EMNLP CODI-CRAC 2025

The 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)

Proceedings of the Workshop

November 9, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-344-9

Preface

Welcome to the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025).

DISRPT is a shared task on discourse processing across formalisms, for a variety of languages and genres, with three subtasks this year: Task 1: discourse segmentation, Task 2: connective detection, and Task 3: relation classification.

We provided training, development, and test datasets from all available languages in RST, SDRT, PDTB, DEP (discourse dependencies), the ISO framework, and the Enhanced Rhetorical Structure Theory (eRST), using a uniform format. Because different corpora, languages, and frameworks use different guidelines, the shared task aims at promoting the design of flexible methods for dealing with various guidelines, to propose a joint evaluation of discourse parsing approaches and to push forward the discussion on converging standards for discourse units and relations.

DISRPT 2025 is part of the joint CODI-CRAC 2025 workshop, a venue that brings together researchers working on all aspects of discourse in Computational Linguistics and NLP. We hope that the next CODI workshops will also feature shared tasks on discourse analysis, as the domain needs more research promoting thorough and diversified evaluation as well as more consistent standards and expansions to languages and text types not yet covered in the field.

We thank the CODI organizers, and the reviewers who helped improve the papers and reproduce the participating systems. Finally we would like to thank the EMNLP 2025 workshop chairs Sunipa Dev, Maja Popović, and Eleftherios Avramidis who organized the EMNLP workshop program.

The DISRPT 2025 Organizers,

Chloé Braud, Chuyuan Li, Yang Janet Liu, Philippe Muller and Amir Zeldes

Program Committee

Program Committee

Chloé Braud, IRIT, CNRS
Chuyuan Li, The University of British Columbia
Yang Janet Liu, University of Pittsburgh
Philippe Muller, IRIT, University of Toulouse
Amir Zeldes, Georgetown University
Robin Pujol, IRIT, University of Toulouse

Table of Contents

Program

Sunday, November 9, 2025

10:15 - 10:30 *Opening Remarks*

The DISRPT 2025 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification
Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu and Philippe Muller

11:00 - 12:00 *Poster Session - Shared Task papers*

DisCuT and DiscReT: MELODI at DISRPT 2025 Multilingual discourse segmentation, connective tagging and relation classification
Robin Pujol, Firmin Rousseau, Philippe Muller and Chloé Braud

CLaC at DISRPT 2025: Hierarchical Adapters for Cross-Framework Multilingual Discourse Relation Classification
Nawar Turk, Daniele Comitogianni and Leila Kosseim

DeDisCo at the DISRPT 2025 Shared Task: A System for Discourse Relation Classification
Zhuoxuan Ju, Jingni Wu, Abhishek Purushothama and Amir Zeldes

HITS at DISRPT 2025: Discourse Segmentation, Connective Detection, and Relation Classification
Souvik Banerjee, Yi Fan and Michael Strube

SeCoRel: Multilingual Discourse Analysis in DISRPT 2025
Sobha Lalitha Devi, Pattabhi Rk Rao and Vijay Sundar Ram

The DISRPT 2025 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification*

Chloé Braud
UT-IRIT/CNRS/ANITI
chloe.braud@irit.fr

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Chuyuan Li
University of British Columbia
chuyuan.li@ubc.ca

Yang Janet Liu
University of Pittsburgh
jal787@pitt.edu

Philippe Muller
UT-IRIT/CNRS
philippe.muller@irit.fr

Abstract

In 2025, we held the fourth iteration of the DISRPT Shared Task (**Dis**course **R**elation **P**arsing and **T**reebanking) dedicated to discourse parsing across formalisms. Following the success of the 2019, 2021, and 2023 tasks on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification, this iteration added 13 new datasets, including three new languages (Czech, Polish, Nigerian Pidgin) and two new frameworks: the ISO framework and Enhanced Rhetorical Structure Theory, in addition to the previously included frameworks: RST, SDRT, DEP, and PDTB. In this paper, we review the data included in DISRPT 2025, which covers 39 datasets across 16 languages, survey and compare submitted systems, and report on system performance on each task for both treebanked and plain-tokenized versions of the data. The best systems obtain a mean accuracy of 71.19% for relation classification, a mean F_1 of 91.57 (Treebanked Track) and 87.38 (Plain Track) for segmentation, and a mean F_1 of 81.53 (Treebanked Track) and 79.92 (Plain Track) for connective detection. The data and trained models of several participants can be found at https://huggingface.co/multilingual-discourse-hub.[1]

1 Introduction

Automatic discourse analysis consists in identifying semantic and pragmatic links between text segments that organize a monologue or dialogue into a coherent and meaningful whole. The goal of discourse parsing is to build a discourse structure representing these links, such as the tree in Figure 1 or the graph in Figure 2. Typical discourse relations include *explanation*, *concession*, or *purpose*

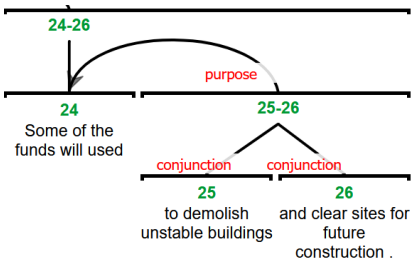

Figure 1: An RST tree example from RST-DT, visualized with rstWeb (Gessler et al., 2019).

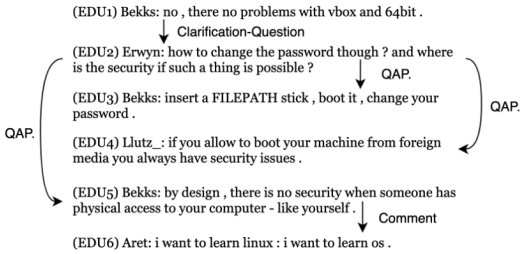

Figure 2: An SDRT graph (Liu and Chen, 2021).

as in Figure 1, but also relations more specific to dialogues such as *Question-Answer Pair* (QAP) in Figure 2. Discourse relations can be triggered by specific expressions, namely discourse connectives, such as *and* for the *conjunction* relation in Figure 1, the relation is then called *explicit*, in contrast with *implicit* relations, which are not explicitly marked.

Discourse relation extraction can be an end task in itself (e.g. find all concessions in a political speech), but discourse information has also been shown to be useful for other tasks, as demonstrated by studies on text style (Yang and Jin, 2023; Zhu et al., 2023), anxiety or emotion (Juhng et al., 2023; Zhang et al., 2023a), and propaganda identification (Chernyavskiy et al., 2024). In addition, discourse attracts renewed interest as current models struggle with long-text modeling and generation (Ivgi et al., 2023; Li et al., 2023; Liang et al., 2023; Feng et al., 2023; Buchmann et al., 2024; Wu et al., 2024a).

As in the last two editions of DISRPT, three

*Discourse Relation Parsing and Treebanking (DISRPT 2025) was held in conjunction with CODI-CRAC at EMNLP 2025 in Suhzou, China and Online (https://sites.google.com/view/disrpt2025/).

[1]The shared task data are also available on our GitHub, as well as the evaluation script: https://github.com/disrpt/sharedtask2025.

Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 1–20

tasks are proposed: **Task 1: discourse segmentation**—identifying the elementary discourse units (EDUs), or more precisely their starting tokens, that may be linked by discourse relations; **Task 2: discourse connective detection**—identifying specific lexical items, called connectives, that can signal a discourse relation (e.g. *while, because, as long as* etc.); **Task 3: discourse relation classification**—identifying a relation label between a pair of attached discourse units. In addition, tasks 1 and 2 have two tracks, depending on whether sentence boundaries and additional morpho-syntactic information is available (`Treebanked`) or not (`Plain`).

The DISRPT shared tasks emerged from the need to evaluate systems for automatic discourse analysis beyond the Penn Discourse Treebank (Prasad et al., 2014, 2019) and RST Discourse Treebank (Carlson et al., 2001), the two most used datasets for discourse relation or connective classification, and discourse segmentation or parsing. Both datasets consist of Wall Street Journal articles in English, from the same period of time. Limiting training/evaluation to these datasets thus greatly restricts applications and understanding of general discourse knowledge of models. Since 2019, the set of datasets composing the DISRPT benchmark has grown in size and representativeness in terms of frameworks, languages, domains, and genres.

But in 2023, with 26 datasets, two problems were raised. First, the total number of labels for relation classification was very large, despite some homogenization that allowed to reduce them from 350 to 191 (Braud et al., 2024). This high number of labels, with almost no overlap between frameworks, prevents efficiently combining the datasets and hinders the development of joint models. Second, the rapid development in NLP sees the emergence of powerful, but computationally expensive models, making the reproduction step, which is crucial in a shared task, more and more difficult, especially with many tasks and datasets. In this new edition, we thus attempt to solve the first issue by proposing **a unified set of 17 discourse labels**, where similar relations are grouped into coarse grained classes. In addition, we imposed two new constraints: each team had to propose **a single model per task** – versus one model per dataset as it was often the case in past editions – with **a limit on the number of parameters at 4B**.

This year, the benchmark has been expanded with 13 new datasets compared to 2023,[2] including datasets from two new frameworks: the ISO framework (Bunt and Prasad, 2016) and the Enhanced Rhetorical Structure Theory (eRST, Zeldes et al. 2025), and new languages (Polish, Czech, Nigerian Pidgin). We also included new dialogic data, with now six datasets including dialogues, vs. two in 2023, and we updated some existing datasets (see Section 4). In total, 39 datasets were made available across six frameworks and 16 languages in a unified format. In the last phase of the shared task, we released six surprise datasets including data for two new languages (Polish and Nigerian Pidgin) and a new framework (ISO). The benchmark also contains six out-of-domain (OOD) datasets for which only `dev`/`test` partitions were available.

Five teams participated in the shared task, with two teams including some of the organizers. Overall, three systems were proposed for Tasks 1 and 2, and five systems for Task 3. For the `Treebanked` track, DiscUT, from the MELODI team, ranked first on the EDU segmentation task and connective detection, with performance very close to the HITS team for the latter. For the `Plain` track, the SeCoRel system, from AU-KBC Research Centre, ranked first on EDU segmentation, and MELODI was first on connective detection. For relation classification, the DeDisCo system, from Georgetown University, ranked first. The results demonstrate that multilingual models are competitive compared to approaches relying on independent, language-specific models used in previous editions, but there is margin for improvements for all tasks, especially for low-resource languages.

2 Related Work

Automatic Discourse Analysis. This is an active domain of research, with many researchers moving toward processing of long-form documents and conversations and taking advantage of the capabilities of contemporary Pretrained Language Models (PLMs). Recent work has shown that discourse information is impactful in varied domains and tasks: In the Question Answering task, answers often require multiple sentences (Prasad et al., 2023; Xu et al., 2023; Zhang et al., 2024), and in summarization or text simplification, outputs must correctly relate to discourse and coreference links, but often fail to do so (Cripwell et al., 2023; Pu et al., 2023;

[2]Two datasets were already included in the DISRPT benchmark release (Braud et al., 2024).

Wu et al., 2024b; Zhang et al., 2023b; Chang et al., 2024; Li et al., 2025). In machine translation too, document-level translation has become an important challenge (Maruf et al., 2021; Pal et al., 2024), and new datasets and metrics are being developed to account for discourse phenomena (Fernandes et al., 2023; Jiang et al., 2023). The study of reasoning in Large Language Models (LLMs) also benefits from data analyzed at the discourse level, which remains challenging for models (Newman et al., 2023; Huang et al., 2024; Sprague et al., 2023; Kim et al., 2024).

The full task of discourse parsing involves identifying the minimal text segments–or Elementary Discourse Units (EDU)–to be linked (segmentation), then a recursive process involves an attachment step between pairs of discourse units (or groups of such units) and the labeling of the discourse relation between these nodes to create either a complete graph (SDRT, (e)RST, and discourse dependencies) or a sparse set of subgraphs (PDTB, ISO), optionally linked to textual triggers such as connectives (PDTB, ISO, eRST).

Most of the existing work on discourse parsing focuses on English, either for monologues (Maekawa et al., 2024) or dialogues (Thompson et al., 2024a), with also some systems developed for Chinese (Hung et al., 2020; Peng et al., 2022b). In order to better understand potential weaknesses or limits of these systems, a long line of work focuses on subtasks, such as segmentation (Marcu, 2000; Muller et al., 2019a) and discourse relation labeling (Dai and Huang, 2018; Xiang and Wang, 2023), but also connective identification, which can provide important clues for identifying discourse relations (Gopalan and Lalitha Devi, 2016; Yu et al., 2019). Again, these studies mostly focus on English, and multilingual or multi-domain comparisons are rare (Li et al., 2014; Liu and Zeldes, 2023; Metheniti et al., 2024). In addition, most work on discourse relation classification focuses on implicit relations (e.g. Liu and Strube 2023; Zhao et al. 2023), which are not triggered by a connective and are therefore harder to identify, thereby hindering our understanding of the difficulty of the task as a whole.

The DISRPT Shared Task. DISRPT was first organized in 2019, with only two tasks: segmentation and connective identification (Zeldes et al., 2019). The third task on relation identification was added in 2021 (Zeldes et al., 2021), and covered 16 datasets across 11 languages. For the last edition, in 2023 (Braud et al., 2023), the benchmark was composed of 26 datasets and 13 languages. In total, 11 teams participated over the three past editions, and additional experiments were presented on the DISRPT benchmark in Braud et al. (2024).

The aim of the shared task has been to promote cross-lingual and cross-framework discourse analysis. The handling of the multilingual aspect of the DISRPT benchmark was done either by using (1) monolingual representations, (2) multilingual representations with systems trained independently on each dataset, or (3) multilingual joint training. For the three past editions, the winning systems for all three tasks were based on option (1) or (2).

In particular, for discourse relations, the best system overall was the one proposed in Gessler et al. (2021), with scores computed on the extended benchmark in Braud et al. (2024): this system relies mostly on monolingual PLMs with additional linguistic features, and models were fine-tuned independently on each dataset. A few attempts have been made to group small datasets per framework, for example, the winning system in 2023 (Liu et al., 2023), or to jointly train over all datasets (Metheniti et al., 2023). Interestingly, one participating system proposed to introduce a relation hierarchy in order to help with label explosion (Varachkina and Pannach, 2021).

For segmentation and connectives, previously two of the winning systems used multilingual embeddings or PLMs, but still learning independent models (Muller et al., 2019b; Metheniti et al., 2023). Again, attempts have been made to group datasets by language families (Kamaladdini Ezzabady et al., 2021) or to transfer from one dataset to another (Dönicke, 2021).

For the 2025 edition, we decided to constrain participants to propose a single model, i.e., one set of parameters and hyper-parameters, that could be evaluated over all the datasets, thereby imposing a multilingual joint approach. Ensemble or pipeline approaches were allowed, as long as the total number of parameters did not exceed 4B parameters. Considering that the very high number of different relation labels was an important obstacle to joint learning, we mapped the annotated relations to a limited set of 17 labels (see Section 4.5).

Existing Mapping Proposals. Previous work has proposed various mappings across a subset of the frameworks and languages covered by DIS-

RPT (Chiarcos, 2012, 2014; Rehbein et al., 2016; Sanders et al., 2021), and applications of the mappings were also limited to a small number of corpora that either mainly contain news data or are primarily in English (Benamara and Taboada, 2015; Bunt and Prasad, 2016; Demberg et al., 2019; Costa et al., 2023). As a result, the generalizability of the proposed mappings is limited.

The ISO (Bunt and Prasad, 2016) proposal for annotation of semantic phenomena gives a set of 20 labels, as well as a mapping from some RST, SDRT, and PDTB corpora. Annotations rely on both a relation label and role labels for arguments, e.g. the *Question-Answer* relation corresponds to the ISO label *Functional Dependence* and a communicative function of *answer* for the second argument. On the other hand, Sanders et al. (2018) proposed to decompose relations into primitive concepts (e.g., *polarity, conditional*). These approaches are interesting, but have never been applied to the range of languages, domains, and frameworks included in DISRPT. In addition, adopting their formats would require substantial work and would change the format of the task too much within the scope of the DISRPT tracks. Moreover, our aim is not to produce annotation guidelines, but rather to allow for cross-framework investigation of the task. However, we took inspiration from the ISO standard when defining our own mapping.

Motivated by previous proposals and the need for generalization in NLP models or LLMs for discourse phenomena, Eichin et al. (2025) develop a unified set of 17 discourse relation labels that enables cross-lingual and cross-framework discourse analysis using the DISRPT 2023 shared task data (Braud et al., 2023), covering four frameworks (RST, PDTB, SDRT, and DEP) and 13 languages from 23 corpora. While the proposed unified label set in Eichin et al. (2025) is thorough, it does not cover the newly introduced framework and datasets. We thus propose a unified label set also taking inspiration from this proposal, but that differs in certain relation collapses. Section 4.5 presents and discusses the development of the unified label set.

3 Tasks and Tracks

This year, not all datasets have data annotated for discourse relation classification (Task 3): the French SUMM-RE dataset (Hunter et al., 2024; Prévot et al., 2025) is only annotated for segmentation. For connective detection (Task 2), only the datasets within the PDTB and ISO frameworks have annotations, while the others have annotations for discourse unit segmentation (Task 1).

For Tasks 1 and 2, two tracks were proposed:

- `Treebanked`: documents are split into sentences or speech turns, morpho-syntactic information and syntactic parses are provided – either gold when available or obtained from an automated tool;
- `Plain`: plain tokenized documents, without sentence split nor morpho-syntactic information. The tokenization is provided by the authors of the corpora.

In addition, we added two constraints for this edition: for each task, each team has to propose a single model that can be evaluated on all datasets, and the total number of parameters of the model should not exceed 4B. These constraints make the replication work more feasible as larger models can be too large for our computational capacity. More importantly, it allows to simplify the practical use of such a model and to evaluate the robustness of the proposed approaches.

4 DISRPT 2025 Data

4.1 Data Format

The shared task aims at providing an unified format across varied annotations projects. Three types of files are provided: the `conllu` and `tok` files contain the data for segmentation and connective identification in the CoNLL-U format with one line per token and the last column containing the label. The `conllu` files indicate both sentence and document boundaries, while the `tok` files have only the latter. The `rels` files correspond to the relation classification task, with one pair of discourse units per line, and additional information such as the corresponding sentences, the type of relation when available, the original relation name, and the DISRPT label in the last column. More information on the DISRPT format can be found in Braud et al. (2024).

4.2 Summary of the Datasets

DISRPT 2025 includes 39 datasets, where a dataset is a unique combination of a language, a framework, and a corpus name; a multilingual corpus such as TEDm thus corresponds to several datasets, one for each language. In total, six frameworks are represented, now including the new eRST framework (Zeldes et al., 2025) created as an extension

of RST, and the ISO framework (Tomaszewska et al., 2024). Data are available for 16 languages, compared to 13 in 2023, with new datasets for Czech, Polish, and Nigerian Pidgin. The datasets also vary in terms of genres and domains, still including news, wiki, or scientific documents, but also more conversations (LUNA, DiscoNaija, and SUMM-RE), online speech such as vlogs, podcasts, and eSports (GUM, GENTLE), and even medical, legal, and poetry writing (Basque RST-TB, GENTLE).

The increase in dialogue data raises issues on how to build the files used for segmentation or connective detection: the notion of sentence is often unclear in dialogues, and some datasets consider speech turns as a way to split documents into smaller units for the `Treebanked` track (i.e. the `conllu` files). For SUMM-RE (Hunter et al., 2024; Prévot et al., 2025), a corpus of spoken dialogues, we discussed with the authors to find an optimal way of splitting the dialogues. One issue is how to deal with back-channeling elements (e.g. *mm*), as they were transcribed, but usually overlapped the other speaker's turn. Corpus creators suggested a fixed list of short turns overlapping longer turns.

We provide general statistics of all the datasets in Table 5 and statistics based on the data partitions in Table 6 in Appendix A.

4.3 Dataset Updates

Compared to the last release of the benchmark in 2024 (Braud et al., 2024), we have implemented several updates to the datasets. These changes limit direct comparisons but are crucial to maintain high-quality data. One important change concerns the Russian RST dataset (`rus.rst.rrt`): it has been substantially reduced, as an author indicated that the Science section contained faulty annotations, and this part of the dataset was thus removed.

Another important change is the modification of the English PDTB v3 splits. This change is motivated by the overlap between the English PDTB v3 and the English RST-DT train/test sections: two corpora annotated over a common set of the Wall Street Journal articles. Since we constrain participants to jointly train on all datasets, maintaining a split where test files from one are present in the other's training set was not possible.

More precisely, we decided to follow the partition proposed in RST-DT and use the same for PDTB v3. It does not lead to the exact same set of files in each split, since PDTB v3 contains more articles than RST-DT, but also four files annotated within the RST-DT do not appear in the PDTB v3. The RST-DT partitions are not based on sections, as for PDTB v3, but articles from different sections are mixed within each set. Using the exact same set of files as in RST-DT to build PDTB v3 `dev` and `test` sets is not enough: we thus add all files from sections 21 and 22 to the PDTB v3 `test` set, as these sections were used as `test` in previous studies (i.e. the so-called Ji and Eisenstein split, Ji and Eisenstein 2015); all files from sections 00 and 24 are used as development, 24 being usually used as `dev` while 00 is generally ignored. Our final partition for PDTB v3 is shown in Table 1: it leads to a larger evaluation set, thus making for a more robust evaluation, and has no conflicts with RST-DT. The exact composition of each split is available on GitHub.[3]

Split	DISRPT23		DISRPT25	
	# tokens	# files	# tokens	# files
train	1,061,229 / 91.75%	1,992 / 92.14%	961,757 / 83.15%	1,805 / 83.49%
dev	39,768 / 3.44%	79 / 3.65%	96,068 / 8.31%	177 / 8.19%
test	55,660 / 4.81%	91 / 4.21%	98,832 / 8.54%	180 / 8.33%

Table 1: Comparison of the distributions of the train / dev / test splits for the English PDTB v3 in DISRPT 2023 and 2025.

