DiscoGeM 2.0: A Parallel Corpus of English, German, French and Czech Implicit Discourse Relations

Frances Yung[1], Merel Scholman[1,2], Šárka Zikánová[3], Vera Demberg[1]
[1]Saarland University, [2]Utrecht University, [3]Charles University
Saarbrücken, Germany; Utrecht, Netherlands; Prague, Czech Republic
{frances, m.c.j.scholman, vera}@coli.uni-saarland.de, zikanova@ufal.mff.cuni.cz

Abstract

We present DiscoGeM 2.0, a crowdsourced, parallel corpus of 12,834 implicit discourse relations, with English, German, French and Czech data. We propose and validate a new single-step crowdsourcing annotation method and apply it to collect new annotations in German, French and Czech. The corpus was constructed by having crowdsourced annotators choose a suitable discourse connective for each relation from a set of unambiguous candidates. Every instance was annotated by 10 workers. Our corpus hence represents the first multi-lingual resource that contains distributions of discourse interpretations for implicit relations. The results show that the connective insertion method of discourse annotation can be reliably extended to other languages. The resulting multi-lingual annotations also reveal that implicit relations inferred in one language may differ from those inferred in the translation, meaning the annotations are not always directly transferable. DiscoGem 2.0 promotes the investigation of cross-linguistic differences in discourse marking and could improve automatic discourse parsing applications. It is openly downloadable here: https://github.com/merelscholman/DiscoGeM.

Keywords: implicit discourse relations, distributional soft labels, multilingual

1. Introduction

Discourse relations are semantic links between text segments (Hobbs, 1979; Sanders et al., 1992). They can either be marked explicitly, through connectives such as *because* and *nevertheless*, or they can be implicit (i.e., not marked linguistically through the use of a 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. While discourse parsers show good performance on explicit relation classification (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 72% on four-way classification (Liu et al., 2021a; Xiang et al., 2022, 2023; Wang et al., 2023). One of the reasons is that implicit discourse relations are highly ambiguous and can have various interpretations (Sanders et al., 1992; Rohde et al., 2016), which are better represented and learned by soft labels (Yung et al., 2022; Pyatkin et al., 2023). Such soft labels consist of probability distributions over the various labels given by the annotators to a single instance. These labels thus retain information regarding the different interpretations that can be inferred for ambiguous discourse relations.

Discourse parsers also typically do not perform well in languages other than English (Braud et al., 2023) and in out-of-domain settings (Atwell et al., 2022; Gessler et al., 2021; Liu et al., 2021b; Scholman et al., 2021). This is largely due to the lack of substantial amounts of annotated data in the different domains and languages. While there are some discourse-annotated resources for the languages contained in our new dataset, such as the French *Annodis Corpus* (Afantenos et al., 2012), the German *Potsdam Commentary Corpus* (Bourgonje and Stede, 2020), and the Czech *Prague Discourse Treebank* (Synková et al., 2022), these are not comparable in terms of genre and topic and are limited in size. To date, there is one discourse-annotated parallel dataset covering several languages (TED-MDB, Zeyrek et al., 2020), but it is rather small, covering around 200 implicit relations per language.

The current work addresses the lack of multilingual data by extending an existing multi-genre corpus of English implicit discourse relations (DiscoGeM 1.0, Scholman et al., 2022a) to three additional languages: DiscoGeM 2.0 is a parallel corpus with German, French and Czech annotations of the same relations previously available in English. German stems from the same Germanic language family as English, whereas French (Romance) and Czech (Slavic) represent different language families. The corpus can thus provide interesting perspectives on cross-linguistic differences between language families.

The relations in DiscoGeM are labelled by 10 annotators, and thus provide a soft label distribution that better captures the relation interpretations. Cortez and Jacobs (2023a,b) and Yung

LREC-COLING 2024, pages 4940–4956
20-25 May, 2024.

et al. (2022) show that soft labels are useful for discourse relation representation and classification. The distributional representation also allows more thorough evaluation of soft model predictions than evaluation against a single gold label (Ru et al., 2023).

In order to facilitate the expansion of the English annotation interface to German, French and Czech annotation interfaces, we modified the annotation methodology used to create DiscoGeM 1.0. The current work shows how annotation methodologies can be adapted to different languages and evaluates the effect of such adaptations on the output. Our main contributions are the following:

- We design an approach to annotate implicit discourse relations in German, French and Czech. This approach can be easily extended to other languages as well.
- We present a parallel corpus of 12,834 inter-sentential implicit discourse relations, with English, German, French and Czech data, each annotated with PDTB3-style labels by at least 10 annotators through crowdsourcing. The corpus is freely downloadable and is an invaluable resource for training and evaluating implicit discourse relation parsers with soft-labels, for adapting parsing models to different domains and languages, and for empirical studies of discourse relations in translation and human discourse relation interpretation in different languages.

2. Related work

2.1. Obtaining parallel discourse annotations

Annotating data is a resource-intensive undertaking. A possible alternative to obtain annotations for multilingual resources would be to project annotations from English to another language in parallel texts (Laali and Kosseim, 2017; Meyer et al., 2011; Sluyter-Gäthje et al., 2020). Annotation projection relies on the assumption that *discourse relations are preserved in translation*. However, this assumption may not always hold: the discourse relation can be changed in the process of translation such that the same overall content is expressed, but the discourse relation sense is not identical anymore. This can in part be attributed to the discourse marking of the relations changing during translation: the relation can be translated without the connective that was originally present (i.e. implicitated), or the connective may be translated as a more ambiguous variant (Crible et al., 2019; Yung et al., 2023), depending also on text genres (Zikánová et al., 2019). When a relation is implicitated, a sense shift can occur between the languages (Zeyrek et al., 2022). This means that readers of the translated texts may infer a different discourse relation than the readers of the original source texts. We need further insight into the degree to which relations are preserved across translations on a larger scale (i.e. looking at translations in different domains and in different translation settings) before we can fully rely on a projection approach. The corpus produced for the current work provides data that can be used to study this issue.

