The DISRPT 2019 Shared Task on Elementary Discourse Unit Segmentation and Connective Detection *

Amir Zeldes
Georgetown University
az364@georgetown.edu

Debopam Das
University of Potsdam
ddas@sfu.ca

Erick Galani Maziero
Federal University of Lavras
erick.maziero@ufla.br

Juliano Desiderato Antonio
Universidade Estadual de Maringa
jdantonio@uem.br

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

Abstract

In 2019, we organized the first iteration of a shared task dedicated to the underlying units used in discourse parsing across formalisms: the DISRPT Shared Task on Elementary Discourse Unit Segmentation and Connective Detection. In this paper we review the data included in the task, which cover 2.6 million manually annotated tokens from 15 datasets in 10 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

The past few years have seen substantial advances in both the development of new discourse annotated corpora for diverse languages (e.g. Iruskieta et al. 2013, Zhou et al. 2014, Afantenos et al. 2012) and approaches to automatic discourse parsing relying on neural and other architectures (Braud et al. 2017, Wang and Lan 2015, Li et al. 2016, Perret et al. 2016). Across frameworks, most work producing substantial amounts of data in multiple languages has been developed within Rhetorical Structure Theory (Mann and Thompson, 1988), the Penn Discourse Treebank's framework (Prasad et al., 2014) and Segmented Discourse Representation Theory (Asher, 1993).

At the same time, there is reason to believe that performance on discourse parsing still has a substantial way to go (Morey et al., 2017), with scores on deep discourse parsing for well studied and homogeneous resources such as the English RST Discourse Treebank (Carlson et al., 2003) still well behind human annotators, and results for other datasets, especially in less studied and lower resource languages lagging much farther behind. To make matters worse, the vast majority of deep discourse parsing papers work with gold segmented discourse units, which allow for easier comparisons of scores, but represent an unrealistically easy scenario. In their recent survey of discourse parsing results, Morey et al. (2017, 1322) point out that "all the parsers in [their] sample except [two] predict binary trees over manually segmented EDUs",[1] meaning that we have very limited information on the accuracy of discourse parsing in realistic settings. In order for discourse parsing to come closer to the reliability of syntactic parsing, a similarly reliable state of the art (SOA) for segmentation into terminal units must be reached.

The comparison with work on syntax parsing brings another point of interest into focus: the recent success of Universal Dependencies (Nivre et al., 2017) as a standard bringing together resources from different languages has been instrumental in creating generic NLP tools that are flexible and applicable to a variety of tasks. This is not only due to converging cross-linguistic annotation guidelines and the codification of a uniform format based on the CoNLL shared task datasets, but also due to the community building afforded by the organization of joint workshops which bring together researchers from a range of domains.

Within this landscape, the first multilingual and cross-framework task on discourse unit segmentation and connective detection aims to promote the development of reliable tools for working with the basic building blocks of discourse annotation. Although it is clear that there are substantial dif-

* Discourse Relation Parsing and Treebanking (DISRPT): 7th Workshop on Rhetorical Structure Theory and Related Formalisms (https://sites.google.com/view/disrpt2019) was held in conjunction with Annual Conference of the NAACL 2019 in Minneapolis, MN.

[1] EDUs or Elementary Discourse units are non-overlapping "minimal building blocks of a discourse tree" (Carlson et al., 2003). EDUs are, mostly, (sentences or) clauses, except for complement and restrictive clauses.

