Signals as Features: Predicting Error/Success in Rhetorical Structure Parsing

Martial Pastor
Centre for Language Studies,
Radboud University,
The Netherlands
martial.pastor@ru.nl

Nelleke Oostdijk
Centre for Language Studies,
Radboud University,
The Netherlands
nelleke.oostdijk@ru.nl

Abstract

This study introduces an approach for evaluating the importance of signals proposed by Das and Taboada in discourse parsing. Previous studies using other signals indicate that discourse markers (DMs) are not consistently reliable cues and can act as distractors, complicating relations recognition. The study explores the effectiveness of alternative signal types, such as syntactic and genre-related signals, revealing their efficacy even when not predominant for specific relations. An experiment incorporating RST signals as features for a parser error / success prediction model demonstrates their relevance and provides insights into signal combinations that prevents (or facilitates) accurate relation recognition. The observations also identify challenges and potential confusion posed by specific signals. This study resulted in producing publicly available code and data, contributing to an accessible resources for research on RST signals in discourse parsing.

1 Introduction

Discourse parsing has sparked significant interest in recent NLP applications. This task goes beyond the conventional scope of sentences and may extend to encompass the identification of Coherence Relations (relations between segments of text) at the discourse level. One of the most popular formalisms for representing coherence relations is Rhetorical Structure Theory (RST; Mann and Thompson, 1988), which has spurred the construction of various datasets that are now used for hierarchical discourse parsing. This last task is challenging and discourse parsers have not achieved the same level of success as other tasks at the sentence level. Moreover, analyzing failure cases, especially in deep learning-oriented parsers, proves difficult.

Concurrently, research on Coherence Relations has also been struggling with identifying the exact linguistic elements that signal them. At the discourse level, a diverse array of signals may occur, making it challenging to discern typical signals for specific relations and their underlying motivation. This has been addressed in the work of Das (2014), where the author provides a comprehensive overview of signals present in the RST-DT dataset (Carlson et al., 2001) and subsequently annotate the signals at play for every relation in this corpus. This process ultimately has resulted in the development of the RST Signaling corpus (Das and Taboada, 2018a). See the Appendix for a comprehensive list of individual signals and signal types for all relations in this corpus.

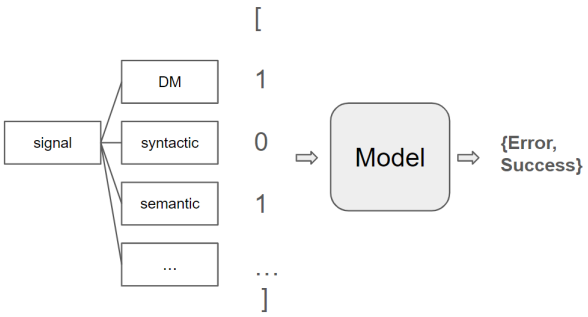

Figure 1: Flow diagram for the predictive model for error / success analysis of the DMRST parser. The predicted labels are SUCCESS for a successful parse while ERROR is where the parser fails. The features here are the signals from Das and Taboada's Signaling Corpus (Das and Taboada, 2018a). They are encoded in a binary feature vector.

In the present paper, our aim is to assess the relevance of Das and Taboada's signals in RST discourse parsing and understand how they contribute to the errors or success of a state-of-the-art parser.

We first describe an experimental set up where we replicate Liu et al.'s DMRST discourse parser (Liu et al., 2021) which achieves state-of-the-art results for coherence relation recognition and then

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 139–148

align the results with the RST Signaling corpus. We then show that although discourse markers (DMs) are prevalent in various sorts of relations, they are not necessarily effective signals (see, for example, the DM *when* for the TEMPORAL relation in example (1)) , unlike other types of signal such as syntactic signals (for example, the *nominal modifier* in example (2)) or genre signals.[1]

(1) "[representing investor clubs from around the U.S were attending] $\xleftarrow[\text{gold:temporal}]{\text{pred:background}}$ [when the market started to slide Friday.]" **wsj_2686**

(2) "[Negotiable, bank-backed business credit instruments] $\xleftarrow[\text{gold:elaboration}]{\text{pred:elaboration}}$ [typically financing an import order.]" **wsj_0602**

Next, we validate the initial analysis by incorporating these signals into a model, using them as features for a predictive model that distinguishes between errors and successes of the DMRST parser (see the flow diagram in Figure 1). We first observe that these signals could serve as relevant features in the context of Discourse Parsing before delving into a more detailed analysis of the signals influencing the parser's predictions of errors or success.

