eRST: A Signaled Graph Theory of Discourse Relations and Organization

Amir Zeldes[1], Tatsuya Aoyama[1], Yang Janet Liu[2], Siyao Peng[2*], Debopam Das[3], and Luke Gessler[4**]

[1]Georgetown University, Department of Linguistics
amir.zeldes@georgetown.edu, ta571@georgetown.edu
[2]MaiNLP, Center for Information and Language Processing, LMU Munich
y.liu1@lmu.de, siyaopeng@cis.lmu.de
[3]Åbo Akademi University, Department of English Language and Literature
Debopam.Das@abo.fi
[4]Indiana University, Department of Linguistics
lgessler@iu.edu

In this article we present Enhanced Rhetorical Structure Theory (eRST), a new theoretical framework for computational discourse analysis, based on an expansion of Rhetorical Structure Theory (RST). The framework encompasses discourse relation graphs with tree-breaking, non-projective and concurrent relations, as well as implicit and explicit signals which give explainable rationales to our analyses. We survey shortcomings of RST and other existing frameworks, such as Segmented Discourse Representation Theory, the Penn Discourse Treebank, and Discourse Dependencies, and address these using constructs in the proposed theory. We provide annotation, search, and visualization tools for data, and present and evaluate a freely available corpus of English annotated according to our framework, encompassing 12 spoken and written genres with over 200K tokens. Finally, we discuss automatic parsing, evaluation metrics, and applications for data in our framework.

1. Introduction

Natural language documents are more than just an ordered list of equally important and self-contained sentences: They form complex structures that can often be divided into more or less prominent sections and subsections, which together give rise to meanings that are not necessarily localizable to individual propositions by themselves. Identifying these structures and the meanings associated with them is the task of **discourse parsing**, in which arbitrary documents are assigned an analysis within a theoretical parsing

* Work done partly at Georgetown University.
** Work done partly at Georgetown University and University of Colorado Boulder.

Action Editor: Min Zhang. Submission received: 22 November 2023; revised version received: 18 March 2024; accepted for publication: 19 July 2024.

https://doi.org/10.1162/coli_a_00538

framework that defines the types of combinatory semantic and pragmatic meanings to be recognized, and the structures that components of a document can create.

While not discussed in the field of NLP as often as syntactic parsing or entity recognition, discourse parsing has been one of the "textbook" examples of Natural Language Understanding (Jurafsky and Martin 2024, pages 536–540) for a long time, with implementable frameworks being suggested as early as Mann and Thompson's (1988) Rhetorical Structure Theory (RST). Most approaches to discourse parsing involve (at least) recognizing spans of text that are connected by one of a set of predetermined discourse relation types (Hovy 1990), such as CAUSE or CONCESSION, and naming the relation and configuration in which those parts appear in the text, which can take on many linguistic forms. For instance, in example (1) from Asher and Lascarides (2003, page 136), both formulations in a. and b. are typically interpreted to mean that the predicate *pushed* is the cause of the predicate *fell*, and that the pushing preceded the falling in time, although these events are related in chronological order in b., but in counter-chronological order in a.

(1) a. Max fell. John pushed him.

b. John pushed Max. He fell.

The exact nature and inventory of such relations, sometimes called "coherence relations," "prominence relations," or also "rhetorical relations," as well as the structures they form, vary across theoretical accounts.

Like other areas of NLP, discourse parsing has benefited from increasingly accurate scores following the introduction of large pre-trained language models, with scores approaching human performance on some subtasks, such as discourse unit segmentation (Gessler et al. 2021), recognition of explicitly signaled relations (Knaebel 2021), as well as hierarchical parsing, especially for English in the news domain (Guz, Huber, and Carenini 2020; Liu, Shi, and Chen 2021; Kobayashi et al. 2022).

By contrast, less progress has been made in advancing our theories of discourse relations and their organization. After the introduction of RST and subsequent projects to construct datasets using the theory (Carlson, Marcu, and Okurowski 2001), several alternative frameworks were proposed to address some of its shortcomings (surveyed below in Section 2.1), with the main strands resulting in implemented datasets including Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003, Section 2.2), the Penn Discourse Treebank framework (PDTB, Prasad et al. 2006, Section 2.3), and the Cognitive Approach to Coherence Relations (CCR, Sanders, Spooren, and Noordman 1992, Section 2.4). These frameworks each improve on certain problems identified quite early on in RST, including notably:

- Tree-breaking, non-projective structures (SDRT)
- Distinguishing implicitly and explicitly signaled relations, with the latter being more reliably identifiable (PDTB)
- Support for multiple concurrent relations (mainly SDRT, but to some extent all of the above)
- Identification of relation hierarchies or subtypes based on formal markers (PDTB) or feature structures (CCR)
- Explicit support for nested relations (SDRT)

Although there has been substantial work in each framework, including refinements to guidelines or covered phenomena, and development of new annotated resources, little has changed in the landscape of implemented theoretical models of discourse relations since the inception of PDTB almost two decades ago. However, this stability should not be taken as a sign that our theoretical models are now completely satisfactory: Each of the theories mentioned above has shortcomings, such as inability to model hierarchical structure in PDTB, or lack of relative prominence marking in SDRT.

In this article we aim to push the development of discourse representation theories further, by proposing a new formalism that draws on insights from several frameworks in an attempt to keep the most useful parts of the original formulation of discourse parsing as envisioned by Mann and Thompson (1988), while incorporating solutions to problems from over three decades of work in the field. Since our formalism is "backwards compatible" with RST, we designate it *Enhanced Rhetorical Structure Theory* (eRST), in the hopes of drawing researchers already familiar with RST and harnessing existing resources for its development (in this sense it can be viewed as an optional "enhanced" representation, similar to Enhanced Dependencies for more basic Universal Dependencies in syntax, Nivre et al. 2020). At the same time, our framework offers important additional expressive mechanisms that should appeal to researchers engaged with other frameworks, specifically supporting:

- Multiple relations between the same nodes
- Non-projective, tree-breaking structures
- Maintaining RST's recursive prominence hierarchy despite the above
- Marking categorized and subtyped discourse relation signals, including implicit and explicit connectives, as well as alternative lexicalization mechanisms
- Use of a hierarchical relation taxonomy
- Supporting new NLU applications by linking relations to implicated spans of text fulfilling specific relation participant roles

We would like to stress that while the last point is of interest to us, the primary motivation for eRST is not improving performance on any particular NLP tasks compared with RST, but simply to provide a more comprehensive and detailed representation of discourse relations in text across any genre, which can recover relations that are present, but not currently covered by RST analyses, along with the rationale or evidence supporting and sub-categorizing their occurrences.

eRST, with its advanced set of features, can support the inquiry of numerous discourse phenomena. Some general research questions we envisage eRST would help us investigate include: How do discourse relations and their signals are distributed across texts and text types or genres? What correlations exist between relation and signal types? Are there semantic or pragmatic correlates of the amount and type of signaling observed for relation types? When and how often does natural discourse violate strict tree constraints? And to what extent can discourse relation identification be completely motivated by localizable signals? Complementing the theoretical framework proposed in this article, we also release data and tools to support development work, which are

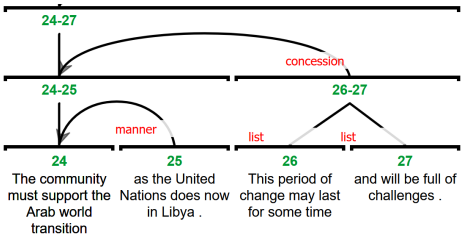


Figure 1
RST Fragment from GUM (Zeldes 2017). The most central point is the nucleus in [24], to which other units are direct or indirect satellites (MANNER and CONCESSION). Symmetrical relations such as LIST are multinuclear nodes ([26]–[27]).

meant to cover a wide range of scenarios and types of text. The main contributions of this article are therefore:

[1] A new framework for discourse relation and discourse structure annotation

[2] Extending a robust freely available annotation tool to create gold standard data

[3] A corpus of over 200K tokens covering 12 spoken and written English text types[1]

[4] A corresponding XML format, annotation, and conversion tools for the freely available search and visualization tool ANNIS (Krause, Leser, and Lüdeling 2016)

[5] A newly defined discourse parsing task including metrics and an official scorer

[6] A baseline system using a contemporary neural architecture and scores

2. Related Work

Mann and Thompson's (1988) formulation of discourse parsing understood relations to hold recursively between adjacent and contiguous spans of text, which covered entire arbitrary documents down to the level of basic propositions known as Elementary Discourse Units (EDUs), thereby forming a hierarchical labeled tree, as in Figure 1. Labeled RST trees are directed and assume a distinction between more prominent "nucleus" units, and less prominent "satellite" units at each level of the tree.

The recursive nature of RST trees was particularly appealing to early research on automatic summarization (Marcu 1997; Teufel and Moens 2002) and dialog planning (Moore and Paris 1993; Taboada and Lavid 2003), since removing satellites and their descendants could be used to obtain extractive summaries of arbitrary passages (Liu

1 Since submission of this paper, the corpus has grown even larger and now covers 16 genres, supplemented by test data in 8 additional genres, for a total of 246K tokens in 24 genres; see Section 4 for more details.

2024) (e.g., [24] is the best extractive summary unit for the entire tree in Figure 1) and a recursive tree could be used to track complex bifurcating topics in a long conversation.

In the years since the proposal of RST, a number of competing frameworks, which will be surveyed below in more detail, have suggested both limiting and expanding the scope of discourse relation identification. For example, according to Sanders, Spooren, and Noordman's (1992, page 2) CCR, relations should be identified by the presence of meanings "of two or more discourse segments that cannot be described in terms of the meaning of the segments in isolation," without necessarily assuming a hierarchy or coverage of the text, and are distinguished using a set of binary attributes (e.g., basic vs. non-basic ordering in Example (1) above, see Hoek, Evers-Vermeul, and Sanders 2019). Asher and Lascarides's SDRT, proposed that segments could participate in multiple relations, addressing early criticism of RST's strict tree constraint (Moore and Pollack 1992), and forming a graph rather than a tree, with elementary units that are also allowed to nest. SDRT distinguishes subordinating and coordinating relations, rather than distinguishing satellites from nuclei, with some consequences for the structures postulated by the theory.

Moving in the opposite direction and more similarly to CCR, the framework of the PDTB (Prasad et al. 2006) proposed to identify relations as projections of explicit or implicit discourse markers called **connectives**, such as the word "because," whose presence (or possible presence when omitted) indicates a causal relation. PDTB analyses are also called shallow discourse parses (Xue et al. 2016), since they do not assume a hierarchical tree or graph structure for documents, but also add more complex facilities by associating each relation with a type of connective, employing a hierarchical label taxonomy, and allowing relations to connect discontinuous/overlapping segments.

Despite progress on new datasets in the frameworks listed above and many refinements to their guidelines, comparatively little progress has been made on discourse relation representation since the publication of PDTB. Because a full survey of the literature on computationally implementable theories of discourse relations and discourse organization is unfeasible in the scope of the current article,[2] we focus here on a synopsis and comparison of the main formalisms used in the field, for which substantial annotated corpus data exists: RST, SDRT, PDTB, CCR, and Discourse Dependencies.

2.1 Rhetorical Structure Theory

RST covers the most languages and datasets for discourse relations (12/26 datasets and 9/13 languages in the recent cross-formalism DISRPT shared task came from RST data [Braud et al. 2023, 2024]). The theory distinguishes itself from other frameworks in its strong assumption of a tree constraint on all graphs, which must cover the entire text of a document, and the distinction of satellite vs. nucleus nodes (cf. Figure 1).

The first large scale implementation of RST was the RST Discourse Treebank, annotating newswire material from the Wall Street Journal (WSJ) corpus (Marcus, Santorini, and Marcinkiewicz 1993), with over 200K tokens in 385 documents, and one of the largest inventories of relations ever implemented, with 78 relations,[3] as well as a pseudo-relation type called SAME-UNIT, used to connect parts of discontinuous units, as shown in [33–35] in Figure 2.

2 See Stede (2012) for an in depth overview.

3 These include subtypes and variants accounting for different nuclearity patterns, which are often collapsed into 16 coarse classes in automatic discourse parsing work, cf. Hernault et al. (2010, page 6).

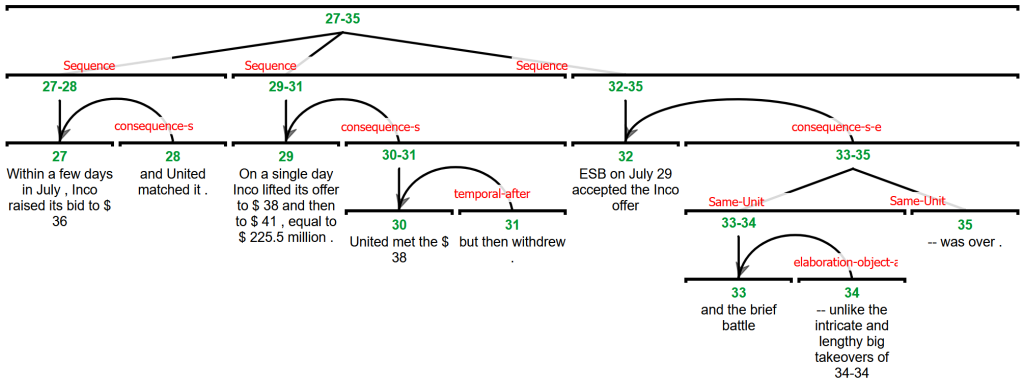


Figure 2
RST fragment from RST-DT: Satellites point to nuclei (e.g., [28] is a CONSEQUENCE of [27]) while the symmetrical SEQUENCE relation connects equally prominent nodes. [33] and [35] form a discontinuous SAME-UNIT.

The figure also demonstrates two shortcomings of RST, which fed into the development of subsequent work. The first is that upon closer inspection, we may notice discourse relations that are unexpressed in the tree: Unit [31] contains two discourse markers explicitly signaling different relations. The word "then" corresponds to the annotated relation TEMPORAL-AFTER, while the word "but" corresponds to no relation in the tree, but probably indicates the existence of a concurrent CONCESSION relation (see Moore and Pollack 1992). The second shortcoming is the lack of a distinction between such explicitly marked relations, for which we can supply simple textual evidence as a rationale (e.g., the existence of "then"), and implicit ones, such as the CONSEQUENCE satellite relation in [30–31], which is not indicated by a word like "then" or "but."

A first attempt to address the latter shortcoming in RST was undertaken in the RST Signaling Corpus (RST-SC; Das and Taboada 2018), which added signal type annotations to relations in the English RST-DT corpus, but did not anchor them to tokens. Thus for [31], the presence of explicit marking was annotated, but the word "then" was not identified as its locus. Liu and Zeldes (2019) presented a pilot study on anchored signals for RST-DT, which was extended to four genres from an early version of the GUM RST treebank (Zeldes 2017), anchoring signals to specific tokens (Liu 2019)—the present work develops this idea further in Section 3 below.

2.2 Segmented Discourse Representation Theory

SDRT (Asher and Lascarides 2003), is the most similar framework to RST in assuming graphs covering entire documents, and discourse units connected recursively using relations defined independently of formal marking. As in RST, EDUs also coalesce to form complex discourse units, which are in turn joined with others to create larger units. SDRT is also notable in producing resources that explore discourse structure for multiparty dialogue, such as the STAC corpus (Asher et al. 2016) and the Molweni corpus (Li et al. 2020), which focus on multiparty chat as part of an online game and in Ubuntu chat forums, respectively.

However, several differences distinguish SDRT, which also aligns with a specific formal semantic representation (DRT; Kamp, Van Genabith, and Reyle 2011) and defines

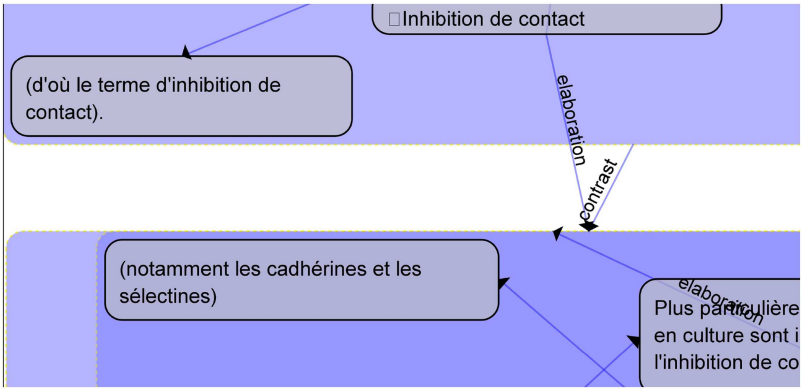


Figure 3
Fragment of an SDRT graph in the Glozz tool. The large blue discourse unit on the bottom has two incoming relations, CONTRAST from the large blue unit at the top, and ELABORATION from the gray EDU with the text "Inhibition de contact" at the top.

relations as part of a formal logic. Notably, SDRT allows multiple relations between units, as in example (2) from Lascarides and Asher (2007) and non-projective graphs, as shown in Figure 3 using Glozz (Widlöcher and Mathet 2012), the most commonly used interface for SDRT annotation.

(2) π_1: John bought an apartment. π_2: But he rented it.

In (2), Lascarides and Asher posit that unit π_2 forms both a NARRATION relation and a CONTRAST relation to π_1. SDRT also distinguishes coordinating relations, such as CONTRAST from subordinating ones, such as ELABORATION (Asher and Vieu 2005), but both can occur concurrently, as in Figure 3 for a French text from the ANNODIS corpus (Afantenos et al. 2010): The bottom complex unit (in blue) has incoming ELABORATION and CONTRAST relations, one from an EDU (in gray) and one from another complex unit. SDRT relations therefore do not reflect an RST-like notion of nuclearity or prominence. Units are also allowed to nest in each other, further complicating the data model.

