The distribution of discourse relations within and across turns in spontaneous conversation

S. Magalí López Cortez **Cassandra L. Jacobs**
Department of Linguistics
University at Buffalo
Buffalo, NY, USA
solmagal;cxjacobs@buffalo.edu

Abstract

Time pressure and topic negotiation may impose constraints on how people leverage discourse relations (DRs) in spontaneous conversational contexts. In this work, we adapt a system of DRs for written language to spontaneous dialogue using crowdsourced annotations from novice annotators. We then test whether discourse relations are used differently across several types of multi-utterance contexts. We compare the patterns of DR annotation within and across speakers and within and across turns. Ultimately, we find that different discourse contexts produce distinct distributions of discourse relations, with single-turn annotations creating the most uncertainty for annotators. Additionally, we find that the discourse relation annotations are of sufficient quality to predict from embeddings of discourse units.

1 Introduction

Discourse relations (DRs) such as Elaboration, Background and Explanation, hold between discourse units contributing to the coherence of a text. Annotation of discourse relations has received attention for its relevance to discourse parsers, with applications in question answering systems (e.g. Jansen et al., 2014), text summarization (e.g. Liu and Chen, 2019), sentiment classification (e.g. Kraus and Feuerriegel, 2019), and machine translation (e.g. Meyer and Popescu-Belis, 2012). However, most of the annotated data and systems have focused on written language, with a few exceptions (e.g., Tonelli et al., 2010; Zeldes, 2017; Scholman et al., 2022). In spoken dialogue or multiparty conversation, participants must quickly juggle a variety of tasks, such as responding to another person to solve a problem (Levinson and Torreira, 2015) or negotiating the question under discussion (Roberts, 2012), often under considerable time pressure that is less present in written production. In addition to these time demands, it is unclear whether spontaneous conversation demonstrates the same patterns of discourse relations as observed in written language (see Crible and Cuenca, 2017, for a discussion of spoken vs. written use of discourse markers).

Perhaps unsurprisingly, the vast majority of work on discourse relations has focused either on written texts, especially news text (Carlson et al., 2003; Prasad et al., 2008, 2018), or highly structured conversations that are constrained by a particular game (Afantenos et al., 2015; Asher et al., 2016). Some recent corpora contain spoken monologues (Scholman et al., 2022), and spoken conversations (Tonelli et al., 2010; Zeldes, 2017), but the field still largely lacks annotated corpora of spontaneous dialogue.

Thus, our goal is to present the first efforts towards an annotated corpus of DRs for spontaneous spoken conversation, with particular attention to relations across different contexts within a conversation. We analyze the patterns of DR annotation within and across speakers and within and across turns and test the coherence of annotators' decisions.

2 Related Work

Most currently available corpora annotated with DRs have focused on written language or spoken monologues. An exception is the Georgetown University Multilayer (GUM) corpus (Zeldes, 2017), which has a set of conversations annotated within Rhetorical Structure Theory (RST, Mann and Thompson, 1987), following the guidelines of the RST Discourse Treebank (RST-DT, Carlson et al., 2003). But it is an open question whether the DRs that have been identified for news texts are appropriate for conversational data. Tonelli et al. (2010) adapt the PDTB framework to annotate a subset of a corpus of Italian conversations about software and hardware troubleshooting, and suggest modifications to the framework to account for

Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023), pages 156–162

spoken data.

Discourse relations corpora have usually been annotated by experts, but some recent corpora have been annotated by novice annotators, such as university students, in the case of the GUM corpus (Zeldes, 2017), or crowdsourced workers, in the case of the DiscoGEM corpus (Scholman et al., 2022). GUM was annotated using RST as part of a Corpus Linguistics class, while DiscoGEM was annotated following the Penn Discourse Treebank (PDTB, Prasad et al., 2008, 2018) framework, using a method for crowdsourcing annotations introduced in Yung et al. (2019), and using a multi-label approach. The present work deviates from prior work in its focus on conversational data and the use of Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003) alongside the STAC corpus (Asher et al., 2016) guidelines.

