Human Label Variation in Implicit Discourse Relation Recognition

Frances Yung,[1,*] **Daniil Ignatev,**[2,*] **Merel Scholman,**[2]
Vera Demberg,[1] **Massimo Poesio**[2,3]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]Utrecht University, Utrecht, The Netherlands
[3]Queen Mary University of London, London, UK
{frances,vera}@coli.uni-saarland.de
{d.ignatev,m.c.j.scholman,m.poesio}@uu.nl
[*] Contributed equally to this submission.

Abstract

There is growing recognition that many NLP tasks lack a single ground truth, as human judgments reflect diverse perspectives. To capture this variation, models have been developed to predict full annotation distributions rather than majority labels, while perspectivist models aim to reproduce the interpretations of individual annotators. In this work, we compare these approaches on Implicit Discourse Relation Recognition (IDRR), a highly ambiguous task where disagreement often arises from cognitive complexity rather than ideological bias. Our experiments show that existing annotator-specific models perform poorly in IDRR unless ambiguity is reduced, whereas models trained on label distributions yield more stable predictions. Further analysis indicates that frequent cognitively demanding cases drive inconsistency in human interpretation, posing challenges for perspectivist modeling in IDRR.

Keywords: perspectivist modeling, annotation variation, implicit discourse relation recognition

1. Introduction

Discourse annotation is a complex task that is prone to significant disagreement even among trained annotators (Artstein and Poesio, 2008; Spooren and Degand, 2010). This is particularly true for the annotation of implicit discourse relations (DRs) which are not marked by explicit discourse connectives, such as *because* for *reason* or *whereas* for *contrast* (Prasad et al., 2017; Hoek et al., 2021).

(1) Arg1: *Our planet will not have enough room for two faction, both of which strive to dominate the whole world.* Arg2: One of them must give way.

Annotating DRs involves high-level understanding of textual semantics and pragmatics in natural language. Disagreements arise from various factors, including linguistic ambiguity. For instance, the analysis of Example (1) may vary depending on whether *"must"* implies an *obligation* (prioritizing the evaluative aspect) or *inevitability* (prioritizing causality). This allows for multiple readings of the same passage and suggests that adopting a single ground-truth reading would lead to information loss.

Nonetheless, implicit discourse relation recognition (IDRR) has traditionally been a single-label classification task in NLP. The largest discourse annotated resource, the Penn Discourse Treebank (PDTB, Prasad et al., 2019), is primarily annotated with one label per sample and includes alternative labels in less than 5% of the samples, even though disagreement between two experts is expected to be around 44% (Hoek et al., 2021). In recent years, the performance of novel IDRR systems appears to have plateaued (SOTA on PDTB-3: F1 55% ACC 64% Wang et al., 2025). This may indicate that a more flexible approach that permits concurrent interpretations is necessary to ensure a more complete solution. Several studies have reported the benefits of training from multiple annotations in IDRR (Yung et al., 2022; Costa and Kosseim, 2024; Long et al., 2024).

Disagreement-aware approaches in data collection and machine learning have earned recognition in various NLP tasks (Basile, 2020; Uma et al., 2021; Röttger et al., 2022; Cabitza et al., 2023; Xu and Jurgens, 2026), particularly in tasks involving subjectivity and linguistic ambiguity. Specialized architectures have been proposed to model individual annotator stances instead of the ground truth in such domains as hate speech detection and toxicity detection (Davani et al., 2022; Kanclerz et al., 2022; Ngo et al., 2022; Deng et al., 2023; Mokhberian et al., 2024). This approach could tentatively apply to crowd-sourced discourse annotations, as labelers can interpret the pragmatic function of a text subjectively; however, it may entail problematic implications, as human-produced data is known to be prone to errors or deliberate attempts at brute-forcing the task by annotators (Poesio and Artstein, 2005; Plank et al., 2014; Plank, 2022). The applicability of perspectivism is therefore task-dependent. In particular, since variation in DR interpretation often stems from individual differences in cognitive ability (Sanders et al., 1992; Spooren and Degand, 2010), the applicability of existing perspectivist models tailored for subjective tasks remains

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4942–4954
11–16 May 2026.

an open question.

This study examines the performance of various IDRR classifiers for English with different levels of awareness of label disagreement, including classical approaches that are optimized to predict the most probable label, models that are trained to predict label distributions, and perspectivist models that make annotator-specific prediction. We make use of the DiscoGeM dataset Scholman et al. (2022); Yung and Demberg (2025), a large collection of unaggregated crowdsourced labels for English implicit DRs. While training on label distributions has been shown to improve single-label prediction performance (Yung et al., 2022; Pyatkin et al., 2023; Costa and Kosseim, 2024), the robustness and generalizability of these models across different tasks remain unclear. For instance, it is not yet known whether the probability outputs of a single-label prediction model naturally align with label-distribution prediction, or whether annotator-specific prediction models can effectively learn IDRR given a limited number of annotator-specific training instances. Our experiments aim to address these open questions through systematic cross-task evaluation. Our contributions include:

1. We confirmed that disagreement-aware approaches improve the accuracy of majority-label prediction.

2. We find that the effectiveness of annotator-specific models strongly depends on label granularity (5 vs. 17 classes).

3. We demonstrate that learning from label distributions provides the most reliable performance for predicting fine-grained label distributions, when the specific annotators are unknown.

