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*

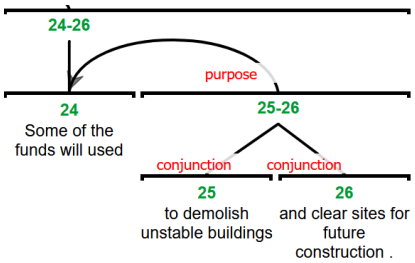

Figure 1: An RST tree example from RST-DT, visualized with rstWeb (Gessler et al., 2019).

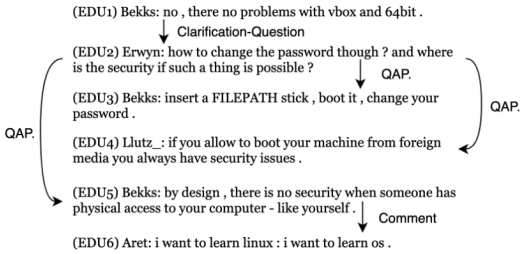

Figure 2: An SDRT graph (Liu and Chen, 2021).

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`.

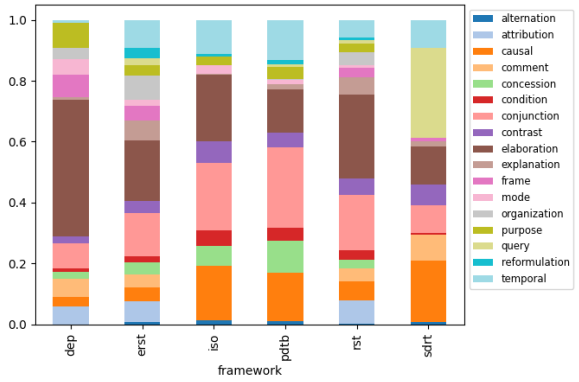

Figure 3: Distribution of the unified label set per framework in the DISRT 2025 datasets.

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.