Incorporating Annotator Uncertainty into Representations of Discourse Relations

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

Abstract

Annotation of discourse relations is a known difficult task, especially for non-expert annotators. In this paper, we investigate novice annotators' uncertainty on the annotation of discourse relations on spoken conversational data. We find that dialogue context (single turn, pair of turns within speaker, and pair of turns across speakers) is a significant predictor of confidence scores. We compute distributed representations of discourse relations from co-occurrence statistics that incorporate information about confidence scores and dialogue context. We perform a hierarchical clustering analysis using these representations and show that weighting discourse relation representations with information about confidence and dialogue context coherently models our annotators' uncertainty about discourse relation labels.

1 Introduction

Discourse relations (DRs) are those relations such as Elaboration, Explanation, Narration, which hold between discourse units. The task of labeling DRs is known to pose difficulties for annotators (Spooren and Degand, 2010), as sometimes more than one interpretation may be possible (Scholman et al., 2022; Webber, 2013).

Recent studies have shown that allowing for multiple labels in annotation can improve the performance of discourse parsers (Yung et al., 2022). Scholman et al. (2022) test different label aggregation methods in a crowdsourced corpus annotated by 10 workers and find that probability distributions over labels better capture ambiguous interpretations of discourse relations than majority class labels. (1) shows an example from their corpus, where the relation between the second and third sentences (in italics and bold, respectively), was interpreted as Conjunction by four annotators and Result by five annotators.

(1) It is logical that our attention is focused on cities. *Cities are home to 80% of the 500 million or so inhabitants of the EU.* **It is in cities that the great majority of jobs, companies and centres of education are located.** (adapted from DiscoGeM, Europarl genre; Scholman et al., 2022, italics and bolding are ours.)

Annotating the discourse relation between these two sentences with both Conjunction and Result captures different possible interpretations of the relation between these segments. For example, the two sentences may contain two conjoined facts about cities, but can also be perceived as describing a causal relation between the first and second sentence (i.e., as cities are home to the largest part of the population, most jobs, companies and educational institutions are located there).

In this work, we investigate which relations are distributionally similar or co-occurring in multilabel annotations of spontaneous conversations. We are particularly interested in how novice annotators interpret discourse relation categories when annotating spoken conversational data. We collect annotations of DRs from Switchboard telephone conversations (Godfrey et al., 1992), allowing for multiple labels, and ask for confidence scores. We find that confidence scores vary significantly across dialogue contexts (single turn vs. pairs of turns produced by the same speaker vs. pairs of turns produced by different speakers). We incorporate information about these three dialogue context types and confidence scores into distributed representations of discourse relations. A clustering analysis shows that discourse relations that tend to occur across speakers cluster together, while discourse relations which tend to occur within a speaker, either in the same turn or different turns, form their own cluster.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 530–537
September 11–15, 2023.

2 Annotation of Discourse Relations

Our analyses are built on the dataset collected in López Cortez and Jacobs (2023), who selected 19 conversations from Switchboard[1], a corpus consisting of telephone conversations between pairs of participants about a variety of topics (e.g. recycling, movies, child care). We chose this corpus because it contains informal, spontaneous dialogues, and because it has been used within linguistics in various studies on conversation (Jaeger and Snider, 2013; Reitter and Moore, 2014).

2.1 Discourse Units

An initial set of turns for annotation was selected by using spaCy's dependency parser (Honnibal et al., 2020, version 3.3.1) to select turns with two or more ROOT or VERB tags. We define a turn as each segment of dialogue taken from Switchboard. We note that an utterance produced by one speaker (A) may take place during a continuous utterance by another speaker (B). Switchboard splits A's utterance into two turns in these cases. We return to this point in the Discussion.

We manually segmented these turns into elementary discourse units (EDUs). The main criteria for segmenting turns into EDUs was that the unit performs some basic discourse function (Asher and Lascarides, 2003). By default, finite clauses are considered EDUs, as well as comment words like "Really?" or acknowledgments such as "Uh-huh" or "Yeah." Cases of interruptions and repairs were segmented if they constituted a turn in Switchboard, as in example (2a), and when they contained a verb, as in example (2b). Cases of repetition as in (2c) were not considered separate EDUs. We segmented disfluencies ("uh") and some non-verbal communication ("[laughter]") but we did not select these for discourse relation labeling.

