# Label distributions help implicit discourse relation classification

**Frances Yung** and **Kaveri Anuranjana** and **Merel Scholman** and **Vera Demberg**
Language Science and Technology, Saarland University
{frances, kaveri, m.c.j.scholman, vera}@coli.uni-saarland.de

## Abstract

Implicit discourse relations can convey more than one relation sense, but much of the research on discourse relations has focused on single relation senses. Recently, DiscoGeM, a novel English multi-domain corpus, which contains 10 crowd-sourced labels per relational instance, has become available. In this paper, we analyse the co-occurrences of relations in DiscoGeM and show that they are systematic and characteristic of text genre. We then test whether information on multi-label distributions in the data can help implicit relation classifiers. Our results show that incorporating multiple labels in parser training can improve its performance, and yield label distributions which are more similar to human label distributions, compared to a parser that is trained on just a single most frequent label per instance.

## 1 Introduction

Much of the research on the discourse relations (DR) assumes (implicitly or explicitly) that only one relation can hold between two spans of text, particularly in the computational field. However, recent work has shown that discourse relations, and implicit relations in particular, can allow for multiple interpretations (e.g., Rohde et al., 2016; Scholman and Demberg, 2017). Prior work has also shown that certain relation senses tend to co-occur in newspaper text, such as ARG2-AS-DETAIL with REASON and CONTRAST with CONCESSION (Demberg et al., 2019). The current study investigates whether the co-occurrence of implicit relation senses depends on text genre.

We furthermore explore whether the performance of a state-of-the-art discourse relation classifier can be improved by training it on the distribution of human-annotated labels, as opposed to training it on only the top label.

The contributions of the current paper are twofold: (1) We analyse the specific patterns of DR co-occurrences in different genres, showing that the distribution of sense co-occurrences are dependent on genre. (2) We train an implicit DR classifier on distributionally upsampled DR labels for each relation instance. This approach leads to better performance according to a soft evaluation metric (cross entropy), demonstrating that the natural distribution of human interpretations can be captured better when training on the distribution of labels.

## 2 Previous work

### 2.1 Co-occurring discourse relations

The assumption that a pair of relational arguments can convey only one single relation sense has led to discourse-annotated resources typically containing one annotated sense per instance. However, recent studies have shown that many relation instances can express multiple senses (e.g., Rohde et al., 2016), and that certain senses co-occur frequently, because their meanings are similar or compatible (e.g., Demberg et al., 2019).

For example, Demberg et al. (2019) showed that newspaper texts annotated as part of both the PDTB (Prasad et al., 2008) and the RST-DT (Carlson et al., 2003) showed consistent patterns of disagreement: CONTRAST and CONCESSION were confused frequently with each other (cf. Robaldo and Miltsakaki, 2014), which can be attributed to the relation senses being similar and difficult to distinguish. Further, the data showed disagreement on INSTANTIATION, LEVEL-OF-DETAIL and REASON relations (cf. Scholman and Demberg, 2017). The confusion between INSTANTIATION and LEVEL-OF-DETAIL relations can be attributed to the relation senses being similar in meaning. The co-occurrence of these senses and REASON is because these relations tend to be multi-interpretable. In the current paper, we extend this line of work by examining whether and how the co-occurrence of relation types differs between genres.

*Proceedings of 3rd Workshop on Computational Approaches to Discourse*, pages 48–53
October 16 – 17, 2022. ©2022 International Conference on Computational Linguistics

### 2.2 Multilabel annotation

An instance of a discourse relation can be annotated with multiple labels under two scenarios. In the first scenario, one annotator (or multiple) identifies multiple co-occurring senses and annotates all identified senses. In the second scenario, each annotator only identifies one discourse sense depending on their perspective. The annotators may agree or disagree with each others.

In the current work, we include data from two corpora: the Penn Discourse Treebank 3.0 (PDTB, Webber et al., 2019) and the DiscoGeM corpus (Scholman et al., 2022). Multilabel relations annotated in the PDTB belong to the first scenario. However, the annotation does not necessarily specify all co-occurring senses (Prasad et al., 2007). In practice, only about 5% of the relations received multiple labels.