3 Discourse relation annotation

In this work, we focus on a subset of 19 dialogues from the Switchboard Corpus (Godfrey et al., 1992). This corpus contains informal language and has been the subject of study of numerous analyses of dialogue within linguistics (Jaeger and Snider, 2013; Reitter and Moore, 2014). In it, two strangers are presented with a topic (e.g., childcare) that they must discuss with each other, but the dialogues are otherwise not tightly constrained. Annotating Switchboard will provide us with a more complete understanding of the use and generality of discourse relations across linguistic contexts and genres.

Following the annotation procedure in the STAC corpus (Asher et al., 2016), we identified a subset of suitable elementary discourse units (EDUs) for annotation by parsing each turn into a dependency structure and included only those turns with at least two roots or verbs. Then, we segmented each of these turns into their respective EDUs. Using these segmentations, we identified EDU candidates for discourse relations that were either within-turn (same speaker) or across two turns (different speakers, or the same speaker), where the two turns were adjacent in the case of different speakers, or only interrupted by one turn, in the case of same speaker. We provide a representative set of these pair types in Table 1 under the Explanation, Comment, and Result examples, respectively.

3.1 Elementary Discourse Units

Elementary discourse units (EDUs) are typically defined as non-overlapping text spans (Mann and Thompson, 1987), which perform some basic discourse function (Asher and Lascarides, 2003), typically at the level of clauses. However, conversational EDUs may not necessarily contain a main verb (e.g., clarification questions: “Saginaw?”) or may be incomplete or interrupted (e.g., “and so–”). So, we define EDUs in Switchboard similarly to written text, with some modifications to account for variability due to spoken language. In particular, our modifications account for noise; non-linguistic communication (e.g., laughter); restarts; and disfluencies (e.g., “uh” or “um”). Additionally, we use complex discourse units (CDUs), which are combinations of EDUs which function together as an argument to a DR (Asher and Lascarides, 2003).

3.2 Relation categories

Discourse relations (DRs) were selected from Segmented Discourse Representation theory (SDRT, Asher and Lascarides, 2003), following the annotation manual for the STAC corpus (Asher et al., 2012). 11 out of 16 relation labels used in Asher et al. (2012) were selected, based on a pilot annotation. We selected the most common relations in an attempt to minimize the number of choices presented to annotators, but the set is non-exhaustive. An "Other" category was added for cases in which none of the selected labels applied. Table 1 shows the list of DRs together with representative examples.

3.3 Annotators

The present study recruited 114 students enrolled in a computational linguistics course grouped into 19 teams consisting of approximately 5 members who annotated the dyads. Each team received a conversation for annotation. Annotations were performed individually, but groups then discussed their work and submitted a report as a team. One team was excluded because they completed their annotations together and submitted a single set of labels. Students were trained to identify discourse relations using a short quiz and live training with the instructor of the course. Annotators were provided with guidelines to which they could refer back, and they had read and annotated the conversation in three previous tasks before annotating discourse relations, to ensure that they were familiar with the

Relation	Discourse Units
Acknowledgement	A: ‖ *it starts recording now.* ‖ B: ‖ **Okay.** ‖
Background	A: ‖ *I'm, we're originally from another state* ‖ **and I know** ‖ **in the state we were from that they did that t-, similar type thing.** ‖
Clarification Question	A: ‖ *We live in the Saginaw area.* ‖ B: ‖ **Saginaw?**‖
Comment	B: ‖ *They seem to be having a real good response.* ‖ A: ‖ **That's pretty good.** ‖
Continuation	A: ‖ *I work off and on just temporarily* ‖ **and usually find friends to babysit,** ‖
Contrast	A: ‖ *I don't work, though,* ‖ **but I used to work and,** ‖
Elaboration	A: ‖ *in the state we were from that they did that t-, similar type thing.* ‖ **The city brought ought,** ‖ **you know,** ‖ **set tr-, separate trash cans** ‖ **and you separated your stuff** ‖
Explanation	A: ‖ *and they discontinued them* ‖ **because people were coming and dumping their trash in them.** ‖
Narration	A: ‖ *and you put it in there* ‖ **and they took it,** ‖
Question-Answer Pair	B: ‖ *Saginaw?* ‖ A: ‖ **Uh-huh.**‖
Result	B: ‖ *No,* ‖ *I just, I noticed* ‖ *it Iowa and other cities like that, it's a nickel per aluminum can.* ‖ A: ‖ Oh. ‖ B: ‖ **So you don't see too many thrown out around the** ‖ **[laughter]** ‖ **streets.** ‖
Other	None of the labels applies

Table 1: Representative discourse unit pairs for annotated discourse relations. The first argument to the discourse relation is shown in *italics* and the second one in **bold**. A and B correspond to speakers, and double pipes (‖) represent boundaries between elementary discourse units.

topics and speakers in each dyad.

