DISRPT: A Multilingual, Multi-domain, Cross-framework Benchmark for Discourse Processing

Chloé Braud[1,2,3], **Amir Zeldes**[4], **Laura Rivière**[1], **Yang Janet Liu**[4], **Philippe Muller**[1,3], **Damien Sileo**[2,5], **Tatsuya Aoyama**[4]
[1]IRIT, University of Toulouse ; [2]CNRS ; [3]ANITI ; [4]Georgetown University ;
[5]Univ. Lille, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[1]firstname.lastname@irit.fr
[4]{amir.zeldes,yl879,ta571}@georgetown.edu
[5]damien.sileo@inria.fr

Abstract

This paper presents DISRPT, a multilingual, multi-domain, and cross-framework benchmark dataset for discourse processing, covering the tasks of discourse unit segmentation, connective identification, and relation classification. DISRPT includes 13 languages, with data from 24 corpora covering about 4 millions tokens and around $250,000$ discourse relation instances from 4 discourse frameworks: RST, SDRT, PDTB, and Discourse Dependencies. We present an overview of the data, its development across three NLP shared tasks on discourse processing carried out in the past five years, and the latest modifications and added extensions. We also carry out an evaluation of state-of-the-art multilingual systems trained on the data for each task, showing plateau performance on segmentation, but important room for improvement for connective identification and relation classification. The DISRPT benchmark employs a unified format that we make available on GitHub and HuggingFace in order to encourage future work on discourse processing across languages, domains, and frameworks.

Keywords: discourse, corpora, multilingual, transfer

1. Introduction

Computational approaches to discourse processing often reveal the implicit organization of texts through semantic-pragmatic relations, such as *explanation* or *contrast*, which link spans of text and form possibly hierarchically ordered subparts of longer pieces of discourse. Various frameworks exist to describe this organization, underlying several annotation projects. While having similar objectives, these frameworks differ in the way they define discourse units, relation labels, and structures of discourse. As a result, annotated corpora according to these frameworks present important discrepancies, which tend to split the domain between approaches dedicated to a specific framework only (see Demberg et al. 2019). In a sense, this situation increases the data scarcity issue we face for work on computational models of discourse. In addition, even within the same framework, specific choices made by annotation teams lead to important differences. This hinders the development of multilingual systems and prevents robust evaluation across languages or domains.

In order to address these issues, we present the DISRPT dataset (**DIS**course **R**elation **P**arsing and **T**reebanking), an effort toward converting existing discourse corpora within a unified format. DISRPT can be seen as a benchmark currently consisting of 28 datasets – from 24 corpora[1] – with annotations for three tasks related to discourse analysis. The benchmark covers 4 frameworks, 13 languages, and multiple domains, and its unified format has been developed within the context of an international shared task held in 2019,[2] 2021,[3] and 2023,[4] with new corpora or tasks included for each edition. Contrary to previous work where unification was mostly done via label mappings (e.g. Benamara and Taboada, 2015; Braud et al., 2017), the goal is to provide unified formats while remaining as faithful as possible to the original annotations, to allow cross-framework investigation.[5]

Currently, the benchmark consists of three tasks: (1) discourse segmentation, (2) discourse connective identification, and (3) discourse relation classification. In addition, for (1) and (2), there are two tracks: (a) treebanked: documents are split into sentences, and dependency syntax information is given (either gold if available or predicted), (b) plain: documents are only tokenized. Finally, the last shared task introduced *out-of-domain* (OOD) datasets, providing some smaller evaluation-only

[1]A dataset is considered a combination of one corpus, one language, and one framework, meaning that we derive several datasets from e.g. the multilingual TED corpus (Zeyrek et al., 2018), which corresponds to 3 datasets (English, Portuguese, and Turkish) in the DISRPT benchmark.

[2]https://sites.google.com/view/disrpt2019/

[3]https://sites.google.com/georgetown.edu/disrpt2021

[4]https://sites.google.com/view/disrpt2023/

[5]Note that discourse structure, for which some work has proposed unification process (e.g. Yi et al., 2021), is not part of the current release of the benchmark.

sets, in order to test systems' transfer abilities.

This work heavily relies on all the work done by the DISRPT shared task organizers in proposing a unified format, but the goal and contributions of this paper are different.

First, we thoroughly expose and explain the conversion process from original annotations to the unified format, which was not explored in previous papers that were mainly centered around the systems comparisons. Compared to the last edition of the shared task (Braud et al., 2023), we propose some modifications to the data after having spotted mistakes, and taking into consideration possible consistency improvements (e.g. lower-casing all relation labels reduces the label space without deviating from original annotations). We also provide a more detailed description of existing frameworks, making it easier for anyone not familiar with all these theories to understand the conversion. These analyses should help future researchers to understand the purpose, the difficulties and the limitations of the conversion process and to understand where improvements can still be made to this benchmark.

Furthermore, we provide descriptive information about the datasets to highlight similarities and discrepancies between annotation projects, with the latter also corresponding to potential obstacles for automatic systems. By highlighting sources of heterogeneity within corpora, our goal is to provide new insights to guide future discourse annotation projects.

Additionally, we augment the benchmark with a new out-of-domain dataset for English: GENTLE (Aoyama et al., 2023), a challenging corpus consisting of varied and unusual genres, which is particularly helpful to test robustness. An additional layer of the English GUM corpus (Zeldes, 2017) is also made available for the first time, the annotation of discourse connectives, making it the first dataset to support all three tasks.

Finally, we present experiments with the two highest scoring state-of-the-art systems for all three tasks: **DisCut** for discourse segmentation and connective identification (Metheniti et al., 2023) and **DisCoDisCo** for discourse relation classification (Gessler et al., 2021).[6]

With the DISRPT benchmark, our goal is to encourage future work on automatic discourse analysis for multiple languages and domains, and to promote convergence of resources. By describing in more detail the composition and format of the datasets, we want to make it clear that this benchmark streamlines the evaluation of discourse analysis on many languages and domains. We hope this effort towards unifying corpora will enhance the understanding of the variations between frameworks and their effects on automatic systems. In order to make the DISRPT benchmark easier to use, we develop a new version of the data that will be updated over time and that can be directly uploaded to HuggingFace within the Discourse Hub[7] in addition to a GitHub repository.[8]

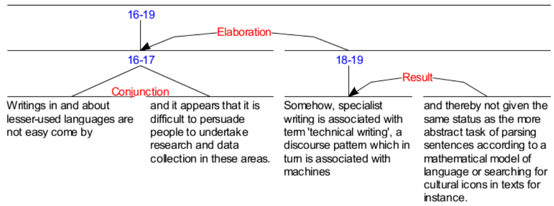

Figure 1: RST Tree (Iruskieta et al., 2015).

2. An Overview of Discourse Annotation

Each discourse framework has different aims and presents specific features in the way they describe their constitutive elements. We briefly present existing frameworks and their main differences. Examples of discourse structures in different representations are given in Figures 1 and 2.

