DiscoExplorer: An Open Interface for the Study of Multilingual Discourse Relations

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Abstract

The relations connecting propositions in discourse such as CAUSE (A because B) or CONCESSION (A although B) are a subject of intense interest in Computational Linguistics and Pragmatics, but challenging to study and compare across languages. Recent progress in standardizing discourse relation inventories across datasets offers the potential to facilitate such studies, but is hindered by the complexity of relevant data and the lack of easily accessible interfaces to analyze it. In this paper we present DiscoExplorer, a new open source web interface, capable of running on local computers, which we use to make datasets from the DISRPT Shared Task on discourse relation classification publicly available, covering 16 different languages. We present the query language, search and visualization facilities for relations and signaling devices such as connectives, as well as some example studies.

1 Introduction

Discourse relations are the implicit and explicit semantic/pragmatic connections that arise when multiple propositions are juxtaposed in a text or conversation. For example, in (1), the explicit connective 'when' indicates a TEMPORAL relation between the two arguments 1 and 2, while the CAUSAL relation between 1 and 3 is understood implicitly (Jin is upset *because* Kim left).

(1) [*Kim left*]$_1$ [*when Jin arrived.*]$_2$ [*Jin is upset now.*]$_3$

A variety of theories have attempted to describe discourse relations and construct datasets for their study, including Rhetorical Structure Theory (RST, Mann and Thompson 1988), Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003), the Penn Discourse Treebank (PDTB, Prasad et al. 2014) and discourse dependencies (Morey et al., 2018). However because each theory and dataset has tended to use distinct relation inventories and data structures (for example hierarchical trees, graphs, or pairs of text spans), comparisons across languages or even datasets in the same language have been challenging.

More recently, the DISRPT shared task (Braud et al., 2024) has made progress in unifying data from such formalisms by focusing on what they have in common: the postulation of relations between parts of a text, and optional inclusion of information about signaling devices, for example the distinction between implicit and explicit relations above. In its most recent edition the shared task also unified relation labels across 38 datasets in 16 different languages (Braud et al., 2025), facilitating cross-linguistic comparisons for the first time, similarly to initiatives to consolidate labels for describing multilingual syntactic functions in projects such as Universal Dependencies (UD, de Marneffe et al. 2021). However what has been missing compared to projects like UD is an easily accessible interface to search and compare data, identify errors, and visualize patterns in datasets. The main contributions of this short paper aim to fill this gap:

- We provide a high performance, open source, client-side interface in pure JavaScript that can be run on any PC
- We make the datasets from the DISRPT shared task searchable online for the public
- We propose a simple, flexible query language to facilitate access for new users

2 Related work

While many local graph search tools exist for linguistic data, such as Semgrex, Ssurgeon or Semgrex-Plus (Tamburini, 2017; Bauer et al., 2023), almost all are limited to searching within sentence boundaries, and are therefore not capable of representing relations across entire texts. Several

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 12–23

online interfaces have facilitated search in syntactically and even semantically annotated treebanks (Guibon et al., 2020; Amblard et al., 2022), but dedicated interfaces for discourse relations are rare, and have generally been fitted to a single resource and theory, such as the Spanish (da Cunha et al., 2011) and Basque (Iruskieta et al., 2013) RST treebank interfaces. Converters for RST data exist to enable searching through data using ANNIS (Krause and Zeldes, 2016), a generic multilayer corpus search interface. However the system is considerably heavier, slower and has a complex query language which is not tailored to discourse relations, and currently cannot import data from other discourse formalisms or the DISRPT format.

Our work takes its primary inspiration from the Grew Match search interface for UD treebanks (Guibon et al., 2020), which leverages the consistent format and label inventory of the UD project to allow access to treebanks using a consistent query language and architecture.

3 DiscoExplorer

3.1 Architecture

Our architecture is designed with three goals in mind: 1. minimizing compute costs to prevent needing a dedicated (and expensive) server; 2. making it possible to run the interface locally for users with proprietary data that cannot be exposed online; and 3. running a fast and responsive search with minimal dependencies. To achieve these goals, we implemented a client-side solution in JavaScript using React, without a database backend, no dedicated indexing (e.g. Meilisearch) or visualization libraries (e.g. D3.js). Instead, we focus on using pure JavaScript, HTML and CSS wherever possible to ensure stability and longevity of the software.

Our data model focuses on discourse relations as the instance to be searched over, where relations are aligned to token positions in documents and span over two possibly discontinuous argument spans (e.g. the cause and effect for CAUSAL relations). Relations that do not cover entire sentences are also associated with context spans indicating words before, after or between the arguments within the same sentences, ensuring that full sentence context is provided with each match. Finally, relations carry labels, a direction (1>2 or 1<2) and possibly a list of typed and subtyped signal tokens, for datasets marking connectives or other signal types.

3.2 Basic interface

The web interface is arranged around two areas: the query form at the top of Figure 1, and the results area at the bottom, which can display concordances for qualitative searches, or switch to a 'frequencies' tab for quantitative analysis. The interface was initially tested with students in a seminar on computational models of discourse at Georgetown University (LING-8415), and based on student feedback, an additional tab was added to perform comparisons between datasets. We are also planning to collect feedback from CODI attendees and the DISRPT community to develop additional features.

