The DISRPT 2021 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification *

Amir Zeldes Georgetown University az364@georgetown.edu

Philippe Muller IRIT Toulouse philippe.muller@irit.fr Yang Janet Liu Georgetown University yl879@georgetown.edu

Mikel Iruskieta University of the Basque Country mikel.iruskieta@ehu.eus

Chloé Braud IRIT-CNRS-ANITI Toulouse chloe.braud@irit.fr Sonia Badene Linagora-IRIT sbadene@linagora.com

Abstract

In 2021, we organized the second iteration of a shared task dedicated to the underlying units used in discourse parsing across formalisms: the DISRPT Shared Task (Discourse Relation Parsing and Treebanking). Adding to the 2019 tasks on Elementary Discourse Unit Segmentation and Connective Detection, this iteration of the Shared Task included for the first time a track on discourse relation classification across three formalisms: RST, SDRT, and PDTB. In this paper we review the data included in the Shared Task, which covers nearly 3 million manually annotated tokens from 16 datasets in 11 languages, survey and compare submitted systems and report on system performance on each task for both annotated and plain-tokenized versions of the data.

1 Introduction

Building on rapid progress in NLP for discourse parsing in the past decade (e.g. Iruskieta et al. 2013, Zhou et al. 2014, Afantenos et al. 2012, Braud et al. 2017, Wang and Lan 2015, Li et al. 2016, Perret et al. 2016), the past two years since the DISRPT 2019 Shared Task (Zeldes et al., 2019) have seen unprecedented performance on benchmark datasets for discourse parsing (e.g. Guz and Carenini 2020; Liu et al. 2020; Kurfali 2020; Zhang et al. 2021b). Following the 2019 Shared Task, our focus in 2021 remained on the most established frameworks for discourse relation treebanking: Rhetorical Structure Theory (Mann and Thompson, 1988), the Penn Discourse Treebank's framework (Prasad et al., 2014), and Segmented Discourse Representation Theory (Asher, 1993).

With the progress achieved on discourse segmentation and connective detection since 2019, this year we decided to extend the competition to a new task: discourse relation classification across frameworks. Although work on relation classification is not new, work has generally been separated into several somewhat independent strands of work: relation classification with explicit connectives (see Kido and Aizawa 2016; Nie et al. 2019), implicit relation classification (Wang and Lan, 2016; Kim et al., 2020), and full discourse constituent (Guz and Carenini, 2020; Zhang et al., 2021b) or dependency parsing (Morey et al., 2018). We believe that these strands of research can come together and benefit from data across frameworks, but only if we are able to formulate a common denominator which poses the problem of discourse relation recognition in similar terms, independently of the underlying theoretical assumptions.

In addition to the new relation classification task, we also updated and expanded the two tasks from 2019 and added a new surprise dataset (and language) which became available in the interim: the Persian RST Corpus (Shahmohammadi et al., 2021).¹ While it remains impossible to place some data openly online due to licensing constraints, all annotations, as well as text for open access datasets, is available from the Shared Task repository,² as well as scripts for reconstructing the underlying text for proprietary datasets.

In the next section, we will describe the tasks for this year, starting with the new task of Relation Classification across Frameworks, followed by the existing tasks, followed by Section 3, which provides an overview of the shared task data. Section 4 briefly discusses our effort addressing the ethical and transparency issues related to NLP shared task, inspired by Escartín et al. (2021). Section 5 then describes the Shared Task system submissions and results, followed by a discussion and conclusion.

²https://github.com/disrpt/ sharedtask2021

^{*}Discourse Relation Parsing and Treebanking (DISRPT 2021) was held in conjunction with CODI at EMNLP 2021 in the Dominican Republic and Online (https://sites.google.com/georgetown.edu/disrpt2021).

¹https://github.com/hadiveisi/
PersianRST

2 Tasks

2.1 Relation Classification across Frameworks

A major challenge to the goal of enabling relation classification across frameworks is formed by the structural differences between theories: while RST conceptualizes discourse in the form of single rooted, usually projective and hierarchical constituent trees, SDRT assumes an also hierarchical but less constrained graph structure (including multiple relations between units), while PDTB-style corpora assume no overarching macro-structure, with relations applying between independent, possibly overlapping spans of text, which need not cover an entire document. Despite these differences, all frameworks postulate *directed*, *labeled discourse relations*, and often come to similar and highly correlated insights (cf. Demberg et al. 2019).

In order to represent data across frameworks on an equal footing, we opted to initially convert discourse relation information from RST, PDTB, and SDRT datasets into a set of discourse unit pairs for which a discourse relation (either explicit or implicit) is known to apply. Because only PDTBstyle data marks connectives, we chose to integrate these into the argument span within which they were syntactically integrated and restrict the spans involved in the relation to exactly two: unit1 and unit2 in text order. However, since PDTB-style data does order argument spans based on relations regardless of text order (to represent cause and effect, etc.), and since SDRT and RST relations are also directed, we retained the direction of the relation in a dedicated field, indicating whether it was a left-to-right relation (1>2) or right-to-left (1<2).

Additionally, since we wanted to give systems a chance to use as much context as desired, we explicitly included the sentences containing each unit in shared task data, as well as giving stand-off running token IDs for both units and their sentences, which are linked to the numbered tokens from the existing segmentation and connective detection tasks. The resulting format, found in the task's .rels files, is shown in Table 1.