EDU segmentation

corpus	language	framework	sentences	tokens	documents	units
deu.rst.pcc	German	RST	2,193	33,222	176	3,018
eng.rst.gum	English	RST	5,274	98,615	114	7,311
eng.rst.rstdt	English	RST	8,318	205,824	385	21,789
eng.sdrt.stac	English	SDRT	10,020	47,741	41	11,531
eus.rst.ert	Basque	RST	1,660	35,313	140	2,910
fra.sdrt.annodis	French	SDRT	1,318	32,411	86	3,709
nld.rst.nldt	Dutch	RST	1,707	24,920	80	2,371
por.rst.cstn	Portuguese	RST	1,950	54,656	136	4,734
rus.rst.rrt	Russian	RST	12,513	272,664	178	19,906
spa.rst.rststb	Spanish	RST	2,136	58,591	267	3,349
spa.rst.sctb	Spanish	RST	478	16,512	50	744
zho.rst.sctb	Mandarin	RST	563	14,442	50	744
Connective detection						
corpus	language	framework	sentences	tokens	documents	units
eng.pdtb.pdtb	English	PDTB	48,630	1,156,648	2,162	26,048
tur.pdtb.tdb	Turkish	PDTB	31,196	496,355	197	8,397
zho.pdtb.cdtb	Mandarin	PDTB	2,891	73,314	164	1,660

Table 1: Datasets in the DISRPT 2019 shared task.

ferences in guidelines and goals across different formalisms and datasets, we hope that the shared task will contribute to a broad discussion of discourse annotation standards and goals, and put less studied resources in focus, next to more frequently addressed corpora such as PDTB (Prasad et al., 2008) and RST-DT (Carlson et al., 2003). Additionally, the release of the DISRPT 2019 shared task dataset[2] in a uniform format, modeled on the CoNLL-U format used by Universal Dependencies, is meant as a first step in creating a multilingual testing grounds for discourse parsing systems, starting with the basic task of identifying the minimal locus at which discourse relations apply: discourse units and connectives.

2 Shared task data

The DISRPT 2019 shared task dataset comprises 15 datasets in 10 languages, 12 of which target elementary discourse unit segmentation, and 3 dedicated to explicit connective annotation. Table 1 gives an overview of the datasets. Of the 15 datasets, 14 were released approximately 1.5 months before the shared task deadline, while the final one, connective annotations from the Turkish Discourse Bank, was released as a 'surprise' dataset/language together with dev and test sets just two weeks before the announced deadline. For four of the datasets, licensing constraints prevented online publication of the underlying texts (e.g. Wall Street Journal material), meaning that the public repository contains only annotations for those corpora, with tokens replaced by underscores. 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.

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., 2003).
- eng.sdrt.stac - Strategic Conversations corpus (Asher et al., 2016).
- eus.rst.ert - Basque RST Treebank (Iruskieta et al., 2013).
- 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).

[2] https://github.com/disrpt/sharedtask2019.

```
# sent_id = GUM_interview_stardust-28
# text = Yes [see below].
1    Yes    yes    INTJ   UH     _                             0   root       _   BeginSeg=Yes
2    [      [      PUNCT  -LSB-  _                             3   punct      _   BeginSeg=Yes|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  [      _  _  _  _  _  _  _  BeginSeg=Yes
413  see    _  _  _  _  _  _  _  _
414  below  _  _  _  _  _  _  _  _
415  ]      _  _  _  _  _  _  _  _
416  .      _  _  _  _  _  _  _  _
```

Figure 1: Data formats: treebanked (*.conll, top) and plain (*.tok, bottom)

- spa.rst.rststb - RST Spanish Treebank (da Cunha et al., 2011).
- spa.rst.sctb - RST Spanish-Chinese Treebank (Spanish) (Shuyuan et al., 2018).
- tur.pdtb.tdb - Turkish Discourse Bank (Zeyrek et al., 2010).
- zho.pdtb.cdtb - Chinese Discourse Treebank (Zhou et al., 2014).
- zho.rst.sctb - RST Spanish-Chinese Treebank (Chinese) (Shuyuan et al., 2018).

As Table 1 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.

Data was released for all corpora in two formats, corresponding to two scenarios: Treebanked data (`*.conll`), which included an (ideally gold) dependency parse, including gold sentence splits and POS tags, and unannotated, plain tokens (`*.tok`). 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 specific tags), a morphology field with unlimited morphological annotations, and a secondary dependency field (only used in the Dutch dataset). 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.