Another way to address the lack of multilingual resources is to annotate more data. The current work does this using crowdsourced annotators. Crowdsourcing platforms provide access to a larger pool of annotators, which makes it easier for each instance to be annotated by multiple people. This allows us to obtain a distribution of labels that better captures the range of possible interpretations of an instance. This is especially beneficial in the context of discourse relation annotation, because discourse relations can be interpreted differently based on the reader's perspective.

To make it possible to crowdsource discourse relation annotations, Yung et al. (2019) propose the two-step methodology, where the workers first freely insert a connective to label the discourse relation between two consecutive sentences, and then disambiguate this choice from a list of unambiguous connectives, which are dynamically generated based on the first connective. This approach was applied to create DiscoGeM 1.0.

In the current work, the two-step approach is redesigned into a one-step approach and applied to different languages. In doing so, we need to be aware of potential biases that are introduced by changes to the annotation methodology, as prior work has shown that different methods can produce different annotations. Specifically, Pyatkin et al. (2023) compared the output of the discourse connective method used for DiscoGEM 1.0 with an alternative crowdsourcing discourse relation annotation method (a Question-Answer based approach, Pyatkin et al., 2020). They found that both methods are valid, as similar sets of labels are produced for many instances, but also that both methods have unique biases.

2.2. DiscoGeM 1.0 corpus

The current work presents an expansion of the DiscoGeM corpus (Scholman et al., 2022a, further referred to as DiscoGeM 1.0), which is a crowdsourced corpus of 6,505 inter-sentential discourse relations. DiscoGeM 1.0 was created using the two-step DC annotation method (described in the previous section) and with the help of 199 crowdsourced annotators (see Scholman et al., 2022a, for a full description of the procedure). The data

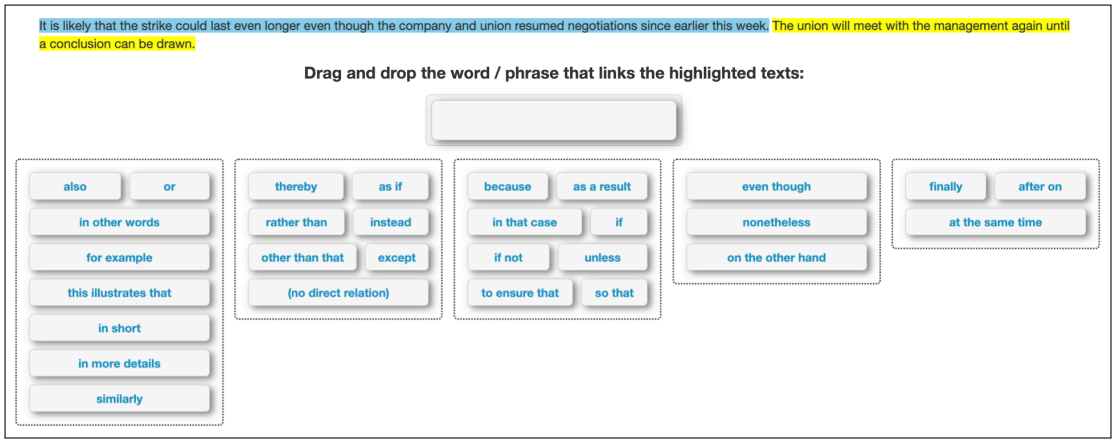


Figure 1: The English interface we used in the verification experiment. The interfaces in other languages that are used in the actual crowdsourcing tasks are shown in Figure 6,7, and 8 in the Appendix.

stems from four genres: European Parliament speeches (Europarl), literature from a collection of twenty novels, Wikipedia texts, and the Wall Street Journal[1].

DiscoGeM 1.0 contains only English text, part of which was translated to or from other languages: German, French and Czech. In this work, we first align the English sentences with the corresponding sentences in the other languages. We then crowdsource annotations for these German, French and Czech discourse relations, thereby creating a novel parallel corpus of inter-sentential implicit discourse relations.

3. Methodology

3.1. Challenges in transferring the connective insertion method to other languages

The crowd-sourcing annotation method used on DiscoGem 1.0 consists of two steps (Yung et al., 2019). In the first step, annotators freely insert a connective that they think fits well between the arguments. In a second step, this connective is disambiguated by providing a selection of connectives that can express the same relation as the connective inserted in step 1 but are less ambiguous. The interface asks participants to drag and drop the connective between the relational arguments, such that it connects them. This works relatively well for English, as the insertion of a connective has little effect on the syntax of the relational arguments. This is however different for languages such as German and Czech, where the insertion of a connective can have various repercussions on word order. This in turn would lead to ungrammatical sequences for the insertion of many connectives, which in turn might mean that participants would choose these connectives less in order to avoid ungrammatical sequences. We therefore decided to adjust the interface and the method, as described below.

3.2. One-step connective insertion task

The adapted interface that we propose in this contribution puts less emphasis on choosing a connective that "fits" between the arguments, and more emphasis on choosing a connective that expresses the semantic relation that holds between the arguments, irrespective of whether syntax would need to be adjusted. The target location to which the connective should be drawn was placed below both arguments, such that the annotators can focus more on the meaning rather than the grammatical compatibility of the phrases, as shown in Figure 1. In order not to overwhelm participants with a large choice of connectives, we decided to only provide one connective per relation to choose from, and to arrange them into five boxes, corresponding to causal, temporal, comparison, positive expansion (e.g., `Conjunction`) and negative expansion (e.g., `Exception`) relation types.

The order of the five boxes as well as the order of connective options within the boxes were fixed for each participant, but randomized among the participants. This is to balance the potential preference of the phrases close to the answer box.