4. Our analysis reveals that annotator inconsistency in cognitively demanding cases, which are common in IDRR, substantially reduces model performance, particularly in ambiguous fine-grained settings.

2. Related work

Disagreement and perspectives in annotation NLP tasks, such as sentiment analysis, toxicity detection, and hate speech detection, involve subjective evaluation of a parameter by annotators. This subjectivity challenges the existence of a ground truth (*alias* gold standard) in the respective datasets. Divergent annotations may need to be treated as a meaningful signal rather than noise (Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019; Larimore et al., 2021; Plank, 2022). Specifically, these annotations can reflect the multitude of the labelers' backgrounds and therefore offer a more representative image of subjective data compared to the aggregated gold standard (Díaz et al., 2018; Garten et al., 2019; Akhtar et al., 2019; Ferracane et al., 2021). Several studies delved into the reasons behind disagreement in detail, identifying various pattern dependent on task (Nie et al., 2020; Jiang and de Marneffe, 2022; Jiang et al., 2023; Sandri et al., 2023). For example, annotators' labeling of toxicity and hate speech vary notably depending on their identity, age, personality-related variables, beliefs and stereotypes (Sap et al., 2022; Sang and Stanton, 2022; Davani et al., 2023, see also Frenda et al. (2024) for an overview).

Annotator-specific label prediction models To embrace the distinct points of view on subjective tasks and model perspectives in annotation, researchers call for the preservation of diverging labels in data collection (Aroyo and Welty, 2015; Poesio et al., 2019; Basile, 2020; Prabhakaran et al., 2021; Cabitza et al., 2023). Existing perspectivist models enhance transformer-based classifiers with trainable annotator-specific components, such as classification heads (Davani et al., 2022), or embeddings (Deng et al., 2023). These architectures enable the integration of annotator-level features, including basic embedded annotator identifiers, and also relevant sociodemographic variables (Mokhberian et al., 2024). We will further explain the models used in our experiments in Section 3.2.

Nonetheless, human label variation does not solely stem from valid differences in perspective but can also result from annotation errors caused by misinterpretation. Weber-Genzel et al. (2024) propose an additional round of explanation annotation and subsequent analysis to tease apart annotation errors and genuine perspectivist label variations. In this work, we conduct a manual analysis to estimate the proportion of annotation variation attributable to differences in interpretation versus annotation errors in IDRR, using third-person annotations of the cues underlying each assigned label.

Perspectives in IDRR Low agreement in implicit DR annotation has been widely reported in existing work (Das et al., 2017; Hoek et al., 2021), also (Poláková et al., 2013). Unlike in other subjective NLP tasks, variation in DR interpretation has not been linked to annotators' demographic factors. Instead, disagreement has been shown to arise from relation-specific ambiguities, such as whether the relation contains a 1st/2nd person conceptualizer (Sanders et al., 1992; Spooren and Degand, 2010). Additional sources of disagreement include textual ambiguity, differences in background knowledge, and ambiguities in relation definitions (Zikánová, 2024; Zikánová et al., 2025; Hewett and Stede,

2025; Ignatev et al., 2025). Several existing resources provide consistent multiple annotations designed to study disagreement in discourse interpretation (Hewett and Stede, 2025; Zeldes et al., 2025; Zikánová et al., 2025). In this work, we use DiscoGeM, a dataset focused on **implicit** DRs.

Recent studies have explored learning from label distributions (Yung et al., 2022; Pyatkin et al., 2023; Costa and Kosseim, 2024; Long et al., 2024), and have consistently shown advantages over majority-label training. However, perspectivist models that explicitly predict the annotations of individual workers have not yet been applied to IDRR, which is a gap this work seeks to address.

3. Experiments

The objective of our experiments is to evaluate how well models can predict disagreement in English IDRR at different levels of granularity. Specifically, we consider three tasks: **single-label prediction**, **label-distribution prediction**, and **annotator-specific label prediction**. We systematically compare models optimized for each of these tasks and evaluate their performance across all three tasks.

3.1. Data

To train IDRR models that account for label variation, we use the DiscoGeM 1.5 dataset Scholman et al. (2022); Yung and Demberg (2025), which is the only available resource of PDTB-style **implicit** DRs that provides multiple annotations for each instance. DiscoGeM is an English multi-genre corpus, containing excerpts from European Parliament proceedings, Wikipedia, and novels. Labels were collected through crowd-sourcing, in which annotators with no presumed knowledge of DRs select a connective that fits between sentences 1 and 2 (Arg1 and Arg2) from a predefined list. The workers can opt for the "*(no direct relation)*" label if they cannot identify any relation between the arguments.

DiscoGeM follows the sense classification of the PDTB, which is structured as a three-level hierarchy, with four coarse-grained Level-1 sense categories and more fine-grained senses for each of the subsequent levels (Webber et al., 2019). The annotation interface allows for annotation to the finest-grained Level-3 senses (29 classes), each represented by a connective choice. On the other hand, our experiments adopt the widespread practice of Level-1 (5 classes) and Level-2 (17 classes) DR prediction (Xiang and Wang, 2023).

Table 1 compares the statistics of DiscoGeM with other multi-annotator corpora previously used for perspectivist modeling. We observe that the mean number of items per worker in DiscoGeM is comparable to- and even surpasses MDA. At the same time, the complexity of the task can make it hard to learn the workers' individual stances even from that many examples. In particular, at Level-3 granularity, the average count of distinct labels per instance amounts to 6.3, leading to a relatively low agreement rate of 0.404 and high entropy of 0.86.[1] Such high level of label variation is not observed in other disagreement datasets.