Other minor changes were also necessary: GUM, which grows annually, was updated to its latest version; the Thai corpus was reparsed by the authors; the STAC corpus was reprocessed entirely based on its lastest version; For the English PDTB v3, some missing relations, or relations with a wrong type, were added back; for the Basque RST dataset, one relation with the label *definitu-gabeko erlazioa* ('undefined') was removed because we were unable to find its definition; the Chinese GCDT was reparsed; for the Italian LUNA dataset, speech turn segmentation was corrected, the entire dataset was reparsed, and all instances of the relation *Interrupted* were removed, as they only involve one argument. For the DiscoNaija dataset, we found some errors in the annotations where arguments were overlapping, and thus these examples were not included in the current version of the dataset (130 explicit or implicit instances ignored in total).

4.4 Segmentation and Sentence Splitting

Comparing the beginning of sentences in the `conllu` files and the label indicating the beginning of an EDU, we found a large number of instances

[3] https://github.com/disrpt/sharedtask2025

(27.27%) where the start of a new sentence is not annotated as a new segment in the English STS corpus (`eng.rst.sts`). Having examined these cases, we found that, in some documents, the corpus has very long, multi-sentence segments, that might be longer than 2000 tokens. We expect large error rates for this dataset on the segmentation task, and systems could struggle when trying to identify relations with very long arguments.

Other cases of discrepancies come from errors of the automated tools used to segment into sentences, as in the previous edition. The datasets containing errors of this type are: ANNODIS (`fra.sdrt.annodis`, 6.12%), the Basque ERT (`eus.rst.ert`, 3.01%), the English OLL (`eng.rst.oll`, 1.91%), the Russian RRT (`rus.rst.rrt`, 1.26%), and the Portuguese CSTN (`por.rst.cstn`, 0.39%). We plan to provide new sentence splitting for these datasets, using updated tools, for next editions.

4.5 A Unified Set of Relation Labels

For DISRPT 2025, we propose a unified set of labels, in order to push forward the development of cross-framework and cross-lingual systems for relation classification. The choice of the labels is inspired by previous cross-framework mapping proposals (See Section 2), but we cannot adopt any of the existing ones directly. Since we have to integrate all the existing DISRPT datasets as well as the new ones, we are forced to take into account the variety of granularity: for example, some datasets have a vague *Temporal* label, and we would need to reannotate the data in order to keep the finer-grained distinction between synchronous and asynchronous relations existing in many datasets.

Moreover, since different annotation projects made different choices in what counts as a discourse relation, it is clear that some labels will not be represented in some datasets. For example, *Attribution* is annotated in RST-style corpora, but it is not considered a relation in the PDTB-style ones: we need to keep this label, which corresponds to a clear definition that could not be merged with other types of relation, and thus some datasets will have this label missing. In a similar vein, there are relations defined for dialogic phenomena, such as *Question-answer*, while some monologic datasets also include similar relations, e.g. *Hypophora* in PDTB v3, and these relations could be mixed with labels less specific to dialogs, such as *Solutionhood*. It is however clear that the distribution for such a class will be very different between monologic and dialogic datasets.

The final mapping has been established by five experts in discourse after discussions considering all the 306 different labels found in the DISRPT 2023 data. They then checked that this set was also able to integrate labels for datasets added in 2025, and in the end all the new relations were possible to integrate. They also considered the coverage of this label set, aiming at having most of the labels represented in all the datasets. The final 17 labels are: ALTERNATION, ATTRIBUTION, CAUSE, COMMENT, CONCESSION, CONDITION, CONJUNCTION, CONTRAST, ELABORATION, EXPLANATION, FRAME, MODE, ORGANIZATION, PURPOSE, QUERY, REFORMULATION, and TEMPORAL. All datasets contain between 9 and 17 labels, eight datasets have the whole 17 labels represented, 18 have 15 labels or more. The final labels are shown in Table 7 in Appendix B with examples of corresponding original labels from different datasets.

In the end, our mapping is rather close to the ISO standard: ATTRIBUTION, FRAME, COMMENT, and ORGANIZATION were added, and the first two have explicitly no corresponding mapping in Bunt and Prasad (2016), while the last two cover relations that seem to be considered as *elaboration* in ISO; TEMPORAL covers the finer-grained distinction between *synchrony* and *asynchrony*, a choice dictated by the variety of annotations in DISRPT.[4] The same goes for the distinction between *condition* and *negative-condition* that could not be kept.

Compared to Eichin et al. (2025), we reorganized their *structuring* class, considering that relation labels such *joint* or *list* should be together (CONJUNCTION), and separated from relations such as *alternation* or *disjunction* (ALTERNATION), and from relations describing some textual organization such as *preparation*, *progression*, *summary* or *heading* (ORGANIZATION). We also kept the distinction between CONTRAST and CONCESSION which is well-established in the datasets. We restricted the COMMENT class to commentaries, and defined a QUERY class to cover several dialog phenomena (e.g. *acknowledgment, clarification question*) but also relations that can be found in monologues, such as *interpretation*, *evaluation* or *problem-solution*.

[4] Corpora with no temporal distinctions: `eng.dep.covdtb`, `zho.pdtb.cdtb`, `por.pdtb.crpc`, `zho.dep.scidtb`, `eng.dep.scidtb`, `fas.rst.prstc`.

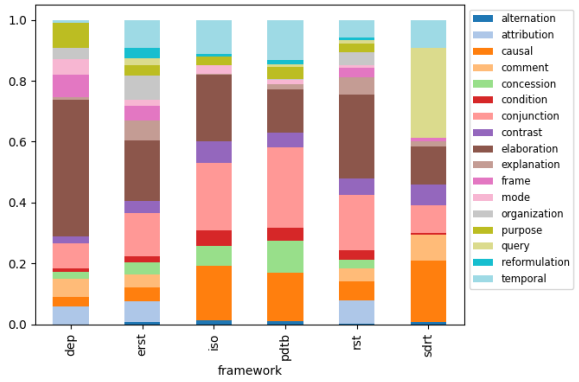

Figure 3: Distribution of the unified label set per framework in the DISRT 2025 datasets.

The final mapping is not completely satisfying, in the sense that several fine-grained distinctions are lost, but we believe that it is necessary if one wants to investigate discourse in a cross-framework setting. Note that, however, the DISRPT format does keep the direction of the relation as an additional feature of the relation, allowing to retrieve, for example, distinctions between *cause* and *result* relations. In addition, Figure 3 shows that, overall, the label distribution is still unbalanced, and the distribution is framework dependent.

5 Participating Systems

CLaC: The CLaC team from Concordia University participated in Task 3: discourse relation classification. Their baseline systems include fine-tuning multilingual PLMs based on Transformers with different encoding of the direction of the relation and different amounts of frozen layers, and prompting a generative model in zero- and few-shot settings. Their best system, called HiDAC (Hierarchical Dual-Adapter Contrastive), achieves the highest performance while relying on a parameter-efficient fine-tuning strategy: the backbone PLM is split into two parts: the lower layers learning representations of the arguments using LoRA (Hu et al., 2022) and a contrastive loss, and the upper ones learning task-specific representations using Mixture-of-Experts LoRA adapters and cross-entropy loss. The whole model is optimized based on a combination of the two losses. In their paper, they report experiments on the public part of the benchmark, excluding the datasets under restrictive licenses. Their model was retrained on all datasets by the organizers during the reproducution phase. Their strategy allows to obtain similar or even better results than full fine-tuning, at a lower computational cost. On the other hand, the prompt-based approaches underperformed compared to fine-tuning.

DeDisCo: The DeDisCo team from Georgetown University also participated only in Task 3. After experimenting with both the encoder-decoder (mt5-based) and decoder-only approaches, the team opted for the latter and used Qwen3-4B (Yang et al., 2025) as a base model. Since the model was just over the maximum parameter count (4.02B), the team first distilled a version with under 4B parameters using the layer pruning approach proposed by Men et al. (2024). They then experimented with different prompts and finally selected a strategy incorporating not only dataset and language specific encoding, but also relation direction, argument and context delimitation, and linguistic feature encoding (a subset of the 'DisCoDisCo' features from Gessler et al. 2021). The final system was fine-tuned end-to-end on all training datasets, which were supplemented by data augmentation on some of the smaller languages. Specifically, they machine-translated the most similar English datasets to six smaller language datasets in Basque, Czech, Dutch, French, German, and Persian in order to create more training data in those languages. Their paper contains an ablation analysis of each of these components (augmentation and different kinds of features) on each dataset.

DiscUT and DiscReT: The MELODI team presented systems for all the three tasks, DiscReT (Discourse Relation Tagger) for relation classification and DiscUT (Discourse Unit Tagger) for segmentation and connective identification for both tracks. The models all rely on an architecture based on Transformers, with a multilingual PLM fine-tuned on all the datasets. For the `Plain` track, documents were segmented using SaT (Frohmann et al., 2024). They experimented with different ways of combining the data by language groups or frameworks, in order to allow the models to gradually learn from more similar groups of annotations, using sequential fine-tuning: the model is first fine-tuned on a specific group, then the fine-tuning continues on another group. They also introduced features representing the framework and the language, and, for relation classification, the direction of the relation and its locality. For all tasks, best performance was obtained when training on the full concatenation of all datasets and all features except locality, using XLM-RoBERTa-large for relation classification, and InfoXLM for the other tasks.

HITS: The HITS team participated in all three tasks with distinct systems. For Task 1, the team fine-tuned mT5-xl (3.7B parameters) using LoRA, then applied weighted loss to compensate for class imbalances (most tokens are not segmentation points) and adverserial training using the Fast Gradient Method (FGM) to boost robustness. For Task 2, the team combined three multilingual encoders (RemBERT, XLM-RoBERTa and mDeBERTa-v3), integrating POS tags and dependency features, and using a CRF layer with a focal loss and label smoothing to combat label imbalance. Finally for Task 3, the team took a two-stage approach, using Rationale-Enhanced Curriculum Learning, using a gemma-2-2b student model to output `json` representations of labels with LoRA fine-tuning, and then using a much larger Qwen2.5-72B-Instruct model as a tutor, which was used to extract verbal rationales for cases the learner model failed to classify correctly. These rationales were fed back to the student learner in a second training procedure to produce the final model.

SeCoRel: The SeCoRel team, from the AU-KBC Research Center, participated in all three tasks and both tracks. For all three tasks, they proposed an approach relying on the fine-tuning of a multilingual PLM based on the Transformer architecture of a relatively small size (XLM-RoBERTa base) and optimized the hyper-parameter values. In order to deal with the `Plain` track, the documents were segmented into sentences using heuristic rules that are not detailed in the paper.

6 Results

Results for each track/dataset are in Tables 2–4.

Task 1: Discourse Segmentation (Table 2). For discourse segmentation on the `Treebanked` track, the best results were obtained by the MELODI team (DiscUT) with at best 91.57 mean F_1 over the 39 datasets. As shown in the previous editions, performance on this task seems to have reached a plateau, with a similar score in 2023 (91.87 F_1). These scores tend to demonstrate that a joint approach can be as effective as models trained separately on each datasets, and that the newly introduced datasets are not harder than the existing ones. However, when looking at the scores in detail, we observe a decrease of around 2 points for GUM and STAC and 1 point for the RST-DT, and the Basque ERT. On the other hand, some datasets seem to benefit from the joint training, such as the Portuguese CSTN (+1.8) and the Dutch NLDT (+1.4 point). For the Russian RRT, the removal of a problematic section greatly improves the score (92.50 against 85.58 in 2023). In addition, some scores are still under 90%, e.g. for the new French SUMM-RE dataset, but also for datasets included for a longer time, such as the French ANNODIS, the Spanish SCTB or the Chinese SciDTB and SCTB, datasets which future work should study more.

We observe some variance in the scores, with very close performance reported for MELODI and HITS, and up to 1.4 points of increase during reproduction. Unfortunately, we were not able to report on several runs, due to time and computational constraints, but it would be important to test this variance more thoroughly for all the three tasks.

For `Plain`, only two teams participated, and SeCoRel ranked first with 87.38 mean F_1. The difficulty of this track is that documents are not split into sentences, but PLMs all have input size limits, preventing use of the full documents as inputs. Both teams pre-processed data to split documents into sentences: SeCoRel used heuristic rules while DiscUT (MELODI) relied on a pre-trained model. The results demonstrate that the tool used did not make a big difference, and that heuristics can perform even better, though scores are clearly lower compared to the `Treebanked` track (87.38 against 91.57), indicating some issues with the sentence splitting.

When comparing the individual results between the two tracks, we can see a large drop in performance for some datasets. For the MELODI system, this drop happens either for low-resource languages / domains or smaller datasets: the Chinese SCTB (-19.7) and SciDTB (-16), the English STS (-8), GENTLE (-6.4) and OLL (-3.7), the Spanish SCTB (-5.1), the Portuguese CSTN (-4.1), the Basque ERT (-3.8), the Czech CRDT (-3.1); for dialogues datasets for which the tool is not adapted: the French SUMM-RE (-14.9), the English MSDC (-10.6) or STAC (-5); and for datasets containing initially gold sentences: the English RST-DT (-2.85). Future research should focus on this more practical setting by improving pre-processing or evaluating solutions to take a larger context into account.

Task 2: Connective Identification (Table 3). For this task, MELODI ranked first in both tracks, but results are very similar between MELODI and

	Treebanked Track									Plain Track					
	DiscUT (MELODI)			HITS			SeCoRel (AU-KBC)			SeCoRel (AU-KBC)			DiscUT (MELODI)		
Dataset	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ces.rst.crdt	94.26	91.92	**93.08**	94.27	91.92	**93.08**	92.50	91.93	92.21	90.18	91.30	**90.74**	86.28	93.78	89.88
deu.rst.pcc	97.87	93.89	**95.84**	94.08	91.52	92.78	94.85	93.56	94.20	94.86	93.90	**94.38**	92.18	95.93	94.01
**eng.dep.covdtb	91.66	93.52	**92.58**	88.99	94.78	91.80	86.46	94.10	90.12	86.81	93.93	90.23	88.97	95.45	**92.10**
eng.dep.scidtb	95.05	95.46	**95.26**	94.42	95.13	94.77	94.00	94.75	94.38	94.00	94.75	94.38	93.77	95.46	**94.61**
eng.erst.gentle	94.30	93.88	**94.09**	94.75	88.33	91.43	93.46	83.65	88.28	92.51	75.07	82.89	88.07	87.29	**87.68**
eng.erst.gum	95.38	92.07	**93.70**	95.86	90.17	92.93	93.70	87.07	90.27	93.58	90.05	**91.78**	91.38	91.31	91.34
eng.rst.oll	92.17	89.93	**91.03**	83.71	90.97	87.19	78.59	89.24	83.58	78.79	90.28	84.14	84.19	90.62	**87.29**
eng.rst.rstdt	97.32	96.07	96.69	96.59	97.78	**97.18**	95.84	94.37	95.10	95.38	93.35	**94.36**	92.71	94.97	93.83
eng.rst.sts	86.74	89.58	**88.14**	79.44	83.93	81.62	78.65	83.33	80.92	65.10	69.94	67.43	80.48	79.76	**80.11**
eng.rst.umuc	93.65	84.70	**88.95**	86.03	88.34	87.17	86.04	87.19	86.61	86.04	87.19	86.61	87.23	87.57	**87.40**
eng.sdrt.msdc	96.99	94.65	**95.81**	96.95	93.84	95.37	96.10	93.46	94.76	96.33	93.06	**94.67**	93.58	78.24	85.23
eng.sdrt.stac	90.75	95.23	92.93	91.64	94.02	**92.81**	87.72	95.32	91.36	85.33	90.73	87.95	83.50	92.98	**87.98**
eus.rst.ert	92.45	89.45	**90.93**	90.04	91.62	90.82	87.76	90.14	88.93	88.06	89.73	**88.89**	86.43	87.83	87.13
fas.rst.prstc	93.75	94.17	**93.96**	93.15	93.28	93.21	91.64	94.93	93.26	91.63	94.78	**93.18**	92.33	91.64	91.98
fra.sdrt.annodis	89.42	87.54	**88.47**	87.87	86.73	87.30	85.96	79.29	82.49	89.41	80.58	84.77	89.64	86.89	**88.24**
fra.sdrt.summre	93.84	84.10	**88.70**	62.23	63.05	62.64	56.83	90.16	69.71	57.45	97.04	72.17	75.44	72.38	**73.84**
nld.rst.nldt	99.09	96.74	**97.90**	95.55	95.27	95.41	96.76	97.04	96.90	97.04	96.08	**96.56**	93.96	96.74	95.33
por.rst.cstn	95.45	96.07	**95.76**	94.82	95.75	95.28	92.74	96.08	94.38	92.74	90.22	91.46	87.05	96.73	**91.64**
rus.rst.rrt	94.84	90.27	**92.50**	93.67	91.36	**92.50**	92.43	91.64	92.03	92.22	91.60	**91.91**	90.51	90.09	90.30
spa.rst.rststb	92.24	93.04	**92.64**	91.84	93.04	92.44	92.46	90.65	91.55	92.00	90.89	91.44	90.31	93.26	**91.74**
spa.rst.sctb	88.75	84.52	86.58	85.55	88.09	**86.80**	87.42	82.74	85.02	86.54	80.36	**83.33**	79.21	83.92	81.50
zho.dep.scidtb	80.34	99.14	**88.76**	80.07	97.45	87.91	83.15	94.47	88.45	83.15	94.47	**88.45**	57.75	98.29	72.75
zho.rst.gcdt	91.96	86.99	89.41	92.18	88.46	**90.28**	87.32	92.67	89.92	86.82	91.12	**88.92**	85.75	89.28	87.48
zho.rst.sctb	60.29	95.83	**74.02**	56.63	94.05	70.69	54.14	93.45	68.56	53.24	88.10	**66.37**	38.16	94.04	54.29
mean	91.61	92.03	**91.57**	88.35	90.79	89.31	86.94	90.88	88.46	86.22	89.52	**87.38**	84.54	90.18	86.57
in paper	-	-	90.19	-	-	90.09	-	-	86.36[5]	-	-	88.00[5]	-	-	86.89

Table 2: **Results for Task 1: discourse segmentation, `Treebanked` and `Plain` tracks.** The table contains the reproduced scores per dataset and average, and we also report the scores for the system's paper when available.

	Treebanked Track									Plain Track					
	DiscUT (MELODI)			HITS			SeCoRel (AU-KBC)			DiscUT (MELODI)			SeCoRel (AU-KBC)		
Dataset	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
deu.pdtb.pcc	78.00	82.97	80.41	93.92	92.27	**93.09**	84.69	88.30	86.46	79.38	81.91	80.62	84.69	88.30	**86.46**
***eng.pdtb.gentle	90.36	84.54	**87.36**	86.12	77.57	81.62	89.41	85.19	87.25	90.56	82.40	86.29	87.17	86.05	**86.61**
eng.pdtb.gum	91.34	81.92	**86.37**	81.91	80.30	81.09	87.64	74.92	81.78	90.44	81.20	**85.57**	87.89	72.38	79.38
eng.pdtb.pdtb	94.17	93.50	**93.83**	93.56	79.17	85.76	92.05	89.47	90.74	95.33	92.41	**93.84**	87.58	86.30	86.93
**eng.pdtb.tedm	83.82	74.02	78.62	79.78	75.53	77.60	84.24	74.03	**78.80**	86.36	74.02	**79.72**	84.24	74.03	78.80
ita.pdtb.luna	65.15	71.64	68.24	87.51	88.62	**88.06**	47.90	74.33	58.26	65.72	62.45	**64.04**	47.63	73.18	57.70
pcm.pdtb.disconaija	84.07	77.57	**80.69**	72.27	60.92	66.11	72.25	71.13	71.69	76.60	80.15	**78.33**	73.73	70.88	72.27
pol.iso.pdc	72.53	70.65	71.58	94.35	92.46	**93.40**	68.76	71.83	70.26	70.41	67.60	68.98	68.37	72.07	**70.17**
por.pdtb.crpc	81.59	79.04	**80.29**	93.02	48.58	63.83	80.41	78.49	79.44	83.22	73.49	78.05	79.70	77.94	**78.81**
**por.pdtb.tedm	82.77	85.22	**83.98**	85.00	73.59	78.89	78.30	81.77	80.00	79.70	79.31	79.50	78.04	82.27	**80.10**
tha.pdtb.tdtb	88.95	92.71	**90.79**	75.34	65.26	69.94	81.84	84.15	82.98	87.81	91.89	**89.80**	78.95	83.47	81.15
tur.pdtb.tdb	91.96	95.19	**93.55**	82.14	77.06	79.52	80.33	80.20	80.26	90.40	93.79	**92.06**	87.84	78.46	82.89
**tur.pdtb.tedm	92.80	52.22	66.83	91.64	81.72	**86.39**	63.01	39.68	48.69	91.48	52.22	**66.49**	80.33	39.68	53.12
zho.pdtb.cdtb	91.43	75.32	82.60	92.43	83.91	**87.96**	70.68	55.49	62.17	88.93	74.67	**81.18**	86.47	57.37	68.98
zho.pdtb.ted	75.03	80.98	**77.89**	77.03	75.30	76.15	63.01	68.86	65.81	71.66	77.24	**74.35**	58.75	67.59	62.86
mean	84.26	79.83	**81.54**	85.73	76.82	80.63	76.30	74.52	74.97	83.20	77.65	**79.92**	78.09	74.00	75.08
in paper	-	-	80.11	-	-	81.00	-	-	72.32[5]	-	-	79.79	-	-	71.98[5]

Table 3: **Results for Task 2: Discourse Connective Detection, `Treebanked` and `Plain` Tracks.** The Table contains the reproduced scores per dataset and average, and we also report the average score for the system's paper.

HITS for `Treebanked`. It is interesting to note that when one system is better than the other on a dataset, it is often by a very large margin, e.g. +21.8 for HITS on the Polish PDC, +19.8 on the Italian LUNA, +19.6 on the Turkish TEDm, but +20.8 for MELODI on Thai TDTB, +16.5 on Portuguese CRPC, and +14.6 on Nigerian Pidgin DiscoNaija. This indicates that these systems operate differently and some sort of combination could be beneficial and should be investigated.

