GDTB: Genre Diverse Data for English Shallow Discourse Parsing across Modalities, Text Types, and Domains

Yang Janet Liu[2,3,†*] **Tatsuya Aoyama**[1*] **Wesley Scivetti**[1*] **Yilun Zhu**[1*]
Shabnam Behzad[1] **Lauren Elizabeth Levine**[1] **Jessica Lin**[1] **Devika Tiwari**[1] **Amir Zeldes**[1]
[1]Corpling Lab, Georgetown University
[2]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[3]Munich Center for Machine Learning (MCML)
y.liu1@lmu.de {ta571,wss37,yz565}@georgetown.edu
{sb1796,lel76,yl1290,dt719,az364}@georgetown.edu

Abstract

Work on shallow discourse parsing in English has focused on the Wall Street Journal corpus, the only large-scale dataset for the language in the PDTB framework. However, the data is not openly available, is restricted to the news domain, and is by now 35 years old. In this paper, we present and evaluate a new open-access, multi-genre benchmark for PDTB-style shallow discourse parsing, based on the existing UD English GUM corpus, for which discourse relation annotations in other frameworks already exist. In a series of experiments on cross-domain relation classification, we show that while our dataset is compatible with PDTB, substantial out-of-domain degradation is observed, which can be alleviated by joint training on both datasets.

1 Introduction

Language in discourse is more than an ordered list of sentences or clauses: parts of a text expressing events, states, facts, or propositions are often connected by discourse relations, such as CAUSE (one part of a text specifies the cause for another), CONCESSION (one part states content which a speaker or author expects recipients to overlook) etc. Such relations may be marked *explicitly*, for example by *connectives*, which are conjunctions or adverbials such as 'because' or 'nevertheless' in English; or they may be *implicit*, requiring recipients to interpret the text more actively.

Given an arbitrary natural language text as input, shallow discourse parsing is the task of identifying pairs of text spans connected by a discourse relation, in some scenarios focusing mainly on explicit, implicit, or other subtypes of relations (Xue et al., 2016), as well as the means by which they are signaled, resulting for example in connective disambiguation (e.g. 'since' is a connective which can express either a CAUSE or TEMPORAL relation). Most commonly, shallow discourse parsing systems use the inventory of relations defined by the Penn Discourse Treebank (PDTB, currently version 3; see Section 2.1).

Systems and data for shallow discourse parsing can be used for a variety of downstream applications, including relation extraction (identification of relations given two spans, Braud et al. 2024), instruction fine-tuning or pretraining of language models (Ein-Dor et al., 2022), study of argumentation and persuasiveness (Rehbein, 2019), and cross-linguistic lexicography of discourse connectives (Scheffler and Stede, 2016; Das et al., 2018; Kurfalı et al., 2020). When finding specific relation types is desired, shallow discourse parsing also forms an end task in itself: for example, finding all CONCESSION relations in a large corpus of speeches by a politician or political party for Computational Social Science studies.

Although work on shallow discourse parsing has expanded to a range of languages (e.g. Chinese, Zhou and Xue 2014; Czech, Synková et al. 2024, German, Sluyter-Gäthje et al. 2020; Italian, Tonelli et al. 2010; Thai, Prasertsom et al. 2024, Turkish, Zeyrek and Kurfalı 2017, Nigerian Pidgin, Saeed et al. 2024), less progress has been made on expanding data to new and diverse domains (see Section 2.2). A major cause of this bottleneck is the effort associated with manual construction of high quality data covering a broad range of domains from scratch.

In this paper we suggest overcoming this hurdle by not starting from scratch: we target the freely available English GUM corpus (which is also available as part of the Universal Dependencies project, de Marneffe et al. 2021), which covers a broad range of 16 spoken and written English genres and for which annotations are available in hierarchical discourse parsing frameworks: RST and eRST (see Section 2.1). Although these frameworks are sub-

[*] equal contribution; [†] work done while at Georgetown.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12287–12303
November 12-16, 2024

stantially different from PDTB, they provide sufficient information to obtain a high quality starting point for semi-automatic conversion of data into the PDTB v3 framework. As an added advantage, we also develop a mapping of (e)RST to PDTB relations, allowing for cross-framework comparisons along the lines proposed by Demberg et al. (2019) (see Zhu et al. 2021 for a similar argument and approach to converting coreference datasets).

In the subsequent sections of this paper, we will first briefly survey the discourse relation frameworks involved in this project (Section 2), and then we will present our data, its creation process, and an evaluation of its quality (Section 3). This will be followed by a set of experiments on cross-corpus and joint-training relation classification to evaluate both the compatibility of our data with PDTB, and the degree of cross-corpus (and by extension, cross-domain) performance degradation.

2 Related Work

2.1 Discourse Relation Frameworks

A number of frameworks have been proposed for the computational modeling of discourse relations. The Penn Discourse TreeBank (PDTB; Prasad et al., 2014), as briefly outlined above, is a lexically grounded shallow discourse parsing framework, which proposes that texts contain any finite amount of discourse relations (including possibly zero) from an inventory of 36 relations (as of v3) presented in Appendix A, which hold between potentially overlapping spans of text.

PDTB's lexical grounding means that each relation is associated with a kind of triggering device allowing for its identification: ***explicit*** relations correspond to a (possibly multi-word) lexical item which in English is either a subordinating conjunction ('because'), a coordinating one ('but'), or an adverbial, including adverbs ('however') and prepositional phrases ('at the same time'). By contrast, ***implicit*** relations are identified by the potential insertability of a connective, which is not actually present in the text, and generally hold either between consecutive sentences in the same paragraph, or between a small set of additional constructions (e.g. purpose infinitives, for which we can insert an implicit 'in order (to)'; see Section 3.2). PDTB further includes some relations using non-connective expressions:

- **alternative lexicalizations** (ALTLEX): non-connective words such as 'this causes' (instead of 'because')
- **alternative lexicalization constructions** (ALTLEXC): constructions with connective-like functions such as auxiliary inversion in 'had I gone' (instead of 'if')
- **entity relations** (ENTREL): elaborating relations mediated by coreferring entities.
- **hypophora**: the relation between questions and their answers

Adjacent sentence pairs in the same paragraph not mediated by these relations are tagged as NOREL. Relations in PDTB are hierarchical (e.g. COMPARISON.CONTRAST is a type of COMPARISON) and either symmetrical (e.g. COMPARISON.SIMILARITY) or specify a direction using a third level of hierarchy (e.g. COMPARISON.CONCESSION.ARG2-AS-DENIER specifies which argument span is being conceded).

The two other most prominent discourse relation frameworks for which substantial implemented corpora exist are Rhetorical Structure Theory (RST Mann and Thompson 1988) and Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003), which both assume that texts can be completely segmented into elementary discourse units (EDUs, roughly equivalent to propositions), and that EDUs always connect to form a hierarchical graph (in the case of RST, a projective, single rooted tree). Since English SDRT corpora are limited in genre and domain, primarily covering videogame chat (Asher et al., 2016; Thompson et al., 2024) and help forum discussions (Li et al., 2020), we focus here on RST, which has been applied to a broad range of domains (see da Cunha et al. 2011; Liu and Zeldes 2023) and languages (e.g. Basque, da Cunha and Iruskieta 2010, Chinese, Peng et al. 2022, Russian, Pisarevskaya et al. 2017 and more). An example of RST discourse annotation is illustrated for a text fragment in Figure 1 (disregarding blue edges and highlighted words, see below).