Col	DG1.5	MDA	HSB	GHC
Train	4.5/52k	6.59/33k	0.78/4.7k	22/69k
Dev	0.6/7k	-	-	-
Test	1.2/14k	3/11k	0.16/1k	5.5/17k
#W	164	334	6	18
#A/#W	448	160	1120	4807
#A/#I	10	5	6	3.13
#L	17/5	2	2	2

Table 1: Statistics of DiscoGeM 1.5 compared to Multi-Domain Agreement, MDA (Leonardelli et al., 2021); Hate Speech-Brexit, HSB (Akhtar et al., 2021); and GHC (Kennedy et al., 2022); counts reported as items/annotations. #W: worker count; #A/#W: average annotation/worker ratio; #A/#I: average annotation/instance; #L: unique label count.

3.2. Models

We compare three classes of models across the three IDRR tasks. To ensure consistency and to facilitate comparison with the existing IDRR pipelines (Long et al., 2024; Costa and Kosseim, 2024), we use RoBERTa-base (Liu et al., 2019) as a backbone for all architectures. Separate models are trained for Level-1 and Level-2 DR classification.

Single truth (ST) model Our baseline is a RoBERTa-base classifier trained on the majority label of each instance. We derive the model's normalized probability distribution (**ST.logit**) to represent its single-label prediction for the label-distribution task. For annotator-specific label prediction, the model's predicted single label (**ST.top1**) is considered as the model's prediction of every perspective for the given instance.

Soft label models We experiment with two models that allow soft-label prediction. The first method performs multi-label classification in the form of a binary classification for each label type. Specifically, we use a Binary Cross-Entropy (BCE) loss averaged over batch and labels, which is a standard approach for multi-label text classification (e.g., He

[1] Considering the 10 collected labels per instance but not the complete set of 29 Level-3 labels, for which most have 0 votes.

and Xia, 2018; Huang et al., 2021). Following Pyatkin et al. (2023), labels with 20% or more distribution in each item of DiscoGeM are selected as the multiple gold labels during training. Similarly to the ST model, we use probability distributions (**multi.logit**) to represent predictions for the label-distribution task.

The other soft prediction model we explored directly predicts the full annotation distribution (**label-dist**) of each sample based on a soft loss, specifically the Kullback-Leibler divergence loss, following Fornaciari et al. (2021); Uma et al. (2021). For single label and annotator-specific label prediction tasks, labels with the highest probability (**multi.top1** and **dist.top1**) are used as the single-label prediction.

Perspectivist models We compared two perspectivist models in our experiments: the multi-task-based multi-annotator model (**MT**, Davani et al., 2022) and the Annotator Embeddings model (**AE**, Deng et al., 2023). Both are trained on paired annotator IDs and their specific labels provided in DiscoGeM and are applied to predict the annotations of specific annotators in the test set.

The **MT** model learns multiple perspectives as separate tasks, using a separate head for each annotator in the dataset on top of a shared encoder. During training, the model is jointly optimized against all labels assigned to a single instance, whereby Cross-Entropy is computed separately for each head and then summed. The **AE** model models worker-specific annotation patterns using distinct annotator and annotation embeddings, which are integrated with the original text embeddings through a weighted sum. The statistics of the original data used in the MT and AE models are shown in Table 1.

We additionally compare against majority-vote baselines (**MT.maj** and **AE.maj**), which assign the most frequently predicted label per item to all annotators. These outputs are also used in the single-label prediction setting. For the label-distribution prediction task, we average the predicted probability distributions of the ten annotators in each test sample, producing one aggregated label distribution per item (**MT.logit** and **AE.logit**).

4. Results

Single-label prediction Table 2 presents the results for the single-label prediction task, evaluated using standard macro-F1 and accuracy against the single majority gold label provided in DiscoGeM. The results replicate previous findings claiming that training on label distributions can improve single-label prediction performance (Yung et al., 2022; Pyatkin et al., 2023; Costa and Kosseim, 2024). The relatively lower performance of the multi-label prediction model on Level-2 DR prediction may be attributed to the increased randomness involved in selecting the top label among a larger set of predicted labels. In contrast, the perspectivist AE model further outperforms the soft-label approaches on Level-1 prediction, but this advantage does not extend to Level-2, suggesting that a higher number of classes may constrain model performance.

	$F1_{Lv1}$	Acc_{Lv1}	$F1_{Lv2}$	Acc_{Lv2}
ST.top1	0.48	0.58	0.24	0.44
multi.top1	0.47	0.57	0.21	0.29
dist.top1	0.49	0.6	**0.28***	**0.47**
MT.maj	0.36	0.39	0.03	0.03
AE.maj	**0.56***	**0.7***	0.21	0.4

Table 2: Results of the **single label prediction task** evaluated by accuracy and macro-F1. Performance averaged per 3 runs. The highest score shown in bold. * marks significant advantage over ST, $p < 0.05$. Here and further on, significance is measured with a resampled paired t-test (Dietterich, 1998).