The overall best average F_1 on the `Treebanked` track is 1 point better than in 2023 (81.54 against 80.47), and is very similar for the `Plain` track

[5]This system was not originally trained and evaluated on the licensed datasets, explaining differences between the reported and reproduced scores.

Dataset	DeDisCO	HITS	DiscReT	CLAC	SeCoRel
ces.rst.crdt	**56.08**	53.38	47.97	47.97	43.92
deu.pdtb.pcc	**67.53**	63.92	63.92	63.92	53.61
deu.rst.pcc	**64.10**	59.71	52.75	46.52	47.25
*eng.dep.covdtb	**71.46**	71.31	69.22	70.46	65.27
eng.dep.scidtb	**84.29**	81.78	78.22	80.31	78.22
*eng.erst.gentle	**68.30**	62.42	53.53	54.00	50.08
eng.erst.gum	**76.50**	67.32	64.21	62.14	58.81
*eng.pdtb.gentle	**67.30**	64.89	64.25	63.10	55.47
eng.pdtb.gum	**73.48**	67.88	69.31	66.15	63.71
eng.pdtb.pdtb	**83.54**	79.95	75.06	76.11	70.43
*eng.pdtb.tedm	**68.95**	64.96	61.54	61.54	57.83
eng.rst.oll	**62.73**	58.30	47.23	53.87	46.49
eng.rst.rstdt	**73.09**	64.92	60.93	64.08	60.46
eng.rst.sts	**54.27**	**54.27**	42.68	41.77	36.28
eng.rst.umuc	**65.91**	63.84	59.09	56.20	56.82
eng.sdrt.msdc	**90.00**	89.60	85.64	85.79	85.03
eng.sdrt.stac	**77.04**	75.89	69.50	69.68	67.91
eus.rst.ert	50.10	54.02	**54.43**	50.93	52.58
fas.rst.prstc	59.29	**59.80**	57.60	55.41	52.20
fra.sdrt.annodis	**60.06**	57.00	57.97	53.78	55.23
ita.pdtb.luna	**72.00**	68.53	66.67	60.27	60.00
nld.rst.nldt	**67.38**	64.92	59.69	56.31	54.15
pcm.pdtb.disconaija	59.88	**60.37**	57.72	56.34	56.05
pol.iso.pdc	**72.01**	**72.01**	60.03	54.78	52.76
por.pdtb.crpc	78.61	76.12	**79.09**	76.28	74.12
*por.pdtb.tedm	**70.33**	65.11	65.66	62.91	61.81
por.rst.cstn	**71.32**	70.22	68.01	69.12	63.60
rus.rst.rrt	**73.93**	72.58	66.68	66.43	62.49
spa.rst.rststb	**69.25**	65.49	61.50	58.22	57.04
spa.rst.sctb	**80.50**	74.21	67.92	66.04	63.52
tha.pdtb.tdtb	**97.10**	95.68	97.02	96.95	96.28
tur.pdtb.tdb	**68.65**	66.03	65.80	61.76	56.29
*tur.pdtb.tedm	58.68	59.50	**60.88**	60.06	57.02
zho.dep.scidtb	**75.35**	70.23	69.77	73.49	66.05
zho.pdtb.cdtb	**89.97**	81.79	77.57	82.32	73.35
zho.pdtb.ted	**75.64**	70.75	67.74	64.14	58.80
zho.rst.gcdt	**75.13**	71.46	61.91	62.54	58.87
zho.rst.sctb	**75.47**	57.86	55.35	60.38	54.09
mean	**71.19**	67.84	64.32	63.48	60.10
in paper	71.28	66.78	64.01	67.46[5]	55.29[5]

Table 4: **Reproduced results for Task 3: Discourse Relation Classification.** The Table also contains average scores as reported in each system's paper.

(79.92 against 79.36). It demonstrates once again the effectiveness of the joint approach. However, many datasets still obtain a performance lower than an average F_1 of 80, again for small or OOD datasets (English, Chinese), or low-resource languages / domains (Turkish TEDm, Italian LUNA, Polish PDC), indicating room for improvement.

As for segmentation, performance is lowered for the `Plain` track, but the drop is less significant, the sentence segmentation being less relevant for this task. For some datasets, performance is a bit higher within this track (e.g. English TEDm), but we also observe large drops (e.g. Italian LUNA, Polish PDC, Turkish TEDm), which requires a more detailed error analysis for future improvements.

Task 3: Relation Classification (Table 4). For this task, DeDisCo ranked first, with a mean accuracy of 71.28, much higher than the performance obtained by the best system in 2023 (62.36) or the best scores reported in 2024 (62.21), although this is not an apples-to-apples comparison as the label set is different. We note that the scores for almost all teams are also above the previous results, suggesting that joint learning is effective for the task. The experiments presented in the DeDisCo paper demonstrate the effectiveness of the decoder-only architecture with instruction learning, and the ablation study shows that some of the features used really boost the performance (especially the direction and context). The results on data augmentation are less clear, with increased performance for some target languages but not all of them, while it improved the results of source datasets overall. The model proposed in the end has a rather high computational cost, nearly fully using the allowed 4B parameters, and future studies could investigate additional methods to lower this cost while relying on large generative models.

We observe that scores are still low, under or just above 60% for several datasets: English STS, Basque ERT, Farsi PRSTC, Turkish TEDm, French ANNODIS and Nigerian Pidgin DiscoNaija. Most of these datasets correspond to small datasets or to a low-resource language, e.g., Nigerian Pidgin is close to English but with many lexical and syntactic differences, and there are likely almost no documents in this language included in the pretraining data of the PLMs used here, demonstrating that future effort should focus on this issue. For English STS, the problem could come from very long arguments, spanning multiple sentences, that may require a special processing or handling.

7 Conclusion

In this paper we present the data, systems, and results for the 2025 edition of the DISRPT shared tasks on discourse relation segmentation, classification and connective detection. The 2025 edition advances multilingual processing of discourse by providing new data, launching a new unified label set, and proposing a single-model, multilingual setup for each track. With five teams participating and a range of new SOTA scores, we are looking forward to applications using models from the shared task, and to proposals to further develop the benchmark in future tasks in the coming years.

Acknowledgements

We would like to thank Souvik Banerjee, Robin Pujol, Abhishek Purushothama, Firmin Rousseau,

Jingni Wu, and Zhuoxuan Nymphea Ju who helped with the system reproduction, Kate Thompson, Elena Chistova, and Peter Bourgonje for corpora preparation discussion, as well as all the authors of the corpora used in DISRPT for their help in converting their data into the shared task format.

This work is partially supported by the AnDiaMO project (ANR-21-CE23-0020) and the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as part of France's "Investing for the Future — PIA3" program.

Chloé Braud and Philippe Muller are part of the programme DesCartes and are also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

Chuyuan Li acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

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

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'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of Starsem*.

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.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

Jan Buchmann, Max Eichler, Jan-Micha Bodensohn, Ilia Kuznetsov, and Iryna Gurevych. 2024. Document structure in long document transformers. In *Proc. EACL 2024*, pages 1056–1073, St. Julian's, Malta.

Harry Bunt and Rashmi Prasad. 2016. Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54.

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*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Yi Cheng and Sujian Li. 2019. 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.

Alexander Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2024. Unleashing the power of discourse-enhanced transformers for propaganda detection. In *Proc. EACL 2024*, pages 1452–1462, St. Julian's, Malta.

Christian Chiarcos. 2012. Towards the unsupervised acquisition of discourse relations. In *Proceedings of ACL*.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

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*, pages 344–352, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings of ACL 2023*, pages 13190–13206, Toronto.

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.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Vera Demberg, Merel Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10:87–135.

Tillmann Dönicke. 2021. Delexicalised multilingual discourse segmentation for DISRPT 2021 and tense, mood, voice and modality tagging for 11 languages. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 33–45, Punta Cana, Dominican Republic. 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.

Shangbin Feng, Zhaoxuan Tan, Wenqian Zhang, Zhenyu Lei, and Yulia Tsvetkov. 2023. KALM: Knowledge-aware integration of local, document, and global contexts for long document understanding. In *Proc. ACL 2023*, pages 2116–2138, Toronto.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proc. ACL 2023*, pages 606–626, Toronto.

Igor Frohmann, Markus Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. 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).

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luke Gessler, Yang Liu, and Amir Zeldes. 2019. A discourse signal annotation system for RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 56–61, Minneapolis, MN. Association for Computational Linguistics.

Sindhuja Gopalan and Sobha Lalitha Devi. 2016. BioDCA identifier: A system for automatic identification of discourse connective and arguments

from biomedical text. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 89–98, Osaka, Japan. The COLING 2016 Organizing Committee.

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 *ICLR*. OpenReview.net.

Kung-Hsiang Huang, Philippe Laban, Alexander Richard Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proc. NAACL 2024*, pages 570–593.

Shyh-Shiun Hung, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. A complete shift-reduce Chinese discourse parser with robust dynamic oracle. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Online. Association for Computational Linguistics.

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.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proc. ACL 2023*, pages 7853–7872, Toronto.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H. Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proc. ACL 2023*, pages 1500–1511, Toronto.

Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual discourse segmentation and connective identification: MELODI at disrpt2021. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zae Myung Kim, Kwang Hee Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of subtlety: Detecting machine-generated texts through discourse motifs. *arXiv preprint arXiv:2402.10586*.

Chuyuan Li, Austin Xu, Shafiq Joty, and Giuseppe Carenini. 2025. Topic-guided reinforcement learning with llms for enhancing multi-document summarization. *arXiv preprint arXiv:2509.09852*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 577–587, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Xianming Li, Zongxi Li, Xiaotian Luo, Haoran Xie, Xing Lee, Yingbin Zhao, Fu Lee Wang, and Qing Li. 2023. Recurrent attention networks for long-text modeling. In *Findings of ACL 2023*, pages 3006–3019, Toronto.

Xiaobo Liang, Zecheng Tang, Juntao Li, and Min Zhang. 2023. Open-ended long text generation via masked language modeling. In *Proc. ACL 2023*, pages 223–241, Toronto.

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 *Proc. ACL 2023*, pages 15696–15712, Toronto.

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

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proc EACL 2024*, pages 2803–2815, St. Julian's, Malta.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *Preprint*, arXiv:2403.03853.

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 Metheniti, Philippe Muller, Chloé Braud, and Margarita Hernández Casas. 2024. Zero-shot learning for multilingual discourse relation classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17858–17876, Torino, Italia. ELRA and ICCL.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019a. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019b. Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking (NAACL)*.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proc. EMNLP 2023*, pages 3194–3212.

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.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

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.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proc. ACL 2023*, pages 15000–15025, Toronto.

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, 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*.

Dongqi Pu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proc. ACL 2023*, pages 5574–5590, Toronto.

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

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the pdtb and ccr frameworks. In *Proceedings of LREC*.

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.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.

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*.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024a. Llamipa: An incremental discourse parser. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics.

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. 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.

Aleksandra Tomaszewska, Purificação Silvano, António Leal, and Evelin Amorim. 2024. ISO 24617-8 applied: Insights from multilingual discourse relations annotation in English, Polish, and Portuguese. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 99–110, Torino, Italia. ELRA and ICCL.

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).

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)*, pages 46–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024a. Spinning the golden thread: Benchmarking long-form generation in language models. *arXiv preprint arXiv:2409.02076*.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024b. Less is more for long document summary evaluation by LLMs. In *Proc. EACL 2024*, pages 330–343, St. Julian's, Malta.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Comput. Surv.*, 55(12).

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proc. ACL 2023*, pages 3225–3245, Toronto.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Dingyi Yang and Qin Jin. 2023. Attractive storyteller: Stylized visual storytelling with unpaired text. In *Proc. ACL 2023*, pages 11053–11066, Toronto.

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 133–143, Minneapolis, MN. Association for Computational Linguistics.

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, 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, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 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).

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023a. DualGATs: Dual graph attention networks for emotion recognition in conversations. In *Proc. ACL 2023*, pages 7395–7408, Toronto.

Longyin Zhang, Bowei Zou, and Ai Ti Aw. 2024. Empowering tree-structured entailment reasoning: Rhetorical perception and LLM-driven interpretability. In *Proc. LREC-COLING 2024*, pages 5783–5793, Torino.

Shiyue Zhang, David Wan, and Mohit Bansal. 2023b. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. In *Proc. ACL 2023*, pages 2153–2174, Toronto.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition. In *Proc. ACL 2023*, pages 6477–6492, Toronto.

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

Xuekai Zhu, Jian Guan, Minlie Huang, and Juan Liu. 2023. StoryTrans: Non-parallel story author-style transfer with discourse representations and content enhancing. In *Proc. ACL 2023*, pages 14803–14819, Toronto.

A DISRPT 2025: Statistics by Partition

Table 5 provides general information for each datasets, such as the domains covered and overall stats. Specifically, '#Docs', '#Sents', '#Tokens', and '#EDUs' correspond to the total number of documents, sentences (Treebanked track), tokens, and EDUs. '#Conn' is the number of connectives, and 'Vocab' is the number of unique tokens. '#Labels' is the size of the label set, and '#Rels' to the total number of relations annotated.

Table 6 gives a detailed overview of statistics for each partition in every dataset. Datasets with 0 training tokens are test-only. Note the the #Units column refers to the number of EDUs in segmentation datasets (top part of the table), and the number of connectives in connective detection datasets (bottom part of the table).

B DISRPT 2025: The Unified Label Set

For DISRPT 2025, we defined a mapping from all the original relations annotated to 17 classes. We indicate in Table 7 some of the relations covered in each framework by each class.

Corpus	Domain	#Docs	#Sents	#Tokens	Vocab	#EDUs	#Conn	#Labels	#Rels	References
Tasks 1 and 3: EDU Segmentation and Relation Classification										
ces.rst.crdt	journalistic texts	54	835	14,664	6,065	1,453	-	17	1,249	Czech RST Discourse Treebank 1.0 (Poláková et al., 2023)
deu.rst.pcc	newspaper commentaries	176	1,944	32,836	8,260	3,111	-	16	2,665	Potsdam Commentary Corpus (Stede and Neumann, 2014)
**eng.dep.covdtb	scholarly paper abstracts on COVID-19	300	2,343	60,907	8,293	5,705	-	11	4,985	COVID-19 Discourse Dependency TB (Nishida and Matsumoto, 2022)
eng.dep.scidtb	scientific articles	798	4,202	102,534	8,700	10,986	-	14	9,903	Discourse Dependency TB for Scientific Abstracts (Yang and Li, 2018)
eng.erst.gentle	multi-genre	26	1,334	17,979	4,133	2,716	-	17	2,552	Genre Tests for Linguistic Evaluation (GENTLE) (Aoyama et al., 2023)
eng.erst.gum	multi-genre	255	14,158	254,890	29,323	32,428	-	17	30,747	Georgetown University Multilayer corpus V11 (Zeldes, 2017)
eng.rst.oll	online learning discussions	327	2,156	46,471	4,821	3,079	-	17	2,751	Online Learning Corpus (Potter, 2008)
eng.rst.rstdt	news	385	8,318	208,912	19,160	21,789	-	17	19,778	RST Discourse TB (Carlson et al., 2001)
eng.rst.sts	scholarly debate	150	2,591	71,206	7,675	3,208	-	17	3,058	Science, Technology, and Society corpus (Potter, 2008)
eng.rst.umuc	diplomatic speeches	87	2,424	61,590	5,684	5,421	-	15	4,997	Potsdam Multilayer UNSC Corpus (Zaczynska and Stede, 2024)
eng.sdrt.msdc	dialogues	440	14,744	231,352	2,589	23,160	-	10	27,848	The Minecraft Structured Dialogue Corpus (Thompson et al., 2024b)
eng.sdrt.stac	dialogues	1,101	7,394	52,271	3,734	12,552	-	11	12,271	Strategic Conversations corpus (Asher et al., 2016)
eus.rst.ert	medical, terminological and scientific	164	2,380	45,780	13,662	4,202	-	16	3,632	Basque RST Treebank (Iruskieta et al., 2013)
fas.rst.prstc	journalistic texts	150	2,179	66,926	7,880	5,853	-	14	5,191	Persian RST Corpus (Shahmohammadi et al., 2021)
fra.sdrt.annodis	news, wiki	86	1,507	32,699	7,513	3,429	-	12	3,321	ANNOtation DIScursive (Afantenos et al., 2012)
fra.sdrt.summre	meeting transcripts	67	21,695	295,392	10,506	35,907	-	-	-	SUMM-RE (Hunter et al., 2024; Prévot et al., 2025)
nld.rst.nldt	expository texts and persuasive genres	80	1,651	24,898	4,935	2,343	-	16	2,264	Dutch Discourse Treebank (Redeker et al., 2012)
pol.iso.pdc	multi-genre	556	9,142	156,980	37,833	5,115	-	12	8,543	Polish Discourse Corpus (Ogrodniczuk et al., 2024; Calzolari et al., 2024)
por.rst.cstn	news	140	2,221	63,332	7,786	5,537	-	15	4,993	Cross-document Structure Theory News Corpus (Cardoso et al., 2011)
rus.rst.rrt	blog and news	234	13,131	262,495	48,691	28,634	-	15	25,095	Russian RST Treebank (Toldova et al., 2017)
spa.rst.rststb	multi-genre	267	2,089	58,717	9,444	3,351	-	16	3,049	RST Spanish Treebank (da Cunha et al., 2011)
spa.rst.sctb	multi-genre	50	516	16,515	3,735	744	-	16	692	RST Spanish-Chinese Treebank (Spanish) (Cao et al., 2018)
zho.dep.scidtb	scientific	109	500	18,761	2,427	1,407	-	14	1,297	Chinese Dependency TB for Scientific Abstracts (Cheng and Li, 2019)
zho.rst.gcdt	multi-genre	50	2,692	62,905	9,818	9,706	-	17	8,413	Georgetown Chinese Discourse Treebank (GCDT) (Peng et al., 2022b,a)
zho.rst.sctb	multi-genre	50	580	15,496	2,973	744	-	17	692	RST Spanish-Chinese Treebank (Chinese) (Cao et al., 2018)
Tasks 2 and 3: Connective Detection and Relation Classification										
deu.pdtb.pcc	newspaper commentaries	176	2,193	33,222	8,359	-	1,116	11	2,109	Potsdam Commentary Corpus 2.2 (Bourgonje and Stede, 2020)
eng.pdtb.gentle	multi-genre	26	1,334	17,979	4,133	-	466	12	786	Genre Tests for Linguistic Evaluation (Aoyama et al., 2023)
eng.pdtb.gum	multi-genre	255	14,158	254,890	29,323	-	8,191	13	13,879	Georgetown University Multilayer corpus v11 (Zeldes, 2017)
eng.pdtb.pdtb	news	2,162	48,630	1,173,379	48,937	-	26,048	13	47,792	Penn Discourse Treebank (Prasad et al., 2014; Webber et al., 2019)
**eng.pdtb.tedm	TED talks	6	381	8,185	1,881	-	341	13	529	TED-Multilingual Discourse Bank (English) (Zeyrek et al., 2018, 2019)
ita.pdtb.luna	speech	60	3,750	25,242	2,457	-	1,071	11	1,525	LUNA Discourse Data Set (Tonelli et al., 2010; Riccardi et al., 2016)
pcm.pdtb.disconaija	transcribed spoken data	176	9,242	140,729	5,063	-	4,025	13	9,903	DiscoNaija (Scholman et al., 2025)
por.pdtb.crpc[6]	news, fiction, and didactic/scientific texts	302	5,194	186,849	22,208	-	5,159	12	11,327	Portuguese Discourse Bank (CRPC) (Mendes and Lejeune, 2022; Généreux et al., 2012)
**por.pdtb.tedm	TED talks	6	394	8,190	2,162	-	305	13	554	TED-Multilingual Discourse Bank (Portuguese) (Zeyrek et al., 2018, 2019)
tha.pdtb.tdtb	news	180	6,534	256,523	11,789	-	10,864	12	10,861	Thai Discourse Treebank (Prasertsom et al., 2024)
tur.pdtb.tdb	multi-genre	197	31,197	496,358	90,108	-	8,748	13	3,176	Turkish Discourse Bank (Zeyrek and Kurfalı, 2017)
**tur.pdtb.tedm	TED talks	6	410	6,286	2,771	-	382	13	574	TED-Multilingual Discourse Bank (Turkish) (Zeyrek et al., 2018, 2019)
zho.pdtb.cdtb	news	164	2,891	73,314	9,085	-	1,660	9	5,270	Chinese Discourse Treebank (Zhou et al., 2014)
zho.pdtb.ted	TED talks	72	8,671	181,910	14,197	-	5,958	15	13,308	TED-Multilingual Discourse Bank (Chinese) (Zeyrek et al., 2018, 2019)

Table 5: **DISRPT 2025 dataset stats**: ** indicates an OOD dataset, new dataset are in boldface, and surprise datasets are underlined.