2.3 Penn Discourse Treebank

The PDTB adopts a "lexically grounded" approach where discourse relations are annotated as senses of their associated discourse connectives (Prasad, Webber, and Joshi 2014). For instance, Figure 4 shows the same two concurrent relations from Figure 2, identified by the two connectives, *but* and *then*: COMPARISON.CONCESSION.ARG2-AS-DENIER and TEMPORAL.ASYNCHRONOUS.PRECEDENCE. Such relations are called **explicit** relations in PDTB-style corpora. On the other hand, since there is no connective between the first two sentences in the figure, there is no explicit relation annotation. However, an implicit connective "then" can be inserted between the two sentences ("*...Inco raised its bid... **Then** on a single day Inco lifted...*"), and therefore an **implicit** relation instance is identified and annotated as TEMPORAL.ASYNCHRONOUS.PRECEDENCE.

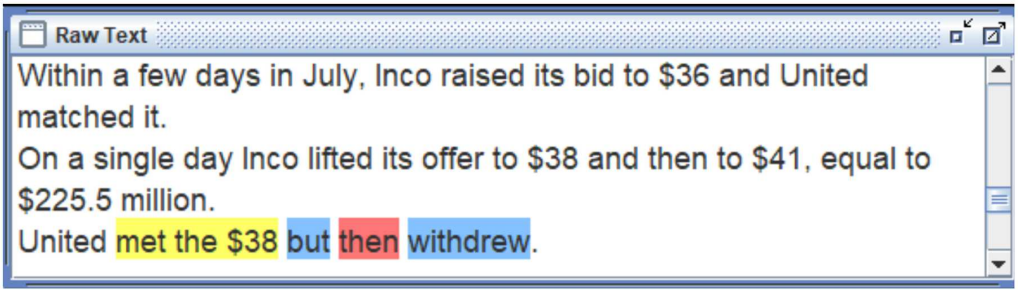


Figure 4
PDTB annotation interface for the same fragment from Figure 2. Two concurrent relations are recognized, corresponding to *but* and *then*, respectively.

In addition to **explicit** and **implicit** relations, several other types are recognized in the English PDTB v3.0: 1) Alternative Lexicalizations (AltLex), AltLexC, EntRel, Hypophora, and NoRel (Prasad, Webber, and Joshi 2014). AltLex items are expressions not considered to be connectives by PDTB's syntactic guidelines, which limit connectives to subordinating or coordinating conjunctions, prepositional phrases, and adverbs (for example "let alone," which can mark an EXPANSION.CONJUNCTION). AltLexC items are lexico-syntactic constructions which indicate relations, such as inverted auxiliaries marking a CONTINGENCY.CONDITION (e.g., "*Had I done it...*").

Arguments associated with identified relations in PDTB follow the principle of minimality: Only the minimal text needed for a given relation will be selected, which can be sentences, clauses, nominalizations, verb phrases, and so forth. (Prasad, Webber, and Joshi 2014). Additional text that is relevant but not necessary for the interpretation can be selected as supplementary information during annotation.

A major shortcoming of PDTB is the lack of higher-level structure over the relations between text spans (compare this with the RST annotation of the same fragment in Figure 2, which constructs an overarching nested structure). However, the lack of high-level structure makes annotation easier compared to RST and SDRT, as high-level structures are considered more challenging (Peng 2023). Thus, PDTB has allowed the creation of large corpora in a variety of languages such as Chinese (Zhou et al. 2014), Turkish (Zeyrek and Kurfalı 2017), Portuguese (Mendes and Lejeune 2022), and Italian (Tonelli et al. 2010) as well as for multilingual versions of TED talks (TED-MDB, Zeyrek, Mendes, and Kurfalı 2018; Zeyrek et al. 2019).

2.4 Cognitive Approach to Coherence Relations

CCR (Sanders, Spooren, and Noordman 1992), unlike most other discourse frameworks, offers a psycholinguistic account of discourse relations and discourse signalling, focusing on discourse comprehension. CCR defines discourse comprehension as a psychological mechanism that creates a coherent representation of the text content based on the ways text segments are linked with each other by discourse relations. CCR characterizes relations as a configuration of five dimensions, each decomposed into binary attributes (Sanders et al. 2021):

- Polarity: positive or negative discourse relations
- Basic operation: causal (strongly linked) or additive (weakly linked) relations

- Source of coherence: objective (semantic) or subjective (pragmatic) relations[4]
- Implication order: basic (antecedent-consequent) or non-basic (reverse) order
- Temporality: temporal or non-temporal relations

As an example, consider the relation in (3), from Sanders et al. (2021, page 11), which CCR decomposes as follows: The relation expresses a denial of expectation,[5] and hence, represents a *negative causal* relation (CONCESSION in RST). The relation links two segments that express facts; so, it is an *objective* relation. The implication order is *basic* since the antecedent segment precedes the consequent segment. Furthermore, the linear sequence of the segments represents a chronological progression, which makes the relation a *temporal* one.

(3) Although [they were officially assured the police would not be involved in the census] [many people are afraid of reprisals . . .]

CCR considers discourse signals (connectives/cue phrases) as processing instructions guiding the reader to infer the relation between segments (Sanders, Land, and Mulder 2007). In the absence of such signals, CCR postulates that relation identification may require additional cognitive energy and longer processing time, affecting text comprehension. Evidence to support these claims comes from both psycholinguistic and corpus-based studies (see Kleijn, Pander Maat, and Sanders 2019; Sanders et al. 2021). CCR annotation, like PDTB's, targets local-level relations and their connectives. CCR corpora, albeit fewer in number, are available for English (Rehbein, Scholman, and Demberg 2016) and Dutch (Vis, Sanders, and Spooren 2012).

2.5 Discourse Dependency Structure

DDS (Li et al. 2014; Morey, Muller, and Asher 2018) deviates from RST's constituency structure (a.k.a. *c-tree*) and connects EDUs using binary and asymmetrical dependency relations to facilitate parsing (i.e., *d-trees*). DDS aligns with widely used syntactic dependency structures such as Universal Dependencies (UD; Nivre et al. 2020) and offers a simple and transparent tree structure for annotating document-level discourse relations (Morey, Muller, and Asher 2018).

Only a few discourse treebanks are annotated natively in DDS, including SciDTB (Yang and Li 2018), SciCDTB (Cheng and Li 2019), and COVID19-DTB (Nishida and Matsumoto 2022). Most DDS data is converted from corpora in other frameworks—for example, Hirao et al. (2013) and Li et al. (2014) designed transformations from RST-DT to obtain parent-child relations between EDUs for summarization and discourse parsing. Both approaches produce binarized, asymmetric dependency trees translating nuclearity to headedness, while differing in the handling of multinuclear relations. Morey, Muller, and Asher (2018) further add head-ordering to preserve the scope of satellite modifications and render conversions between constituent and dependency

4 Similarly, RST relations are sometimes classified as either *subject matter* or *presentational* relations.
5 For a *positive causal* relation, the implication would be: police not involved → no need to fear reprisals.

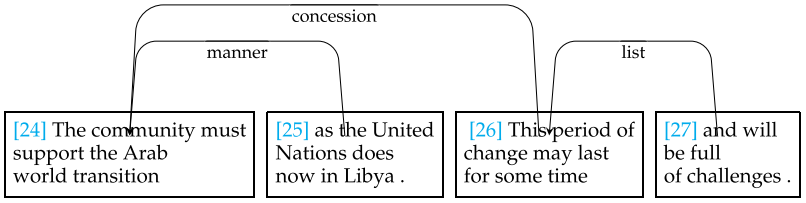


Figure 5
Head-ordered DDS converted from the RST fragment in Figure 1.

trees bijective. Figure 5 presents a converted head-ordered DDS equivalent to the fragment in Figure 1. The multinuclear *list* relation is transformed into a right-to-left dependency arc, with a chain modifying the first EDU in [24].

DDS datasets have also been converted from SDRT and PDTB data, with the latter complemented by automatically predicted higher-level relations (Stede et al. 2016; Yi, Sujian, and Yueyuan 2021). Due to the lack of large-scale DDS-native treebanks, discourse dependency parsing models are either trained on converted datasets (Yi, Sujian, and Yueyuan 2021) or through zero/few-shot learning and bootstrapping (Cheng and Li 2019; Nishida and Matsumoto 2022).

2.6 Multiple Frameworks

Some multilayer datasets have been developed that contain analyses in multiple frameworks in parallel. RST-DT and PDTB contain overlapping material from the WSJ corpus, allowing for some framework comparisons (Demberg, Asr, and Scholman 2019). Stede and Neumann (2014) and Bourgonje and Stede (2020) added connective annotations for explicit relations to the German RST-annotated Potsdam Commentary Corpus (PCC). Sun and Wang (2022) constructed a corpus of 500 Chinese "run-on" sentences annotated with both RST and PDTB-style analyses. However, to the best of our knowledge, this article is the first attempt at producing a new theory incorporating insights and advantages from the multiple frameworks described above, in which complete RST-style trees with nuclearity are anchored to connectives and other signals, while allowing tree-breaking relations as postulated in SDRT. The next section defines the scope of the formalism, before presenting data and parsing experiments implementing our analyses.

3. Formalism

Analyses in eRST aim to retain the benefits of RST trees (nuclearity and its applications to recursive summarization and information extraction, dialogue planning, etc.), while addressing weaknesses discussed in Section 2.1. Analyses consist of three components which we discuss below:

[1] A primary, single-rooted, labeled projective n-ary constituent tree over non-overlapping EDUs, which cover the text

[2] A possibly empty set of secondary labeled, directed, and possibly cyclic and/or non-projective edges, which are licensed under specific conditions

[3] A possibly empty set of categorized signals associated with a set of tokens and a single primary or secondary edge from [1] or [2]

Although in the study below we expand on a corpus with existing primary trees, the formalism is intended to be applicable to the analysis of plain text, for which a primary tree would then be prepared as part of the eRST analysis.

3.1 The Primary Tree

A primary tree G is defined, as in traditional RST, as a directed, single-rooted and fully connected labeled tree. Let V be a set of terminal and non-terminal vertices with a subset of ordered terminals S that are segments covering the tokens of the text T, and edges E between vertices with labels from the set L:

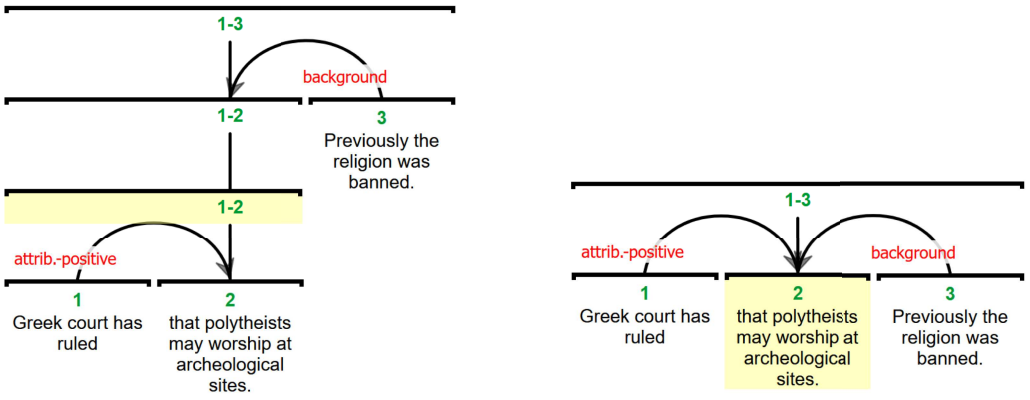
$$\begin{aligned} G &= \langle V, S, E, L, T \rangle \\ V &= \{v_1, v_2, \ldots, v_n\} \\ S &\subseteq V \\ E &= \{\langle v_i, v_j \rangle \mid v_i \in V \wedge v_j \in V \wedge v_i \text{ is the parent of } v_j\} \\ L &= \{l_1, l_2, \ldots, l_m\} \\ T &= \{t_1, t_2, \ldots, t_k\} \end{aligned} \tag{1}$$

Note that S is actually in almost all cases a *proper* subset of V: The sole exception is the degenerate case where there is only a single EDU, which produces $S = V$. The tree is further constrained to be projective. All tokens belong to exactly one terminal segment (i.e., EDU), and there is a single unique label for each edge:

$$\begin{aligned} &\forall\, t \in T[\exists!\, s \in S[s \text{ contains } t]] \\ &\forall\, v \in V[\exists!\, l \in L[l \text{ labels } v]] \end{aligned} \tag{2}$$

In addition, each node in V is classified as a satellite or nucleus node, and for each non-terminal node in V, at least one child node is a nucleus. eRST allows n-ary branching trees, though binarization via Chomsky Normal Form is possible as a trivial conversion for use with binary parsers.

The criteria for building trees are the same as in RST (see Mann and Thompson 1988; Taboada and Mann 2006), and will not be discussed in depth due to space. Briefly, propositions are grouped together based on the function they serve, with more prominent or necessary units being assigned the nucleus status, and less necessary or omittable units serving as satellites. Labels are defined based on relations' effect on the reader or hearer, and are assigned based on the perceived intention of the author or speaker to have such an effect, for example, a group of EDUs which is perceived as an explanation supplying evidence for a claim in another group of EDUs may be analyzed as a satellite to the latter group, with a label such as EXPLANATION-EVIDENCE. While eRST as a theory does not necessarily prescribe a specific set of relation labels, the labels in this article will come from the inventory of the GUM corpus, which has 32 total labels, including the label SAME-UNIT to connect parts of discontinuous EDUs (see Appendix A for the full inventory).



(a) "no empty hierarchy" (b) "no satellite ties"

Figure 6
Violations of additional constraints in eRST.

eRST primary trees are thus largely identical to RST trees, with some constraints made more explicit than in previous implementations. Specifically, eRST trees must define an explicit word *tokenization* of EDU contents to allow for the alignment of signals (see Section 3.3 below); it is assumed that there is no empty hierarchy, that is, each non-terminal has at least two children, and hierarchy is strictly ordered without ties. These constraints mean that there are no unary derivations, and no two satellite children for the same node. Violations of both constraints are illustrated in Figure 6.[6]

On the left, a redundant span has no child relations (the lowest span labeled [1–2]); on the right, [2] has two satellites which are not hierarchically ordered—instead eRST requires that [1] scopes over [2–3] (meaning the ATTRIBUTION contents of what the court ruled include the other two units, or that [3] scopes over [1–2], forming BACKGROUND to both. In this case, the latter option is the correct one.

3.2 Secondary Edges

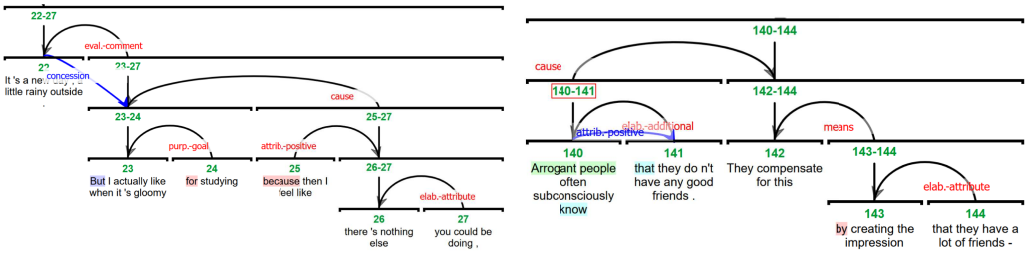
As noted earlier, some discourse relations occur in texts which cannot be expressed in a projective, acyclic tree as defined in Section 3.1. To represent such relations, we define a subset of edges, called **secondary edges**, which are not constrained by limitations on projectivity or cycles.[7] Secondary edges are permitted to connect any two nodes in the primary tree, including nodes which are already connected by a primary edge, subject to the following constraints. A secondary edge:

[1] may only be added if it is supported by a sufficient **signal**

[2] may only connect two nodes which are not already connected by a secondary edge with the same directed path

6 These constraints are often applied in RST trees in practice, but are occasionally violated in most datasets, and would cause problems for some of the algorithms we use for aligning signals below.

7 The term "secondary edge" is inspired by the "secedges" in the German Tiger Treebank (Brants et al. 2002), where similar edges where added to a primary syntax tree for tree-breaking dependencies.



(a) Secondary CONCESSION with orphan "but" (b) Secondary ATTRIBUTION

Figure 7
Secondary edges licensed by an orphan DM and reported speech. On the left, a secondary CONCESSION in blue points from 22 to the sentence containing 23–24; on the right, a secondary positive ATTRIBUTION points from 140 to 141.

[3] may not connect a node to itself

[4] may not require the definition of additional nodes

Constraints [2–3] mean that any two nodes v_1, v_2 in the tree can, at most, be connected via three edges: a primary edge, a secondary edge $v_1 \rightarrow v_2$, and a secondary edge $v_2 \rightarrow v_1$. This places an upper bound on the complexity of the formalism (see Section 3.4).

Constraint [1] lies at the core of our proposal for additional relations: Since agreement on discourse relations is already challenging, we want to limit additional edges only to clearly signaled cases. While it is conceivable that a variety of definitions could be used for "clear signals," we limit our proposal to two kinds of signals: Discourse markers (DMs) like "but" or "then"[8] for which no corresponding associated parent relation can be found, which we refer to as **"orphan DMs"**; and unambiguous morphosyntactic signals, in our implementation specifically restricted to either **reported speech** that is not already captured in a primary ATTRIBUTION relation,[9] and adnominal clauses that are not already captured using an embedded adnominal ELABORATION or PURPOSE relation, for example, relative clauses not interpreted as a primary ELABORATION.[10] Figure 7 demonstrates these two types of licensing conditions.