Lv1	CE	JSD	MD	ED
ST.logit	1.708	0.748	0.570	0.566
multi.logit	1.520*	0.153*	0.630	0.356*
label-dist	**1.341***	**0.118***	0.547*	0.304*
AE.logit	1.402*	0.123*	**0.367***	**0.216***
Lv2	CE	JSD	MD	ED
ST.logit	3.071	0.371	1.126	0.551
multi.logit	2.788	0.342	1.021*	0.440*
label-dist	**2.265***	**0.269***	**0.878***	**0.347***
AE.logit	2.947	0.450	1.187	0.488

Table 3: Results of the **label distribution prediction task** evaluated by 4 soft metrics. Results based on 3 random seeds. The lowest score shown in bold. * marks significant advantage over ST, $p < 0.05$.

Label-distribution prediction Table 3 compares the results where the predicted distributions are evaluated against the gold annotation distributions of each instance using four soft metrics: Cross-Entropy (CE), Jensen–Shannon Divergence (JSD), Manhattan Distance (MD), and Euclidean Distance (ED). Lower values indicate greater similarity between the predicted and reference distributions. The results show that the **ST.logit** model lags behind the others, suggesting that the probability distribution of the single-truth model does not effectively capture the underlying label distribution, and that training with soft labels is beneficial. The **label-dist** model achieves the best overall performance in this task, which is expected given that it is explicitly optimized for distributional prediction. The perspectivist models perform comparably well, but only at

	$F1_{Lv1}$	Acc_{Lv1}	$F1_{Lv2}$	Acc_{Lv2}
ST.top1	0.35	0.44	0.15	0.28
multi.top1	0.35	0.44	0.14	0.20
dist.top1	0.35	0.46*	0.17*	0.29*
MT	0.33	0.34	0.08	0.10
AE.maj	0.37*	0.50*	0.11	0.21
AE	**0.75***	**0.84***	**0.23***	**0.44***

Table 4: Results of the **annotator-specific label prediction task** evaluated by accuracy and macro-F1, considering each annotator-specific annotation as one instance. Performance averaged per 3 runs. The highest score shown in bold. * marks significant advantage over ST, $p < 0.05$.

Level-1, again indicating that their advantage diminishes as the number of classes increases.

Annotator-specific label prediction Finally, Table 4 presents the results of annotator-specific label prediction, averaged over 3 runs. The models are evaluated using *global* accuracy and macro F1-score, where each annotator-specific annotation is counted as an individual instance (Deng et al., 2023). The results suggest that the perspectivist **AE** model consistently outperforms all other models by a substantial margin. Although the performance on Level-2 prediction is substantially lower than on Level-1, as also observed in the previous tasks, it still remains approximately twice as high as that of the other models. The low performance of **AE.maj** highlights the disagreement between the annotators and demonstrates the importance of annotator-specific prediction. The **MT** model, in contrast, appears unable to predict annotation perspectives in this task. Notably, the original model was tested on 2-way and 6-way classification with fewer annotations per instance (3 vs 10), and high number and diversity of labels could have exacerbated the weaknesses of multi-task learning, such as conflicting gradient updates (Yu et al., 2020).

To summarize, the results show that: (1) learning from label distributions or annotator-specific labels improves performance in single-label prediction, whereas single-label strategies perform poorly on distributional prediction tasks due to the high disagreement and ambiguity in IDRR; (2) models optimized for soft-label prediction (i.e., to learn per-sample label distributions without annotator information) and perspectivist models that are optimized for annotator-specific predictions each achieve the best performance on their respective tasks; and (3) perspectivist models can also successfully predict label distributions, but their effectiveness decreases as the number of classes increases.

	$F1_{Lv1}$	Acc_{Lv1}	$F1_{Lv2}$	Acc_{Lv2}
$AE_{\#E+}$	0.75	0.84	**0.23**	**0.44**
$AE_{\#E-}$	**0.76**	**0.85**	**0.23**	0.43
$AE_{Low-\kappa}$	0.68	0.81	0.14	0.25
$AE_{Med.-\kappa}$	0.74	0.83	0.22	**0.46**
$AE_{High-\kappa}$	**0.76**	**0.85**	**0.24**	0.44

Table 5: Results of the **worker-specific label prediction task** in different subsets of workers. Performance averaged per 10 runs. The highest scores shown in bold.

5. Analysis

The AE model outperforms the label-dist model only on Level-1 in distribution predction (Table 3); its performance on annotator-specific prediction also drops more significantly from Level-1 to Level-2 prediction compared with other models (Table 4). These results show that the perspectivist model struggles to predict fine-grained annotation of specific annotators. In this section, we aim to identify the conditions under which the perspective-aware models can accurately predict a worker's individual label. In addition, we will analyze the different types of perspectives and biases in IDRR.

5.1. When do perspectivist models perform better?

During perspective-aware training, models can be biased in favor of more prolific annotators. To assess the effect of this factor, we divide the annotators into two groups based on the counts of their items in the train set: less or equal to the median count ($AE_{\#E-}$, mean count=93) or above the median ($AE_{\#E+}$, mean count=540). The model performance on these subsets of workers is shown in the upper half of Table 5. Surprisingly, comparing the metrics across the two groups reveals no significant difference, suggesting that the per-annotator training samples are enough in the dataset and the difference in model performance is not due to the prolificness of the workers.