The multilabel annotations in DiscoGeM belong to the second scenario. Annotators were asked to identify the strongest relation sense they inferred, but this interpretation could vary between annotators. Every relation instance was annotated by ten participants, thus multiple perspectives were included in the relation sense annotation (that is, the different annotations for one instance were aggregated into one multilabel annotation). As in the PDTB annotation, it is not guaranteed that the multilabels include all co-occurring senses; in the extreme case, the same prominent sense could have been chosen by all workers while the other co-occurring sense does not appear in the multilabel.

### 2.3 Multilabel evaluation

Two forms of evaluation have been used in the context of multilabel data: (i) traditional *hard evaluation metrics* such as accuracy or F1 can be used, but these ignore the information provided by the distribution of senses; (ii) *soft evaluation metrics* can be applied in scenarios where it is assumed that the instances do not involve a single true label; these leverage the information provided by distributions, and seem best suited to our task.

Uma et al. (2021) considered a variety of soft evaluation approaches, including cross entropy (Peterson et al., 2019), Jensen-Shannen divergence (Lin, 1991), and entropy similarity and correlation. Their analyses of these metrics' results on six datasets shows that the relative performance of training methods is critically affected by the chosen evaluation. They recommend to use cross-entropy to compare the output of a system to a soft label; we follow this recommendation in the current work. Cross entropy captures how confident the model is in its top prediction compared to humans and the reasonableness of its distribution over alternative categories.

## 3 Data

We include data from the PDTB 3.0 (Webber et al., 2019) and the DiscoGeM corpus (Scholman et al., 2022).

**PDTB** The PDTB 3.0 is a news text corpus containing articles from the Wall Street Journal. The corpus contains 15,544 inter-sentential (between sentences) and 6,188 intra-sentential (within sentences) implicit relations. We follow the split suggested by Ji and Eisenstein (2015): sections 2-20, 0-1, and 21-22 are used as training, validation, and test sets.

**DiscoGeM** DiscoGeM is a crowdsourced mixed-genre corpus of 6,505 inter-sentential implicit relations, containing text from Europarl (prepared spoken text, argumentative genre), literature (narrative genre), and Wikipedia (informative genre) (see Scholman et al., 2022, for more details on the genres and data selection process). DiscoGeM contains texts from both original English sources as well as translated English.

The annotations were crowdsourced using a connective insertion task, that allowed the authors to extract PDTB 3.0 labels. Every instance received annotations from 10 crowdworkers to represent the multiple interpretation of the discourse relation. The Cohen's Kappa between the multiple labels annotated by the crowdworkers and trained annotators was 0.79 (see original paper for further details.) The labels used in the current study are a single majority vote label (randomly sampled in case two senses received an equal number of votes) and a soft distribution label based on the raw votes.

The DiscoGeM data was split[1] into 70% for train, 20% for test, and 10% for dev for the implicit relation classification reported in Section 5 and for future studies training and testing on DiscoGeM. In splitting the data, we first set aside texts for testing only (i.e., unseen texts). These texts consist of 26 out of 197 turns of speech in the Europarl genre,

[1] The splits can be found in DiscoGeM's online repository: https://github.com/merelscholman/DiscoGeM.

2 out of 20 books in the Novels genre, and 4 out of 69 Wikipedia texts. They are balanced between original English and translated English. For the remainder of the data, we ensured that each text[2] is split proportionally into train, dev and test sets. The dev and test portions are always consecutive, such that the training portions have minimal gaps, but either one comes first randomly and is randomly picked from different locations of the article. The distribution of the relation types in the training data we used is shown in Table 3.