The basic query form is meant to be user friendly by exposing the available datasets and labels in each dataset as drop down filters. Negation of a filter is realized as a simple checkbox – for example, selecting the label CONDITION and the negation box for any signal type in `eng.erst.gum` (data from the GUM corpus, Zeldes 2017) yields examples of implicit conditionals, as in (2). More complex queries must utilize the DiscoExplorer query language (DEQL) described in 3.3.

(2) [*you take this painting,*]$_1$ [*I want that recorder*]$_2$ (=if you take this painting)

3.3 Query language – DEQL

Our query language aims to be simple but powerful, meaning on the one hand, it should respond as expected to simply typing words in the search box, while on the other hand allowing users to do exact sequence or flexible match queries, queries restricted to the first/second or source/target spans of the relation, as well as leveraging token annotations. Since DISRPT data is released with accompanying UD annotations, we expose the UD POS tags, dependency labels and lemmas directly for querying. All queries can be restricted by the UI to a specific relation label chosen from a drop down list (either the universal DISRPT label, or each dataset's original labels, or both), specific signal types or subtypes if available in the data (e.g. explicit connectives), and relation directions. The exact query can be saved and reproduced via a shareable link.

As an example of simple text based queries and their interactions with argument spans, consider the differences between the following, all executed with the 'exact sequence' match turned off and the CONDITION label selected:[1]

[1]UI filters are indicated in red and are not part of the query string, but are stored in reproducible shareable query links.

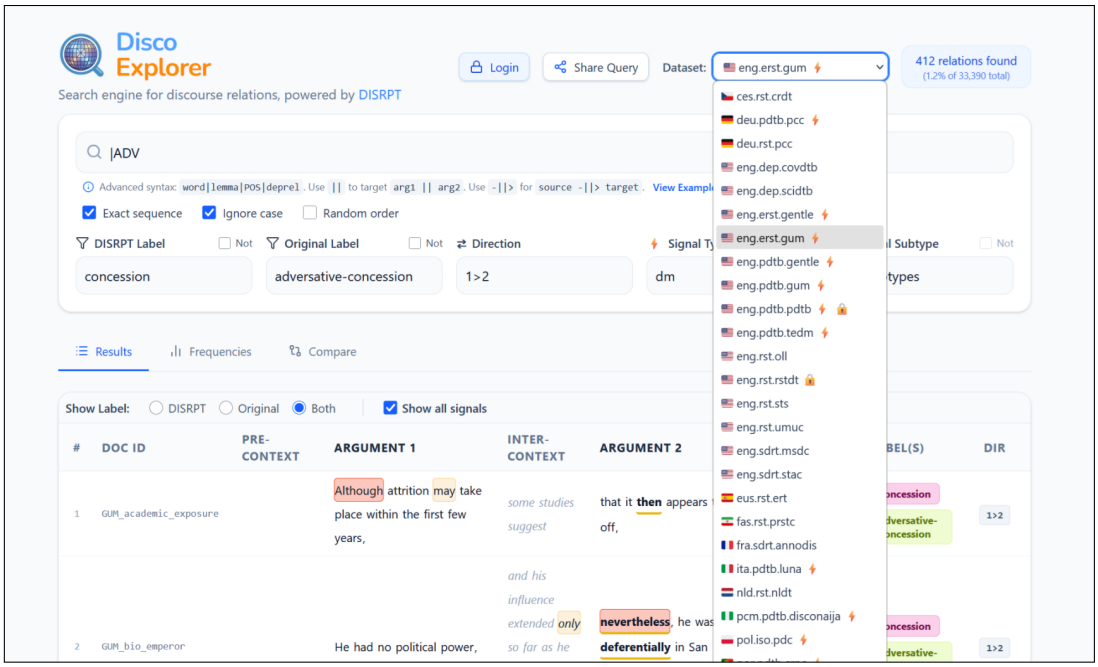


Figure 1: DiscoExplorer search interface: Users can input a query and select filters. Underlines show query matches and signals are highlighted (e.g. red for discourse markers, yellow for lexical signals).

(3) CONDITION `if then` (finds CONDITION relations with 'if' and 'then' anywhere)

(4) CONDITION `if || then` (same, but ensures 'if' and 'then' are in arg1 and arg2)

(5) CONDITION `if -||> then` (same, but 'if' must be in the relation source and 'then' in the target, regardless of text order)

While in (3) we only guarantee that 'if' and 'then' appear somewhere, in (4) we require that they appear in that text order, one in each argument. By contrast, (5) requires that 'if' appears in the source of the relation (the protasis) and 'then' in the target (the apodosis), regardless of their text order.

