Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3281–3290
Marseille, 20-25 June 2022

DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations

Merel C.J. Scholman, Tianai Dong, Frances Yung, Vera Demberg
Saarland University
Campus C7.2, 66123, Saarbrücken
{m.c.j.scholman,tdong,frances,vera}@coli.uni-saarland.de

Abstract

We present DiscoGeM, a crowdsourced corpus of 6,505 implicit discourse relations from three genres: political speech, literature, and encyclopedic texts. Each instance was annotated by 10 crowd workers. Various label aggregation methods were explored to evaluate how to obtain a label that best captures the meaning inferred by the crowd annotators. The results show that a significant proportion of discourse relations in DiscoGeM are ambiguous and can express multiple relation senses. Probability distribution labels better capture these interpretations than single labels. Further, the results emphasize that text genre crucially affects the distribution of discourse relations, suggesting that genre should be included as a factor in automatic relation classification. We make available the newly created DiscoGeM corpus, as well as the dataset with all annotator-level labels. Both the corpus and the dataset can facilitate a multitude of applications and research purposes, for example to function as training data to improve the performance of automatic discourse relation parsers, as well as facilitate research into non-connective signals of discourse relations.

Keywords: discourse annotations, implicit relations, genre, crowdsourcing, label aggregation

1. Introduction

Discourse relations are semantic links between text segments (Hobbs, 1979; Sanders et al., 1992). A common distinction is made between *explicit relations*, which are marked by a discourse connective such as "because" or "however", and *implicit relations*, which do not contain a specific discourse connective. Understanding the discourse relations that hold between segments in natural language is crucial to many NLP applications, such as text generation, dialogue understanding, and question-answering systems.

Shallow discourse parsers aim to predict the sense of the discourse relation. While parsers show good performance on explicit relations (Pitler and Nenkova, 2009; Lin et al., 2014; Knaebel and Stede, 2020), performance on labelling implicit relations is significantly lacking, with the current state-of-the-art achieving an F1 around 64% on a four-way classification (Ji et al., 2016; Lan et al., 2017; Liang et al., 2020; Shi and Demberg, 2019a). Moreover, it has been shown that the performance of parsers degrades strongly if used on a different domain (Scholman et al., 2021). Therefore, creating additional resources for implicit relations in different genres is necessary. In addition, resources on implicit relations can facilitate research investigating how implicit relations are signalled.

However, obtaining manually annotated data is a costly and time-consuming process. To contribute to this data collection challenge, we used a crowdsourcing methodology to create a corpus of 6,505 annotated intersentential implicit discourse relations. We included data from three genres to foster cross-genre studies: political speech taken from the Europarl corpus, literary texts taken from 20 novels, and encyclopedic texts stemming from Wikipedia. The relations are annotated with the same annotation scheme, PDTB 3.0 (Webber et al., 2019), which makes it possible to study genre differences in a comparable framework (Webber, 2009).

This research effort also addresses a more theoretical point: whether relations, and in particular implicit relations, can express more than one meaning (Rohde et al., 2016; Scholman and Demberg, 2017b). Consider Example (1), taken from the novel *Animal Farm*. The double slash represents the break between the first and second argument.

(1) I have little more to say. I merely repeat, remember always your duty of enmity towards Man and all his ways. // Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

The second argument can be interpreted as providing detail about the "duty of enmity towards man", which is an ARG2-AS-DETAIL relation. It can also be interpreted as the resulting interpretation of the duty, which would be a RESULT relation. Indeed, Example (1) received an equal number of votes for both relation senses in the current study. Such ambiguous cases frequently occur in natural language, and assigning a single sense to those instances would not do justice to the ambiguity of the relation. By crowdsourcing many observations per relation, we obtain a distribution of relation senses per relation that might better represent the meaning inferred by multiple annotators.

We evaluate the results of a simple majority vote, an item-response model that predicts a single true label (Passonneau and Carpenter, 2014), and CrowdTruth's metrics (Dumitrache et al., 2018), which provide a

"soft" label with probabilities of different relation senses. This is a crucial step for understanding the optimal method to extract discourse relation labels from crowdsourced annotations.
The main contributions of the present research are the following:

- We present a PDTB3-style corpus of 6,505 inter-sentential implicit discourse relations and their various aggregated labels, along with all participant-level annotations (65,863 observations). Both the corpus and the raw dataset are available online.[1]
- We examine genre differences, showing that the distributions of implicit relation types differs greatly between genres.
- We evaluate various aggregration methods and label types to determine which aggregated label optimally represents relations in our data.