RST enforces a single tree hierarchical structure over an entire document, assuming that every smallest unit of analysis (i.e. EDU) is related to another unit or subtree by one of the proposed discourse relations such as CAUSE, BACKGROUND, or CONTRAST (see Appendix B for the relation inventory used in GUM). Importantly, such relations

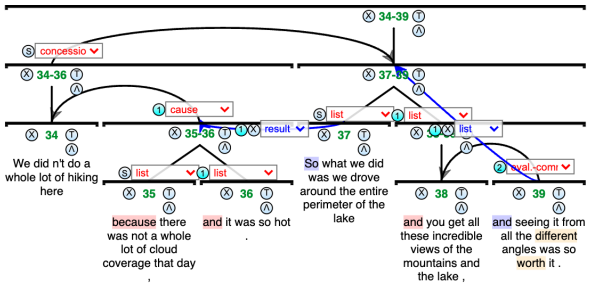


Figure 1: A Discourse Analysis in RST (disregarding blue edges and highlighted words) and eRST.

are annotated in the spirit of *plausibility judgments* (Mann and Thompson, 1988, 246) from the perspective of the writer, independent of the words in the text, meaning that it is fundamentally pragmatic in orientation, rather than lexically grounded. Relations in RST are either directed, from a less prominent satellite to a more prominent nucleus, or symmetrical, forming multinuclear units. However, RST does not mark connectives or other expressions indicative of relations, and is incapable of expressing multiple, concurrent relations on the same nodes, or tree-breaking relations.

Recently, Zeldes et al. (2024) proposed an enhanced version of RST called eRST, which adds additional, tree-breaking relations on top of RST, as well as signaling annotations, which indicate the rationale for relation annotations using 45 signal categories arranged into 8 classes. Zeldes et al. (2024) also released a re-annotated version of the multi-genre English GUM RST treebank (see Section 3) containing enhanced RST trees with relation signals. Fortunately for the present purpose, one of the signal classes in this dataset corresponds to connectives based on PDTB definitions, facilitating the conversion process we will outline below. Figure 1 highglights the additional structures of the eRST graph, in which tree-breaking relations are marked by blue edges, and signal token spans are highlighted in different colors based on their classes. Connectives corresponding to regular relations are marked in red, and those corresponding to tree-breaking edges are marked in blue.

2.2 Datasets

Due to its size, work on shallow discourse parsing in English has focused on PDTB, which, as of version 3, contains over 53K discourse relation instances, the majority of which are either explicitly or implicitly associated with connectives (about 24K and 21K each). Despite its impressive size, PDTB is limited as a resource for text covering language other than newswire, since its underlying data comes exclusively from the Wall Street Journal (WSJ) corpus, containing WSJ articles from 1989. As a result, the data is missing both a range of contemporary thematic content (e.g. words like 'cellphone', 'European Union', or 'website' are absent) and diverse modes of communication (e.g. PDTB's HYPOPHORA relation, indicating question-answer pairs, occurs only 141 times, and spoken language connectives such as 'cause'/'cuz' are unattested).

The few other available datasets which cover English discourse relations in the PDTB-style are newer, but much smaller, and cover few domains: The TED Multilingual Discourse Bank (TED-MDB, Zeyrek et al. 2019) follows PDTB in including not only explicit and implicit relations, but also ALTLEX, ENTREL, and NOREL annotations for TED talks translated into six languages, for a total of 3,649 relations. However, only 661 of those annotations cover English data, and the corpus has not been updated to conform to the PDTB v3 guidelines.

A much larger dataset, Edina-DR (Ma et al., 2019), contains 27,998 implicit relations from conversational data. However, the corpus is annotated fully automatically and only at the top level of the PDTB relation hierarchy, thereby distinguishing only 4 relation labels (COMPARISON, EXPANSION, CONTINGENCY, and TEMPORAL).

DiscoGeM (Scholman et al., 2022) contains 6,505 relations from Wikipedia texts, European Parliament proceedings and literature, and comes closest to the goal of the resource presented here in offering diverse text types with detailed manual relation annotations. Version 2.0 of the corpus (Yung et al., 2024) also adds parallel data for a subset of relations for three languages: Czech, French, and German. However, the corpus only covers intersentential *implicit* relations, thereby limiting the scope of the task substantially, and still contains no conversational spoken data, academic writing, YouTube data etc., which we aim to cover with GDTB. With this in mind we present the contents of our corpus in the next section.

3 GDTB

3.1 Contents

The dataset presented in this paper is the GUM Discourse Treebank (GDTB), a multi-genre PDTB v3-style corpus for English semi-automatically con-

	GDTB	PDTB v3
Tokens	228,399	1,156,308
Docs	235	2,161
Genres	16	1
AltLex	224	1,498
AltLexC	13	140
EntRel	553	5,538
Explicit	7,202	24,238
Hypophora	465	146
Implicit	4,503	21,781
Norel	662	287
All	13,622	53,628

Table 1: Relation Type Counts: GDTB vs. PDTB v3.

verted from the GUM corpus (Zeldes, 2017).[1] GUM is a growing multilayer corpus of English containing, among other things, discourse parses with aligned connective annotations in eRST, Universal Dependencies syntax trees, entity and coreference annotations, and more.

After conversion of the data, the process for which is described below, the final GDTB benchmark based on GUM v10 contains 13.6K relation annotations, a little more than a quarter of the size of PDTB v3, but stemming from much more diverse and up to date materials. Table 1 compares the two datasets. Note that because sentences in some genres are shorter than in news text, GDTB is less than $1/4$ the size of PDTB in tokens, but denser in discourse relations; at the same time, shorter paragraphs in many genres mean the proportion of implicit relations is lower. Some relation types are also more frequent in GDTB; in particular, HYPOPHORA, which corresponds to questions, is common in many genres but rare in newspaper language. The underlying data in GUM is regularly expanded and currently covers 16 genres, where data collection for four of these is still ongoing ('growing' genres). Table 2 gives an overview of the data, with the four growing genres at the bottom. Genres cover both spoken (e.g. conversations, courtroom transcripts, YouTube vlogs) and written modalities (incl. news, academic, how-to guides from wikiHow) from various open licensed sources, which should make models trained on GDTB more robust to open domain data (see Section 4).

[1]Our data is made available at https://github.com/gucorpling/gum2pdtb under a Creative Commons license in accordance with the original GUM license. Data from the Reddit genre (Behzad and Zeldes, 2020) is released without underlying text, but a script is provided to reconstruct the data using an API. We plan to include future versions of GDTB with new documents as part of the main GUM corpus releases, as the GUM corpus grows.

Genre	Docs	Tokens	Relations
academic	18	17,169	815
bio	20	18,213	868
conversation	14	16,391	1,113
fiction	19	17,510	1,281
interview	19	18,196	1,188
news	23	16,146	724
reddit	18	16,364	1,146
speech	15	16,720	913
textbook	15	16,693	936
vlog	15	16,864	1,415
voyage	18	16,514	799
how-to	19	17,081	1,331
court	6	7,069	478
essay	5	5,750	348
letter	6	5,982	365
podcast	5	5,737	359

Table 2: Genre Breakdown for GDTB. The bottom four 'growing' genres are still being collected for GUM and counts represent sizes as of GUM v10.

3.2 Dataset Conversion

As mentioned above, GUM v10 contains several annotation layers that describe linguistic phenomena at various levels, including eRST trees, but also gold syntax and coreference annotations, which we harness to create GDTB.

Sense Mapping Our approach to creating GDTB uses a cascade of relation conversion modules, and manual annotation for some types of error-prone cases in the entire corpus (all relations in the test set are also manually annotated). All modules rely on a mapping of allowable output relations, adapted from the PDTB v2 proposal in Demberg et al. (2019), which had to be modified in several ways (see also Costa et al. 2023 on mapping v3 data). PDTB v3 introduced finer-grained Level-3 sense distinctions, which are mostly concerned with relation directionality. Because what PDTB calls *Arg1* and **Arg2** is determined by the syntactic configuration and their linear order in the text, its interaction with the order-dependent Level-3 senses is not straightforwardly mappable from RST relations, where directionality is based on labels (e.g. CAUSE vs. RESULT) and nuclearity or relative prominence. That said, in many cases a deterministic mapping can be achieved (e.g. what an RST CONCESSION relation concedes is reliably the opposite argument span of the .ARGX-AS-DENIER argument in PDTB v3).