Next, we compare model performance between workers with more conventional interpretations of DRs and those with less typical ones. We capture this difference by comparing the worker's annotation with the majority. Specifically, we compute the normalized Pointwise Mutual Information (nPMI) between each of the 17 Level-2 labels assigned by the worker (L_w) and the majority (L_{maj}) in the same subset of samples. A positive nPMI(L_w,L_{maj}) indicates that a worker tends to assign L_w to samples for which the majority label is L_{maj}, suggesting an association between this pair of worker and majority labels. Conversely, a negative nPMI value suggests that the worker tends *not* to assign L_w when the majority label is L_{maj}, reflecting a negative corelation between this label pair.

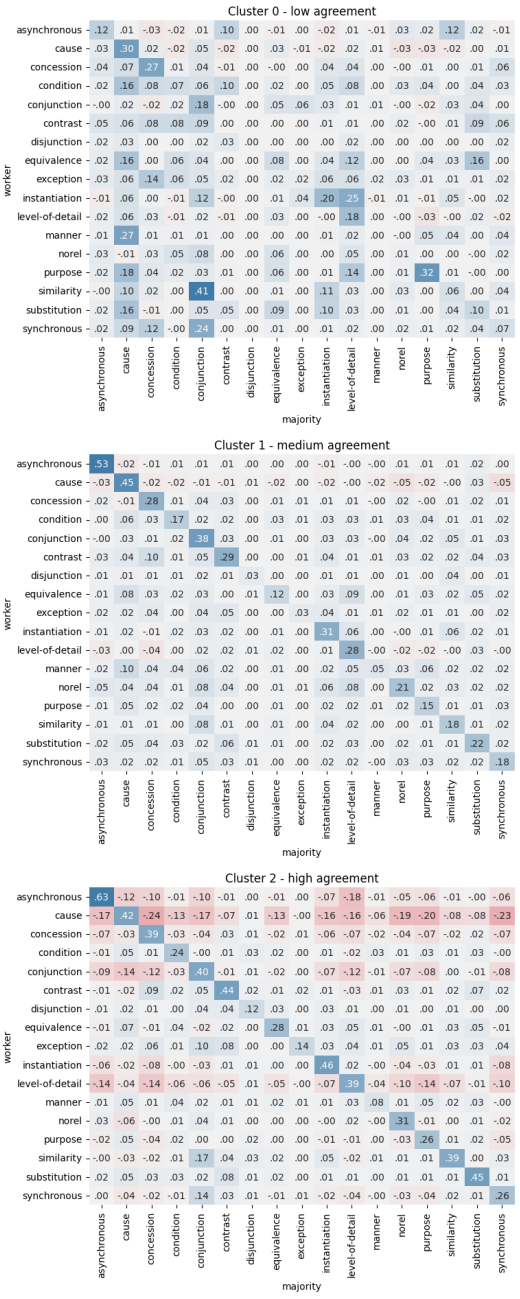

Figure 1: Mean nPMI values between the worker and majority labels for the three clusters of workers with different levels of agreement with the majority: Cluster 0 (low agreement), Cluster 1 (medium agreement), Cluster 2 (high agreement).

We then apply K-means clustering ($k = 3$) to group workers based on their vectors of nPMI values. Figure 1 illustrates the mean nPMI matrices for the three clusters. As shown, the workers in Cluster 0 show low agreement, frequently assigning labels that differ from the majority. In contrast, Cluster 2 workers exhibit the highest agreement with the majority vote, characterized by a dark blue main diagonal (indicating strong positive correlations) and light or red off-diagonal cells (indicating weak or negative correlations). Cluster 1 also demonstrates relatively high agreement, although the absence of strong red cells indicates occasional deviations from the majority consensus.

The lower half of Table 5 presents the performance of the AE model across these clusters of workers with different levels of agreement with the majority ($\text{AE}_{Low-\kappa}$, $\text{AE}_{Med.-\kappa}$, and $\text{AE}_{High-\kappa}$). The results reveal that the model performs notably worse in predicting the annotations of workers who have low agreement with the majority, implying that the labeling biases and perspectives of this group are harder for the model to learn and predict. We further examine specific cases to understand when the model can effectively capture annotator preferences and when it fails to do so. Figure 2 presents the nPMI matrices for 4 workers, comparing their labels with the majority labels. Figure 3 shows the corresponding confusion matrices of the predictions of the AE model against the actual labels of these workers. For clearer visualization, all values are aggregated at the Level-1 category.

Annotator A exhibits a clear preference for the *COMparison* relation, whereas Annotator B shows a bias toward *CONtingency*, as indicated by the prominent blue rows in their respective nPMI matrices (Figure 2). Importantly, the AE model successfully captures these tendencies: it correctly predicts when these workers are likely to choose these relations and when they are not, as reflected in the darker main diagonal indicating alignment between the predicted and actual labels in Figure 3.

In contrast, Annotators C and D display an apparent bias toward the *NO-Relation label*, yet the AE model struggles to model this behavior accurately. Instead, it tends to under-predict the *NO-Relation* label; the model predicts other classes while the workers actually had selected *NO-Relation*. Manual inspection of these cases reveals that Annotators C and D often resort to the *NO-Relation* label when the DR is particularly ambiguous or difficult to interpret (examples in Section 5.2). This suggests that their use of *NO-Relation* is not a consistent labeling bias but rather a fallback strategy employed in response to uncertainty.

The analysis demonstrates that modeling individual perspectives in IDRR is particularly challenging when the workers demonstrate inconsistent behaviors in uncertain, difficult cases. To gain deeper insight into how such cases might be mitigated, we conduct a manual analysis of the frequency and nature of different types of disagreement in IDRR.