## 4 Analysis of relation co-occurrence

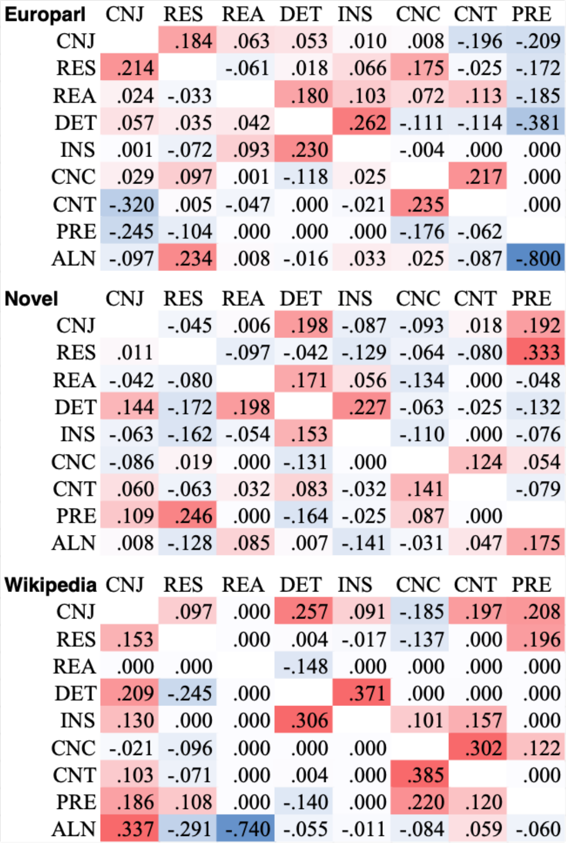
We examine what types of DRs are often interpreted by the crowdworkers at the same time. For each annotated instance in DiscoGeM, we select the TOP1 label as the label that received the majority vote (randomly sampled in case of a tie). The TOP2 label is the label that received the second highest vote. In case where there's less than 30% agreement on the TOP2 label, we consider TOP1 to be a single label (i.e., the TOP2 label = ALONE). This allows us to focus the analysis on co-occurrences of generally accepted interpretations.

From the distributions of the TOP1 $\sim$ TOP2 pairs of the entire corpus, we extract the marginal probabilities $P(Top1)$ and $P(Top2)$, as well as expected joint probabilities $P_{exp}(Top1, Top2)$. We compare the observed joint probabilities $P(Top1, Top2)$ per genre by normalized pointwise mutual information (NPMI), as defined below:

$$NPMI(Top1; Top2) = \frac{PMI(Top1; Top2)}{-\log P_{exp}(Top1, Top2)}$$

Figure 1 shows the NPMI of each unique pair of DRs for the most frequent level-2 relation types.[3] A value of 1.0 means the relations always occur together, $-1.0$ means they never occur together, and 0.0 means that they co-occur as much as would be expected to happen by chance.

The results confirm the overall co-occurrence patterns established in previous work: CONTRAST and CONCESSION co-occur frequently, and INSTANTIATION, LEVEL-OF-DETAIL and REASON also co-occur in every genre. Interestingly, these co-occurrences are more likely when INSTANTIATION or LEVEL-OF-DETAIL is the TOP1 label, and not REASON. This indicates that INSTANTIATIONS and LEVEL-OF-DETAIL relations can often also convey an argumentative relation, but this is less likely vice versa.



| Europarl | CNJ | RES | REA | DET | INS | CNC | CNT | PRE |
|---|---|---|---|---|---|---|---|---|
| CNJ | | .184 | .063 | .053 | .010 | .008 | -.196 | -.209 |
| RES | .214 | | -.061 | .018 | .066 | .175 | -.025 | -.172 |
| REA | .024 | -.033 | | .180 | .103 | .072 | .113 | -.185 |
| DET | .057 | .035 | .042 | | .262 | -.111 | -.114 | -.381 |
| INS | .001 | -.072 | .093 | .230 | | -.004 | .000 | .000 |
| CNC | .029 | .097 | .001 | -.118 | .025 | | .217 | .000 |
| CNT | -.320 | .005 | -.047 | .000 | -.021 | .235 | | .000 |
| PRE | -.245 | -.104 | .000 | .000 | .000 | -.176 | -.062 | |
| ALN | -.097 | .234 | .008 | -.016 | .033 | .025 | -.087 | -.800 |