2 Related Work

2.1 Discourse Markers and beyond

In the broader context of literature focusing on text comprehension and cognitive linguistics, investigations into the cognitive aspects of coherence relations reveal that the presence of discourse markers (DMs), or connectives as they are sometimes referred to, tends to facilitate the processing of textual information (Gaddy et al., 2001). This particular line of research has primarily delved into recognizing and categorizing coherence relations using DMs. However, a limitation of this approach is its failure to address relations that seem unmarked due to the absence of DMs.

While DMs are commonly considered the most effective indicators for identifying coherence relations, studies on signaling show that a significant proportion of relations occurs in text without the presence of DMs (Das, 2014). Das and Taboada (2018b) explore the nature of relations traditionally considered implicit or unmarked. They reveal that relations exclusively signaled by DMs constitute only 18.21% of the RST Signaling corpus. This suggests that the signaling of coherence relations is more intricate than previously perceived. The researchers then propose their own taxonomy of various signals, ultimately contributing to the development of the RST Signaling corpus (Das and Taboada, 2018a), which we use for the experiments presented in this article.

2.2 Signals and Discourse Parsers

In early studies researching the effectiveness of linguistic elements for Discourse Parsing, several investigations have explored the importance of DMs (Pitler et al., 2008). For instance, the DM *if* in example (3) is usually considered to make the CONDITION relation easy to identify.

(3) "[If I sell now,] $\xrightarrow[\text{gold:condition}]{\text{pred:condition}}$ [I'll take a big loss.]" **wsj_2386**

The role of DMs has been emphasized, particularly in the context of shallow discourse parsing with the Penn Discourse Treebank (PDTB; Prasad et al., 2008). Previous studies suggest that in a shallow parsing context, which is distinct from RST as it focuses solely on local relations in text and disregards paragraph-level structures, explicit relations are the most straightforward to recognize. Moreover, there is a widely held consensus that the sole signals involved in explicit relations are discourse markers (DMs). Studies, such as the one conducted by Knaebel (2021), demonstrate the efficacy of neural shallow parsers utilizing contextualized embeddings in identifying relations explicitly marked by DMs, achieving an F1 score of 62.75% for explicit and 40.71% for implicit relations on Section 23 of PDTB v2 (Prasad et al., 2014). Additionally, the best performing system in the relation classification task in the shared initiative established by Zeldes et al. (2021) reported a mean accuracy of 79.32% for explicit relations and 50.86% for implicit relations in the 2023 edition (Braud et al., 2023).

Although certain corpus linguistics investigations have examined DMs in the RST dataset (Das and Taboada, 2018b; Stede and Neumann, 2014), only the work conducted by Liu et al. (2023) delves into the particular role of DMs in RST parsing and begins to question their pervasiveness as effective signals. After examining both the RST-DT corpus and the GUM dataset (Zeldes, 2017) which have

[1] pred here corresponds to the predicted label by the DMRST parser presented in section 3.2 and gold corresponds to the label annotated in the gold RST-DT dataset.

been annotated with DMs and other signals, they found that, although DMs have a notable impact, their significance is overshadowed by certain intra-sentential characteristics when predicting relation labels. While this confirms the relatively easier classification of explicit relations, the subsequent analysis by the authors indicates that explicitness is not confined exclusively to discourse markers; it also extends to other intra-sentential elements. This emphasizes the need for additional research into textual elements that explicitly signal coherence relations.

2.3 Predicting Parsing Errors

When it comes to constructing models to understand parsing performance, our reference is primarily Liu et al.'s investigation, which focuses on the prediction of parsing errors (Liu et al., 2023). Liu et al. replicate several parsers and given a coherence relations and its signals they predict the number of parsers that make errors. These parsers serve the purpose of detecting those cases in which the relation label assignment is likely or potentially at fault. Following an analysis of the essential features in their predictive model for error analysis, they note the significance of syntactic signals. This underscores the importance of determining whether an Elementary Discourse Unit (EDU) holds a typical intra-sentential role, such as nominal modifier or adjunct, as such roles are more likely to be predicted accurately. Additional influential features include EDU length, with shorter EDUs more likely to have comparable instances in the training data compared to longer ones, and genre, as certain genres present greater difficulty in parsing.