3.4 Annotation procedure

Annotators were presented with pairs representing either an EDU or CDU (π_1) and another EDU or CDU (π_2). Annotators were shown two spans of text π_1 and π_2 with π_1 presented in italics and π_2 presented in bold face font in the annotation software Prodigy (Montani and Honnibal, 2018), with two preceding and two subsequent turns for context. Annotators were asked to determine the relation between π_1 and π_2 from a list of the DR categories in Table 1. If annotators thought that no relation was present, they were told to reject the item and move on to the next pair. Critically for our research question, annotators could mark several relations for a pair of EDUs simultaneously. In addition to labeling discourse relations, annotators were also asked to provide a confidence rating on a scale from 1-5, but we leave these analyses for future work. In total, each annotator provided judgments for an average of 25 EDU pairs across 464 total pairs.

In the next section, we test whether annotators show greater uncertainty about discourse relations in different discourse contexts. We analyze the distribution of their labels to assess whether discourse relations in conversation vary in their contexts of use.

4 Uncertainty in the annotation of discourse relations

Different EDU pairs in the present annotation task were drawn either from the same turn, or across turns but within or across speakers. Thus, we can assess how much discourse relations vary by the placement of an utterance in a dialogue. Given the complex dynamics in dialogue, we expect to find significant differences in discourse relation use across different discourse contexts. We visualize the distribution of the relations in Figure 1.

Annotators generally selected more discourse relations per EDU pair in the single-turn case, with an average of 8.16 relations per team or 1.60 per annotator. When EDUs spanned turns within a single speaker, groups selected significantly fewer relations (average = 7.29, $t(302) = -2.16$, $p < .05$). Groups likewise selected even fewer relations for EDUs between two speakers (average = 6.51, $t(314) = -2.54$, $p < .05$). On its face, this pattern appears surprising, because it suggests that annotators find more relations appropriate for single-speaker productions. However, an alternative interpretation of these results is that annotators may instead have been uncertain about the distinctions between the different discourse relations. This second interpretation is corroborated by post-hoc poll data from 35 annotators, of whom 32 (91.4%) stated that the selection of discourse relations was best suited to annotating cross-speaker

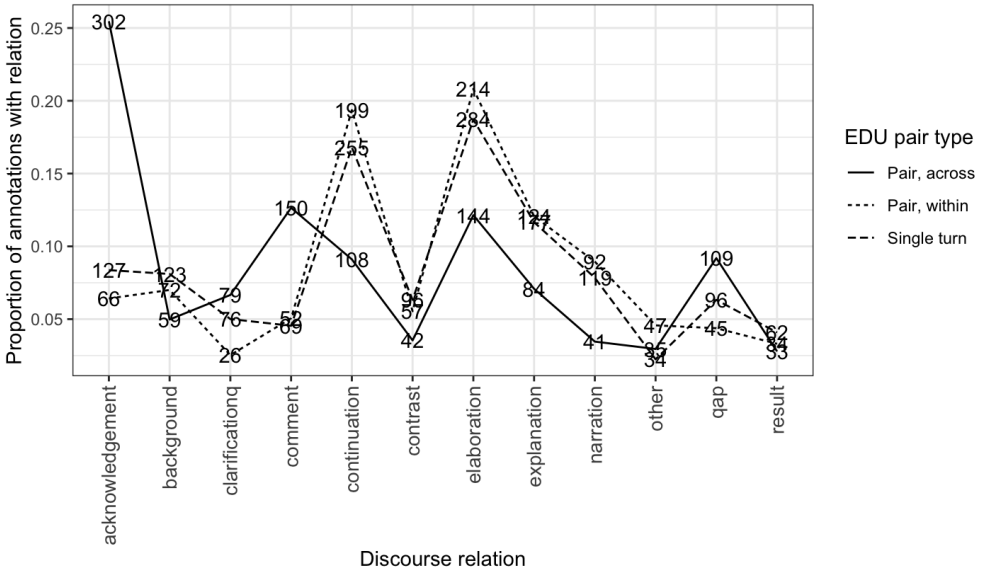

Figure 1: Distribution of discourse relations across three EDU pair types. y axis represents proportions of EDU pairs with a given label; numbers represent the count of a label within a discourse context category.