Corpus	Train					Dev					Test				
	#Docs	#Toks	#Types	#Units	#Rels	#Docs	#Toks	#Types	#Units	#Rels	#Docs	#Toks	#Types	#Units	#Rels
Tasks 1 and 3: EDU Segmentation and Relation Classification															
ces.rst.crdt	48	11,766	5,080	1,152	978	3	1,346	777	140	123	3	1,552	874	161	148
deu.rst.pcc	142	26,517	6,988	2,534	2,349	17	3,117	1,446	282	260	17	3,202	1,413	295	273
eng.dep.covdtb	0	0	0	0	0	150	29,405	5,466	2,754	2,399	150	31,502	5,505	2,951	2,586
eng.dep.scidtb	492	62,488	6,715	6,740	6,060	154	20,299	3,540	2,130	1,933	152	19,747	3,341	2,116	1,910
eng.erst.gentle	0	0	0	0	0	0	0	0	0	0	26	17,979	4,133	2,716	2,552
eng.erst.gum	191	193,740	24,681	24,756	23,465	32	30,435	6,311	3,897	3,708	32	30,715	6,828	3,775	3,574
eng.rst.oll	293	37,265	4,330	2,511	2,217	17	4,601	1,276	280	263	17	4,605	1,131	288	271
eng.rst.rstdt	309	169,321	17,016	17,646	16,002	38	17,574	4,000	1,797	1,621	38	22,017	4,808	2,346	2,155
eng.rst.sts	135	57,203	6,816	2,581	2,446	7	7,129	1,859	291	284	8	6,874	1,755	336	328
eng.rst.umuc	77	49,727	5,085	4,333	3,988	4	6,005	1,475	565	525	6	5,858	1,606	523	484
eng.sdrt.msdc	307	166,719	2,199	16,285	19,598	32	17,926	782	1,860	2,232	101	46,707	1,239	5,015	6,018
eng.sdrt.stac	887	42,582	3,355	10,159	9,912	105	5,149	972	1,239	1,231	109	4,540	761	1,154	1,128
eus.rst.ert	116	30,690	10,217	2,785	2,533	24	7,219	3,316	677	614	24	7,871	3,528	740	485
fas.rst.prstc	120	52,497	6,884	4,607	4,100	15	7,033	2,005	576	499	15	7,396	2,061	670	592
fra.sdrt.annodis	64	22,515	5,712	2,255	2,177	11	5,013	1,722	556	523	11	5,171	1,823	618	621
fra.sdrt.summre	47	210,398	8,816	25,532	0	7	28,176	2,675	3,515	0	13	56,818	3,707	6,860	0
nld.rst.nldt	56	17,562	3,911	1,662	1,608	12	3,783	1,227	343	331	12	3,553	1,283	338	325
por.rst.cstn	114	52,177	6,856	4,601	4,148	14	7,023	1,639	630	573	12	4,132	940	306	272
rus.rst.rrt	188	208,982	42,193	22,839	20,014	19	24,490	8,161	2,555	2,266	27	29,023	9,588	3,240	2,815
spa.rst.rststb	203	43,055	7,648	2,472	2,240	32	7,551	2,240	419	383	32	8,111	2,338	460	426
spa.rst.sctb	32	10,253	2,642	473	439	9	2,448	971	103	94	9	3,814	1,271	168	159
zho.dep.scidtb	69	11,288	1,795	871	801	20	3,852	970	301	281	20	3,621	918	235	215
zho.rst.gcdt	40	47,639	8,192	7,470	6,454	5	7,619	2,166	1,144	1,006	5	7,647	2,061	1,092	953
zho.rst.sctb	32	9,655	2,195	473	439	9	2,264	838	103	94	9	3,577	1,137	168	159
Task 2: Connective Detection and Relation Classification															
deu.pdtb.pcc	142	26,831	7,071	934	1,723	17	3,152	1,460	88	192	17	3,239	1,424	94	194
eng.pdtb.gentle	0	0	0	0	0	0	0	0	0	0	26	17,979	4,133	466	786
eng.pdtb.gum	191	193,740	24,681	6,240	10,519	32	30,435	6,311	972	1,682	32	30,715	6,828	979	1,678
eng.pdtb.pdtb	1,805	975,544	44,249	21,484	39,524	177	97,449	12,391	2,178	3,973	180	100,386	12,323	2,386	4,295
eng.pdtb.tedm	0	0	0	0	0	2	2,616	842	110	178	4	5,569	1,354	231	351
ita.pdtb.luna	42	16,209	1,846	671	944	6	2,983	708	139	206	12	6,050	1,156	261	375
pcm.pdtb.disconaija	138	111,843	4,454	3,268	7,834	18	14,561	1,140	369	1,052	20	14,325	1,336	388	1,017
pol.iso.pdc	459	129,689	33,063	4,226	7,040	49	13,923	5,769	463	760	48	13,368	5,735	426	743
por.pdtb.crpc	243	147,594	18,821	3,994	8,794	28	20,102	5,243	621	1,285	31	19,153	4,903	544	1,248
por.pdtb.tedm	0	0	0	0	0	2	2,785	934	102	190	4	5,405	1,549	203	364
tha.pdtb.tdtb	139	199,135	10,462	8,277	8,274	19	27,326	3,107	1,243	1,243	22	30,062	3,188	1,344	1,344
tur.pdtb.tdb	159	398,515	77,245	7,063	2,444	19	49,952	17,476	831	311	19	47,891	16,748	854	421
tur.pdtb.tedm	0	0	0	0	0	2	2,159	1,073	135	211	4	4,127	1,957	247	363
zho.pdtb.cdtb	125	52,061	7,049	1,034	3,657	21	11,178	2,806	314	855	18	10,075	2,698	312	758
zho.pdtb.ted	56	144,581	12,382	4,701	10,649	8	17,809	2,913	589	1,329	8	19,520	3,255	668	1,330

Table 6: **Dataset statistics by partitions.** #Units refers to EDUs for segmentation datasets and connectives for connective detection datasets. #Types gives the total vocabulary size of unique token forms.

DISRPT25 labels	ISO	PDTB	(e)RST/DEP	SDRT
ALTERNATION	*disjunction*	*alternative, expansion.disjunction*	*joint-disjunction, alternativa*	*alternation*
ATTRIBUTION	-	-	*attribution, attribution-negative*	*attribution*
CAUSE	*cause*	*contingency.cause.result / reason*	*consequence, cause-result*	*result*
COMMENT	*expansion*	-	*comment, topic-comment*	*comment*
CONCESSION	*concession*	*comparison.concession (+speechact)*	*concession, comparison*	
CONDITION	*condition*	*conditional, contingency.condition*	*contingency-condition, hypothetical*	*conditional*
CONJUNCTION	*conjunction*	*expansion.conjunction*	*joint-list, topic-drift, topic-shift*	
CONTRAST	*contrast, substitution*	*expansion.exception, contrast, exception.substitution*	*antithesis, adversative-contrast*	*correction, contrast*
ELABORATION	*elaboration*	*expansion.instantiation, expansion.level-of-detail*	*example, elaboration-process, definition*	*e-elaboration, q_elab, elaboration*
EXPLANATION	*cause*	*contingency.cause, explanation-motivation, evidence, justify*	*explanation*, explanation*	
FRAME	-	*expansion.background*	*background, bg-goal, bg-compare*	*frame, background*
MODE	*manner, similarity*	*expansion.manner, comparison.similarity*	*manner, means, preference*	-
ORGANIZATION	*elaboration*	*progression, Expansion*	*organization-heading, summary*	-
PURPOSE	*purpose*	*contingency.goal, purpose*	*purpose, enablement*	*goal*
QUERY	*functional dependence*	*hypophora*	*interpretation, problem-solution, question-answer*	*acknowledgment, clarification_question*
REFORMULATION	*restatement*	*expansion.restatement, repetition*	*restatement*	-
TEMPORAL	*synchrony, asynchrony*	*temporal.asynchronous / synchronous*	*temporal-after, sequence, context-circumstance*	*narration, flashback, temploc*

Table 7: **DISRPT 2025 Label Set**. The class defined for the shared task are indicated in the first column, the other columns are examples of relations covered from different datasets for each framework.

DisCuT and DiscReT: MELODI at DISRPT 2025

Multilingual discourse segmentation, connective tagging and relation classification

[1]Robin Pujol* and [1]Firmin Rousseau* and [1,3]Philippe Muller and [1,2,3]Chloé Braud
[1]UT - IRIT ; [2]CNRS ; [3]ANITI
[1]firstname.lastname@irit.fr

Abstract

This paper presents the results obtained by the MELODI team for the three tasks proposed within the DISRPT 2025 shared task on discourse: segmentation, connective identification, and relation classification. The competition involves corpora in various languages, in several underlying frameworks, and datasets are given with or without sentence segmentation. This year, for the ranked, closed track, the campaign adds as a constraint to train only one model for each task, with an upper bound on the size of the model (no more than 4B parameters). An additional open track authorizes any size of, possibly non public, models that will not be reproduced by the organizers and thus not ranked. We compared several fine-tuning approaches either based on encoder-only transformer-based models, or auto-regressive generative ones. To be able to train one model on the variety of corpora, we explored various ways of combining data – by framework, language or language groups, with different sequential orderings –, and the addition of features to guide the model.

For the closed track, our final submitted system is based on XLM-RoBERTa large for relation identification, and on InfoXLM for segmentation and connective identification. Our experiments demonstrate that building a single, multilingual model does not necessarily degrade the performance compared to language-specific systems, with at best 64.06% for relation identification, 90.19% for segmentation and 81.15% for connective identification (on average on the development sets), results that are similar or higher that the ones obtained in previous campaigns. We also found that a generative approach could give even higher results on relation identification, with at best 64.65% on the dev sets.[1]

1 Introduction

Discourse parsing, the task consisting in finding semantic and rhetorical relations between spans of text (clauses, sentences, or paragraphs) is a well-known, yet challenging problem in computational linguistics, and has been shown to help in other NLP tasks, such as summarization (Zhang et al., 2023; Cripwell et al., 2023), question-answering (Fernandes et al., 2023; Jiang et al., 2023a), explainability (Devatine et al., 2023) or reasoning (Sharma et al., 2025). These relations reflect the argumentation structure or presentational choices, and have also been generalized to conversation, where dialog-specific phenomena such as adjacency pairs can be represented as relations between utterances – e.g., answer to a question, acknowledgment to a statement.

The field is characterized by significant divergence among different theoretical frameworks, with various views on the proper units of the structure, distinct typologies of relation, and heterogenous formats for annotations. The DISRPT shared tasks have been aiming at a standardization of discourse-related tasks since 2019, by providing a unique format for data representation, and aligning intermediate objectives, such as discourse segmentation into basic units.

The 2025 edition goes a few steps beyond, by proposing a unified typology of 17 relations – over around 350 distinct labels originally and 191 in 2023 –, allowing to build systems across the varied annotation projects. The new campaign also integrates a few new corpora and new languages (Czech, Polish, Nigerian Pidgin), and take into account some datasets updates (e.g. the English PDTB split is modified to avoid overlap with the RST DT). Note that the label unification does not solve all discrepancies between datasets. A same pair of segments can be annotated with 2 distinct labels, as in (a) below: the same pair of sentences,

*Equal contribution.

[1]Code segmentation / connective: https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut_disrpt25 ; relation: https://gitlab.irit.fr/melodi/andiamo/discourserelations/discret_disrpt25. Trained models are available on HuggingFace within the Discourse Hub organization: https://huggingface.co/multilingual-discourse-hub

Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 21–35

from the same document, is annotated in both the RST DT and the PDTB, but with different relations that also correspond to different labels in the DISRPT label set – resp. *contrast* and *alternation*. This could come from differences between the frameworks, the annotation project or the annotators training. In addition, the divergence between annotations is also reflected in the segmentation decisions, and we may have two slightly different pair of segments with the same label, as in (b), which could also be confusing for the model. These examples demonstrate that combining discourse datasets remains hard for automatic systems.

(a) [Call it a fad.] [Or call it the wave of the future.] RST DT : *contrast* ; PDTB : explicit *expansion.disjunction* (wsj_0633)

(b) [who was derided as a "tool-and-die man"] [when GE brought him in **to clean up Kidder in 1987**] RST DT *temporal* ; PDTB : explicit *temporal.synchronous* (wsj_0604)

This year adds a few constraints to reflect the recent evolution of NLP: instead of allowing for separate models, it is mandatory to cast each of the 3 sub-tasks (discourse unit segmentation, connective recognition, relation labelling) into a multilingual approach, with only one model trained on all datasets available for the task. This makes for more general model, and acknowledge the existence of more and more non-monolingual models. With the advent of "Large" pretrained Language Models (PLM), questions of reproducibility or accessibility of resources are now also very important, and the official track caps the size of pretrained model to 4B parameters for what are expected to be the main approaches: fine-tuning a large model or using an in-context learning approach in a generative context.

In this work, we describe various comparisons made between some of those choices, involving fine-tuning moderately sized encoders or sequence-to-sequence models, adding specific features to guide the models, and considering different strategies to manage the variety of data during training. For all tasks, our best approach relies on fine-tuning a PLM with a concatenation of the datasets with additional features with the following overall results: 90.19% (parsed) / 86.89% (plain) for segmentation, 80.11% / 79.79 for connective detection, and 64.06% for relation, all on the development sets. We also showed with the generative model (open track) how a multilingual approach proves better than fine-tuning separate models for relation labelling (64% on average on dev sets), confirming the importance of multilingual pretrained models.

2 Related work

Segmentation and connective identification have long been considered as easy and solved tasks, but most existing studies were on English newswire from the RST DT (Carlson et al., 2001). However, performance drop for under-represented languages and domains, or for dialogues, or when gold information such as sentence split is not given (Braud et al., 2017a; Scholman et al., 2021). Preliminary studies made use of lexico-syntactic features, e.g. (Lin et al., 2014) for connective or e.g. (Fisher and Roark, 2007; Braud et al., 2017b) for segmentation, while more recent approaches rely on transformer architectures, mostly varying the pretrained language model (PLM) used (Bakshi and Sharma, 2021; Metheniti et al., 2023; Lu et al., 2023).

Discourse relation classification is possibly the most studied task, especially on the English PDTB (Miltsakaki et al., 2004; Webber et al., 2019) with a specific focus on implicit relations – where no explicit connective such as – e.g. *when, because, if..then* – is used to mark the relation. In first studies, the emphasis was on improving the representation using e.g. linguistic features (Lin et al., 2009), or data augmentation, especially based on connective information, e.g. (Qin et al., 2017; Shi et al., 2017). PLMs have allowed to increase performance, especially for domain transfer (Shi and Demberg, 2019), but there is still a large room for improvements and many strategies have been proposed, relying on extending contextual information and external knowledge e.g. (Dai and Huang, 2018; Liu et al., 2020; Dai and Huang, 2019), leveraging explicit data or sense hierarchy, possibly with contrastive learning e.g. (Kim et al., 2020; Liang et al., 2020; Long and Webber, 2022), with also attempts relying on additional pre-training of PLMs (Kishimoto et al., 2020). Current state-of-the-art on implicit discourse relation in PDTB3 is around 60% in F1 with a dedicated model relying on the hierarchical organization of senses and the presence of implicit connectives in this dataset (Jiang et al., 2023b). Some attempts have been made to use large generative models such as ChatGPT for the task, with, for now, very low scores in zero- or few-shot settings (Yung et al., 2024).

DISRPT shared tasks allow to build and evaluate models on a large range of languages, discourse

frameworks, domains and genres within an unified format (Zeldes et al., 2019, 2021; Braud et al., 2023, 2024). Since the first edition, the benchmark has grown in size and representativeness with now 39 datasets, 5 frameworks, 16 languages, several domains and both monologues and dialogues (represented in 6 datasets in 2025 against 2 in 2023).

In all previous editions, the winning systems involved a fine-tuning approach of an encoder model on separate datasets. In 2023, the winning team for segmentation and connective identification (Metheniti et al., 2023) did use a multilingual model (XLM-RoBERTa large) but it was fine-tuned separately on each dataset. Their best results on the 2024 extended benchmark (Braud et al., 2024) are 92.14 in F_1 for segmentation (treebanked track[2]) and 82.73 for connective identification.

For relation classification, the same team presented results for a joint training over all datasets with, however, performance behind the best system (Liu et al., 2023) where several models were fine-tuned – a single one for large datasets, and a joint training on datasets from the same frameworks for the others. The best results for relations are still the ones obtained with the 2021 winning system, DisCoDisCo (Gessler et al., 2021), also reported in an extended version of the benchmark in (Braud et al., 2024), with, at best, 62.21 mean accuracy. Note that these results are not directly comparable to ours, since the relation set has changed.

In 2023, one participating team proposed a generative approach (Anuranjana, 2023) with, however, results far behind the other systems. Recently, the DISRPT benchmark has been used to test discourse understanding of large language models (Eichin et al., 2025) using an unified set of relations – different from the ones proposed in DISRPT 2025 –, but even very large models, with more that 10B parameters, struggle with the task, with performance under 60% mean accuracy when using only a linear probe on top of the pretrained frozen models.

3 Data

We train and evaluate our models using all the datasets provided by the shared task organizers.[3] In total, the benchmark is composed of 39 datasets, covering 13 languages and 6 frameworks. All the corpora are listed in Appendix A.4. The format and pre-processing steps are described in (Braud et al., 2024). This year, 13 new datasets were added, with now data for 3 new languages – Polish, Czech and Nigerian Pidgin –, and 6 frameworks in total – PDTB, RST, SDRT, eRST and ISO.

4 Global Approach

Our approach for all tasks relies on a pretrained language model, based on a Transformer architecture, fine-tuned on a concatenation of the datasets in different steps. For each task, we report, for reference, the scores obtained with the same language model fine-tuned on the separate datasets.

Full concatenation: Our reference / baseline system corresponds to the concatenation of all datasets, the model randomly draws instances from all datasets to form each batch during training. We then test different approaches meant to help the model to handle variations between annotations.

Framework/language-based learning: Here we merge the data in different steps, to test if a model could be better by learning separately the tasks for a single framework or group of languages before being introduced to a new one. The datasets for one framework / language group are concatenated and then learning is done sequentially on each group. The final performance correspond to the ones obtained at the end of sequential training, when all groups have been seen. We didn't have enough time to test all possible orders, but a few different options were investigated, as described below for each task. We also tested the injection of a subset of some datasets at the end of the first fine-tuning step, especially targeting datasets with very low results, as is done in certain continual pretraining techniques (Prabhu et al., 2023) to avoid catastrophic forgetting.

Feature augmentation: We implement some of the features tested in (Metheniti et al., 2024) to guide the model during joint training, more specifically language and framework information. Additional features for relation identification are indicated in Section 5.1. These information are given as additional tokens in the input, and ignored in the loss computation for sequential tasks.

Fine-tuning a generative model (open track experiments): We also compare our approach for relations to a quantized 4B LLM fine-tuned using a LoRA adapter (Hu et al., 2022), where the head

[2]Results on the Plain track are not indicated in the 2024 paper, but there were very similar to the treebanked track in 2023.

[3]https://github.com/disrpt/sharedtask2025

of the language model is restricted to 17 tokens corresponding to the discourse relations to predict.

5 Segmentation

The segmentation task is a sequential learning task, with two possible labels – 'B' and 'O', resp. beginning or not an EDU (Elementary Discourse units, the minimal text segments to be linked by discourse relations). Two tracks are proposed in DISRPT 2025: the Treebanked track where sentence split and morpho-syntactic information are given – gold or predicted –, and the Plain track where the full documents have to be segmented from the provided tokenization. We optimized our approach on the Treebanked track, and the best system was retrained for the Plain track, with the preprocessing indicated below.

5.1 Settings for segmentation

Preprocessing For both tracks, we test the addition of features representing the language and the framework.

Additionally, for the Plain track (segmentation and connective), we need to split the documents into subsequences small enough for our models. Sentence information is very important for segmentation, so we tested a recent, high-performing sentence splitter, SaT – Segment any Text (Frohmann et al., 2024), available on HuggingFace.[4] This model is multilingual, and designed for robustness across domains via a new pretraining scheme and additional fine-tuning. We found that the smaller version SaT-1L struggles with French quotation marks, a problem already noticed with Stanza, and thus chose the SaT-3L version. For the French corpus Annodis, when comparing the beginning of new sentences and the labels indicating the beginning of a new segment, we found that around 6% of the sentences do not correspond to a new segment in the original conllu data – split with stanza –, while only 2% are still not well segmented when using SaT3L. Even if the input is tokenized, some remapping was necessary afterwards to correspond to the input tokenization.

For dialogic datasets, this tool is far from perfect, so we adopted a simple cut-off strategy for LUNA as done in (Metheniti et al., 2023), but we experiment with SaT3L for the others.

Model architecture For segmentation and connective identification, contrary to discourse relation classification, we compare "small" models, under 560M parameters, as we consider these tasks as lower levels and thus believe that a faster and cheaper model should be favored.

Training dataset construction When combining datasets in the same language or group of languages, we tested sequential learning with the following order: CHINESE > ROM > GERM. The other datasets, and also some small datasets from these groups, are not part of the training set and predicted in zero-shot (e.g. for Farsi or Dutch) as we noticed it gave better results. CHINESE included all corpora in mandarin Chinese, ROM all romance languages (French, Spanish, Italian, Portuguese), GER all germanic languages (English, German).[5]

Comparisons As a (likely) upper bound, we fine-tuned one single system per dataset, as done in previous DISRPT campaigns, but using a multilingual pretrained model.

The reference system corresponds to the full concatenation, and we compared different multilingual pretrained models, all uploaded from HuggingFace: multilingual BERT base or large (Devlin et al.), RemBERT (Chung et al., 2021), XLM-Align (Chi et al., 2021b), and InfoXLM large (Chi et al., 2021a). Information on the size of the models is given in Table 1.

We compare models with or without features (language and framework), and a full concatenation *vs* framework grouping.

In a fully concatenated training scenario using mdeberta-v3-base, we tested models with language and/or framework features, as well as without any features.

Model	Size	# parameters	# layers
mBERT base	0.7 Go	110 M	12
XLM-Align	0.6 Go	125 M	12
InfoXLM large	2.7 Go	559 M	24
RemBERT	2.2 Go	575 M	36

Table 1: Size of the models tested for sequence classification: segmentation and connective identification.

Hyper-parameters Our models are fine-tuned with 30 epochs, a batch-size of 4, a learning rate of 10^{-5}, the ADAMW optimizer, and training will

[4] https://huggingface.co/segment-any-text

[5] The datasets only predicted in zero-shot are: eng.rst.oll, deu.rst.pcc, ces.rst.crdt, eus.rst.ert, fas.rst.prstc, nld.rst.nldt, por.rst.cstn, spa.rst.rststb.

early stop if the monitored metric does not improve strictly for 10 consecutive evaluation steps. When using InfoXLM, the first 6th layers are frozen (out of 24). The models are optimized based on the mean F_1 over all datasets. The training for one model on all the datasets is around 2 hours and 40 minutes for the connective identification task, and 2 hours and 20 minutes for the segmentation task. The training has been done on a L40S GPU.

5.2 Results and analysis

All models are compared based on the results on the development set and fine-tuned with all datasets, including the surprise datasets.