3.3. Validation of one-step task

To test whether the one-step method yields results that are comparable to the previously used two-step method, we invited the annotators who took part in the creation of DiscoGeM 1.0 to use the one-step method to annotate the same set of 18 items (i.e. one item constitutes the two arguments

[1]These relations are sampled from the PDTB corpus (Webber et al., 2019)

that need to be labelled) that they had annotated with the two-step method 1.5 years earlier. This enables a precise comparison of the methodology on the same group of annotators.

76 workers completed the task again with the one-step method. The majority labels obtained by both methods are the same for all of the 18 items in the selection task, except for one item, where the majority labels of the two methods correspond to the two alternative readings of the relation. This confirms the comparability of the two versions of crowdsourcing task.

3.4. Creation of connective lists

Following DiscoGeM 1.0, 28 relations types defined in PDTB 3.0 are distinguished, including `no relation`. We use the same set of relations for German, French and Czech. Applying the one-step annotation method in different languages necessitated the creation of sets of connectives (one per relation that should be distinguished) to be included as choices for the crowdworkers. The connectives need to be chosen such that they are as unambiguous as possible, but at the same time well understood by crowd workers. We based the choice of which connectives to include on the connective lexica that exist for Czech (Mírovskỳ et al., 2021), French (Roze et al., 2012) and German (Stede, 2002; Scheffler and Stede, 2016; Stede et al., 2018), and additionally consulted with native speakers. Table 1 presents the connective∼relation mapping for the most frequent relation types in DiscoGeM 2.0. The complete mapping of the 28 relation classes is shown in Table 3 in the Appendix.

To validate whether the workers connectives are indeed chosen by workers for the relations for which we selected them, we pretested the connectives on 10 workers in each language with 20 short examples of discourse relations, which were created to demonstrate either an unambiguous (e.g. *Fiona loves horror movies // She dislikes action movies.* `contrast`), or ambiguous (e.g. *Mary arrived the party // Owen left immediately.* `precedence`, `synchronous` or `result`) relation. The pretest results will be discussed in Sections 4.1, 5.1 and 6.1.

3.5. Data included in DiscoGeM 2.0

DiscoGeM 1.0 contains PDTB3-style annotations of inter-sentential implicit discourse relations in either original English text or translated English text. The included relations stem from different genres. In this work, we add multilingual annotations[2] to two of the genres in DiscoGeM 1.0, namely *Europarl* and *literature*. The included languages are English, German, French and Czech.

Table 2 presents details on the corpus size per genre. The other two genres included in DiscoGeM 1.0, *Wikipedia* and *WSJ*, are not included because only English versions of these texts were available.

The items of the **Europarl genre** in DiscoGeM 1.0 were taken from the Europarl corpus (Koehn, 2005; Cartoni and Meyer, 2012), which contains prepared speech of the European Parliament. We aligned the original English relations included in DiscoGeM 1.0 with their German and French translations[3], and we aligned DiscoGeM 1.0's translated English relations with their German, French or Czech originals.

The texts of the **literature genre** come from a total of twenty novels originally written in each of the four languages (5 novels per language). The list of novels is indicated in Table 4 in the Appendix. Novels consist of narrative writing, typically presented in a sequence of events, and therefore represent a different genre to formal spoken language in EuroParl. DiscoGeM 2.0 includes the German, French and Czech translations of the originally English relations from DiscoGeM 1.0, as well as the originals of the translated English relations.

Table 2 shows the number of annotated relations in DiscoGeM 2.0 in different translation directions. The Europarl and literature genres include parallel relations that are translated to or from English (but not, for example, translations from German to French or Czech). In total, DiscoGeM 2.0 consists of 12,834 relations.The number of relations in German, French and Czech is smaller than their English counterparts because a total of 420 translated English relations were not alignable to their original German, French or Czech counterparts and were thus excluded (see Section 3.6).

3.6. Data preparation

The German and French raw texts were sentence-splitted and tokenized using spaCy (Honnibal et al., 2020). Since spaCy doesn't include a model for Czech, the Czech texts were split into sentences based on punctuation (full stops, question marks and exclamation marks) and into tokens based on white spaces. Although the data from the Europarl corpus is originally sentence-aligned (using the classic Gale-Church algorithm, Gale et al.,

[2]That is, the items included in DiscoGeM 2.0 consist of the DE, FR, and CZ translations of the original EN relations included in DiscoGeM 1.0, or the original DE, FR, and CZ texts that correspond to the translated EN relations included in DiscoGeM 1.0.

[3]Czech translations of the original English relations in DiscoGeM 1.0 are, unexpectedly, not available in the Europarl corpus.

Relation sense		English	German	French	Czech
TEMPORAL	PRECEDENCE	then	dann	ensuite	potom
	SUCCESSION	after	davor,	après que	předtím
	SYNCHRONOUS	at the same time	gleichzeitig	en même temps	zároveň
CAUSAL	REASON	because	weil	parce que	protože
	RESULT	therefore	daher	c'est pourquoi	proto
COMPARISON	ARG2-AS-DENIER	nonetheless	trotzdem	néanmoins	přesto
	CONTRAST	on the other hand	andererseits	d'autre part	na druhou stranu
EXPANSION	CONJUNCTION	also	darüberhinaus	en plus	také
	ARG2-AS-INST.	for example	zum Beispiel	par exemple	například
	ARG2-AS-DETAIL	in more detail	genauer gesagt	plus précisement	konkrétně
NO RELATION		(no direct relation)	(keine direkte Beziehung)	(pas de relation directe)	(bez přímého vztahu)

Table 1: The connective mapping for the eight most frequent relation types in DiscoGeM 2.0. The full list is available in Table 3 in the Appendix

	Europarl			
orig. ↓ / data lang. →	EN	DE	FR	CS
English (EN)	418	**417**	**414**	—
German (DE)	701	**701**	—	—
French (FR)	739	—	**727**	—
Czech (CS)	700	—	—	**697**
Subtotal	2558	**1118**	**1141**	**697**
	Literature			
orig. ↓ / data lang. →	EN	DE	FR	CS
English (EN)	800	**787**	**758**	**777**
German (DE)	800	**683**	—	—
French (FR)	780	—	**729**	—
Czech (CS)	680	—	—	**526**
Subtotal	3060	**1470**	**1487**	**1303**
Total	5618	**2588**	**2628**	**2000**

Table 2: Breakdown of the 12,834 parallel implicit discourse relations in DiscoGeM 2.0. Numbers in bold are obtained as part of the DiscoGeM 2.0 expansion.