On the left, the secondary edge captures the relation corresponding to the orphan DM "but" (highlighted in dark blue), which has no corresponding relation in the primary tree—instead, the annotator perceives the main function of [23–27] as an EVALUATION of how a rainy day isn't too bad; note also that the nuclearity of the EVALUATION goes in the opposite direction of the CONCESSION, and that they do not scope over the same part of the text (the secondary edge connects [27] with [23–24]).

On the right, the fact that arrogant people subconsciously know they have no friends ([140–141]) is seen as the CAUSE of compensating for this. Although "know" is

8 We use the term "discourse marker" with largely the same definition and items as PDTB connectives, but with the difference that the spans they connect correspond to RST EDUs or complex discourse units, rather than argument spans following PDTB guidelines.

9 We note that ATTRIBUTION is especially well known to co-exist alongside other relations (Potter 2019), and has merited concurrent treatment in PDTB as well.

10 We do not rule out that other types of reliable signals could be added to license further secondary edges in future work.

a typical ATTRIBUTION verb, the annotator has analyzed the nucleus of the causal predicate to be *knowing* that they have no friends ([140]), rather than the fact that they have no friends in itself ([141]), forcing them to make [141] an ELABORATION to [140]. Nevertheless, syntactically [140–141] unambiguously follows the reported speech/cognition verb pattern, licensing a secondary ATTRIBUTION signaled by the predicate "know" and the complementizer "that" (syntactic signals highlighted in cyan), and the attribution source "Arrogant people," a semantic signal highlighted in green.

As shown in these examples, the relation inventory for secondary edges is assumed to be the same set of labels L used for the primary tree, though we note that if the inventory distinguishes a multinuclear and a satellite version of the same relation, these collapse and become indistinguishable for secondary edges, which do not indicate nuclearity by nature (though they do indicate directionality (e.g., a secondary CONCESSION still has a conceded part and a claim the concession contrasts with). In this way, eRST offers a partial remedy to previous criticism of RST in situations in which a single nuclearity choice may not express everything we want to know about discourse structure (Moore and Pollack 1992; Stede 2008), by allowing the expression of concurrent relations; however, the cost of retaining the advantages of the primary tree is that nuclearity itself is still kept unambiguous—secondary edges express only relations, and not overall prominence in the discourse structure.

As an annotation practice, we therefore recommend that secondary edges should only be added after the complete primary tree has been annotated, so that the most prominent relations can determine nuclearity without considering the presence of orphaned DMs. The first example in Figure 7 illustrates why this is important: If primary and secondary edges are annotated concurrently, annotators may be tempted to select the unmarked relation as primary, and utilize the resulting orphan "but" to establish the second relation, so that both can be marked. If the relation not corresponding to a DM is deemed more prominent from a functional perspective, as in Figure 7(a), then this is what we want; but that means the primary relation must always be established first, or else we may compromise our standards for determining nuclearity. In other words, the primary tree in eRST should be the same tree as in plain RST.

Finally, we note that not all potential DM items will necessarily receive a corresponding edge. Although we may expect any DM without a corresponding primary edge to automatically receive a secondary one, this will be ruled out in two cases: (1) when there are two orphan DMs which can be associated with secondary edges along the same path, due to constraint [2]; and (2) when the necessary spans for the edge to connect do not exist, conflicting with constraint [4]. While we did not encounter the first situation in our data (see Section 4.1), the second issue has occurred, especially when the necessary spans do not exist due to segmentation guidelines, as in Example (4).

(4) [**If** you live in or near a big city,]$_{<condition>}$ it is easier to attract enough customers .. [**than if** you live in a sparsely populated rural area.]$_{<antithesis>}$

In this example, there are three subordinating conjunctions: Two "if's" and one "than." The first "if" marks a primary CONDITION relation, and "than" marks an ANTITHESIS. Although the second "if" clearly has a conditional sense, it is not a condition of the main clause ("easier to attract customers"), but rather a condition for the implied elliptical clause that might have followed "than" ("easier .. than *it would be*"). However, because

such a clause does not appear, EDU segmentation guidelines prevent the existence of the necessary argument span for a secondary relation, and the orphaned second "if" remains without a corresponding edge.

3.3 Signals

Like PDTB, eRST anchors relations to markers in a text called "signals" (Liu 2019): This allows one not only to know which relation is indicated by which signal(s), but also to pinpoint exactly which words/phrases/constructions in the text contribute to the signalling of the relation. However, we assume a broader perspective on signalling than PDTB, encompassing much more than DMs and similar expressions. Following the taxonomy proposed by RST-SC (Das and Taboada 2014), we divide non-DM signals into seven types, corresponding to: *graphical*, *lexical*, *morphological*, *numerical*, *reference*, *semantic*, and *syntactic* features.

These types are divided into further subtypes, shown with examples in Table 1. For example, the *reference* type indicates that cohesion is signaled by anaphoric reference to an entity, with four subtypes: *personal* (anchored to a personal pronoun and its antecedent), *demonstrative* (an anaphoric NP headed or determined by a demonstrative), *comparative* (a comparative or relative expression, e.g., "(an)other" anaphora), and *propositional* reference (e.g., NPs referring back to a verbal phrase). Some signal types are anchored to a lexical indicative expression such as a word (e.g., "nice" signaling an EVALUATION), phrase, or other alternate expression, with the latter corresponding to

Table 1
Non-DM signal types and subtypes, with examples highlighting in red the signal tokens which indicate the relation of the unit in square brackets.

signal type	subtypes	example
graphical	colon, dash, semicolon	[*Let me tell you a story* **:**] $_{<organization-preparation>}$
	layout	[*Introduction*] $_{<organization-heading>}$
	items in sequence	*1. wash* [*2. cut*] $_{<joint-list>}$
	parentheses, quotation marks	*it rained* [**(***and snowed a bit***)**] $_{<elaboration-additional>}$
	question mark	[*Did you***?**] $_{<topic-question>}$ *No.*
lexical	alternate expression	*He agreed.* [**That is** *he said yes*] $_{<restatement-repetition>}$
	indicative word/phrase	*They planned a party!* [*That's* **nice/Can't wait!**] $_{<evaluation-comment>}$
morphological	mood	**Go** *with them* [*I think you should*] $_{<explanation-motivation>}$
	tense	*I* **started** *an hour ago,* [*now I'***m** *resting*] $_{<joint-sequence>}$
numerical	same count	[**Two** *reasons***.**] $_{<organization-preparation>}$ *First...*
reference	comparative	[*I don't want* **it**] $_{<adversative-antithesis>}$ *I want* **another one**.
	demonstrative / personal	*They met* **Kim**. [**This person** / **she** *was...*] $_{<elaboration-additional>}$
	propositional	**They met Kim**. [**This encouner** *was...*] $_{<elaboration-additional>}$
semantic	antonymy	*Beer is* **cheap**, [*wine is* **expensive**] $_{<adversative-contrast>}$
	attribution source	[**Kim** *said*] $_{<attribution-positive>}$ *they would*
	lexical chain	*it was* **funny** [*so they* **laughed**] $_{<causal-result>}$
	meronymy	**The house** *was big,* [**the door** *two meters tall*] $_{<elaboration-additional>}$
	negation	*Kim danced,* [*Yun did***n't** *dance*] $_{<adversative-contrast>}$
	repetition/synonymy	*They met* **Dr. Kim**. [**Dr. Kim**/**The surgeon** *was...*] $_{<elaboration-additional>}$
syntactic	infinitival/relative clause	*a plan* [**to** *win*] $_{<purpose-attribute>}$
	interrupted matrix clause	[*I meant* –] $_{<orgnization-phatic>}$ *I mean,*
	modified head	*a* **plan** [*to win*] $_{<purpose-attribute>}$
	nominal modifier	*articles* [**explaining** *chess*] $_{<elaboration-attribute>}$
	parallel syntactic construction	**it's all** *tasty* [**it's all** *pretty*] $_{<joint-list>}$
	past/present participial clause	*Kim appeared* [**dressed** *in black*] $_{<elaboration-attribute>}$
	reported speech	[*Kim said*] $_{<attribution-positive>}$ **that** *they* **would**
	subject auxiliary inversion	*I would have* [**had** *I known*] $_{<contingency-condition>}$

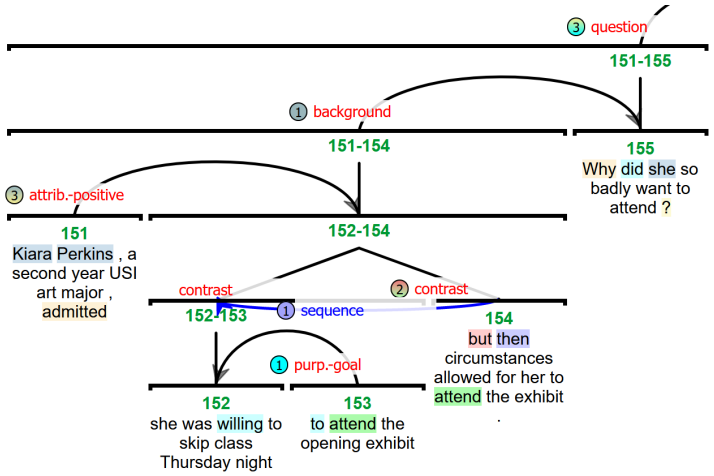


Figure 8
A larger fragment annotated in eRST.

PDTB's inventory of AltLex signals. Other signals may be anchored to tokens which are only signals in very specific contexts, such as negations accompanying an adversative relation with the same predicate in positive and negative environments, which can be crucial to interpreting contrasts (cf. Webber 2013). Others still refer to paired tokens, such as brackets or quotation marks, while some are not anchored to any tokens, such as graphical layout (a heading identified by its font, size, and separate appearance), or placement in a sequence of indented blocks or bullet points.[11]

Figure 8 shows a larger eRST graph fragment, which differs from a corresponding basic RST tree only in the addition of highlighted signals (background colors for tokens) and tree-breaking secondary edges (blue arrow edge type). Units [151–154] function as BACKGROUND to a question "Why did she so badly want to attend?" ([155]), which carries 3 signals (notice the number "3" next to the relation QUESTION): A *lexical* signal in dark yellow ("Why"), a *graphical* one in light yellow (the "?"), and a syntactic auxiliary inversion (in cyan, anchored to "did" in [155]). The BACKGROUND relation has a personal *reference* signal in gray ("Kiara Perkins . . . she," indicating the background relates to this person), and the contents of the BACKGROUND is attributed to Perkins in [151], signaled by a *semantic attribution source* (another signal anchored to the span "Kiara Perkins"), an attribution verb "admitted" (*lexical*, in yellow), and a complement clause headed by "willing" (*syntactic, reported speech*, in cyan). The PURPOSE-GOAL clause in [153] is signaled by a to-infinitive (*syntactic, infinitival clause*, anchored to "to"), and the CONTRAST is marked by "but" (a DM, in red) and the lexical chain "attend . . . attend." Finally, an orphan "then" (in blue) indicates the presence of the secondary SEQUENCE relation. More details on the annotation interface used for the visualization are given in Section 4.1.

11 Following acceptance of this article, we have also been discussing the possibility of incorporating implicit connectives, following PDTB definitions, as a type of signal not anchored to specific tokens, or even as a possible trigger for another type of secondary edges. We leave this idea for exploration in further work.

3.4 Complexity and Effort

Although eRST graphs as shown in Figure 8 are quite complex, the computational complexity of derivations in eRST is not very much higher than in RST. Because we retain the premise of a primary tree, and we constrain secondary edges to connecting at most each pair of nodes in each direction once, parsing secondary edges in the worst case scenario adds a single step in quadratic time. As for signal detection, although each span of tokens can be used multiple times with multiple signal types or for multiple relations (see the example "Kiara Perkins" above, which serves as both an *attribution source* and a *personal reference* signal for different relations), the problem of signal detection can be approached as token-wise multilabel classification, where each token must be classified for a complex signal type (we can consider, e.g., reference:personal_reference to be a single label) and a pointer to a corresponding relation from the closed set of relations available in the tree. In practice, automatic eRST parsing can be substantially less costly using a pipeline, as we will show in Section 5, or potentially a joint model.

In terms of annotation effort, an anonymous reviewer has brought up the relatively high cost of a manual RST analysis and its extensions using eRST, while another reviewer has asked why the starting point of the analysis is RST-like, rather than PDTB-like. To address the first point, we certainly agree that building primary trees is labor-intensive, an aspect in which eRST does not differ from RST; however, the amount of effort associated with the addition of secondary edges and signals can be reduced by relying on preprocessing from NLP tools for tasks that show promising performance, such as connective detection (Gessler et al. 2021; Liu, Fan, and Strube 2023), provided that these can be aligned to trees (see Section 5.3). Unalignable signals could then be inspected as indicators that secondary edges may be needed. Although the cost for eRST annotation at the moment is high, we would argue that it is unavoidable for a computationally implemented theory accounting for discourse relations and the devices used to mark them, since these will inevitably include multiple, concurrent instances. We are hopeful that with increasing performance of NLP models, much of the task could be done automatically, and we give some first numbers for complete automatic eRST annotation in Section 5.3.

Regarding the use of RST, rather than PDTB as a starting point, eRST focuses on incorporating two of PDTB's greatest advantages into a hierarchical discourse representation: (1), the inclusion of multiple concurrent relations, and (2), providing a text-anchored rationale for relations (DMs and other signal types). However, strictly building on top of PDTB as a framework would bring in its shortcomings (see Section 2.3): Lack of the usually implicit high level relations (e.g., between paragraphs), lack of hierarchical structure and nuclearity (recognizing that documents consist of more or less important but coherent parts and subparts), and a focus on a very limited inventory of signals (the latter is addressed by building on the RST-SC inventory instead).

4. Data

To confront our formalism with real data and provide testing grounds for linguistic and computational research using the theory, we extend the RST annotations in the English Georgetown University Multilayer corpus (GUM, Zeldes 2017). GUM is a growing corpus, created through a classroom annotation project in which students learn to annotate a text across a semester of coursework using multiple formalisms, resulting in a rich set of annotations for each document. The corpus covers morphosyntactic annotations according to Universal Dependencies guidelines (de Marneffe et al. 2021), nested

Table 2
Genres and Documents in the GUM Corpus, version 9.

genres	source	docs	tokens	EDUs
Interviews	Wikinews	19	18,190	2,410
News stories	Wikinews	23	16,145	1,779
Travel guides	Wikivoyage	18	16,514	1,792
How-to guides	wikiHow	19	17,081	2,395
Academic	various	18	17,169	1,981
Biographies	Wikipedia	20	18,213	2,071
Fiction	various	19	17,510	2,474
Web forums	Reddit	18	16,364	2,263
Conversations	SBC	14	16,416	2,810
Speeches	various	15	16,720	1,914
Vlogs	YouTube	15	16,864	2,436
Textbooks	OpenStax	15	16,693	2,027
total		**213**	**203,879**	**26,352**

entity annotations, coreference and bridging anaphora (see Zeldes 2022), complete RST trees, and more. At the time of this paper's submission, GUM (v9) encompassed 213 documents, which come from 12 different spoken and written genres, detailed in Table 2. These form the data analyzed in this article; however, we note that since that time, the GUM corpus has grown to encompass four more genres in version 10 (courtroom transcripts, essays, letters and podcasts), and has been expanded with a test partition-only corpus called GENTLE (**GEN**re **T**ests for **L**inguistic **E**valuation, Aoyama et al. 2023) with eight challenging genres.[12] These data sources have at the time of publication also been annotated in the eRST framework, bringing the total available data up to 246K tokens in 24 genres. To browse these analyses, see the eRST Website.[13]

With over 26K EDUs in v9 (or 32K in the recently released v10 + GENTLE), GUM is the largest English RST corpus, followed by RST-DT, with 21,789 EDUs. GUM data is available from the corpus Website (`https://gucorpling.org/gum/`), with underlying text under respective licenses from each source, and annotations under a Creative Commons Attribution license (CC-BY 4.0).

Due to the centrality of the WSJ corpus in past studies of RST, for this article we also partly annotate RST-DT for eRST annotations, focusing on the test-set of 38 documents, to which we add full token-anchored annotations of discourse markers (including orphans based on the primary tree) and corresponding secondary edges according to our guidelines. Due to the licensing restrictions on RST-DT, we release these annotations separately, without the underlying text.

4.1 Annotation Process

Since primary RST trees were already available for GUM and RST-DT, eRST annotations were divided into three main parts: (1) identifying and associating DMs with trees; (2) adding secondary edges where DMs were left over as orphans or syntactic triggers were identified; and (3) adding non-DM signals using semi-automatic methods.

12 Specifically: dictionary entries, live eSports commentary, legal, medical, poetry, mathematical proofs, syllabi, and threat letters.
13 `https://gucorpling.org/erst/`.

Table 3
DM detection and attachment performance Untyped = connective detection; Typed = detection + classification: DM or orphan; Sourced = detection + classification + association with the correct relation.

	untyped	typed (total)	typed (DM)	typed (orphan)	sourced
P	89.31%	72.73%	86.39%	23.11%	46.39%
R	78.83%	64.18%	68.68%	33.93%	40.88%
F	83.74%	68.16%	76.51%	27.50%	43.46%

4.1.1 DM, Orphan, and Secondary Edge Annotation. For DM identification, we preprocessed the corpus with the best-performing English connective detector trained on PDTB v3, the DisCoDisCo system (Gessler et al. 2021), winner of the DISRPT 2021 shared task on Connective Detection (Zeldes et al. 2021). This step was undertaken to ensure high recall, high conformity with PDTB connective definitions, and high consistency (it has repeatedly been shown that correcting state of the art NLP outputs outperforms purely manual annotation due to tool consistency; see Mikulová et al. 2022).