5.2. Causes of disagreement

We selected a subset of 100 samples that were consistently annotated by 3 workers, referred to here as Annotators X, Y, and Z. For each Level-2 label assigned by the workers, we examined the nature

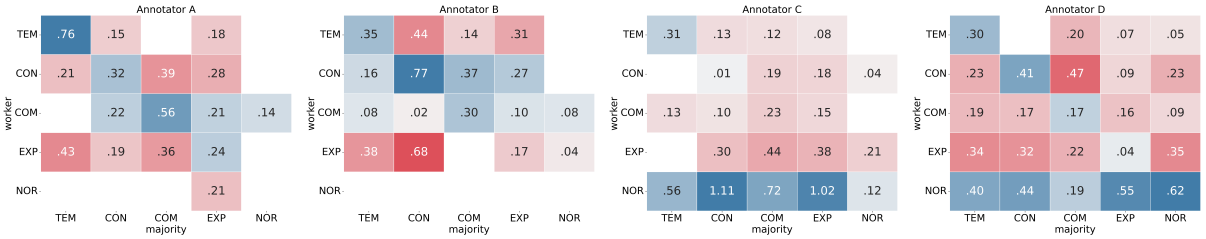

Figure 2: Bias of four workers compared with the majority label based on nPMI. **darker** colors mean more divergence from the majority, where **blue** means higher tendency (positive nPMI) and **red** means lower tendency (negative nPMI). Abbreviations: temporal, contingency, comparison, expansion, no relation.

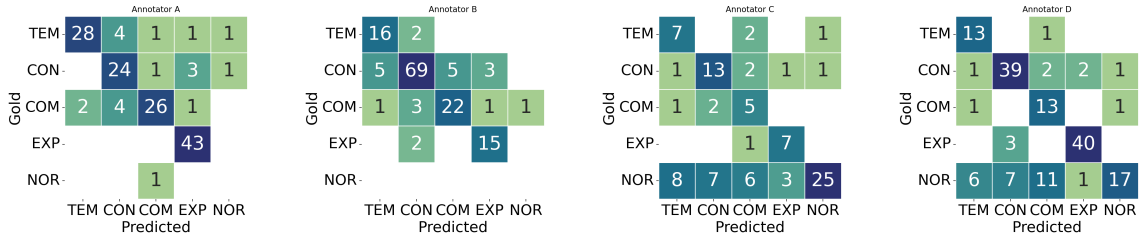

Figure 3: Confusion matrices of worker-specific prediction of the above workers. **Gold** refers to the original labels. Abbreviations: temporal, contingency, comparison, expansion, no relation.

of the disagreement and categorized it according to the potential underlying reason for divergence. When a worker's label matched one or more of the others, it was counted as a case of agreement. Table 6 presents the distribution of disagreement categories, while Table 7 summarizes the three most frequently assigned labels by each worker within this subset. We discuss these categories and their implications in detail below.

agreement	35%
disagreement	
inherent ambiguity	31%
cognitive demands:	
unidentified cues	5%
sentence complexity	10%
annotation error	11%
task design	8%

Table 6: Distribution of disagreement categories in 300 annotations (100 items × 3 workers)

	top1	top2	top3
X	conj.(20%)	asynch.(13%)	inst. (10%)
Y	lev-of-det (22%)	cond. (11%)	synch. (9%)
Z	cause (23%)	lev-of-det (11%)	goal (9%)

Table 7: Top 3 annotated labels by Annotators X, Y and Z in the 100 analyzed items. Abbreviations: conjunction, asychronous, instantiation, condition, sychronous.

Table 6 shows that over one-third of the annotations fall under **agreement**, where the assigned label matches at least one of the other workers for the same item. Another one-third of the annotations were found to differ due to the **inherent ambiguity** of the DRs. This ambiguity arises from a variety of causes. A primary source of ambiguity stems from differences in how workers interpret or emphasize **different spans** within the relation arguments. In Example (2), Annotator Z focused only on the beginning of Arg2, interpreting the relation as *concession* (i.e., the man opened the window even though it was raining). In contrast, Annotator Y considered the full span of Arg2 and identified the DR as *synchrony*, emphasizing the simultaneous nature of the two events.

(2) Arg1: *What was the man about to do? He lowered the window on our side.* Arg2: A heavy rain was now falling, and, by a gesture, the man expressed his annoyance at his not having an umbrella...

Another source of disagreement arises from the **lack of contextual information**. Because DiscoGeM is designed to capture the organic, local interpretation of DRs, workers are instructed to label relations given only one to two sentences of context preceding Arg1 and following Arg2. Consequently, the interpretation of a DR can vary depending on each worker's inferred or assumed broader context. In the narrative fragment of Example (3), the relationship between the characters and the nature of the events remain unclear without access to the wider context of the novel, leading to annotator disagreement.

(3) Arg1: *Pianelli was constantly looking for a scoop, but I'd often wondered what lay behind his fascination*

with this particular case. Arg2: From what I remembered, he and Vinca had never been close. They'd never spent time together...