5.2.1 Findings from preliminary experiments

Learning with features or in sequence: We report results of different comparisons using BERT in Table 2. As we can see from the Table, the use of additional features corresponding to the language and the framework seems to help the model, and thus they were kept for the subsequent experiments. On the other hand, grouping datasets per languages lead to a drop in performance (85.75) compared to full concatenation (89.85). Data reinjection, meaning continuing the fine-tuning but only on datasets corresponding to low performance,[6] only leads to small improvement (86.12).

Model	no feat	lang	lang+fmk	grpd
mBERT base	89.85	90.10	**90.48**	85.75

Table 2: Different segmentation settings with mBERT base fine-tuned on the concatenation of all datasets: 'no feat': no additional features; 'lang': a token added to represent the language; 'lang+fmk': additional tokens representing the language and the framework; or with sequential learning per language without features: 'grpd': grouping datasets as described in Section 5.1.

Models comparison We compare different models for the setting giving the best performance on BERT – the full concatenation. We also use features indicating the language and framework when datasets are concatenated. As we can see in Table 3, all the models tested give very similar results, between 90.39 and 91.08% mean F_1, the best model being InfoXLM. The results obtained are also very close to the upper bound corresponding to single models, with at best 91.59. As noted in (Braud et al., 2024), the performance seems to have reached a plateau on this task, and using separate or joint models give similar results when averaged over the datasets.

	BERT large	RemBERT	XLM-Align	InfoXLM
		Single		
mean	91.05	91.59	91.07	91.54
mean23	91.00	92.08	91.70	91.82
		Concatenation		
mean	90.17	90.76	90.39	**91.08**
mean23	90.98	90.96	90.97	91.25

Table 3: Segmentation: Comparison between different models with single models and full concatenation, mean F_1 score over the datasets in the training+surprise release. The 'mean23' corresponds to the average score considering only the datasets present in DISRPT 2023.

5.2.2 Final results

Final results per dataset for segmentation are given in Table 4, with a full concatenation, language and framework features, and InfoXLM as pretrained language model. Even though the average between the 2023 and 2025 campaigns is the same, we can see a lot of differences in the datasets common to both campaigns. It's unclear why some corpora seem to benefit from the multilingual training while others are best when trained in isolation.

6 Connective identification

Connective identification is a sequential learning task, with three labels ('B', 'I', 'O', resp. beginning, inside, or outside a connective). We only test the best setting from the Segmentation task i.e. a full concatenation, with the language / framework features.

As for segmentation, our best results are obtained with InfoXLM (at best 80.47% mean F_1 in average (on dev). We found that freezing layers gave the best results with 6 layers frozen (vs 78.70 with all layers kept and 80.04 with 12 layers frozen). We obtained lower results with RemBERT (at best 79.18) and XLM-Align (at best 79.97). Our final results per dataset are given in Table 5. Again the averages for 2023 and 2025 are similar, with a lot of variance across common datasets.

7 Relation labeling

Discourse relation labeling is a multi-class classification task. The unification of the relation sets make it possible to jointly train over all datasets,

[6] The datasets reinjected are: fra.sdrt.annodis, spa.rst.sctb, zho.rst.gcdt, eng.rst.sts, zho.dep.scidtb, rus.rst.rrt, zho.rst.sctb.

Dataset	Treebanked			Plain		
	Dev	**Test**	Test23	Dev	**Test**	Test23
ces.rst.crdt	92.42	94.04	-	91.35	89.94	-
deu.rst.pcc	96.48	93.75	**96.01**	93.31	92.89	**94.24**
*eng.dep.covdtb		90.13	**92.13**	91.32	92.00	**92.13**
eng.dep.scidtb	95.41	94.60	**95.07**	94.65	**94.52**	94.49
eng.erst.gentle	-	89.45	-	-	87.33	-
eng.erst.gum	95.12	90.76	**95.50**	93.20	93.74	**94.46**
eng.rst.oll	94.12	89.27	-	85.45	86.19	-
eng.rst.rstdt	95.80	95.99	**97.62**	94.75	94.55	**97.74**
eng.rst.sts	79.05	87.52	-	72.55	82.34	-
eng.rst.umuc	90.69	88.21	-	89.46	88.33	-
eng.sdrt.msdc	96.33	93.98	-	86.96	85.41	-
eng.sdrt.stac	93.73	91.01	**95.22**	88.31	87.49	**90.67**
eus.rst.ert	90.43	**91.08**	89.93	87.82	87.14	**91.09**
fas.rst.prstc	93.73	**93.76**	93.40	92.75	93.34	**93.36**
fra.sdrt.annodis	90.39	86.38	**88.21**	89.29	87.22	**90.89**
fra.sdrt.summre		88.67	-	76.20	75.01	-
nld.rst.nldt	96.48	**96.73**	96.54	93.17	93.93	**97.19**
por.rst.cstn	92.71	**94.79**	93.98	91.26	91.53	**94.36**
rus.rst.rrt	92.41	**92.56**	85.58	90.50	**90.86**	85.41
spa.rst.rststb	94.52	92.04	**95.53**	92.68	91.06	**93.70**
spa.rst.sctb	88.68	**88.48**	85.63	72.52	80.67	**84.21**
zho.dep.scidtb	90.34	83.15	**89.07**	83.31	76.08	**90.04**
zho.rst.gcdt	89.30	90.29	**92.55**	87.99	88.73	**91.74**
zho.rst.sctb	64.52	67.91	**81.84**	53.76	55.09	**78.55**
mean	91.08	90.19	91.25	86.63	86.89	91.43

Table 4: Segmentation: Final results per datasets on the Treebanked and Plain tracks, on both dev and test sets. Results from DISRPT 2023 are reported from (Braud et al., 2023). The model is InfoXLM fine-tuned on a full concatenation of the datasets, with features indicating the language and framework in the input.

Dataset	Treebanked			Plain		
	Dev	Test	Test23	Dev	Test	Test23
deu.pdtb.pcc	85.56	76.84	-	85.41	76.60	-
**eng.pdtb.gentle	-	87.69	-	-	86.41	-
eng.pdtb.gum	88.72	88.49	-	88.10	88.36	-
eng.pdtb.pdtb	92.52	93.59	**93.66**	93.04	**93.88**	91.64
*eng.pdtb.tedm	79.25	**79.09**	78.36	78.22	**78.80**	75.83
ita.pdtb.luna	73.94	64.65	**65.85**	67.42	61.92	**71.60**
pcm.pdtb.disconaija	71.03	78.51	-	68.27	77.68	-
pol.iso.pdc	67.45	70.01	-	63.84	70.08	-
por.pdtb.crpc	85.02	80.04	**80.66**	81.48	78.00	**79.49**
*por.pdtb.tedm	80.91	**80.30**	80.29	75.49	**80.29**	79.45
tha.pdtb.tdtb	92.18	**90.36**	85.66	90.42	**89.70**	69.92
tur.pdtb.tdb	89.67	91.62	**92.77**	88.93	**92.85**	91.12
*tur.pdtb.tedm	62.67	**65.30**	64.10	65.09	**65.12**	64.78
zho.pdtb.cdtb	79.74	80.00	**89.00**	76.52	82.25	**90.38**
zho.pdtb.ted	77.89	75.10	-	79.01	74.93	-
mean	80.47	80.11	81.15	78.66	**79.79**	79.36

Table 5: Connective: Final results per datasets on the Treebanked and Plain tracks, on both dev and test sets. Results from DISRPT 2023 are reported from (Braud et al., 2023). The model is InfoXLM-large fine-tuned on a full concatenation of the datasets, with features indicating the language and framework in the input.

limiting label scarcity issues. However, some group of relations may be heterogeneous, the distribution of labels is imbalanced, with large differences between datasets. For this task, we present two set of experiments: the fine-tuning of an encoder-only model (closed track) and the fine-tuning with LoRA of a generative language model (open track).

7.1 Settings for relation classification

Preprocessing Within the closed track, we test the addition of features to inform the model with the language and framework of a specific instance. In addition, we encode the direction with specific tokens rather than switching units, as it was proved more efficient in (Metheniti et al., 2024). We also experiment with features indicating if the relation is inter or intra sentential by adding a token **`local`** or **`non local`**. With direction, locality and language/framework features, our input looks like, e.g. (from eng.dep.covdtb) : **`English dep local`** *this qualitative case study has investigated six issues* **<** *related to preparedness and response to MERS and poliomyelitis :* **{**

The input of our systems are pairs of segments, that could be longer than the maximal length of our models inputs. We truncate the input pairs if too long, by considering the whole length of the pair of arguments. Once tokenized, we compute the length of each unit, and truncate if the total length, combining both units, is larger than the max length of our model: if only one unit exceeds half of the max length, we truncate this unit at (max length - length of the other unit); if both units are longer than half of the max length, they are both truncated at the half of the max length.

Model architecture Contrary to the other tasks, we aim at testing bigger models, and we thus evaluate the fine-tuning of different pretrained multilingual models in the XLM-RoBERTa family until reaching the limitation of 4B parameters.

For the open track, we test only a large 4B model, and use LoRA for a faster training. More precisely, we use a decoder-only quantized 4B model (Qwen3-4B), finetuned with LoRA. The language model head is restricted to 17 tokens standing for each relation, where tokens are characters (from 'A' to 'Q'), to avoid issue with over generation. In inference model, we look at the probability given to these 17 tokens. The prompt is an instruction to pick a relation among the given list, given two textual segments. We found just keeping the instruction in English with no explicit mention of input languages worked better (example in Appendix A.2). We tested LoRA with rank 32 and 64, a batch of 64 for Nk steps (N=$3-5$ < 1 epoch), represent-

ing 65M of trainable parameters

Training dataset formation When combining datasets based on frameworks, we test the following orders:

- PDTB > SDRT > RST+ERST+DEP+ISO
- PDTB > SDRT+RST+ERST+DEP+ISO.

When combining datasets based on languages, we test the following order: ROM > GER > SLAV > fas > eus > zho > tur > tha, where upper letters indicate a group of languages and lower case indicate a single language (possibly covering several datasets). The list of languages in each group in indicated in Table 10 in Appendix.

Note that, due to time constraints, we only report results on the datasets in the regular release for these variations, not on the surprise datasets.

Baselines / comparisons As an upper bound, we trained separate models on each dataset with XLM-RoBERTa-base, to understand how a joint multilingual model fares against a specialized fine-tuning.

The reference joint system corresponds to the full concatenation, and we compare different sizes of the multilingual model XLM-RoBERTa, with possibly some layers frozen due to computational and time limitations: XLM-RoBERTa base (125M parameters), XLM-RoBERTa large (561M) and XLM-RoBERTa XL (3.48B). Our comparison are mainly done on the base model.

Hyper-parameters The XLM-RoBERTa models are fine-tuned with 10 epochs, a learning rate of 1^{-5}, the ADAMW optimizer, and early stopping with a patience of 10 and a minimun delta of 0 and we evaluate it every 2000 steps. The first 6 layers are frozen when using XLM-RoBERTa large and XLM-RoBERTa base, and the first 18 with XLM-RoBERTa XL. We also adapt our batch size and gradient accumulation steps forXLM-RoBERTa large and XLM-RoBERTa base : we get a training batch size of 4 and a gradient accumulation of 4 but for XLM-RoBERTa XL we have a training batch size of 1 and a gradient accumulation of 16. The models are optimized based on the mean F_1 over all datasets. The training time is approximately 6 hours with XLM-RoBERTa base and 11 with large, and 62 with the XL version. We use one L40S GPU cards.

7.2 Results and analysis

Sequential learning *vs* references As described in Section 4, we compare a joint training to a sequential learning approach based on grouping datasets either by languages ('LANG') or frameworks ('FMK', order 1). Results are given in Table 6 in two settings: without any additional features, or with features indicating the language, the framework and the direction, see Section 4.

	Without feature				With feature			
	Single	Concat	FMK	LANG	Single	Concat	FMK	LANG
Mean	55.06	62.42	56.97	56.57	55.19	**62.54**	57.01	57.04
Mean23	55.81	63.33	58.40	56.60	55.90	**63.66**	58.37	56.88

Table 6: Relation classification: mean accuracy of XLM-RoBERTa base (regular release, dev set). Systems are tested with and without features (framework, language, direction). Single: one model is fine-tuned on each dataset separately; Concat: all datasets are concatenated together; FMK: sequential learning based on frameworks (PDTB > SDRT > RST+ERST+DEP) ; LANG: sequential learning based on languages (ROM > GER > SLAVE > fas > ASIAN).

The 'Single' model was tested as an upper bound, since a separate model is fine-tuned, specialized on each dataset, as in most previous approaches to the task. In both configurations, the joint model reaches higher performance (at best 62.54% acc.) than the separate models (at best 55.19% acc.), demonstrating that datasets from different frameworks and languages can help each other. The unification of relation sets probably helps a lot here.

The sequential approaches do not outperform the full concatenation, neither by grouping frameworks nor languages. When looking at the performance at each step of sequential learning, the systems seem to forget crucial information for the datasets introduced at the beginning, with performance lowering for the initial groups. For the per framework approach, we also test a sequence PDTB > other frameworks, and reach 59.19 in accuracy against 57.01 with three groups of frameworks and 62.54 with a simple concatenation: making bigger groups only closes the gap with the full joint training.

The two sequential approaches give similar results, but we notice that the mean accuracy is better with the per framework approach when considering only the datasets present in 2023 ('mean23'). We tested with a reinjection, using XLM-RoBERTa base, of the Czech dataset – on which the performance are very low - at the end of the sequential learning: however, while with 1 additional epoch, the score on the Czech dataset increases – 43.9 vs 42.28 –, the performance drops if we continue to 10 epochs – 38.21 –, and the overall performance

are lower (10 epochs: 61.39 against 63.66).

	Without features		Lang+Fmk+Dir		Lang/Fmk+Dir		Lang+Fmk+Dir+Loc	
	FMK	LANG	FMK	LANG	FMK	LANG	FMK	LANG
Mean	56.97	56.57	57.01	57.04	56.12	56.35	**57.23**	56.51
Mean23	58.40	56.60	58.37	56.88	57.12	56.35	**58.43**	56.59

Table 7: Relation classification: mean accuracy of XLM-RoBERTa base (regular release, dev set). Systems without features or using different set of features: 'lang': language, 'fmk':framework, 'dir': direction, 'lang/fmk': lang (resp. fmk) features for sequential learning on frameworks (resp. languages), 'loc': location.

Feature set As we can see in Table 6, the additional features do not seem very helpful, with only limited gain for both reference and sequential approaches (+0.1%). In addition to tokens representing the language ('lang'), the framework ('fmk') and the direction of the relation ('dir'), we thus investigated the use of another feature representing the distance between the arguments – expressed as *local* for intra-sentential and *non local* for inter-sentential relations ('loc'). We also test the use of a feature indicating the language when learning per framework, and the other way around ('lang/fmk'). However, while adding features generally improve performance of our sequential approaches, the improvement remains very limited with the whole set of features tested, see Table 7.

Size of the model In the end, our best model was obtained with the full concatenation of datasets and features representing the language, the framework and the direction. Within this setting, we compare different sizes of the pretrained model. The results are presented in Table 8. As expected, the performance improves with larger models, but only slightly (+0.5% acc. between the base and large version), and even decreases with the largest one, maybe because of excessive freezing.

Including surprise datasets The final results are given in Table 9. We evaluate a 0-shot setting: the model trained only on the regular release is evaluated on the surprise datasets. As expected, the performance drop: from 62.54 to 59.34 for XLM-RoBERTa base, and from 63.04 to 60.15 for the large version. Unsurprisingly, the datasets with the lowest accuracy are the new ones. Some of them also correspond to new languages or frameworks, e.g. ces.rst.crdt, pol.iso.pdc and pcm.pdtb.disconaija. For others, it is more surprising, as the benchmark already contains similar data, e.g. deu.pdtb.pcc, eng.rst.oll, eng.rst.sts and zho.pdtb.ted. It could come from a lack of robustness of our system, or specific features of these datasets. Interestingly, performance are better in 0-shot for deu.rst.pcc (at best 55 against 45) which could indicate a form of over-fitting. Overall, when fine-tuned on the new set of data, scores are improved, reaching 64.06 mean accuracy with XLM-RoBERTa large on the dev set and 64.01 on the test set. We tested 2 runs with the large model, and

	XLM-RoBERTa-base	XLM-RoBERTa-large	XLM-RoBERTa-XL
Mean	62.54	**63.04**	62.38
Mean23	63.66	**63.76**	63.75

Table 8: Relation classification: mean accuracy of XLM-RoBERTa base (regular release, dev set). Full concatenation and different model sizes.

Dataset	DEV					TEST	
	Zero-shot		Full FT			Full FT	
	Base	Large	Base	**Large**	XL	**Large**	Test23
ces.rst.crdt	42.28	50.41	43.09	**55.28**	54.47	48.65	-
deu.pdtb.pcc	34.9	32.29	56.25	57.81	57.81	**62.89**	-
deu.rst.pcc	53.08	**55.0**	42.69	49.62	45.0	50.18	26.92
eng.dep.covdtb	69.57	**71.99**	68.7	65.65	63.86	67.25	41.3
eng.dep.scidtb	78.74	77.39	78.43	**79.93**	79.46	78.74	67.56
eng.erst.gentle	-	-	-	-	-	**55.13**	-
eng.erst.gum	56.45	57.85	56.31	60.25	56.55	**64.1**	-
eng.pdtb.gentle	-	-	-	-	-	**61.96**	-
eng.pdtb.gum	64.27	65.46	64.09	67.18	66.11	**68.18**	-
eng.pdtb.pdtb	71.36	71.96	71.86	**74.33**	73.57	73.71	69.25
eng.pdtb.tedm	58.99	58.43	60.11	60.67	60.11	**65.53**	19.94
eng.rst.oll	54.75	54.75	54.75	57.03	**57.79**	46.86	-
eng.rst.rstdt	62.12	**63.29**	59.1	60.33	60.27	60.65	49.98
eng.rst.sts	44.37	**45.77**	40.49	43.66	41.2	38.72	-
eng.rst.umuc	56.0	57.71	57.14	59.05	56.76	**60.33**	-
eng.sdrt.msdc	84.45	84.99	84.63	**86.02**	84.59	85.16	-
eng.sdrt.stac	62.71	65.07	62.63	68.72	60.76	**70.74**	56.89
eus.rst.ert	53.26	54.89	53.75	**57.0**	56.03	54.23	51.77
fas.rst.prstc	55.91	53.31	53.31	55.51	56.31	**57.26**	50.34
fra.sdrt.annodis	61.95	58.51	59.66	**62.52**	61.19	59.74	44.96
ita.pdtb.luna	58.25	59.22	61.65	60.68	64.56	**65.6**	58.42
nld.rst.nldt	55.89	52.87	54.98	59.21	55.59	**62.15**	43.69
pcm.pdtb.disconaija	25.57	35.84	50.57	54.94	47.81	**57.92**	-
pol.iso.pdc	35.53	38.42	53.16	**58.55**	56.58	58.41	-
por.pdtb.crpc	68.79	71.52	70.58	73.39	73.15	**77.96**	72.76
por.pdtb.tedm	58.42	57.89	57.89	63.16	62.63	**67.86**	54.95
por.rst.cstn	63.7	62.83	64.05	66.49	**68.41**	66.91	62.87
rus.rst.rrt	62.71	61.87	62.58	64.12	63.77	**66.75**	61.52
spa.rst.rststb	64.23	63.71	64.75	**69.19**	65.8	62.44	58.22
spa.rst.sctb	63.83	67.02	64.89	**69.15**	67.02	66.04	33.33
tha.pdtb.tdtb	95.49	96.14	95.25	96.14	95.9	**96.73**	95.24
tur.pdtb.tdb	50.8	52.73	56.27	58.84	52.73	**64.85**	49.05
tur.pdtb.tedm	**60.19**	56.4	56.87	56.4	55.92	59.23	49.73
zho.dep.scidtb	60.85	61.21	61.57	62.99	59.07	**67.44**	**67.44**
zho.pdtb.cdtb	80.7	**82.22**	79.06	81.29	**82.22**	78.63	69.13
zho.pdtb.ted	41.46	43.19	62.9	67.12	65.16	**67.67**	-
zho.rst.gcdt	63.02	64.81	62.13	**65.31**	61.13	62.85	55.72
zho.rst.sctb	61.7	58.51	**62.77**	58.51	56.38	52.83	49.06
Mean	59.34	60.15	61.36	**64.06**	62.38	64.01	-
Mean 2023	63.57	63.72	63.15	65.36	63.75	**66.17**	54.4

Table 9: Relation classification with full concatenation and base features (lang, fmk, dir): Mean accuracy including the surprise datasets, scores are on the development (DEV) or test (TEST) set. 'Zero-shot' is a system trained only on the regular release ; 'Full FT' is a system fine-tuned on all datatest (regular+surprise). The pre-trained language model is XLM-RoBERTa version base, large or XL. In bold, the best score per dataset.

obtained stable results (mean= 63.88, std= 0.18).

We tested an even larger model – XLM-RoBERTa XL – without improving these scores, but note that, due to computational and time constraints, we froze 18 layers, possibly impeding its performance. Our final scores are a bit lower compared to 2023 (-1.5% on dev), due to the introduction of new, challenging datasets, such as the Czech ces.rst.crdt (at best 54.4% in acc), the English eng.rst.sts (at best, 45.77), and other datasets that remain difficult – deu.rst.pcc (at best 55).

Results for the open track: generative model The results obtained with the generative approach are reported in Table 11. While the model we used is fully open, can be fine-tuned locally, and is under the parameter count constraint, we consider it in the open track because reproducing the training will be difficult without a recent high-end GPU (not necessarily for RAM constraints, as it needs only 8GB but for various configuration issues; we ran it with an L40S GPU). The model has been uploaded to the huggingface hub,[7] and an inference script is provided to verify predictions on all corpora. Training code is available. Mean average accuracy over the dataset (dev set only) is about 1% higher than our best model based on a decoder-only model, demonstrating the potential of this approach. Notably, training converges after less than one epoch over the concatenation of all datasets.