Secondly, the RST framework adopted in Demberg et al. (2019) is based on the RST-DT corpus (Carlson et al., 2003), which uses a set of relation labels slightly different from that of GUM v10. This incompatibility was resolved in consultation with the original RST-DT relation descriptions (Carlson et al., 2003) and the description of GUM v10 discourse relations.[2] Finally, since the resulting label mapping is still often many-to-many, each module employs different strategies to disambiguate potential PDTB senses, which are detailed below. Figure 4 in Appendix G presents some GDTB examples spawned by RST annotations given our conversion process described below.

Explicit Module Explicit relation candidates are generated with simple heuristics: (1) for each eRST relation in GUM, add the relation to the candidate list if it is signaled by a connective; (2) for each relation in the candidate list, determine the allowable PDTB labels based on the connective and the RST relation; and (3) take the target and source EDU spans and convert them into PDTB argument spans based on another set of rules (see **Argument Span Module** below for more details).

For step (1), we use the gold GUM eRST framework annotations from Zeldes et al. (2024) to determine if a given relation is explicitly signaled by a connective. For (2), we refer to Appendix A of the PDTB v3 annotation guidelines (Webber et al., 2019) to obtain a list of corresponding connectives and PDTB senses they may signal. For simplicity, rare combinations, such as secondary senses (e.g. TEMPORAL.SYNCHRONOUS|COMPARISON.CONTRAST) and speech act variants XYZ+SPEECHACT (which only make up 0.4% of the PDTB explicit data, or 121 cases) are not considered. For cases where multiple outcomes are possible, we train DisCoDisCo (Gessler et al., 2021), a discourse relation classification system which remains state-of-the-art on the DISRPT shared task benchmark for relation classification (Braud et al., 2024), on PDTB v3, and use its predictions to disambiguate data in our training and development sets. The test set is completely manually corrected to allow for the evaluation in Section 4.

Implicit Module Implicit relations are also handled in a three-step approach: (1) identify every junction allowing an implicit relation (viz. between sentences, before purpose infinitives and participial adverbial clauses, and between zero-coordinated clauses); (2) predict the connective given existing RST relations, and (3) map the connective and relation onto a PDTB relation. For (1) we use the gold syntax trees and RST relations, which allow us, for example, to identify sentence boundaries, purpose infinitives, etc. Implicit relations are only allowed in the absence of an explicit relation in the same span, with one exception: RST SEQUENCE relations signaled by an explicit EXPANSION.CONJUNCTION connective (e.g. 'and'), are allowed a second TEMPORAL relation with implicit 'then', matching PDTB's policy, as in example (1).

(1) I cut my losses **and** (*then*) ran - 'and'
Expl. EXPANSION.CONJUNCTION + '(then)'
Impl. TEMPORAL.ASYNCHRONOUS.PRECEDENCE

For step (2), we train a connective prediction model to output a list of hypothetical connectives for each relation. This model is trained on implicit relations from the PDTB training set. We also supply the model with information about the possible PDTB relation senses that are compatible with existing RST relations at that juncture as part of the input. Specifically, we fine-tune `flan-t5-large`[3] (Chung et al., 2024) for 25 epochs for this task and select the best-performing model on the dev set. See Appendix C for task performance and comparison to a majority baseline, and Appendix D for example prompts used in this task.

We manually validate the entire test set and establish that the process is generally reliable for well-mappable relations (e.g. RST CONDITION or CAUSE relations are easy to map), but less reliable for RST relations with no specific equivalents. In particular, we identify a high error rate for RST CONTEXT-BACKGROUND and JOINT-OTHER, which we manually correct for the entire corpus.

Following connective prediction we use the same mapping of connectives and RST relation combinations as the explicit module to select the most likely PDTB relation. In ambiguous cases we again rely on DisCoDisCo predictions, except for relations corresponding to RST CONTEXT-BACKGROUND and JOINT-OTHER relations, or in the test set, which we manually correct.

[2] https://wiki.gucorpling.org/gum/rst

[3] https://huggingface.co/google/flan-t5-large

AltLex Module In some instances, there is no explicit connective present to signal a discourse relation, but the insertion of an implicit connective appears redundant due to an existing expression in the spans under consideration. In such cases, PDTB recognizes this expression as an *alternative lexicalization* (ALTLEX) of the discourse relation, annotating the span of the ALTLEX expression, its relation, and its corresponding argument spans.

As there is no specific syntactic requirements for ALTLEX expressions, their detection is challenging. For this corpus conversion, we adopt a conservative approach for the annotation of potential ALTLEX expressions, adopting a pattern-matching approach similar to that outlined in Knaebel and Stede (2022), catching only cases which are attested in PDTB v3. The ALTLEX module is only consulted in the absence of an Explicit relation.

AltLexC Module ALTLEXC is a subtype of ALTLEX, where the relation is expressed by syntactic constructions within a sentence, as shown in (2). The ALTLEXC module is only consulted in the absence of an Explicit or ALTLEX relation. The list of acknowledged constructions for ALTLEXC in PDTB v3 is closed, making a rule-based approach based on syntax trees straightforward. To identify these, we first extract the seven syntactic constructions listed in the PDTB v3 annotation guidelines (Webber et al., 2019), such as Auxiliary Inversion, where **Arg2** signals the CONTINGENCY.CONDITION sense.

(2) **Had it happened five hours earlier or four hours earlier**, *I think the death toll would have been more than a thousand.*

We then verify that the matching syntax trees are contained in spans connected with compatible RST relations based on the mapping. For example, the original RST relation CONTINGENCY-CONDITION in (2) is mapped onto the PDTB relation CONTINGENCY.CONDITION. Although this process only captures 13 instances of ALTLEXC in the corpus, these were error free based on manual inspection, and additional searches in the syntax annotations suggest that these have been exhaustively identified in the corpus.

Hypophora Module The hypophora relation type is straightforwardly generated for each RST TOPIC-QUESTION relation in the source annotations, since these correspond exactly to questions, the category covered by hypophora.

EntRel Module At the juncture of each two adjacent, same-paragraph sentences for which no other relation has been generated, we check the GUM gold coreference and RST annotations to see whether any kind of JOINT or ELABORATION relation applies, and if so, whether it corresponds to coreference from a definite or pronominal expression in the second sentence referring back to the first sentence. If so, we generate an ENTREL relation, and otherwise, mark the span as NOREL, following PDTB guidelines. This annotation type is manually corrected only in the test set.

Argument Span Module In order to make target and source EDU spans conform to PDTB-style argument spans, we first apply the argument labeling convention described in the PDTB v3 annotation manual (Webber et al., 2019, Section 3.1) to each pair of RST EDU spans, which are a set of rules based on syntactic configuration and linear text order. Once the argument labeling has determined which span is *Arg1*/**Arg2**, we refine the sense labels by adding the Level-3 sense information to restore directionality. In addition, we emulate PDTB's *Minimality Principle*, which states that only the minimal text needed for a given discourse relation should be included in the argument spans (Prasad et al., 2014). As a result, we also adjust the corresponding EDU spans by clipping EDU spans to a single sentence that contains just the head EDU if they are multi-sentential, and the exact span dominated by a relation source or target intra-sententially. Attribution spans which scope over an argument nucleus are also removed, in accordance with PDTB guidelines (e.g. *[X said] [A happened] [because B]* results in the removal of 'X said' for the CAUSE relation). Argument span accuracy is evaluated below.