3 Experimental Setup

3.1 Datasets

Our current study uses two RST corpora. One is the RST-DT dataset (Carlson et al., 2001) which is widely used for English RST parsing and has been a standard choice for evaluating RST parsers. Additionally, we here incorporate the RST Signaling corpus by Das and Taboada (2018a), which is essentially an extension of the original RST-DT dataset. The signaling dataset contains additional annotations pertaining to the linguistic elements that signal coherence relations within the original RST corpus.

3.1.1 RST Discourse Treebank

The RST-DT is known for its hierarchical tree structures and was initially annotated with 76 coherence relations. The relations investigated here come from the RST-DT test set, which contains a total of 38 documents. As for the relations labels, we currently employ the harmonized set of 18 labels as described by Braud et al. (2017).

3.1.2 RST Signaling Corpus

In the RST Signaling corpus, every single relation in the RST-DT has been annotated for the linguistic element(s) that signal the relation. In this corpus, a total of 50 different signals are identified (Das and Taboada, 2019). The authors distinguish between three main classes, viz. single, combined and unsure. The single signals belong to one of the following types: DM, reference, lexical, semantic, syntactic, graphic, genre, and numerical. With combined signals multiple (single) signals co-occur. "unsure" is used a signal label with those relations where the annotators were either unsure or were unable to identify any specific signal .

Regarding Liu et al.'s remarks about difficulties exploiting data from the RST signaling corpus, it is important to note that the data indeed offers an alignment of the annotations with specific tokens. However, an error in the calculation of token positions in the annotations scheme was identified and subsequently rectified. Following the recalculation of positions, we are now able to align the RST signals from Das and Taboada 2018a with the RST-DT test set.[2]

3.2 DMRST Discourse Parser

The experimental setup first replicates the DMRST parser developed by Liu et al. (2021). This parser, based on XLM-RoBERTa-base (Conneau et al., 2020), is a top-down multilingual system that concurrently handles EDU segmentation and RST tree parsing. Its suitability for our purposes lies in its state-of-the-art performance in relation label prediction. The authors have provided access to a well-trained model through a readily available model checkpoint optimized for inference. This particular model underwent training on a multilingual collection of RST discourse treebanks, offering native support for six languages: English, Portuguese, Spanish, German, Dutch, and Basque.

[2]The code for aligning RST signals and for the experiments can be found here: https://github.com/metabolean5/signals-as-features

We use this model to predict the labels of the 2306 relations in the RST-DT test set and obtain an accuracy of 0.67 using the RST-Parseval metrics (Marcu, 2000).

It is worth noting that, although the parser can predict tree structure and discourse relations directly from raw text, our study opts to utilize gold EDU segmentation. In our experimental configuration, we input both the raw text from the original RST-DT test set and the segment breaks based on gold EDU segmentation.

4 Analysis

4.1 Preliminary Analysis of RST Signals

Here, we present an initial analysis of the signal distribution across the RST-DT test set. While we previously delved into Das and Taboada's analysis in the related works section, we now wish to underscore additional aspects of their signal annotation work. Notably, a significant disparity exists among various types of relations and their corresponding signals. For example, the ATTRIBUTION relation, the most successfully recognized relation by the DMRST parser, has only one relation which is signaled by a DM out of 343 relations. The ELABORATION relation, accounting for 796 instances in the test set, is signaled by a diverse array of signals (29 different signals), with only 24 cases attributed to a DM. Additionally, the SAME-UNIT relation is exclusively indicated by a singular syntactic signal, namely the *interrupted matrix clause* (127 cases).

Nonetheless, DMs continue to serve as the main signal type for certain relations. In the case of CONTRAST relations within the RST-DT test set, a DM is used to signal 112 out of 144 instances. Additionally, for CONDITION relations, 41 DMs are used in 48 cases, and for TEMPORAL relations, 47 DMs in 73 cases.

4.2 Signal Analysis of Discourse Parser Performance

4.2.1 DMs

In this section, we examine the specific performance of the DMRST parser for certain relations. The complete statistics for this section are available in the Appendix.