(2) a. B: || So you don't see too many thrown out around the || [laughter] || streets. ||
A: || Really ||
B: || Or even bottles. ||

b. B: || I think, || uh, || I wonder || if that worked. ||

c. A: || What kind of experience do you, do you have, then with child care? ||

[1]We discarded the annotations from one conversation because the annotators did not follow the guidelines.

Because many EDUs are very short, we selected pairs of elementary discourse units and complex discourse units (CDUs) for discourse relation annotation. CDUs consist of two or more EDUs that constitute an argument to a discourse relation (Asher and Lascarides, 2003). We use the term *discourse units* (DUs) to refer to both EDUs and CDUs.

2.2 Dialogue Contexts

We manually selected items for annotation across three different contexts: within a single turn, across two turns within a speaker, and across two immediately adjacent turns (two speakers). (3) shows an example for each context kind, with the first DU in italics and the second in bold. Example (3a) shows two discourse units within a speaker's turn. (3b) shows two discourse units uttered by the same speaker but that span across two different turns, interrupted by one turn. We did not include any constraint for the length of the interrupting turn. (3c) shows two DUs uttered by speakers in adjacent turns. We leave for future work the annotation of pairs of discourse units that may have a longer-distance relation with more turns in between DUs.

(3) a. A: || *and they discontinued them* || **because people were coming and dumping their trash in them.** ||

b. B: || No, || *I just, I noticed* || *in 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**.

c. A: || *We live in the Saginaw area.* ||
B: || **Saginaw?** ||

2.3 Taxonomy of Discourse Relations

The DRs chosen to annotate our corpus were adapted from the STAC corpus manual (Asher et al., 2012, 2016). STAC is a corpus of strategic multiparty chat conversations in an online game. Table 1 shows the taxonomy used. We selected 11 DRs based on a pilot annotation by the first author, and added an "Other" category for relations not included in the list of labels. We focused on a small taxonomy to minimize the number of choices presented to our novice annotators. We refer readers to López Cortez and Jacobs (2023) for details and examples of each relation in the taxonomy. Future work will include revising the taxonomy used.

Acknowledgement	Elaboration
Background	Explanation
Clarification Question	Narration
Comment	Question-Answer Pair
Continuation	Result
Contrast	Other

Table 1: Taxonomy of discourse relations.

2.4 Annotation Procedure

The annotation of discourse relations was done by students enrolled in a Computational Linguistics class. Students were divided into 19 teams of approximately 5 members each, and each team was assigned a conversation. The annotation was performed individually, but teams then discussed their work and wrote a report together. Annotators were trained using written guidelines, a quiz-like game, and a live group annotation demo.

We used the annotation interface Prodigy (Montani and Honnibal, 2018). Each display presented the two target discourse units plus two context turns before and two after. Annotators also had access to the entire conversation throughout the annotation task. Below the text, the screen showed a multiple choice list of discourse relations plus the "Other" category. We allowed for the selection of multiple labels following previous findings that allowing for multiple labels better captures ambiguous interpretations of discourse relations (Scholman et al., 2022) and improves the performance of discourse parsers (Yung et al., 2022).

Each display also asked for confidence scores in the range 1-5, corresponding to least to most confident. We did not pursue label-specific confidence scores but rather the confidence in the label(s) as a whole in the interest of minimizing annotator overhead. The results of this work show that per-label confidence scores or a slider-based approach may be informative and is a topic for future work. We include an example annotation item in Appendix C.