2.1. Discourse Frameworks

RST: One of the earliest discourse frameworks proposed in studying computational discourse modeling is Rhetorical Structure Theory (RST, Mann and Thompson 1988), where structures form hierarchical, projective (connecting only adjacent nodes), labeled constituent trees (Figure 1). Discourse units exhaustively cover a text and may either form the nucleus (central proposition) or satellite (supporting proposition) of a larger unit, which enters into a similar relation with other units recursively. The definitions of the relations are based on authors' or speakers' intents: for example, EVIDENCE relations connect satellite units presented by a speaker or writer with the intent of increasing the hearer/reader's belief in the content of the corresponding nucleus. RST has led to several corpora and to the largest number of discourse parsers (e.g. Sporleder and Lascarides 2004; Joty et al. 2015; Ji and Eisenstein 2014; Wang et al. 2017; Liu et al. 2021; Kobayashi et al. 2022).

[6] Note that DisCoDisCo did not participate in DISRPT 2023, but their scores in 2021 were higher than the 2023 winner HITS (Liu et al., 2023) on the common corpora, it thus corresponds at the moment to the state-of-the-art on the relation task.

[7] https://hf.co/datasets/multilingual-discourse-hub/disrpt

[8] https://github.com/disrpt/latest

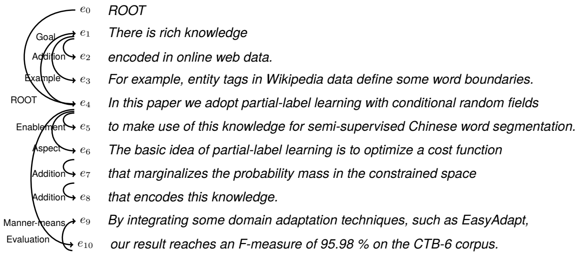

Figure 2: Dependency Tree (Yang and Li, 2018).

SDRT: Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003) is more recent and adds two main differences: structures are graphs rather than trees (no adjacency constraint), and relations are based on semantic and pragmatic constraints expressed with formal logics. Only a few corpora have been annotated in SDRT, but the STAC corpus (Asher et al., 2016) led to many discourse parsers dedicated to dialogues such as Shi and Huang (2019); Wang et al. (2021); Liu and Chen (2021); Li et al. (2023).

DEP: Building upon several studies proposing to encode discourse structures as dependencies (Hirao et al., 2013; Muller et al., 2012), Li et al. (2014) proposed to annotate discourse graphs using pure dependency structures, with no non-terminal nodes (Figure 2), while often keeping relations and segmentation rules from RST— this is abbreviated here as the DEP framework. This framework is for now limited to 3 corpora with associated parsers (Li et al., 2014; Yang and Li, 2018; Nishida and Matsumoto, 2022).

PDTB: The Penn Discourse Treebank (PDTB, Prasad et al. 2005) proposes a lexically grounded approach: rather than trying to produce full structures over entire documents, connectives (e.g. *because, while ...*) are identified along with the textual spans they connect (i.e. their *arguments*), and a sense label is then applied to the connective and spans. This process is extended to annotate relations that are not explicitly marked (i.e. *implicit* relations) but where a connective could have been inserted (i.e. human annotators would accept insertion of *because* etc.) based mainly on adjacent sentences within PDTB2 (Prasad et al., 2008), and applied to some intra-sentential relations in PDTB3 (Webber et al., 2019). The relation senses are organized within a 3-level hierarchy corresponding to coarse-to-fine-grained distinctions. With sparser annotations, as not every proposition implies a discourse relation, this framework has produced the largest corpora that are mainly used for the tasks of discourse connective and relation identification (Knaebel and Stede, 2023; Long and Webber, 2022).

2.2. Discourse Elements

Most discourse corpora follow one of the four frameworks presented above, and the annotations cover different discourse elements.

Elementary Discourse Unit (EDU): the minimal span of text to be linked by a discourse relation, in general a clause, and usually at most a sentence; the set of EDUs fully covers a document without any overlap. EDUs are segmented in corpora annotated within RST, SDRT, and DEP but not PDTB where the notion of arguments (of a connective or relation) is used instead. These arguments combined are not supposed to give a full coverage and may overlap for multiple relations, making the notion less structurally constrained. For this reason, PDTB-style datasets are not included in the segmentation task of DISRPT.

Another distinction exists between RST and DEP compared to SDRT-based corpora: in the former, nested EDUs are segmented separately as shown in Example 1 where 3 EDUs are identified. Then a pseudo-relation (*same-unit*) connects EDUs 1 and 3 to indicate they are in fact the same unit. Within SDRT, annotators would directly annotate embedded EDUs, which would lead to only 2 EDUs in this example: one consisting of EDUs 1 and 3, and one covering only EDU 2.

(1) $[$But maintaining the key components $(\ldots)]_1$ $[$– a stable exchange rate and high levels of imports –$]_2$ $[$will consume enormous amounts $(\ldots).]_3$ (Carlson and Marcu, 2001)

The task associated with this level of information is **discourse segmentation**: the goal is to identify EDU boundaries, and data from RST, DEP, and SDRT therefore looks the same: the beginning point of units 1, 2, and 3 must be identified in all three frameworks. To have homogenous representations with the other frameworks, SDRT embedding units are split at the location of embedded units.

Discourse Connective: a word or expression that can be used to trigger a discourse relation, e.g. *but*, *as soon as*. These lexical elements are the basis of annotation in PDTB-like corpora, but are rarely annotated within other frameworks (but we release here the connective annotations for the GUM corpus).[9] Connectives can consist of a single token (*while*) or multiple tokens (*as soon as*), and the same string can sometimes be used as a discourse connective and sometimes

[9]There exist annotations of discourse signals for English RST-DT (e.g. Das and Taboada 2017) and German PCC (Stede and Neumann, 2014), but these lack some alignment information and are not included in DISRPT.

not (e.g. *and* connecting sentences vs. connecting nouns). Moreover, connectives can be discontinuous, e.g. *if...then*. Finally, connectives can be modified, e.g. the expression *18 months after* is annotated as an explicit trigger of a temporal relation in the English PDTB, with *after* the head connective modified by *18 months*. The task here is **connective identification** (sometimes called detection or disambiguation): one needs to decide whether an expression is used as a discourse marker.

Discourse Relation: the label of the semantic-pragmatic relation that holds between two or more discourse units. Relations are defined using different criteria depending on the underlying frameworks (Section 2.1), and each annotation project proposes its own label set, possibly modified from existing corpora depending on the goal of the annotation or the genre of the text. Examples of typical discourse relations include causal, comparative, conditional, and temporal types. The corresponding task is **discourse relation classification**: the goal is to find the right label associated with a pair of textual segments among a typically large set of possible labels.