EDU pairs. Future work will require recruiting greater numbers of annotators to be able to distinguish between these two hypotheses.

4.1 Inter-annotator agreement

We computed measures of inter-annotator agreement for multilabel tasks using Marchal et al. (2022). This approach uses bootstrap sampling to estimate the chance frequencies of DRs in a multilabel dataset to provide a baseline for agreement between annotators.

We summarize the results of this analysis in Table 2. Following Marchal et al. (2022), we computed observed, expected and adjusted agreement for six measures. Soft-match agreement uses the intersection of labels selected by two annotators; boot-match corrects the expected agreement by using the bootstrapping method (as opposed to ignoring non-intersecting labels); augmented kappa uses DR labels weighted according to the number of labels annotated for each item; precision and recall are calculated as the proportion of intersecting DR labels over the set of labels selected by the first and second annotator, respectively; F1 is the usual harmonic mean between precision and recall.

Both observed and adjusted agreement metrics were well above chance using the bootstrapping method proposed by Marchal et al. (2022). Agreement is in general modest (Landis and Koch, 1977), which may be partly due to the challenging nature of the DRs annotation task (Spooren and Degand, 2010), and partly due to annotators' uncertainty on DR labels across different context types.

	observed	expected	adjusted kappa
`soft-match`	0.43	0.11	0.36
`augmented`	0.27	0.11	0.18
`boot-match`	0.43	0.21	0.27
`boot-rec.`	0.33	0.14	0.22
`boot-prec.`	0.36	0.17	0.23
`boot-F1`	0.32	0.13	0.21

Table 2: Outputs of Marchal et al. (2022) inter-annotator agreement analysis.

Relation	Intercept	Different speaker	Within turn
Background	-2.73	1.96	0.24
Clarification Q.	0.02	0.41	-0.43
Comment	-1.69	0.84	0.46
Continuation	-1.60	2.35	0.31
Contrast	-2.23	2.07	0.17
Elaboration	-0.78	1.95	0.13
Explanation	-1.40	2.00	0.09
Narration	-3.29	2.68	0.53
Other	-3.54	2.20	1.12
Q-A Pair	-0.81	0.64	-0.01
Result	-2.35	1.63	-0.00

Table 3: Coefficient estimates from a multiclass logistic regression predicting each annotation label.

4.2 Predicting relation selection

We use a model comparison approach to understand the contributions of discourse context (within/across speakers and within/across turns) to relation annotation by first constructing a null model that estimates the base rates of each dis-

course relation. Then, we constructed a multiclass logistic regression model containing the discourse context variables of interest, which significantly improved fit to the annotation data ($X^2(22) = 447.98$, $p < .001$). This improvement in fit suggests that the distribution of discourse relations that are identified by annotators is distinct across contexts. Adding the annotator group/topic also significantly improved fit beyond the model containing the contextual variables alone ($X^2(198) = 900.06$ $p < .001$). We summarize the results of this final model in Table 3.[1]

An informal evaluation of the coefficients suggests that discourse relations are not uniformly distributed across contexts. Intuitively, Acknowledgements, Clarification Questions, Comments, and Question-Answer Pairs are more likely across speakers than within. Additionally, Continuations, Elaborations, Explanations, and Narrations are more likely to occur within a single speaker. The pattern of results is more unclear when comparing EDUs that are produced by a single speaker but which occur either within or across turns. For example, relations such as Clarification Questions are less likely to occur within a turn than across turns.

4.3 Classifier for relations

To validate the quality of the annotations, we built a model to classify EDU pairs into discourse relations. We reasoned that if annotators are following the guidelines and use information about the EDU pairs, then a classifier should be able to predict DR labels. We encoded the first EDU or CDU (π_1) and the second (π_2) as the two "sentences" in the next sentence prediction architecture of BERT (Devlin et al., 2019). This enables the classifier to represent the π_1 and π_2 components somewhat separately.

