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).