Discourse Structure: the attachment links between discourse units, forming a constituent tree in RST, a dependency tree in DEP, and a graph in SDRT. For these frameworks, the annotation goal was to build a full structure covering the entire documents where discourse units are linked together. The annotation consists in linking / attaching discourse units, then labelling the type of link using discourse relations. Note that this is not a goal in PDTB-like corpora where full coverage of the text via discourse units or relations is not guaranteed. This aspect of discourse annotation is not yet implemented as a task in the DISRPT shared task, though the entire graph structure of a document is represented in the information used for the tasks above (discourse unit locations, and which ones are connected to which/with what labels).

2.3. Original File Formats

Several formats exist for discourse annotations:

- PDTB format: a pipe delimited format representing a stand-off annotation for discourse connectives and relations
- RST formats: different types of files exist (`dis`, `lisp`, `rs3`, `rsd`), either in a bracketed plain text or an XML encoding of the discourse trees
- SDRT format: XML encoding or specific textual format of the full discourse graph
- DEP format: distributed as XML, JSON, or in a tabular format (`rsd`)

3. Conversion Process

The rules for the conversion are to produce a unified format covering different frameworks, and to remain as faithful as possible to the original annotations. A few modifications were necessary in order to homogenize annotations across corpora.

3.1. Proposed Formats

The proposed format has been designed to be easy to use. There are two types of files: the CoNLL-U format, used for connective identification and EDU segmentation, is adopted from the Universal Dependencies project (de Marneffe et al., 2021), which has already been widely used in the community, and a dedicated tabular `rels` format for relation classification.

Segmentation and Connectives: CoNLL-U files The segmentation and connective annotations are both token-level annotations and cast within a `BIO` scheme: a token either starts an EDU or not, or is part of a discourse connective or not. More precisely, segmentation only includes `B` labels (i.e. initial boundary of an EDU) and `O` labels for all the other tokens. For connectives, a token can be labeled `B` if it marks the beginning of a connective. The `I` label is used for multi-word connectives: for example, *in the meantime* corresponds to a sequence of `B I I`. Label `O` is used for all other tokens. We modify the existing label format to be closer to a `BIO` scheme, with a pair of key=value conforming to the CoNLL-U format (e.g. `BeginSeg=Yes` becomes `Seg=B-seg`). The exact labels are given in Appendix A.

The final format is a CoNLL-U file where each line corresponds to a token and the label is given in the last column (see data in the repository for examples). Meta data is used to indicate the start of a new document via a CoNLL-U hashtag comment line. Finally, note that there are 2 tracks for these tasks: the *treebanked* track, where sentence boundaries, morpho-syntactic, and syntactic information are made available (either gold or predicted); and a *plain* track, where neither morpho-syntactic information nor sentence boundaries are available. For the latter, the files have the extension `.tok` instead of `.conllu`, but the format is identical, except that morpho-syntactic columns are filled with underscores instead of POS tags etc.

Discourse Relations: `rels` files For the relations classification task, a different format was proposed, where each line corresponds to a pair of text spans and the associated relation, with additional information: token ids of each pair, the sentence to which each argument belongs, and the direction of the relation. The last column contains the label to be predicted within the shared task, and the penultimate column the original labels before conversion (see Section 3.2.3).

In the shared task data, the arguments are presented in the order of the text: `unit1` and `unit2` are linearly ordered, i.e. `unit1` appears before `unit2`. However, previous work on relation classification uses annotations where the pairs are ordered following the *direction* of the relation. Some discourse relations are indeed asymmetrical / oriented: *cause(`unit1`,`unit2`)* means that `unit1` is the cause for `unit2`, while it is reversed for *cause(`unit2`,`unit1`)*. The decision to propose arguments ordered linearly, with an additional column indicating the direction, makes for a more realistic scenario since this information is not known by discourse parsers, but probably corresponds to a more difficult task: existing systems act as if they knew the direction, while predicting it could be hard. In order to encourage work on this aspect, the current release includes both options, as the HuggingFace interface allows to choose whether to encode relation direction in a column and serialize the connected units in text order, or to use the serialization order to indicate the direction.

3.2. Modifications

3.2.1. Segmentation

SDRT corpora have embedded EDUs, while this is not the case for RST/DEP corpora (Section 2.2). The shared task organizers decided to reduce the EDU segmentation to a binary task for all corpora, thereby transforming discontinuous EDUs into separated EDUs in SDRT datasets, while keeping the RST ones unchanged. It can be seen as a simplification of the task, and this information should be retrieved in order to perform full discourse parsing. Note that the arguments of the relations are given in full form: split EDUs are merged to correspond to the full arguments of the relations.

3.2.2. Discourse Connective

The overall rare discontinuous connectives, such as *if ... then*, are modeled as two separate connective spans for simplicity. These connectives are very infrequent in the English PDTB, and most previous studies focused on detecting the first part (Lin et al., 2010), but further studies are needed to investigate their frequency for other languages.

Additionally, for some corpora, the annotation covers both the head connective and its modifiers: in English PDTB, one has to identify expressions such as *18 months after* or *at least partly because*, while in English GUM the task is limited to head connectives, i.e. *after* and *because*. Keeping modifiers could be seen as more realistic, since it is the whole expression that triggers the relation, it is also faithful to the original annotation. That said, it leaves some heterogeneity in the task as annotations were not done this way in all PDTB-style corpora (see connective sets in Appendix Table 8).

Finally, note that in this new release, we correct an error on the encoding of discontinuous connectives for one dataset (thai.pdtb.tdtb), resulting in many more connective instances (see Table 1).

3.2.3. Discourse Relations

Non-binary Relations: Relations are not all binary (e.g. *list* in RST-DT), but they are binarized following standard practice (Soricut and Marcu, 2003).[10]

Complex Discourse Units: Relations can hold between EDUs or involve a complex discourse unit, i.e. a discourse unit consisting of sub-units linked by discourse relations: the algorithm to retrieve head units is based on the nuclearity principle (RST/DEP corpora) and relation types (SDRT subordinating and coordinating relations) in order to always have relations between EDUs.

Label Sets: A few modifications to the relation labels have been made for the shared task to homogenize the different label sets. Originally, we count 370 distinct labels in total; the 2023 edition of the shared task had 191 labels for classification. Our additional modifications lead to 152 labels.

- as usually done, labels of the English RST-DT are reduced to coarse-grain classes described in Carlson et al. (2001), and only level-2 relations are used for PDTB-3 annotations;
- labels in other languages are all translated to English: e.g. *testuingurua* in Basque becomes *background*;
- labels corresponding to a spelling error or a minor change in unusual spelling of a label are modified (e.g. *backgroun* becomes *background*, *topichange* becomes *topic-change*)

In addition, the following steps have also been taken in this release:

1. lower-case all labels (from 370 to 316 original labels in total):
2. remove 1 relation in ita.pdtb.luna, that does not correspond to a label (4 instances overall);

[10] Relations are not always binary; we follow the common practice in discourse parsing of binarizing all relations by creating an additional instance for each subsequent member of, e.g. an *n*-way *contrast* relation.

