 2014a. Chinese Discourse Treebank 0.5 LDC2014T21.

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014b. Chinese Discourse Treebank 0.5 LDC2014T21.

Zhi Min Zhou, Man Lan, Zheng Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the SIGDIAL 2010 Conference*, pages 139–146, Tokyo, Japan. Association for Computational Linguistics.

A Data

We train and evaluate our models using all the datasets provided by the shared task organizers.* In total, the benchmark is composed of 39 datasets, covering 13 languages and 6 frameworks. These datasets were obtained from the following corpora: the Czech RST Discourse Treebank 1.0 (Poláková et al., 2023b), the Potsdam Commentary Corpus (Stede and Neumann, 2014b; Bourgonje and Stede, 2020), the COVID-19 Discourse Dependency Treebank (Nishida and Matsumoto, 2022b), the Discourse Dependency TreeBank for Scientific Abstracts (Yang and Li, 2018b; Yi et al., 2021b; Cheng and Li, 2019b), the Genre Tests for Linguistic Evaluation corpus (Aoyama et al., 2023b), the Georgetown University Multilayer corpus (Zeldes, 2017), the RST Discourse Treebank (Carlson et al., 2001), the Science, Technology, and Society corpus (Potter, 2008b), the University of Potsdam Multilayer

*https://github.com/disrpt/sharedtask2025

UNSC Corpus (Zaczynska and Stede, 2024b), the Minecraft Structured Dialogue Corpus (Thompson et al., 2024b), the Strategic Conversations corpus (Asher et al., 2016b), the Basque RST Treebank (Iruskieta et al., 2013b), the Persian RST Corpus (Shahmohammadi et al., 2021b), the ANNOtation DIScursive corpus (Afantenos et al., 2012b), the SUMM-RE corpus (Hunter et al., 2024a; Prévot et al., 2025), the Dutch Discourse Treebank (Redeker et al., 2012b), the Polish Discourse Corpus (Ogrodniczuk et al., 2024b; Calzolari et al., 2024), the Cross-document Structure Theory News Corpus (Cardoso et al., 2011b), the Russian RST Treebank (Toldova et al., 2017b), the RST Spanish Treebank (da Cunha et al., 2011b), the RST Spanish-Chinese Treebank (Cao et al., 2018c), the Georgetown Chinese Discourse Treebank (Peng et al., 2022d,b), 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., 2018b, 2019), the LUNA Corpus Discourse Data Set (Tonelli et al., 2010; Riccardi et al., 2016), the Portuguese Discourse Bank (Mendes and Lejeune, 2022; Généreux 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., 2014b).