More experienced users who are familiar with UD annotations may also want to use token annotations to restrict queries. To enable this we use the format `word|lemma|pos|deprel`, where each of these elements may be lacking. If less than three annotations are specified, the system uses the search values to implicitly identify the key, since POS and deprel have closed vocabularies. Thus the following searches find:

(6) PURPOSE EXACT `to|PART |VERB|advcl -||>` (PURPOSE relation with a to-infinitive)

(7) TEMPORAL EXACT `when |ADJ|advcl -||>` (TEMPORAL with 'when' followed by a reduced adjectival adverbial clause)

The example in (6) will find VERBs heading an adverbial clause (UD *advcl*) immediately preceded by the word 'to' tagged as PART. The interface automatically detects that VERB is a POS tag value and *advcl* is a dependency relation. In (7) we find reduced temporal clauses of the type 'when possible', since the word 'when' must be followed immediately by an adjective heading an adverbial clause. The final operator `-||>` ensures that both searches only consider the source span of the relation, regardless of text order.

3.4 Frequencies interface

The frequencies tab gives raw counts, percentages and plots of a category or numerical variable selected by the user from the Breakdown drop down (see Figure 2). Categorical variables include DISRPT labels, original labels, relation direction, signal type/subtype and any available metadata (for example genre, if known). If filters are selected for any of these in the query, a binary yes/no breakdown of the selected feature is also available. Updating the query instantly updates matches, numbers and plots, and raw results are also downloadable as a .tsv file.

A second drop down called 'Cross-tabulate' allows users to select a second dimension from the same options and generate a cross table, coupled with a chi-squared residual plot indicating combinations that appear more or less than expected, as

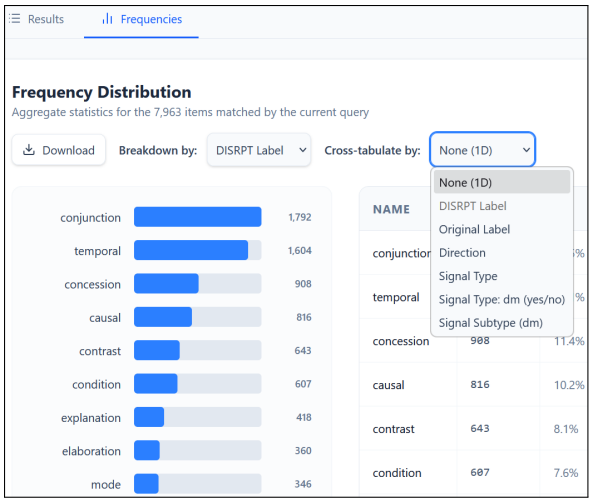


Figure 2: Frequency breakdown of DISRPT labels.

well as displaying significance codes. For example, Figure 3 shows an association plot of explicit connective signals vs. DISRPT label in the English PDTB corpus, showing that while CONCESSION, CONDITION and CONJUNCTION are mostly explicit, CAUSAL relations are more often implicit, while CONTRAST relations are more balanced.

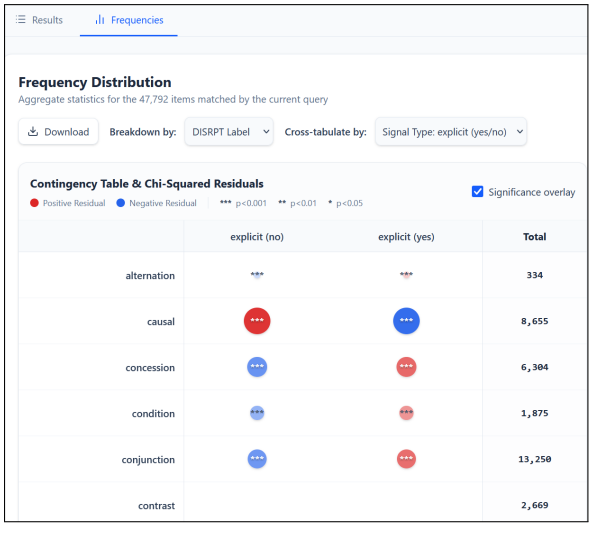


Figure 3: Association of explicitness vs. label in PDTB.

If a numerical variable is chosen for breakdown, the interface will plot a boxplot (for a single variable) or scatterplot (two cross-tabulated numerical variables) or multiple boxplots (numerical cross-tabulated with categorical). Available numerical variables are currently argument length in tokens and percentile position in document (for argument 1 or 2 in text order), the same for the source or target argument (regardless of text order), distance in tokens between arguments, and the number of signals for the relation (if available).

3.5 Comparison interface

Based on student feedback, comparing datasets is a desirable capability, and we implement this in a similar way to cross-tabulation, where, instead of using a categorical variable, we use dataset identity. However, since each dataset has its own distribution for each variable, we display results for each value side-by-side, with the primary selected dataset in blue and the comparison in orange with pairwise plots, as shown for a categorical variable (label type) with barplots in Figure 4 for a comparison between the eRST GUM corpus and the eRST GENTLE corpus (Genre Tests for Linguistic Evaluation, Aoyama et al. 2023), which follows the same annotation scheme but includes 8 challenging genres such as medical texts, poetry and even course syllabuses.