| Novel | CNJ | RES | REA | DET | INS | CNC | CNT | PRE |
|---|---|---|---|---|---|---|---|---|
| CNJ | | -.045 | .006 | .198 | -.087 | -.093 | .018 | .192 |
| RES | .011 | | -.097 | -.042 | -.129 | -.064 | -.080 | .333 |
| REA | -.042 | -.080 | | .171 | .056 | -.134 | .000 | -.048 |
| DET | .144 | -.172 | .198 | | .227 | -.063 | -.025 | -.132 |
| INS | -.063 | -.162 | -.054 | .153 | | -.110 | .000 | -.076 |
| CNC | -.086 | .019 | .000 | -.131 | .000 | | .124 | .054 |
| CNT | .060 | -.063 | .032 | .083 | -.032 | .141 | | -.079 |
| PRE | .109 | .246 | .000 | -.164 | -.025 | .087 | .000 | |
| ALN | .008 | -.128 | .085 | .007 | -.141 | -.031 | .047 | .175 |

| Wikipedia | CNJ | RES | REA | DET | INS | CNC | CNT | PRE |
|---|---|---|---|---|---|---|---|---|
| CNJ | | .097 | .000 | .257 | .091 | -.185 | .197 | .208 |
| RES | .153 | | .000 | .004 | -.017 | -.137 | .000 | .196 |
| REA | .000 | .000 | | -.148 | .000 | .000 | .000 | .000 |
| DET | .209 | -.245 | .000 | | .371 | .000 | .000 | .000 |
| INS | .130 | .000 | .000 | .306 | | .101 | .157 | .000 |
| CNC | -.021 | -.096 | .000 | .000 | .000 | | .302 | .122 |
| CNT | .103 | -.071 | .000 | .004 | .000 | .385 | | .000 |
| PRE | .186 | .108 | .000 | -.140 | .000 | .220 | .120 | |
| ALN | .337 | -.291 | -.740 | -.055 | -.011 | -.084 | .059 | -.060 |

Figure 1: NPMI per genre of co-occurrences of the most frequent relation senses. The x-axis represents the TOP1 label; the y-axis the TOP2 label.

There are also a number of interesting divergences between genres. For example, relations in the Wikipedia genre are more likely to co-occur with CONJUNCTIONS than relations in other genres; CONJUNCTIONS are also more likely to occur alone in Wikipedia. This is likely due to the expository nature of Wikipedia texts. Further, we can see differences in the patterns of RESULT and PRECEDENCE relations. These are likely to co-occur in novels in particular, as well as Wikipedia to some degree, but not in Europarl data. Rather, in Europarl, RESULT relations occur as a single label (ALN) more frequently compared to the other genres. This matches the argumentative nature of the political genre.

We can conclude from this analysis that there are not only differences in the relation distributions

[2] Europarl: turn of speech; Novels: paragraphs; Wiki: articles.

[3] CNJ: CONJUNCTION; RES: RESULT; REA: REASON; DET: LEVEL-OF-DETAIL; INS: INSTANTIATION; CNC: CONCESSION; CNT: CONTRAST; PRE: PRECEDENCE; ALN: alone

between genres, but there are also differences in the distribution of relation co-occurrences between genres. Such fine-grained differences might impact parsers. In the remainder of this paper, we will study whether incorporating multiple labels in parser training, instead of only the TOP1 label can improve performance.

## 5 Implicit Relation Classification

### 5.1 Model

We implemented the BMGF-RoBERTa model (Liu et al., 2021), which is the current state-of-the-art model for 11-way implicit relation classification on the PDTB 2.0 corpus. The model combines three modules – contextualized word representations, bilateral text span multi-perspective matching and a global gated multi-head attention module. Liu et al. (2021) reported an accuracy of 58.13% on the PDTB 2.0 (11-way).

As discussed in Section 4, the distribution of DRs varies across genres. To capture this information, we prepend a genre token to the input. This token makes the model aware of the genre while learning DR prediction from the text. We also experimented with adding the genre token to the dense layers, but adding it to the input gave better results.

We train the model for 50 epochs on the combined datasets of PDTB 3.0 and DiscoGeM (see Section 3 for the splits) based on the code implemented by Liu et al. (2021).[4] For the results, we select the epoch with the highest accuracy on the combined test set (PDTB+DiscoGeM) and report the accuracy and F1 scores on the individual test sets from that epoch.

### 5.2 Incorporating multiple labels in training

Peterson et al. (2019) demonstrate that models generally trained on a single label task generalize poorly on out-of-domain data because the distributions learnt across the labels do not reflect human uncertainty. We incorporate information on label co-occurances as well as human uncertainty in predicting DRs by proportionally upsampling labels from the DiscoGeM human annotations. We implement three models to study the effect of incorporating multiple labels in training:

(i) Single label model (**single**): The model is trained on the instances in the dataset using only the majority votes.