3.3 Evaluation

To assess the quality of the annotations in GDTB, we evaluate system outputs against the manually corrected test set (1531 relations), as well as conducting an inter-annotator agreement study. For the first experiment, we compare the fully corrected test data to the same test data but with only corrections done on the entire corpus, such as inspection of BACKGROUND and OTHER relations. Following an initial training session with joint adjudication, data was corrected by a team of nine Computational Linguistics graduate students and faculty with for-

Relation Scores (exact label and span match)			
type	**P**	**R**	**F1**
altLex	0.9500	0.7600	0.8444
altLexC	1.0000	1.0000	1.0000
EntRel	0.7593	0.8913	0.8200
Explicit	0.9812	0.9874	0.9843
Hypophora	0.8750	0.8537	0.8642
Implicit	0.8784	0.8205	0.8485
NoRel	0.7887	0.9180	0.8485
micro-avg.	0.9277	0.9161	**0.9218**
Span Scores (incl. relation type but not sense)			
altLex	0.9500	0.7600	0.8444
altLexC	1.0000	1.0000	1.0000
EntRel	0.7778	0.9130	0.8400
Explicit	0.9935	1.0000	0.9967
Hypophora	0.8750	0.8537	0.8642
Implicit	0.9824	0.9176	0.9489
NoRel	0.7887	0.9180	0.8485
micro-avg.	0.9678	0.9554	**0.9616**

Table 3: Test Set Accuracy (manual correction).

mal training in discourse parsing formalisms as part of a research project. Following previous work we evaluate on Level-2 relations in two scenarios: ***exact match***, where the label, argument span, and relation type must match, and ***span-only match***, meaning the relation type was identified and argument spans are correct but the label may not be.

As Table 3 shows, the overall quality of the corpus is very high, with a micro-F1 score of 92, above our initial expectations given human agreement scores reported for PDTB annotation. Although there are no comprehensive numbers available for PDTB v3 annotation, Prasad et al. (2008) reported 84% accuracy (exact match between annotators) on v2 senses, which did not include the more challenging intra-sentential implicit relations, ALTLEXC, or Hypophora, and Zeyrek et al. (2019) similarly reported 79% agreement. Bourgonje and Stede (2020) reported a Cohen's Kappa of 0.74 on all relations, again excluding intra-sentential relations.[4] On the lower end, Scholman et al. (2022) reported 60% agreement and κ=.45 on Level-3 senses using the v2 inventory with crowd workers, while Yung et al. (2024) emphasized the importance of collecting multiple labels and reporting confusion matrices.

As the bottom half of Table 3 shows, argument spans are relatively unproblematic compared to sense prediction, especially for implicit cases, where span matching achieves almost 0.95, but exact match F1 including sense is just below 0.85. This is not unexpected, given that human judgments on insertion of an unexpressed connective vary considerably. For more details on the kinds of labels that human annotators corrected, and detailed confusion matrices, see Appendix E. We further double-annotated 8 documents from the test set focusing on implicit instances. A Cohen's kappa of 0.79 was achieved on connectives, κ=0.77 on Level-3 senses, and κ=0.83 on Level-2 using PDTB v3 inventory, indicating excellent agreement.

4 Experimental Setup

To test the utility of our corpus and its compatibility (or redundancy) with the existing PDTB v3, we again train the DisCoDisCo relation classifier (Gessler et al., 2021) using the standard DISRPT version of PDTB v3 relations, which simply provides the spans of the two arguments including connectives, and their containing sentences, without a separate connective field, and as a result treats explicit, implicit, and other relation types uniformly,[5] though we also report separate scores on different relation types and overall.

To investigate the effects of both data size and data diversity, we evaluate in three training setups: **within-corpus** (e.g. train and test on PDTB v3, and the same for GDTB respectively); **cross-corpus** (train on PDTB v3 and evaluate on the GDTB test set, and vice versa); and **joint training** (train on both training sets, evaluate on each test set). See implementation details in Appendix F.

5 Results

Table 4 gives an overview of within-, across-, and joint-corpus overall relation classification accuracy. The overall scores show that GDTB is the more challenging corpus when training is done jointly, and cross-corpus degradation is non-negligible, with around 10 points degradation for training on PDTB v3 and testing on GDTB, and even more so in the opposite direction. Although joint training slightly under-performs within-corpus numbers in both directions, the relatively low level of degradation for the joint model compared to the corre-

[4] We considered reporting kappa for our data as well, but this requires the set of relations for annotation to match and carry only one label each. Using argument spans to align instances, we can compute kappa for the 93.7% of relations which have exactly one sense in both the predicted and corrected data – for these we obtain κ=0.913.

[5] DISRPT datasets do not contain instances of ENTREL.

	Test Set	
Training	**GDTB**	**PDTB v3**
within-corpus	0.6447	0.7572
cross-corpus	0.5660	0.4457
joint-training	0.6440	0.7390

Table 4: Overall Accuracy Scores (within-corpus=train set is from the corpus of the test set; cross-corpus=train set from opposite corpus; joint=train on both).

Train	Test	Explicit	Implicit	altLex	altLexC	Hypophora
GDTB	GDTB	0.7645	0.4579	0.4400	1	0.8780
	PDTB v3	0.6114	0.2842	0.3333	0.5000	0.7500
PDTB v3	GDTB	0.6794	0.4048	0.3600	1	0.5854
	PDTB v3	0.8817	0.6020	0.8986	0.9167	0.8750
GDTB & PDTB v3	GDTB	0.7374	0.4908	0.4400	1	0.9512
	PDTB v3	0.8679	0.5683	0.8261	0.8333	0.8750

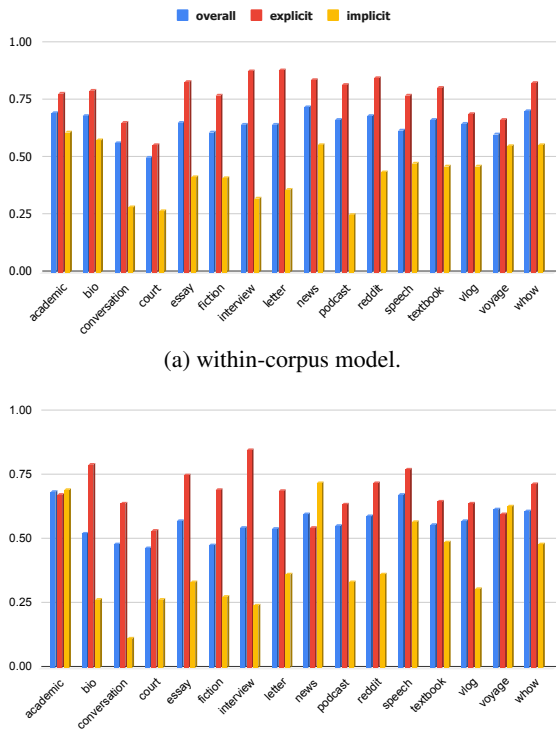
Table 5: Accuracy by Relation Types.

sponding within-corpus numbers suggests that the joint model is a much better choice for training a system to tag truly unseen, open domain data.

However, the different proportions of explicit relations (which are easier to tag) and implicit ones mean that Table 4 does not give the entire picture. Thus, we also report accuracy scores for different relation types in Table 5, which shows that the relation type which benefits most from joint training is HYPOPHORA, which is rare in PDTB v3; without GDTB data, the PDTB-trained model degrades 29 points out of domain. For HYPOPHORA, we see that the joint model out-performs or performs as well as single corpus training and testing. The joint model is unsurprisingly superior in the cross-corpus macro-average, which is a better proxy for realistic applications to unseen data in the wild.

Implicit relations in particular show massive cross-corpus degradation for PDTB, which is again unsurprising: while connectives remain more or less constant across datasets (i.e. 'but' usually signals COMPARISON.CONTRAST or COMPARISON.CONCESSION in both datasets), in implicit settings, the relations between surrounding lexical items must be learned, which vary more substantially by genre. Although the GDTB model has seen some news data from GUM *news*, the quantity of the material is insufficient to achieve comparable scores to the PDTB-trained model, which has 1.2M tokens of WSJ data to learn from.

If we zoom in on genres in GDTB, Figure 2 shows scores for subsets of GDTB `test` from both



(a) within-corpus model.

(b) cross-corpus model.