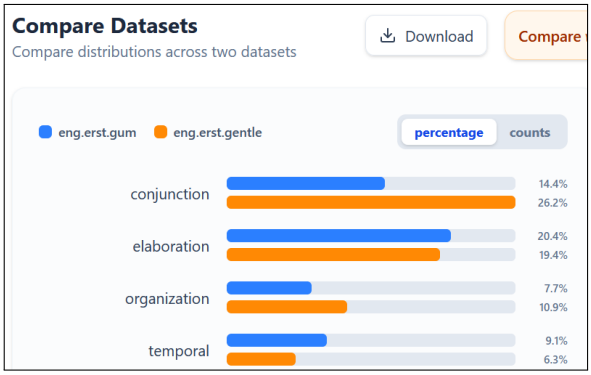


Figure 4: Relation labels in GUM vs. GENTLE.

The figure shows that CONJUNCTION is more common in GENTLE (in orange), which is primarily due to genres containing many lists, such as medical notes and syllabuses. The ELABORATION label, but contrast, is very similar in prevalence.

As with frequencies, numerical variables receive side-by-side boxplots. Figure 5 shows the number of signals per relation, this time filtered to show just MODE relations (manner and means). These have significantly fewer signals in GENTLE, largely owing to data from the poetry and medical genres.

4 Evaluation

Data We import the 38/39 DISRPT 2025 datasets which contain discourse relations (the remaining dataset contains only discourse unit segmentation information, without labels). The datasets come from five different frameworks: RST, PDTB, SDRT, eRST (Zeldes et al., 2025) and discourse dependencies. In total, these cover over 300,000 relations across 5 million tokens in almost 10,000

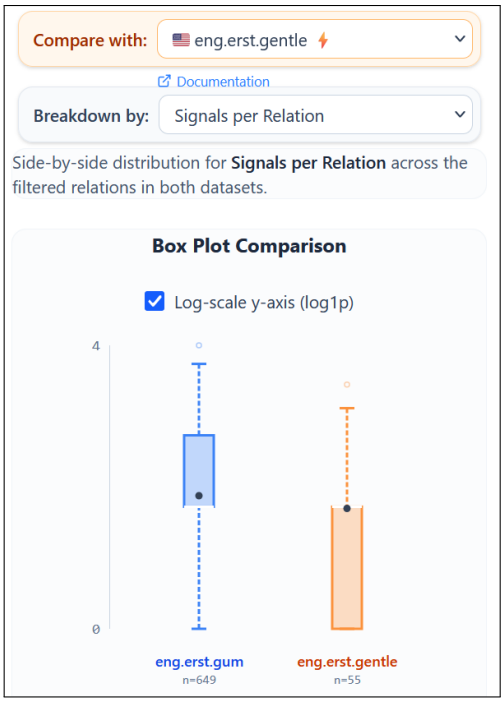


Figure 5: Signals per MODE relation compared.

documents (see Table 2 in Appendix A for full details). The largest dataset is English PDTB (Prasad et al., 2014) with over 47K relations and 1.1M tokens, as well as five signal types (explicit or implicit connectives, alternative lexicalizations and constructions, and a special hypohphora type for questions). Datasets in the eRST framework also distinguish signal subtypes, in a taxonomy of 8 major types and over 40 subtypes.

Performance While it is difficult to benchmark our system due to a lack of directly comparable alternatives, we conduct a simple timing experiment using the GUM data in comparison to its publicly available version in ANNIS. We note that ANNIS can perform much more elaborate searches than DiscoExplorer, such as dependency graph queries between tokens (e.g. checking that a token is the subject of a specific verb), as well as querying other annotation layers, such as entity annotations; here we limit comparison to simple searches for tokens and discourse relations on a consumer laptop. Since our data is loaded into main memory, query response times are close to instantaneous, with the only added latency of loading the dataset once, which ANNIS does not have (Table 1).

Query type	DEQL	DiscoExp.	ANNIS	Hits
(load)	–	2.820s	–	–
tok	think	0.022s	3.98s	291
tok+pos+deprel	think|VERB|advcl	0.027s	4.68s	17
rel+tok+pos	CONJUNCTION think|VERB	0.030s	3.01s	12
neg-rel+tok+pos	NOT CONJUNCTION think|VERB	0.028s	4.41s	410
rel	ELABORATION	0.003s	3.45s	6812

Table 1: Query latency compared with ANNIS.

5 Discussion and Conclusion

This short paper presented DiscoExplorer, a new browser based search interface for multilingual discourse relation datasets based on the DISRPT benchmark. The interface offers a simple user-friendly way to search for discourse relations and examine their distributions using filters, as well as a more complex query language to restrict matches by tokens and their annotations. Quantitative results can be tabulated, plotted and downloaded.

A comparison of query run times with ANNIS showed that although the interface requires an initial load time for each dataset, the approach using main memory search in JavaScript is very fast. While some of the complex searches a system like ANNIS would allow are not supported, tailoring the interface to the DISRPT data model, centered around discourse relation instances, allows for a simple query language and data structure, and also means we do not need a backend or any compute resources to offer the system to the public.

A further advantage of relying on the DISRPT data model is the abundance of data already available in the shared task format (currently 38 datasets), and the likely release of further public data in the format of the shared task, which has been running for four iterations as of 2026.