In addition, workers also exhibit **individual preferences** when labeling ambiguous DRs. In Example (4), both Arg1 and Arg2 refer to the same entity pair (Fanny and "I"), yet the DR between them is not obvious. This type of relation corresponds to what the PDTB framework classifies as an *entity relation* (Prasad et al., 2017), which is typically implicit in natural text and often resists representation by an explicit connective. In this case, Annotator X labeled the relation as *conjunction*, whereas Annotator Y chose *level-of-detail*. Both interpretations are reasonable and highlight distinct annotation preferences toward particular relation types, preferences that are also reflected in the individual label distributions shown in Table 7.

(4) Arg1: *...Fanny was the first girl I 'd ever dated, back when I was fourteen and in the ninth grade.* Arg2: We went to see a Saturday matinee of Rain Man...

The types of ambiguity discussed above are also commonly observed in other annotation tasks, often leading to similar patterns of disagreement among annotators (Jiang and de Marneffe, 2022; Sandri et al., 2023). We additionally identified types of disagreement that emerge from the **cognitively demanding** nature of IDRR. First, some workers appear to be more sensitive to or linguistically skilled in generating discourse inferences based on subtle **cues**, while others may overlook these cues or rely less on them in their interpretations (Graesser et al., 1994; Scholman et al., 2020; Zwaan and Radvansky, 1998). In Example (5), the specification of a year in the text prompted Annotators X and Y to select an *asynchronous* relation. In contrast, Annotator Z focused on the event development, labeling the relation as *cause*, which is also justifiable given the context.

(5) Arg1: *...I lived in fear and in a constant state of mental anguish that led me to disastrously fail my final exams.* Arg2: By the summer of '93, I had already left the Côte d' Azur for Paris, where ... I enrolled in a second-rate business school.

Another cognitive challenge in IDRR arises from the difficulty of interpreting **complex sentence** structures, which obscure the intended semantics of the discourse arguments. In Example (6), Arg2 begins with the connective *after*, which actually signals an intra-argument temporal relation within Arg2 itself. However, this connective appears to have prompted Annotator X to label the relation between Arg1 and Arg2 as *asynchronous*. The DR should instead concern the contrast between Maxime and "I", as identified by Annotator Y.

(6) Arg1: *...While it had devastated me and stopped me in my tracks, for Maxime, it had broken down barriers, released him from a straitjacket, and left him free to write his own story.* Arg2: After what happened, I was never the same...

In our manual analysis, we also identified a portion of labels that we could not justify and therefore categorized as **annotation errors**. In Example (7), Annotator Y labeled the relation as *equivalence*, which is not reasonable given the context. It is possible that they were uncertain about the correct relation and selected a label at random. Here, Annotator X assigned the label *no-relation*, which is appropriate in this case. However, since this worker tends to select *no-relation* in a range of cases, such as Example 1, it can also be argued that they may be using this label as a fallback strategy when unsure of the correct relation.

(7) Arg1: *From what I remembered, he and Vinca had never been close... they had almost nothing in common.* Arg2: Vinca's mother was Pauline Lambert, an Antibes-born actress with close-cropped red hair...

Finally, we observe that some of the disagreements in DiscoGeM may have arisen from the **task design**, specifically, from the choice of connectives to represent each DR type. In example (8), *condition* is a reasonable interpretation (i.e., If the report is on the table, we are determined to adopt it). However, the connective associated with this relation, *in that case*, does not fit naturally in the context, discouraging workers from choosing it. In other words, the annotation outcome depends on the workers' flexibility in performing the connective insertion task; some may prioritize selecting a connective that sounds more natural in the given context, even if it does not perfectly represent the underlying relation.

(8) Arg1: *...The report has nevertheless been passed by the Committee for Foreign Affairs, it is on the table and it is essential to vote on the report...* Arg2: Our group is fully determined to debate and adopt it ...

Overall, much of the observed variation stems from the cognitively demanding nature of IDRR, which leads to substantial disagreement that is difficult to model using perspectivist approaches due to inconsistencies in annotator behavior. The nature of shallow discourse parsing, where DRs are determined at the level of sentence pairs, allows room for individual differences in interpretation.

Nonetheless, these divergent perspectives are accounted for by existing theoretical principles. For instance, clause-level discourse can support multiple attachments to different anaphoric clauses (Webber et al., 2003), or interact with presuppositional structure (Asher and Lascarides, 2003). In

Example (2), the interpretation depends on whether “was falling" attaches to “lowered" or “expressed." Similarly, in Example (4), “went to see (a movie)” may be interpreted either a “date" or as a separate event.

When aggregated to the coarser Level-1 relation classes, there remains a higher likelihood that noisy labels overlap with consciously assigned labels. However, at the more fine-grained Level-2 classes, the noise appears to amplify disagreement, making individual perspectives harder to model consistently and resulting in markedly lower performance.

6. Conclusion

IDRR is a task where disagreement arises not only from the inherent ambiguity of language and differences in background knowledge but also from individual variation in the cognitive ability to interpret DRs. In this work, we evaluated existing approaches to model disagreement in IDRR with varying levels of precision. We found that models designed to predict specific annotator perspectives experience substantial performance drops when the prediction classes are fine-grained. In such cases (i.e., when the number of classes is large), modeling general per-instance label distributions offers a more realistic and reliable estimation.