3. use full labels instead of top-level class for GUM (as for GCDT, using the exact same labels); similarly, full Level-2 labels (*Temporal.synchrony*) are used for the English PDTB instead of single senses (*synchrony*);

4. use the first sense annotated instead of the least frequent. The initial choice was made to reduce sparsity and promote semantically rich relations, but it is not the most common setting, possibly hindering direct comparisons. This leads to 3 fewer relations in total, all from the por.pdtb.crpc dataset, as they only appear as a second sense: *hypophora* $(1$ occurrence in train), *qap.hypophora* $(14$ instances), and *qap* (21).[11] It is crucial for future work to understand the distribution of these multiple annotations and which annotation to choose.

3.3. Preprocessing

The proposed CoNLL-U format includes some preprocessing. First, data is tokenized: this is made necessary by the `BIO` encoding where labels are associated with specific tokens. In addition, sentence boundaries, morpho-syntactic, and syntactic annotations are provided, as well as annotations of multi-word expressions. This information is either gold if available, or predicted: in that case, it either comes with the original corpora or was added by the shared task organizers. In the latter case, Stanza (Qi et al., 2020) was the main tool used. We provide the preprocessing information for each dataset in Table 7 in Appendix D.

The *plain* track allows to evaluate discourse segmentation and connective identification in a realistic scenario, from a tokenized raw text. Having tokenized data makes the comparison between different automatic tools (sentence splitters, syntactic parsers) difficult, while they could have important influence on performance (Gessler et al., 2021; Metheniti et al., 2023). Note also that the absence of sentence boundaries has an effect on evaluation for segmentation: since sentence boundaries are always EDU boundaries, most of existing studies on the task only evaluated the intra-sentence segmentation, thus considering the sentence segmentation as a solved task, while performance is still low especially for languages other than English, or specific domains. To help comparisons, we provide an evaluation script including intra-sentential scores when sentences are gold.

[11] Note that the original relations are still present in the `rels` files in the penultimate column.

4. DISRPT Benchmark

4.1. Data Composition

The DISRPT benchmark consists of 28 datasets converted from 24 corpora covering 4 frameworks, 13 languages, and multiple genres or domains. Table 1, modified from Braud et al. (2023), provides detailed statistics on all DISRPT datasets regarding their sizes and properties. Each dataset is associated with a name normalized based on the name of the original corpus, the language, and the framework. The RST Discourse Treebank, for example, is called **eng.rst.rstdt**, and the English PDTB is called **eng.pdtb.pdtb**. The list of abbreviations for all covered languages is given in Table 4 in Appendix B.

Compared to the 2023 edition of the Shared Task, this benchmark consists of an additional corpus, **eng.rst.gentle**, a small corpus covering different genres but limited to an evaluation set (Aoyama et al., 2023), and a new annotation layer of GUM, here called **eng.pdtb.gum**, corresponding to $6,515$ connectives for now without the corresponding relations, an important effort to better understand the links between different frameworks. In addition, changes can be observed in some label sets and instance counts, due to the modifications described in Section 3.2.

4.2. Data Statistics

Datasets vary in many aspects. First, the size of the datasets goes from about 6k to 8k tokens for the OOD TED datasets to more than 1 million for the largest one, the English PDTB. Two frameworks cover most of the datasets: 13 for RST and 10 for PDTB. We count more RST corpora but they make for less data when considering the total number of tokens $(1,283,530$ tokens against $2,413,112$ for PDTB). Note however that half of the data for the PDTB framework comes from the English one, the other PDTB corpora are more comparable with the RST ones. In terms of languages, English and Chinese are well-represented (resp. 9 and 4 datasets), but we have some variety with 11 other languages covered, including some low resource ones such as Thai and Farsi. Many genres and domains are covered, but we note that dialogues and speech only correspond to very small datasets, and there is a need for more resources for these text types.

Concerning EDU segmentation, as mentioned earlier, sentence boundaries are always EDU boundaries, but annotation rules vary a lot when it comes to intra-sentential boundaries. It is striking in Figure 3 that *eng.sdrt.stac* and *eng.rst.gum* almost have the same number of sentences but vary considerably in terms of number of EDUs, with