The table shows an entry from the English GUM corpus (Zeldes, 2017), part of an interview with a biologist studying ants, in which use of a lot of (ant) colonies is brought as evidence (pointing right-to-left, hence dir is 1<2) that the results of a study are very reliable. The unit_txt fields

give the space separated tokens in the head unit of each block of discourse units involved in the relation. The first unit 'The results are very reliable' is annotated as discontinuous in the source corpus, as indicated by the reserved token '<*>', which stands in for the missing words 'we got', which can be seen in the full sentence containing this unit, under unit1_sent. The second unit is also part of a longer, complex block of units, including a coordinated second piece of evidence (the experiment was blind), and a further nested discourse unit ('wherever possible'), both of which have similar entries for their respective discourse relations elsewhere in the .rels data.

As this example shows, context can help to identify relations: in fact, without the containing sentences, we would not know that these are ant colonies, which makes the setting 'experimental study' easier to recognize. Since we did not want to restrict systems to just head units, and conceivably wanted to allow systems to use the entire document if desired, the token offset fields (unit1_toks, s2_toks, etc.) were given and indicated the position of the relation in the larger text, represented in sequence in the .tok and .conllu files from the EDU Segmentation and Connective Detection tasks.

For datasets with large numbers of relations and/or very rare relations, we used established collapsing procedures to reduce the number of target relations and ensure that dev and test sets only included relations attested in the training data. In such cases, the original labels are available in the orig_label column, and target labels for the shared task are available in the final label column.

2.2 EDU Segmentation and Connective Detection

As in the 2019 Shared Task, EDUs and connectives were represented in each corpus in two formats, corresponding to two scenarios: Treebanked data (now named *.conllu, previously named .conll in 2019), which included an (ideally gold) dependency parse, including gold sentence splits and POS tags, and unannotated, plain tokens (*.tok). As in 2019, for datasets that had Universal POS tags and/or UD dependencies, including these was preferred, though we followed the CoNLL-U format's convention of allowing two POS tag fields (UPOS for universal tags, XPOS for language spe-

field	value
doc	GUM_interview_ants
unit1_toks	954-955,958-961
unit2_toks	962-967
unit1_txt	The results <*> are very reliable.
unit2_txt	We had a lot of colonies
s1_toks	954-961
s2_toks	962-982
unit1_sent	The results we got are very reliable.
unit2_sent	We had a lot of colonies containing a lot of ants, and wherever possible we conducted the experiment blind.
dir	1<2
orig_label	evidence
label	evidence

Table 1: An Example of Columns in the Tabular *.rels Data Format for the Relation Classification Task.

	id = = Yes	_		_stardus	st-28				
1	Yes	yes	INTJ	UH	_	0	root	_	BeginSeg=Yes
2	[[PUNCT	-LSB-	-	3	punct	_	BeginSeg= <mark>Yes</mark> SpaceAfter=No
3	see	see	VERB	VBP	Mood=Ind Tense=Pres VerbForm=Fin	1	parataxis	_	_
4	below	below	ADV	RB	_	3	advmod	_	SpaceAfter=No
5]]	PUNCT	-RSB-		3	punct	_	SpaceAfter=No
6	·	•	PUNCT		-	1	punct	-	-
411	Yes								BeginSeg=Yes
412	ſ	-	-	-	—	-	_	-	BeginSeg=Yes
413	see	_	-	_	—	_	_	_	
414	below	_	-	_	—	-	-	-	-
415	1	-	-	-	-	-	-	-	-
416	1	_	-	_	-	-	-	-	—
410	•	_	-	-	-	-	-	-	—

Figure 1: Segmentation Format: treebanked (*.conllu, top) and plain (*.tok, bottom)

cific tags), a morphology field with unlimited morphological annotations, and a secondary dependency field for enhanced dependencies where available. The tenth column (MISC in CoNLL-U) was used for gold standard labels and additional annotations (e.g. SpaceAfter to indicate whitespace in underlying data), which all followed the CoNLL-U key=value format: BeginSeg=Yes for EDU segmentation and BI tags for connectives, Seg=B-Conn and Seg=I-Conn, versus "_" for unannotated tokens. The second scenario included no annotations except for tokenization and the same document boundary annotations found in the treebanked files. No sentence splits were provided in this scenario. Figure 1 illustrates both formats.

. . .

Evaluation scripts were provided for both segmentation and connective detection, with positive class, span-based F-score being the target. As in 2019, for connective detection the evaluation targets only exact matches, meaning that precision and recall are calculated out of the total connective spans (not tokens) available in the gold data, without partial credit or fuzzy matching. As a result, systems identifying the span in (1) below are given one precision error and one recall error, since the prediction misses the gold span and invents one not present in gold data. (1) Gold: In/B-Conn order/I-Conn to/_ Pred: In/B-Conn order/I-Conn to/I-Conn