Figure 2: GDTB Scores by Genres and Relation Types.

the within-corpus and the cross-corpus models.[6] Overall, the best-performing genres for each model are *news* and *academic* respectively, while the worst-performing genre is *court* for both models. The bottom four places in implicit relations within-corpus are all occupied by spoken genres (*podcast, court, conversation, interview*). The *conversation* genre is particularly bad for cross-corpus, i.e. when training on PDTB, with an implicit score of just 11.43%. Unsurprisingly, *news* scores highest on cross-corpus implicit relation prediction (72.22%), though *academic* scores higher overall (68.60%) (due to frequent explicit relations), and *interview* scores highest for cross-corpus explicit (84.85%), likely due to the frequent use of easy connectives such as TEMPORAL 'when' in questions, and 'then' in sequential narration in answers.

One of the major motivations of this work is releasing genre-diverse, complete PDTB v3-style data to facilitate cross-framework and cross-domain shallow discourse parsing. While, strictly speaking, there is no other dataset annotated like PDTB v3 to evaluate on, other existing datasets do contain a subset of PDTB v3 annotations. For instance, Table 6 shows the accuracy scores of the English portion of the TED-Multilingual Discourse

[6]Full results and confusion matrices are in Appendix H.

	GDTB-trained	PDTB-trained	joint-training
TED-MDB (English)	0.5214	0.5556	0.5641

Table 6: Accuracy Scores of TED-MDB (English).

Bank (TED-MDB, Zeyrek et al. 2018, 2020) across three training scenarios. As the table shows, we observe out-of-domain gains on the TED talks in the corpus when using the jointly trained model. However, it is worth pointing out that since TED-MDB only contains a subset of PDTB v3 annotations, these scores are not directly comparable to the scores reported in the tables above.

6 Conclusion and Outlook

In this paper, we present GDTB, a PDTB-style dataset covering 16 English spoken and written genres for open-domain shallow discourse parsing, which we create primarily using a cascade of conversion modules leveraging enhanced RST annotations. The data covers all aspects of PDTB v3 annotation, including explicit and implicit inter-/intra-sentential relations as well as alternative lexicalizations, entity relations, and hypophora.

We show that RST relations lead to reliable PDTB-style annotations, particularly for explicit relations. Using state-of-the-art fine-tuned sequence-to-sequence models and the (e)RST relations as inputs, we are also able to obtain high quality predictions for implicit relations, which we correct in whole for the test set and in part for the remaining data's most unreliably convertible RST relation types (e.g. CONTEXT-BACKGROUND and JOINT-OTHER).

Our experiments show that there is substantial degradation in cross-corpus PDTB-style relation classification in both directions, demonstrating PDTB's current inadequacy for relation classification in open domain settings. We show that jointly training a relation classification system on both PDTB v3 and GDTB leads to much greater cross-corpus stability without sacrificing much performance on PDTB v3. We are therefore confident that GDTB can be a valuable resource for improving out-of-domain performance of PDTB-style English shallow discourse parsing systems.

In future work, we believe the same pipeline presented in this paper can be applied to additional corpora annotated with RST in general and eRST in particular. Specifically, the recent addition of eRST annotations to the GENTLE corpus (**GEn**re **T**ests for **L**inguistic **E**valuation), an extension corpus applying GUM's annotation scheme to 8 more unusual English genres (Aoyama et al., 2023), should allow for more GDTB-like data to be produced with ease. This data would cover the additional GENTLE genres, which encompass dictionary entries, eSports video commentary, legal documents, medical notes, poetry, mathematical proofs, course syllabuses, and threat letters.

Finally, beyond discourse parsing, we also see great promise in using GDTB for theoretical studies of discourse relation variation across genres, and for the comparison of alignments between PDTB-style relations and RST or eRST annotations. In particular, we believe GDTB and GUM-RST will prove to be informative for future comparisons between theoretical frameworks, along the lines proposed by Demberg et al. (2019).

Limitations

In this work, we use the gold eRST relations present in GUM as a starting point for our PDTB-style annotations. It is likely that annotating the same underlying data from scratch in the PDTB style would yield a slightly different result than our approach here, particularly regarding implicit relations. In PDTB, implicit relations are posited between all adjacent sentences but only between adjacent sentences which have an eRST relation that connects the two sentences in GDTB. This means that while implicit relation precision in GDTB has high quality, our recall is likely to be partial. We believe that due to the prioritization of pragmatically prominent relations in RST and the possibility of multiple tree-breaking relations in eRST, the most salient relations in documents should already be included in our data. In support of this claim, we note that previous work has found that PDTB-style relations for Czech seldom violate RST tree projectivity constraints (Poláková et al., 2021), especially if multiple concurrent relations are permitted (see also Polakova et al. 2024 on annotating RST relations for Czech PDTB-style data).

Another limitation is the noise inherent to a conversion approach which uses automatic processes. In our case, this is especially true for the automatically generated implicit connectives, which may diverge somewhat from the most natural connectives chosen by humans. As stated previously, this issue is particularly problematic for implicit relations

spawned from RST relations without good PDTB mappings, such as JOINT-OTHER and CONTEXT-BACKGROUND, which were therefore manually corrected in our entire dataset (including correction of connectives), next to the correction of all connectives in the test set. That said, our evaluation shows that, by relying on multiple sources of information for our final predictions, the final product is substantially better than an automatically created dataset tagging just discourse relations from plain text, producing a resource that is close to gold-standard quality, or at a minimum, significantly 'better-than-silver' (cf. Gessler et al. 2020). Future work could improve the quality of additional predicted connectives to bring the complete corpus closer to gold-standard accuracy.

In addition, this work explores the possibility of converting an English RST-style discourse treebank to a PDTB-style one based on the English PDTB v3, leading to several limitations in the applicability of our methods. Primarily, the work is limited to English as a target language, and does not address the lack of diverse data in other languages. That being said, we believe that the resource created here is valuable for facilitating multilingual shallow discourse parsing, as recently experimented in Bourgonje and Demberg (2024) where state-of-the-art model originally developed for English discourse relation classification was extended to a multilingual setting, and by employing some simple yet effective learning techniques, the discourse relation classification performance becomes more generalizable and robust across both languages and domains.

Lastly, our methods assume the existence of RST and connective annotations for the source material. While there are many RST datasets which could be converted following the methods proposed here, nearly no RST dataset also contains connective annotations (the German PCC corpus, Bourgonje and Stede 2020, is an exception). Nevertheless, our methods could be applied to other RST corpora using automatic or manual connective annotation, especially in languages for which connective lexical resources and/or some PDTB-style data exist.

Ethics Statement

This work, like most work in Computational Linguistics, can enable the creation of Natural Language Processing systems which may cause harm. However, we believe that the lack of diverse data for training systems is a greater potential harm than making additional data available, which will hopefully allow systems to behave in less biased and more generalizable ways.

In addition, this work employs deep learning architectures whose training involves carbon emissions. While these should not be ignored, we assess them to be of a modest scope, given that we are relying on existing pre-trained models, which are only fine-tuned on small amounts of data using limited resources. Finally, all human labor involved in this paper was carried out by paid university employees, including funded graduate students as part of their research work. No unpaid volunteers or low-paid crowd workers were involved in the creation of this data.

Acknowledgments

We recognize the support for Yang Janet Liu through the ERC Consolidator Grant DIALECT 101043235.

References

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

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

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

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

Peter Bourgonje and Vera Demberg. 2024. Generalizing across languages and domains for discourse relation classification. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 554–565, Kyoto, Japan. Association for Computational Linguistics.

Peter Bourgonje 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, Marseille, France. European Language Resources Association.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

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

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. Mapping explicit and implicit discourse relations between the RST-DT and the PDTB 3.0. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 344–352, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Iria da Cunha and Mikel Iruskieta. 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5):563–598.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, Melbourne, Australia. Association for Computational Linguistics.

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

Vera Demberg, Merel Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10:87–135.

Liat Ein-Dor, Ilya Shnayderman, Artem Spector, Lena Dankin, Ranit Aharonov, and Noam Slonim. 2022. Fortunately, discourse markers can enhance language models for sentiment analysis. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10608–10617. AAAI Press.