1994), we chose to re-align the sentences cross-lingually using the Vecalign method (Thompson and Koehn, 2019), which uses sentence embeddings from LASER (Artetxe and Schwenk, 2019) for performance reasons (F1 0.90 vs 0.72 for DE-FR alignments). We note that sentences and their translations often cannot be aligned in a one-to-one manner, as translations may change the order in which things are mentioned, or may choose to split information across sentences in a different way than the original text.

We excluded relations from the dataset in cases where the two arguments of the English relations were aligned to the same sentence in the other language (i.e., the relation is no longer inter-sentential); or, frequently occurring in the literature data, if the translation was not sentence-by-sentence (i.e., the two arguments of the English relation are not translated to or from two consecutive segments in the other language). In total, 20 ($< 1\%$) and 400 (13%) relations were not alignable in the Europarl and novel genres, respectively.

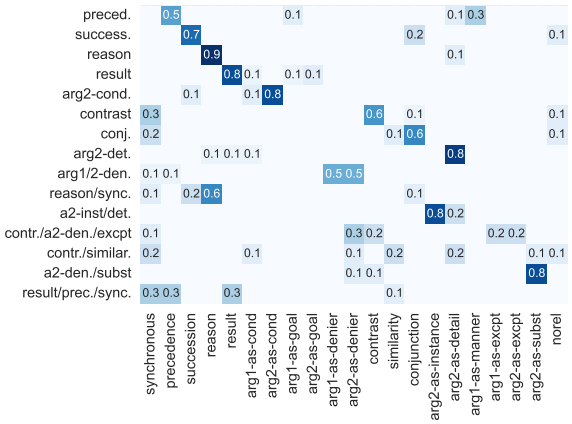
The relations that were included in DiscoGeM 1.0 were implicit, but it is possible that the aligned discourse relations in German, French or Czech include a connective (due to explicitation or implicitation during translation). However, an analysis of a subset of the data reveals that such cases are rare[4]. Moreover, implicit discourse relations can also be interpreted in the presence of explicit connectives (Rohde et al., 2018). Therefore, these relations are also annotated as part of our method.

3.7. Procedure

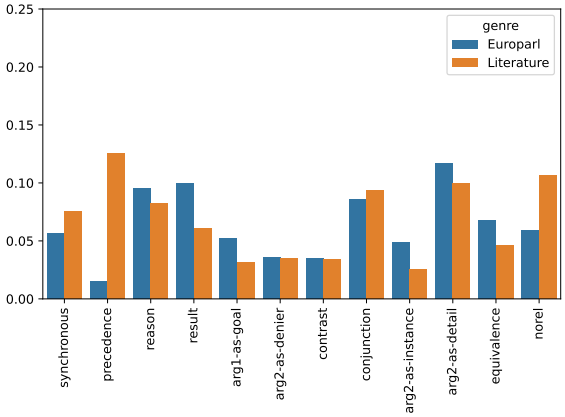
We recruited adult crowd workers registered on Prolific whose native language is either German, French or Czech (depending on the task they take part in). Following Scholman et al. (2022b), we first ran a selection task to evaluate the quality of the workers' annotations. This task also contained a feedback component to implicitly train the workers. All workers who scored more than 50% agreement with the gold labels on the selection task were included in our pool of final annotator candidates. From the selected workers, 90 German workers, 87 French workers and 37 Czech workers took part in the annotation. The number of Czech workers is smaller because the Czech speaker pool on the crowdsourcing platform was smaller.

The relations were presented to the annotators in batches of $\sim$20 items. The annotators were allowed to annotate more than one batch, but could only annotate each batch once. Participa-

[4]We analyzed 2691 argument pairs in the EN-DE data (including those that are not annotated for other reasons) using automatic word alignment (Dou and Neubig, 2021) and explicit connective identification (Bourgonje, 2021). There were only 60 cases (2.2%) in which the German inter-sentential connectives were not aligned to any words in the English text (explicitated or implicitated).



(a) Confusion matrix of the workers' labels (X-axis) vs. gold relations (Y-axis) in the validation experiment.



(b) Distribution for the most frequ. 12 labels in German.

Figure 2: Statistics of German subset.

tion ranged from 1 batch per participant to a total of 92 batches. In total, we collected data for 349 batches. Data collection took place in 2023 (average run time including selection: three weeks per language). Every batch was completed by at least 10 participants. Participants were awarded 1.8 British pounds for each batch of annotation.

4. German Subcorpus

4.1. Validation Experiment

Figure 2a compares the annotations by the German crowd workers (x-axis) to the gold labels of the validation samples (y-axis). Each row shows the normalized distribution of the workers' labels.[5]

The chosen labels largely converge on the intended senses of the unambiguous samples and spread over the multiple possible senses of the ambiguous samples. We however observe some variation across different items: for example, 90% of the workers agreed on the sense of the `reason` relation, while only 50% agreed on the sense of the `precedence` relation (1st row). For the ambiguous item textttresult/precedence/synchronous (last row), the workers' labels were evenly distributed among the different senses of the relation, while for another ambiguous relation, a preference towards `arg2-as-denier` was observed (2nd last row). While these annotations are consistent with expectations, we were surprised to see several `arg1-as-manner` for the `precedence` relation. We therefore inspected this item in more detail. The arguments of the `precedence` relation were *Sam hat neue Wandfarbe gekauft // Er hat das Wohnzimmer gestrichen. (Sam bought new paint. He painted the living room)*. The linking phrase for `arg1-as-manner` is *hiermit (thereby)*, which, when inserted into this item, takes on a non-connective meaning "*with that*", indicating an annotator mistake. Nevertheless, we decided to stick to *hiermit* as a marker for the `arg1-as-manner` relation, as alternative markers are similarly ambiguous. We therefore recommend to manually check `arg1-as-manner` annotations to check their validity. Also note that markers for the `synchronous` relation tend to be applicable to `contrast` relations in many languages, including German (connective *gleichzeitig (at the same time)*), see the first column of Figure 2a.