(ii) Multi-label model (**multi**.): The model is trained on duplicated instances; the repetition is proportional to the distribution of relation senses provided by the crowdworkers (labels with less than 20% votes were excluded to eliminate noisy samples).

(iii) Uncertainty model (**uncert**.): This setting is similar to multi., but all the repetitions are instances of the majority label (labels with less than 20% votes were excluded).

Consider an instance that received the following annotations from 10 crowdworkers: 5×A, 3×B, 1×C, and 1×D. In the single label model, this instance is included once in training and is classified as A. In the multi-label model, this instance is included 8 times; 5 times it is classified as A and 3 times as B. In the uncert. model, the instance is also included 8 times, but always classified as A.

The uncert. model thus takes into account the uncertainty of human annotation because instances with lower agreement are represented less in the training data. The multi-label model considers both uncertainty and sense co-occurrence, while the single model considers neither.

To keep the ratio of DiscoGeM to PDTB data as in the the single label condition, multi. and uncert. models are also upsampled. For the multi. model, if a second label is present, we treat the label distribution as a 50-50 split between the two labels. Otherwise the distribution consists only of the one label with a 100% weight. We upsample the PDTB instances by a factor of 6 (i.e., every PDTB instance is included six times in training), which is roughly the same factor by which the DiscoGem corpus was upsampled on average after excluding labels with less than 20% of the votes. For the uncert. model, we upsample only the majority label 6 times for all instances.

### 5.3 Results

The results are presented in Table 1; we report each metric averaged across 5 runs to account for variation in the results. Let us first consider the results on the DiscoGeM corpus – both as a complete dataset as well as per genre. When evaluated against a single gold label, the performance of the model trained on the distribution of labels (multi) is on par with the model trained just on a single

[4] The hyperparameters are identical to Liu et al. (2021) but batch size is 64.

label. However, the multi model reaches a better performance when evaluated against the distribution of human labels, compared to the single model, as shown in Table 2. Additionally, the certain-label model is substantially under-performing in both evaluation settings. This means that the improvement from training on the distribution of labels indeed comes from the information of how often each label is chosen and which labels co-occur with each other, and not simply from down-sampling difficult instances.

| | DG-all | EP | Novel | Wiki | PDTB 3.0 |
|---|---|---|---|---|---|
| single | 48.66 (0.54) | 53.25 (1.20) | 45.31 (0.34) | **45.58 (2.76)** | **55.28 (1.04)** |
| | 22.33 (1.37) | **25.88 (3.47)** | 23.10 (1.71) | 24.02 (2.90) | 37.18 (1.91) |
| multi. | **49.91 (0.64)** | **54.45 (0.47)** | **47.13 (0.78)** | 44.34 (2.08) | 54.82 (0.73) |
| | **23.66 (1.19)** | 25.44 (3.27) | **25.56 (1.07)** | **25.58 (3.91)** | **38.07 (2.25)** |
| uncert. | 48.43 (0.54) | 53.32 (1.34) | 44.95 (0.85) | 44.65 (4.23) | 54.70 (0.61) |
| | 21.62 (1.69) | 22.08 (2.92) | 23.10 (2.05) | 24.65 (0.88) | 32.48 (1.97) |

Table 1: 11-way accuracy and macro F1 (in grey) of the single, multi. and uncert. models, evaluated against the single (majority) label. We report mean and (standard deviation) based on five runs.

| | DG-all | EP | Novel | Wiki | PDTB 3.0 |
|---|---|---|---|---|---|
| single | 1.86 (.03) | 1.79 (.02) | 1.87 (.04) | 2.08 (.04) | 1.48 (.04) |
| multi. | **1.79 (.03)** | **1.73 (.02)** | **1.8 (.05)** | **2.03 (.06)** | **1.36 (.02)** |
| uncert. | 1.83 (.05) | 1.77 (.07) | 1.84 (.05) | 2.01 (.06) | 1.48 (.02) |