Luke Gessler, 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 the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

René Knaebel and Manfred Stede. 2022. Towards identifying alternative-lexicalization signals of discourse relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 837–850, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Murathan Kurfalı, Sibel Ozer, Deniz Zeyrek, and Amália Mendes. 2020. TED-MDB lexicons: Tr-EnConnLex, pt-EnConnLex. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 148–153, Online. Association for Computational Linguistics.

Jiaqi Li, 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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the*

17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Mingyu Derek Ma, Kevin Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019. Implicit discourse relation identification for open-domain dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 666–672, Florence, Italy. Association for Computational Linguistics.

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

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

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, A. Nasedkin, S. Nikiforova, I. Pavlova, and A. Shelepov. 2017. Towards building a discourse-annotated corpus of Russian. In *Computational Linguistics and Intellectual Technologies: 23rd International Conference on Computational Linguistics and Intellectual Technologies "Dialogue"*, pages 194–204.

Lucie Polakova, Jiří Mírovský, Šárka Zikánová, and Eva Hajicova. 2024. Developing a Rhetorical Structure Theory treebank for Czech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4802–4810, Torino, Italia. ELRA and ICCL.

Lucie Poláková, Jiří Mírovský, Šárka Zikánová, and Eva Hajičová. 2021. Discourse relations and connectives in higher text structure. *Dialogue & Discourse*, 12(2):1–37.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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

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

Ines Rehbein. 2019. On the role of discourse relations in persuasive texts. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 144–154, Florence, Italy. Association for Computational Linguistics.

Muhammed Saeed, Peter Bourgonje, and Vera Demberg. 2024. Implicit Discourse Relation Classification For Nigerian Pidgin.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1008–1013, Portorož, Slovenia. European Language Resources Association (ELRA).

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1044–1050, Marseille, France. European Language Resources Association.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague discourse treebank 3.0. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1270–1279, Torino, Italia. ELRA and ICCL.

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

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, 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)*, Valletta, Malta. European Language Resources Association (ELRA).

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*.

Nianwen Xue, 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, Berlin, Germany. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Sarka Zikanova, and Vera Demberg. 2024. DiscoGeM 2.0: A parallel corpus of English, German, French and Czech implicit discourse relations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4940–4956, Torino, Italia. ELRA and ICCL.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2024. eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, pages 1–47.

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

Deniz Zeyrek, 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.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED multilingual discourse bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Lang. Resour. Eval.*, 54(2):587–613.

Deniz Zeyrek, 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuping Zhou and Nianwen Xue. 2014. The Chinese Discourse TreeBank: A Chinese Corpus Annotated with Discourse Relations. *Language Resources and Evaluation*, 49:397 – 431.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A PDTB v3 Sense Hierarchy

Table 7 presents the PDTB v3 Sense Hierarchy, reproduced from the PDTB v3 annotation manual (Webber et al., 2019, Section 4). This work reproduces the complete Level-3 sense labels, with the exception of the rare +BELIEF and +SPEECHACT variants, which are collapsed to their corresponding basic Level-3 variants.

Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE SUCCESSION
CONTINGENCY	CAUSE	REASON RESULT NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST ARG2-AS-SUBST

Table 7: An Overview of the PDTB v3 Sense Hierarchy.

B RST Relation Inventory in GUM

Table 8 presents the RST relation inventory used in GUM, both fine-grained relation labels as well as the corresponding coarse-grained relation classes are provided. Note that SAME-UNIT is not a true discourse relation but instead a label used to mark discontinuous spans in RST as a result of RST's EDU segmentation. We include it here for completeness purposes.

C Implicit Connective Prediction Performance

Regarding the evaluation of implicit connective predictions, we calculated two types of accuracy of predicted connective in GDTB against the human evaluations on the GDTB test set: ***exact match*** and ***fuzzy match***. Overall, for exact match the connectives were judged as natural by our evaluators at an

GUM v10 Classes	GUM v10 Relations	GUM v10 Classes	GUM v10 Relations
ADVERSATIVE	ADVERSATIVE-ANTITHESIS	JOINT	JOINT-DISJUNCTION
	ADVERSATIVE-CONCESSION		JOINT-LIST
	ADVERSATIVE-CONTRAST		JOINT-SEQUENCE
ATTRIBUTION	ATTRIBUTION-POSITIVE		JOINT-OTHER
	ATTRIBUTION-NEGATIVE	MODE	MODE-MANNER
CAUSAL	CAUSAL-CAUSE		MODE-MEANS
	CAUSAL-RESULT	ORGANIZATION	ORGANIZATION-HEADING
CONTEXT	CONTEXT-BACKGROUND		ORGANIZATION-PHATIC
	CONTEXT-CIRCUMSTANCE		ORGANIZATION-PREPARATION
CONTINGENCY	CONTINGENCY-CONDITION	PURPOSE	PURPOSE-ATTRIBUTE
ELABORATION	ELABORATION-ATTRIBUTE		PURPOSE-GOAL
	ELABORATION-ADDITIONAL	RESTATEMENT	RESTATEMENT-PARTIAL
EXPLANATION	EXPLANATION-EVIDENCE		RESTATEMENT-REPETITION
	EXPLANATION-JUSTIFY	TOPIC	TOPIC-QUESTION
	EXPLANATION-MOTIVATION		TOPIC-SOLUTIONHOOD
EVALUATION	EVALUATION-COMMENT	SAME-UNIT	SAME-UNIT

Table 8: RST Relation Inventory in GUM v10.

accuracy of 79%. While accuracy at matching the connective exactly is somewhat low, many of the model predicted connectives were compatible with the correct PDTB sense, but were changed by annotators to add additional fluency and naturalness. The system predicts a reasonable connective, that is, a connective that is valid for the gold PDTB v3 sense, at an accuracy of 89%. We call this more lenient scoring method the *fuzzy match* accuracy, compared to the *exact match* accuracy where the connective must be identical to what the annotator ultimately decided on. We also report a genre breakdown for both scoring scenarios in Table 9. Overall, the higher *fuzzy match* scores indicate that the vast majority of automatically generated connectives are at least reasonable, even though annotators sometimes decide that a more natural sounding connective is possible.

For comparison, we also compute a majority baseline for each RST relation in GDTB test. The baseline predicts the most frequent connective given the RST relation that spawns the GDTB relation. We find that this majority baseline has an exact match score of 51%. We find that the majority baseline produces a fuzzy match score of 88%, indicating that the RST relation is a very strong signal towards the set of PDTB-valid connectives. However, we find that our model provides substantially more natural connectives, leading to higher exact match scores. A genre breakdown for the majority baseline is reported in Table 10 below.

D Implicit Connective Prediction Prompt

Table 11 provides example prompts that were used to train Flan-T5 for the connective prediction task. The selected examples come from implicit relations in the training split of PDTB v3 used in the DISRPT shared task (Braud et al., 2023), which are what the model was trained on. For consistency, we label the argument spans as 'Sentence 1' and 'Sentence 2', regardless of whether the relation is inter- or intra-sentential.

Genres	Exact Match Accuracy	Fuzzy Match Accuracy
academic	0.710	0.871
bio	0.814	0.907
conversation	0.885	0.962
court	0.846	0.923
essay	0.727	0.818
fiction	0.679	0.839
interview	0.735	0.882
letter	0.900	0.950
news	0.882	0.941
podcast	0.875	0.875
reddit	0.811	0.865
speech	0.634	0.829
textbook	0.756	0.854
vlog	0.958	0.958
voyage	0.789	0.921
how-to	0.878	0.939

Table 9: Exact and Fuzzy Match Accuracy of Connective Prediction by Genres (Fine-tuned Connective Prediction Model).