2. Related work

2.1. Annotation methods

Traditionally, discourse relation annotation is performed by a set of trained coders, and every item receives a label from one or two experts. The underlying assumption is that a pair of arguments can express a single relational sense. However, prior research has shown that this is not necessarily the case, since some relations tend to be more ambiguous and allow for multiple readings (Rohde et al., 2016; Scholman and Demberg, 2017b). This holds particularly for implicit relations (Hoek et al., 2021). Enforcing a single ground truth in discourse annotation will therefore sacrifice valuable nuances encoded in the relations (Aroyo and Welty, 2013).
Rather than a single label, a distribution of labels per relation can better capture the range of possible interpretations of a relation. However, in order to obtain a relational distribution for every item, one would need many annotations per item, which is often not feasible for traditional annotation projects.
Crowdsourcing can provide a solution: it gives access to a large number of participants that can quickly provide annotations. The obtained annotations are independent and do not rely on implicit expert knowledge. Crowdsourcing has been used in various efforts to obtain discourse relation interpretations (Kawahara et al., 2014; Pyatkin et al., 2020; Rohde et al., 2016; Scholman and Demberg, 2017a). We here use a two-step insertion method, proposed by Yung et al. (2019). In this approach, PDTB3-relation labels are inferred from connectives that participants insert into the text (see Section 3).

[1] https://github.com/merelscholman/DiscoGeM

In a series of crowdsourced annotation studies, Yung et al. (2019) showed that the method can be successfully used to reproduce the original PDTB (Prasad et al., 2008) and RST-DT (Carlson et al., 2003) labels for implicit relations, and that the obtained annotations are robust and replicable. Moreover, the method captured the ambiguity of relations by providing a distribution of relation senses, which more accurately represented the relations' true meaning.

2.2. Label aggregation

It has become common to crowdsource annotations for various classification tasks for NLP. To deal with inter-annotator agreement, many efforts have aggregated the different answers into a supposedly true answer. One common method is to obtain a silver label by taking the majority vote as the true label; that is, consensus is enforced by taking the label that received the most votes for an individual item.
Such majority-based silver labels can be considered noisy, as they do not take into consideration annotator quality and biases. More advanced methods correct for this. For example, the Dawid-Skene model, implemented by Passonneau and Carpenter (2014) and referred to here as IRT, is a probabilistic annotation model that uses unsupervised learning to estimate the probability of labels for every item and coder. It adjusts for noisy annotators and provides a "soft" label, i.e. a probability distribution over the labels provided by the annotators, with an estimate of a single true label.
CrowdTruth (Aroyo and Welty, 2013; Dumitrache et al., 2018) is another method that creates a soft label by harnessing inter-annotator disagreement. Crucially, CrowdTruth rejects the notion that there is a single true label for each item. This is important to discourse relation classification, as these relations tend to be ambiguous. Instead of enforcing agreement between annotators, CrowdTruth captures the ambiguity inherent in semantic annotation through the use of disagreement-aware metrics. CrowdTruth's probability distributions are therefore less centered on a single label than IRT's distributions.
In the current contribution, we will evaluate the performance of these different metrics on discourse relation classification, comparing the labels obtained through simple majority crowd aggregation, Passonneau and Carpenter (2014)'s Dawid-Skene model and Dumitrache et al. (2018)'s CrowdTruth 2.0 approach.

2.3. Genre and discourse relations

Text genre crucially affects the distribution of discourse relations (Rehbein et al., 2016; Webber, 2009). For example, Rehbein et al. (2016) observed a substantial difference in the frequency of discourse relation types between spoken and written genres, as well as between two spoken subgenres (broadcast interviews versus telephone conversations). However, most discourse-annotated data that are available come from

	EP	Lit	Wiki	Total
No. DRs	2,800	3,060	645	6,505

Table 1: Corpus size in number of discourse relations per genre and in total.

a few restricted genres: newspaper text (e.g., Penn Discourse Treebank, (Prasad et al., 2008; but see Webber (2009)), biomedical text (e.g., BioDRB, (Prasad et al., 2011)), and spoken text (Rehbein et al., 2016; Zeyrek et al., 2019) (an exception is the GUM corpus, Zeldes (2017)).

In order for NLP tools to be more robust and generally applicable, training data from various genres and domains are necessary. Scholman et al. (2021) showed that the performance of connective identification models drops significantly when applied to genres other than newspaper text. This likely extends to discourse relation classification parsers. Indeed, Shi and Demberg (2019b) show that training on in-domain data improves performance of discourse relation parsers.

The DiscoGeM corpus addresses these gaps by including data from three different genres: political speech (from the Europarl corpus), literary texts (from 20 novels), and encyclopedic texts (from Wikipedia). We include the speaker tags for Europarl and authorship for the novels, to allow for investigations into inter-speaker variation. Further, we include English text from different source languages (English original or translated from German, French, or Czech). We plan to extend the corpus with parallel annotations of these other languages, to facilitate cross-linguistic studies.