Following connective detection, a script associated each predicted connective with the nearest compatible relation in the tree hierarchy, prioritizing the outgoing relation of the EDU containing it, followed by recursively searching for a larger span containing the original EDU until finding a relation compatible with the connective, based on the PDTB guidelines and the PDTB-RST relation mapping from Demberg, Asr, and Scholman (2019). If no relation is found, then the connective is flagged as an orphan.

Table 3 summarizes performance for connective detection and alignment. The higher precision for all metrics (except orphans) is due to the aligner searching for a compatible *outgoing* relation, but not *incoming* relations. Allowing attachment to any compatible incoming relation increases recall but degrades precision substantially. The current outputs were deemed sufficient as a starting point for manual correction.

After this preprocessing, five annotators went over the entire dataset manually to correct DM identification and alignment to relations, adding secondary edges for true orphans, whose relations were not expressed in the primary tree. Manual annotation was done using rstWeb (Zeldes 2016), an open source Web interface for RST annotation which was extended to support signal marking by Gessler, Liu, and Zeldes (2019), and which we extend further for this article with support for secondary edges.[14] The annotation process was repeated for the RST-DT test set.

We conducted an inter-annotator agreement study of DM identification and relation association for 36 GUM documents and the RST-DT test set, and report mutual F1 scores. For RST-DT, three annotators double-annotated 38 documents in the test partition, amounting to about 21K tokens. A mutual F1 score of 95 was obtained for identifying DMs, and 88.8 for relation association. For GUM, two annotators double-annotated 36 documents (32K tokens) and obtained an F1 score of 92.3 for DM identification and 90 for relation association.

For secondary edges, a first inter-annotator agreement experiment on the GUM dev set (24 documents) showed substantial disagreements, with S/N/R scores of .311, .279, and .223, corresponding to % agreement on edge attachment points (regardless

14 Available at `https://gucorpling.org/rstweb/info/` under the MIT license.

Table 4
Twenty additional GUM DMs not attested in PDTB as connectives.

type	marker	by analogy to
subordinating	cuz	because
	cause	because
	whilst	while
	as far as	as long as
	into	marks RESULT, e.g., "trick someone **into** thinking"
	that	marks PURPOSE, e.g., "I longed for nets, **that** I might capture them"
	whither	where
	wherever	whenever
adverbial	as such	marks RESULT
	apart from	aside from
coordinating	or else	not attested in PDTB; used for disjunction, analogous to "otherwise"
	/	standing for "or"
prepositional	by the end	in the end
	in essence	in short
	at the time	at the same time
	around the same time	at the same time
	to that end	to this end
	to wit	not in PDTB, similar to "for example" or "specifically"
	in brief	in short
	since then	since (used adverbially in isolation)

of source/target), exact edge path (including directionality), and the assigned relation. This is despite the fact that annotators achieved an F-score of .642 on detecting orphan DMs (i.e., agreeing that a secondary edge was called for, and where the DM tokens were). Inspection revealed that disagreements hinged either on whether an orphan candidate was a connective (especially for spoken sentence initial "And" or "So") and what the exact scope of the relation included (e.g., including or omitting trailing bibliographical citations in academic data). After refining the guidelines to be more explicit, a second experiment on an additional 12 documents produced much better results of S/N/R = .529, .49, and .412, respectively, indicating agreement levels just 16 points below the human agreement score on primary relation R of .571 (Morey, Muller, and Asher 2017). We consider this to be substantial agreement, especially given that secondary relations involve some of the most challenging ones, and their scores do not benefit from common, easy cases such as relative or other adnominal clause attachments.

Finally, with the entire corpus in hand, one of the authors of the article reviewed all annotations for a final consistency pass and finalized the list of possible DMs, which for RST-DT was a subset of the PDTB connectives. The list of DMs in GUM required some expansions due to items that are not attested in PDTB or related corpora, such as TED-MDB, probably due to the different genres in the corpus.[15] Table 4 provides the added items, along with the rationale for adding them by analogy to existing PDTB items.

We note that "aside from" is attested in PDTB only once as an AltLex governing a noun phrase, but in GUM we have "apart from thinking," which fits the DM definition

15 Although TED-MDB contains spoken data, it does not cover dialogue, nor data from the web, such as GUM's Reddit data. Such user-generated content often contains unique words or spellings, such as "/" or "cuz" for "or" and "because," and may require adaptation of guidelines; see Sanguinetti et al. (2020, 2022).

by governing a VP. The item "as such" is attested in PDTB as AltLex as well, but should be an explicit connective since it is a relational prepositional phrase, similar to "at the same time." All other DMs in GUM beyond the 20 in Table 4 are attested in PDTB, amounting to a total of 211 distinct types, disregarding connective modifier variants such as "*two/three months* later," which we consider to belong to the type "later."

4.1.2 Non-DM Signals. Due to limited resources, we did not annotate or correct all non-DM signals fully manually, and did not annotate them in the RST-DT data. However, thanks to the rich annotations available in GUM, we were able to induce many signal types fully automatically from the gold syntax trees and coreference annotations, and were able to manually annotate or correct many other cases, and evaluated accuracy and agreement manually on a subset of documents (see below).

Graphical and *reference* signals were tagged automatically based on token forms (parentheses, question marks, etc.) and gold coreference chains for eligible relations. For example, any QUESTION relation whose sentence contained a question mark was assumed to be signaled by that question mark, and any ELABORATION relation containing pronominal anaphora in a satellite pointing to an antecedent in the nucleus was automatically admitted as a *reference* signal. The list of relations eligible for each such signal type was obtained by consulting the RST-SC corpus.

Some *morphological* and *syntactic* signals were also tagged automatically using the Python dependency tree-editing library DepEdit (Peng and Zeldes 2018): Relative or infinitival clauses conveying adnominal attribute relations are easy to identify, as are reported speech for ATTRIBUTION, imperative mood for MOTIVATION, modals in a CONDITION, and some tense markers (e.g., past perfect marking BACKGROUND relations). Harder cases required manual verification, such as change of tense to signal a SEQUENCE (e.g., past followed by present or present followed by future, fully verified manually), parallel syntactic constructions (annotated completely manually), or semantic attribution sources. For the latter, the subject or external subject (for nested or coordinate verb phrases) of each attribution predicate was identified using (enhanced) UD trees, and remaining cases for which the source could not be identified were annotated manually, such that every ATTRIBUTION in the corpus has a source signal.

Among the trickiest categories to annotate were *lexical* signals, which require a large inventory of candidate items, and *semantic* lexical chains, which consist of related, non-co-referring word or phrase pairs that signal a relation, and are open-ended. For the former, we took the combined list of AltLex expressions in PDTB, a manually filtered list of the top 100 items most associated with each relation by chi square statistics, and additional items that were noticed during the annotation, all restricted to relevant POS tags. We observe that automatic annotation of all such items as signals for their associated relation was close to error free: This is intuitively not very surprising, since, if a word associated with EVALUATION, such as "good," appears in such an already manually annotated gold standard relation, it is highly likely to be signaling that relation. A total of only 10 errors contradicting this approach were noticed during quality controls. To illustrate the process, consider example (5), from a short story about a boy leading his developmentally disabled sister, Cara, out of a store after an unpleasant incident.

(5) I herd Cara towards the front of the store, mouthing sorry at the front cashier. [She's kind of **pretty**.]$_{<evaluation-comment>}$ She smiles at me. [**Nice** big brother with retarded sister.]$_{<evaluation-comment>}$

The first EVALUATION contains two words listed as evaluative: "kind" and "pretty"; however the instance of "kind" is in the wrong part of speech ("kind" is listed as evaluative only as an adjective), so only "pretty" is selected as a signal. In the second EVALUATION about the cashier's smile, "Nice" is correctly identified; a false positive, adjectival "big" would also be included as a signal, and constitutes one of the 10 errors noticed in the data and removed manually.

For lexical chains, we were concerned about creating a subjective list of associated word pairs from the corpus, and instead decided to use a large existing inventory of word associations, opting for MIT's ConceptNet (Speer, Chin, and Havasi 2017), which contains over 34 million conceptual relations between words. We allowed a script to suggest lexical chains of two or more items in the sentence containing each connected satellite and nucleus, or in two connected clauses for intra-sentential relations, and filtered the result manually. Since ConceptNet does not connect items to themselves, but lexical repetitions or variations on the same stem (e.g., "participate"..."participant") can be signals, especially for RESTATEMENT and PREPARATION relations, we also allowed candidates based on stem matching using the Snowball stemmer for English (Porter 2001). Examples of both types of chain appear in (6)–(7):

(6) He had no political **power**, [and his **influence** extended only so far as he was humored by those around him]$_{<elaboration-additional>}$

(7) [Have a realistic but **exaggerated** setup.]$_{<organization-preparation>}$ The opening of the joke—or setup—should have a basis in the real world so your audience can relate to it, but it should also include **exaggeration**

In (6) "influence" is recognized as a type of "power," mirrored by a ConceptNet "is-a" relation between the two words, while in (7), the identical stem "exaggerat-" helps to identify how the initial PREPARATION satellite prepares the reader for the subsequent nucleus.

The final list of lexical chains amounted to 1,280 manually verified instances, covering 2,825 tokens in the corpus, meaning approximately 1.3% of corpus tokens are part of a lexical chain. We note that the total would have been much higher due to plain repetition of nouns, for example, in ELABORATION or BACKGROUND relations, but many of these were rejected from the chain proposing script, not because they were irrelevant, but because they were already captured under coreference-related signal types, and were therefore ineligible to be lexical chains (e.g., the "setup" mentioned twice in (7), is excluded as a lexical chain, because it is already part of a coreference-based signal instead).

To evaluate the accuracy of our annotations, and humans annotators' ability to agree on non-DM signals, we conducted two experiments: annotations from all signal categories were manually corrected for 12 documents in the test set (one from each genre), and four of these were double annotated. Table 5 gives mutual precision/recall and F1-scores for the humans (hum-v-hum, 4 docs) and for the automatic annotation performance compared to the human annotation (cpu-v-hum). We evaluate in two scenarios: token-anchored (signals only match if their type, subtype, covered tokens and associated relation match) and unanchored (token spans may differ), and we report micro and macro-averaged scores (across documents average).

Table 5
Human vs. human and system vs. human agreement for all signal types on a subset of documents.

	anchored			unanchored		
hum-v-hum	P	R	F1	P	R	F1
Micro	0.813	0.798	0.805	0.865	0.837	0.851
Macro	0.809	0.801	0.804	0.859	0.839	0.848
cpu-v-hum	P	R	F1	P	R	F1
Micro	0.841	0.920	0.879	0.868	0.950	0.907
Macro	0.837	0.922	0.877	0.865	0.954	0.906

Comparing human vs. human scores (top half of the table) to automatic system scores (bottom) is not strictly possible, since human agreement is computed on a smaller subset of documents; that said, we can observe that the system performs about as well as humans (on precision) or better (on recall), and that the gap between anchored and unanchored scenarios is similar as well. We suspect that the reason why the system has the upper hand in recall is that two humans inevitably generate additional disagreeing signals, while they are less likely to remove a predicted signal unless it is truly wrong. As a result, when compared with adjudicated output containing only the more unanimously recognized signals, the system does not miss less certain cases which one human might flag but not another.

Looking at prediction errors more qualitatively, we note that lexical chain disagreements are by far the most common in both human and system performance, followed by indicative words, while syntactic and coreference based signals are almost always correct. To understand why, we consider the typical example type in (8)–(9), compared with the very uncommon syntactic error type in (10).

(8) The Beavertail **cactus** can grow to be about 24 inches . . . has **pads** that look like beavertails

(9) There are **holes** in the center of the base of each pot to allow for **drainage**

(10) A: Do you need a partner? B: **To** go there?

In (8), an annotator recognizes "cactus" and "pads" as a meaningful lexical chain for a LIST relation, a pair of related terms since pads are the leaves of a cactus; however the second annotator does not recognize this chain, and they are not included as related terms in ConceptNet. In (9), one annotator recognizes "holes" and "drainage" as a chain indicating a PURPOSE relation, which is again not detected by another annotator or ConceptNet. Finally in (10), we see an unusual case of a syntactic signal across speakers, where a human annotator marks the infinitive "to" as a syntactic signal indicating PURPOSE, but the system misses the signal, since these are separate sentences and there is therefore no syntactic tree relation to indicate the signals—such examples of syntactic signal corrections are very rare. Overall, we feel these results indicate a high level

Table 6
DMs and secondary edges in RST-DT test, GUM, and the GUM genres.

genre	dms	orphans	dms+orphans	relations	markers_per_rel	secedges	%secedge
RST-DT	406	87	493	2,580	0.191	87	3.37%
GUM	6,025	895	6,920	29,950	0.231	1,008	3.37%
news	334	38	372	1,933	0.192	43	2.22%
academic	403	55	458	2,069	0.221	61	2.95%
bio	392	38	430	2,303	0.186	53	2.30%
conversation	484	127	611	3,341	0.182	131	3.92%
fiction	611	85	696	2,899	0.240	97	3.35%
interview	514	73	587	2,698	0.217	83	3.08%
reddit	622	89	711	2,606	0.272	100	3.84%
speech	446	70	516	2,179	0.236	76	3.49%
textbook	440	47	487	2,201	0.221	54	2.45%
vlog	675	181	856	2,942	0.290	193	6.56%
voyage	397	50	447	2,062	0.216	64	3.10%
whow	707	42	749	2,717	0.275	53	1.95%

of reliability for the additional signal types in the corpus, while also indicating that subtypes such as lexical chains and other indicative lexical signals warrant more study in order to arrive at dependable operationalizations that do not rely solely on a lexical resource like ConceptNet.

4.2 eRST Annotations across Genres

Table 6 gives an overview of the prevalence of secondary edges in GUM as a whole, as well as by genre, and in comparison to our annotations of the RST-DT test-set. As the table shows, ∼13% of discourse markers in GUM are orphans (895/6,920), fewer than in RST-DT, which has ∼17% (87/493). At the same time, the proportion of secondary edges is identical in both datasets at 3.37%, and there are only slightly more DMs per relation in GUM (.231 on average per relation, compared to .191 in RST-DT). The latter differences suggest different amounts of unmarked relations in both corpora, and slightly more relations with multiple markers in GUM.[16]

Looking at GUM genres, we see considerable variation. *News* unsurprisingly comes very close to the RST-DT values for DMs, but far below for secondary edges, perhaps because GUM news stories are shorter than some of the long and complex texts in the RST-DT test set. Other genres are even stronger outliers: Secondary edges are rarest in how-to guides (*whow*, 1.95%) and most common in *vlog* (6.56%), the latter due to frequent linking of sentences with sentence initial *And*, and to some extent *So* (only counting cases in which "so" is actually a DM). *Academic* is surprisingly below average in DMs per relation, despite common assumptions in the literature about the explicitness of academic text (e.g., Hughes 1996; see Biber and Gray 2010 for criticism).

However, it would be wrong to say that *vlogs* are more explicit in their discourse relations than *news*, since DMs/orphans are not the only means of signaling relations.

16 An anonymous reviewer has noted <4% of relations being secondary may mean that they are close to negligible, but we note that depending on our interests they can be very important: Over 14% of GUM RESULT relations are secondary, as are over 10% of CONCESSION relations (see Section 6 for more statistics). These are highly relevant to semantic applications, research on political speech, and more, and demonstrate the added value of eRST in covering the full breadth of relations in texts.

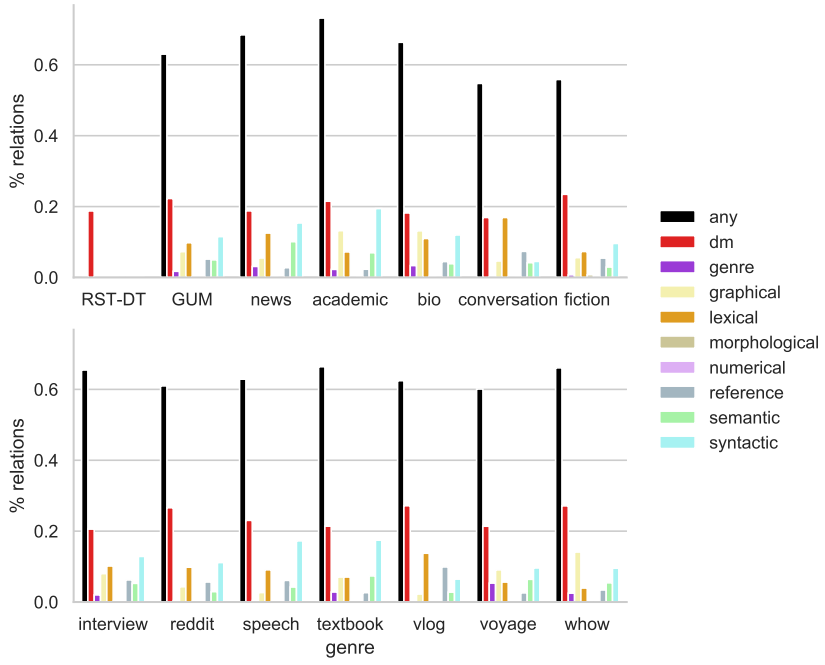


Figure 9
Distribution of signal types across genres vs. GUM and the RST-DT test set.

Turning to all signal types, Figure 9 shows the percentage of each major signal class across all GUM genres vs. the entire corpus, including the proportion of relations signaled by any means in black ("any"). RST-DT is shown for comparison but only has data for DMs (the "dm" bar includes orphans, and secondary edges are counted in relation totals).