3 Shared Task Data

The DISRPT 2021 Shared Task data comprises 16 datasets in 11 languages, 13 of which target elementary discourse unit segmentation, and 3 dedicated to explicit connective annotation, and all 16 being included in the relation classification task. Table 2 gives an overview of the datasets. Of the 16 datasets, training data for 15 were released in their updated 2021 version approximately 3 months before the shared task deadline, while the final one, RST annotations from the Persian RST Corpus, was released as a 'surprise' dataset/language together with all test sets just one month before the announced deadline. For five of the datasets, licensing constraints prevented online publication of the entire underlying texts (e.g. Wall Street Journal material), meaning that the public repository contains only annotations for those corpora, with tokens replaced by underscores. Of these datasets, four contain no text at all (English PDTB and RST-DT, Chinese CDTB, and Turkish TDB), and one (English GUM) contains data for 11/12 genres, but requires download of the text for the 12th genre, Reddit forum discussions.

corpus	language	framework	sentences	tokens	documents	units	segStyle	relations	relTypes	discontinuous
deu.rst.pcc	German	RST	2,193	33,222	176	3,040	EDU	2,164	26	No
eng.rst.gum	English	RST	8,292	152,856	168	19,268	EDU	13,897	23	Yes
eng.rst.rstdt	English	RST	8,318	205,829	385	21,789	EDU	16,002	17	Yes
eng.sdrt.stac	English	SDRT	11,087	52,354	45	12,588	EDU	9,580	16	No
eus.rst.ert	Basque	RST	2,380	45,780	164	4,202	EDU	2,533	29	Yes
fas.rst.prstc	Persian	RST	2,179	66,926	150	5,855	EDU	4,100	17	Yes
fra.sdrt.annodis	French	SDRT	1,507	32,699	86	3,429	EDU	2,185	18	Yes
nld.rst.nldt	Dutch	RST	1,651	24,898	80	2,343	EDU	1,608	32	No
por.rst.cstn	Portuguese	RST	2,221	63,332	140	5,537	EDU	4,148	32	Yes
rus.rst.rrt	Russian	RST	23,044	473,005	332	41,542	EDU	28,868	22	Yes
spa.rst.rststb	Spanish	RST	2,089	58,717	267	3,351	EDU	2,240	28	Yes
spa.rst.sctb	Spanish	RST	516	16,515	50	744	EDU	439	24	Yes
zho.rst.sctb	Mandarin	RST	580	15,496	50	744	EDU	439	26	Yes
corpus	language	framework	sentences	tokens	documents	units	segStyle	relations	relTypes	discontinuous
eng.pdtb.pdtb	English	PDTB	48,630	1,156,658	2,162	26,048	Conn	43,920	23	Yes
tur.pdtb.tdb	Turkish	PDTB	31,197	496,358	197	8,748	Conn	2,451	23	Yes
zho.pdtb.cdtb	Mandarin	PDTB	2,891	73,314	164	1,660	Conn	3,657	9	Yes

Table 2: Datasets in the DISRPT 2021 Shared Task.

A script included in the shared task repository was provided in order to reconstruct the data, which requires users to have access to the original LDC releases of the underlying corpora in the case of PDTB, RST-DT, and CDTB or the original Turkish texts available from Middle East Technical University (METU) by request. Missing data for GUM is downloadable directly by running the script and consenting to non-commercial use conditions.

The short names for every dataset begin with an ISO 639-3 three letter code for the language, a framework designation (RST/SDRT/PDTB) and an acronym for the corpus. The names correspond to the following included corpora:

- deu.rst.pcc Potsdam Commentary Corpus (Stede and Neumann, 2014).
- eng.pdtb.pdtb Penn Discourse Treebank (Prasad et al., 2014).
- eng.rst.gum Georgetown University Multilayer corpus (Zeldes, 2017).
- eng.rst.rstdt RST Discourse Treebank (Carlson et al., 2001).
- eng.sdrt.stac Strategic Conversations corpus (Asher et al., 2016).
- eus.rst.ert Basque RST Treebank (Iruskieta et al., 2013).
- fas.rst.prstc Persian RST Corpus (Shahmohammadi et al., 2021).
- fra.sdrt.annodis ANNOtation DIScursive (Afantenos et al., 2012).
- nld.rst.nldt Dutch Discourse Treebank (Redeker et al., 2012).
- por.rst.cstn Cross-document Structure Theory News Corpus (Cardoso et al., 2011).
- rus.rst.rrt Russian RST Treebank (Toldova et al., 2017).

- spa.rst.rststb RST Spanish Treebank (da Cunha et al., 2011).
- spa.rst.sctb RST Spanish-Chinese Treebank (Spanish) (Cao et al., 2018).
- tur.pdtb.tdb
 Turkish Discourse Bank (Zeyrek and Webber, 2008; Zeyrek and Kurfalı, 2017).
- zho.pdtb.cdtb Chinese Discourse Treebank (Zhou et al., 2014).
- zho.rst.sctb RST Spanish-Chinese Treebank (Chinese) (Cao et al., 2018).

As Table 2 shows, these datasets range from small (under 15,000 tokens for the smallest corpus, zho.rst.sctb), to the larger RST corpora (over 200,000 tokens for RST-DT and the Russian RST Treebank), to the largest PDTB-style datasets (almost half a million tokens for Turkish, and over a million for the English PDTB). The variability in sizes, languages, frameworks, and corpus-specific annotation guidelines were expected to challenge systems, but also promote architectures which can be extended to more languages in the future, and ideally stay robust for low resource settings.

4 Ethical Considerations and Transparency in DISRPT 2021

In the interest of transparency and with inspiration from Escartín et al. (2021), we made efforts to address several ethical and transparency matters in the DISPRT 2021 Shared Task. We created a FAQ page³ on our main website that laid out information about the Shared Task, participation, equity, and the evaluation criteria etc., which are briefly summarized below.