The shared task repository also contained an evaluation script to score systems on each dataset. For both evaluations, we opted to compute precision, recall and F1 score on discourse unit segmentation and connective detection, micro-averaged within each dataset, and macro-averaged results across all corpora for each system in each scenario (treebank/plain tokens). Similarly to evaluation of NER performance, scores reward only the positive classes, i.e. precision and recall of segmentation is judged purely based on identification of segmentation points, with no reward for recognizing negative cases.

For connective detection, the evaluation targets exact span retrieval, meaning that precision and recall are calculated out of the total connective spans (not tokens) available in the gold data. This means that partial credit was not given: a system identifying the span in Example (1) is given one precision error and one recall error, since it 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

Dataset	ToNy			GumDrop			DFKI RF			IXA			Mean
(treebanked)	P	R	F	P	R	F	P	R	F	P	R	F	
deu.rst.pcc	95.22	94.76	**94.99**	93.33	90.48	91.88	95.33	83.33	88.93	90.91	91.84	91.37	91.86
eng.rst.gum	95.84	90.74	93.21	96.47	90.77	**93.53**	97.96	83.71	90.27	95.52	88.61	91.94	92.38
eng.rst.rstdt	95.29	96.81	**96.04**	94.88	96.46	95.67	93.65	85.47	89.37	94.56	94.93	94.75	93.99
eng.sdrt.stac	94.34	96.22	95.27	95.26	95.39	**95.32**	97.65	91.94	94.71	92.51	90.71	91.60	94.24
eus.rst.ert	89.77	82.87	**86.18**	90.89	74.03	81.60	92.77	60.54	73.27	91.19	80.27	85.38	82.40
fra.sdrt.annodis	94.42	88.12	**91.16**	94.38	86.47	90.25	94.04	81.18	87.13	91.10	90.50	90.79	89.96
nld.rst.nldt	97.90	89.59	93.56	96.44	94.48	**95.45**	98.38	88.08	92.95	90.91	93.02	91.95	93.60
por.rst.cstn	92.78	93.06	**92.92**	91.77	89.92	90.84	93.18	77.36	84.54	93.01	92.38	92.69	90.37
rus.rst.rrt	86.65	79.49	**82.91**	83.47	75.52	79.30	82.79	67.51	74.37	73.22	74.11	73.67	77.75
spa.rst.rststb	92.03	89.52	**90.74**	89.02	81.80	85.26	93.01	76.54	83.99	85.68	87.94	86.80	86.86
spa.rst.sctb	91.43	76.19	**83.12**	89.76	67.86	77.29	95.28	60.12	73.72	93.22	65.48	76.92	79.20
zho.rst.sctb	87.07	76.19	81.27	80.95	80.95	80.95	88.81	75.60	**81.67**	90.37	73.57	81.11	81.54
mean	92.73	87.80	**90.11**	91.38	85.34	88.11	93.57	77.61	84.58	90.18	85.28	87.41	87.84

Table 2: EDU segmentation results on treebanked data.

3 Results

We report precision, recall and F1 for systems in the two tasks, each consisting of two scenarios: EDU segmentation and connective detection, with treebanked and plain tokenized data. Four systems were submitted to the shared task, all of which attempted the EDU segmentation task, and three of which also approached the connective detection task for at least some datasets. For teams that submitted multiple systems, we selected the system that achieved the best macro-averaged F-score across datasets as the representative submission.

3.1 EDU segmentation

The main results for EDU segmentation on the test sets are given in Table 2 for treebanked data, and in Table 3 for plain tokenized data. No one system performs best on all corpora, suggesting that the different approaches have different merits in different settings. Overall, ToNy (Muller et al., 2019) performs best on the most datasets, and on average has the highest F-scores (90.11, computed by averaging five runs of the system, since GPU training was not deterministic). The next best systems by average F-score are GumDrop (Yu et al. 2019, 88.11 F_1), IXA (Iruskieta et al. 2019, 87.18 F_1) and DFKI RF (Bourgonje and Schäfer 2019, 84.56 F_1).