This corpus is the first to fully enable an extensive genre analysis of implicit relation distributions. Comparing distributions in existing corpora has as a major drawback that those corpora have all been annotated by different annotators, which limits their comparability. For example, the Chinese Discourse Treebank (Xue, 2005) contains text from the same genre as the PDTB and is annotated using the same framework, yet the distributions of relations in these corpora are different. This is likely not only due to the difference in language or in text source, but also to the different sets of annotators.

3. Methodology

3.1. Data

Table 1 presents details on the corpus size per genre.

Europarl The Europarl genre consists of political speech, oftentimes prepared. The nature of political discourse can be described as more argumentative than the other two genres included in DiscoGeM. The relations originally uttered in English were gathered from the Europarl corpus (Koehn, 2005), the translated English relations from Europarl Direct (Cartoni and Meyer, 2012), which is a directional parallel corpus with more available meta-data. For every source language, the corpus contains 700 implicit relations, totalling 2,800 English implicit relations. However, due to an error in sampling the English original relations, only 296 items of the intended English original data are verified to be English original.[2]

Literature This genre consists of narrative writing, which can be described as an account of a sequence of fictional or nonfictional events, usually in chronological order. The literary genre will therefore likely have a higher rate of temporal relations than the other genres in DiscoGeM. The corpus contains a total of twenty books and 160 relations per book.[3] Of the 20 books, 5 were originally English and the remaining 15 were translations (5 per source language). The online appendix contains a list of the titles of the books that were included.

Wikipedia The Wikipedia genre is classified as written informative (encyclopedic) texts, which explain known facts about specific topics. This genre will therefore likely have a higher rate of CONJUNCTION and ARG2-AS-DETAIL relations. DiscoGeM includes the first section (i.e. the summary) of Wikipedia texts on a total of 69 topics. For these texts, the reference labels were available (see Section 3.5). A maximum of 20 items per text were included in our corpus, or fewer if there were less than 20 implicit relation candidates in the summary. The first item of every new Wikipedia text was presented with a note regarding the Wikipedia entry title ("*[Context: This is a text on X]*").

3.1.1. Data preparation

Implicit relation candidates were identified automatically from the text files. Candidates were defined as two consecutive sentences within the same paragraph for which the second sentence did not contain an explicit connective in the first five words of the sentence (determined through string match).[4] We excluded relations with arguments shorter than five words or longer than 50 words. This was done as a quality control measure; arguments longer than that tend to be demotivating to participants and therefore receive noisier labels. Finally, for the literary genre, we additionally excluded candidates from dialogues (i.e. "He said X. She said Y."). Such cases were very frequent and considered to be less informative relations, as they are simply a narration of responses between two speakers. Had they been included, the proportion of Temporal relations in the novels genre would likely have been higher.

Relations were provided with additional context to facilitate the annotation. The first arguments of items

[2] In the Europarl corpus, a proportion of the data was untagged. During sampling, we mistakenly assumed these were English instances, but in reality they can also be translated text. The unverified instances are marked in DiscoGeM.

[3] Three books did not have up to 160 candidates.

[4] See `https://github.com/merelscholman/DiscoGeM` for the list of connectives used to exclude candidates.

from Europarl and literature consisted of on average three sentences, and the second arguments of on average 2 sentences, unless the speaker ended their turn (Europarl) or a paragraph ended (novels) after one sentence. For the Wikipedia genre, the first arguments of items consisted of all preceding sentences within the same paragraph, and the second arguments consisted of two sentences following the argument break (i.e. Arg2 consisted of two sentences), unless the paragraph ended after one sentence.

3.2. Task design

Participants were shown a text passage containing a blank between two text segments. They completed two steps for every item. In the first step, they were asked to type in the blank a connective that they thought best expressed the relation between the textual arguments. They were also given the option to type *nothing* if they thought no phrase could fit between the segments.

The freely inserted connectives were often ambiguous. Consequently, it would not be possible to infer a specific discourse relation label from these free insertions. The second step addressed this: participants were asked to choose from a list of at most 10 connectives that disambiguated the connective they inserted in the first step. The selection of the connectives was determined dynamically from their choice in the first step (see Section 3.2.1). They could choose the *none of these* option if they thought none of the given options fit.

When the insertion in the first step did not match any of the entries in our connective bank, or when the participant typed *nothing*, participants were presented with a default list of twelve connectives that that can express a variety of relations. This list of default connectives consisted of *for example, more generally, more specifically, to provide background information, in addition, also, despite this, even though, by contrast, therefore, subsequently,* and *due to*.

3.2.1. Connective bank

The mapping between the free insertion in the first step and the choices provided to the participants in the second step stems from a connective bank. This bank was created specifically for this method, based on existing discourse resources. All connectives were manually annotated for the different senses they can express. The set of relational labels is based on the sense hierarchy of PDTB3.[5] We extensively tested the coverage of the connective bank in pretests and earlier studies (Yung et al., 2019), in order to capture the possible connectives used by the participants.