³https://sites.google.com/georgetown. edu/disrpt2021/faq

Participation & Organization Generally speaking, we allowed anyone to participate, including the organizers. However, in the interest of transparency and fairness, results from teams which overlap some of the organizers or annotators of any of the included datasets are denoted as such. Though as Escartín et al. (2021) pointed out, there is an understanding that organisers have privileged access to test sets, we pledge that the test data is under strict quarantine until its release and strictly follow this practice. All systems were always rerun and reproduced by evaluators who do not overlap the submission team, and all datasets' original publications had already been made independently of the Shared Task. There are always scenarios where there is overlap between the Shared Task organizers, dataset creators, and participants, and excluding organizer teams may discourage people from organizing future shared tasks, while unfairly penalizing students of organizers who are likely to form the next generation of researchers in the field.

Evaluation & Reproducibility We published the evaluation scripts in the Shared Task GitHub repository⁴ at the beginning of the Shared Task for participants to use. We also announced at the beginning of the Shared Task how the overall system rankings in each category will be determined, which is described on the main website. For teams that submitted multiple systems, the best scoring system by macro-averaged F-score on all datasets was selected to represent the team. Moreover, because we believe that code openness and reproducibility are very crucial components in NLP practice, we state up front that all systems must include code to retrain the system from scratch, so that evaluators can test aspects of systems' performance and reproduce reported scores, as well as a detailed README file explaining how to train the system. Systems which cannot be run in the evaluation phase are not be accepted. All submitted systems followed these guidelines, and we were able to retrain and reproduce reported scores falling within the range of multiple random training runs.

Data Access & Resources We are aware that not all teams may have LDC subscriptions. Thus, in the interest of promoting equity regardless of access and funding status, we evaluated submitted systems on the closed LDC datasets for the participants who

⁴https://github.com/disrpt/ sharedtask2021/tree/main/utils do not have access to LDC data. We reported scores obtained during the evaluation phase to authors so that they can add them to their papers for the camera-ready version. In addition, resources about each dataset (e.g. papers, annotation manuals etc.) are provided in the README file in each data directory for participants who would like to know more about the corpora used in the shared task.

Negative Results We believe that negative results can bring the field forward, especially when they are accompanied by insightful analysis about why a certain approach does not work. Thus, we encourage such submissions, and their acceptance is subject to the contribution that their analyses can provide to the field. In response to one of the Escartín et al. (2021)'s shared task checklist questions, we allow participants to opt to exclude or anonymize their results in the overall ranking to avoid withdrawal or fear of negative results, though no submission opted for this possibility.

Identification of Annotation Errors We are aware that current machine learning models are capable of catching annotation errors in the original datasets. Thus, we encouraged participants to report such errors to us if they identify any in their error analysis study so that we could pass them along to the original authors of the corpora. In addition, during the data preprocessing stage, we identified and recorded several annotation errors in several corpora. We reported some of these to the original authors earlier this year as they needed to be corrected in order for such annotation instances to be included in the Shared Task. We intend to compile the rest of the annotation errors after the workshop and report them to corpus creators accordingly.

5 Results

For EDU segmentation and connective detection, we report precision, recall, and F1 scores for systems in the two tasks, each consisting of two scenarios: EDU segmentation and connective detection, with treebanked and plain tokenized data; for relation classification the accuracy score is used. The overall system rankings in each category are determined by the macro-average score across treebanks, where each treebank score is decided by the micro-averaged metric.

Five systems were submitted to the shared task, four of which attempted the EDU segmentation

corpus	corpus TMVM		D	DiscoDisco*			disCut*			SegFormers			
(treebanked)	Р	R	F1	Р	R	F1	Р	R	F1	Р	R	F1	
deu.rst.pcc	94.57	88.78	91.58	97.07	94.15	95.58	98.91	92.52	95.61	93.71	96.26	94.97	94.44
eng.rst.gum	94.16	79.86	86.42	93.9	94.43	94.15	93.27	93.65	93.46	91.47	95.67	93.52	91.89
eng.rst.rstdt	87.18	65.52	74.81	96.39	96.89	96.64	96.16	95.99	96.08	96.15	98.04	97.09	91.16
eng.sdrt.stac	98.73	85.52	91.65	96.25	93.63	94.91	97.41	92.37	94.82	96.73	95.67	96.20	94.40
eus.rst.ert	87.23	66.49	75.46	93.42	87.73	90.46	90.04	83.11	86.44	87.23	88.65	87.94	85.08
fas.rst.prstc	88.28	74.18	80.62	92.79	93.10	92.94	93.54	90.75	92.12	90.53	94.18	92.32	89.50
fra.sdrt.annodis	82.13	49.84	62.03	89.43	90.65	90.02	87.26	88.67	87.96	86.76	88.03	87.39	81.85
nld.rst.nldt	95.38	85.50	90.17	97.50	94.50	95.97	94.15	95.27	94.71	98.16	94.67	96.39	94.31
por.rst.cstn	96.76	68.30	80.08	93.18	95.56	94.35	90.31	94.44	92.33	92.60	94.12	93.35	90.03
rus.rst.rrt	86.00	64.08	73.44	85.57	86.89	86.21	88.19	81.10	84.50	84.93	85.18	85.06	82.30
spa.rst.rststb	88.73	82.17	85.33	92.53	91.96	92.22	92.04	93.04	92.54	89.12	94.35	91.66	90.44
spa.rst.sctb	79.34	57.14	66.44	83.44	81.55	82.48	85.39	90.48	87.86	92.31	78.57	84.89	80.42
zho.rst.sctb	79.43	66.67	72.49	90.30	77.38	83.34	92.48	73.21	81.73	79.89	82.74	81.29	79.71
mean	89.07	71.85	79.27	92.44	90.65	91.48	92.24	89.58	90.78	90.74	91.24	90.93	88.12