In cases where the relation is signaled by a DM, the DMs prove helpful for some relations: for example, 83% of the CONDITION relations signaled by DMs were correctly predicted. However, they do not necessarily make the identification easier. For CONTRAST, 73% of the relations signalled by DMs are successfully predicted and only 33% for TEMPORAL.

As for BACKGROUND relations, where DMs are still predominant but not as overwhelmingly so (53 relations signalled by DMs out of 111 cases), the parser correctly predicts 53% of them. We also observe that for the 796 ELABORATION relations, which the parser usually gets right (79% of them being successfully predicted), only 50% of relations indicated by a DM (12 out of 24 cases) are correctly predicted. The most effective signals here being syntactic.

We also note here, that 9 of the 12 cases which were not predicted correctly for this relation were either JOINT (5) or CONTRAST (4) which are relations where DMs are widely present. The confusion induced by specific DMs can offer valuable insights into the nature of distractors, a concern addressed in Liu et al. (2023). An example is the DM *and* which is typical of JOINT relations, and which might function as a distractor despite its intended role as a signal for ELABORATION. A similar kind of confusion arises with the discourse marker *when* in example (1), frequently causing the parser to misclassify temporal relations as background relations and vice versa. Similarly, we also observe that the DM *but*, while predominant in the CONTRAST relation, is also present with lower frequency in various other relations such as BACKGROUND, JOINT, ELABORATION, or CAUSE and causes comparable confusions.

(4) "[Yet another political scandal is racking Japan.] $\xrightarrow[\text{gold:cause}]{\text{pred:contrast}}$ [But this time it's hurting opposition as well as ruling-party members.]" **wsj_1189**

What emerges from this picture, is that in cases where DMs are typical of certain relations (CONDITION and CONTRAST), the model picks up on these DMs and they do play a role in correct relation label recognition. However, this is not observed for TEMPORAL relations, where DMs offer little or no assistance. Then again, TEMPORAL relations are generally hard to predict. Finally, when it comes to other relations where DMs are involved, we observe that they are not very reliable as signals and that they tend to create confusion with other relations typically signaled by DMs as seen in examples (1) and (4).

4.2.2 Other signals

Consistent with Liu et al. prior findings, our utilization of the RST signaling dataset demonstrates the effectiveness of syntactic signals for the majority of relations successfully predicted by the DMSRT parser. Similar to observations with DMs, relations typically signaled by specific syntactic cues exemplify this pattern. Notably, ATTRIBUTION, with a parser accuracy of 97%, shows 337 out of 343 relations indicated by the *reported speech* signal. ENABLEMENT follows, where 40 out of 46 relations are signaled by the *infinitival clause*, with the parser achieving 85% accuracy. The final noteworthy example is SAME-UNIT relations (127 cases), exclusively signaled by an *interrupted matrix clause*, predicted by the parser with 95% accuracy.

In the case of relations such as JOINT or ELABORATION, which are signaled by a variety of signals, syntactic ones, while not dominant, contribute to the parser's accuracy. For instance, with ELABORATION relations, those signaled by a *relative clause* (142 cases out of 796) show a 99% success rate in parser predictions.

(5) "[he hoped for unanimous support for a resolution] $\xleftarrow[\text{gold:elaboration}]{\text{pred:elaboration}}$ [he plans to offer tomorrow]"
wsj_1189

Similarly, in JOINT relations, 80% of accurately predicted *parallel syntactic constructions* (representing 30 out of 212 relations for this label) demonstrate a comparable pattern.

This implies that, unlike DMs, syntactic signals remain reliable even when not predominant. This is attributable to the specificity of syntactic structures, which are closely tied to individual relations and are not as ambiguous as DMs. Of note, we see that syntactic specificity cannot just be explained by the fact that syntactic signals, unlike DMs, belong to a set of repeated sequences or lexicalized forms. Though that may be the case for the *reported speech* signal with verba dicendi (verbs like 'say', 'report', and 'declare'), we can see that even when the relative pronoun *that* is dropped in example (5), the relation is still systematically correctly predicted.

In a similar manner, although not prominently featured in the entire RST-DT signaling corpus, the *genre* category stands out as an effective signal for various types of relationship Notably, 83% of the relations signaled by this category are accurately predicted.