Further analysis reveals that annotators behave inconsistently when faced with cognitively demanding cases; e.g., they may use the *no-relation* tag when unable to identify a suitable connective. To reduce such noise, annotators could be given the option to skip uncertain cases during data collection; however, this approach is often difficult to implement reliably in a crowd-sourcing setting, as it may be prone to misuse. Another possible direction is to distinguish genuine interpretive variation from annotation errors (Klie et al., 2023; Weber-Genzel et al., 2024); particularly, annotation models may prove of use (Hovy et al., 2013; Passonneau and Carpenter, 2014; Ivey et al., 2025). However, our findings suggest that in IDRR, the inherent ambiguity may blur the boundary between the two as well. In future work, we plan to explore methods that integrate annotation error detection with perspectivist modeling to more effectively capture human label variation in highly ambiguous discourse interpretation.

7. Acknowledgments

This project is supported by the German Research Foundation (DFG) under Grant SFB 1102 (“Information Density and Linguistic Encoding", Project- ID 232722074); and by NWO under the AINed Fellowship Grant NGF.1607.22.002 (“Dealing with Meaning Variation in NLP").

8. Limitations

All of the modes compared in our paper are standard architectures using RoBERTa-base (Liu et al., 2019) as a backbone. They may not reflect the current state of the art in neural approaches to IDRR. We pick RoBERTa as a generic encoder-type model for the sake of demonstration. The same results may not hold using other advanced IDRR models.

The subset of data that we used for quantitative analysis may be insufficient to cover all the peculiarities of individual annotations. However, given that all DiscoGeM annotators worked on randomly batched items, parallel data is sparse, and we sampled the largest existing subset thereof. We consider it representative of the original perspectives, as it covers a diverse set of labels and elicits meaningful insights into the annotators' decision making. We leave further analysis into reasons for disagreement to future work.

9. Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *International Conference of the Italian Association for Artificial Intelligence*, pages 588–603. Springer.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Nelson Filipe Costa and Leila Kosseim. 2024. Exploring soft-label training for implicit discourse relation recognition. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 120–126, St. Julians, Malta. Association for Computational Linguistics.

Debopam Das, Manfred Stede, and Maite Taboada. 2017. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 11–19, Santiago de Compostela, Spain. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. *Cognitive science*, 43(1):e12701.

Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 250–259. Springer.

Jet Hoek, Merel Scholman, and Ted JM Sanders. 2021. Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6, Tübingen, Germany. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniil Ignatev, Denis Paperno, and Massimo Poesio. 2025. Annotator disagreement in rst annotation schemes. *Society for Computation in Linguistics*, 8(1).

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. Nutmeg: Separating signal from noise

in annotator disagreement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Gabriel Cardenas, Alyzeh Hussain, Austin Lara, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 1:79–108.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wanqiu Long, N. Siddharth, and Bonnie Webber. 2024. Multi-label classification for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Anh Ngo, Agri Candri, Teddy Ferdinan, Jan Kocon, and Wojciech Korczynski. 2022. StudEmo: A non-aggregated review dataset for personalized emotion recognition. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 46–55, Marseille, France. European Language Resources Association.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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 1778–1789, Minneapolis. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer.

Valentina Pyatkin, Frances Yung, Merel Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Merel Scholman, Vera Demberg, and Ted JM Sanders. 2020. Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, 57(10):844–861.

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

Alexandra 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.

Zhipang Wang, Yu Hong, Weihao Sun, and Guodong Zhou. 2025. Using subtext to enhance generative IDRR. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 464–473, Vienna, Austria. Association for Computational Linguistics.

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

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr

NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):1–34.

Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in nlp. *arXiv preprint arXiv:2601.09065*.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.

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.

Šárka Zikánová. 2024. Text structure and its ambiguities: corpus annotation as a helpful guide. In *ITAT*, pages 2–12.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162.

10. Language Resource References

Hewett, Freya and Stede, Manfred. 2025. *Disagreements in analyses of rhetorical text structure: A new dataset and first analyses*. Association for Computational Linguistics. PID https://aclanthology.org/2025.law-1.3/.

Poláková, Lucie and Mírovskỳ, Jiří and Nedoluzhko, Anna and Jínová, Pavlína and Zikánová, Šárka and Hajicová, Eva. 2013. *Introducing the prague discourse treebank 1.0*. Asian Federation of Natural Language Processing.

Prasad, Rashmi and Webber, Bonnie and Lee, Alan and Joshi, Aravind. 2019. *Penn Discourse Treebank Version 3.0*. Abacus Data Network. PID https://hdl.handle.net/11272.1/AB2/SUU9CB.

Scholman, Merel and Dong, Tianai and Yung, Frances and Demberg, Vera. 2022. *DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations*. European Language Resources Association. PID https://aclanthology.org/2022.lrec-1.351/.

Yung, Frances and Demberg, Vera. 2025. *On Crowdsourcing Task Design for Discourse Relation Annotation*. International Committee on Computational Linguistics. PID https://aclanthology.org/2025.comedi-1.2/.

Zeldes, Amir and Aoyama, Tatsuya and Liu, Yang Janet and Peng, Siyao and Das, Debopam and Gessler, Luke. 2025. *eRST: A Signaled Graph Theory of Discourse Relations and Organization*. MIT Press. PID https://aclanthology.org/2025.cl-1.3/.

Zikánová, Šárka and Nedoluzhko, Anna and Mírovskỳ, Jiří and Hajičová, Eva. 2025. *Gold Data and Multiple Understanding of Discourse Relations*. Springer. PID https://link.springer.com/chapter/10.1007/978-3-032-02551-7_22.