Table 3: EDU Segmentation Results on Treebanked Data. Disclosure: Systems marked with * were submitted by teams containing organizers and annotators of shared task datasets (see Section 4).

corpus TMVM		DiscoDisco*			disCut*			SegFormers			mean		
(plain)	Р	R	F1	Р	R	F1	Р	R	F1	Р	R	F1	
deu.rst.pcc	91.34	86.05	88.62	95.15	92.86	93.94	94.67	96.60	95.63	96.81	92.86	94.79	93.25
eng.rst.gum	91.86	64.04	75.46	92.65	92.59	92.61	92.76	89.54	91.13	90.38	94.01	92.16	87.84
eng.rst.rstdt	93.37	69.61	79.76	96.80	95.92	96.35	93.39	94.50	93.94	94.31	96.16	95.23	91.32
eng.sdrt.stac	83.30	58.37	68.64	91.77	92.06	91.91	85.30	87.01	86.14	88.38	87.01	87.69	83.60
eus.rst.ert	84.49	64.05	72.87	92.70	88.38	90.47	91.45	83.78	87.45	86.47	90.68	88.52	84.83
fas.rst.prstc	86.77	64.63	74.08	92.95	92.78	92.86	93.59	89.40	91.45	94.20	89.70	91.90	87.57
fra.sdrt.annodis	86.33	52.10	64.98	87.95	83.79	85.78	89.90	86.41	88.12	89.67	87.06	88.34	81.81
nld.rst.nldt	91.48	82.54	86.78	96.97	92.54	94.69	94.35	93.79	94.07	96.07	94.08	95.07	92.65
por.rst.cstn	94.34	65.36	77.22	93.21	95.03	94.11	93.36	91.83	92.59	92.60	94.12	93.35	89.32
rus.rst.rrt	84.02	61.29	70.88	87.31	84.24	85.74	83.60	84.01	83.80	84.32	84.15	84.23	81.16
spa.rst.rststb	88.86	74.57	81.09	93.30	90.30	91.76	92.19	89.78	90.97	90.95	89.57	90.25	88.52
spa.rst.sctb	80.65	59.52	68.49	83.97	77.98	80.86	78.65	89.88	83.89	79.78	86.90	83.19	79.11
zho.rst.sctb	83.33	59.52	69.44	84.04	70.00	76.21	68.11	75.00	71.39	65.52	79.17	71.70	72.19
mean	87.70	66.28	75.25	91.44	88.34	89.79	88.56	88.58	88.51	88.42	89.65	88.96	85.63

Table 4: EDU Segmentation Results on Plain Tokenized Data.

corpus	TMVM			DiscoDisco*			disCut*			SegFormers			mean
(treebanked)	Р	R	F1	Р	R	F1	Р	R	F1	Р	R	F1	
eng.pdtb.pdtb	85.98	65.54	74.38	92.93	91.15	92.02	93.32	88.67	90.94	89.73	92.61	91.15	87.12
tur.pdtb.tdb	80.00	24.14	37.10	93.71	94.53	94.11	90.55	86.93	88.70	90.42	91.17	90.79	77.68
zho.pdtb.cdtb	30.00	0.96	1.86	89.19	85.95	87.52	84.43	66.03	74.10	85.05	87.50	86.26	62.44
mean	65.33	30.21	37.78	91.94	90.54	91.22	89.43	80.54	84.58	88.40	90.43	89.40	75.75
(plain)	Р	R	F1	P	R	F1	P	R	F1	P	R	F1	mean
eng.pdtb.pdtb	81.06	34.37	48.27	94.29	90.92	92.56	88.84	92.09	90.43	90.37	91.97	91.16	80.61
tur.pdtb.tdb	73.13	28.86	41.38	91.98	95.22	93.56	90.12	88.10	89.10	89.36	91.05	90.20	78.56
zho.pdtb.cdtb	53.33	5.13	9.36	90.27	86.54	88.35	77.40	72.44	74.83	85.25	83.33	84.28	64.21
mean	69.17	22.79	33.00	92.18	90.89	91.49	85.45	84.21	84.79	88.33	88.78	88.55	74.46

Table 5: Connective Detection Results.

task and the connective detection task; two systems tackled the relation classification task. For teams that submitted multiple systems, we selected the system that achieved the best macro-averaged F-score across datasets as the representative submission. For teams that were not able to obtain all the licensed data, we evaluated submitted systems on the closed LDC datasets for those team and reported the scores to the authors.