5 Predictive Model for Success/Error Analysis

In this section we aim to provide a deeper insight of the previous analysis by building an error / success prediction model. Our goal here is to utilize signals from Das and Taboada's Signaling corpus to predict whether the DMRST parser will encounter an error or not. This approach enables us to assess the utility of signals in Discourse Parsing and determine if the presence or absence of these signals is linked to errors or successful parsing outcomes.

The implementation of our predictive model for error/success analysis is based on the XGBoost algorithm (Chen and Guestrin, 2016). This ensemble gradient boosting approach is renowned for its high accuracy. It has the capability to capture arbitrary interactions among features and is well-regularized to avoid overfitting.

The present experiment consists in training an XGBoost model to predict the DMRST parsing errors, the predicted label set being {1,0} where 1 is a correctly predicted error or successful parse and where 0 is where our model fails. The signals from the Signaling Corpus are encoded in a binary feature vector. With this configuration we train XGBoost on the 2306 relations outputs by the DMRST parser and get an 0.78 accuracy for a randomly selected 761 relations test set. Figure 2 presents an analysis of feature importance using classification gain which is often used to estimate feature importance (Shang et al., 2019).

Table 1 gives an overview of the distribution of the coherence relations in the test set, while Table 2 presents the distribution of the signal classes and types. Table 3 details the predicted error/success rate for specific signal types.

5.1 Observations

The most reliable of single signals overall are syntactic ones (91.6% correct+1), genre (83.3% correct+1), graphical (82.8%) and DM (59.0% correct+1). Here we note that (specific) syntactic signals are used with specific coherence relations: in the case of ATTRIBUTION we find *reported speech*, with ELABORATION we find mostly *relative clauses* and *nominal modifiers* and with SAME UNIT it is the *interrupted clause* that is used predominantly to signal the relation. Genre (*inverted*

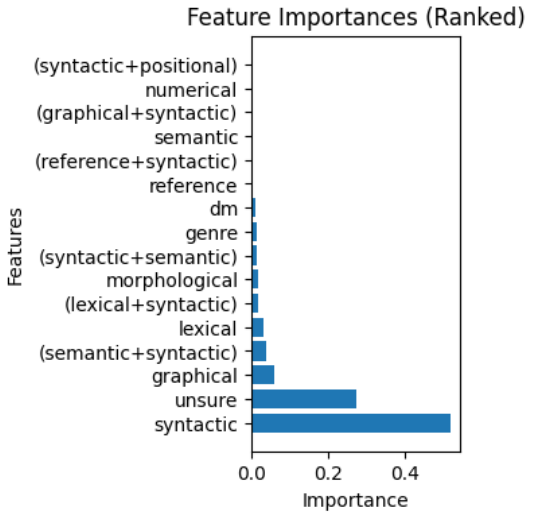

Figure 2: The relative importance of signals as features. Feature importance is based on classification gain which is often used to estimate feature importance (Shang et al., 2019).

Relation	abs. frq. (N)	rel. frq. (%)
attribution	106	13.9
background	36	4.7
cause	33	4.3
comparison	5	0.7
condition	12	1.6
contrast	46	6.0
elaboration	279	36.7
enablement	13	1.7
evaluation	26	3.4
explanation	36	4.7
joint	67	8.8
manner-means	8	1.1
same unit	39	5.1
summary	13	1.7
temporal	26	3.4
textual organization	4	0.5
topic-change	2	0.3
topic-comment	10	1.3
ALL	761	100.0

Table 1: Frequency distribution of coherence relations in the 761 relations test set.

Sign. class	signal type	abs. frq. (N)	rel. frq. (%)
single	DM	144	16.9
	reference	8	1.1
	lexical	26	3.4
	semantic	61	8.0
	morph.	8	1.1
	syntactic	275	36.1
	graphical	58	7.6
	genre	24	3.2
	numerical	0	0.0
combined	sem.+syn.	32	4.2
	lex.+syn.	6	0.8
	syn.+sem.	11	1.4
	syn.+pos.	0	0.0
	grap.+syn.	12	1.6
unsure	unsure	78	10.2
	ALL	761	100.0

Table 2: Signal classes and types in the 761 relations test set.

pyramid scheme) is almost exclusively (17/18 correct+1) used with ELABORATION. Graphical signals, especially *colon* and *dash* are used with ELABORATION, while graphical - *items in sequence* are typically used with JOINT. DMs are effective in signaling the CONTRAST, JOINT and CONDITION relations. The most used DMs here are *but* for CONTRAST, *and* for JOINT, and *if* for CONDITION. With BACKGROUND, CAUSE and TEMPORAL DMs perform really poorly.