Corpus	Domain	#Docs	#Sents	#Tokens	Vocab	#EDUs	#Conn	#Labels	#Rels	References
Tasks 1 and 3: EDU Segmentation and Relation Classification										
deu.rst.pcc	newspaper commentaries	176	2,193	33,222	8,359	3,018	-	26	2,665	Potsdam Commentary Corpus (Stede and Neumann, 2014)
**eng.dep.covdtb	scholarly paper abstracts on COVID-19 and related coronaviruses	300	2,343	60,849	8,293	5,705	-	12	4,985	COVID-19 Discourse Dependency Treebank (COVID19-DTB) (Nishida and Matsumoto, 2022)
eng.dep.scidtb	scientific articles	798	4,202	102,493	8,700	10,986	-	24	9,904	Discourse Dependency TreeBank for Scientific Abstracts (SciDTB) (Yang and Li, 2018)
**eng.rst.gentle	multi-genre	26	1,334	17,797	4,135	2,708	-	31	2,540	Genre Tests for Linguistic Evaluation (GENTLE) (Aoyama et al., 2023)
eng.rst.gum	multi-genre	213	11,656	203,879	19,404	26,252	-	14	24,688	Georgetown University Multilayer corpus V9 (Zeldes, 2017)
eng.rst.rstdt	news	385	8,318	205,829	19,160	21,789	-	17	19,778	RST Discourse Treebank (Carlson et al., 2001)
eng.sdrt.stac	dialogues	45	11,087	52,354	3,967	12,588	-	16	12,235	Strategic Conversations corpus (Asher et al., 2016)
eus.rst.ert	medical, terminological and scientific	164	2,380	45,780	13,662	4,202	-	29	3,825	Basque RST Treebank (Iruskieta et al., 2013)
fas.rst.prstc	journalistic texts	150	2,179	66,694	7,880	5,853	-	17	5,191	Persian RST Corpus (Shahmohammadi et al., 2021)
fra.sdrt.annodis	news, wiki	86	1,507	32,699	7,513	3,429	-	18	3,338	ANNOtation DIScursive (Afantenos et al., 2012).
nld.rst.nldt	expository texts and persuasive genres	80	1,651	24,898	4,935	2,343	-	32	2,264	Dutch Discourse Treebank (Redeker et al., 2012)
por.rst.cstn	news	140	2,221	58,793	7,786	5,537	-	32	4,993	Cross-document Structure Theory News Corpus (Cardoso et al., 2011)
rus.rst.rrt	blog and news	332	23,044	473,005	75,285	41,532	-	22	34,566	Russian RST Treebank (Toldova et al., 2017)
spa.rst.rststb	multi-genre	267	2,089	58,717	9,444	3,351	-	28	3,049	RST Spanish Treebank (da Cunha et al., 2011)
spa.rst.sctb	multi-genre	50	516	16,515	3,735	744	-	25	692	RST Spanish-Chinese Treebank (Spanish) (Cao et al., 2018)
zho.dep.scidtb	scientific	109	609	18,761	2,427	1,407	-	23	1,298	Discourse Dependency TreeBank for Scientific Abstracts (SciDTB) (Yi et al., 2021; Cheng and Li, 2019)
zho.rst.gcdt	multi-genre	50	2,692	62,905	9,818	9,706	-	31	8,413	Georgetown Chinese Discourse Treebank (GCDT) (Peng et al., 2022b,a)
zho.rst.sctb	multi-genre	50	580	15,496	2,973	744	-	26	692	RST Spanish-Chinese Treebank (Chinese) (Cao et al., 2018)
Tasks 2 and 3: Connective Detection and Relation Classification										
eng.pdtb.gum	multi-genre	213	11,656	203,879	19,404	-	6,515	-	-	Georgetown University Multilayer corpus V9 (Zeldes, 2017)
eng.pdtb.pdtb	news	2,162	48,630	1,156,657	48,937	-	26,048	23	47,851	Penn Discourse Treebank (Prasad et al., 2014; Webber et al., 2019)
**eng.pdtb.tedm	TED talks	6	381	8,048	1,881	-	341	20	529	TED-Multilingual Discourse Bank (English) (Zeyrek et al., 2018, 2019)
ita.pdtb.luna	speech	60	3,753	26,114	2,392	-	1,071	15	1,544	LUNA Corpus Discourse Data Set (Tonelli et al., 2010; Riccardi et al., 2016)
por.pdtb.crpc[12]	news, fiction, and didactic/scientific texts	302	5,194	186,849	22,208	-	5,159	19	11,330	Portuguese Discourse Bank (CRPC) (Mendes and Lejeune, 2022; Généreux et al., 2012)
**por.pdtb.tedm	TED talks	6	394	8,190	2,162	-	305	20	554	TED-Multilingual Discourse Bank (Portuguese) (Zeyrek et al., 2018, 2019)
tha.pdtb.tdtb	news	180	6,534	256,523	11,789	-	10,864	21	10,865	Thai Discourse Treebank (TDTB)
tur.pdtb.tdb	multi-genre	197	31,196	487,389	88,923	-	8,748	23	3,185	Turkish Discourse Bank (Zeyrek and Webber, 2008; Zeyrek and Kurfalı, 2017)
**tur.pdtb.tedm	TED talks	6	410	6,143	2,771	-	382	23	577	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)

Table 1: General Statistics of DISRPT Datasets: $**$ indicates an OOD dataset. '#Docs', '#Sents', '#Tokens' and '#EDUs' correspond resp. to the total number of documents, sentences (Treebanked track), tokens, and EDUs. #Conn is the number of tokens starting a connective, and 'Vocab' of unique tokens. '#Labels' is the size of the respective label set and '#Rels' to the total number of pairs annotated.

the latter corresponding to far more intra-sentential EDUs, allegedly leading to a harder task. In that particular case, the difference can be due to the genre of the datasets: *eng.sdrt.stac* contains chat conversations, and the notion of sentences is in fact not exactly the same as in written texts.

As mentioned earlier, connectives are not annotated the same way in all (PDTB) corpora: sometimes modifiers are included, and sometimes they are not, and the type of modifiers can differ. If we count the number of expressions to be identified in each corpus (i.e. single tokens annotated

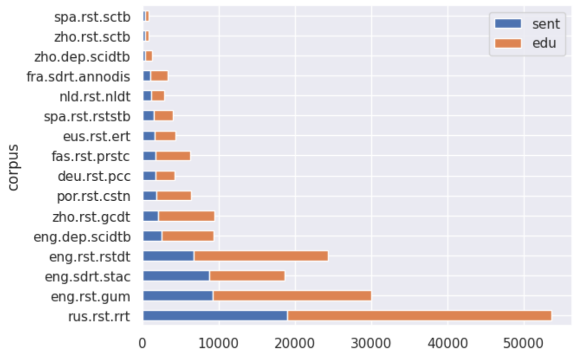

Figure 3: Number of sentences and EDU segments in RST/SDRT/DEP corpora.

with 'B' or sequences of `B I I`), we find indeed large differences: *ita.pdtb.luna* has a lexicon of 61 different forms while *eng.pdtb.pdtb* counts $1,231$ items. This aspect introduces heterogeneity in the task, probably harder with more diversified forms, and work on this task should clearly mention what part of the annotation was taken, e.g., the CoNLL shared tasks only included head connectives (Xue et al., 2015). The exact counts are given in Appendix E.

5. Experiments

We test the best systems for the three tasks of the shared task using the latest release of the DISRPT benchmark, thereby presenting SOTA results on these tasks.

DisCut (Metheniti et al., 2023) was the winning system of the 2023 Shared Task (Braud et al., 2023) on discourse segmentation, and its performance on connective detection was on par with the winning system. DiscoDisco (Gessler et al., 2021) won the 2021 competition for the discourse relation classification task (Zeldes et al., 2021) and was not tested during the 2023 edition, but the reported scores are better than the 2023 winning system on common corpora, justifying its use in this paper.

5.1. Experimental Setting

DisCut is based on a Transformer architecture with an additional linear layer for token classification. The aim was to provide a single model for all corpora by using a multilingual language model, and the version used is based on `XLM-RoBERTa-large` (Conneau et al., 2020), with the first 6 layers being frozen. The only modification needed is due to a change in the labels for segmentation and connectives, as we introduce labels closer to the `BIO` scheme (e.g. `BeginSeg=Yes` becomes `Seg=B-seg`). This modification is fully reversible, but we decided to modify the code of the system, and we release the modified version.[13] For eng.rst.gentle, the model was trained on eng.rst.gum.

DisCoDisCo is also a Transformer-based system which consists of a feature-rich, encoder-less sentence pair classifier for the relation classification task, enhanced with hand-crafted features. Specifically, a language-specific pretrained BERT model, and a linear projection and softmax layer is used on the output of the pooling layer to predict the label of the relation. Because DisCoDisCo did not participate in the 2023 Shared Task, we have to adapt the system to the new datasets introduced in the latest release. Specifically, for the two new languages, `XLM-RoBERTa-base` was used. As for hand-crafted features, newly introduced datasets that do not have an existing dataset in the same language and framework (i.e. eng.dep.covdtb, eng.dep.scidtb, ita.pdtb.luna, por.pdtb.crpc, por.pdtb.tedm, tha.pdtb.tdtb, and zho.dep.scidtb) only used the baseline setup (i.e. no hand-crafted features were used). For the other new datasets, the hand-crafted features of the corresponding datasets from the same framework and language in DISRPT 2021 were adopted. Finally, note that OOD corpora may contain labels that do not exist at training time, which is the case for eng.dep.covdtb: we thus mapped the relations in eng.dep.scidtb based on Nishida and Matsumoto (2022) and retrain the model. For the other OOD datasets, no preprocessing is done.