3 Dialogue Context as a Predictor of Confidence Scores

First we sought to understand how discourse relations and dialogue context (as defined above) influence annotator confidence. Because our confidence ratings data has multiple observations for each annotator, each team and each DU, it is hierarchical and thus benefits from being analyzed using hierarchical mixed effects models. Due to the ordinal nature of the ratings data, we use the cumulative link approach (CLMM; Liddell and Kruschke, 2018; Howcroft and Rieser, 2021) rather than model confidence scores as real-valued in linear regression. We first built a null model containing only random intercepts by annotator and compared it to a model containing an additional fixed effect and random slope by annotator for dialogue context type: single turn, across turns within speaker and across speakers (*kind*, dummy coded). A likelihood ratio test revealed a significant improvement in fit by adding kind as a predictor ($\chi^2(7) = 126.64, p < 0.001$). Adding random intercepts for DU pairs to account for annotation difficulty across DU pairs also led to a significant improvement in model fit beyond the model containing dialogue context kind ($\chi^2(1) = 195.01, p < 0.001$). This suggests that our annotators' confidence scores are sensitive to the context of DU pairs.

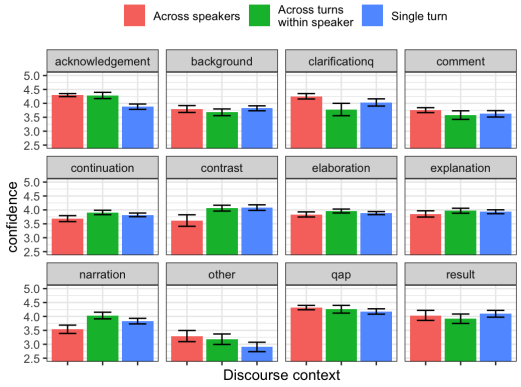


Figure 1: Confidence scores per context kind across discourse relations. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.

Figure 1 shows mean confidence scores per context kind across discourse relations. Confidence scores within a speaker both across and within turns received similar confidence ratings ($\beta = -0.13$, $z = -0.56$, $p =$ n.s.[2]), while annotators were significantly more confident for relation annotation across speakers ($\beta = 0.63$, $z = 3.05$, $p < .01$). The CLMM revealed that annotators used confidence scores between 3 and 5 overall, except for the label "Other", for which they selected lower confidence scores. Background received lower confidence scores overall. Continuation, Contrast and Narration received higher scores for contexts

[2]Not statistically significant.

within speaker. Comment and Result received higher scores for turns across speakers and single turn. For Elaboration and Explanation, mean confidence scores are very similar across the three contexts, with slightly higher scores for single turn and pairs of turns within speaker. Acknowledgment, Clarification Question ("clarificationq") and Question-Answer Pair ("qap") received higher scores for turns across speakers, which makes sense given the dialogic nature of these relations. However, these relations also received rather high confidence scores for single turn and pairs of turns within speaker, which is a bit surprising. We suspect this might be due to the context turns included for each pair of DUs, which might have led annotators to choose relations between discourse units other than for the pair of highlighted DUs. Future analysis will look closer at this aspect.

4 Distributed Representations from Discourse Relation Annotations

To model the similarity between discourse relations as perceived by annotators, we computed embedding representations of discourse relations. We extracted each n individual annotation containing **r**elation-**c**onfidence (r, c) tuples selected by a given annotator for a pair of DUs. We concatenate bag-of-relation vectors with one-hot encoded features representing the dialogue context kind, and multiply the count vector of annotated relations (either 1 or 0 for each relation) by the confidence score (1-5) for that pair of DUs. This weighting learns more from high confidence; an ideal reweighting may be possible with additional parameter search, possibly in conjunction with the CLMM outputs.