EDU Segmentation The main results for EDU segmentation on the test sets are given in Table 3 for

treebanked data, and in Table 4 for plain tokenized data. The results indicate improvements across the board for datasets that were represented in the 2019 Shared Task, including new state of the art scores on the benchmark English RST dataset RST-DT (97.09 using gold parses and sentence splits from the system SegFormers, and 96.35 for segmentation from plain text, without sentence boundary information, by the system DisCoDisCo). Improvements are also seen on small datasets, such as Chinese RST, now scoring 83.34 and 76.21 for treebanked / plain text data respectively (both by Dis-

CoDisco), compared to 81.67 and 73.13 in 2019.

The discrepancy between the scores in the two scenarios indicates that gold treebanked data still provides a substantial benefit to EDU segmentation. This may most obviously be the case due to the inclusion of gold sentence splits in the datasets that provide them, which usually coincide with EDU boundaries, though slightly higher scores from approaches using other gold features were also observed (see Section 6 for system analysis).

Connective Detection The main reuslts for connective detection on the test sets are given in Table 5 for both treebanked data and plain tokenized data. As with EDU segmentation, the results show substantial improvements on all datasets compared to 2019, with means climbing from the low 80s to the low 90s. The most notable improvement is on Chinese connective detection, which rises from high 70s to scores around 90, and Turkish, from low 80s to low 90s. The rise in scores for several Transformer-based system scenarios suggests that this may be at least in part a result of improvements in language models for languages other than English.

Relation Classification The main results for the relation classification task are given in Table 6. Since the relation classification task is new, it is difficult to evaluate the quality of the scores obtained in the results. For some datasets, such as PDTB, existing scores have been previously reported on the same underlying data, but generally in different settings, where implicit and explicit relations were scored separately. To make matters worse, most previous work on PDTB uses the older version 2 (Prasad et al., 2008), making scores again completely non-comparable. Previous scores on PDTB-V2 (level 2 relations, the same hierarchical level used for the Shared Task) with explicitly specified connectives have reached scores above 90 for a while (Kido and Aizawa, 2016); more recently, results on PDTB-V3 implicit relation classification have reached an accuracy of 64.83 (Kim et al., 2020). By comparison, we see a best score of 74.44 (DisCoDisCo) on all PDTB-V3 relations (including implicit and explicit), but the task is both easier and harder, since in both cases no connective is specified (for uniformity with other formalisms), but the relation direction is specified in the data (since the Shared Task targets a scenario of labeling unlabeled dependency structures).

corpus	DiscRel	DisCoDisCo*	mean
deu.rst.pcc	35.38	39.23	37.31
eng.pdtb.pdtb	50.02	74.44	62.23
eng.rst.gum	47.06	66.76	56.91
eng.rst.rstdt	55.41	67.10	61.26
eng.sdrt.stac	54.11	65.03	59.57
eus.rst.ert	44.25	60.62	52.44
fas.rst.prstc	58.95	52.53	55.74
fra.sdrt.annodis	45.76	46.40	46.08
nld.rst.nldt	45.40	55.21	50.31
por.rst.cstn	48.53	64.34	56.44
rus.rst.rrt	61.34	66.44	63.89
spa.rst.rststb	45.07	54.23	49.65
spa.rst.sctb	69.18	66.04	67.61
tur.pdtb.tdb	48.10	60.09	54.10
zho.pdtb.cdtb	88.65	86.49	87.57
zho.rst.sctb	70.44	64.15	67.30
mean	54.23	61.82	58.02

Table 6: Relation Classification Results.

Results on EDU datasets are equally difficult to compare to previous work due to the focus on full discourse parsing for frameworks like RST, whereas the relation classification task has focused solely on labeling gold graph structures. However, we are encouraged to see best scores above 67 for benchmark datasets such as RST-DT, which beat state-of-the-art discourse dependency parsers' labeled attachment scores (LAS), currently 51.8 (Zhang et al., 2021a); while it is understandable that LAS will always be lower due to the need to predict both attachment and labeling correctly, the Shared Task results indicate that progress on unlabeled attachment (currently 70.2 on RST-DT, ibid.) can lead to substantial gains in labeled accuracy when combined with the approaches tested in this Shared Task.

6 Analysis of Systems

EDU Segmentation and Connective Detection All four systems tackling the segmentation tasks addressed both EDU Segmentation and Connective Detection with a unified architecture, with the partial exception of DisCoDisCo, which varied in using a CRF layer for Connective Detection (in order to model B- and I- label sequence probabilities), but not for EDU Segmentation, where a CRF layer was found to be harmful.

Of the four systems, the three highest scoring ones relied on a Transformer-based neural sequence labeling approach backed by large, pretrained language models, while the fourth system, TMVM, focused on grammatical features extracted from gold and predicted trees in order to capture clause boundaries across languages. TMVM's submission is also unique in reporting scores when training on each individual dataset and predicting on other datasets using those features, many of which are general across languages. Of the datasets cross-tested by TMVM, training on a different dataset was beneficial for one English case (training on GUM and testing on RST-DT, leading to a 5 point gain from 75 to 80), and the two small Spanish and Chinese datasets (a 3 point gain from 66 to 69 for Spanish when training on the larger Spanish RST STB, and a more modest but surprising 2 point gain from 72 to 74 when training on Basque and testing on Chinese, see Dönicke 2021). The latter two cases are probably owing to the small size of the target corpora. The disCut paper takes a related approach and trains on multiple corpora from the same language or language family, but does not obtain superior results (see Ezzabady et al. 2021), possibly because the single-corpus training results are already so strong.