As we can see GUM *news* is very comparable to RST-DT in DM prevalence. We also observe that *academic* is not actually low in signals: It marks the most relations at 73.2%, followed by *news* at 68.4%, *textbook* (66.4%), and *whow* (66.1%), mainly owing to syntactic cues in the first three, and to some extent frequent graphical signals as well (esp. *whow*). The overall "any" signal proportion in GUM is 63%, much lower than the figure by Das and Taboada (2017) for RST-DT at 92.7%, though we note that some different signal types were included in that study, and that GUM *news* is closer to that mark.[17] By contrast, the finding of around 20% DM marking in RST-DT test is in line with a similar estimate of 18.21% in Das and Taboada (2017, 26).

Another set of contrasts obscured by looking just at signal types can be seen by comparing how each coarse relation class tends to be signaled, which is very heterogeneous.[18] Figure 10 gives the proportion of signals for each relation class. We can see

17 The lack of token-level searchable annotations in RST-SC complicates tracking down causes for these differences quantitatively, but data inspection suggests the inclusion of much more wide ranging lexical chains constitutes most of the difference.

18 The same can be said for fine-grained relations, which we disregard here for space reasons.

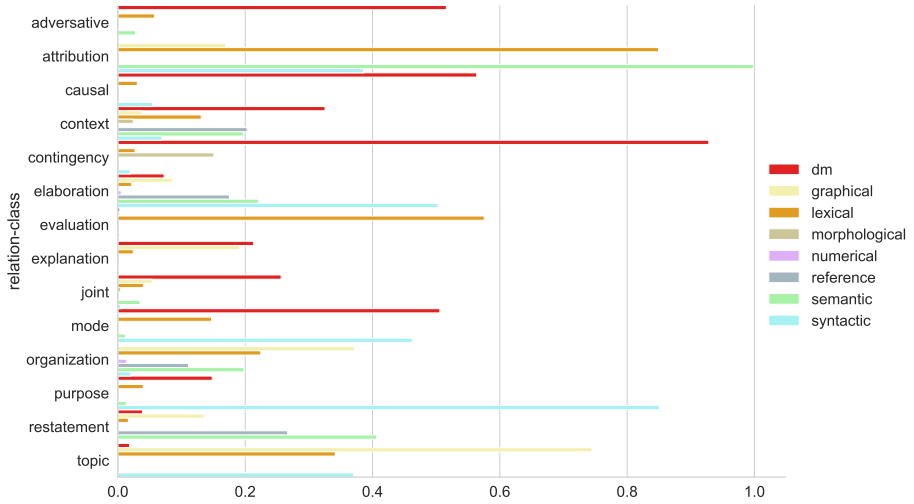


Figure 10
Distribution of signal types for each coarse relation class in GUM.

some relations are almost never signaled by a DM (for example ATTRIBUTION, ORGANIZATION), while others nearly always have one (CONTINGENCY, normally marked by *if*, but sometimes by other means, such as syntactic inversion or morphological mood), and combinations are quite common (cf. Crible 2022). The PURPOSE class is marked almost always by a syntactic signal (usually a *to*-infinitive), but can have other signals (e.g., the DM *in order*). Relations which can be signaled by recurring mentions, such as ELBORATION, RESTATEMENT and CONTEXT, show such *semantic* and *reference* signals often (especially *semantic repetition*). We can also see that ATTRIBUTION is always marked (not least because a *semantic source* for the attribution should be present by definition), while the least marked class overall is JOINT, containing, for example, temporal SEQUENCE (often marked only by implicit chronological order), LIST, and other coordinate structures.

The most lexically marked classes are EVALUATION (e.g., positive and negative adjectives) and ORGANIZATION, the latter primarily due to back-channeling and preparation markers in conversation ("Uh-huh," "you know,", "I mean,"). Since our data delivers not only statistics on relation signaling types, but actual aligned token indices for each relation's signals, we can extract the most frequent DMs and lexical expressions used to convey each relation, which are given in Table 7.

As the table shows, some of the most frequent DMs are unsurprisingly polysemous ("and," "so," and "as" occur in several classes), and next to DMs we find lexical signals for every class, which often work in tandem with the DMs, for example, the DM "then" in the JOINT class can co-occur with lexical items such as "today" to mark temporal sequences, or with adverbs not recognized as DMs by PDTB, such as "too" for a JOINT-LIST. And while ATTRIBUTION is only marked by a DM twice, as in (11), it is usually accompanied by a speech or cognition verb such as "said" or "think," providing a clear lexical signal.

(11) [**As** Heald told The Huffington Post,]$_{<attribution>}$ US surface ozone has dropped partly due to the Clean Air Act.

Table 7
Top DMs and lexical markers per coarse relation class.

relation	freq	% signaled	top DMs	top lexical
adversative	2,405	55.88%	but (641), however (101), though (59), and (59)	may (35), only (24), might (14), actually (13)
attribution	1,592	99.94%	as (2)	said (162), think (152), know (113), say (73)
causal	1,240	63.31%	and (221), because (167), so (167), as (22)	due to (7), result (6), caused (6), cause (3)
context	2,317	79.59%	when (311), as (96), after (81), while (48)	never (17), always (17), often (12), following (11)
contingency	518	96.72%	if (367), if then (35), when (31), unless (12)	depending (4), based on (3), every time (2), in the case (2)
elaboration	5,339	89.32%	and (208), also (55), for example (28), with (25)	including (36), too (18), e.g., (17), especially (14)
evaluation	1,047	57.21%	and (5), so (4), so that (1)	good (45), very (40), important (19), great (19)
explanation	1,650	41.94%	so (104), because (69), as (42), and (22)	see (22), shown (5), based on (3), considering (3)
joint	8,922	34.17%	and (1701), also (211), then (202), or (127)	now (39), too (34), again (24), today (12)
mode	512	78.71%	by (98), as (76), without (32), as if (14)	using (44), based on (19), according to (5), guided by (2)
organization	1,805	75.73%	thus (1)	you know (94), yeah (57), oh (54), I mean (42)
purpose	904	94.03%	for (39), so (30), in order (24), so that (18)	stop (7), achieve (6), prevent (5), avoid (4)
restatement	1,213	55.56%	and (23), in other words (7), or (7), in short (2)	that is (6), aka (6), i.e. (5), known as (2)
topic	486	82.51%	so (6), if (3)	what (80), how (42), why (24), who (11)

The data in the table only begins to scratch the surface of how relations are marked, and much more remains to be learned by examining the ways in which relations can conceivably be marked, and ways in which the same items may occur without signaling the relation with which they are associated, a topic we leave to future research.

4.3 Search and Visualization

To support exploration of the data, we add support for eRST annotations to the existing RST search and visualization facilities in ANNIS (Krause and Zeldes 2016, available open source under the Apache 2.0 license), an open source search platform for multilayer corpus data.[19] ANNIS supports search across all annotation layers in GUM, meaning queries can combine syntactic structures, coreference links, RST relations, and more. For instance, the ANNIS Query Language example in (12) searches for a non-terminal `group` (a complex discourse unit) dominating a terminal `segment` (an EDU) with some ADVERSATIVE relation type, using a regular expression, with a representative result shown below it (node colors in the query correspond to covered text color in the example).

(12) `kind="group" >[relname=/adversative.*/] kind="segment"`

a. [*The value of Airbnb is approximately $30 billion.*] [*Compare this market value to Hilton 's market capitalization of $19 billion and Marriott's of $35 billion.*]$_{<adversative-contrast>}$

b. [*Not with your gloves or anything .*]$_{<adversative-antithesis>}$ [*Find something else to pick it up with.*]

Using the expansion to the functionality to support eRST, we can also limit results to ones in which a DM signal is available, and anchored to a word with a particular POS tag, for example, a coordinating conjunction (POS tag `CC`), marked in green in (13). The operator `_i_` indicates that the second node *includes* the POS node, and the expression

19 Available from `https://corpus-tools.org/annis/`.

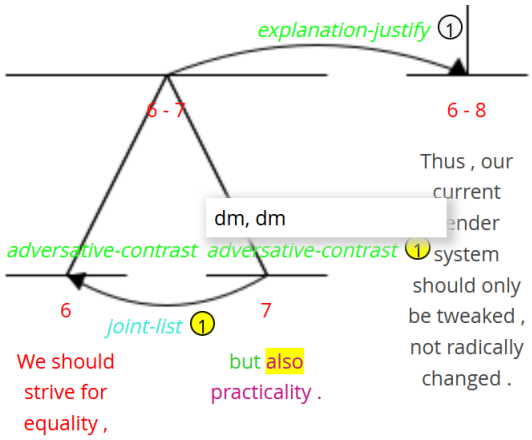


Figure 11
ANNIS visualization for the second result of the query in (13). The visualization closely follows the view in rstWeb, except for editing facilities.

`signal_type="dm" > #3` indicates that a DM signal dominates the token declared as the third node, which carries the POS tag `CC`.

(13)
```
kind="segment" >[relname=/adversative.*/] kind="segment" _i_
pos="CC" & signal_type="dm" > #3
```

a. [*They have ideas*] [*but they can not formulate them in the right way.*]$_{<adversative-contrast>}$

b. [*We should strive for equality*] [*but also practicality.*]$_{<adversative-contrast>}$

In the last example, where the `CC` is matched by the word "but," we also see a second marker "also" which is not highlighted in the query result—this corresponds to a secondary edge orphan DM, whose edge can be seen in the ANNIS visualization for this search result in Figure 11. Each signal can be highlighted by hovering over the counter button next to each relation (showing "1" for the CONTRAST relation). The secondary edge corresponding to "also" has the relation JOINT-LIST connecting [6] and [7]. We release the code for the new visualization and search capabilities, and make the annotated corpus freely available for search via ANNIS at `https://gucorpling.org/annis`.

5. Parsing eRST

5.1 Task Definition and Metrics

Using notation from Section 3.1, the goal of conventional RST parsing is to produce the tree G given the textual tokens T and the EDUs S, which partition T into contiguous

spans.[20] RST parsing is usually evaluated with the Parseval metrics, and we follow previous work in considering only binarized trees and using the original Parseval scoring scheme instead of the older RSTParseval (Black et al. 1991; Morey, Muller, and Asher 2017). Each non-terminal vertex can be seen as the product of a **parsing decision**, where two vertices $\{v^a, v^b\}$ are joined by a relation with nuclearity n and label l. We refer to the unordered pair $\{v^a, v^b\}$ as the decision's associated span s. For any well-formed sequence of parsing decisions $D = \langle\langle s_1, n_1, l_1\rangle, \ldots, \langle s_m, n_m, l_m\rangle\rangle$, there is exactly one tree that may result, so evaluating decisions is equivalent to evaluating the tree.

To evaluate parser output, consider the gold parsing decision sequence D, and the parser's decision sequence $\hat{D}$, and let $d.s$, $d.n$, and $d.l$ correspond to span, nuclearity, and label of a single parsing decision d. The four Parseval metrics can now be defined as precision[21] metrics over the sets D and $\hat{D}$:

$$\begin{aligned}
\textbf{S}\text{pan}(D, \hat{D}) &= \frac{\#\{\hat{d} \mid \hat{d} \in \hat{D} \wedge \exists d \in D[d.s = \hat{d}.s]\}}{\#\hat{D}} \\
\textbf{N}\text{uclearity}(D, \hat{D}) &= \frac{\#\{\hat{d} \mid \hat{d} \in \hat{D} \wedge \exists d \in D[d.s = \hat{d}.s \wedge d.n = \hat{d}.n]\}}{\#\hat{D}} \\
\textbf{R}\text{elation}(D, \hat{D}) &= \frac{\#\{\hat{d} \mid \hat{d} \in \hat{D} \wedge \exists d \in D[d.s = \hat{d}.s \wedge d.l = \hat{d}.l]\}}{\#\hat{D}} \\
\textbf{F}\text{ull}(D, \hat{D}) &= \frac{\#\{\hat{d} \mid \hat{d} \in \hat{D} \wedge \exists\, d \in D[d.s = \hat{d}.s \wedge d.n = \hat{d}.n \wedge d.l = \hat{d}.l]\}}{\#\hat{D}}
\end{aligned} \tag{3}$$

Note that all metrics depend first on checking whether some predicted span exists in the gold tree. The Span metric does only this, and the remaining three metrics add criteria: Nuclearity and Relation metrics also require the span's nuclearity and relation label, respectively, to match the corresponding span in the gold tree, and the Full metric requires matching span, nuclearity, and relation.

eRST introduces signals and secondary edges, which must be scored as well. First, let us extend our formalization so that in addition to the vertices joined by the edges v_i, v_j, each member of E also carries a binary indicator variable σ which is true only for secondary edges. Further, we expand G with Λ, a vocabulary of signal labels, and Z, the signals, where each signal may be defined as $\langle e, \lambda, \tau\rangle$: e is the associated edge,[22] λ is the signal type label of the signal,[23] and τ is a set of tokens associated with the signal.

Let us define signal precision $\mathbf{S_P}$ and signal recall $\mathbf{S_R}$, which assess the quality of the predicted signals without considering associated tokens:

$$\begin{aligned}
\mathbf{S_P}(G, G') &= \frac{\text{SUM}(\{\#(\hat{\zeta} \cap \zeta) \mid \hat{\zeta} \subseteq \hat{Z} \wedge \zeta \subseteq Z \wedge \forall z \in \zeta[\forall \hat{z} \in \hat{\zeta}[\hat{z}.e = z.e \wedge \hat{z}.\lambda = z.\lambda]]\})}{\#\hat{Z}} \\
\mathbf{S_R}(G, G') &= \frac{\text{SUM}(\{\#(\hat{\zeta} \cap \zeta) \mid \hat{\zeta} \subseteq \hat{Z} \wedge \zeta \subseteq Z \wedge \forall z \in \zeta[\forall \hat{z} \in \hat{\zeta}[\hat{z}.e = z.e \wedge \hat{z}.\lambda = z.\lambda]]\})}{\#Z}
\end{aligned} \tag{4}$$

20 Some parsers relax the requirement that S be given at prediction time, but most assume S as an input—i.e., the parser receives gold EDUs and segmentation is considered a separate task.

21 We may add metrics for recall/F1, but these would only differ from precision if the assumption of tree projectivity were dropped.

22 We will consider two edges equal if the terminal vertices of both the source and the target nodes are identical, i.e., they span the same EDUs.

23 This label may be decomposed into λ_1, λ_2, etc., if desired; we use a single label item here for simplicity, e.g., "semantic:lexical_chain" can be a monolithic signal type.

Informally, these equations group signals by the combination of their label and associated edge. This group is computed for both the gold and predicted tree, and the overlap between the gold and the predicted group for the label–edge combination is noted. The size of this overlap is summed across all groups, and the sum is divided by the total number of predicted signals (precision) or gold signals (recall). This slightly complicated formulation of precision and recall is necessitated by the fact that there could be potentially many signals which share the same edge and label (but then not the same tokens, which are however ignored in the above metric).

We would also like to have a quantitative *anchored* metric of how well the parser performed at predicting the actual tokens (word forms) associated with a signal. Due to the complication just noted above, formulating precision and recall metrics is not entirely straightforward. Consider the case where for some given edge e and signal label λ, the gold tree has some signals $\zeta \subseteq Z$ and the predicted tree has some signals $\hat{\zeta} \subseteq \hat{Z}$. It is possible that even if all signals in $\hat{\zeta}$ and ζ are associated with edge e and are labeled λ, ζ may not be equal in size to $\hat{\zeta}$. Moreover, even if they are the same size, it is not immediately clear how to put signals from the two sets into pairwise correspondence so that their tokens may be compared.

Our solution is to allow the evaluation procedure to find the optimal pairing for predicted and gold signals for each label–edge signal group. While in principle this is an expensive operation with computational complexity $\mathcal{O}(Z!)$, we expect that a label–edge group in the 99th percentile by size would contain no more than 5 signals, and moreover, heuristics would likely be able to make the optimal pairing search more efficient as needed. Let us therefore define an $\textsc{optimal-pair}(\lambda, e, Z, \hat{Z})$ function which makes label–edge groups from the signal sets and produces a set of optimal pairings $\{\langle z_1, \hat{z}_1\rangle, \ldots, \langle z_m, \hat{z}_m\rangle\}$ that maximizes the total number of overlapping tokens across the two label–edge groups. With the assistance of OPTIMAL-PAIR, we may now define signal token precision $\mathbf{W_P}$ and recall $\mathbf{W_R}$ as follows:

$$\begin{aligned}
\mathbf{W_P}(G, G') &= \frac{\textsc{sum}(\{\#(\hat{z}.\tau \cap z.\tau) \mid \lambda \in \hat{\Lambda} \wedge e \in \hat{E} \wedge \langle z, \hat{z}\rangle \in \textsc{optimal-pair}(\lambda, e, Z, \hat{Z})\})}{\textsc{sum}(\{\#\hat{z}.\tau \mid \hat{z} \in \hat{Z}\})} \\
\mathbf{W_R}(G, G') &= \frac{\textsc{sum}(\{\#(\hat{z}.\tau \cap z.\tau) \mid \lambda \in \Lambda \wedge e \in E \wedge \langle z, \hat{z}\rangle \in \textsc{optimal-pair}(\lambda, e, Z, \hat{Z})\})}{\textsc{sum}(\{\#z.\tau \mid z \in Z\})}
\end{aligned} \tag{5}$$

Intuitively, think of these metrics as proceeding in the following way: For every relation in the predicted and gold trees, group signals such that each group shares some edge e and some signal label λ. (Perhaps, for example, the signals all share the label DM.) Note that (unlabeled) edge correctness is a prerequisite for signals to be deemed correct, since associated edges are identified by the EDU yield of their source and target nodes. Now, for each signal label–edge group, find an optimal pairing between predicted and gold signals such that overlap in each pair's associated tokens is maximized. Finally, count the number of overlapping tokens across all pairs, and divide by the total number of token associations in either the predicted (precision) or gold (recall) signals.