5.2. Results

Table 2 provides scores (averaged over 3 runs for each dataset) on the three tasks on all 28 datasets shown in Table 1. For relations, the mean accuracy excluding eng.rst.gentle (as it was not available during the shared task) is 62.43, which is a little bit higher than HITS (62.36), the best-performing system in 2023 (Liu et al., 2023). It is also worth noting that DisCoDisCo's score on the English PDTB (eng.pdtb.pdtb) dataset is very close (75.14) to the one reported in 2021 (74.44), suggesting that using the rarest or the first annotated sense does not have a huge impact on the overall performance. Finally, the score on eng.rst.gum (64.12) is lower than the one reported in 2023 (68.19), which likely resulted from switching from predicting coarse relation classes to fine-grained labels.

Results on segmentation and connective iden-

[12] In this version of the corpus, 15 documents are missing compared to the original dataset due to pre-processing issues.

[13] https://github.com/phimit/jiant-discut

dataset	Rel (acc)	Seg (F1)	Conn (F1)
deu.rst.pcc	35.77	96.31	-
eng.dep.covdtb	76.68	92.01	-
eng.dep.scidtb	74.78	95.50	-
eng.pdtb.gum	-	-	91.30
eng.pdtb.pdtb	75.14	-	92.40
eng.pdtb.tedm	57.83	-	80.19
eng.rst.gentle	56.26	93.00	-
eng.rst.gum	64.12	95.53	-
eng.rst.rstdt	66.08	97.71	-
eng.sdrt.stac	63.51	96.60	-
eus.rst.ert	60.32	91.16	-
fas.rst.prstc	53.38	93.80	-
fra.sdrt.annodis	46.88	89.20	-
ita.pdtb.luna	46.24	-	68.47
nld.rst.nldt	51.69	97.15	-
por.pdtb.crpc	74.84	-	81.60
por.pdtb.tedm	58.24	-	77.56
por.rst.cstn	61.52	93.94	-
rus.rst.rrt	65.99	85.48	-
spa.rst.rststb	57.75	92.85	-
spa.rst.sctb	67.92	85.04	-
tha.pdtb.tdtb	86.63	-	90.75
tur.pdtb.tdb	60.74	-	91.90
tur.pdtb.tedm	44.96	-	65.27
zho.dep.scidtb	63.26	89.53	-
zho.pdtb.cdtb	86.72	-	87.88
zho.rst.gcdt	61.91	92.53	-
zho.rst.sctb	60.38	81.20	-
mean	62.21	92.14	82.73

Table 2: Results for Relation Classification (Rel) using DisCoDisCo and Discourse Segmentation (Seg) and Connective detection (Conn) using DisCut (Treebanked data) of DISRPT 2023 Datasets.

tification are also close to the ones presented in 2023 (Metheniti et al., 2023). We note that performance on Thai for connectives are largely improved ($+5\%$) thanks to the correction in the data (`I` labels without immediately preceding `B`).

Furthermore, for the newly introduced datasets, results are good for eng.pdtb.gum on connective detection (91.30), with results on par with the larger eng.pdtb.pdtb (92.40). The corpus eng.rst.gentle is shown to be indeed challenging for relations, but scores for segmentation are rather high for a small, OOD dataset.

6. Conclusion

In this paper, we presented a benchmark for discourse processing including 28 datasets covering 13 languages, 4 frameworks, and multiple domains. We have detailed the conversion process and the modifications introduced to produce a unified format. The aim of this benchmark is to encourage work on transfer learning for discourse processing across languages, domains, and frameworks. We have also highlighted some aspects that should be discussed further in the community about 1) the way embedded discourse units are encoded; 2) the differences in annotation of discourse connectives (i.e. with or without modifiers); 3) the huge divergences on label sets, which seem sometimes artificial (e.g. *alternative* vs *alternation*); 4) the problem of choosing which label to predict when multiple labels are annotated; and 5) the issue of encoding the direction of relations. The performance of SOTA systems on the benchmark demonstrates that there is still large room for improvement on relation classification, a typically hard task as well as the connective identification task for specific text types (e.g. dialogues in ita.pdtb.luna) or OOD data (e.g. the TED datasets). On the other hand, performance on segmentation has reached a plateau, which requires more in-depth analyses to better understand what kind of errors the systems are still making.

Acknowledgements

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.

This work is also partially supported by the SLANT project (ANR-19-CE23-0022) and the ANR grant SUMM-RE (ANR-20-CE23-0017). 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.

7. Limitations

This paper and the associated benchmark data have several limitations in both discourse representation and possible bias in the data, which have not been explored in this paper. The first major limitation is that although we ensure that data is available in a homogeneous format and made some efforts to harmonize datasets using common conversion tools and conceptual frameworks, there remain fundamental differences between underlying discourse frameworks (e.g. the concept of discourse relations in RST vs. PDTB), individual corpora and their guidelines (even for the same language) and the specific meanings of discourse relation labels, which may recur across datasets with subtly different meaning.

Additionally, we have not explored how bias may feature in many of the datasets presented here, which are products of specific times, data sources and sampling strategies, which may be skewed in a variety of ways towards specific author/speaker demographics, topics, and more. The existence of multiple datasets for several languages in the collection offers a first step towards facilitating an evaluation of cross-corpus degradation which may result from biased data, but much work remains to be done. These are all issues we would like to address in future work.

8. Bibliographical 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. 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).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

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

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

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.

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. CSTNews - 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 and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, University of Southern California Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Debopam Das and Maite Taboada. 2017. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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(1):87–135.

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.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Mikel Iruskieta, Iria Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49(2):263–309.

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

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Rene Knaebel and Manfred Stede. 2023. Discourse sense flows: Modelling the rhetorical style of documents across various domains. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14462–14482, Singapore. Association for Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2022. A simple and strong baseline for end-to-end neural RST-style discourse parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6725–6737, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloe Braud, and Giuseppe Carenini. 2023. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2562–2579, Dubrovnik, Croatia. Association for Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. Technical report, National University of Singapore.

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.

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.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

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.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Amália Mendes and Pierre Lejeune. 2022. Crpc-db a discourse bank for portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 79–89, Berlin, Heidelberg. Springer-Verlag.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India. The COLING 2012 Organizing Committee.

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.

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.

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

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a Dutch text corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'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.