Among the Transformer-based systems, although DisCoDisCo performs best overall, other systems score highest on particular datasets in different scenarios, with SegFormers attaining a new SOTA score on RST-DT from gold trees (97.09) next to DisCoDisCo's new SOTA score on the same data without gold trees (96.35 in the plain text scenario), and disCut showing impressively strong performance on German (95.61 from gold trees, and an even better 95.63 in the plain text scenario). Overall best scores for EDU Segmentation in the top 3 systems are within a point of each other, but the consistent slight lead of DisCoDisCo may be due to the fact that it was the only segmentation system using per-language transformers and making use of grammatical features, which in the plain text scenario were induced using a SOTA Transformerbased parser and sentence splitter as well.

For Connective Detection, DisCoDisCo's lead becomes slightly larger, about 3 points above Seg-Formers and almost 7 points above disCut, perhaps due to a combination of features (which give a minor boost of just a few points across datasets in their paper, see Gessler et al. 2021), the use of a CRF layer and the combination of both contextualized and static word embeddings as well as character embeddings, which provided a variety of views on the data and allowed for both fine tuning and fixed generalization in the test set. Choice of language models is also an important factor, and here again disCut and SegFormers rely on multilingual LMs, while DisCoDisCo uses individual per-language models, which may also be a reason for higher scores on Chinese and especially Turkish, which may be less well-represented by multilingual LMs.

Finally we note that in the plain text scenario, all systems except for DisCoDisCo used noncontextualized tools for sentence splitting and/or automatic parsing (SegFormers: CoreNLP; disCut: stanza; TMVM: SpaCy); DisCoDisCo used the tranformer-based sentence splitter from the AMAL-GUM corpus (Gessler et al., 2020) and DiaParser (Attardi et al., 2021), both with language-specific transformers, with the result that plain text numbers are very close to gold numbers for DisCoDisCo (and in fact insignificantly better for plain Connective Detection: 91.49 on average vs. 91.22 for gold treebanked data). This echoes results from the 2019 task (see Yu et al. 2019) which showed the crucial importance of high quality preprocessing in general, and sentence splitting in particular.

Relation Classification Although both systems tackling the Relation Classification task use word embeddings, their approaches are rather different: while the best system, DisCoDisCo, relies on a similar approach to their segmentation task entry, which encodes each sequence using a transformer, the DiscRel system uses whole sentence embeddings from the Sentence-Transformers library (SBERT, Reimers and Gurevych 2020) to compute Euclidean distance between discourse units, as part of a two-level feature-based Random Forest classifier. Both systems also use the provided discourse relation direction, which is meant to approach the case of labelling an unlabelled discourse dependency graph.

In terms of encoding features, DisCoDisCo uses a sentence pair classifier architecture feeding the underlying transformer embeddings in the form of the BERT classification ([CLS]) token embeddings, while injecting categorical features at two levels: using pseudo tokens, encoded as usual by the transformer, which indicate the relations direction; and using an inserted pseudo token next to the CLS token, which encodes all sequence-level categorical features, such as genre, gold speaker information and more (see Gessler et al. 2021). These features are then read and trained on by the transformer block alongside the word embeddings.

The DiscRel system takes a two level approach to relation classification, using two stacked oneversus-rest Random Forest classifiers, which first attempt to distinguish the coarse class of discourse relation (for example collapsing relations such as EVALUATION, INTERPRETATION, JUSTIFY), and then sub-classify the coarse classes into the final target labels using the second classifier (see Varachkina and Pannach 2021). For large datasets, the first classifier is trained on the train set, while the second is trained using the development data, but for smaller datasets, training data is duplicated along with the development set for the second classifier.

DiscRel's approach performs substantially better for both Chinese datasets, the small Spanish dataset, and Farsi. A commonality in this set is its inclusion of the smallest datasets (the parallel zho.rst.sctb and spa.rst.sctb), which is perhaps unsurprising given the robustness of tree-based classifiers on small data. However, the gains on the Farsi data, which is not very small, are also substantial, and the system scores over 2 points better than Dis-CoDisCo on CDTB, suggesting either that SBERT has better embeddings for these languages, or in the case of CDTB, possibly that the small number of target relations (only 9, the smallest relation set) benefits from the stacked classifier approach.

For the remaining 12 datasets, DisCoDisCo scores highest, in many cases by a wide margin of over 10 points above DiscRel (including all English corpora, with +14 on PDTB, +18 on GUM and +12 on RST-DT, and +11 on STAC), leading to a final mean gain of over +7 points. Feature ablations in the DisCoDisCo paper suggest that categorical features confer a minor boost in scores, especially for datasets rich in gold metadata and annotations, but that the primary driver of accuracy is the use of contextualized word embeddings. We also note that, as with segmentation, DisCoDisCo uses languagespecific LMs for each dataset, and not always the same LMs as for segmentation, while DiscRel uses one of two multilingual SBERT models for all datasets. For choice of LMs, DisCoDisCo's paper indicates that transformers pretrained with the Next Sentence Prediction task (NSP) are particularly suited to relation classification, i.e. models such as BERT (Devlin et al., 2019), but not the RoBERTa family of models (Liu et al., 2019).