8 Conclusion

The MELODI team submitted systems for the DISRPT 2025 campaign for all tasks and setups: segmentation, connective identification, and relation classification. We explored various fine-tuning strategies for both encoder-only (closed track) and generative decoder-only (open track) models, and methods for combining diverse datasets across languages and frameworks. We show that training only one model on all data can achieve performance close to separate fine-tuning on each dataset, with even better results in the case of relation labelling. Given the time constraints a lot of potentially interesting ideas have not been fully explored and might be avenues for further progress.

Acknowledgments

This work is supported by the AnDiaMO project (ANR-21-CE23-0020). Our work has benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the French "Investing for the Future – PIA3" program under the Grant agreement n°ANR-19-PI3A-0004. Chloé Braud and Philippe Muller are part of the program DesCartes and are also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. The work was also supported by the ANR grant SUMM-RE (ANR-20-CE23-0017).

[7] https://huggingface.co/philippemuller/disrpt_sft_qwen3

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'12)*, 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. The Association for Computational Linguistics.

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. Association for Computational Linguistics.

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'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.

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, Ophélie Lacroix, and Anders Søgaard. 2017a. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of ACL*.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017b. Does syntax help discourse segmentation? not so much. In *Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442.

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.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multidomain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

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*.

Yi Cheng and Sujian Li. 2019. 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.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2976–2987, Hong Kong, China. Association for Computational Linguistics.

Nicolas Devatine, Philippe Muller, and Chloé Braud. 2023. An integrated approach for political bias prediction and explanation based on discursive structure. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11196–11211, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arxiv:1810.04805 [cs].

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.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Seeger Fisher and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings ACL*.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. 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).

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023a. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2023b. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8048–8064, Toronto, Canada. Association for Computational Linguistics.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language*

Processing, pages 343–351, Singapore. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled End-to-End Discourse Parser. *Natural Language Engineering*, 20(2):151–184.

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.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of IJCAI*.

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. 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.

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, and Philippe Muller. 2024. Feature-augmented model for multilingual discourse relation classification. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 91–104, St. Julians, Malta. Association for Computational Linguistics.

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.

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.

Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K. Dokania, Philip H. S. Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. 2023. Computationally budgeted continual learning: What does matter? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 3698–3707. IEEE.

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.

Laurent Prévot, Roxane Bertrand, and Julie Hunter. 2025. Segmenting a large french meeting corpus into elementary discourse units. In *Proceedings of SIGDIAL*.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

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

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 Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

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*.

Krish Sharma, Niyar R Barman, Akshay Chaturvedi, and Nicholas Asher. 2025. Dimsum: Discourse in mathematical reasoning as a supervision module. *arXiv preprint arXiv:2503.04685*.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Triet Thai, Ngan Chu Thao-Ha, Anh Vo, and Son Luu. 2022. UIT-ViCoV19QA: A dataset for COVID-19 community-based question answering on Vietnamese language. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 801–810, Manila, Philippines. Association for Computational Linguistics.

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'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.

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.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

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, 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.

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.

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 Ogrodniczuk. 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*.

Shiyue Zhang, David Wan, and Mohit Bansal. 2023. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2153–2174, Toronto, Canada. Association for Computational Linguistics.

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

A Appendix

A.1 Language groups

Table 10 shows which language we grouped for some of the experiments.

Abbrev.	language codes
ROM	spa, por, ita, fra
GER	nld, deu, eng
SLAV	ces, rus

Table 10: Group of languages used in the sequential learning experiments for discourse relation identification. The datasets corresponding to Farsi, Basque, Chinese, Turkish and Thaï are considered as specific groups.

A.2 Prompt example for the generative model

Here you have two discourse units of text prefixed with <argument1> and <argument2>:

<argument1>: Cuando calienta el Sol

<argument2>: En este artículo daremos una descripción general del Sol , de las capas que lo componen , de las estructuras que se observan y estudian en cada una de ellas , y de algunos de los fenómenos solares que inciden más directamente en nuestro planeta .

Classify the discourse relation between the two arguments using the following labels:

class A: elaboration
class B: attribution
class C: conjunction
class D: temporal
class E: explanation
class F: contrast
class G: causal
class H: purpose
class I: comment
class J: concession
class K: condition
class L: mode
class M: organization
class N: frame
class O: query
class P: reformulation
class Q: alternation
SOLUTION
The correct answer is: class M

A.3 Generative approach: scores per dataset

The scores per dataset with the generative approach described in Section 7.1 are indicated in Table 11.

corpus	dev	test
Mean	64.65	65.54
ces.rst.crdt	47.97	50.00
deu.pdtb.pcc	60.42	65.46
deu.rst.pcc	42.11	54.58
eng.dep.covdtb	65.05	67.52
eng.dep.scidtb	83.13	81.26
eng.erst.gum	62.01	65.00
eng.pdtb.gentle	—	63.99
eng.pdtb.gum	67.60	68.71
eng.pdtb.pdtb	76.31	75.18
eng.pdtb.tedm	60.39	64.96
eng.rst.oll	55.70	47.97
eng.rst.rstdt	62.77	63.62
eng.rst.sts	47.36	46.34
eng.rst.umuc	61.00	61.36
eng.sdrt.msdc	88.15	86.61
eng.sdrt.stac	70.22	70.92
eus.rst.ert	52.03	51.34
fas.rst.prstc	56.71	56.76
fra.sdrt.annodis	64.53	60.23
ita.pdtb.luna	68.21	69.07
nld.rst.nldt	58.16	59.69
pcm.pdtb.disc	58.03	60.37
pol.iso.pdc	59.01	59.62
por.pdtb.crpc	74.16	77.80
por.pdtb.tedm	63.16	66.76
por.rst.cstn	68.59	71.32
rus.rst.rrt	65.38	66.68
spa.rst.rststb	70.24	65.02
spa.rst.sctb	69.68	69.18
tha.pdtb.tdtb	95.01	96.73
tur.pdtb.tdb	54.82	61.76
tur.pdtb.tedm	59.72	61.98
zho.dep.scidtb	70.11	70.23
zho.pdtb.cdtb	78.95	76.78
zho.pdtb.ted	66.62	67.97
zho.rst.gcdt	66.15	60.97
zho.rst.sctb	57.98	61.01

Table 11: Relation classification (open track): Results obtained with a generative model (Qwen4B) fine-tuned with LoRA. Results on the dev set on average for 2 runs of training. Test set was done only once on the last trained model (64.71 acc on the dev for this one). This is considered in the open track, but trained model can be found on huggingface hub under user philippemuller, with a notebook reproducing the test inference and evaluation.

A.4 Language Resources

The datasets were obtained from the following corpora: the Czech RST Discourse Treebank 1.0 (Poláková 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; Généreux et al., 2012), the Thai Discourse Treebank (Thai et al., 2022), the Turkish Discourse Bank (Zeyrek and Webber, 2008; Zeyrek and Kurfalı, 2017), and the Chinese Discourse Treebank (Zhou et al., 2014).

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 parsing[1]: 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.

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-

[1] https://sites.google.com/view/disrpt2025

Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 36–47

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 Face[4]: *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 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-learn[5] which preserves label distribution, we then randomly selected 4 groups (≈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 repository[6]) without examples and tested both natural-order

[2] https://sites.google.com/georgetown.edu/disrpt2021/results
[3] https://sites.google.com/view/disrpt2023/results
[4] https://huggingface.co/
[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 ≈1k 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

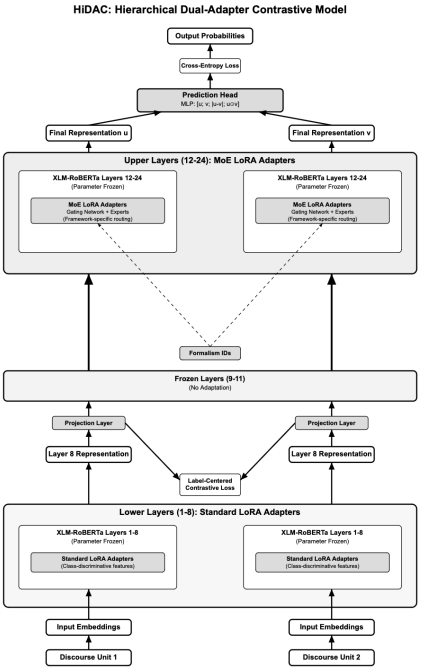

Figure 1: Overview of the HiDAC architecture. The contrastive loss is applied to an intermediate layer (8) to structure the embedding space, while the cross-entropy loss is applied to the representation at the last layer for the final classification.

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 $L_{\text{total}} = \lambda_{\text{ce}} \cdot L_{\text{ce}} + \lambda_{\text{cl}} \cdot L_{\text{lcl}}$, 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:

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

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

	Label	Train (≈170k)	Dev (≈28k)
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.

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 Transformers[8] 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

		BERT-based Models																	
		mBERT						XLM-R-Base						XLM-R-Large					
Category	**Size (%)**	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																			
pdtb	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%	59.5%	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 (≈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.

		Prompt-based Models					Custom Model		
		Zero-Shot		Few-Shot			HiDAC		
Category	**Size (%)**	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									
pdtb	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 (≈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 $\lambda_{\text{ce}} = 1.0$ and $\lambda_{\text{cl}} = 0.3$. 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-adapter 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 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 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. Multilayer 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 Ogrodniczuk. 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.

DeDisCo at the DISRPT 2025 Shared Task: A System for Discourse Relation Classification

Zhuoxuan Ju, Jingni Wu, Abhishek Purushothama, Amir Zeldes
Corpling Lab
Georgetown University
{zj153, jw2175, ap2089, amir.zeldes}@georgetown.edu

Abstract

This paper presents **DeDisCo**, Georgetown University's entry in the DISRPT 2025 shared task on discourse relation classification. We test two approaches, using an mt5-based encoder and a decoder based approach using the openly available Qwen model. We also experiment on training with augmented dataset for low-resource languages using matched data translated automatically from English, as well as using some additional linguistic features inspired by entries in previous editions of the Shared Task. Our system achieves a macro-accuracy score of 71.28, and we provide some interpretation and error analysis for our results.[1]

1 Introduction

Recent computational work on discourse relations has introduced richer multilingual corpora (Liu et al., 2024; Zeldes et al., 2025) and advanced transformer-based methods for implicit and explicit discourse relation classification (e.g. Li et al. 2024; Metheniti et al. 2024). While most previous studies focus on implicit relations within a single language (Liu and Strube, 2023), the DISRPT shared-task setting requires handling both explicit and implicit relations across multiple languages and annotation frameworks (Braud et al., 2024). This means that for example explicit CAUSAL relations as in 1, where the relation can be identified thanks to the explicit connective 'because', are targeted next to implicit ones, as in 2, where it is implied that it's hard for the speaker to think about food (because) they are not hungry (both example taken from `eng.erst.gum`).

(1) [I bought it]$_{unit1}$ [**because** it was funny]$_{unit2}$

(2) [I'm so not hungry right now,]$_{unit1}$ [it's hard for me to think about food.]$_{unit2}$

In DISRPT 2025 (see Braud et al. 2025), we designed and tested two architectures – one based on an encoder model and one based on a decoder – and submitted the decoder-only system, which is called DeDisCo (**De**coder-based **Dis**course **Co**gnoscente), which was trained via supervised fine-tuning with instruction-style prompts, tailored for multilingual discourse relation classification. Given the shared task constraint of a single multilingual model, supervised fine-tuning allowed us to efficiently fine-tune a compact decoder architecture while maintaining robust generalization across languages. Although we also explored encoder-based alternatives for methodology comparison, the decoder model formed the core of our official submission.

2 Related Work

2.1 Discourse Frameworks and Datasets

Multiple frameworks have been used to analyze and organize natural language discourse such as RST (Mann and Thompson, 1988) and eRST (Zeldes et al., 2025), PDTB (Marcu and Wong, 2004), SDRT (Lascarides and Asher, 2007), Discourse Dependencies (Stede et al., 2016), and ISO 24617-8 (Tomaszewska et al., 2024). These frameworks apply different structural and semantic guidelines for this purpose. However all of these frameworks utilize components such as discourse units and relations to build either shallow, or hierarchical/graph based discourse structures. Discourse units are then combined with *discourse relations labels* to describe the complex discourse structures present in texts.

Considerable efforts have been dedicated to building discourse corpora in multiple languages containing various text types. Combined with the variety of frameworks, many corpora contain fea-

[1] Full disclosure: our team includes both organizers and dataset annotators from the shared task. Results were reproduced independently by other organizers. All code is available at https://github.com/gucorpling/disrpt25-task.

Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 48–62

tures specific to languages, genres, and frameworks: for example, in RST, most datasets use clauses as discourse units, but only some datasets allow for discontinuous units; in PDTB, some datasets focusing on implicit relations only allow relations between sentences, while in others, smaller units are included. Additionally, the nature of the text will make some relations more or less prevalent, such as questions (the QUERY relation), which appear almost only in dialogic text types. Hence, both language and genre-specific considerations can be applied in many instances, leading to heterogeneity in the data parameterized by the language, framework, corpus, and genre. Previous DISRPT Shared tasks (Zeldes et al., 2019, 2021; Braud et al., 2023) have brought together collections of these varied datasets, with the previous iteration (DISRPT 2023 Braud et al., 2023) containing 26 corpora annotated across 4 frameworks and 13 languages, and the 2025 edition bringing these up to 39 datasets in coming from 6 frameworks and 16 languages. New in this edition is also the unification of the label set to the same 17 labels across all datasets; however, labels may exhibit subtle differences in usage across corpora, meaning once again that encoding dataset specific input may be crucial for high performance across benchmarks.

2.2 Discourse Parsing and Relation Classification

The task of discourse parsing refers to constructing discourse structures (in a given scheme) from natural language text. This involves sub-tasks such as segmenting discourse units, connecting them in order, identifying lexical units that signal relations (in PDTB and eRST), and assigning relation labels. In addition to the inherent challenges of these tasks, heterogeneity leads to systems that are useful for only a small subset of the corpora. With neural models, this challenge is more severe due to the need for substantial amounts of training data to build such systems. The multilingual and multi-track nature of the DISRPT shared task may alleviate some of these problems by posing the problem as a multilingual one across all datasets, allowing low-resource languages to gain in performance from the additional data available in high-resource languages. The tracks on discourse unit segmentation, connective identification, and relation classification also allow systems to focus on these problems in isolation – in this paper we target only the latter task of relation classification, given the units connected by the relation.

Regardless of the labels used, which are commonly framework-specific (Hovy, 1990), relation classification can be considered a single-label classification task over two ordered, non-overlapping textual input units. Multiple methods have been applied to the task, including in the context of full hierarchical discourse parsing, such as shift-reduce parsing (Marcu, 1999), feature based and neural parsers (Ji and Eisenstein, 2014), and simple span-based transformer encoder (Gessler et al., 2021; Metheniti et al., 2024), and decoder (Anuranjana, 2023) models.

2.3 Feature Encoding and Data Augmentation

Several past top systems Yu et al. (2019); Gessler et al. (2021); Metheniti et al. (2023) used hand-crafted features to mitigate discourse parsing challenges, even with transformer-based models. These range from corpus-level features such as language and framework, document-level features such as length or genre, and sample- or unit-level features such as lexical overlap or position. Unique features like whether two units share the same speaker were also crafted in corpus-specific systems. Features can be represented categorically (e.g., one-hot), as dense embeddings in neural architectures, or as plain text in LLM prompts.

Data augmentation has been used to improve both dataset and language-specific performance. Task-specific augmentation entails transforming or synthesizing data that is similar to the target dataset based on a distribution that also adheres to the dataset design. e.g. Liu et al. (2023) grouped together data for multiple languages with smaller corpora for such augmentation. For language-specific augmentation, the same or similar source for different languages can be translated into the target language, providing higher training data for the target language. This is popularly known as the translate-train paradigm (Conneau et al., 2018), which we employ below.

3 Data and Approach

The Discourse Relation Classification task of the DISRPT 2025 aims to classify discourse relations across a diverse set of languages and annotation frameworks. The relation classification task includes 38 of the joint task's 39 corpora, spanning 16 languages and 6 different frameworks, with a unified set of 17 labels provided. For convenience, the language, corpora, and framework are part of

the listed in appendix D (Table 6). The majority of the data are annotated using Rhetorical Structure Theory (RST; Mann and Thompson, 1988) and the Penn Discourse Treebank framework (PDTB; Marcu and Wong, 2004). A smaller subset of the corpora is annotated using SDRT, discourse dependencies (DEP), eRST, and the ISO frameworks.

For our system, we experimented with both encoder-based and decoder-based models. After evaluating their performance, we selected the stronger decoder-based model for our official submission. Below, we described how each model was implemented with our feature set and data augmentation.

3.1 Features

Language Corpus Framework (LCF) As noted, the training data spans multiple annotation frameworks, corpora, and languages, which poses a challenge for generalization since relations defined under different schemes or languages are not always directly comparable. To address this diversity, we incorporate metadata into the input to help the model distinguish and generalize across datasets. We refer to these elements collectively as LCF features (Language, Corpus, Framework). For example, the dataset identifier `zho.rst.gcdt` indicates a Chinese corpus (`zho`) named GCDT (Peng et al., 2022b), annotated using the RST framework.

DiscoDisco Features The DiscoDisco system (Gessler et al., 2021) introduced several handcrafted discourse features extracted from the data for the relation classification task (see Table 1), which were later shown to be highly effective (Metheniti et al., 2023). We selectively incorporated a subset of these features into our models, see Table 5 in appendix B for a detailed list of which DiscoDisco features are used in the decoder and the encoder separately. For exact details of these features we refer to the original DiscoDisco paper and system implementation.

Direction Discourse relations are marked between two segments of text, referred to as Arg1 and Arg2. In the DISRPT datasets, these segment pairs follow the text's original sequence, and an extra column specifies the intended argument direction for the annotated relation (e.g., `1>2`). For example in a cause relation, the cause points to the result, but may appear first (`1>2`) or second (`1<2`). This directional information is incorporated into both of our models, but encoded in different ways (see below).

Context Text context beyond the two units being classified plays a crucial role in discourse relation classification. Notably, Dai and Huang (2018) demonstrated that paragraph-level context significantly enhances the prediction of implicit discourse relations. In our models, we explore adding context to the model input and the impact of varying context window sizes surrounding the target sentence.

To clarify the experimental setup, Table 4 in appendix B provides an overview of which features, context, and data augmentation used in each model configuration.

3.2 Data Augmentation

Given the limited size of training data for certain languages, many of which also exhibit lower model accuracy, we apply targeted data augmentation to enhance performance. We focus on six low-resource languages: Czech (14.6K tokens), Dutch (24.9K), French (32.7K), Basque (45.7K), German (66K), and Persian (67K), covering a total of seven datasets. While these six corpora are not strictly the smallest by token count, they exemplify low-resource conditions due to the combination of restricted training data for the entire language, and comparatively weak baseline performance. We therefore targeted them for augmentation to mitigate these weaknesses. Our augmentation strategy involves translating English training instances from a source corpus into the target languages using API calls to the ChatGPT 4.1 model (OpenAI et al., 2024), and providing this data for system training and replication following the shared task rules (our system does not access ChatGPT in any way at training or testing time).

To ensure compatibility with the target language, we select English data for translation based on four criteria: annotation framework, discourse relation label distribution, genre alignment, and overall dataset size. The detailed correspondences are provided in Table 2. For each target corpus, we generated augmented data equivalent to approximately 75% of its original size. This ratio was chosen to enrich the training set without overshadowing the signals from the original in-domain examples. To ensure the quality and relevance of the synthetic data, we implemented a multi-faceted filtering strategy. First, we maintained a label distribution in the augmented set that closely mirrored the original. Second, we aligned the data's genre; for instance, as the German `deu.rst.pcc` contains mostly editorial texts and news, we primarily drew

Feature	Type	Ex.	Description
Genre	Cat.	reddit	Document genre (e.g., eng.erst.gum)
Children	Num.	2	Child units each discourse unit has
Discontinuous	Cat.	false	Unit tokens not contiguous
Is Sentence	Cat.	true	Unit is a complete sentence
Length Ratio	Num.	0.3	Token length ratio (u1 vs. u2)
Same Speaker	Cat.	true	Same speaker for u1 and u2
Doc. Length	Num.	214	Document length in tokens
Position	Num.	0.4	Unit position in doc (0–1)
Distance	Num.	7	Other units between u1 and u2
Lexical Overlap	Num.	3	Shared non-stoplist words

Table 1: Features used in 2021 DiscoDisco system.

Target Corpus	Source Corpus	Selected Source Genres
ces.rst.crdt	eng.erst.gum	essay, news
deu.pdtb.pcc	eng.pdtb.gum	essay, news, speech
deu.rst.pcc	eng.erst.gum	essay, news, speech
eus.rst.ert	eng.erst.gum	textbook, academic
fra.rst.prstc	eng.erst.gum	news, academic
nld.rst.nldt	eng.rst.(oll, sts)	bio, news, letter
fas.rst.prstc	eng.rst.rstdt	all

Table 2: Source–Target Genre & Framework Alignment for Translation-Based Data Augmentation

source material from the 'essay' and 'news' genres from the English GUM corpus, supplementing it with a small amount of 'speech' data to reach the target volume. Finally, we tried to enforce annotation guideline consistency. For example, we observed that German RST annotations do not segment relative clauses, unlike its counterpart in English. Therefore, we excluded any source examples with these incompatible structural patterns from the German augmentation set.

3.3 Pruned Qwen3-4B Decoder Only

For our decoder-only approach, we frame discourse relation classification as a generative task. Specifically, we feed a prompt to a decoder-only model, instructing it to directly select the correct label from a predefined label set included within the prompt itself. We employ the Qwen3-4B model[2] (Yang et al., 2025), chosen for its strong multilingual capabilities, supporting over 100 languages and dialects, which aligns well with the multilingual classification task. We apply supervised fine-tuning with instruction-style prompts to improve its task-specific performance.