As a result of submissions to the task, and especially the transition to multilingual models trained on all datasets, it is also increasingly possible to generate predicted datasets following the DISRPT label scheme in a variety of languages. The current best performing system for predicting relation labels, DeDisCo (Ju et al., 2025), achieves 76.13% accuracy on the DISRPT test set across languages, and 75.55% on Chinese RST (Peng et al., 2022), 79.17% on Portuguese (Mendes and Lejeune, 2022), or 71.39% on the PDTB framework Georgetown Discourse Treebank (GDTB, Liu et al. 2024), suggesting that automatically tagged corpora that are useful for research on discourse relations across languages may not be far and could easily be made searchable using this system.

We release the code and place the system online for public use together with this paper, available at https://gucorpling.org/discoexplorer.

Limitations

This paper only presents and evaluates a search interface in terms of responsiveness and does not conduct a user study, though we hope to gather some feedback on the system from participants at the CODI workshop and the community using DISRPT data. The data used by the system is provided by DISRPT as-is, and we make no claims regarding the accuracy of particular annotations within those datasets. AI was not used in any way to write this paper, though AI coding assistants were used in the creation and debugging of the system itself.

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).

Maxime Amblard, Bruno Guillaume, Siyana Pavlova, and Guy Perrier. 2022. Graph querying for semantic annotations. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 95–101, Marseille, France. European Language Resources Association.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

John Bauer, Chloé Kiddon, Eric Yeh, Alex Shan, and Christopher D. Manning. 2023. Semgrex and ssurgeon, searching and manipulating dependency graphs. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 67–73, Washington, D.C. Association for Computational Linguistics.

Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proceedings of the 12th Web as Corpus Workshop*, pages 50–56, Marseille, France. European Language Resources Association.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 1–20, Suzhou, China. Association for Computational Linguistics.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

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 C. F. Cardoso, Erick G. Maziero, Maria Lucía R. Castro Jorge, Eloize M. R. Seno, Ariani Di Felippo, Lucia H. M. Rino, Maria das Graças V. Nunes, and Thiago A. S. Pardo. 2011. CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Anais do III Workshop "A RST e os Estudos do Texto"*, pages 88–105, Cuiabá, MT, Brasil. Sociedade Brasileira de Computação.

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, Gerardo Sierra, Luis Adrián Cabrera-Diego, Brenda Gabriela Castro Rolón, and Juan Miguel Rolland Bartilotti. 2011. The RST spanish treebank on-line interface. In *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pages 698–703. RANLP 2011 Organising Committee.

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

Luke Gessler, Yang Liu, and Amir Zeldes. 2019. A discourse signal annotation system for RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 56–61, Minneapolis, MN. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2012. The RST Basque TreeBank.

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 *Anais do IV Workshop "A RST e os Estudos do Texto"*, pages 40–49, Fortaleza, CE, Brasil. Sociedade Brasileira de Computação. 21–23 October 2013.

Zhuoxuan Ju, Jingni Wu, Abhishek Purushothama, and Amir Zeldes. 2025. DeDisCo at the DISRPT 2025 shared task: A system for discourse relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 48–62, Suzhou, China. Association for Computational Linguistics.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024. GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

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

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. 2024. Polish discourse corpus (PDC): Corpus design, ISO-compliant annotation, data highlights, and parser development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12829–12835, Torino, Italia. ELRA and ICCL.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Towards building a discourse-annotated corpus of russian. In *Proceedings of the 23rd International Conference on Computational Linguistics and Intellectual Technologies "Dialogue-2017"*, Moscow, Russia.

Andrew Potter. 2008. Interactional coherence in asynchronous learning networks: A rhetorical approach. *The Internet and Higher Education*, 11(2):87–97.

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.

Ponrawee Prasertsom, Apiwat Jaroonpol, and Attapol T. Rutherford. 2024. The Thai Discourse Treebank: Annotating and classifying Thai discourse connectives. *Transactions of the Association for Computational Linguistics*, 12:613–629.

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).

Merel CJ Scholman, Marian Marchal, AriaRay Brown, and Vera Demberg. 2025. DiscoNaija: A discourse-annotated parallel Nigerian Pidgin-English corpus. *Language Resources and Evaluation*, pages 1–37.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian rhetorical structure theory. *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).

Fabio Tamburini. 2017. Semgrex-plus: a tool for automatic dependency-graph rewriting. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 248–254, Pisa, Italy. Linköping University Electronic Press.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024. Discourse structure for the Minecraft corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4957–4967, Torino, Italia. ELRA and ICCL.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, Aravind Joshi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. 2010. LUNA Corpus Discourse Data Set.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Cheng Yi, Li Sujian, and Li Yueyuan. 2021. Unifying discourse resources with dependency framework. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1058–1065, Huhhot, China. Chinese Information Processing Society of China.

Karolina Zaczynska and Manfred Stede. 2024. Rhetorical strategies in the UN security council: Rhetorical Structure Theory and conflicts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 15–28, Kyoto, Japan. Association for Computational Linguistics.

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

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

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

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. Ted multilingual discourse bank (tedmdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, 54(2):587–613.