The final version of our connective bank contains over 2,000 entries, which include typical discourse connectives (e.g. *because*), variations of connectives (e.g. *largely because*), combinations of connectives (e.g. *and because*), frequent typos (e.g. *becuase*), and "alternative lexicalizations" (e.g. *the reason is that*).

[5]We cover each Level-3 sense in PDTB 3.0, except the belief and speech-act relations. These relations cannot be distinguished reliably with their non-belief and non-speech-act versions by means of the inserted connective.

Relation sense	Most common free insertion in Step 1	Most common connective in Step 2
TEMPORAL		
PRECEDENCE	subsequently	subsequently,
CAUSE		
REASON	because	the reason(s) is/are that
RESULT	as a result	consequently,
COMPARISON		
ARG2-AS-DENIER	however	despite this,
CONTRAST	however	by comparison,
EXPANSION		
CONJUNCTION	also	in addition,
ARG2-AS-INST.	for example	for instance,
ARG2-AS-DETAIL	in more detail	in more detail,

Table 2: The connective mapping for the eight most frequent relation types in DiscoGeM. The full list is available in the online appendix.

The list given in Step 2 contains connectives that mark the relation senses that we want to distinguish as unambiguously as possible. We determined these connectives using Knott (1996)'s connective hierarchy and PDTB's connective lists. Table 2 presents the connective mapping for the eight most common type of relations in DiscoGeM; the complete list can be found in the online appendix.

3.3. Crowdsourced annotators

We recruited participants registered on Prolific who matched the following prerequisites: their native language must be English, their current country of residence must be the UK or Ireland, the highest completed education level must be an undergraduate degree, their minimum approval rate must be 95%, their minimum number of previous submissions must be 150.

Following Scholman et al. (2022), we first ran a selection task to obtain annotations that allowed us to evaluate the accuracy of the participants. This task also contained a feedback component to implicitly train participants: if their answer did not match the reference label, participants were presented with an explanation of what was expected and why that interpretation was expected. Further, a self-selection component was included: upon completion, participants were asked to rate how much they enjoyed the task and if they would like to continue. All participants that scored higher than 50% agreement with the gold labels for this text and indicated that they wanted to continue were included in our pool of final participant candidates. The standard of 50% agreement was determined as an acceptable level of agreement on this particular text in pretests.

Of 310 participants that took part in the selection task, 199 were included in our final participant pool (mean

age: 41 years; age range: 19-77 years; 144 female). These participants were invited to take part in the studies. They were allowed to participate in more than one study, with a maximum of five studies per day. Participation ranged from 1 study per participant to 41 studies in total, with the average participant having taken part in 17 studies.

Quality of the annotations was checked on two occasions during the data collection phase. We calculated for every participant, and for every batch they participated in, what percentage of quality checks they passed, based on three measures: (i) provide more than 7 unique connective types per batch of 20 items in Step 1; (ii) indicate that "no connective fits" less than 5 times per batch in Step 2; (iii) at least 25% agreement with the reference label, if available. A score of 1 was assigned if they passed the check; 0 if they failed. We then created a composite score by averaging the scores across all batches that a participant completed. Participants who scored less than 90% on our composite measure were not invited for further studies, but their data was included in the final dataset. In other words, we did not exclude any data, but we did aim to retain the best performing participants that provided the most informative labels. In total, 16 participant were excluded after the first check, and an additional 10 participants were excluded from further participation after the second check. The raw dataset with participant-level annotations contains information on the participants' accuracy scores.

3.4. Procedure

The implicit relations were divided into batches. A batch only contained data from the same text type. Relations within a batch were presented sequentially (i.e. according to the original order in the text) to allow participants to benefit from the context. Every batch contained approximately 20 relations.

In total, we collected data for 329 batches. Data collection took place in December 2021. Every batch was completed by 10 participants.[6] The task was implemented on LingoTurk (Pusse et al., 2016).

Participants were awarded with 1.88 British pounds for each batch of annotation. On average, participants spent 20 seconds on one item. Participants could take part in at most five batches per day. The batches that were uploaded per day came from the different genres, to encourage higher annotator engagement due to more variability in text type.

3.5. Reference annotations

The reference annotations for Wikipedia texts were available to compare the crowdsourced labels against. These reference labels were created by two trained, expert annotators, and adjudicated by another annotator.

The reference annotations were done using a traditional approach to annotation; that is, they did not annotate using the two-step interface but rather assigned labels directly. The annotators were allowed to annotate multiple labels per relation if they inferred more than one sense, to properly reflect the different interpretations that can be inferred.

The two annotators showed frequent disagreements (60% agreement on a level three distinction; κ=.45).[7] However, under the assumption that the implicit relations can often signal more than one relation sense, disagreements do not necessarily reflect incorrect labels (Aroyo and Welty, 2013). This is why a third annotator adjudicated the two annotations to create the reference label. Accordingly, disagreements between the two annotators could result in a gold label consisting of multiple labels, or in the adjudicator selecting one annotation and rejecting the other. In total, 31% of data received a single reference label, 56% of data was labeled with two senses, 12% with three senses and 1% with four senses. The most frequent co-occurring senses were CONJUNCTION and ARG2-AS-DETAIL.