4.2. Relation distribution for German

To study to what extent the distribution of discourse relations depends on the genre, we turn to Fig. 2b. The distributions show the proportions of relation types per genre. The observed patterns echo the finding of DiscoGem 1.0 on the English relations: `precedence` relations are prevalent in literature, while causal relations such as `reason`, `result`, `condition` and `purpose` are more prevalent in Europarl. The difference reflects the contrast between the condensed logical reasoning in argumentative texts and the free narrative nature of literature. This genre effect can also be observed in the French and Czech subsets, as we see in the following sections.

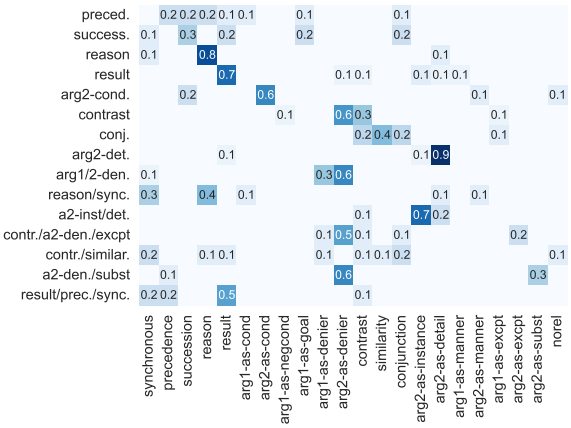
5. French Subcorpus

5.1. Validation Experiment

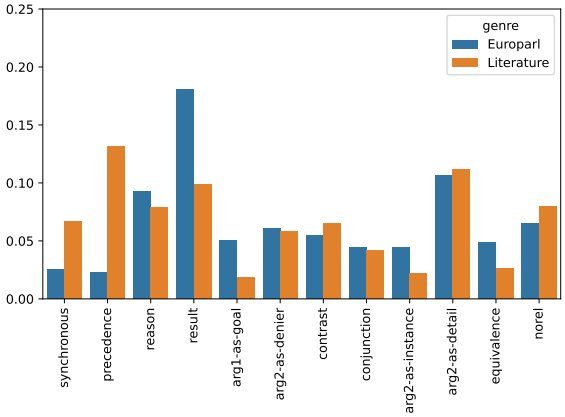
Next, we look at the co-occurrence of the French linking phrases in the validation experiment, as shown in Figure 3a. The workers' choices still overlap with the gold labels, but compared with German, the proportion of agreement is lower (i.e. there are fewer dark color cells in the matrix).

Specifically, the workers confused `precedence` and `succession` with a `result` or `purpose` rela-

[5]The numbers of some rows do not add up to one due to rounding.



(a) Confusion matrix of the workers' labels (X-axis) vs. gold relations (Y-axis) in the validation experiment.



(b) Distribution for the most frequent 12 labels in French.

Figure 3: Statistics of French subset.

tion instead. For example, the `succession` item was *Michael a mis sa tenue de bal // Il a pris une douche. (Michael put on his prom suit // He took a shower.)* In addition to the temporal sequence, the fact that he put on the suit could also be viewed as the reason or purpose for taking a shower (i.e. wanting to be clean when wearing a new suit). Causal relations have been hypothesized to be a default interpretation when people try to make sense of a discourse (Sanders, 2005).

5.2. Relation distribution for French

The resulting relation distribution for French shows that the method results in a preference for `result` relations (Figure 3b). The bias might be even stronger in the Europarl genre, in which causal relations are frequent and thus more expected. Similar frequent distributions of `result` are also found in the Czech Europarl subset but less in the German subset. To gain further insight into the difference between the German and French annotations, we inspected the distributions of the German Europarl items that align with French items with high proportion of `result` labels. We found that German annotators also chose a causal relation sense for these instances (specifically `arg1-as-goal`), as well as `conjunction` and `arg2-as-detail` (see Figure 5), which are more neutral in nature. The choice could depend on the preference of individual annotators as well as finer-grained differences between the connectives expressing these relation senses in German and French. The cross-lingual comparison of the annotations will be discussed in Section 7.

6. Czech Subcorpus

6.1. Validation Experiment

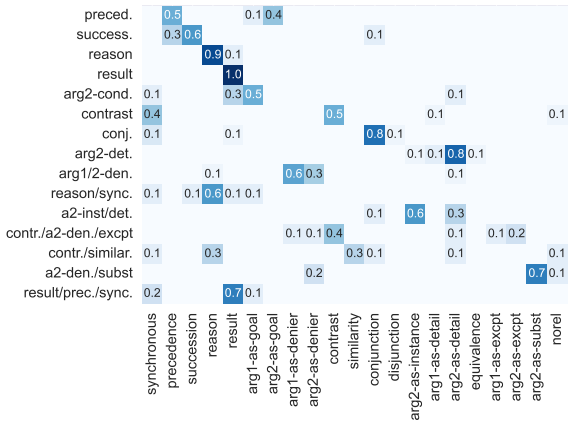
The results of the validation task in Czech in Figure 4a show that, apart from the common co-occurence of `synchronous` annotations with `contrast` relations (as already discussed in Section 4.1 for German), the annotators' choices converge nicely on the gold senses of the synthetic relation samples. This confirms that the chosen connectives reflect the workers' interpretation of discourse relations well.