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

A Dataset details

Table 2 gives an overview of the datasets that are currently searchable using the system. Datasets marked by an asterisk (∗) require LDC licenses; annotations for this data and a small subset of `eng.erst.gum` coming from Reddit (Behzad and Zeldes, 2020) can be obtained from the DISRPT shared task repository, along with scripts to reconstruct the underlying text.

Additional datasets can be added to the system as long as they conform to the DISRPT shared task format, meaning that relations are serialized in the `.rels` format and token annotations are available in a corresponding `.conllu` file. In particular, the DISRPT format assumes that relations apply between flat spans of text, meaning that hierarchical information as found in formalisms such as RST is lost. Instead, relations are interpreted as a dependency conversion of constituent structures, as illustrated in Figure 6.

The figure shows the system representation of a single relation, in this case a CONCESSION between two head units:

(8) a. [*this is a terrific opportunity*]
b. [***but*** *we will have to wait until after the event*]

These units are only parts of the sentences they come from, as shown in the eRST graph fragment. Modifiers of those units which appear in the same sentences are represented in DiscoExplorer as pre-, inter- and post-context, depending on whether they appear before the first argument, between the two arguments, or after the second.

Note also that while the eRST graph on the left highlights multiple signals, only the red discourse marker "but" is attached to the CONCESSION relation, and that same signal is visualized in the DiscoExplorer search results. The cyan highlighted syntactic signals for PURPOSE relations ('opportunity .. to improve' and 'wait .. to assess') belong to those respective relations, and would be highlighted in queries actually retrieving the associated relation, rather than being highlighted in a query retrieving a different relation that happens to overlap the same text.

Corpus	Language	Framework	Labels	Relations	Sentences	Tokens	Documents	Signals
ces.rst.crdt	Czech	RST	17	1,249	835	14,664	54	–
deu.pdtb.pcc	German	PDTB	11	2,109	2,193	33,222	176	types
deu.rst.pcc	German	RST	16	2,882	1,944	32,836	176	–
eng.dep.covdtb	English	dependencies	11	4,985	2,343	60,907	300	–
eng.dep.scidtb	English	dependencies	14	9,903	4,202	102,534	798	–
eng.erst.gentle	English	eRST	17	2,552	1,334	17,979	26	subtypes
eng.erst.gum	English	eRST	17	30,747	14,158	254,890	255	subtypes
eng.pdtb.gentle	English	PDTB	12	786	1,334	17,979	26	types
eng.pdtb.gum	English	PDTB	13	13,879	14,158	254,890	255	types
∗eng.pdtb.pdtb	English	PDTB	13	47,792	48,630	1,173,379	2,162	types
eng.pdtb.tedm	English	PDTB	13	529	381	8,185	6	types
eng.rst.oll	English	RST	17	2,751	2,156	46,471	327	–
∗eng.rst.rstdt	English	RST	17	19,778	8,318	208,912	385	–
eng.rst.sts	English	RST	17	3,058	2,591	71,206	150	–
eng.rst.umuc	English	RST	15	4,997	2,424	61,590	87	–
eng.sdrt.msdc	English	SDRT	10	27,848	14,744	231,352	440	–
eng.sdrt.stac	English	SDRT	11	12,271	7,394	52,271	1,101	–
eus.rst.ert	Basque	RST	16	3,632	2,380	45,780	164	–
fas.rst.prstc	Farsi	RST	14	5,191	2,179	66,926	150	–
fra.sdrt.annodis	French	SDRT	12	3,321	1,507	32,699	86	–
ita.pdtb.luna	Italian	PDTB	11	1,525	3,750	25,242	60	types
nld.rst.nldt	Dutch	RST	16	2,264	1,651	24,898	80	–
pcm.pdtb.disconaija	Naija	PDTB	13	9,903	9,242	140,729	176	types
pol.iso.pdc	Polish	ISO	12	8,543	9,142	156,980	556	types
por.pdtb.crpc	Portuguese	PDTB	12	11,327	5,194	186,849	302	types
por.pdtb.tedm	Portuguese	PDTB	13	554	394	8,190	6	types
por.rst.cstn	Portuguese	RST	15	4,993	2,221	63,332	140	–
rus.rst.rrt	Russian	RST	15	25,095	13,131	262,495	234	–
spa.rst.rststb	Spanish	RST	16	3,049	2,089	58,717	267	–
spa.rst.sctb	Spanish	RST	16	692	516	16,515	50	–
tha.pdtb.tdtb	Thai	PDTB	12	10,861	6,534	256,523	180	–
∗tur.pdtb.tdb	Turkish	PDTB	13	3,176	31,197	496,358	197	–
tur.pdtb.tedm	Turkish	PDTB	13	574	410	6,286	6	types
zho.dep.scidtb	Mandarin	dependencies	14	1,297	500	18,761	109	–
zho.pdtb.cdtb	Mandarin	PDTB	9	5,270	2,891	73,314	164	–
zho.pdtb.ted	Mandarin	PDTB	15	13,308	8,671	181,910	72	types
zho.rst.gcdt	Mandarin	RST	17	8,413	2,692	62,905	50	–
zho.rst.sctb	Mandarin	RST	17	692	580	15,496	50	–
Total	16	6	17	311,796	257,705	5,139,564	9,890	14 datasets

Table 2: DISRPT 2025 datasets searchable in DiscoExplorer (∗ marks datasets requiring an LDC license).