For our calculations of inter-annotator agreement between the adjudicator and the two annotators, we consider a partial match as agreement in order to account for the multiple possible interpretations. The agreement between the first annotator and the reference label was κ=.82 (88% agreement), and between the second annotator and the reference label was κ=.96 (97% agreement).

4. Results

4.1. Obtaining an aggregated label from the crowdsourced insertions

We calculated the following aggregated labels:

- **Majority-single**: sense that received majority agreement from the crowd workers. In case the majority label consisted of more than one sense, a single sense was randomly chosen.
- **IRT-single**: sense with the highest probability, based on the Dawid-Skene model (Passonneau and Carpenter, 2014). The model was run separately on each genre.
- **Majority-distribution**: combination of all senses that reached a threshold of 20% agreement. In cases where no sense reached the threshold for an instance, the single sense was assigned.

[6] 27 batches contain observations from 11 participants due to erroneous automatic list assignment and two batches contain more observations because extra data was collected for a related project.

[7] As a general guideline, Spooren and Degand (2010) consider a kappa of .7 to signal good IAA for DR annotation. However, IAA on implicit relations is known to be lower than on explicits, see, e.g., (Demberg et al., 2019; Hoek et al., 2021; Kishimoto et al., 2018).

- **CrowdTruth-distribution**: combination of all senses that reached a threshold of 0.2 probability based on CrowdTruth 2.0 (Dumitrache et al., 2018).

The single measures speak to the traditional assumption that discourse relations convey a single relational sense, whereas the distribution measures represent the idea that relations are ambiguous and can convey multiple senses. First, we look at the distribution of relation senses for the single measures, to determine whether the labels show distinct patterns or biases in their label assignment. Next, we look at the number of senses per label for the distribution measures, to determine how many senses a label includes when using distribution aggregation labels.

Single measures Figure 1 displays the distribution of relation senses across the various aggregation methods. For ease of comparison, this visualization includes the labels that received the highest probability in CrowdTruth. The visualization indicates that the IRT measure corrects for the prevalence of categories in the data by assigning the two most frequently occurring senses - RESULT and CONJUNCTION - less often. The agreement with the reference labels will reveal whether IRT's prevalence correction is desired behaviour.
When looking at agreement between the majority measure and IRT, we find that the measures come to different conclusions for some of the data (κ=.79; %=83). A large portion of the disagreements (55%) are cases where the majority response is CONJUNCTION or ARG2-AS-DETAIL, and IRT assigned a more informative label. Note that, in many of these cases, both senses are included in the majority distribution and CrowdTruth distribution labels. Consider Example (2), taken from the novel *In Search of Lost Time*.

(2) He's a crafty customer, always sitting on the fence, always trying to run with the hare and hunt with the hounds. What a difference between him and Forcheville. // There at least you have a man who tells you straight out what he thinks. Either you agree with him or you don't.

This item was interpreted as ARG2-AS-DETAIL by five annotators. Under this reading, the second argument provides additional detail on what the difference between "him and Forcheville" is. An additional three annotators interpreted Example (2) as REASON, whereby the second argument provides the reason for the speaker uttering the first argument. The majority response for this item is therefore ARG2-AS-DETAIL, but both senses are included in the distribution labels. IRT, however, assigned the highest probability to the REASON sense, likely correcting for the prevalence of the ARG2-AS-DETAIL relation sense.

Distribution measures Table 3 presents an overview of the average number of senses per measure. It shows

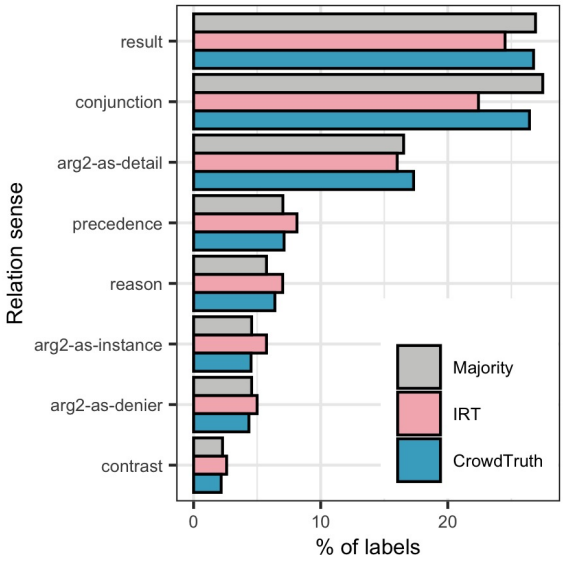

Figure 1: Distribution of the most frequent relation senses in DiscoGeM.