For the treebanked scenario, the best configuration for ToNy (using contextualized Bert embeddings, Devlin et al. 2018), receives the highest F-score on 8 datasets, the next best system, GumDrop, does so on 3 datasets, and DFKI's system on one: the Chinese RST dataset, which is notably the smallest one in the shared task with around 14,000 tokens.

Results for all systems show clearly that precision is usually higher than recall across the board. This suggests that some 'safe' strategies, such as assuming segment boundaries at the beginnings of sentences (which are gold standard split in most cases), yield good results, with the challenge being much more the identification of non-obvious segmentation points within sentences. Another obvious trend is the comparatively high performance on datasets that are large and gold-treebanked. The counterexample to the generalization that large corpora fare well is rus.rst.rrt, which can be explained by the lack of gold parses for this dataset, as well as some tricky conventions, such as handling segmentation differently within academic abstracts and bibliographies.

For the established RST benchmark dataset, RST-DT, two systems exceed the previous state of the art score (93.7, Bach et al. 2012), suggesting substantial progress (ToNy: 96.04; GumDrop: 95.67) compared to results previous to the shared task. For other languages, previous benchmark results using different corpora include F-scores of 80 for Spanish (Da Cunha et al., 2010), 73 for French (Afantenos et al., 2010), 83 for Basque (Iruskieta and Zapirain, 2015) and between 88 and 93 for German (Sidarenka et al., 2015).

For automatically parsed data, two systems submitted results, and results were extracted for a third system by shared task testers. The two systems that included results for this scenario in their papers were conincidentally also the top scoring systems overall, suggesting that numbers may represent the state of the art for this task. Inria's system ToNy achieves top performance on all but one dataset, and the best average F-score, possibly owing to the document-level model adopted by the system, in addition to the use of contextualized embeddings (see Section 4). Both top sys-

Dataset	ToNy			GumDrop			DFKI RF			Mean
(plain)	P	R	F	P	R	F	P	R	F	
deu.rst.pcc	94.88	94.49	**94.68**	91.99	89.80	90.88	94.20	71.77	81.47	89.35
eng.rst.gum	92.28	82.89	**87.33**	94.03	77.22	84.80	90.29	64.17	75.02	83.11
eng.rst.rstdt	93.60	93.27	**93.43**	89.56	91.43	90.49	45.96	35.85	40.28	74.87
eng.sdrt.stac	87.56	80.78	**83.99**	84.24	77.45	80.70	80.21	50.30	61.82	76.34
eus.rst.ert	87.43	80.94	**84.06**	90.06	73.36	80.86	88.21	58.01	69.99	79.21
fra.sdrt.annodis	94.31	89.15	**91.65**	94.46	85.29	89.64	93.47	67.35	78.29	87.07
nld.rst.nldt	94.81	89.97	**92.32**	94.72	88.41	91.45	95.14	68.12	79.39	88.26
por.rst.cstn	93.04	90.72	**91.86**	92.95	85.08	88.84	90.82	67.17	77.22	86.41
rus.rst.rrt	83.37	78.44	**80.83**	82.06	74.84	78.28	57.27	42.11	48.53	69.53
spa.rst.rststb	89.11	90.09	**89.60**	87.50	79.82	83.49	89.23	63.60	74.26	82.97
spa.rst.sctb	87.16	76.79	**81.65**	85.27	65.48	74.07	88.35	54.17	67.16	75.57
zho.rst.sctb	66.26	64.29	65.26	76.97	69.64	**73.13**	85.71	57.14	68.57	69.66
mean	88.65	84.31	**86.38**	88.65	79.82	83.89	83.24	58.31	68.5	80.19

Table 3: EDU segmentation results on plain tokenized data.

tems exceed the previous SOA of 89.5 on unparsed RST-DT: Georgetown's system GumDrop reaches 90.49, and ToNy achieves a remarkable 93.43, almost as high as previous results on gold parsed data. GumDrop performs better by a wide margin on the small Chinese dataset, but is overall well behind on many of the larger datasets, and about 2.5 F-score points lower on average than the best system, ToNy.