B Additional use cases

In addition to exploring relation labels, DiscoExplorer can be used to study the distribution of relation signals (in datasets with signal annotations, indicated in the dataset list by a lightning bolt icon in Figure 1), relation directions and signal subtypes. Figure 8a shows a breakdown of relation

Figure 7 shows the disparity of right-to-left versus left-to-right temporal relations signaled by 'when' in English, where source of the relation tends to precede the target, but not always – the proportion is about 7:3 in favor of placing the TEMPORAL clause first.

Figure 8 shows two panels demonstrating the breakdown of signaling devices. Panel 8a shows major signal types signaling three relations classes: ALTERNATION relations (such as 'A or B') are typically signaled by a discourse marker (dm) such as '(either) or', '(or) else', 'alternatively' etc.; ATTRIBUTION relations, indicating the source of information, are typically signaled semantically by the presence of a speaker noun phrase, or lexically through a speech verb, but very rarely using a dm (for example '**As** Smith had it, ...') or with no signals (implicitly). By contrast CAUSAL relations usually appear with a dm, and appear with no explicit signaling more often than the other relations (column 'None').

Finally, Panel 8b shows discourse marker subtypes used to signal three original labels for EXPLANATION relations in the data, using a cutoff to show

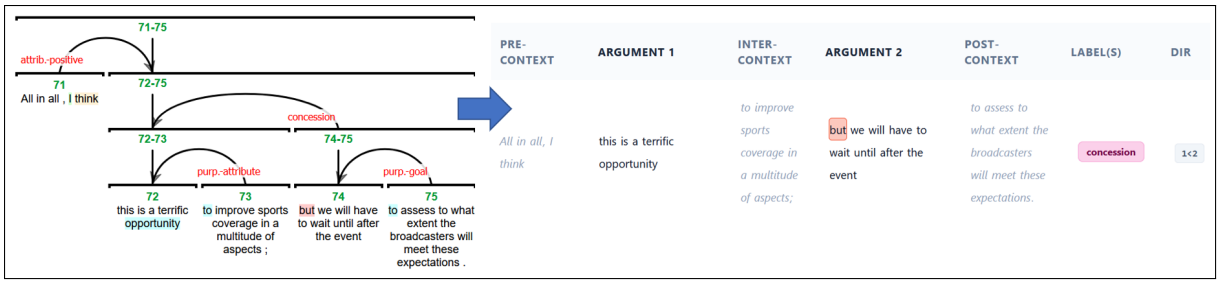


Figure 6: Original eRST graph fragment for a CONCESSION relation, visualized using rstWeb (Gessler et al., 2019) and the corresponding output in DiscoExplorer.

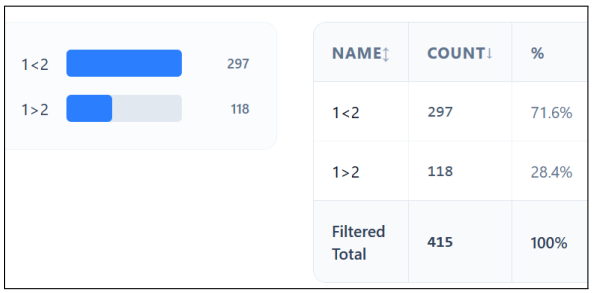


Figure 7: Frequencies of TEMPORAL 'when' clauses in left-to-right vs. right-to-left directions

only items appearing 20 or more times. Although the DISRPT relation labels do not include subtypes, we can get breakdowns for relation subtypes if the original labels of the underlying dataset include them, which is the case here. The biggest disparity is the preference of EXPLANATION-EVIDENCE relations to be marked by 'for example' and 'as'. By contrast, EXPLANATION-JUSTIFY favors the use of 'and', as in (9).

(9) [*The record is replete with case law that says exactly that,*]$_1$ [***and** I'm not here to dispute that today.*]$_2$

Meanwhile EXPLANATION-MOTIVATION relations appear disproportionally often with 'so', attempting to convince someone to do something using a supporting argument, as in (10).

(10) [*Good jokes have a lot of details and personality,*]$_1$ [***so** don't be afraid to embellish.*]$_2$

However using the interface, it is easy to find examples of 'so' as a discourse marker with any of the three labels.