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

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of AAAI*, volume 33, pages 7007–7014.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT-NAACL 2003*, pages 149–156, Edmonton.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 43–49, Geneva, Switzerland. COLING.

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).

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).

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3943–3949. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. 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.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of CoNLL*.

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.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–27.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources*.

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

9. Language Resource References

Afantenos, Stergos and Asher, Nicholas and Benamara, Farah and Bras, Myriam and Fabre, Cécile and Ho-dac, Mai and Draoulec, Anne Le and Muller, Philippe and Péry-Woodley, Marie-Paule and Prévot, Laurent and Rebeyrolles, Josette and Tanguy, Ludovic and Vergez-Couret, Marianne and Vieu, Laure. 2012. *The ANNODIS corpus*. self. PID http://redac.univ-tlse2.fr/corpus/annodis/.

Aoyama, Tatsuya and Behzad, Shabnam and Gessler, Luke and Levine, Lauren and Lin, Jessica and Liu, Yang Janet and Peng, Siyao and Zhu, Yilun and Zeldes, Amir. 2023. *GENTLE: A Genre-Diverse Multilayer Challenge Set for English NLP and Linguistic Evaluation*. self. PID https://gucorpling.org/gum/gentle.html.

Asher, Nicholas and Hunter, Julie and Morey, Mathieu and Farah, Benamara and Afantenos, Stergos. 2016. *STAC: Strategic Conversation Corpus*. self. PID https://www.irit.fr/STAC/corpus.html.

Cao, Shuyuan and da Cunha, Iria and Iruskieta, Mikel. 2018. *The RST Spanish-Chinese Treebank*. self. PID http://ixa2.si.ehu.eus/rst/zh/index.php.

Paula Christina Figueira Cardoso and Erick Galani Maziero and Maria Lucía del Rosario Castro Jorge and M. Eloize and R. Kibar Aji Seno and Ariani Di Felippo and Lucia Helena Machado Rino and Maria das Graças Volpe Nunes and Thiago Alexandre Salgueiro Pardo. 2011. *The CSTnews Corpus*. self. PID http://nilc.icmc.usp.br/CSTNews/login/?next=/-CSTNews/.

Lynn Carlson and Daniel Marcu and Mary Ellen Okurowski. 2001. *RST Discourse Treebank*. LDC, ISLRN 299-735-991-930-2.

da Cunha, Iria and Torres-Moreno, Juan-Manuel and Sierra, Gerardo. 2011. *The RST Spanish Treebank*. self. PID http://www.corpus.unam.mx/rst/index_es.html.

Mikel Iruskieta and María Jesús Aranzabe and Arantza Diaz de Ilarraza and Itziar Gonzalez-Dios and Mikel Lersundi and Oier Lopez de Lacalle. 2012. *The RST Basque TreeBank*. self. PID http://ixa2.si.ehu.eus/diskurtsoa/en/.

Mendes, Amália and Lejeune, Pierre. 2022. *CRPC-DB a Discourse Bank for Portuguese*. ELRA. PID https://www.clul.ulisboa.pt/en/recurso/portuguese-discourse-bank2.

Nishida, Noriki and Matsumoto, Yuji. 2022. *COVID-19 Discourse Dependency Treebank*. self. PID https://github.com/norikinishida/biomedical-discourse-treebanks.

Peng, Siyao and Liu, Yang Janet and Zeldes, Amir. 2022. *GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing*. self. PID https://github.com/logan-siyao-peng/GCDT.

Redeker, Gisela and Berzlánovich, Ildikó and van der Vliet, Nynke and Bouma, Gosse and Egg, Markus. 2012. *Multi-Layer Discourse Annotation of a Dutch Text Corpus*. ELRA. PID https://research.rug.nl/en/publications/multi-layer-discourse-annotation-of-a-dutch-text-corpus.

Sara Shahmohammadi and Hadi Veisi and Ali Darzi. 2021. *The Persian RST Corpus*. self. PID https://github.com/hadiveisi/PersianRST.

Manfred Stede and Arne Neumann. 2014. *Potsdam Commentary Corpus 2.0: Annotation for Discourse*. ELRA. PID http://angcl.ling.uni-potsdam.de/resources/pcc.html.

Toldova, Svetlana and Pisarevskaya, Dina and Ananyeva, Margarita and Kobozeva, Maria and Nasedkin, Alexander and Nikiforova, Sofia and Pavlova, Irina and Shelepov, Alexey. 2017. *Ru-RSTreebank*. self. PID https://rstreebank.ru/.

Tonelli, Sara and Riccardi, Giuseppe and Prasad, Rashmi and Joshi, Aravind and Stepanov, Evgeny A. and Chowdhury, Shammur Absar. 2010. *LUNA Corpus Discourse Data Set*. ELRA. PID http://universal.elra.info/product_info.php?cPath=37_38&products_id=1832.

Webber, Bonnie and Prasad, Rashmi and Lee, Alan and Joshi, Aravind. 2022. *The Penn Discourse Treebank 3.0*. LDC, ISLRN 977-491-842-427-0.

Yang, An and Li, Sujian. 2018. *SciDTB: Discourse Dependency TreeBank for Scientific Abstracts*. self. PID https://github.com/PKU-TANGENT/SciDTB.

Yi, Cheng and Sujian, Li and Yueyuan, Li. 2021. *Unifying Discourse Resources with Dependency Framework*. self. PID https://github.com/PKU-TANGENT/UnifiedDep.

Amir Zeldes and Lauren Levine. 2017. *GUM: The Georgetown University Multilayer Corpus*. self, ISLRN 421-566-418-865-2.

Zeyrek, Deniz and Mendes, Amália and Grishina, Yulia and Kurfalı, Murathan and Gibbon, Samuel and Ogrodniczuk, Maciej. 2022. *TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style*. LDC. PID https://github.com/MurathanKurfali/Ted-MDB-Annotations.

Zeyrek, Deniz and Webber, Bonnie and Kurfalı, Murathan. 2008. *TDB 1.1: Turkish Discourse Bank*. self. PID http://medid.ii.metu.edu.tr/theCorpus.html.

Yuping Zhou and Jill Lu and Jennifer Zhang and Nianwen Xue. 2014. *Chinese Discourse Treebank 0.5*. LDC. PID https://catalog.ldc.upenn.edu/LDC2014T21.

A. Labels for Segmentation and Connective Tasks

Segmentation and connective identification are encoded using a `BIO` scheme. Note that, compared to the original shared task format, the label sets for these tasks are modified in order to propose label names closer to a `BIO` scheme (with a pair of key=value conforming to the CoNLL-U format), as described in Table 3.

Moreover, when multi-word expressions are annotated, the label is associated to the first token of the expanded multi-word expression (e.g. *I'm* is expanded as *I am* and the label is on the pronoun *I*), the original contracted form holds a meaningless label '_' that is ignored during evaluation.