For secondary edges, the four Parseval metrics are directly applicable: a secondary edge has all the salient properties of a primary edge, although we note that the term "nuclearity" should properly be replaced by "direction," since secondary edge source and target designations imply only a direction and not necessarily a higher prominence for the target vis-a-vis the source (such prominence is only represented via the primary tree, to maintain RST's unambiguous nuclearity property). The only outstanding question for how to perform Parseval evaluation for eRST trees, then, is how to combine metrics

that are respectively computed for the primary and secondary edges. We expect that in general it could be useful to consider both in isolation and also to consider the metrics for both kinds of edges combined. However, in the latter case, since secondary edges are rare, they would not change a metric pooling both very much, so in the interest of space we report on each type separately in our experiments below. We publicly release our scorer with the code and data for this article.

5.2 Model Architecture

Although the main objective of this article is to present eRST as a framework, and providing a comprehensive NLP system for its parsing is outside of the current scope, we present an initial experiment in eRST parsing in this section, extending existing architectures. Conventional RST parsers take either a top–down or a bottom–up approach: Top–down begins with the entire document and decomposes it recursively into sections, which may coincide (or be forced to coincide) with paragraphs, sentences, and so forth (Feng and Hirst 2014; Kobayashi et al. 2020). Recent approaches rely on end-to-end neural architectures: The DMRST parser (Liu, Shi, and Chen 2021) used a pointer network as its decoder and maintains a stack by top–down, depth-first span splitting; Zhang, Kong, and Zhou (2021) utilized adversarial learning to distinguish gold versus incorrect trees.

Bottom–up approaches are perhaps closer to human RST annotation practices (Shen et al. 2022), beginning by connecting related clauses and sentences, then larger structures. This approach currently wins on span and nuclearity identification scores, but not on relation classification. Guz, Huber, and Carenini (2020) provided a transition-based neural shift-reduce parser using SpanBERT embeddings and Yu et al. (2022) proposed a second EDU-level pretraining on top of sentence-level training for next EDU prediction as well as discourse marker prediction.

After briefly considering ways of implementing novel end-to-end approaches to the task above, we quickly realized that substantial additional research would be needed in order to not only add model components to predict signals and secondary edges, but also to perform at near-state of the art (SOTA) levels for primary tree parsing. eRST involves aspects not only of the RST parsing literature surveyed above, but also of connective detection (see Yu et al. 2019; Gessler et al. 2021; Metheniti et al. 2023 for recent work), explicit and implicit relation recognition (Rutherford, Demberg, and Xue 2017; Dai and Huang 2018; Kim et al. 2020; Scholman et al. 2021), and discourse relation classification (Liu, Fan, and Strube 2023), which remain challenging even for recent neural models (Qin et al. 2017; Kurfalı and Östling 2021; Braud et al. 2023). As a starting point, we therefore decided to adapt existing SOTA models for predicting primary trees and explicit connectives, and to construct a baseline approach on top of those systems.[24] Our approach consists of the following four components:

Primary Tree Parsing. After testing several off-the-shelf parsers, we chose the top–down DMRST (Liu, Shi, and Chen 2021), which remains SOTA for the RST Relation metric for GUM and is efficient and easy to run. The system produces projective, binary, labeled

24 We do not mean to say that a unified, end-to-end approach to the task is a bad idea: Our approach is merely motivated by the observation that our initial attempts to do so resulted in unusable primary tree prediction accuracy. We believe there is great potential in jointly learning the related subtasks in eRST, similarly to successful work in multitask learning and pretraining for RST parsing (Braud, Plank, and Søgaard 2016).

trees, which we use as an input for signal prediction and association, as well as the basis for the available non-terminal nodes for secondary edge prediction. For comparison, we also provide numbers using the best bottom–up shift-reduce-style parser from Guz and Carenini (2020) in the next section.

Connective Detection. We use DisCoDisCo (Gessler et al. 2021), the highest scoring connective detection system on the DISRPT benchmark. The system is trained on the eRST training set's contiguous DM and orphan token spans, with discontinuous spans split into two BIO-encoded connective instances. This means that discontinuous connectives (in accordance with PDTB's definitions, e.g., *if...then*) must be re-merged later, based on a closed list of discontinuous items attested in the training data. We also note that non-DM signals cannot be predicted using this system, since they often overlap. A DM-lexicon-based baseline is also provided for comparison in the next section.

Morphosyntax and Coreference. We use the AMALGUM pipeline (Gessler et al. 2020), which is designed to predict the same annotations present in GUM, including UD parses, entity annotations, and coreference resolution. For testing we then use the same pre-processing scripts that feed the manual annotation for the eRST corpus described in Section 4.1, except with predicted, rather than gold standard, syntax trees and coreference, which can then be used to predict morphological, syntactic, semantic, and reference signals, and with no manual correction.

Association and Secondary Edges. Here we propose a new transformer-based text classifier, which receives two text spans known to be connected by a relation (based on the input primary tree), one of which contains a DM. The system predicts whether the relation between the spans is signaled by the DM, which is marked in the input by surrounding "**" characters. The spans are either the head EDUs of the relation (for intra-sentential relations) or the two sentences containing the head EDUs (for inter-sentential relations). We further embed the relation label, the distance between head EDUs and the direction of the relation in the input, and, for secondary edges, the relation label of any primary edge with the same source and target of the secondary relation, if available. The serialization is exemplified in (14).

(14) ANTITHESIS (LIST) *left 1*: past studies have tended to avoid this task » and have ****instead**** used samples of researchers

In (14), the input suggests that a left-to-right secondary ANTITHESIS edge may exist between the given textual units, which are adjacent (direction and distance: *left 1*), for which a primary LIST edge already exists, and which is marked by the orphan "instead." Note that the DM "and," which also appears in the example (in fact, it is the DM for the primary LIST relation), is not targeted, as implied by the "**" notation, which singles out the word "instead."

In order to predict secondary edges at test time, the system also generates secondary edge prediction candidates for all primary edge paths attested in the predicted input tree, for all relation labels compatible with any predicted DM they contain, as well as relations between any adjacent pair of sentences, again provided that they contain a compatible DM. This compatibility is based on a DM-to-relation mapping obtained from the training data. Finally, the system ranks edges by binary classification probability

Table 8
eRST graph metrics for our system with different inputs (LX = Lexicon-based connective detection; DD = DisCoDisCo connective detector; AM = UD trees and coreference from the AMALGUM parser; DMRST and G&C are the top–down and bottom–up SOTA RST parsers cited above).

		primary				secondary			
RST	**NLP**	**S**	**N**	**R**	**F**	**S**	**N**	**R**	**F**
gold	gold	1.000	1.000	1.000	1.000	0.389	0.270	0.205	0.184
gold	LX+AM	1.000	1.000	1.000	1.000	0.210	0.142	0.113	0.091
gold	DD+AM	1.000	1.000	1.000	1.000	0.369	0.256	0.195	0.174
DMRST	LX+AM	0.620	0.545	0.492	0.482	0.055	0.044	0.027	0.022
DMRST	DD+AM	0.620	0.545	0.492	0.482	0.101	0.061	0.030	0.030
G&C-RST	LX+AM	0.595	0.530	0.470	0.457	0.022	0.016	0.022	0.016
G&C-RST	DD+AM	0.595	0.530	0.470	0.457	0.055	0.037	0.037	0.028

and chooses the top possible relation to associate with each input DM (as predicted by DisCoDisCo). If the predicted relation is secondary then the DM is classified as an orphan, and the secondary edge is added to the graph.

For the transformer embeddings we tested several options, trained on all true examples in the training set, enriched with an equal number of negative examples, and halting on dev set performance for early stopping. We compared base-sized versions of BERT, DeBERTa v3, XLNet, and Electra, and chose Electra-base-discriminator as the highest performing model on the dev set.

5.3 Results

Table 8 gives results for eRST graph structures in seven scenarios. In the first row, we provide gold primary trees, syntax trees, and connective positions, and the transformer model only predicts secondary edges and signal associations (*RST = gold, NLP = gold*). This is an upper bound for the system performance, when no cascading errors from other components affect its accuracy.

In the other scenarios, we use AMALGUM tools for automatic syntax parsing and coreference resolution, and vary how primary trees and DMs are predicted. RST trees are either gold, or predicted using one of two RST parsers: the state of the art top-down parser DMRST (Liu, Shi, and Chen 2021) and, for comparison, the slightly less accurate, best bottom–up parser from Guz and Carenini (2020) (abbreviated G&C). DMs are predicted either using a DM lexicon as a baseline (LX), or DisCoDisCo (DD), the SOTA system for DM detection. For the lexicon-based DM detection baseline we simply create a lexicon containing any string which is a DM in the training set more than 50% of the time and assume that it should always be predicted to be a DM, regardless of context (including multi-token DMs).

In the bottom four scenarios in the table, only EDU segmentation and word-tokenization are given as inputs.[25] All tools are trained on the official GUM V9 training partition (165 documents), using the development partition for early stopping

25 If these are not provided, numbers become very hard to compare due to segmentation conflicts; however we assume that both tasks can be performed automatically with high accuracy in production settings.

(24 documents) and the test set for the final scores (24 documents). As the table shows, secondary edge prediction is challenging even when gold RST trees and NLP preprocessing are given, especially for the FULL metric. This is because correct prediction of a secondary edge requires the identification not only of a discourse relation (e.g., that two parts of the text stand in CONTRAST), but also that there is not already a primary edge corresponding to the relation, and that there is a sufficient trigger, such as an orphan DM or syntactic environment allowing for the secondary edge. Even with this information correctly recognized, the system must still choose the correct attachment points for the edge source and target in the hierarchical tree, as well as the edge direction and the label. Seen from this perspective, and considering the little training data available (fewer than 1K secondary edges), the SPAN score of 0.389 is actually rather high, while the Relation score of 0.205 is not much less than half the R score of a primary predicted parse 0.492 (using DMRST, rows 4–5). This is despite the fact that primary parses gain score from easy wins, such as correct attachment of relative clauses and other explicit intrasentential relations—secondary edges can be expected to be trickier cases.

Turning to the impact of predictions by baselines or previous SOTA tools, we see that automatically predicted NLP, including connective detection, does not produce substantial degradation in secondary edge predictions when DD is used, since the gold primary tree is still just as useful in determining whether a secondary edge is missing given existing primary ones, and connective detection is a relatively high performance task, often scoring over 90% for English (Braud et al. 2023). Using the LX baseline produces a very substantial degradation of almost half the score (second row). Predicted syntax trees could mainly impact prediction of syntactically motivated secondary edges (missing ATTRIBUTIONS from complement clauses, or ELABORATION relative clauses), yet these are not only rare, but also easy to predict correctly using a SOTA syntax parser.

The situation in the last four rows is very different: Switching to predicted RST trees is catastrophic for secondary edge prediction, since, even if a relation missing from the primary tree is recognized, it could very well be an error in the primary parse: If the secondary edge detector correctly identifies and adds a real relation, the score will actually be impacted worse if that relation was a primary one in the gold data, since the detector then incurs both a precision and a recall error. Using a slightly less good parser does not matter as much as it does for the primary tree, but still degrades the Full metric (F). Switching to the LX baseline for DM detection is unsurprisingly catastrophic, especially when compounded with automatic primary RST parsing.

Moving on to the second part of the eRST graph prediction task, Table 9 shows performance on signal detection (identifying the signal types associated with each relation in the graph) and signal anchoring (also identifying the exact token span of each signal) broken down by major signal types, in the same scenarios. In each predicted scenario, the same scripts are used to identify the non-DM signal types for which automatic prediction is feasible, but the inputs are changed, for example, a syntactic relative clause signal is still predicted based on the syntax annotation, but in the predicted syntax setting, it uses automatic dependency parses, unlike in the gold data which we release with this article.

As the table shows, here too DM results are quite good as long as the gold RST tree is provided and DD is used; the LX baseline produces substantially worse numbers for DMs/orphans and overall. With predicted primary RST trees, DM and orphan identities can again be swapped (if a primary/secondary relation pair are swapped in the prediction, what should be an orphan becomes a DM and vice versa), and in general, orphan prediction is challenging, since it only has a chance of being correct if

Table 9
eRST signal type detection and anchoring scores per signal category.

		SIGNAL DETECTION									
RST	**NLP**	**all**	**dm**	**orphan**	**graph**	**morph**	**num**	**lex**	**sem**	**ref**	**syn**
gold	gold	0.925	0.915	0.176	1.000	0.936	0.429	0.992	0.845	0.991	0.989
gold	LX+AM	0.756	0.724	0.080	0.886	0.957	0.0	0.870	0.686	0.802	0.961
gold	DD+AM	0.824	0.915	0.188	0.886	0.957	0.429	0.881	0.687	0.802	0.961
DMRST	LX+AM	0.450	0.351	0.030	0.416	0.500	0.0	0.435	0.271	0.137	0.838
DMRST	DD+AM	0.483	0.433	0.044	0.416	0.500	0.286	0.436	0.272	0.137	0.838
G&C-RST	LX+AM	0.431	0.334	0.005	0.408	0.571	0.0	0.411	0.242	0.119	0.822
G&C-RST	DD+AM	0.459	0.398	0.010	0.408	0.571	0.0	0.414	0.242	0.119	0.822
		SIGNAL ANCHORING									
gold	gold	0.915	0.889	0.147	1.000	0.871	0.429	0.994	0.882	0.994	0.944
gold	LX+AM	0.555	0.679	0.055	0.972	0.900	0.0	0.837	0.459	0.537	0.898
gold	DD+AM	0.591	0.886	0.159	0.970	0.900	0.429	0.852	0.459	0.537	0.898
DMRST	LX+AM	0.298	0.331	0.017	0.567	0.386	0.0	0.416	0.138	0.088	0.786
DMRST	DD+AM	0.314	0.422	0.030	0.564	0.386	0.286	0.418	0.137	0.088	0.786
G&C-RST	LX+AM	0.290	0.308	0.000	0.577	0.480	0.0	0.389	0.118	0.086	0.780
G&C-RST	DD+AM	0.304	0.386	0.000	0.577	0.480	0.0	0.392	0.118	0.086	0.780

the secondary edge was predicted correctly as well. We can also see that the penalty for switching to G&C as the RST parser is fairly limited, but noticeable, mainly for DMs.

For non-DM signals too, predicted primary trees mean that the required relation for alignment may often not exist. This is especially clear for "easy" signal types, such as graphical ones, which include unambiguous punctuation and layout factors, such as bullet points marking a LIST relation—if the structure of a LIST is predicted correctly, signal identification may be trivial, but an incorrect parse leads to a signal detection error as well.

Beyond these findings, we note that some signal types are challenging to get right even for gold trees, such as *numerical* signals, which require matching numerical expressions to quantities of things mentioned, or *morphological* ones, such as sequence of tenses. The latter signal type is predicted for any sequential temporal relation when units in sequence have succeeding tenses (past then present, present then future, etc.), but these morphological cues somewhere within the span of a sequence of events do not always indicate the sequence itself, as shown in Example (15), where a present tense direct speech predicate is uttered after a past tense narrative sentence, but the tense change is not actually a signal of the sequence.

(15) [I **pulled** the bike to a halt (…)] ["I **think** I've got a fairy stuck up my nose.."]$_{<sequence>}$

NLP prediction quality, too, can matter considerably, even when gold RST trees are provided, for types such as *reference* and *semantics*, since automatic coreference resolution substantially underperforms the gold coreference information delivered in the gold NLP scenario. The most reliably predictable signal type is unsurprisingly *syntactic*, for two reasons: (1) it depends on form-based NLP inputs for which reliable tools exist (syntactic dependency parsing), and (2) it is associated with some of the easiest relations to infer in the RST tree: Relative and other adnominal clauses, which both RST parsers usually parse correctly.

In sum, these results demonstrate that while eRST parsing is a challenging task, it is not a hopeless one, especially in an era in which computational linguistics takes on increasingly complex tasks—our system is a very rough proof of concept, and we are certain that better ones can be developed, even with current base-sized LMs, let alone much larger sequence-based LLMs. At the same time, it is clear that signal detection and secondary edge prediction is primarily feasible if we are confident we have the right primary tree: Without that tree, scores on the remaining tasks suffer from cascading errors very substantially. The same applies to a lesser extent to DM detection: With a better system for this subtask, scores on signals and secondary edges will rise, as made clear by the comparison between the LX baseline and DD.

6. Applications

Although eRST parsing will require further research before we can expect to leverage reliable automatic analyses for practical applications, it is worth considering what information the formalism exposes and how it could be used in practice. Since eRST graphs can easily be reduced to primary unsignaled RST trees, it goes without saying that they provide the same benefits as those trees, for example, proposition extraction (=EDU segmentation); a built-in, recursive ability to extract the most prominent units in any document (or subspan) for extractive summarization, central discourse unit identification (Atutxa et al. 2019), topic segmentation (Xing, Huber, and Carenini 2022) or related tasks; and identification of specific relations of interest (e.g., parsing all speeches of a public figure or political party and extracting all CONCESSION relations made by them for inspection), which can also be used for representation learning in downstream tasks (Huber and Carenini 2022; Pu, Wang, and Demberg 2023). Since relation spans and associated discourse markers are exposed by the graph, it is also possible to extract shallow discourse parses and use them to disambiguate connective senses or perform other tasks relying on shallow parsing, such as sentiment analysis or opinion mining. In fact, we are planning to leverage the information in eRST to generate training data compatible with current shallow discourse parsing frameworks as an additional resource in appropriate formats.