6.2. Relation distribution for Czech

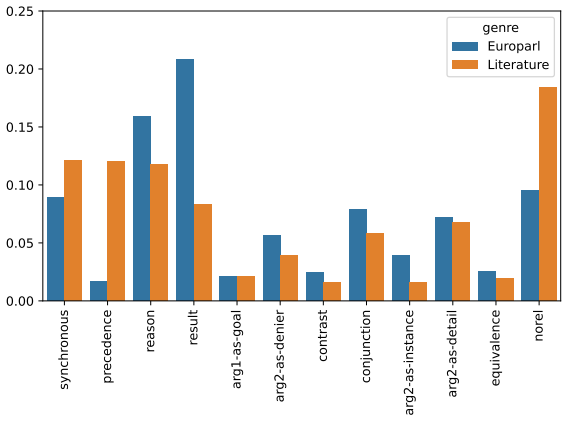
As explained in Section 3.7, fewer annotators took part in the annotation of the Czech items than in the annotation of the German and French items. This could be a possible explanation for the labels of the resulting annotations, as seen in Figure 4b, which are less evenly distributed (more distinct peaks) compared to German and French. For example, there is a high proportion of `no relation` labels which can actually be attributed to four workers who annotated many `no relation` labels with respect to the total number of items they annotated.

7. Discussion

Depending on the purpose of usage, the data in DiscoGem 2.0 can be used to treat individual bias in different ways. For example, the aggregated labels can be used as a reference interpretation that reflects the general interpretation of the language users. In DiscoGem 2.0, we provide labels aggregated by seven algorithms, including majority voting, D&S (Passonneau and Carpenter, 2014), MACE (Hovy et al., 2013), Wawa (Limited, 2023), MMSR (Ma and Olshevsky, 2020), ZBS (AI, 2023) and GLAD (Whitehill et al., 2009), which were obtained using the Crowd Kit tool (Ustalov et al., 2021). In addition, DiscoGeM 2.0 includes the annotator information of each label, which can



(a) Confusion matrix of the workers' labels (X-axis) vs. gold relations (Y-axis) in the validation experiment.



(b) Distribution for the most frequent 12 labels in Czech.

Figure 4: Statistics of the Czech subset.

be used to train models directly from the crowdsourced labels (Rodrigues and Pereira, 2018; Chu et al., 2021). The annotator information can also be used to develop models of individual bias. The goal of these models (such as Davani et al., 2022) is to predict the labels annotated by a specific annotator, taking into account their bias.

In this section, we provide an initial analysis of the cross-lingual annotation based on the aggregated and distributional labels of the different language versions of each item.

7.1. Comparing the aggregated labels

Figure 5 shows the confusion matrices of the labels annotated independently on the German, French and Czech versions of the discourse arguments. Each matrix reflects the statistics of the subset of items where annotations are available for both languages on the X and Y axes (see Table 2). The labels are aggregated by majority voting and combined on the second level of the PDTB sense hierarchy. The most frequent 8 level-2 labels are displayed.

The results displayed in Figure 5 indicate that the annotations generally agree cross-lingually; a darker diagonal line can be observed across all the matrices. Patterns of collocation or confusion of DR types can be identified in line with previous works, such as that between CONTRAST and CONCESSION (Robaldo and Miltsakaki, 2014; Yung et al., 2022) and between CAUSE and LEVEL-OF-DETAIL (Scholman and Demberg, 2017); these patterns do occur to different extents in different languages. In addition, there are some language-specific characteristics, such as the higher tendency to annotate CAUSE in French, as mentioned in Section 5, and the higher rate of confusion between SYNCHRONOUS and ASYNCHRONOUS in Czech. We plan to carry out more in-depth analysis in future work.

7.2. Comparing the label distributions

Next, we look at the extent to which the label distributions of items are similar between different languages. To fully explore the information contained in our crowdsourced data, we compare the item-wise label distribution across languages (see also Uma et al., 2021).

We use Jensen-Shannon Divergence (JSD) to quantify the difference between two label distributions. Lower JSD indicates higher similarity between the distributions. JSD equals 0 if the two distributions are exactly the same, and 1 if all votes belong to one label in one distribution and another label in the other distribution.

We compare three JSD values. 1) The *actual cross-lingual* JSD, which is calculated between the actual label distributions of each pair of aligned relations, such as English-German, English-French, or German-French. 2) The *expected cross-lingual* JSD, which is calculated between the label distributions of two languages after randomly shuffling the item order of each language separately. It represents the divergence we expect between the annotations of two languages given their overall label distributions. 3) The *lower-bound intra-lingual* JSD, which is calculated between two distributions sampled from the specific label distribution of that item and language version. It represents the expected variance of the labels annotated to the same language version of the same item.

The actual cross-lingual JSDs averaged across all items were found to be similar for all language pairs, ranging from 0.64 to 0.71, while the *expected cross-lingual* JSD over all items and language pairs was 0.83, which is higher. This suggests that annotators of different languages do agree more on the discourse interpretations of a specific item than what we expect between two randomly picked annotations from each language. On the other