Table 2: Cross entropy (mean and standard deviation) of the single, multi and uncert. models, evaluated against all labels provided by the DiscoGeM crowdworkers. Smaller numbers are better.

| | DG-all | EP | Novel | Wiki | PDTB 3.0 |
|---|---|---|---|---|---|
| Conjunction | 27.54 | 23.87 | 27.26 | 43.52 | 20.96 |
| | 53.36 | 51.32 | 51.37 | 62.11 | 52.76 |
| Cause | 33.75 | 43.83 | 28.20 | 18.50 | 26.50 |
| | 57.42 | 65.75 | 48.31 | 16.27 | 62.89 |
| Level-of-detail | 17.58 | 17.33 | 17.95 | 16.86 | 14.92 |
| | 40.84 | 41.31 | 43.07 | 27.34 | 38.93 |
| Asynchronous | 7.73 | 1.18 | 13.49 | 7.76 | 5.79 |
| | 52.91 | 20.20 | 56.40 | 33.50 | 56.70 |
| Instantiation | 4.45 | 5.05 | 3.51 | 6.3 | 6.69 |
| | 19.73 | 25.80 | 7.45 | 21.50 | 60.80 |
| Concession | 5.11 | 6.23 | 4.43 | 3.75 | 6.83 |
| | 20.19 | 13.6 | 24.8 | 18.3 | 46.00 |
| train size | 26287 | 11030 | 12511 | 2746 | 102306 |

Table 3: Label distribution (%, white rows) of the training data and averaged F1 (in grey) of the multi. model for the most frequent relation types. The last row shows the size of the upsampled training data.

Table 3 presents the distributions of the six relation types that most frequently occur in the training data, along with the per-class F1 score of the multi. model (i.e. after upsampling). These results show that the model's performance follows a similar pattern as the distribution per relation sense. For example, the performance on CAUSE relations is particularly high in the EP genre, likely because causal relations are abundant in the EP training data. The model shows divergent results for the LEVEL-OF-DETAIL (DET) relations: performance is particularly poor on Wiki data, even though the distribution is similar across genres. This can be explained by the co-occurrence patterns presented in Section 4. As seen in Figure 1, DET often co-occurs with CONJUNCTIONS (CNJ) in Wiki, and so the training data contains many instances annotated with DET and CNJ at the same time. Given the frequent distribution of CNJ in Wiki, the parser is likely to classify many of these instances as CNJ. For instances where DET is the majority label, the prediction of CNJ is credited by cross-entropy evaluation but penalized by the F1 score. This also explains the lower than expected performance of the CAUSE relations in the novel genre, where they frequently co-occur with ASYNCHRONOUS.

In sum, these results show that the parser is affected by genre distributions, and that when the parser diverges from the majority label, it can actually predict a related label that the annotation often co-occurs with. This indicates that our proposal of adding genre information and including multiple annotation in the training data was successful.

## 6 Discussion and Conclusion

We presented a simple, yet effective approach to take into account information on multi-label distributions in implicit relation classifiers. Our results show that adding multi-label information leads to results that are on par with single label models when using a hard evaluation metric. The multi-label model performs even better than single label models when using a soft evaluation metric. This means that the multi-label model generates labels that are more reasonable given the distributions of labels provided by humans. This was the case even for the PDTB test set, despite the PDTB being designed as a primarily single label dataset.

In future work, we also plan to study the impact of co-training with a large corpus such as the PDTB on the DiscoGeM genres. Furthermore, will explore methods of incorporating human uncertainty in discourse relation classification with a specific focus on out-of-domain data.

## Acknowledgments

FY and KA are supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074). MS and VD are supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme, Grant 948878 ("Individualized Interaction in Discourse").

## References

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.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue & Discourse*, 10(1):87–135.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3830–3836.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626.

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, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Robaldo Robaldo, and Bonnie Webber. 2007. The penn discourse treebank 2.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Livio Robaldo and Eleni Miltsakaki. 2014. Corpus-driven semantics of concession: Where do expectations come from? *Dialogue & Discourse*, 5(1):1–36.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher NL Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58.

M C J Scholman and Vera Demberg. 2017. Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse*, 8(2):56–83.

Merel C. J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France. European Language Resources Association (ELRA).

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

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