The gaps between the systems given the reproduction and multiple run averages suggest that the two approaches capture some orthogonal information, indicating that incorporating insights from both systems may be worth pursuing: sentence embeddings and Euclidean distance could be incorporated in a neural system, and the features used by DisCoDisCo could be used in a two stage approach such as the one taken by DiscRel.

7 Conclusion

We conducted the second DISRPT shared task for Discourse Relation Parsing and Treebanking across frameworks, resulting in a number of new state of the art scores on benchmark datasets, as well as the publication of a new multilingual benchmark for discourse parsing sub-tasks. With the introduction of the new classification task in 2021, the DISRPT data now covers discourse unit segmentation, connective detection, and relation classification for unit pairs in 16 datasets, coming from 11 languages and containing approx. 3,000,000 tokens.

With the exception of the distinction between Connective Detection and EDU Segmentation datasets, we were able to bring all datasets to a common denominator by relying on a dependency representation of the underlying discourse annotations. Especially in the case of Relation Classification, this means that we can potentially harness the Shared Task data for training and evaluation of discourse relation information using multiple datasets for the same language (where available), and learn from multiple frameworks at the same time. Publishing uniformly evaluated results using a single scorer, input format, and feature set, with a stable reference gold parse for each dataset, we hope to facilitate the conversation and collaboration between researchers coming from different frameworks, such as RST, SDRT, and PDTB.

We are optimistic that the future for computational work on discourse relations is bright, with scores on Segmentation and Connective Detection now looking much higher than in previous years (finally around the low 90s), and Relation Classification slowly catching up, but still offering a substantial challenge with scores in the low 60s. With some luck, we may soon see automatic discourse analysis methods become close to, or even as reliable, as morphosyntactic tagging and parsing have become in recent years. In order to realize that goal, we are certain that we will need new datasets covering previously unexplored languages and genres, as well as in depth analysis of the kinds of discourse relations we should recognize and the challenges in inducing them automatically and reliably across all types of text.

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).
- Nicholas Asher. 1993. *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.
- 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).
- Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for EnhancedUniversal dependencies. In Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021), pages 184–188, Online. Association for Computational Linguistics.
- 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.
- 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, 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.

- 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.
- Vera Demberg, Fatemeh Torabi Asr, and Merel Scholman. 2019. How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tillmann Dönicke. 2021. Delexicalised Multilingual Discourse Segmentation for DISRPT 2021 and Tense, Mood, Voice and Modality Tagging for 11 Languages. In Proceedings of Discourse Relation Parsing and Treebanking 2021 (DISRPT 2021), Punta Cana, Dominican Republic.
- Carla Parra Escartín, Teresa Lynn, J. Moorkens, and Jane Dunne. 2021. Towards transparency in NLP shared tasks. *ArXiv*, abs/2105.05020.
- Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual Discourse Segmentation and Connective Identification: MELODI at DISRPT2021. In *Proceedings of Discourse Relation Parsing and Treebanking 2021 (DISRPT 2021)*, Punta Cana, Dominican Republic.
- Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. Dis-CoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection. In *Proceedings of Discourse Relation Parsing and Treebanking 2021 (DISRPT* 2021), Punta Cana, Dominican Republic.
- Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. AMALGUM

 a free, balanced, multilayer English web corpus.

 In Proceedings of the 12th Language Resources and Evaluation Conference, pages 5267–5275, Marseille, France. European Language Resources Association.
- Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In Proceedings of the First Workshop on Computational Approaches to Discourse, pages 160–167, Online. 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.
- Yusuke Kido and Akiko Aizawa. 2016. Discourse relation sense classification with two-step classifiers. In *Proceedings of the CoNLL-16 shared task*, pages 129–135, Berlin, Germany. 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.
- Murathan Kurfali. 2020. Labeling explicit discourse relations using pre-trained language models. In *Text, Speech, and Dialogue. TSD 2020*, pages 79–86, Brno, Czech Republic. Springer.
- Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 362–371, Austin, Texas. 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 the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3830–3836. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.
- William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.
- Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

- Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, California. 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, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.
- 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).
- Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the* 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4512–4525, Online. Association for Computational Linguistics.
- Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.
- Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the 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.
- Hanna Varachkina and Franziska Pannach. 2021. A Unified Approach to Discourse Relation Classification in nine Languages. In *Proceedings of Discourse Relation Parsing and Treebanking 2021 (DISRPT* 2021), Punta Cana, Dominican Republic.
- Jianxiang Wang and Man Lan. 2015. A refined endto-end discourse parser. In *Proceedings of the Nine*-

teenth Conference on Computational Natural Language Learning - Shared Task, pages 17–24, Beijing, China. Association for Computational Linguistics.

- Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for English and Chinese in CoNLL-2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40, Berlin, Germany. Association for Computational Linguistics.
- Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. Gum-Drop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT* 2019), pages 133–143, Minneapolis, MN.
- 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 DIS-RPT 2019 shared task on elementary discourse unit segmentation and connective detection. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.
- 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 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*.
- Liwen Zhang, Ge Wang, Wenjuan Han, and Kewei Tu. 2021a. Adapting unsupervised syntactic parsing methodology for discourse dependency parsing. 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 5782–5794, Online. Association for Computational Linguistics.
- Longyin Zhang, Fang Kong, and Guodong Zhou. 2021b. Adversarial learning for discourse rhetorical structure parsing. 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 3946–3957, Online. Association for Computational Linguistics.
- Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese Discourse Treebank 0.5 LDC2014T21.