Pruning The public Qwen3-4B model originally contains 4.02 billion parameters, slightly exceeding the shared task's 4B parameter limit. To address this, we adopt a pruning strategy based on layer removal as proposed by Men et al. (2024), which identifies redundant layers by measuring the similarity between their input and output representations. We determined that removing a single, most redundant layer was sufficient to meet the parameter requirement. After fine-tuning, the resulting pruned model[3] achieved performance on par with its unpruned counterpart.

Supervised Fine-Tuning Our methodology involves full-parameter supervised fine-tuning of the model on the task-specific dataset, which is reformulated into an instruction-style prompt format. Each instruction is enriched with a comprehensive set of features, including LCFs, direction, context, and selected DiscoDisco features (e.g., same speaker, position), as detailed in Section 3.1. The context is constructed from the sentence immediately preceding the first argument, the sentence(s) containing both arguments, and the sentence immediately following the second argument. We experimented with two distinct styles for prompt design: Verbose Instructional Prompt and a Structured Templated Prompt. The verbose prompt, illustrated in Figure 1, uses natural language to explicitly define the model's role, the task objective, the various input components, and decision-making guidelines. This contrasts with the structured prompt, which organizes the raw inputs into a compact, delimiter-separated format (e.g., ... $$ Arg1 $$ > ## Arg2 ## ...), resembling the input format for encoder models (Section 3.4). Although a large part of the verbose instructions is repeated identically in all samples, and may therefore be considered redundant, our experiments consistently showed that the verbose instructional prompts yielded superior performance, improving model accuracy by approximately 1–2% compared to the more concise, structured variants.

[2] https://huggingface.co/Qwen/Qwen3-4B

[3] https://huggingface.co/JuNymphea/Georgetown-qwen3-4B-pruned-for-disrpt2025

Prompt Design

```
## Role and Goal:
You are an expert in discourse analysis, tasked with identifying the discourse relation between two sentence units based on the provided label. Your goal is to accurately determine the relationship between these two units.

## Guidelines:
1. You will receive Unit1 and Unit2. Unit1 appears before Unit2 in the original text.
2. You will also be informed about the language of these units.
3. You will also be informed of the corpus from which the data is drawn, which may help guide your analysis.
4. The framework for analysis will be provided, outlining the structure used for discourse analysis.
5. You will be informed whether Unit1 and Unit2 are spoken by the same speaker.
6. You will also be given the distance between Unit1 and Unit2.
7. You will be provided with the percentage position of Unit1 and Unit2 in the original document.
8. You will be given the context in which these two units appear.
9. The direction of the relationship between these two units will be given.
10. You will be provided with a set of labels representing possible discourse relations. Choose one label that best fits the relationship between Unit1 and Unit2, and output only the chosen label.

## Labels:
contrast, condition, mode, organization, frame, temporal, concession, reformulation, comment, query, attribution, alternation, purpose, explanation, elaboration, causal, conjunction

## Language:
eng

## Corpus:
gum

## Framework:
erst

## Same Speaker:
True

## Distance Between Unit1 and Unit2:
51

## Percentage Position of Unit1:
0.0

## Percentage Position of Unit2:
0.7979

## Context:
Aesthetic Appreciation and ... on visitor visual behaviour .

## Direction:
From Unit1 to Unit2.

## Unit1:
Aesthetic Appreciation and Spanish Art :

## Unit2:
In this study we used eye - tracking in the first stage
```

Figure 1: Illustration of the Verbose Instructional Prompt used in Qwen3-4B experiments.

3.4 mT5 Encoder

We also experimented with mT5 (Xue et al., 2021), a multilingual T5 variant pretrained on a Common Crawl-based corpus covering 101 languages. Specifically, we selected the mT5-XL variant, comprising 3.7 billion parameters, which comes closest to the shared task 4B limit. For our classification task, we used only the encoder and added a classification head. We explored two strategies for incorporating metadata and discourse features: (1) encoding them as special input tokens, and (2) using separate embedding layers concatenated with the encoder input.

Feature Injection via Input Tokens We prepended LCF features as special input tokens (e.g., `LANG_eng`, `FW_erst`, `CORP_gum`), so that the model can incorporate this metadata directly in its tokenized input sequence. Since mT5 is trained to interpret prompt-like text, this approach naturally lets the model condition on task-specific context. This makes metadata injection especially effective in our setting, where the goal is to classify discourse relations across varied domains and annotation schemes.

In addition to metadata, we also applied this injection strategy to categorical features from the DiscoDisco feature set (Gessler et al., 2021), which capture properties such as whether the units are full sentences, whether the relation is discontinuous, and whether the two units share the same speaker. We encoded these features as explicit key-value tokens, for example `IS_SENTENCE_1`, `DISCONTINUOUS_0`, and `SAME_SPEAKER_1`. This design differs from the method of Metheniti et al. (2024), who append only raw feature values (e.g., `0.3`, `0.5`) in a fixed order, with each position implicitly corresponding to a particular feature. While their strategy reduces vocabulary size, it ties interpretation to positional indexing, making it less robust to reordering. By contrast, our approach makes the semantics of each feature explicit and order-independent, which aligns more naturally with mT5's training paradigm and improves interpretability.

Furthermore, we implemented pseudo-directional features from DiscoDisco. Specifically, for relations labeled as left-to-right (1>2), we inserted direction using the tokens } and > before and after the first argument span to signal directional flow. For right-to-left (1<2) relations, the inverse

markers were used. These directional cues are lightweight but informative, and help disambiguate argument structure across instances, especially in genres with flexible syntax or conversational turn-taking.

The resulting input sequence is organized as follows: metadata (LCF features), followed by categorical DiscoDisco features, and finally the target argument span:

(3)
```
LANG_eng FW_erst CORP_gum [SEP]
IS_SENTENCE_1 DISCONTINUOUS_0
SAME_SPEAKER_1 GENRE_academic
[SEP] } Aesthetic Appreciation
and Spanish Art: > Arg2: In this
study we used eye-tracking in the
first stage
```

Feature Embedding(s) We hypothesized that treating argument spans (arg1 and arg2) together as a single sequence separated by a special token [SEP] might better leverage mT5's native relative positional embeddings and attention dynamics. Meanwhile, surrounding context (Pre/Post) and metadata features were proposed to be embedded separately, since context can be long and sparse, which may attenuate positional signal strength if concatenated directly with argument spans. Thus, our proposed embedding schema is:

(4)
```
Concat(Embed(meta + features) +
Embed(pre-context) + Embed(arg1 [SEP]
arg2) + Embed(post-context))
```

This setup preserves the association between argument spans, while mitigating positional confusion or noise from long context sequences.

Exploratory experiments on a smaller development subset indicated that this structure offered modest conceptual clarity but did not substantially outperform simply prepending LCF and categorical DiscoDisco features as special tokens.

Corpus	Decoder (DeDisCo)	Encoder
ces.rst.crdt	52.70	51.35
deu.pdtb.pcc	67.01	56.19
deu.rst.pcc	67.03	49.82
eng.dep.covdtb	68.21	73.05
eng.dep.scidtb	83.66	79.58
eng.erst.gentle	67.08	61.29
eng.erst.gum	73.45	62.98
eng.pdtb.gentle	67.94	61.07
eng.pdtb.gum	71.39	65.20
eng.pdtb.pdtb	83.77	77.32
eng.pdtb.tedm	71.79	61.54
eng.rst.oll	62.73	51.66
eng.rst.rstdt	73.27	62.60
eng.rst.sts	58.54	49.39
eng.rst.umuc	67.36	59.09
eng.sdrt.msdc	90.00	84.11
eng.sdrt.stac	75.80	65.96
eus.rst.ert	54.64	55.67
fas.rst.prstc	60.47	57.77
fra.sdrt.annodis	60.39	52.82
ita.pdtb.luna	70.13	66.13
nld.rst.nldt	68.62	53.85
pcm.pdtb.disconaija	59.39	41.40
pol.iso.pdc	74.02	55.05
por.pdtb.crpc	79.17	75.64
por.pdtb.tedm	68.41	64.84
por.rst.cstn	70.22	69.85
rus.rst.rrt	74.85	68.95
spa.rst.rststb	69.72	64.55
spa.rst.sctb	83.02	76.73
tha.pdtb.tdtb	96.73	96.80
tur.pdtb.tdb	64.13	65.08
tur.pdtb.tedm	59.23	54.55
zho.dep.scidtb	80.00	68.37
zho.pdtb.cdtb	88.65	86.54
zho.pdtb.ted	75.49	66.24
zho.rst.gcdt	75.55	65.37
zho.rst.sctb	74.21	66.67
Macro Average	71.28	64.34
Micro Average	76.13	69.74

Table 3: Accuracy of encoder-only and decoder-only models on the test sets of 38 corpora. We use corpus codes for simplicity (check Table 6 in appendix D for language and framework information), with accuracy scores reported separately for the decoder-only and encoder-only models. Macro and micro averages are reported at the bottom.

4 Results

4.1 Performance Comparison of Encoder-Only and Decoder-Only Model

Table 3 reports the test scores of the encoder-only and decoder-only models across all 38 corpora.

We note several coarse observations at the outset: First, the decoder model outperforms the encoder model in all datasets except four: `eng.dep.covdtb`, `eus.rst.ert`, `tur.pdtb.tdb`, `tha.pdtb.tdtb`. However, the difference is minor (around 1 accuracy point or less) in all but `eng.dep.covdtb`, which is a 'test-only' dataset, meaning the models have seen no data of the same kind – although a few other datasets are also test-only, they do have 'related' datasets (`eng.erst.gentle`, Aoyama et al. 2023, is modeled on `eng.erst.gum`, and the `*.tedm` datasets closely follow recent versions of PDTB, Prasad et al. 2018). Aside from this unique property, we are unsure what sets this dataset apart.

By contrast, in the most extreme case the decoder achieves a 19 gain compared to the encoder on the Polish `pol.iso.pdc`. We can rule out data

contamination with gold data as a reason, since the data was released in the *surprise* test set and was reported to be annotated very recently.

Beyond individual dataset differences, we observed a broader trend with respect to data scale. In lower-resource settings, such as Czech (14.6K tokens), Dutch (24.9K), and French (32.7K), the decoder model consistently shows a substantial advantage over the encoder, with accuracy gains exceeding 10 points in some cases. These datasets have limited supervision and lack related training corpora, making them particularly reliant on pretrained representations. The decoder's autoregressive architecture and larger capacity appear to enable better generalization under these conditions. On the other hand, in larger datasets such as Thai (256K tokens), Turkish (496K), and English PDTB (1.17M), the performance gap narrows. In fact, the encoder slightly outperforms the decoder on Thai PDTB, suggesting that when sufficient labeled data is available, the simpler encoder-only setup can be just as effective, if not more so.

We also observed differences in how the two architectures respond to feature integration. LCF and DiscoDisco features consistently improved the encoder model, but in many cases degraded performance for the decoder. This suggests that encoder-only models can more effectively leverage categorical metadata and structural cues as additional signals, whereas decoder-only models are more sensitive to such injections. In contrast, extending the context window benefited the decoder but often harmed the encoder.

4.2 Decoder Model Ablation Tests

To assess the contribution of each additional features, we conduct an ablation study on the decoder-only model. In each experiment, one specific feature is removed from the input to evaluate its impact on performance. Detailed results are presented in appendix E (Table 7).

We find that direction is the most influential feature. When direction information is removed, 32 out of the 38 corpora experience a drop in accuracy of more than 4%, and 12 corpora suffer a decrease of over 10%. By contrast, for `tha.pdtb.tdtb`, direction contributes only marginal gains, and for `tur.pdtb.tedm`, it even leads to a performance drop. This is likely because Turkish is a free-word-order left-branching language, where in most subordinate constructions the clause with the connective precedes the main clause, causing direction features based on linear order to introduce noise rather than useful signal (Zeyrek et al., 2009).

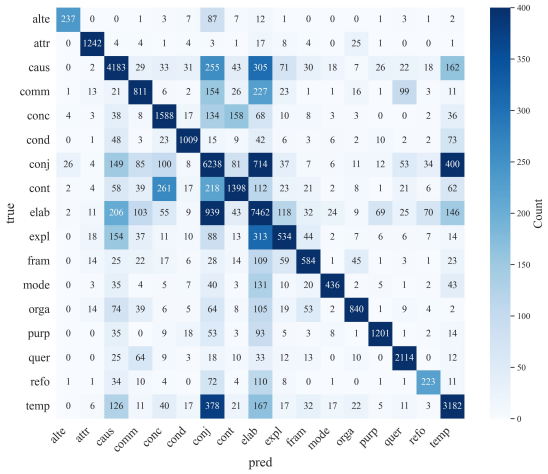

Figure 2: Confusion matrix over the entire dataset.

The second most impactful feature is context. Although prior work (Judge et al., 2024) suggests that adding context may not always be beneficial, we observe that context is a highly effective input for the decoder-only model. However, its impact varies significantly across corpora. For instance, `deu.rst.pcc` and `pcm.pdtb.disconaija` see gains exceeding 15%, and even up to 20% in `pol.iso.pdc`. For most corpora, the improvement is around 3%, though performance actually deteriorates on `tur.pdtb.tdb` and `ita.pdtb.luna`.

By comparison, LCF features tend to have smaller or even negative effects across many corpora. Nevertheless, they yield notable improvements (greater than 5%) for `deu.rst.pcc` and `zho.rst.sctb`.

A surprising finding emerges regarding data augmentation: it does not always improve performance on the target corpus. For example, For example, we observe gains on five corpora, except for `ces.rst.crdt` and `deu.pdtb.pcc`. On the other hand, for the source corpora, all five English datasets show consistent improvements from data augmentation, despite the source samples for the augmented entries also existing in the English training.

As for the DiscoDisco features, their overall impact was less pronounced across most corpora. The most notable exception was `eng.dep.covdtb`. For this unique, test-only dataset with no in-domain training data, removing these features surprisingly boosted performance by nearly 5%, suggesting they may introduce counterproductive noise.

(a) tha.pdtb.tdtb **(b)** ces.rst.crdt

Figure 3: Confusion Matrices for Common Relations in the Highest and Lowest Scoring Datasets.

5 Error analysis

Strengths and Confusions at the Relation Level The confusion matrix in Figure 2 indicates very high performance on key discourse relations including ELABORATION, CONJUNCTION, CAUSAL, and TEMPORAL, demonstrating that the model reliably classifies these central categories. However, we observed minor yet notable confusion between ELABORATION and CONJUNCTION, and this might because that ELABORATION and CONJUNCTION are frequently conflated due to their semantic and structural similarity, especially when explicit lexical signals are absent or ambiguous, which is often the case for ELABORATION, the most over-predicted label in the dataset. Teasing apart CONJUNCTION and TEMPORAL is also challenging, especially in cases where consecutive events are not signaled explicitly, but implicitly form a temporal succession relation.

Strengths and Confusions at the Corpus Level Figure 3 shows the confusion matrices for the highest- and lowest-scoring datasets, `tha.pdtb.tdtb` (TDTB) and `ces.rst.crdt` (CRDT). In CRDT, the model frequently defaults to the majority relation CONJUNCTION, reflecting strong over-prediction of common classes. The label distribution is also highly imbalanced: CONJUNCTION and ELABORATION dominate the dataset, while many relations (e.g., ALTERNATIVE, REFORMULATION, MODE) appear only rarely, making them difficult for the model to learn.

Relations marked by overt lexical cues, such as CONJUNCTION, achieve high accuracy in both datasets. Performance in TDTB is further aided by the fact that this dataset covers only 12/17 possible labels, substantially reducing the possibilities for confusion compared to CRDT, which covers the full set of 17 labels.

6 Conclusion

We present **DeDisCo**, a decoder-only model, using a pruned Qwen3-4B basis, for the multilingual discourse relation classification task in the DISRPT 2025 Shared Task. Our system DeDisCo leverages supervised fine-tuning together with rich features, including metadata and instance-level cues such as unit distance, document position, and gold speaker information. We also incorporated augmented datasets to improve coverage for low-resource languages.

Our results suggest that decoder-only architectures are effective for this task, as their structure allows the model to integrate diverse sources of information (e.g., metadata, features, and context) within a unified text stream. Carefully designed instruction templates and feature injection further improve generalization, and natural prompt styles are helpful despite textual redundancy, even in a full fine-tuning setup. This enables the model to condition on a broad range of cues when making predictions, while adhering to a format the model is familiar with from its initial supervised fine-tuning. Error analysis highlights class imbalance as a persistent challenge, often leading to over-prediction of majority classes. Nonetheless, augmented data yielded measurable gains for low-resource languages, underscoring the importance of data enrichment strategies in this task.

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

Kaveri Anuranjana. 2023. DiscoFlan: Instruction finetuning 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. The Association for Computational Linguistics.

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'16)*, 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) (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.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 1–19, Suzhou, China. The Association for Computational Linguistics.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54. Portoroz, Slovenia.

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*.

Yi Cheng and Sujian Li. 2019. 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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland,

Oregon, USA. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. 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).

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eduard H. Hovy. 1990. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Linden Hall Conference Center, Dawson, Pennsylvania. Association for Computational Linguistics.

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.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Evi Judge, Reece Suchocki, and Konner Syed. 2024. An analysis of sentential neighbors in implicit discourse relation prediction. *Preprint*, arXiv:2405.09735.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. 2024. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176, St. Julians, Malta. Association for Computational Linguistics.

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.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024. 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.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1999. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 365–372, College Park, Maryland, USA. Association for Computational Linguistics.

Daniel Marcu and William Wong. 2004. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 219——222, Lisbon, Portugal. European Language Resources Association (ELRA).

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *Preprint*, arXiv:2403.03853.

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, and Philippe Muller. 2024. Feature-augmented model for multilingual discourse relation classification. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 91–104, St. Julians, Malta. Association for Computational Linguistics.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

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

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*.

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

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, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1051–1058, Portorož, Slovenia. European Language Resources Association (ELRA).

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 '14)*, pages 925–929, Reykjavik.

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.

Aleksandra Tomaszewska, Purificação Silvano, António Leal, and Evelin Amorim. 2024. ISO 24617-8 applied: Insights from multilingual discourse relations annotation in English, Polish, and Portuguese. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 99–110, Torino, Italia. ELRA and ICCL.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 133–143, Minneapolis, MN. Association for Computational Linguistics.

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, 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, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 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, Umit Deniz Turan, Cem Bozsahin, Ruket Cakici, Ayisigi B. Sevdik-Calli, Isin Demirsahin, Berfin Aktas, İhsan Yalcinkaya, and Hale Ogel. 2009. Annotating subordinators in the Turkish discourse

bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 44–47, Suntec, Singapore. Association for Computational Linguistics.

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.

A Data

We train and evaluate our models using all the datasets provided by the shared task organizers.[4] 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., 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; 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., 2014).

[4] https://github.com/disrpt/sharedtask2025

B Feature Utilization in Decoder and Encoder Models

Feature Group	Decoder	Encoder
LCF features	+	+
DiscoDisco features	+	+
Direction	+	+
Context (window size)	+	–
Augmented dataset	+	+

Table 4: Feature groups and dataset augmentation used in decoder vs. encoder models ("+" = included, "–" = excluded).

DiscoDisco Feature	Decoder	Encoder
Genre	–	+
Children	–	–
Discontinuous	–	+
Is Sentence	–	+
Length Ratio	–	–
Same Speaker	+	+
Document Length	–	–
Position	+	–
Distance	+	–
Lexical Overlap	–	–

Table 5: Detailed inclusion ("+") or exclusion ("–") of all DiscoDisco features for each model.

C Experimental Setup of Decoder

The model was trained on four NVIDIA H100 GPUs. Training one epoch took approximately three hours, and evaluation on the test sets required an additional one and a half hours. With a per-device batch size of 1, and using 16 gradient accumulation steps, the effective batch size was 64.

On rare occasions, the generative model produced outputs that were not part of the predefined set of valid labels. In such cases, our evaluation script replaced the output with a randomly selected valid label. This was extremely infrequent, occurring fewer than five times across all evaluations.

D Corpora from DISRPT 2025 Shared Task

Language Name (Code)	Framework	Corpus LFC Short Code
Czech (ces)	RST	ces.rst.crdt
Standard German (deu)	PDTB	deu.pdtb.pcc
Standard German (deu)	RST	deu.rst.pcc
English (eng)	DEP	eng.dep.covdtb
English (eng)	DEP	eng.dep.scidtb
English (eng)	eRST	eng.erst.gentle
English (eng)	eRST	eng.erst.gum
English (eng)	PDTB	eng.pdtb.gentle
English (eng)	PDTB	eng.pdtb.gum
English (eng)	PDTB	eng.pdtb.pdtb
English (eng)	PDTB	eng.pdtb.tedm
English (eng)	RST	eng.rst.oll
English (eng)	RST	eng.rst.rstdt
English (eng)	RST	eng.rst.sts
English (eng)	RST	eng.rst.umuc
English (eng)	SDRT	eng.sdrt.msdc
English (eng)	SDRT	eng.sdrt.stac
Basque (eus)	RST	eus.rst.ert
Persian (fas)	RST	f as.rst.prstc
French (fra)	SDRT	fra.sdrt.annodis
French (fra)	SDRT	fra.sdrt.summre′
Italian (ita)	PDTB	ita.pdtb.luna
Dutch (nld)	RST	nld.rst.nldt
Nigerian Pidgin (pcm)	PDTB	pcm.pdtb.disconaija
Polish (pol)	ISO	pol.iso.pdc
Portuguese (por)	PDTB	por.pdtb.crpc
Portuguese (por)	PDTB	por.pdtb.tedm
Portuguese (por)	RST	por.rst.cstn
Russian (rus)	RST	rus.rst.rrt
Spanish (spa)	RST	spa.rst.rststb
Spanish (spa)	RST	spa.rst.sctb
Thai (tha)	PDTB	tha.pdtb.tdtb
Turkey (tur)	PDTB	tur.pdtb.tdb
Turkey (tur)	PDTB	tur.pdtb.tedm
Chinese (zho)	DEP	zho.dep.scidtb
Chinese (zho)	PDTB	zho.pdtb.cdtb
Chinese (zho)	PDTB	zho.pdtb.ted
Chinese (zho)	RST	zho.rst.gcdt
Chinese (zho)	RST	zho.rst.sctb

Table 6: Collation of the data in DISRPT 2025 Shared task with the corresponding Language and Framework for reference. The frameworks include RST (Mann and Thompson, 1988), PDTB (Marcu and Wong, 2004), DEP (Stede et al., 2016), SDRT (Lascarides and Asher, 2007), eRST (Zeldes et al., 2025), and ISO 24617-8 (Bunt and Prasad, 2016; Tomaszewska et al., 2024). The languages are sorted based on the language code (ISO 639-3[5]). The corpus `fra.sdrt.summre` marked with ′ is not part of the relation classification task.