Genres	Exact Match Accuracy	Fuzzy Match Accuracy
academic	0.548	0.967
bio	0.558	0.930
conversation	0.731	1.0
court	0.308	0.846
essay	0.318	0.818
fiction	0.500	0.928
interview	0.441	0.853
letter	0.550	0.800
news	0.471	0.882
podcast	0.250	0.750
reddit	0.567	0.865
speech	0.537	0.756
textbook	0.463	0.854
vlog	0.625	0.958
voyage	0.447	0.895
how-to	0.612	0.878

Table 10: Exact and Fuzzy Match Accuracy of Connective Prediction by Genres (Majority Baseline).

E Corrected Relation Correspondences

Figure 3 provides the confusion matrix for manually corrected relations in the test set (corrected relations on the y-axis) and their originally predicted relation as outputted by the conversion process (x-axis). The figure indicates that, apart from errors being overall rare, some relation predictions are completely reliable (e.g. Hypophora are trivial to predict given RST's QUESTION-ANSWER annotations). The most frequent error type is inferring a

Input	Output
Sentence 1: In July , the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos . Sentence 2: By 1997 , almost all remaining uses of cancer-causing asbestos will be outlawed . Relations: contingency.cause.reason,contingency.purpose,contingency.cause.result	as a result
Sentence 1: Sales figures of the test-prep materials are n't known , but their reach into schools is significant . Sentence 2: In Arizona , California , Florida , Louisiana , Maryland , New Jersey , South Carolina and Texas , educators say they are common classroom tools . Relations: contingency.condition,comparison.contrast,expansion.level-of-detail,expansion.manner,expansion.conjunction,expansion.instantiation,contingency.negative-condition	for example
Sentence 1: Choose 203 business executives , including , perhaps , someone from your own staff , Sentence 2: and put them out on the streets , Relations: temporal.asynchronous.precedence,temporal.synchronous,temporal.asynchronous.succession,expansion.conjunction	then

Table 11: Example Prompts used for the Connective Prediction Task.

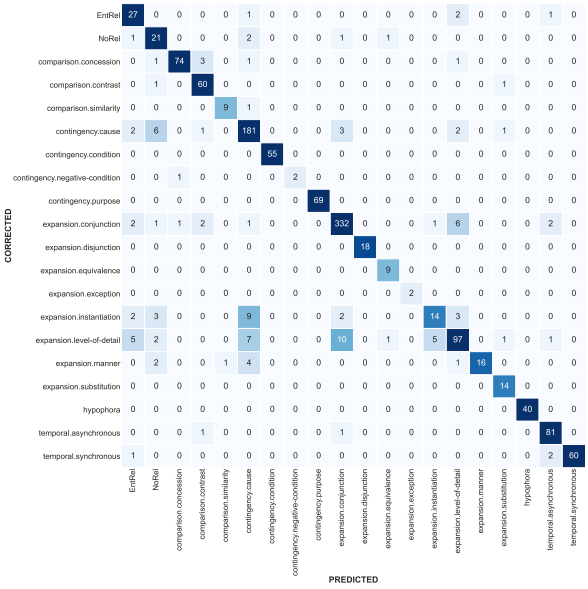


Figure 3: Confusion Matrix for Corrected Relations and their Initially Predicted Labels.

CAUSE relation where annotators felt a less marked EXPANSION was warranted. Generally speaking, correction into an EXPANSION category was the most common type, likely because implicit EXPANSION connectives, such as 'and', 'in fact', or 'specifically' are among the easiest to insert between sentence pairs.

F Implementation Details

The connective prediction task within the implicit module was conducted using NVIDIA RTX A6000 GPUs with 64GB RAM. Experiments related to DisCoDisCo's relation classification task were conducted on 1 NVIDIA Tesla L4 GPU with 24GB GPU Memory on Google Cloud Platform. For DisCoDisCo, overall we followed the original hyperparameters and training settings therein.[7] However, we did not use any hand-crafted features proposed in the original work as such features are not available for GDTB and show degradation for PDTB v3, according to Gessler et al. (2021).

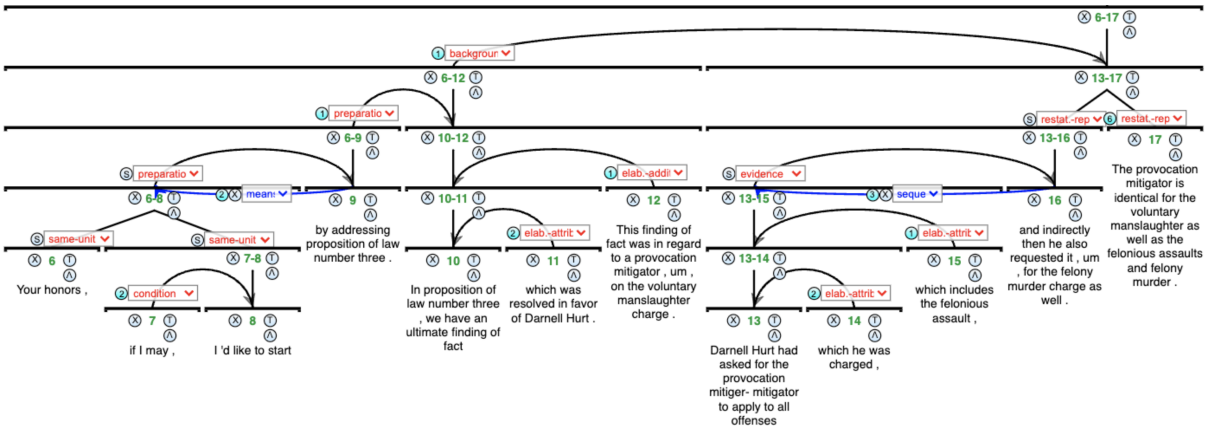
G Examples of GDTB based on the Conversion Process

Figure 4 provides an illustration of an RST fragment from a document in GUM as well as a set of the PDTB-style relations spawned from the gold RST annotations given the conversion process described in Section 3.2, covering explicit, implicit, and entity relations. "–" in the satellite EDU column indicates that the relation is multi-nucleus, where both EDU spans are considered nuclei (such relations still have 2 EDU spans, from which PDTB-style *Arg1* and **Arg2** are spawned). "–" in the connective column means that the relation was not signaled by connectives (and was either signaled by other signals or unsignaled).

H Full Results and Confusion Matrices

Table 12 presents the accuracy of the GDTB test set from both the within-corpus and cross-corpus models by genres and relation types. Since some relation types such as ALTLEX and HYPOPHORA are very rare, their scores are not available. In addition, we provide four confusion matrices in Figure 5 that give a better idea of what PDTB v3 sense labels are prone to errors overall as well as for the major relation types including explicit, implicit, and ALTLEX relations. Unsurprisingly, EXPANSION.CONJUNCTION is often the sense label that models tend to overpredict. In particular, it is commonly confused with CONTINGENCY.CAUSE across the board, but it tends to be more easily confused with TEMPORAL.ASYNCHRONOUS and EXPANSION.LEVEL-OF-DETAIL for implicit relations. For explicit relations, EXPANSION.CONJUNCTION is often confused with EXPANSION.LEVEL-OF-DETAIL.