3.2 Connective detection

The main results for connective detection are given in Table 4. Three systems approached this task, though the DFKI system was not adapted substantially from the segmentation scenario, leading to low performance (Bourgonje and Schäfer, 2019), and did not report results on automatically parsed data.

ToNy again has the highest scores for the most datasets, obtaining the highest mean F-score for the plain tokenized scenario, and coming second to GumDrop only on the Turkish dataset in the gold syntax scenario. The margin for this particular result is however very wide, with GumDrop leading by almost 10 points, resulting in GumDrop obtaining the highest average F-score on gold syntax connective detection (though this score is in fact below the best plain tokenized result). This surprising result remained robust across 5 runs of the ToNy system (GumDrop was deterministically seeded and therefore reproducible in a single run).

Overall the connective detection results demonstrate that syntax is not central to the task (treebanked and plain results are close) and that accuracy is correlated with dataset size, presumably because the inventory of possible explicit connectives and their disambiguating environments is more exhaustively attested as the dataset grows.

4 Analysis of systems

The four systems submitted to the task all use either RNNs with word embeddings (ToNy, IXA), decision tree ensembles on linguistic features (DFKI's best system) or both (GumDrop). For two of the systems approaching both shared tasks, the same architecture is used for both connective detection and EDU segmentation, whereas GumDrop uses a slightly different architecture in each case. The high performance of ToNy on both tasks is remarkable in that a generic sequence labeling approach achieves excellent results despite not using engineered features or a tailored learning approach.

Looking at the internal distribution of scores for each system, we can observe that ToNy performs well on some of the less consistent resources, in particular for the automatically parsed and segmented Russian data, which all other systems degrade on, and on corpora with automatic parses but gold or very high quality sentence splits, such as the Spanish datasets and German. For some of the corpora with gold parses in the gold scenario, GumDrop takes the lead, perhaps thanks to the use of a large number of linguistic features next to character and word embeddings (notably for GUM, which has manually produced dependencies, rather than conversion from constituents in RST-DT).

ToNy's high scores on almost all datasets in the plain tokenized scenario seem to be related not only to contextualized embeddings substituting for missing morphosyntactic information, but also to the whole-document or large chunk approach (see Muller et al. 2019), which makes reli-

Dataset	ToNy			GumDrop			DFKI RF		
(treebanked)	P	R	F	P	R	F	P	R	F
eng.pdtb.pdtb	89.39	87.84	**88.60**	87.91	88.78	88.35	84.84	74.64	79.41
tur.pdtb.tdb	76.89	64.00	69.85	76.69	81.86	**79.19**	72.29	62.63	67.11
zho.pdtb.cdtb	82.67	76.25	**79.32**	81.27	70.22	75.35	73.21	43.22	54.35
mean	82.98	76.03	79.25	81.91	80.21	**80.93**	76.78	60.16	66.96
(plain)	P	R	F	P	R	F	P	R	F
eng.pdtb.pdtb	91.32	87.84	**89.54**	84.56	82.81	83.68	–	–	–
tur.pdtb.tdb	84.06	86.74	**85.37**	76.76	81.74	79.17	–	–	–
zho.pdtb.cdtb	81.64	71.07	**75.99**	80.62	67.31	73.37	–	–	–
mean	85.67	81.88	**83.63**	80.65	77.29	78.77	–	–	–

Table 4: Connective detection results.

able sentence splitting less crucial. At the same time, the performance advantage of the system is not found for the smallest corpus, zho.rst.sctb. DFKI was able to perform substantially better than ToNy for the gold scenario, while the next best system, GumDrop, takes the lead for Chinese on plain data, perhaps thanks to a high accuracy ensemble sentence splitter included in the system. The higher scores on this corpus for both DFKI and GumDrop, which employ Gradient Boosting and/or Random Forests, may suggest that the robustness of tree ensembles against overfitting allows for better generalization to the test data in the lowest resource scenario.