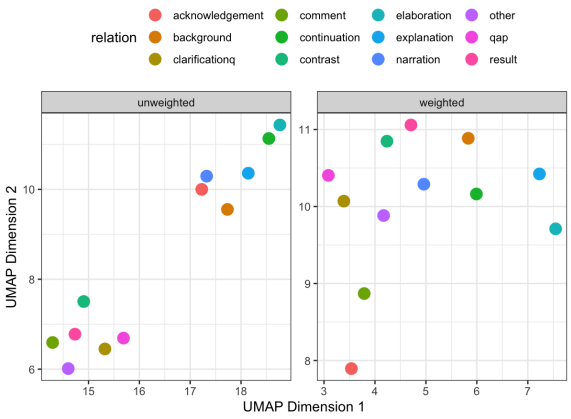
For an $n \times 1$ confidence ratings matrix C, an $n \times 12$ bag-of-relations matrix R, an $n \times 3$ discourse context matrix D for each annotation, we obtain an annotation matrix $A = C \times (R|D)$. We then obtain a square co-occurrence matrix O such that $O = A \cdot A^T$, which we factorize using Principal Component Analysis (without shifting the intercept following Levy and Goldberg, 2014). Each relation is thus represented as a vector that consolidates co-occurrences between all relations within a single annotator that are weighted by confidence score. We then projected these embeddings into two dimensions with UMAP (McInnes et al., 2018) and performed a hierarchical clustering analysis over the resulting coordinates due to the greater discriminability afforded by continuous distance metrics.

Informally, the UMAP coordinates appear more gradient in the representational space when confidence was included (right panel) than when it was not included (left panel). When context is not included, the UMAP coordinates primarily represent the frequency of labels in our corpus, which we include in Appendix A. We visualize the UMAP coordinates in Figure 2a. Figure 2b shows a dendrogram with the output clusters, colored according to the optimal number of clusters ($k = 2$), calculated using average silhouette widths (Levshina, 2022). There are two large clusters, one of which contains two sub-clusters with Background and Continuation, on the one hand, and Elaboration and Explanation on the other. In the other large cluster, Acknowledgement and Comment form a sub-cluster. These are very common relations between pairs of turns across speakers. Clarification Question and Question-Answer Pair form another sub-cluster, also common relations between pairs of turns across speakers, in close proximity to the Other label, which received a sub-cluster of its own. Narration and Contrast and Result, form the last sub-clusters, which we suspect is due in part to the frequencies of these relations (Schnabel et al., 2015). We include a dendrogram with the output clusters of a hierarchical clustering analysis performed with base bag-of-relations vectors (without context kind and confidence scores weight) in Figure 3 in Appendix B for comparison.

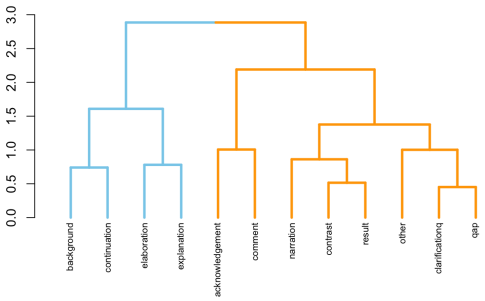
Currently, we provide these results as a proof of concept of the feasibility and interpretability of noisy labels produced by novice annotators. Importantly, annotations weighted by confidence produce coherent clusters of discourse relations. We envision applications of DR embeddings to several domains including dialogue generation, such that appropriate responses to input are partially conditioned on a latent or mixed combination of DRs.

5 Related Work

Annotation of discourse relations is usually done within Rhetorical Structure Theory (Mann and Thompson, 1987), as in the RST-DT (Carlson et al., 2003) and GUM (Zeldes, 2017) corpora, within Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003), as in the STAC (Asher et al., 2016) and Molweni (Li et al., 2020) corpora, or within the Penn Discourse Treebank framework (Prasad et al., 2008, 2014, 2018).



(a) The coordinates obtained with UMAP for all discourse relations plotted in two-dimensional space. The plot on the left shows the unweighted embedding representations and the figure on the right shows the weighted embedding representations.



(b) Dendrogram showing hierarchical clustering of Discourse Relations built from UMAP coordinates. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.

Figure 2: Dimensionality reduction and clustering of relation embeddings.

We use a taxonomy adapted from SDRT, in particular, the STAC corpus.

Annotators are usually trained to identify discourse relations using the framework's taxonomy. Some recent alternatives to explicitly collecting annotation of DRs include crowdsourcing by eliciting connectives (Yung et al., 2019; Scholman et al., 2022) or question-answer pairs (Pyatkin et al., 2020) rather than relations. In this work, we wanted to investigate how annotators perceive discourse relation categories, and therefore a connective insertion task would only provide indirect evidence. We train annotators on DR labeling and ask annotators to choose from a set of discourse relation labels. We allow for multiple labels to investigate what relations are more confusable or perceived as co-occurring (Marchal et al., 2022).