Measure	Mean senses	% single senses	Max. senses
Majority - distribution	2.2	20	5
CrowdTruth - distrib.	1.8	37	4

Table 3: Distribution measures information. *Mean senses*: mean number of senses per label; *% single senses*: percentage of data for which the distribution measure contains a single sense; *Max. senses*: maximum number of senses included in one label.

that the majority distribution measure on average contains more senses per label than CrowdTruth, which is more selective given the chosen threshold of 20%.
The relation senses that co-occur frequently are stable across both measures, as shown in Table 4. This table shows the frequency of sense combinations per aggregation method, out of all instances that have a label consisting of multiple senses. Some of these combinations are not surprising: ARG2-AS-DETAIL and CONJUNCTION are closely related relation senses, as are ARG2-AS-DETAIL and ARG2-AS-INSTANCE, and PRECEDENCE and RESULT. The prevalence of the combination of CONJUNCTION and RESULT, and ARG2-AS-DETAIL and RESULT is more unexpected, but as Example (3) illustrates, it can in fact be quite easy to interpret both readings for a single item.

(3) 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.

Example (3) is taken from the Europarl genre. This item was interpreted as CONJUNCTION by four annotators and RESULT by five annotators. Under the

Frequently co-occurring senses	Maj	CT
CONJUNCTION & RESULT	11	13
ARG2-AS-DET. & CONJUNCTION	9	12
PRECEDENCE & RESULT	5	5
ARG2-AS-DET. & RESULT	4	6
ARG2-AS-DET. & ARG2-AS-INST.	3	4

Table 4: Most frequent senses found to co-occur (% of number of multi-sense labels). *Maj*: majority distribution measure; *CT*: CrowdTruth distribution measure.

Aggregated measure	κ	%	P	R	F1
Majority - single	.55	67	.67	.49	.55
IRT - single	.53	64	.64	.46	.52
Majority - distrib.	**.79**	85	.54	.70	.58
CrowdTruth - distrib.	.75	82	.69	.66	**.59**

Table 5: Agreement statistics per aggregated measure.

CONJUNCTION reading, the speaker presents two facts about cities. However, one can also interpret these facts in a causal manner: because they are home to 80% of the inhabitants, the majority of jobs is located in cities. Both interpretations are therefore valid, which is reflected in the distribution labels.

4.2. Comparing the aggregated labels with the reference annotations

The agreement between the reference labels and aggregated measures was assessed in various ways. We calculated Cohen's Kappa (Cohen, 1960) using the intersection of agreed-upon labels. We corrected multi-label annotations to the intersecting label or otherwise the randomly sampled label. Note that this method underestimates chance agreement, because the chance of agreeing on a sense is higher when a label contains multiple senses. We also calculated agreement in terms of recall, precision and F1. Precision equals the proportion of true positive labels compared to the total set of provided labels; i.e. to what extent the aggregated label contains the reference label senses. Recall equals the proportion of true positive labels compared to the total set of relevant labels; i.e. whether the aggregated measure missed any senses that are part of the reference label. F1 is the weighted average of the two.

Table 5 presents the agreement statistics with the reference labels for Wikipedia data per aggregated label type. These results are comparable to other implicit annotation efforts. For example, Kishimoto et al. (2018) report an F1 of .51 on crowdsourced annotations of implicit relations; Hoek et al. (2021) report a κ of .58 on expert annotations of implicits; and Demberg et al. (2019) find that PDTB and RST-DT annotators agree on 37% of implicit relations.

Of all aggregation measures evaluated in the current study, the majority distribution measure shows highest agreement with the reference labels in terms of intersection kappa, and the CrowdTruth distribution measure in terms of F1. Given that the latter assigned fewer senses on average than the majority-distribution measure, the CrowdTruth distribution measure can be considered a competitive measure.

Naturally, the more senses per label that the aggregated measures have, the higher the chance of agreement. This means that the chosen threshold impacts the agreement statistic. Here, we have chosen a relatively low threshold of 20%. To evaluate the impact of threshold, we also compared the intersection kappa agreement with the reference label for measures at various other increments. Agreement indeed declines when the threshold is raised: at a threshold of 40%, both distribution measures show an agreement of κ=.57 with the reference. For future studies, efforts should evaluate the trade-off between accuracy and number of included senses per label to determine an optimal threshold.

Figure 2 presents the distribution of the three most frequent reference relation senses across the single label measures. For ease of comparison, this figure again contains the single CrowdTruth labels. The visualization shows that the IRT label tends to overcorrect for CONJUNCTION. In other words, we again see that IRT penalizes frequent senses more because it assumes biases in the data. This is not always desired, as in the case of CONJUNCTION relations.