However, eRST graphs go beyond these original applications of discourse parsing to allow for additional, more fine-grained applications, for which we provide some examples here.[26] In this section we would like to start by considering how much added value eRST brings to the core application of RST, namely, relation extraction, before considering some of its more novel applications and implications.

Relation Extraction Compared to RST. At the most basic level, the addition of secondary edges allows analysts to represent multiple concurrent or tree-breaking relations without having to choose a single function label per unit in a text. Although secondary edges are comparatively rare, they are not evenly distributed across labels, and for some label types of interest, they may constitute over 10% of relation instances. Table 10 gives counts and proportions for secondary edges by relation label for labels that have >5

26 A reviewer has asked whether LLMs make such analyses redundant in practice. We do not believe so, for at least two reasons: (1) LLMs benefit from a variety of annotated data types for pre-training and instruction fine-tuning, and eRST data could be used to generate such supervision; and (2) identifying eRST relations with their associated signals is an end-task in itself, which can serve human analysts, e.g., in computational social science (analyzing political speech), quantitative and qualitative humanities research on rhetoric, and more, while offering evidence in support of the relations identified in each text.

Table 10
Proportions of secondary edges for relations with >5 secondary instances.

relation	primary	secondary	total	% secondary
causal-result	437	72	509	14.10%
explanation-justify	455	60	515	11.70%
adversative-concession	759	87	846	10.30%
mode-manner	273	30	303	9.90%
adversative-antithesis	375	35	410	8.50%
causal-cause	588	39	627	6.20%
elaboration-additional	2,326	136	2,462	5.50%
joint-sequence	1,868	87	1,955	4.50%
contingency-condition	446	19	465	4.10%
elaboration-attribute	2,182	85	2,267	3.70%
context-circumstance	968	35	1,003	3.50%
explanation-evidence	729	26	755	3.40%
adversative-contrast	887	28	915	3.10%
joint-other	1,866	51	1,917	2.70%
joint-list	3,707	88	3,795	2.30%
joint-disjunction	305	6	311	1.90%
attribution-positive	1,335	24	1,359	1.80%
restatement-partial	370	5	375	1.30%
context-background	1,071	7	1,078	0.60%

secondary instances in our data, in descending order of secondary edge percentage. The proportions give an idea of the extent of information an eRST parse gains (or a plain RST parse misses) out of the total possible relations recognized in our formalism. As the numbers show, primary trees alone miss substantial amounts of labels like CAUSAL-RESULT, EXPLANATION-JUSTIFY, and ADVERSATIVE-CONCESSION.

Signal-based Relation Subtypes. In addition to the word-sense disambiguation provided by associating DMs with relations (we can tell a temporal *since* from a causal one, as in PDTB), signals can be used to identify subclasses of relations which are more fine-grained than the two-level taxonomy used in our data. For example, RST-DT distinguishes a fine-grained relation OTHERWISE (normally collapsed under the coarse CONDITION class) which is not distinguished in GUM. However, using our anchored signals, it is easy to extract all relations marked by OTHERWISE and obtain their exact satellite and nucleus scopes. It is also possible to define subtypes not found in any other datasets. For example, the non-conditional explanatory *if* found in "I have oregano **if** you want any" characterizes a subtype of EXPLANATION-JUSTIFY relation, which can be retrieved directly using the relation and discourse marker combination, as shown in Figure 12 with the discourse marker in red.

Non-DM signals can also be used to extract subclasses of relation instances, such as identifying temporal relations signaled by explicit date or time expressions, ELABORATIONs discussing meronyms of a nucleus entity, or CONTRAST signaled via antonyms. In all of these cases, explicit signal annotations allow us to access sub-categories of relations, and even extract specific, open-class words, which expose more fine-grained semantic and pragmatic information. The potential of these possibilities is further enhanced by the presence of secondary relations, which can be queried concurrently

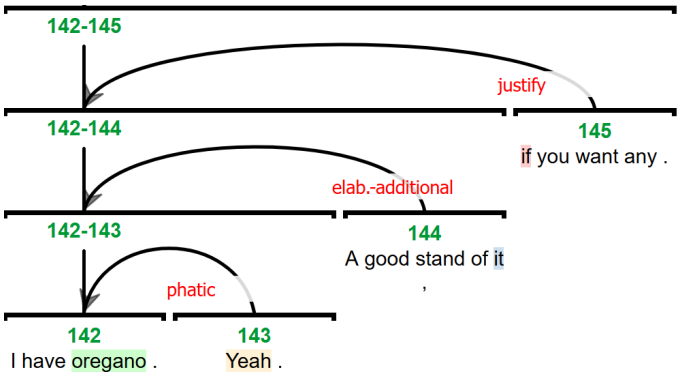


Figure 12
An EXPLANATION-JUSTIFY relation marked by *if* belongs to a special class of justifications which can be retrieved using signals, despite not having a dedicated relation label.

to primary ones (finding all SEQUENCE relations which are also an ELABORATION, or excluding cases that have a concurrent CONTRAST relation, etc.).

Attribution Scope, Source, Mode, and Polarity. Since attributions and their polarity are already identifiable using the relations ATTRIBUTION-POSITIVE and ATTRIBUTION-NEGATIVE (e.g., "officials did not say..."), and since the constituent tree expresses scope, RST data already exposes the span of positively or negatively attributed content. However, the addition of signals for *attribution source* and the *indicative word* (or *phrase*) instantiating the attribution predicate allow us to extract full information on the mode of attribution: Via a speech verb such as *say* or cognitive predicate such as *think*, or no predicate at all in "newspaper style" attribution giving just a quote and a name. The source of the attribution can correspond to a named or non-named entity. Figure 13 illustrates the information exposed by the formalism for a comprehensive extraction of attributions and their components: The *attribution source* is marked in green and lexical predicate signal in cyan.

In the case of the multilayer GUM corpus, the existence of aligned lemmatization, entity recognition, entity linking (Wikification), and coreference resolution layers allow

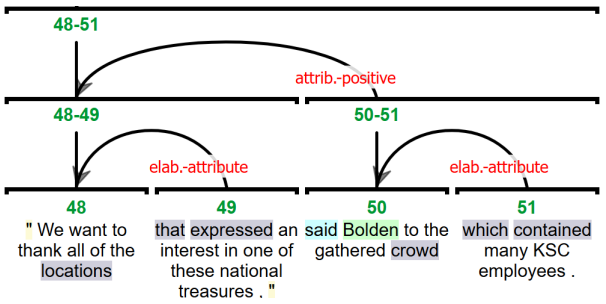


Figure 13
Attribution with anchored signals representing the attribution predicate in cyan and the attribution source in green.

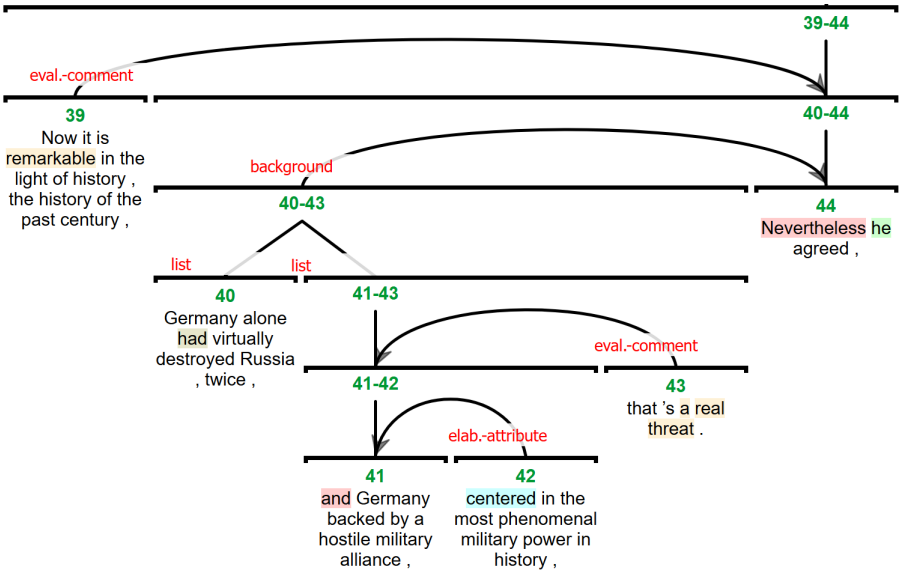


Figure 14
Two evaluations whose scope is marked by the tree, with the evaluation content signaled using indicative words and phrases: *remarkable* and *a real threat*.

us to link attributions not only to an attested verb form (e.g., *said* in the figure, indicating the mode of attribution) and entity span (e.g., *Bolden*), but also to the predicate lemma (*say*, substantially reducing the number of attribution mode predicate types) and the underlying entity identifier if available (e.g., coreference cluster 3, the cluster containing a mention *NASA administrator Charles Bolden*, and also linked to the Wikipedia identifier `https://en.wikipedia.org/wiki/Charles_Bolden`).

Evaluation Content. Similarly to attributions, the relation EVALUATION-COMMENT provides scope for what is being evaluated. However in a regular RST tree, it is not possible to know whether the evaluation is negative or positive (or neither), nor what evaluative terms were being used with respect to the content. The eRST graph improves on this whenever an *indicative word* (or *phrase*) is available, as illustrated in Figure 14 in yellow.

The signaling annotations combined with the tree itself allow us to know that the entire span of units 40–44 is being evaluated as *remarkable*, while a nested evaluation from 43 scoping over 41–42 speaks of *a real threat*. This information is substantially more detailed than what a basic RST tree can provide. Our data indicates that for 62.3% of EVALUATION relations, an associated indicative item is available, amounting to nearly 1,200 tokens with over 200 lemma types. The most common items are in Table 11. Although top adjectives like "good" and intensifiers like "very" dominate the top of the table (amounting to nearly 14% of tokens), frequencies quickly drop to around 1%, reaching a frequency of 5 at rank 56, and single attestations (a.k.a. hapax legomena) at rank 145. This shows the long tailed distribution of evaluative items, which are much less predictable or limited compared with DMs.

Reliability and Explainability. A major challenge for current neural NLP systems lies in reliability and explainability: When predicting structured outputs without a rationale,

Table 11
Frequency-ranked lemmas in lexical indicative signals for EVALUATION.

rank	lemma	frequency	% signals
1	good	90	0.076
2	very	74	0.062
3	feel	57	0.048
4	great	30	0.025
5	important	27	0.022
6	look	27	0.022
7	bad	23	0.019
8	seem	23	0.019
9	mean	22	0.018
10	beautiful	20	0.016
	...		
56	fundamental	5	0.004
	...		
145	inaccurate	1	0.001

downstream applications and users have little way of knowing which predictions are likely to be correct or incorrect, what the rationale is for the prediction, and what we could do to filter out mistakes or improve systems. A complete automatic parse in eRST includes a built-in rationalization mechanism in the form of signals, which can be used for filtering (only use explicitly-signaled relations, or just ones signaled by a DM) and to better understand the predictions being outputted. Even though signal predictions can of course be wrong in themselves, especially explicit connective detection is now a fairly reliable task, and can be used by human analysts to better understand discourse parsing outputs, or as part of the input for downstream tasks which should be made aware of the strength and type of evidence for a system's predictions. By contrast, investigating cases of totally unsignaled relations in gold annotated data can help us to understand the limitations of our signaling annotation scheme, and try to address how human analysts arrive at an analysis in the absence of instances of anchorable, or even any signals of any kind.

7. Conclusion

In this article we presented eRST, a comprehensive theoretical framework representing discourse relations and structure, which expands on the existing Rhetorical Structure Theory, but incorporates insights from previous work in alternative frameworks, such as PDTB and SDRT. In particular, our proposal addresses weaknesses in RST, such as inability to handle tree-breaking and potentially multiple concurrent relations, as well as the failure to address the role of discourse relation marking devices, including, but not limited to, connectives. Going beyond PDTB's model, which is focused on morphosyntactically defined connectives and some additional highly constrained marker types, eRST adopts the view of the RST Signaling Corpus (Das and Taboada 2017) by aspiring to a more exhaustive inventory of discourse relation signals, which we attach directly to relevant tokens in each text. The resulting representation retains advantages of RST, such as a strong commitment to recursive nuclearity and a hierarchical tree structure spanning entire documents, while enabling a more complete analysis covering

previously disregarded relations, as well as the rationale for their identification in text-based terms.

Beyond the potential of eRST data to support more detailed theoretical studies of how discourse structure and meaning are constructed in natural language, we have also demonstrated some of the potential practical applications of eRST. These include not only classic uses of RST, such as searching for discourse relations (e.g., finding CAUSE and RESULT in a parsed collection of texts, or EVIDENCE for a particular claim etc., now including tree-breaking cases), Central Discourse Unit detection, or extractive summarization, but also unique possibilities supported by the availability of signal annotations. The latter include tasks such as detailed ATTRIBUTION extraction, analysis of components of EVALUATION relations, and more.

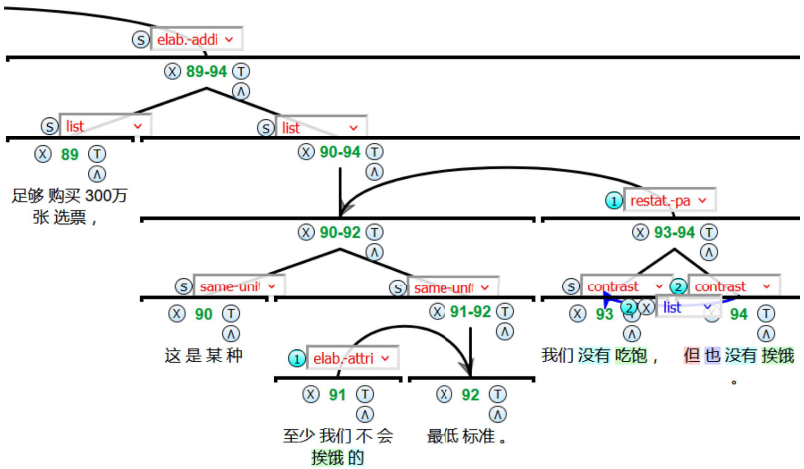
The introduction and definition of a parsing task for eRST, including revised evaluation metrics, a publicly available scorer, and a baseline implementation, provide a new and more comprehensive benchmark for discourse parsing, enriched by the presence of a more complete representation of the total relations available in each document. In particular, we hypothesize that the explicit presence of multiple concurrent relations in our data can shed more light on at least a subclass of parser errors in traditional RST parsing, in cases where parsers fail to predict the primary tree relations, but turn out to predict relations encoded as secondary ones, making errors that are not entirely wrong. We note that some recent initial work in this direction using double annotations in RST-DT seems to suggest that this hypothesis is correct (Liu, Aoyama, and Zeldes 2023).

Additionally, we see potential for using data with the rich annotations present in eRST in probing and improving current LLMs, which can harness textual representations of the relations and signals exposed in our discourse parses for either pre-training tasks or for instruction fine-tuning. Zero or few-shot successes and failures in solving eRST tasks may also teach us about what specific LLMs do more or less well, and what levels of discourse awareness they possess. Conversely, we are optimistic that LLMs can be used to predict aspects of eRST parses, or eventually even complete parses, as some recent work on relation classification using sequence to sequence models has indicated (Anuranjana 2023; Chan et al. 2023). Parsers built on top of LLM outputs may allow us to analyze larger datasets using increasingly accurate automatic parses, and to bootstrap data to tackle difficult, out-of-domain, or perhaps even multilingual scenarios in which manually annotated discourse parsing data is scarce.

Finally, we believe that the data and tools released with this article represent a substantial resource for research. The GUM corpus is now larger than the seminal RST-DT corpus for English (Carlson, Marcu, and Okurowski 2001), with 12 written and spoken genres, showing the applicability of the framework to a broad range of text types. Since GUM continues to grow and cover new genres, we anticipate challenges but also opportunities in applying eRST to new kinds of data. We are also keenly aware of the limitation of richly annotated corpora primarily to English, and hope to be able to extend eRST to more languages in the future, with obvious first targets in the languages that already have RST treebanks which could be extended with eRST—for example, the Georgetown Chinese Discourse Treebank (GCDT) (Peng, Liu, and Zeldes 2022) follows the same RST annotation scheme as GUM, and many of the tools and scripts used for this article could be adapted to enrich it with relative ease (see Appendix B). We hope that the release of the new eRST annotated GUM will encourage others to experiment with the framework and tools, and invite researchers working on discourse relations and representations to test the theory and provide feedback to evolve it further. For updates, annotation samples, and discussion we also refer interested readers to the eRST website at `https://gucorpling.org/erst`.