For connective detection, the best DFKI system using Random Forests does not attain good scores, probably due to the need to memorize sequences of vocabulary items. For English PDTB, ToNy and GumDrop are very close, suggesting that both systems can memorize the inventory of connectives and disambiguate ambiguous cases with similar success. For the smaller datasets, with the exception of the unexpectedly low performance on gold Turkish, ToNy has a more substantial lead. It is also worth noting that in 4/6 scenarios (all but Chinese), GumDrop has higher recall than precision, while ToNy has higher precision than recall in 5/6 scenarios, perhaps pointing to imbalanced learning issues for the latter versus weaker disambiguation capacity for the former.

5 Conclusion

By organizing the first shared task on EDU segmentation and connective detection, we hope to have pushed the field further in terms of bringing together resources and researchers from related fields, and making systems available that are flexible enough to tackle different datatset guidelines, but accurate enough to form the basis for deeper discourse parsing tasks in the future.

One particular point of progress has been making an official scorer and providing data in a uniform format based on the popular CoNLL-U specification used by Universal Dependencies. We expect this will make it easier to provide discourse annotations together with manually treebanked or automatically parsed data, as well as to compare future results with scores from this shared task. We also plan to maintain the DISRPT dataset and possibly extend it for future editions of the workshop.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Ccile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Pry-Woodley, Laurent Prvot, Josette Rebeyrolle, 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 LREC 2012*, pages 2727–2734, Istanbul, Turkey.

Stergos Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning recursive segments for discourse parsing. *arXiv preprint arXiv:1003.5372*.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of LREC 2016*, pages 2721–2727, Portorož, Slovenia.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of SIGDial 2012*, pages 160–168, Seoul, South Korea.

Peter Bourgonje and Robin Schäfer. 2019. Multilingual and cross-genre discourse unit segmentation.

In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, Minneapolis, MN.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST Discourse Parsing. In *Proceedings of EACL 2017*, pages 292–304, Valencia, Spain.

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 Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Text, Speech and Language Technology 22, pages 85–112. Kluwer, Dordrecht.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 1–10, Portland, OR.

Iria Da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. 2010. Discourse segmentation for Spanish based on shallow parsing. In *Mexican International Conference on Artificial Intelligence*, pages 13–23. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding.

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, Kepa Bengoetxea, Aitziber Atutxa, and Arantza Diaz de Ilarraza. 2019. Multilingual segmentation based on neural networks and pretrained word embeddings. In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, Minneapolis, MN.

Mikel Iruskieta and Benat Zapirain. 2015. Euseduseg: A dependency-based EDU segmentation for Basque. *Procesamiento del Lenguaje Natural*, 55:41–48.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of EMNLP 2016*, pages 362–371, Austin, TX.

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. 2017. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT. In *Proceedings of EMNLP 2017*, pages 1319–1324, Copenhagen, Denmark.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, Minneapolis, MN.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lê H`ông, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Ue-

matsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. Technical report, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jerémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of NAACL 2016*, pages 99–109, San Diego, CA.

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 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakesh, Morocco.

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 LREC 2012*, pages 2820–2825, Istanbul, Turkey.

Cao Shuyuan, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese Treebank. In *Proceedings of the Joint Workshop of Linguistic Annotation, Multiword Expression and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, NM.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse segmentation of German texts. *JLCL*, 30(1):71–98.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Language Resources and Evaluation Conference (LREC '14)*, pages 925–929, Reykjavik.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relation markers in Russian RST Treebank. In *Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of CoNLL 2015*, pages 17–24, Beijing.

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, Minneapolis, MN.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Deniz Zeyrek, Işın Demirşahin, Ayışığı B. Sevdik Çallı, and Ruket Çakıcı. 2010. The first question generation shared task evaluation challenge. *Dialogue and Discourse*, 3:75–99.

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