Another observation that can be made from Figure 2 is that when the measures disagree with the reference label, all three often prefer a closely related label. For example, when the reference label is PRECEDENCE, the other measures at times prefer CONJUNCTION or RESULT, but not ARG2-AS-DETAIL or ARG2-AS-INSTANCE. Hence, even when the aggregated labels disagree with the reference labels, the provided labels are relatively similar. This is a positive result and supports the reliability of the method.

4.3. Distribution of implicit relations in various genres

To get a better understanding of the differences between the genres, we look at the distribution of the discourse relations in the data. Figure 3 presents the distribution across the three genres for the eight most frequent relation senses in the data.

There are clear differences in the distributions. First, we see that CONJUNCTION is more prevalent in Wikipedia text than in other genres. This was expected, given that Wikipedia texts typically presents facts about a single topic.

Second, we can see that there are more PRECEDENCE relations in literature than in other genres. Again, this was expected, given that books tend to contain a sequence of events in chronological order. Note that we had excluded relation candidates that contained narrated speech, and so the actual proportion of temporal relations in novels is likely even higher. Also noticeable is that Europarl is characterized by a very low proportion of temporal relations.

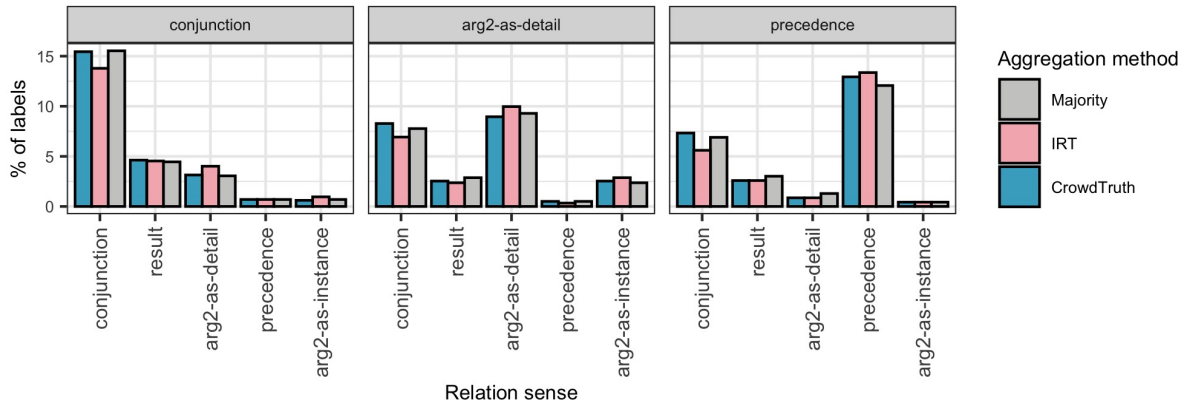

Figure 2: Distribution of the three most frequent reference labels across the single label measures.

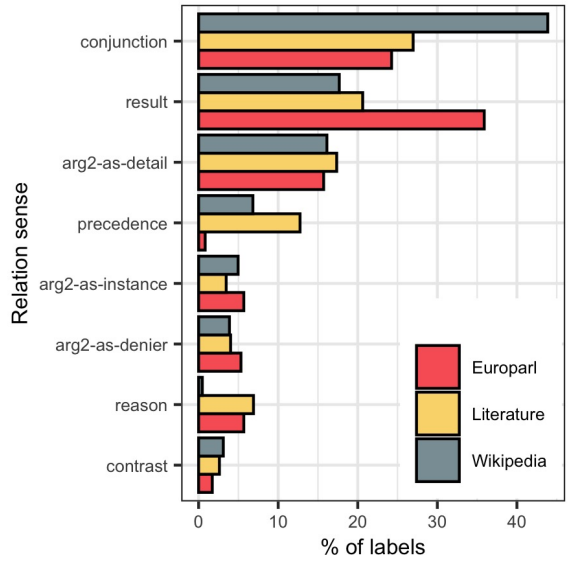

Figure 3: Distribution of the most frequent relation senses across genres; labels from the majority-single measure.

Third, Figure 3 shows that RESULT relations tend to occur more in Europarl than in other genres. We had expected a higher proportion of REASON relations in this genre, but the dominance of RESULT relations is not surprising; it simply reflects a different order of causal arguments. Looking at REASON relations, we do not find a particularly high proportion in the Europarl data. However, we can see that there is a very low proportion of REASON relations in the Wikipedia data. This can be attributed to the encyclopedic nature of the texts.
What we can conclude from this overview is that relation distributions are very different between genres. This highlights the importance of taking genre effects into consideration in discourse analyses.