Table A.12
Relation labels in the GUM corpus.

relation name	nuclearity	definition
ADVERSATIVE-ANTITHESIS	→←	R is meant to prefer N as an alternative to S
ADVERSATIVE-CONCESSION	→←	R is meant to look past an incompatibility of N with S
ADVERSATIVE-CONTRAST	Λ	W presents multiple Ns as incompatible, but of equal prominence
ATTRIBUTION-NEGATIVE	→←	S states that a potential source is NOT a source of the information in N
ATTRIBUTION-POSITIVE	→←	S states a source for the information in N
CAUSAL-CAUSE	→←	S is the cause of N and N is more prominent)
CAUSAL-RESULT	→←	S is the result of N or: N is the cause of S, and N is more prominent)
CONTEXT-BACKGROUND	→←	S provides information to increase R's understanding of N
CONTEXT-CIRCUMSTANCE	→←	S details circumstances often spatio-temporal) under which N applies
CONTINGENCY-CONDITION	→←	N occurs depending on S
ELABORATION-ADDITIONAL	←	is used in all other cases, when S is an elaboration on N as a whole
ELABORATION-ATTRIBUTE	←	is used when S elaborates on a participant within N, rather than on the entire proposition in N
EVALUATION-COMMENT	→←	S provides an assessment of N by W R does not have to share this assessment)
EXPLANATION-EVIDENCE	→←	S provides evidence which increases R's belief in N
EXPLANATION-JUSTIFY	→←	S increases R's acceptance of W's right to say N
EXPLANATION-MOTIVATION	→←	S is meant to influence R's willingness to act according to N
JOINT-DISJUNCTION	Λ	W presents multiple Ns which can be regarded as interchangeable alternatives
JOINT-LIST	Λ	W presents multiple Ns in parallel which can be regarded as additive to one another
JOINT-OTHER	Λ	any other collection of unlike discourse units of equal prominence at the same level of hierarchy
JOINT-SEQUENCE	Λ	Multiple Ns form a temporally ordered sequence of events in order
MODE-MANNER	→←	S indicates the manner in which N happens
MODE-MEANS	→←	S indicates the means by which N happens
ORGANIZATION-HEADING	→	explicit text organizing device such as a heading
ORGANIZATION-PHATIC	→←	W holds the floor, without contributing propositional content
ORGANIZATION-PREPARATION	→	covers all other forms of S units primarily used to signal an upcoming N
PURPOSE-ATTRIBUTE	→←	is used when S gives the purpose of a participant in N, rather than on the entire proposition in N
PURPOSE-GOAL	→←	the proposition in N as a whole is initiated or exists in order to realize S
RESTATEMENT-PARTIAL	←	S partly realizes the same role and content as a previous N
RESTATEMENT-REPETITION	Λ	Multiple Ns realize the same role and content
SAME-UNIT	Λ	indicates a discontinuous discourse unit this is not a discourse relation)
TOPIC-QUESTION	→	N is the answer to the question posed by S
TOPIC-SOLUTIONHOOD	→←	N is a solution to a problem presented by S



EDU	Chinese	English translation
89	足够购买300万张选票，	enough to buy 3 million votes
90	这是某种	this is some kind of
91	至少我们不会挨饿$_{\text{SEM_REP}}$ 的$_{\text{SYN}}$	that$_{\text{SYN}}$ at least we won't starve$_{\text{SEM_REP}}$
92	最低标准。	minimum standard.
93	我们没有$_{\text{SEM_CHAIN}}$ 吃饱$_{\text{SEM_ANT}}$ ，	We are not$_{\text{SEM_CHAIN}}$ full$_{\text{SEM_ANT}}$,
94	但$_{\text{DM}}$ 也$_{\text{ORPHAN}}$ 没有$_{\text{SEM_CHAIN}}$ 挨饿$_{\text{SEM_REP, SEM_ANT}}$ 。	but$_{\text{DM}}$ we are not$_{\text{SEM_CHAIN}}$ starving$_{\text{SEM_REP, SEM_ANT}}$ either$_{\text{ORPHAN}}$.

Figure B.15
GCDT example *gcdt_interview_falkvinge*.

Appendix A. Relation Labels in GUM

Table A.12 gives the full list of relation labels in GUM. Note that SAME-UNIT is not a proper discourse relation, but rather a technical device used to connect multiple parts of a discontinuous EDU. For a fuller description of the labels and the most current GUM annotation guidelines, see `https://wiki.gucorpling.org/gum/rst/`. All definitions refer to the Reader (or hearer) as R, the Writer (or speaker) as W, a nucleus as N, and a satellite as S. The nuclearity of the direction is either ← (for satellite relations that only go left-to-right), → (the opposite), →← (a satellite relation in either direction), or Λ (multinuclear relation). Relation names all have the form <coarse-class>-<fine-grained>, that is, the first three relations in the table belong to the coarse class ADVERSATIVE.

Appendix B. GCDT Example

Figure B.15 provides a sample eRST annotation in Mandarin Chinese using GCDT (Peng, Liu, and Zeldes 2022), along with translations of each EDU. We can observe several discourse signals in this sample, for instance, a discourse marker 但 *dàn* "but", a semantic lexical chain 没有 … 没有… *méi yǒu... méi yǒu* "not have … not have", an instance of semantic repetition 挨饿 … 挨饿 *āi è …āi è* "starve … starving", and a semantic antonym 吃饱 … 挨饿 *chī bǎo …āi è* "full/eat enough … starving". These function quite similarly to the equivalent examples from our English data, though the inventory and distribution of different signal types leaves much to study.

On top of these signals on primary relations, we also note the secondary edge between EDUs 93 and 94, which occurs in circumstances similar to English environments with multiple DMs, in this case, where an orphan 也 *yě* "also" marks a secondary `list` relation.

References

Afantenos, Stergos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning recursive segments for discourse parsing. In *Proceedings of LREC 2010*, pages 3578–3584.

Anuranjana, Kaveri. 2023. DiscoFlan: Instruction fine-tuning and refined text generation for discourse relation label classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28. `https://doi.org/10.18653/v1/2023.disrpt-1.2`

Aoyama, Tatsuya, 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. `https://doi.org/10.18653/v1/2023.law-1.17`

Asher, Nicholas, 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.

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

Asher, Nicholas and Laure Vieu. 2005. Subordinating and coordinating discourse relations. *Lingua*, 115:591–610. `https://doi.org/10.1016/j.lingua.2003.09.017`

Atutxa, Aitziber, Kepa Bengoetxea, Arantza Diaz de Ilarraza, and Mikel Iruskieta. 2019. Towards a top-down approach for an automatic discourse analysis for Basque: Segmentation and central unit detection tool. *PLOS ONE*, 14(9):1–25. `https://doi.org/10.1371/journal.pone.0221639` PubMed: 31483814

Biber, Douglas and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1):2–20. `https://doi.org/10.1016/j.jeap.2010.01.001`

Black, E., S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A

procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California*, pages 308–311.

Bourgonje, Peter and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1061–1066.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–42.

Braud, Chloé, 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. https://doi.org/10.18653/v1/2023.disrpt-1.1

Braud, Chloé, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of COLING 2016*, pages 1903–1913.

Braud, Chloé, 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.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, pages 1–10. https://doi.org/10.21236/ADA460581

Chan, Chunkit, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57. https://doi.org/10.18653/v1/2023.findings-acl.4

Cheng, Yi and Sujian Li. 2019. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29. https://doi.org/10.18653/v1/W19-8104

Crible, Ludivine. 2022. The syntax and semantics of coherence relations: From relative configurations to predictive signals. *International Journal of Corpus Linguistics*, 27(1):59–92. https://doi.org/10.1075/ijcl.19109.cri

Dai, Zeyu and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151. https://doi.org/10.18653/v1/N18-1013

Das, Debopam and Maite Taboada. 2014. RST Signalling Corpus Annotation Manual. Unpublished manuscript.

Das, Debopam and Maite Taboada. 2017. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770. https://doi.org/10.1080/0163853X.2017.1379327

Das, Debopam and Maite Taboada. 2018. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149–184. https://doi.org/10.1007/s10579-017-9383-x

de Marneffe, Marie Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402

Demberg, Vera, Fatemeh Torabi Asr, and Merel Scholman. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135. https://doi.org/10.5087/dad.2019.104

Feng, Vanessa Wei and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521. https://doi.org/10.3115/v1/P14-1048

Gessler, Luke, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir

Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of Discourse Relation Parsing and Treebanking 2021 (DISRPT 2021)*, pages 51–62. https://doi.org/10.18653/v1/2021.disrpt-1.6

Gessler, Luke, Yang Liu, and Amir Zeldes. 2019. A discourse signal annotation system for RST trees. In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, pages 56–61. https://doi.org/10.18653/v1/W19-2708

Gessler, Luke, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. AMALGUM – A free, balanced, multilayer English Web corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5267–5275.

Guz, Grigorii and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167. https://doi.org/10.18653/v1/2020.codi-1.17

Guz, Grigorii, Patrick Huber, and Giuseppe Carenini. 2020. Unleashing the power of neural discourse parsers—A context and structure aware approach using large scale pretraining. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3794–3805. https://doi.org/10.18653/v1/2020.coling-main.337

Hernault, Hugo, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33. https://doi.org/10.5087/dad.2010.003

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

Hoek, Jet, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. 2019. Using the cognitive approach to coherence relations for discourse annotation. *Dialogue and Discourse*, 10(2):1–33. https://doi.org/10.5087/dad.2019.201

Hovy, Eduard H. 1990. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, pages 128–136.

Huber, Patrick and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394. https://doi.org/10.18653/v1/2022.naacl-main.170

Hughes, Rebecca. 1996. *English in Speech and Writing*. Routledge.

Jurafsky, Daniel and James H. Martin. 2024. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd edition.

Kamp, Hans, Josef Van Genabith, and Uwe Reyle. 2011. Discourse representation theory. *Handbook of Philosophical Logic: Volume 15*, pages 125–394. https://doi.org/10.1007/978-94-007-0485-5_3

Kim, Najoung, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414. https://doi.org/10.18653/v1/2020.acl-main.480

Kleijn, Suzanne, Henk L. W. Pander Maat, and Ted J. M. Sanders. 2019. Comprehension effects of connectives across texts, readers, and coherence relations. *Discourse Processes*, 56(5–6):447–464. https://doi.org/10.1080/0163853X.2019.1605257

Knaebel, René. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133. https://doi.org/10.18653/v1/2021.codi-main.12

Kobayashi, Naoki, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down RST parsing utilizing granularity levels in documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8099–8106. https://doi.org/10.1609/aaai.v34i05.6321

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

https://doi.org/10.18653/v1/2022.findings-emnlp.501

Krause, Thomas, Ulf Leser, and Anke Lüdeling. 2016. graphANNIS: A fast query engine for deeply annotated linguistic corpora. *Journal for Language Technology and Computational Linguistics*, 31(1):1–25. https://doi.org/10.21248/jlcl.31.2016.199

Krause, Thomas and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139. https://doi.org/10.1093/llc/fqu057

Kurfalı, Murathan and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19. https://doi.org/10.18653/v1/2021.repl4nlp-1.2

Lascarides, Alex and Nicholas Asher. 2007. Segmented Discourse Representation Theory: Dynamic semantics with discourse structure. In *Computing Meaning*, Studies in Linguistics and Philosophy 3. Springer, pages 87–124. https://doi.org/10.1007/978-1-4020-5958-2_5

Li, Jiaqi, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652. https://doi.org/10.18653/v1/2020.coling-main.238

Li, Sujian, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35. https://doi.org/10.3115/v1/P14-1003

Liu, Wei, Yi Fan, and Michael Strube. 2023. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49. https://doi.org/10.18653/v1/2023.disrpt-1.4

Liu, Yang. 2019. Beyond the Wall Street Journal: Anchoring and comparing discourse signals across genres. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 72–81. https://doi.org/10.18653/v1/W19-2710

Liu, Yang. 2024. *Generalizability and Genre Effects in Discourse Understanding and Parsing in Rhetorical Structure Theory*. Ph.D. thesis, Georgetown University.

Liu, Yang and Amir Zeldes. 2019. Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Proceedings of the Society for Computation in Linguistics*, 2(35):314–317.

Liu, Yang Janet, 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. https://doi.org/10.18653/v1/2023.sigdial-1.3

Liu, Zhengyuan, 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. https://doi.org/10.18653/v1/2021.codi-main.15

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Marcu, Daniel. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, University of Toronto.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics*, 19(2):313–330. https://doi.org/10.21236/ADA273556

Mendes, Amália and Pierre Lejeune. 2022. CRPC-DB a discourse bank for Portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Proceedings*, pages 79–89. https://doi.org/10.1007/978-3-030-98305-5_8

Metheniti, Eleni, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42. https://doi.org/10.18653/v1/2023.disrpt-1.3

Mikulová, Marie, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. 2022. Quality and efficiency of manual annotation: Pre-annotation bias. In

Proceedings of the Language Resources and Evaluation Conference, pages 2909–2918.

Moore, Johanna D. and Cecile L. Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694.

Moore, Johanna D. and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Morey, Mathieu, 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 the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324. https://doi.org/10.18653/v1/D17-1136

Morey, Mathieu, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235. https://doi.org/10.1162/COLI_a_00314

Nishida, Noriki 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. https://doi.org/10.1162/tacl_a_00451

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

Peng, Siyao. 2023. *Cross-Paragraph Discourse Structure in Rhetorical Structure Theory Parsing and Treebanking for Chinese and English*. Ph.D. thesis, Georgetown University.

Peng, Siyao, 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.

Peng, Siyao and Amir Zeldes. 2018. All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 167–177.

Porter, Martin F. 2001. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/introduction

Potter, Andrew. 2019. The rhetorical structure of attribution. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 38–49. https://doi.org/10.18653/v1/W19-2706

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. The Penn Discourse TreeBank 1.0 annotation manual. Technical report, University of Pennsylvania, PDTB Research Group.

Prasad, Rashmi, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950. https://doi.org/10.1162/COLI_a_00204

Pu, Dongqi, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590. https://doi.org/10.18653/v1/2023.acl-long.306

Qin, Lianhui, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017. https://doi.org/10.18653/v1/P17-1093

Rehbein, Ines, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046.

Rutherford, Attapol, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th*

Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 281–291. https://doi.org/10.18653/v1/E17-1027

Sanders, Ted J. M., Vera Demberg, Jet Hoek, C. J. Scholman, Merel, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71. https://doi.org/10.1515/cllt-2016-0078

Sanders, Ted J. M., Jentine Land, and Gerben Mulder. 2007. Linguistic markers of coherence improve text comprehension in functional contexts. *Information Design Journal*, 15(3):219–235. https://doi.org/10.1075/idj.15.3.04san

Sanders, Ted J. M., Wilbert P. M. Spooren, and Leo G. M. Noordman. 1992. Towards a taxonomy of coherence relations. *Discourse Processes*, 15:1–35. https://doi.org/10.1080/01638539209544800

Sanguinetti, Manuela, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250.

Sanguinetti, Manuela, Lauren Cassidy, Cristina Bosco, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2022. Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57:493–544. https://doi.org/10.1007/s10579-022-09581-9

Scholman, Merel, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106. https://doi.org/10.18653/v1/2021.codi-main.9

Shen, Andrew, Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2022. Easy-first bottom-up discourse parsing via sequence labelling. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 35–41.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451. AAAI Press. https://doi.org/10.1609/aaai.v31i1.11164

Stede, Manfred. 2008. RST revisited: Disentangling nuclearity. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *'Subordination' versus 'Coordination' in Sentence and Text: A cross-linguistic perspective*, Studies in Language Companion Series 98. John Benjamins, Amsterdam, pages 33–58. https://doi.org/10.1075/slcs.98.03ste

Stede, Manfred. 2012. *Discourse Processing*, Synthesis Lectures on Human Language Technologies 4. Morgan & Claypool, [San Rafael, CA]. https://doi.org/10.1007/978-3-031-02144-2

Stede, Manfred, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1051–1058.

Stede, Manfred 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.

Sun, Kun and Rong Wang. 2022. Constructing the corpus of Chinese textual 'run-on' sentences (CCTRS): Discourse corpus benchmark with multi-layer annotations. In *International Conference on Natural Language and Speech Processing (ICNLSP)-2022*, Trento, Italy. https://doi.org/10.31234/osf.io/jua9g

Taboada, Maite and Julia Lavid. 2003. Rhetorical and thematic patterns in scheduling dialogues: A generic characterization. *Functions of Language*, 10(2):147–179. https://doi.org/10.1075/fol.10.2.02tab

Taboada, Maite and William C. Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8:423–459. https://doi.org/10.1177/1461445606061881

Teufel, Simone and Marc Moens. 2002. Summarising scientific articles—Experiments with relevance and rhetorical status. *Computational Linguistics*,

28(4):409–445. https://doi.org/10.1162/089120102762671936

Tonelli, Sara, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Vis, Kirsten, José Sanders, and Wilbert Spooren. 2012. Diachronic changes in subjectivity and stance—A corpus linguistic study of Dutch news texts. *Discourse, Context & Media*, 1(2):95–102. https://doi.org/10.1016/j.dcm.2012.09.003

Webber, Bonnie. 2013. What excludes an alternative in coherence relations? In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 276–287.

Widlöcher, Antoine and Yann Mathet. 2012. The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on document engineering*, pages 171–180. https://doi.org/10.1145/2361354.2361394

Xing, Linzi, Patrick Huber, and Giuseppe Carenini. 2022. Improving topic segmentation by injecting discourse dependencies. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 7–18.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 Shared Task*, pages 1–19. https://doi.org/10.18653/v1/K16-2001

Yang, An 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. https://doi.org/10.18653/v1/P18-2071

Yi, Cheng, 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. https://doi.org/10.1007/978-3-030-84186-7_17

Yu, Nan, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280. https://doi.org/10.18653/v1/2022.acl-long.294

Yu, Yue, 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 the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 133–143. https://doi.org/10.18653/v1/W19-2717

Zeldes, Amir. 2016. rstWeb—A browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*, pages 1–5. https://doi.org/10.18653/v1/N16-3001

Zeldes, Amir. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612. https://doi.org/10.1007/s10579-016-9343-x

Zeldes, Amir. 2022. Can we fix the scope for coreference? Problems and solutions for benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62. https://doi.org/10.5210/dad.2022.102

Zeldes, Amir, 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 Discourse Relation Parsing and Treebanking 2021 (DISRPT 2021)*, pages 1–12. https://doi.org/10.18653/v1/2021.disrpt-1.1

Zeyrek, Deniz and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81. https://doi.org/10.18653/v1/W17-0809

Zeyrek, Deniz, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. TED multilingual discourse bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–38. https://doi.org/10.1007/s10579-019-09445-9

Zeyrek, Deniz, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018).

Zhang, Longyin, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3946–3957. https://doi.org/10.18653/v1/2021.acl-long.305

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