As for signals that appear effective in predicting errors in relation label assignment, there were three specifically that stood out. Thus *indicative word* was encountered as a signal with a total of 26 cases out of which 15 cases of EVALUATION were predicted as true error (error+1). *lexical chain* was found with a total of 37 cases out of which 20 cases appear as error+1 (mostly EXPLANATION, CAUSE, and ELABORATION). Finally, *unsure* proved to be a good predictor for error+1. It occurred as a 'signal' with a total of 78 cases, 66 (84.6%, Table3) of which were found to be erroneous which was correctly predicted by our predictive model. *unsure* occurred most frequently with CAUSE (14/33 cases), EXPLANATION (16/36 cases), and ELABORATION (11/279 cases).

6 Conclusion

We have presented an approach for assessing the importance of Das and Taboada's signals within the context of discourse parsing. Our initial observations reveal distinct patterns in the performance of a discourse parser when graphed for specific signals, leading to various implications.

Initially, it is noted that DMs are not consistently reliable signals for all relationships; in fact, they can be viewed as *distractors*, causing confusion between relations signaled by the same DMs. Subsequently, an examination of the effectiveness of alternative signal types, including syntactic, semantic, and genre-related signals, is conducted. The findings demonstrate that, despite certain syntactic signals not being predominant for specific relations, they still prove to be effective.

Subsequently, we conduct an experiment incorporating the modeling of RST signals as features for an parser error or parser success prediction model. The results demonstrate the relevance of

Signal	corr.+1	corr.+0	err.+1	err.+0
DM	**59.0**	3.5	3.5	34.0
reference	0.0	25.0	37.5	37.5
lexical	0.0	15.4	**84.6**	0.0
semantic	11.5	37.7	**44.3**	6.6
morph.	0.0	0.0	100	0.0
syntactic	**91.6**	0.4	0.0	8.0
graphical	**82.8**	1.7	3.4	12.1
genre	**83.8**	0.0	0.0	16.7
numerical	0.0	0.0	0.0	0.0
ref.+syn.	0.0	**50.0**	33.3	16.7
sem.+syn.	**56.3**	0.0	0.0	43.8
lex.+syn.	**100.0**	0.0	0.0	0.0
syn.+sem.	**72.7**	0.0	0.0	27.3
syn.+pos.	0.0	0.0	0.0	0.0
grap.+syn.	16.7	33.3	0.0	**50.0**
unsure	0.0	15.4	**84.6**	0.0

Table 3: Predicted error/success rate (%) for specific signal types used to signal coherence relations. Correct/Error denotes whether the relation label assigned by the DMRST parser was correct, while 1/0 indicates whether the Predictive model was able to predict the accuracy (1=yes, 0=no).

utilizing signals as features, providing valuable insights into the signals (or combination of signals), that facilitate relation recognition. Moreover, our observations also shed light on scenarios where the presence of specific signals might pose challenges or lead to confusion, making it difficult for the parser to accurately discern certain relations.

Finally, we plan on sharing both our code and data, providing a readily accessible resource for research on RST signals within the context of discourse parsing.

7 Limitations

Initially, the examination of imbalances in characteristics constituted a challenge because of the multilingual composition of the training dataset. Furthermore, depending on a single model checkpoint for experimentation introduces the potential for errors influenced by coincidental variations in training. Additionally, we highlight that the corpus is restricted to newswire data, and exploring data from different genres is likely to provide additional insights.

It is also important to mention that in the current study, we specifically examined only those instances of potential signals that were identified as relevant for labeling coherence relations. This approach thus excluded what Liu et al. (2023) refer to as *distractors*.

8 Acknowledgements

This paper was produced as part of the HYBRIDS project, a Marie Skłodowoska-Curie Doctoral Network funded by the European Union under grant no. 101073351 and the UK Research and Innovation (UKRI) Horizon Funding Guarantee.

We express our gratitude to the reviewers for their valuable feedback. Their feedback has helped us addressing key points of interest and implications arising from the current study.

References

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.

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*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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.

Debopam Das. 2014. *Signalling of coherence relations in discourse*. Ph.D. thesis.