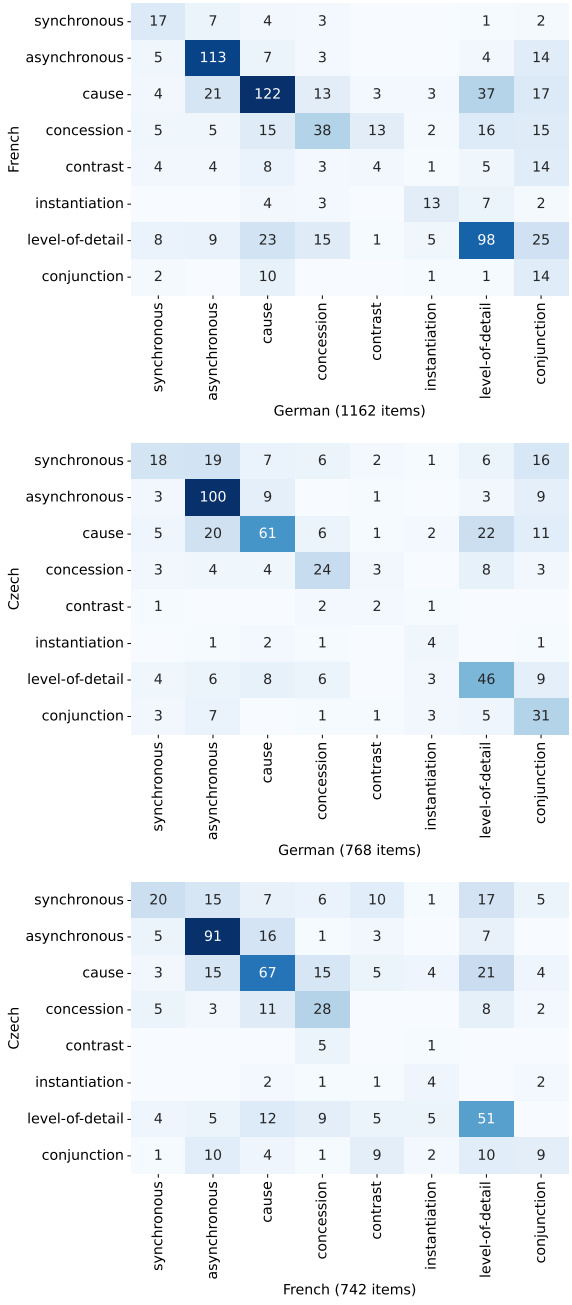


Figure 5: Comparison among the German, French and Czech majority labels of the corresponding subset of data where the annotation of the language pair is available (i.e. all are translations from English). The most frequent 8 level-2 relations are shown.

hand, the resulting expected *intra-lingual* JSD for all items and languages was 0.43 on average. This suggests that the annotation distributions deviate more across languages than they would have if annotated by two groups of workers of the same language.

The observed difference in relation label distributions across languages warrants further investigation in future work. It could stem from differences in translation, where the relational arguments from the source language get translated into the target language in a different order, or where some aspects are slightly changed such that a different (but possibly related) relation is inferred by readers of the target language.

Two examples of differences in cross-lingual annotation are shown in Figure 9 in the Appendix. In these examples, there is not any overlap between the labels annotated to the German original texts and the English translation. In Example 1, the English translation is not completely literal and may have affected the discourse interpretation. In Example 2, however, the English translation is semantically equivalent to the German original and yet different relations are interpreted. We plan to carry out analysis on more examples taking into account more global context.

Our corpus thus opens up new avenues for comparative research in discourse marking and translation, as well as providing new multi-domain resources for training discourse relation classifiers.

8. Conclusion

We presented DiscoGeM 2.0, a freely available multi-lingual parallel resource of implicit discourse relations in two genres. Each sample contains 10 labels that are crowdsourced with our proposed one-step annotation method specifically adapted for different languages. We assessed the applicability of the proposed method and analyzed potential methodological bias. The distributions of the resulting annotations confirmed the genre effects reported in previous works for English, and revealed that the interpretation of implicit discourse relations does not always agree across the original texts and the translations, suggesting that discourse annotations might not always be projectable in parallel texts.

In future works, we plan to use the distributional annotations of DiscoGeM 2.0 to analyze various sources of the multilingual difference. These may include the actual meaning shifts in the translation , especially in terms of discourse signals, a change due to connective meaning, or a difference in discourse processing of speakers of different mother tongues. We also plan to use the soft labels in DiscoGeM 2.0 to train, evaluate, and fine-tune implicit discourse relation classifiers for different languages and genres.

Ethics statement

The crowdsourced annotators were reimbursed equal to German minimum wage of 12 EUR per hour. The data collection was approved by the

Deutsche Gesellschaft für Sprachwissenschaft (DGfS) ethics committee.

Limitations

The parallel data we have used in the annotations has been pre-processed and aligned automatically, which might contain errors. Concerning the quality of the annotation, we have selected workers based on their performance in a selection task. However, their performance has not been further monitored. Finally, the implicit discourse relations to be annotated were based on the samples collected in DiscoGeM 1.0, which did not exhaustively include all relations in the original texts. The distributions thus might not fully represent those of the source materials.

Acknowledgements

We thank Philippine Geelhand for her advice on the selection of the French discourse connectives for the annotation task. We also thank the three anonymous reviewers for their valuable comments and suggestions. The research reported in this paper was supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074) and the Czech Science Foundation (project no. 24-11132S, Disagreement in Corpus Annotation and Variation in Human Understanding of Text); a part of the used data comes from the project no. LM2018101 by the Czech Ministry of Education, Youth and Sports (Digital Research Infrastructure for Language Technologies, Arts and Humanities).