[7] https://github.com/gucorpling/DisCoDisCo

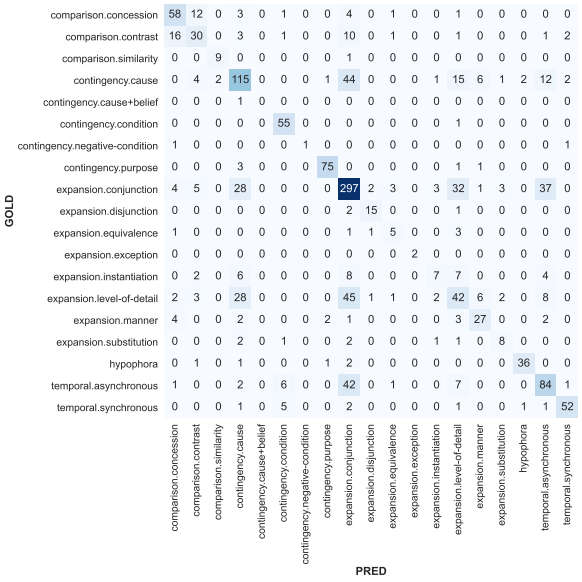


ORIGINAL GUM-RST/eRST ANNOTATIONS			CORRESPONDING GDTB INSTANCES				
RELATION	EDUs (nucleus)	EDUs (satellite)	CONN	SENSE	ARG1	ARG2	RELTYPE
contingency-condition	[8]	[7]	if	contingency.condition.arg2-as-cond	[6, 8]	[7]	explicit
mode-means	[6, 8]	[9]	by	expansion.manner.arg2-as-manner	[6, 7, 8]	[9]	explicit
mode-means	[6, 8]	[9]	by	contingency.purpose.arg1-as-goal	[6, 7, 8]	[9]	explicit
organization-preparation	[10, 11, 12]	[6, 7, 8, 9]	in particular	expansion.level-of-detail.arg2-as-detail	[6, 7, 8, 9]	[10, 11]	implicit
elaboration-additional	[10, 11]	[12]	specifically	expansion.level-of-detail.arg2-as-detail	[10, 11]	[12]	implicit
context-background	[13, 14, 15, 16, 17]	[6, 7, 8, 9, 10, 11, 12]	--	EntRel	[12]	[13, 14, 15, 16]	entrel
joint-sequence	[13, 14, 15, 16]	--	and	expansion.conjunction	[13, 14, 15]	[16]	explicit
joint-sequence	[13, 14, 15, 16]	--	then	temporal.asynchronous.precedence	[13, 14, 15]	[16]	explicit
joint-sequence	[13, 14, 15, 16]	--	also	expansion.conjunction	[13, 14, 15]	[16]	explicit
restatement-repetition	[13, 14, 15, 16, 17]	--	in other words	expansion.equivalence	[13, 14, 15, 16]	[17]	implicit

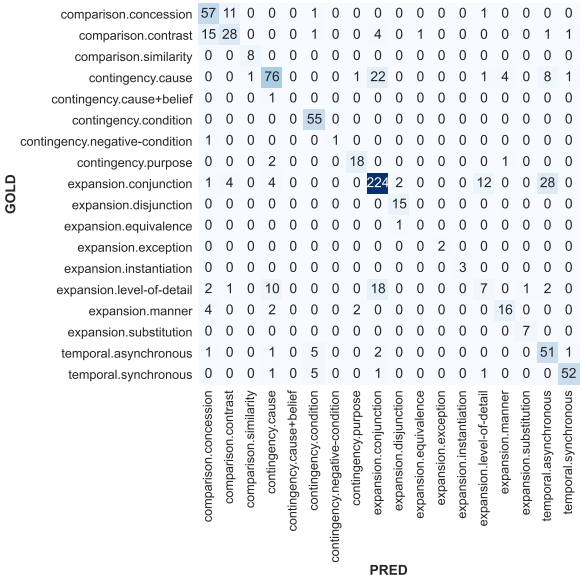
Figure 4: Examples of GDTB based on Gold GUM-RST Annotations and the Corresponding PDTB-style Instances.

	within-corpus							cross-corpus						
	overall	explicit	implicit	altLex	altLexC	hypophora			overall	explicit	implicit	altLex	altLexC	hypophora
academic	0.6977	0.7826	0.6111	0.5	–	–		*academic*	0.686	0.6739	0.6944	0.75	–	–
bio	0.6818	0.7949	0.5778	0.6667	1	–		*bio*	0.5227	0.7949	0.2667	0.6667	1	–
conversation	0.5669	0.6533	0.2857	0	–	0.8667		*conversation*	0.4803	0.64	0.1143	0	–	0.6
court	0.5	0.5581	0.2667	–	–	1		*court*	0.4667	0.5349	0.2667	–	–	0.5
essay	0.6557	0.8333	0.4167	0	–	–		*essay*	0.5738	0.75	0.3333	0	–	–
fiction	0.6098	0.7742	0.4138	1	–	1		*fiction*	0.4797	0.6935	0.2759	0	–	0
interview	0.6477	0.8788	0.3243	0.3333	–	1		*interview*	0.5455	0.8485	0.2432	0.3333	–	0.6667
letter	0.6458	0.8846	0.3636	–	–	–		*letter*	0.5417	0.6923	0.3636	–	–	–
news	0.72	0.8387	0.5556	0	–	–		*news*	0.6	0.5484	0.7222	0	–	–
podcast	0.6667	0.8182	0.25	–	–	–		*podcast*	0.5556	0.6364	0.3333	–	–	–
reddit	0.6833	0.8472	0.439	0	–	0.5		*reddit*	0.5917	0.7222	0.3659	0	–	0.6667
speech	0.6196	0.7727	0.4773	0.5	–	–		*speech*	0.6739	0.7727	0.5682	0.75	–	–
textbook	0.6667	0.8070	0.4634	0.5	–	1		*textbook*	0.5588	0.6491	0.4878	0	–	0
vlog	0.6486	0.6917	0.4615	0.5	–	–		*vlog*	0.5743	0.6417	0.3077	0	–	–
voyage	0.6029	0.6667	0.5526	–	–	–		*voyage*	0.6176	0.6	0.6316	0	–	0
how-to	0.7034	0.8281	0.5556	–	–	–		*how-to*	0.6102	0.7188	0.4815	–	–	–

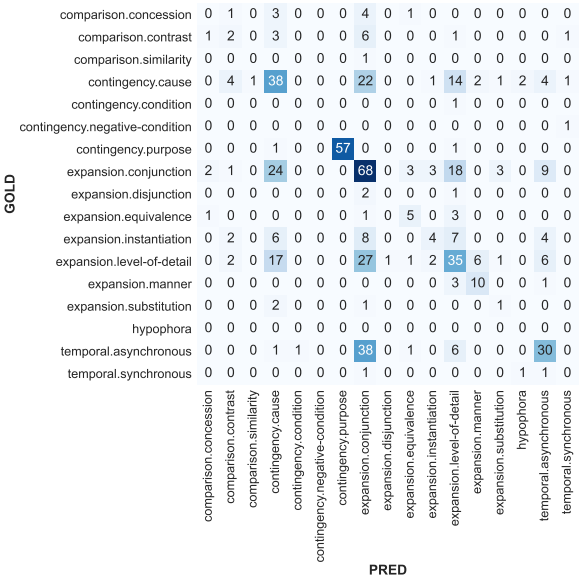
Table 12: Accuracy of GDTB `test` by Genres and Relation Types. "–" means such relation types are not available.



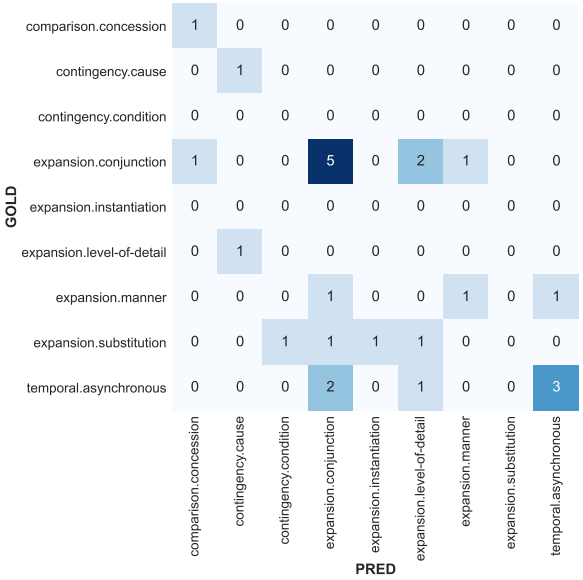
(a) Overall.



(b) Explicit Relations.



(c) Implicit Relations.



(d) AltLex Relations.

Figure 5: Confusion Matrices for GDTB `test` from the within-corpus Experiment.