Debopam Das and Maite Taboada. 2018a. Rst signalling corpus: a corpus of signals of coherence relations. *Language Resources and Evaluation*, 52.

Debopam Das and Maite Taboada. 2018b. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770.

Debopam Das and Maite Taboada. 2019. Multiple signals of coherence relations. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Michelle L Gaddy, Paul van den Broek, and Yung-Chi Sung. 2001. The inxuence of text cues on the allocation of attention during reading. *Text representation: Linguistic and psycholinguistic aspects*, 8:89.

René Knaebel. 2021. Discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, and Amir Zeldes. 2023. What's hard in English RST parsing? predictive models for error analysis. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. 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.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

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.

Erbo Shang, Xiaohua Liu, Hailong Wang, Yangfeng Rong, and Yuerong Liu. 2019. Research on the application of artificial intelligence and distributed parallel computing in archives classification. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1267–1271.

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).

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

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.

Appendix: Summary of the DMRST parser's performance for all signals and relations

Relation	Signal Type	Signal	Correct	Error	Total N
Attribution	DM	DM	0.00	**1.00**	1
	Syntactic	Reported Speech	**0.98**	0.02	337
	Graphical	Colon	0.00	**1.00**	1
	Genre	Newspaper Style Attr.	**0.75**	0.25	4
Background	DM	DM	**0.53**	0.47	53
	Lexical	Indicative Word	0.11	**0.89**	9
	Syntactic	Past Part. Clause	0.00	**1.00**	1
		Present Part. Clause	0.40	**0.60**	5
		Relative Clause	0.33	**0.67**	3
	Morphological	Tense	0.04	**0.96**	23
	Synt.+Positional	Present Part. Clause+Beginning	0.50	0.50	2
	Unsure	Unsure	0.00	**1.00**	16
Cause	DM	DM	0.15	**0.85**	27
	Reference	Reference	0.00	**1.00**	1
		Comparative Reference	0.00	**1.00**	1
	Lexical	Alternative Expression	0.00	**1.00**	1
		Indicative Word	0.00	**1.00**	3
	Semantic	Lexical Chain	0.27	**0.73**	11
	Morphological	Tense	0.00	**1.00**	3
	Syntactic	Infinitival Clause	0.00	**1.00**	2
		Present Part. Clause	**0.75**	0.25	4
	Graphical+Synt.	Comma+Present Part. Clause	**0.75**	0.25	4
	Unsure	Unsure	0.00	**1.00**	29
Comparison	DM	DM	0.36	**0.64**	11
	Reference	Reference	0.33	**0.67**	3
		Comparative Reference	0.33	**0.67**	3
	Lexical	Indicative Word	0.50	0.50	4
	Semantic	Lexical Chain	0.14	**0.86**	7
	Syntactic	Parallel Synt. Constr.	**1.00**	0.00	1
	Synt.+Semantic	Parallel Synt. Constr.+Lex. Chain	**1.00**	0.00	1
	Unsure	Unsure	0.25	**0.75**	4
Condition	DM	DM	**0.83**	0.17	41
	Unsure	Unsure	0.14	**0.86**	7
Contrast	DM	DM	**0.73**	0.27	112
	Semantic	Lex. Chain	0.25	**0.75**	12
	Syntactic	Parallel Synt. Constr.	0.40	**0.60**	5
	Synt.+Semantic	Parallel Synt. Constr.+Lex. Chain	0.40	**0.60**	5
	Unsure	Unsure	0.05	**0.95**	20
Elaboration	DM	DM	0.50	0.50	24
	Reference	Personal Reference	0.44	**0.56**	68
		Propositional Reference	0.00	**1.00**	3
	Lexical	Indicative Word	**0.67**	0.33	3
	Semantic	Meronymy	**0.80**	0.11	18
		Repetition	**0.75**	0.25	61
		Synonymy	**1.00**	0.00	2
	Syntactic	Nominal Modifier	**0.91**	0.09	180
		Adj Modifier	0.00	**1.00**	2
		Infinitival Clause	0.00	**1.00**	4
		Present Part. Clause	**0.62**	0.38	8
		Relative Clause	**0.99**	0.01	142
	Graphical	Colon	**0.89**	0.11	36
		Dash	**0.95**	0.05	41
		Items in Sequence	0.00	**1.00**	2
		Parentheses	**1.00**	0.00	15
	Genre	Inverted Pyramid Scheme	**0.85**	0.15	47
	Graphical+Synt.	Comma+Present Part. Clause	0.57	0.43	7
	Lexical+Synt.	Lexical Chain+Subject NP	**0.78**	0.22	45
	Semantic+Synt.	General Word+Subject NP	0.50	0.50	2
		Meronymy+Subject NP	**0.87**	0.13	15
		Repetition+Subject NP	**0.77**	0.23	48
		Synonymy+Subject NP	**1.00**	0.00	2
	Ref.+Synt.	Personal Ref.+Subject NP	0.46	**0.54**	57
		Proposit. Ref.+Subject NP	0.00	**1.00**	2
	Unsure	Unsure	0.45	**0.55**	33