6 Discussion and Future Work

In this study, we collected multiple annotations of discourse relations from a subset of the Switchboard corpus, together with confidence scores. We found that dialogue context had a significant effect on confidence scores. We computed embedding representations of DRs using co-occurrence statistics and weighted the vectors using context type and confidence scores, and found that these representations coherently model our annotators uncertainty about discourse relation labels.

Discourse units that occur across turns as defined by Switchboard do not necessarily occur across continuous utterances from the speaker's point-of-view. Obtaining information about whether same-speaker pairs of discourse units fall into the same or different utterances may help to explain additional variance in annotator confidence.

Additionally, in this work, we investigated annotators' confidence on the annotation of adjacent turns. In future work, we plan to annotate discourse relations across longer-distance discourse units and to allow for hierarchical annotation. We expect that annotation confidence will also vary across longer-distance units and across different depths of annotation.

In the future, we plan to use this information to run a larger scale annotation study of the Switchboard corpus to analyze discourse relation patterns in spoken dialogues.

Limitations

This work is limited by the size of the dataset and the taxonomy used in the annotation task. While we found that our annotators perceived some of the categories as more similar or confusable, future work can examine annotators' uncertainty in a larger set of discourse relations. The selection of DUs for annotation was also non-exhaustive. In future work, we plan to expand the selection procedure so that we include more distantly related DUs. We also note that the frequency of discourse relation labels and individual differences in confidence levels among annotators may bias the representations. We plan to look into these potential biases in future work.

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. We also thank the students who participated in the annotation task.

References

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.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "spaCy: Industrial-strength Natural Language Processing in Python".

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Natalia Levshina. 2022. Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft*, 41(1):179–205.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

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.

Torrin M Liddell and John K Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348.

S. Magalí López Cortez and Cassandra L. Jacobs. 2023. The distribution of discourse relations within and across turns in spontaneous conversation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 156–162, Toronto, Canada. 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.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

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 Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4):921–950.

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.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

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

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

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.

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

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.

A Frequencies of Discourse Relation Labels

Discourse Relation	Count
Elaboration	636
Continuation	554
Acknowledgement	494
Explanation	383
Comment	265
Background	252
Narration	249
Question-Answer Pair	248
Contrast	191
Clarification Question	179
Result	124
Other	106

Table 2: Raw counts of discourse relation labels in our corpus from most to least frequent.

B Clustering without Context and Confidence Weighting

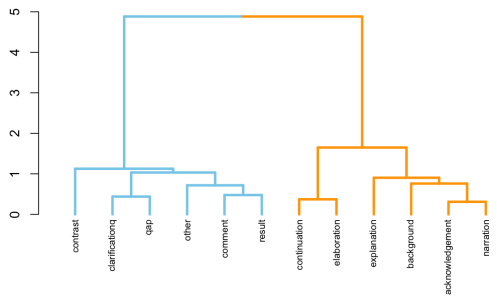


Figure 3: Dendrogram showing hierarchical clustering of Discourse Relations built from UMAP coordinates without context kind and confidence scores weighting. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question. The two main clusters align with the two-dimensional plot of the unweighted UMAP coordinates in Figure 2a