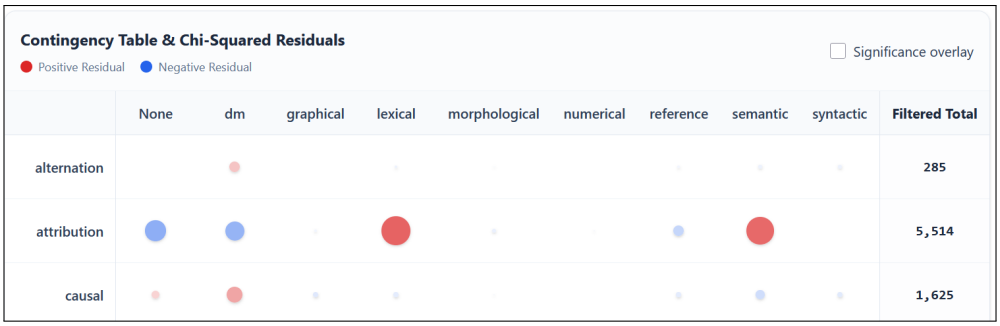
C Resource consumption

While the interface runs very quickly and requires no server-side compute resources to run, an anonymous reviewer has inquired about the memory footprint of loading a large dataset. This is difficult to quantify exactly, since we cannot trivially access browser memory management internals, but to get a rough benchmark, we ran Chrome on a Windows 11 64 bit machine and compared memory usage for the browser with different datasets loaded, as shown in Table 3.

scenario	tokens	relations	memory	Δ idle
Chrome idle	0	0	305.9 MB	0
`eng.erst.gum` *loaded*	273,257	33,390	709.2 MB	+403.3
`eng.pdtb.pdtb` *loaded*	1,173,379	47,792	750.1 MB	+444.2
compare	–	81,182	770.4 MB	+464.5

Table 3: Memory usage in several scenarios using Chrome.

The dataset 'loaded' state by default means that a query runs to retrieve all relations, since no filter has been selected. As the table shows, loading a fairly large and richly annotated corpus such as `eng.erst.gum` requires about 400 MB of RAM. The largest dataset in tokens and relations, `eng.pdtb.pdtb`, takes only slightly more, at 444 MB – we suspect the small difference is due to the amount of space taken up by the signal annotations in the former dataset, which the latter lacks. Using the comparison function on the two datasets does not add much more memory (∼465 MB total). We suspect that this is because a new full result set is not actually loaded into memory - only the information being compared, by default the relation label statistics, is added to the main memory, while indexing specific matching tokens or sentence spans is unnecessary for the comparison data, since no detailed search results are shown for the second dataset.



(a) Residual plot for DISRPT labels versus major signal types.

(b) Residual plot for discourse markers appearing 20+ times versus EXPLANATION relation labels

Figure 8: Contingency tables for signal types and subtypes in `eng.erst.gum`.

D Corpus resources

The system described in this paper would be useless without the datasets it makes searchable. In addition to the datasets and papers cited above, we would like to acknowledge the projects that have produced the remaining datasets in Table 2, all of which can be searched in our publicly available instance of DiscoExplorer and are available under their original licenses from the DISRPT shared task:

- `deu.rst.pcc` and `deu.pdtb.pcc` – the Potsdam Commentary Corpus, Stede and Neumann (2014)
- `eng.dep.covdtb` – the COVID-29 Discourse Treebank, Nishida and Matsumoto (2022)
- `eng.dep.scidtb` – SciDTB, Yang and Li (2018)
- `eng.pdtb.tedm`, `por.pdtb.tedm` and `tur.pdtb.tedm` – TED Multilingual Discourse Bank, Zeyrek et al. (2020)
- `end.rst.oll` and `end.rst.sts` – RST Online-Learning and Science, Technology, and Society corpora, Potter (2008)
- `end.rst.rstdt` – RST Discourse Treebank, Carlson et al. (2003)
- `end.rst.umuc` – University of Potsdam Multilayer UNSC Corpus, Zaczynska and Stede (2024)
- `eng.sdrt.msdc` – Minecraft Structured Dialogue Corpus, Thompson et al. (2024)
- `eng.sdrt.stac` – Strategic Conversation Corpus, Asher et al. (2016)
- `eus.rst.ert` – Basque RST Treebank, Iruskieta et al. (2012)
- `fas.rst.prstc` – Persian RST Corpus, Shahmohammadi et al. (2021)
- `fra.sdrt.annodis` – the ANNODIS corpus, Afantenos et al. (2012)
- `ita.pdtb.luna` – LUNA corpus, Tonelli et al. (2010)
- `nld.rst.nldt` – Dutch DTB, Redeker et al. (2012)
- `pcm.pdtb.disconaija` – DiscoNaija corpus, Scholman et al. (2025)

- `pol.iso.pdc` – Polish Discourse Corpus, Ogrodniczuk et al. (2024)
- `por.rst.cstn` – CST News Corpus, (Cardoso et al., 2011)
- `rus.rst.rrt` – Russian RST Treebank, Pisarevskaya et al. (2017)
- `tha.pdtb.tdtb` – Thai Discourse Treebank, Prasertsom et al. (2024)
- `tur.pdtb.tdb` – Turkish Discourse Bank, Zeyrek and Kurfalı (2017)
- `spa.rst.sctb` and `zho.rst.sctb` – the RST Spanish-Chinese Treebank, Cao et al. (2018)
- `zho.pdtb.cdtb` – Chinese Discourse Treebank, Zhou et al. (2014)
- `zho.dep.scidtb` – Chinese SciDTB, Yi et al. (2021)