Relation	Signal Type	Signal	Correct	Error	Total N
Enablement	DM	DM	0.00	**1.00**	1
	Syntactic	Infinitival Clause	**0.85**	0.15	40
	Unsure	Unsure	0.20	**0.80**	5
Evaluation	DM	DM	0.25	**0.75**	8
	Lexical	Alternative Expression	0.00	**1.00**	5
		Indicative Word	0.10	**0.90**	50
	Graphical	Parentheses	0.00	**1.00**	4
	Unsure	Unsure	0.15	**0.85**	13
Explanation	DM	DM	0.25	**0.75**	8
	Lexical	Alternative Expression	0.5	0.5	4
		Indicative Word	0.00	**1.00**	1
	Semantic	Lexical Chain	0.09	**0.91**	34
	Syntactic	Infinitival Clause	0.00	**1.00**	1
	Unsure	Unsure	0.18	**0.82**	44
Joint	DM	DM	**0.83**	0.17	76
	Lexical	Indicative Word	0.38	**0.62**	8
	Semantic	Lexical Chain	**0.73**	0.27	60
	Syntactic	Parallel Synt. Constr.	**0.80**	0.20	30
	Graphical	Items in Sequence	**0.98**	0.02	41
	Synt+Lexical	Parallel Synt. Constr.+Lex. Chain	**0.85**	0.15	20
	Unsure	Unsure	0.41	**0.59**	17
Manner-Means	DM	DM	0.00	**1.0**	1
	Lexical	Indicative Word	**0.80**	0.20	15
	Syntactic	Present Part. Clause	0.00	**1.00**	4
	Graph.+Synt.	Comma+Present Participle Clause	0.00	**1.00**	2
	Lexical+Synt.	Indicative Word+Part. Clause	**0.86**	0.14	14
	Unsure	Unsure	0.00	**1.00**	7
Same-Unit	Syntactic	Interrupted Matrix Clause	**0.95**	0.05	127
Summary	DM	DM	0.00	**1.00**	1
	Semantic	Lexical Chain	0.00	**1.00**	1
		Repetition	0.00	**1.00**	2
	Graphical	Parentheses	**0.67**	0.33	15
		Colon	0.00	**1.00**	2
		Dash	0.00	**1.00**	1
	Genre	Inverted Pyramid Scheme	0.00	**1.00**	3
	Lexical+Synt.	Lexical Chain+Subject NP	0.00	**1.00**	1
	Semantic+Synt.	Repetition+Subject NP	0.00	**1.00**	1
	Unsure	Unsure	0.00	**1.00**	7
Temporal	DM	DM	0.30	**0.70**	47
	Lexical	Indicative Word	**0.75**	0.25	4
	Semantic	Indicative Word Pair	**1.00**	0.00	1
		Lexical Chain	0.20	**0.80**	5
	Morphological	Tense	0.00	**1.00**	2
	Syntactic	Relative Clause	0.25	**0.75**	4
	Unsure	Unsure	0.18	**0.82**	11
Textual Org.	Genre	Newspaper Layout	**0.78**	0.22	9
Topic-Change	DM	DM	0.00	**1.00**	3
	Genre	Newspaper Layout	**0.80**	0.20	5
	Unsure	Unsure	0.00	**1.00**	5
Topic-Comment	DM	DM	0.00	**1.00**	3
	Lexical	Alternative expression	0.00	**1.00**	1
		Indicative word	0.00	**0.00**	1
	Semantic	Lexical chain	0.00	**1.00**	4
	Unsure	Unsure	0.00	**1.00**	15