Shared Task Label	New Label
Segmentation	
`BeginSeg=Yes`	`Seg=B-seg`
_	`Seg=O`
Connective Identification	
`Seg=B-Conn`	`Conn=B-conn`
`Seg=I-Conn`	`Conn=I-conn`
_	`Conn=O`

Table 3: Labels used for the segmentation and connective identifications tasks.

B. Language Abbreviations

Table 4 present the language abbreviations for all languages represented in the DISRPT benchmark.

Language Code	Language Name
deu	German
eng	English
eus	Basque
fas	Farsi
fra	French
ita	Italian
nld	Dutch
por	Portuguese
rus	Russian
spa	Spanish
tha	Thai
tur	Turkish
zho	Chinese

Table 4: Language Abbreviations.

C. Relation Mapping Details

Table 5 provides the mapping done for the relation labels in addition to translation to English when needed: we here report the information given by the shared task organizers (Braud et al., 2023) and add some missing information (e.g. for deu.rst.pcc) and the modifications proposed in this paper. A few cases of labels were also removed when they did not correspond to a discourse relation. Note that, additionally, labels are translated to English for some corpora such as eus.rst.ert.

In addition, as described in Section 5.1, the predicted relations are modified when they do not correspond to the labels existing in the target dataset for OOD settings.

[14] The -nn / mult part of the label stand for multi-nuclei relations and is ignored.

[15] The mapping for very rare relations was proposed by Manfred Stede, author of the paper presenting this corpus.

Corpus	Original label	Mapped label
eus.rst.ert	anthitesis	antithesis
	motibation	motivation
	solution-hood	solutionhood
	birformulazioa-nn[14]	restatement
spa.rst.rststb	backgroun	background
fas.rst.prstc	topicomment	topic-comment
	topichange	topic-change
	topidrift	topic-drift
	causemult[14]	cause
	contrastmult	contrast
	jointmult	joint
por.rst.cstn	non-volitional-cause	nonvolitional-cause
	non-volitional-cause-e	nonvolitional-cause-e
	non-volitional-result	nonvolitional-result
	non-volitional-result-e	nonvolitional-result-e
deu.rst.pcc[15]	e-elab	e-elaboration
	enablement	background
	justify	reason
	motivation	reason
	otherwise	antithesis
	unless	antithesis
fra.sdrt.annodis	e-elab	e-elaboration
nld.rst.nldt	span	relation removed
eng.dep.scidtb	null	relation removed
ita.pdtb.luna	null	relation removed

Table 5: Relation Mapping used in the DISRPT 2023 Shared Task and additional proposed changes. The other modifications (translation, mapping to RST DT classes) are described in the literature (Carlson and Marcu, 2001; Braud et al., 2017) and in the GitHub repository.

eng.dep.scidtb	eng.dep.covdtb
evaluation	findings
elab-.*	elaboration
bg-.*	background
cause	cause-result
result	cause-result
contrast	comparison

Table 6: Relation mapping performed on the train set of eng.dep.scidtb to eng.dep.covdtb, following Nishida and Matsumoto (2022).

D. Preprocessing

We indicate in Table 7 the preprocessing information for each dataset, corresponding to tokenization, sentence splitting, POS tagging, syntactic analysis and multi-word expression expansion. These information can be either gold, or automatically predicted. In the latter case, the information is either distributed with the corpus ('given') – in which case we indicate, when possible, the tool used to create these annotations –, or performed by the shared task organizers, in general using Stanza. Note that the tokenization step is crucial, since labels for segmentation and discourse connective identification are linked to tokens. It is thus difficult to change the tokenization.

Corpus	Token	Sentence	POS/Synt	MWE
deu.rst.pcc	gold	gold	tnt tagger/ stanza (gsd)	none
eng.dep.covdtb	stanza	stanza	stanza	depedit
eng.dep.scidtb	stanza	stanza	stanza	depedit
eng.pdtb.gum	gold	gold	gold	depedit
eng.pdtb.pdtb	gold	gold	gold[16]	depedit
eng.pdtb.tedm	stanza (gum)	stanza	stanza	stanza
eng.rst.gentle	gold	gold	gold	gold
eng.rst.gum	gold	gold	gold	depedit
eng.rst.rstdt	gold	gold	gold	depedit
eng.sdrt.stac	stanza (ewt)	stanza	stanza	depedit
eus.rst.ert	stanza	stanza	stanza	none
fas.rst.prstc	stanza	stanza	stanza	stanza
fra.sdrt.annodis	spacy	spacy	spacy	none
ita.pdtb.luna	given	given (silence[17].)	stanza	stanza
nld.rst.nldt	stanza	stanza	stanza	none
por.pdtb.crpc	given (LX-center)	given (SentenceChunker)	stanza	given[18]
por.pdtb.tedm	given (LX-center)	given (SentenceChunker)	stanza	given[18]
por.rst.cstn	stanza (bosque)	stanza	stanza	stanza
rus.rst.rrt	stanza (syntagrus)	stanza	stanza	none
spa.rst.rststb	stanza (ancora)	stanza	stanza	none
spa.rst.sctb	stanza (ancora)	stanza	stanza	none
tha.pdtb.tdtb	gold	gold	gold	gold
tur.pdtb.tdb	UDPipe	UDPipe	UDPipe	UDPipe
tur.pdtb.tedm	stanza	stanza	stanza	stanza
zho.dep.scidtb	stanza	stanza	stanza	none
zho.pdtb.cdtb	gold	gold	gold	none
zho.rst.gcdt	stanza	stanza	stanza	none
zho.rst.sctb	stanza (gsdsimp)	stanza	stanza	none

Table 7: Preprocessing information of corpora included in the benchmark: 'given' means that the preprocessing was predicted but distributed with the original corpus, we indicate the tool used when known. The information can also be 'gold', if it comes from a manual annotation. In the other cases, an automatic tool was used, for Stanza we use the default model for the target language, or the one indicated in the first column.

E. Additional Statistics

The expressions annotated as connectives can vary in nature, depending on whether the annotation includes modifiers or not. Table 8 indicates the size of the connective lexicon for each dataset.

dataset	# connectives
ita.pdtb.luna	61
eng.pdtb.pdtb	1231
zho.pdtb.cdtb	274
tur.pdtb.tdb	324
eng.pdtb.tedm	71
por.pdtb.crpc	644
por.pdtb.tedm	66
tur.pdtb.tedm	173
tha.pdtb.tdtb	132
eng.pdtb.gum	143

Table 8: Size of the connective lexicons for PDTB-style datasets.

[16] The syntax trees are, more precisely, a CoreNLP coversion from PTB trees, that could include errors.

[17] LUNA is composed of speech transcriptions where the notion of sentence is not well-defined, the segmentation is based on silence, see Tonelli et al. (2010)

[18] The corpus was given with multi-word expressions already expanded, without the indication of the original contracted forms.