4.4. Descriptive data for the methodology

A total of 3,426 unique insertions were provided by participants in the first step. These could be mapped to 325 entries in the connective bank. For example, "because", "becuase" and "becaue" were all mapped to the same entry in the connective bank: "because". 6.8% of the first step insertions could not be mapped to a connective in the connective bank. These instances were "new" typo's that were not encountered during extensive pretesting, connectives followed or preceded by additional content words (e.g., "yet we may say that"), or other, non-connective insertions (e.g., "which"). In such cases, participants were presented with a default list, which allowed us to recover an annotated label. Finally, participants indicated that none of the provided connectives fit to describe the relation for 3.3% of the responses in the second step. These numbers all converge with the Results obtained in Yung et al. (2019), further supporting the robustness of the method.

5. Conclusion and future work

We have presented DiscoGeM, a crowdsourced corpus of genre-mixed implicit discourse relations. It is equipped with several variations of annotated labels, of which we recommend to use CrowdTruth's soft label or the majority single label, if a single label is preferred.

The corpus can be useful for various research purposes. First, the data can be used to improve the accuracy of implicit relation parsers on different genres, given that the results show that genres differ from each other in the associated sense distributions for implicit relations. Another interesting line of research could be to exploit the translation aspect of the corpus by studying how the corpus can contribute to transfer learning in discourse classification tasks (e.g., train on English translated text and test on original text) (Long et al., 2020).

Further, we release annotator-level annotations and meta-data such as annotator quality based on IRT and CrowdTruth, to allow researchers to further study the effect of annotator noise on the resulting labels. This data can be used to, for example, filter presumably noisy data and train on the remaining data.

Acknowledgements

We are thankful to Marian Marchal for fruitful discussion on the project. This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding".

6. Bibliographical References

Aroyo, L. and Welty, C. (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013 ACM*, 2013(2013).

Carlson, L., Marcu, D., and Okurowski, M. E. (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.

Cartoni, B. and Meyer, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Demberg, V., Scholman, M. C. J., and Asr, F. T. (2019). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, pages 87–135.

Dumitrache, A., Inel, O., Aroyo, L., Timmermans, B., and Welty, C. (2018). CrowdTruth 2.0: Quality metrics for crowdsourcing with disagreement. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management, SAD+ CrowdBias 2018*, pages 11–18. CEUR-WS.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Hoek, J., Scholman, M. C. J., and Sanders, T. J. (2021). Is there less annotator agreement when the discourse relation is underspecified? In *Integrating Perspectives on Discourse Annotation (DiscAnn)*, pages 1–6.

Ji, Y., Haffari, G., and Eisenstein, J. (2016). A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342, San Diego, California, June. Association for Computational Linguistics.

Kawahara, D., Machida, Y., Shibata, T., Kurohashi, S., Kobayashi, H., and Sassano, M. (2014). Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 269–278, Dublin, Ireland.

Kishimoto, Y., Sawada, S., Murawaki, Y., Kawahara, D., and Kurohashi, S. (2018). Improving crowdsourcing-based annotation of Japanese discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Knaebel, R. and Stede, M. (2020). Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse (CoDi 2020)*, pages 65–75, Online, November. Association for Computational Linguistics.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, The University of Edinburgh: College of Science and Engineering.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Lan, M., Wang, J., Wu, Y., Niu, Z.-Y., and Wang, H. (2017). Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark, September. Association for Computational Linguistics.

Liang, L., Zhao, Z., and Webber, B. (2020). Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Long, W., Webber, B., and Xiong, D. (2020). Ted-cdb: A large-scale chinese discourse relation dataset on ted talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803.

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

Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore, August. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):1–18.

Pusse, F., Sayeed, A., and Demberg, V. (2016). Lingo-

Turk: Managing crowdsourced tasks for psycholinguistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 57–61, San Diego, CA.

Pyatkin, V., Klein, A., Tsarfaty, R., and Dagan, I. (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.

Rehbein, I., Scholman, M. C. J., and Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Rohde, H., Dickinson, A., Schneider, N., Clark, C., Louis, A., and Webber, B. (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 (LAW X)*, pages 49–58, Berlin, Germany.

Sanders, T. J. M., Spooren, W. P. M. S., and Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Scholman, M. C. J. and Demberg, V. (2017a). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW)*, pages 24–33, Valencia, Spain.

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

Scholman, M. C. J., Dong, T., Yung, F., and Demberg, V. (2021). Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the First Workshop on Computational Approaches to Discourse (CoDi 2020)*.

Scholman, M. C. J., Pyatkin, V., Yung, F., Dagan, I., Tsarfaty, R., and Demberg, V. (2022). Design choices in crowdsourcing discourse relation annotations: The effect of worker selection and training. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France. European Language Resources Association (ELRA).

Shi, W. and Demberg, V. (2019a). Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden, May. Association for Computational Linguistics.

Shi, W. and Demberg, V. (2019b). Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5790–5796.

Spooren, W. P. M. S. and Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682.

Xue, N. (2005). Annotating discourse connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 84–91, Ann Arbor, MI.

Yung, F., Demberg, V., and Scholman, M. C. J. (2019). Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.

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

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., and Ogrodniczuk, M. (2019). TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–27.