We built a classifier head trained on the resulting embeddings without fine-tuning to predict each individual annotator label. We chose to model each annotator label individually to learn agreement/majority class implicitly because prior studies have shown that this improves generalization (Yung et al., 2022). We use a leave-one-conversation-out training procedure, in which we test a ridge regression classifier on all of the annotations from a single conversation while we train it on all other annotations across the other conversations. This ensures minimal memorization of specific turns within a conversation, which is critical given our multilabel annotation approach.

Strict annotation-level accuracy to predict each selected label from all annotators was quite poor, with macro average precision at .21, recall at .19, and F1 at .19. However, recall was substantially higher when considering whether the top guess belonged to the set of all labels provided by annotators, at .76 overall and .71 averaged by group.

To quantify the uncertainty of the annotators across different contexts, we leverage the classifier to produce a label distribution for a given (π_1, π_2) pair. We then compute the cross-entropy between the model's predictions and annotators' gold label distributions, collapsing across all annotations for an EDU pair. Overall, cross entropy between model predictions and annotator labels was highest for the single-turn case, with (mean $= 0.43$), but lowest for EDUs between two speakers (mean $= 0.38$), suggesting greater uncertainty in label assignment.

5 Discussion

In two experiments, we demonstrated that novice DR label annotations in a single turn are more difficult than across turns. We showed that including discourse context (within/across speaker and within/across turn) to a logistic regression model significantly improves fit to our annotation data. A classifier trained to predict DR labels from embeddings of (π_1, π_2) pairs showed modest success for recall of any of the annotations, but poor precision and recall overall. A comparison of this classifier's predictions and annotators' gold label distributions revealed greater uncertainty for the annotation of discourse relations within a single turn.

These results demonstrate that different conversational contexts are associated with different distributions of discourse relations. The uncertainty of choice of discourse relations within a turn may be due to several factors. DRs that typically occur across adjacent turns and across speakers (e.g., Acknowledgements) might have clearer signals. At the same time, DRs that occur more frequently within speakers, and, in particular, within a turn, might be more ambiguous, or might co-occur with other relations. More work is necessary to disentangle uncertainty about the identity of the best fit relation from whether multiple relations are appropriate.

[1] Due to the multilabel nature of the annotation task and the one-versus-rest training for the multiclass model, coefficients for each DR are not independent, were not estimated jointly, and should be interpreted broadly as representing separate logistic regressions.

Limitations

The current work is limited by the size of the dataset and the nature of spontaneous conversation. While the discourse relations proposed as part of this work were selected to be general and build on categories from the literature, the list is not exhaustive and it is likely that these relations may be culturally, linguistically, and situationally specific. Future work in this area should validate the generality of the discourse relation system used in this work.

The selection of EDUs and CDUs for annotation is also non-exhaustive; additional segments could be included in future work.

Annotation quality is also a practical limitation. Annotation for discourse relations typically results in low-agreement data, even among expert annotators (e.g., DiscoGEM; Scholman et al., 2022). Even though our research questions focus on this disagreement as a positive, other researchers may require greater numbers of annotations in order to obtain a gold label.

Ethics Statement

We are not aware of ethical issues associated with the texts used in this work. Students participated in the annotation task as part of course credit but annotation decisions were not associated with their performance in the course.

Acknowledgements

We would like to thank Jürgen Bohnemeyer and three anonymous reviewers for feedback on a previous version of this paper.

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. 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*. Cambridge University Press.

Nicholas Asher, Vladimir Popescu, Philippe Muller, Stergos Afantenos, Anais Cadilhac, Farah Benamara, Laure Vieu, and Pascal Denis. 2012. Manual for the analysis of settlers data. *Strategic Conversation (STAC). Université Paul Sabatier*.

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.

Ludivine Crible and Maria-Josep Cuenca. 2017. Discourse markers in speech: distinctive features and corpus annotation. *Dialogue and Discourse*, 8(2):149–166.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

T Florian Jaeger and Neal E Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986.

Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6.

Zhengyuan Liu and Nancy Chen. 2019. Exploiting discourse-level segmentation for extractive summarization. In *Proceedings of the 2nd Workshop on New*

Frontiers in Summarization, pages 116–121, Hong Kong, China. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. *Rhetorical Structure Theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *EACL 2012: Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, CONF, pages 129–138.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*.

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 Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

D. Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

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.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

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).

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics.

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