[5] https://www.iso.org/iso-639-language-code

E Ablation Test Results of Decoder

Corpus	Decoder	w/o LCF		w/o DiscoDisco		w/o Direction		w/o Context		w/o Aug	
		abs.	*gain*	*abs.*	*gain*	*abs.*	*gain*	*abs.*	*gain*	*abs.*	*gain*
ces.rst.crdt	52.70	54.73	-2.03	55.41	-2.71	46.62	6.08	50.00	2.70	57.43	-4.73
deu.pdtb.pcc	67.01	63.92	3.09	65.98	1.03	62.37	4.64	62.37	4.64	69.07	-2.06
deu.rst.pcc	67.03	59.71	7.32	64.84	2.19	48.72	18.31	51.28	15.75	64.10	2.93
eng.dep.covdtb	68.21	70.46	-2.25	73.16	-4.95	57.39	10.82	71.04	-2.83	71.81	-3.6
eng.dep.scidtb	83.66	84.29	-0.63	83.87	-0.21	82.57	1.09	81.05	2.61	83.56	0.1
eng.erst.gentle	67.08	62.93	4.15	65.99	1.09	60.27	6.81	64.62	2.46	67.32	-0.24
eng.erst.gum	73.45	71.29	2.16	73.11	0.34	64.72	8.73	66.84	6.61	72.41	1.04
eng.pdtb.gentle	67.94	65.27	2.67	67.05	0.89	63.61	4.33	67.43	0.51	66.79	1.15
eng.pdtb.gum	71.39	68.00	3.39	70.86	0.53	65.38	6.01	68.47	2.92	70.8	0.59
eng.pdtb.pdtb	83.77	83.19	0.58	84.07	-0.30	73.22	10.55	82.61	1.16	83.80	-0.03
eng.pdtb.tedm	71.79	68.95	2.84	71.23	0.56	64.67	7.12	68.09	3.70	69.52	2.27
eng.rst.oll	62.73	60.89	1.84	60.52	2.21	49.45	13.28	59.41	3.32	59.04	3.69
eng.rst.rstdt	73.27	67.38	5.89	73.41	-0.14	68.54	4.73	69.51	3.76	72.99	0.28
eng.rst.sts	58.54	55.49	3.05	56.71	1.83	45.43	13.11	50.00	8.54	54.88	3.66
eng.rst.umuc	67.36	66.53	0.83	66.53	0.83	61.16	6.2	61.16	6.20	62.4	4.96
eng.sdrt.msdc	90.00	89.75	0.25	90.03	-0.03	88.82	1.18	86.14	3.86	89.08	0.92
eng.sdrt.stac	75.80	76.33	-0.53	75.98	-0.18	70.92	4.88	70.66	5.14	74.73	1.07
eus.rst.ert	54.64	56.49	-1.85	51.96	2.68	43.30	11.34	46.60	8.04	52.16	2.48
fas.rst.prstc	60.47	59.12	1.35	59.12	1.35	51.52	8.95	50.84	9.63	59.63	0.84
fra.sdrt.annodis	60.39	56.04	4.35	61.19	-0.80	53.14	7.25	51.53	8.86	58.94	1.45
ita.pdtb.luna	70.13	70.40	-0.27	70.13	0	61.6	8.53	70.93	-0.80	72.53	-2.40
nld.rst.nldt	68.62	69.85	-1.23	69.54	-0.92	55.08	13.54	61.23	7.39	67.69	0.93
pcm.pdtb.disconaija	59.39	60.96	-1.57	61.16	-1.77	51.13	8.26	42.97	16.42	60.18	-0.79
pol.iso.pdc	74.02	73.08	0.94	73.62	0.40	62.99	11.03	53.97	20.05	72.14	1.88
por.pdtb.crpc	79.17	79.09	0.08	78.85	0.32	73.48	5.69	77.48	1.69	77.72	1.45
por.pdtb.tedm	68.41	68.68	-0.27	68.41	0	65.11	3.30	65.38	3.03	67.03	1.38
por.rst.cstn	70.22	70.96	-0.74	71.32	-1.10	69.12	1.10	70.59	-0.37	71.32	-1.1
rus.rst.rrt	74.85	74.81	0.04	75.31	-0.46	66.18	8.67	69.52	5.33	74.46	0.39
spa.rst.rststb	69.72	71.83	-2.11	70.66	-0.94	64.55	5.17	65.96	3.76	70.42	-0.70
spa.rst.sctb	83.02	77.99	5.03	80.50	2.52	70.44	12.58	76.73	6.29	86.16	-3.14
tha.pdtb.tdtb	96.73	96.88	-0.15	96.5	0.23	96.58	0.15	96.73	0	96.50	0.23
tur.pdtb.tdb	64.13	66.03	-1.90	66.75	-2.62	59.86	4.27	64.61	-0.48	66.75	-2.62
tur.pdtb.tedm	59.23	58.4	0.83	59.78	-0.55	59.5	-0.27	58.95	0.28	58.4	0.83
zho.dep.scidtb	80.00	78.6	1.40	77.21	2.79	69.77	10.23	74.42	5.58	76.28	3.72
zho.pdtb.cdtb	88.65	88.52	0.13	90.5	-1.85	83.91	4.74	87.34	1.31	88.79	-0.14
zho.pdtb.ted	75.49	76.09	-0.60	75.86	-0.37	67.97	7.52	71.95	3.54	75.79	-0.30
zho.rst.gcdt	75.55	73.66	1.89	75.13	0.42	62.96	12.59	70.51	5.04	76.71	-1.16
zho.rst.sctb	74.21	67.30	6.91	73.58	0.63	62.26	11.95	70.44	3.77	71.70	2.51
Macro Average	71.28	70.10	1.18	71.21	0.08	63.80	7.49	66.56	4.72	70.82	0.47
Micro Average	76.13	75.15	0.98	76.38	-0.25	69.53	6.60	72.24	3.89	75.86	0.27

Table 7: Accuracy results of the ablation study on the decoder-only model: next to the scores from Table 3, we report scores without LCF features, without DiscoDisco features, without direction, without context and without data augmentation, as well as the "gain" for each (non-ablated score – ablated score).

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).

SeCoRel: Multilingual Discourse Analysis in DISRPT 2025

Sobha Lalitha Devi and Pattabhi RK Rao and Vijay Sundar Ram R
AU-KBC Research Centre
MIT Campus of Anna University
Chennai, India
sobha@au-kbc.org

Abstract

The work presented here describes our participation in DISRPT 2025 shared task in three tasks, Task1: Discourse Unit Segmentation across Formalisms, Task 2: Discourse Connective Identification across Languages and Task 3: Discourse Relation Classification across Formalisms. We have fine-tuned XLM-RoBERTa, a language model to address these three tasks. We have come up with one single multilingual language model for each task. Our system handles data in both the formats .conllu and .tok and different discourse formalisms. We have obtained encouraging results. The performance on test data in the three tasks is similar to the results obtained for the development data.

1 Introduction

This paper describes our system, used in DISRPT2025 shared task "Discourse Relation Parsing and Treebanking (DISRPT)". This is a Shared Task on Discourse Segmentation, Connective and Relation Identification across Formalisms. The shared task has the following three tasks: a) Task 1- Discourse Segment Identification, b) Task 2 – Discourse Connective Identification and c) Task 3 – Relation Identification. The organizers have provided data from different languages and annotations on these data follow different discourse formalisms. One of the main goals is that only one language model has to be developed which will apply to all languages and formalisms.

Discourse relations are the coherence relations between two discourse segments or also called as Elementary Discourse Units (EDUs) that can be realized explicitly or implicitly in a text. Discourse connectives play a role in signaling the relations in a discourse. They connect two discourse units, which may be a sentence, clause or multiple sentences. The relations can be intra sentential or inter sentential i.e. within a sentence or across sentences.

Thus the main objective of the work presented here is to develop a single language model for each of the task which will work for all languages and formalisms. The pre-trained XLM-RoBERTa language model was adapted through fine-tuning. In the following sections, we give a detailed description of our system.

2 Data

We train and evaluate our models using all the datasets provided by the shared task organizers.[1] 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 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;

[1] GitHub: https://github.com/disrpt/sharedtask2025, and HuggingFace: https://huggingface.co/multilingual-discourse-hub.

Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 79–86

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; 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., 2014).

The shared task was held in 2019 (Zeldes et al., 2019), 2021 (Zeldes et al., 2021), 2023 (Braud et al., 2023) and 2025 (), with more information on the data format in (Braud et al., 2024).

3 System Description

Motivated by the works presented in the previous DISRPT 2021 and 2023 workshops, a fine-tuning strategy was chosen. XLM-RoBERTa architecture is considered suitable for multilingual tasks, like question answering, discourse parsing because it employs self-attention mechanisms to effectively capture contextual dependencies within the text.

For discourse relation identification (task 3), the problem is framed as a classification task, in which it will learn to categorize discourse relations between different parts of the text.

Tasks 1 and 2 are handled as a sequence labeling task. This approach aims to identify and label the boundaries of discourse units within a given sequence of text.

XLM-RoBERTa is a transformer network-based model framework which relies on a strong self-attention mechanism to understand and interpret context effectively. This self-attention mechanism allows the model to weigh the significance of different parts of the input sequence, irrespective of their position, leading to a more nuanced understanding of the input data (Conneau et al., 2020).

XLM-RoBERTa-base (XLM-R-B) is a multilingual language model, well-suited for this shared task. XLM-R-B has a relatively smaller parameter size of 2.55B compared to XLMR Large, which translates to fewer computational resources required for processing. This efficiency makes it a practical choice for the present shared task.

3.1 Hyper-parameter Fine Tuning

In our approach, for fine-tuning XLM-RoBERTa we follow on the work of (Wolf et al., 2019), who offered a thorough framework for training for text classification models with Hugging Face's Transformers library. Although their configuration provided a strong basis for training the model, we modified it to better fit the discourse datasets provided in the shared task. Increasing the number of epochs from the initial setting to 10 was a crucial change that enabled the model to go through more thorough training and better absorb the subtleties of the data. In order to achieve effective gradient descent during training and maximize the trade-off between stability and quick convergence, we also changed the learning rate. Refining the batch sizes was another important modification. We set the evaluation batch size at 16 and the training batch size at 8. These modifications were designed to ensure adequate data flow for model learning while managing memory limitations on our hardware. In order to avoid over fitting, we also adjusted regularization parameters like the weight decay. The model's efficiency and generalization were enhanced by these adjusted parameters in conjunction with the monitoring of training and evaluation performance. All these optimizations were same for all three tasks. We were not able to get access to licensed datasets such as pdtb, thus these datasets were trained without words.

4 Results

Evaluation was done on the outputs produced by the system using the evaluation script provided by the organizers. The results are tabulated in the Tables 1, 2, 3, 4 and 5, for each of the tasks on different file formats and languages. Table 1 and 2 display the results obtained for task 1. Table 3 and 4 display the results obtained for task 2, And Table 5 displays the results obtained for task 3.

In task1 and task 2, the major challenge was tokenizing the input data into sentences. The data being multilingual we had to employ a multilingual sentence splitter. We had developed a basic sentence splitter using heuristic rules which handles different language texts. The results for (*.tok) files evidently shows the impact of sentence splitting.

File name	Prec Dev	Prec Test	Rec Dev	Rec Test	F1 Dev	F1 Test
ces. rst.crdt	93.40	89.63	91.40	91.30	92.40	90.46
eng. erst.gum	95.60	91.80	83.00	79.84	88.80	85.40
eng. rst.sts	60.30	71.18	89.30	73.51	72.00	72.32
eng. sdrt.stac	88.60	85.02	91.20	90.00	89.80	87.45
fra. sdrt.annodis	92.13	90.65	82.19	80.00	86.88	85.05
rus. rst.rrt	90.90	91.99	92.25	92.59	91.60	92.29
zho. dep.scidtb	93.72	86.69	94.35	97.02	94.03	91.56
deu. rst.pcc	95.63	97.13	93.26	91.86	94.43	94.42
eng. rst.oll	80.00	85.76	97.14	89.99	87.74	87.79
eng. rst.umuc	87.28	89.73	91.15	85.27	89.17	87.45
eus. rst.ert	92.05	90.71	90.69	89.72	91.36	90.21
nld. rst.nldt	96.00	96.75	97.95	97.04	96.96	96.89
spa. rst.rststb	93.67	90.02	95.46	92.17	94.56	91.08
zho. rst.gcdt	91.57	93.58	84.52	82.87	87.90	87.90
eng. dep.scidtb	96.52	96.49	92.48	92.39	94.46	94.39
eng. rst.rstdt	83.37	82.66	83.41	82.73	83.39	82.70
eng. sdrt.msdc	97.00	96.36	93.97	93.55	95.46	94.94
fas. rst.prstc	92.79	93.15	93.92	93.43	93.35	93.29
por. rst.cstn	91.66	90.96	92.53	95.42	92.10	93.14
spa. rst.sctb	78.04	85.20	93.20	85.71	84.95	85.45
zho. rst.sctb	49.20	55.71	93.20	89.88	64.42	68.72
Mean	**88.00**	**88.00**	**91.00**	**89.00**	**89.00**	**88.00**

Table 1: Evaluation Results for Task 1: Discourse Segmentation (for *.tok files)

For some language files such as ita.pdtb.luna, the sentence splitting was not efficient. In general it is observed that the results obtained for *.tok files are better. One probable reason is that for these files our sentence splitting algorithm worked better.

We observe that there are many false positives in zho.rst.sctb and spa.rst.sctb which has led to high recall and low precision. In the dataset eng.rst.umuc, the system has failed to learn segment start which is with-in the sentence. This has affected both recall and precision. Similar problem is observed in eng.rst.rstdt dataset also.

In task 2, the system has identified single word connectives with high precision and recall. It has poorly identified connectives with multiple words and apostrophe such as 'the same way', 'it would be same thing if', 'because of that', 'years have passed' etc. Improving the tokenization and contextual learning will boost the accuracy of connective identification.

In the relation identification task (task 3), we observed that the major errors are in the identification of 'elaboration' and 'conjunction' relation types. 'Elaboration' relation type is confused with relation types such as 'conjunction', 'organization', 'temporal' and 'frame'. Similarly 'conjunction is confused with 'temporal', 'explanation', 'frame' and 'causal'. We need to address these two relation types for improving the accuracy of the relation identification system. We need to train the system with syntactic features.

5 Conclusion

We have submitted our test runs for all the three tasks of the DISRPT 2025 shared task. We have fine-tuned the XLM-Roberta to handle multilingual and multi-formalism data. The three models and the system runs are available in the following link:

https://drive.google.com/drive/folders/1g3Rcve50OvlEWuqDzr8twiFipP8YLGhC?usp=sharing

Acknowledgments

We thank the organizers of DISRPT 2025 for providing us the datasets and giving us opportunity to participate in this task.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac,

File name	Prec Dev	Prec Test	Rec Dev	Rec Test	F1 Dev	F1 Test
ces. rst.crdt	93.00	91.93	91.00	91.93	92.00	91.93
deu. rst.pcc	96.35	97.13	93.62	91.86	94.96	94.43
eng. dep.scidtb	96.52	96.50	92.44	92.39	94.44	94.40
eng. erst.gum	95.85	96.00	83.60	83.89	89.31	89.54
eng. rst.oll	81.52	85.62	96.07	88.89	88.20	87.22
eng. rst.rstdt	54.17	54.49	24.59	23.53	33.83	32.87
eng. rst.sts	63.40	86.09	93.47	88.39	75.56	87.22
eng. rst.umuc	87.29	90.08	91.15	85.09	89.18	87.51
eng. sdrt.msdc	97.47	96.68	93.12	92.98	95.24	94.80
eng. sdrt.stac	90.26	88.36	92.74	95.41	91.48	91.75
eus. rst.ert	92.46	90.60	90.55	89.86	91.49	90.23
fas. rst.prstc	92.80	93.16	93.92	93.58	93.36	93.37
fra. sdrt.annodis	91.02	87.35	82.01	80.42	86.28	83.74
fra. sdrt.summre	61.06	57.22	87.85	89.17	72.05	69.71
nld. rst.nldt	95.73	96.75	97.96	96.75	96.83	96.75
por. rst.cstn	91.67	90.97	92.54	95.42	92.10	93.14
rus. rst.rrt	91.26	92.28	92.72	92.65	91.98	92.47
spa. rst.rststb	93.43	90.59	94.99	92.17	94.20	91.38
spa. rst.sctb	77.42	86.14	93.20	85.12	84.58	85.63
zho. dep.scidtb	93.73	86.69	94.35	97.02	94.04	91.57
zho. rst.gcdt	93.04	93.88	85.31	84.34	89.01	88.86
zho. rst.sctb	49.75	57.14	96.12	95.24	65.56	71.43
Mean	**85.42**	**86.62**	**88.79**	**87.55**	**86.17**	**86.36**

Table 2: Evaluation Results for Task 1: Discourse Segmentation (for *.conllu files)

File name	Prec Dev	Prec Test	Rec Dev	Rec Test	F1 Dev	F1 Test
deu. pdtb.pcc	80.85	81.31	86.36	78.72	83.51	79.99
eng. pdtb.gum	87.64	87.89	92.69	86.00	90.10	86.93
eng. pdtb.tedm	79.00	83.41	71.81	76.19	75.23	79.63
ita. pdtb.luna	81.37	68.29	59.71	53.63	68.87	60.08
pcm. pdtb.disconaija	64.48	74.09	56.09	73.71	60.00	73.90
pol. iso.pdc	71.20	70.20	57.66	63.61	63.72	66.74
por. pdtb.crpc	82.95	81.62	76.81	71.87	79.76	76.44
por. pdtb.tedm	75.49	77.40	75.49	79.31	75.49	78.34
tha. pdtb.tdtb	-	-	-	-	-	-
tur. pdtb.tedm	78.33	78.43	34.81	32.38	48.20	45.84
Zho. pdtb.ted	78.54	-	82.68	-	80.56	-
Mean	**77.98**	**78.08**	**69.41**	**68.38**	**72.54**	**71.98**

Table 3: : Evaluation Results for Task 2: Discourse Connective Identification (for .tok files)

File name	Prec Dev	Prec Test	Rec Dev	Rec Test	F1 Dev	F1 Test
deu.pdtb.pcc	80.00	81.32	86.36	78.72	83.06	79.99
eng.pdtb.gum	86.55	90.65	83.43	82.23	84.96	86.23
eng.pdtb.tedm	79.00	83.89	71.82	76.62	75.24	80.09
ita.pdtb.luna	78.49	64.11	52.51	51.34	62.93	57.02
pcm.pdtb.disconaija	61.89	70.47	47.97	65.21	54.04	67.74
pol.iso.pdc	71.28	71.39	57.88	65.02	63.88	68.06
por.pdtb.crpc	83.24	81.42	76.81	71.69	79.89	76.25
por.pdtb.tedm	74.76	76.92	75.49	78.82	75.12	77.86
tha.pdtb.tdtb	75.27	77.15	84.23	86.16	79.49	81.40
tur.pdtb.tedm	76.27	78.43	33.33	32.39	46.39	45.84
zho.pdtb.ted	78.67	69.10	82.68	82.03	80.62	75.02
Mean	**76.86**	**76.80**	**68.41**	**70.02**	**71.42**	**72.32**

Table 4: : Evaluation Results for Task 2: Discourse Connective Identification (for .conllu files)

File name	Dev Data Accuracy	Test Data Accuracy
ces.rst.crdt	39.02	42.57
deu.pdtb.pcc	54.17	59.28
deu.rst.pcc	42.31	45.05
eng.dep.covdtb	66.28	68.10
eng.dep.scidtb	81.27	79.84
eng.erst.gentle	-	44.95
eng.erst.gum	43.61	46.89
eng.pdtb.gentle	-	46.31
eng.pdtb.gum	49.23	49.58
eng.pdtb.pdtb	28.29	26.92
eng.pdtb.tedm	49.44	53.28
eng.rst.oll	53.23	41.33
eng.rst.rstdt	10.12	10.90
eng.rst.sts	40.80	35.06
eng.rst.umuc	58.10	59.09
eng.sdrt.msdc	85.80	84.86
eng.sdrt.stac	65.88	67.64
eus.rst.ert	51.30	54.64
fas.rst.prstc	52.10	51.52
fra.sdrt.annodis	59.27	51.85
ita.pdtb.luna	60.68	65.07
nld.rst.nldt	51.06	55.69
pcm.pdtb.discon	54.47	56.54
pol.iso.pdc	47.89	50.07
por.pdtb.crpc	69.11	73.88
por.pdtb.tedm	58.42	64.29
por.rst.cstn	61.78	61.40
rus.rst.rrt	60.11	62.70
spa.rst.rststb	69.19	57.98
spa.rst.sctb	65.96	65.41
tha.pdtb.tdtb	95.66	96.21
tur.pdtb.tdb	25.40	24.94
tur.pdtb.tedm	50.71	49.04
zho.dep.scidtb	65.48	67.44
zho.pdtb.cdtb	60.81	58.92
zho.pdtb.ted	59.74	59.92
zho.rst.gcdt	60.44	55.93
zho.rst.sctb	52.13	55.97
Average	**52.61**	**55.29**

Table 5: Evaluation for Task 3 Discourse Relation Classification

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).

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

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

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 Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Alexis Conneau, KartikayKhandelwal, NamanGoyal VishravChaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, MyleOtt, Luke Zettlemoyer, and VeselinStoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

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 2012)*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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.

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. Multilayer 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).

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.

T Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, and M Funtowicz. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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, 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, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 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.

Author Index