C Annotation Interface

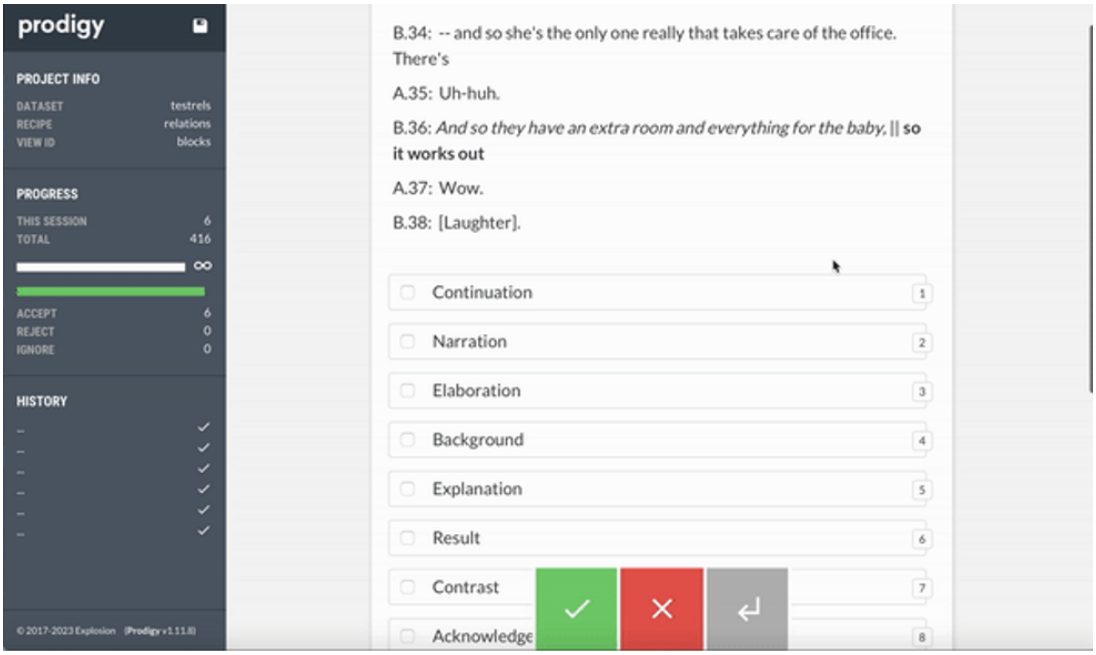


Figure 4: Example annotation task. EDUs to be annotated and discourse relations.

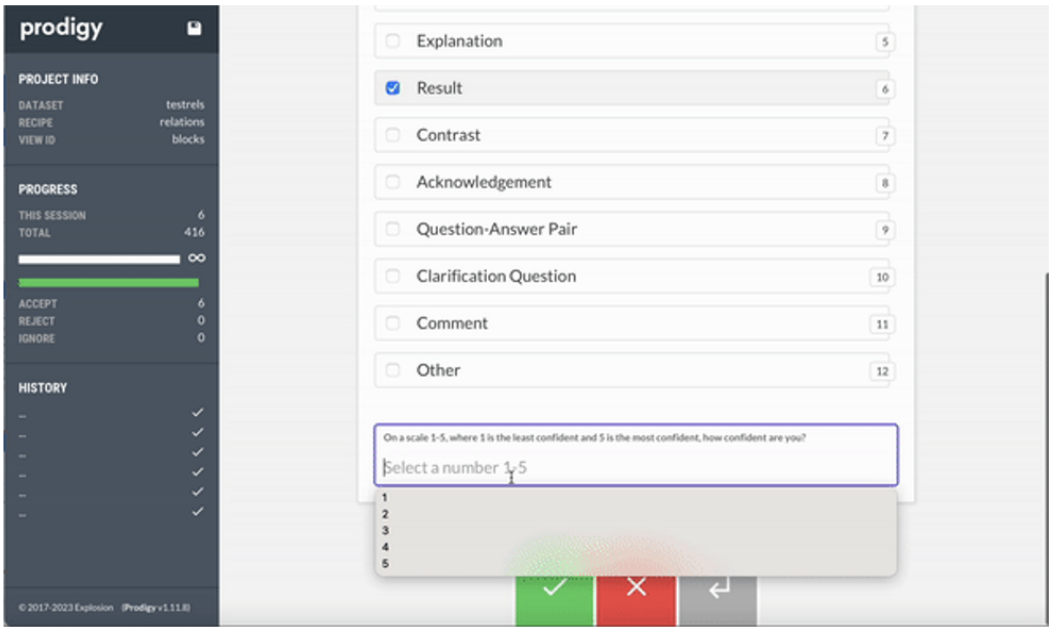


Figure 5: Example annotation task. Discourse relations and confidence score.