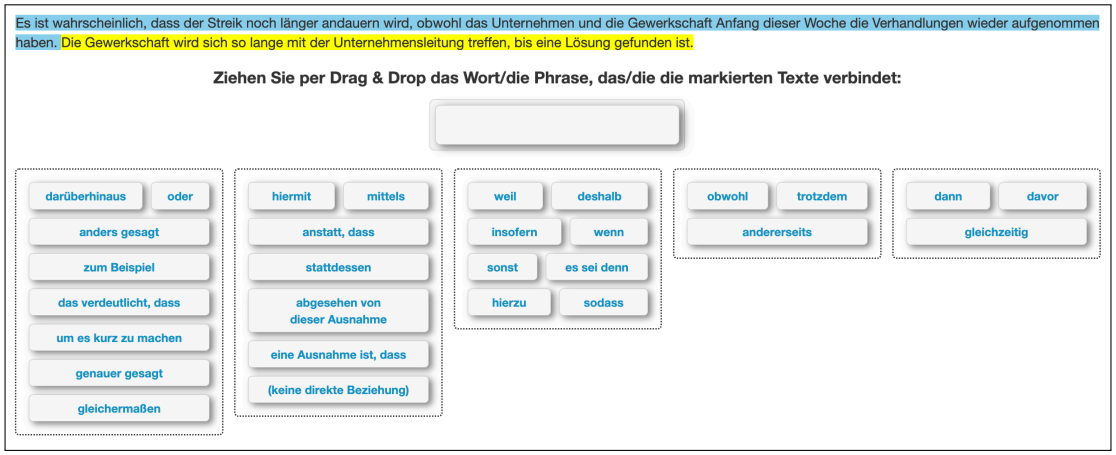


Figure 6: The German crowdsourcing interface

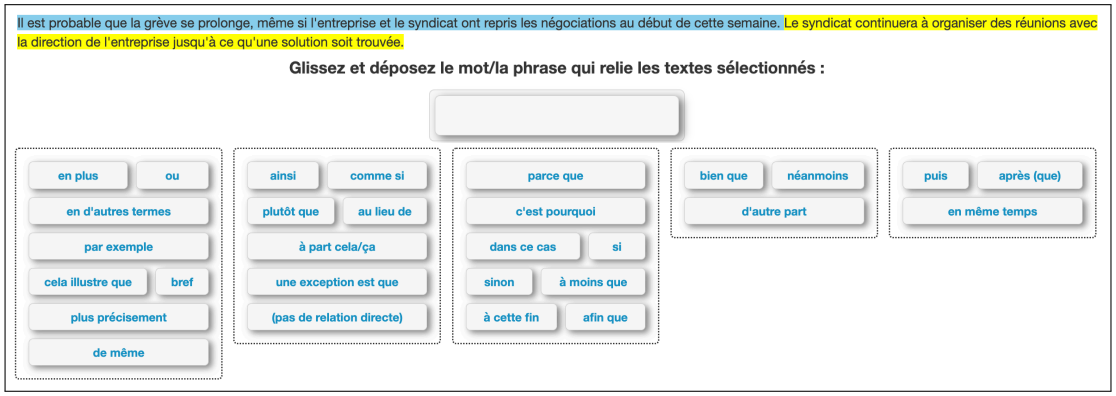


Figure 7: The French crowdsourcing interface

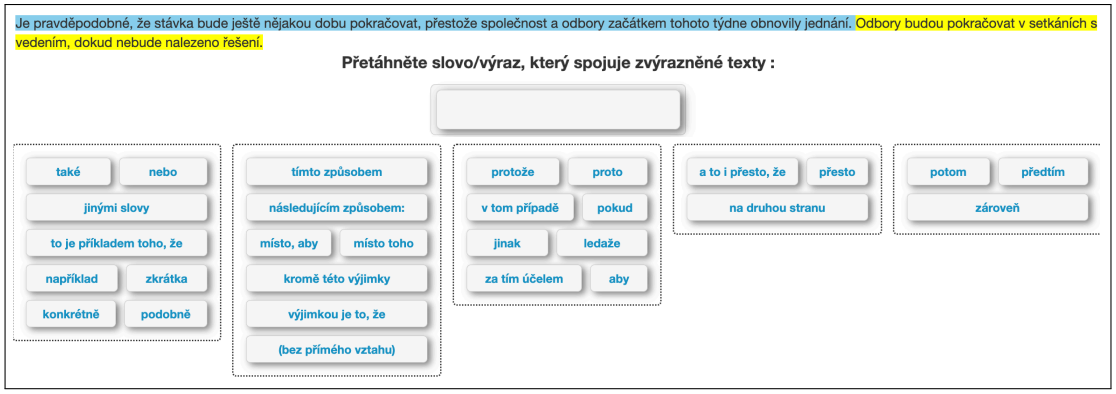


Figure 8: The Czech crowdsourcing interface

Relation sense	English	German	French	Czech
TEMPORAL				
PRECEDENCE	then	dann	ensuite	potom
SUCCESSION	after	davor,	après que	předtím
SYNCHRONOUS	at the same time	gleichzeitig	en même temps	zároveň
CONTINGENCY				
SIMILARITY	similarly	gleichermaßen	de même	podobně
REASON	because	weil	parce que	protože
RESULT	as a result	daher	c'est pourquoi	proto
ARG1-AS-COND	in that case	insofern	dans ce cas	v tom případě
ARG1-AS-NEGC	if not	sonst	sinon	jinak
ARG2-AS-COND	if	wenn	si	pokud
ARG2-AS-NEGC	unless	es sei denn	à moins que	ledaže
ARG1-AS-GOAL	for that purpose	dazu	à cette fin	za tím účelem
ARG2-AS-GOAL	so that	sodass	afin que	aby
COMPARISON				
ARG1-AS-DENIER	even though	obwohl	bien que	a to i přesto, že
ARG2-AS-DENIER	nonetheless	trotzdem	néanmoins	přesto
CONTRAST	on the other hand	andererseits	d'autre part	na druhou stranu
EXPANSION				
CONJUNCTION	also	darüberhinaus	en plus	také
DISJUNCTION	or	oder	ou	nebo
EQUIVALENCE	in other words	anders gesagt	en d'autres termes	jinými slovy
ARG1-AS-EXCPT	other than that	abgesehen von dieser Ausnahme	à part cela/ça	kromě této výjimky
ARG2-AS-EXCPT	an exception is that	eine Ausnahme ist, dass	une exception est que	výjimkou je to, že
ARG1-AS-INST	this illustrates that	das verdeutlicht, dass	cela illustre que	to je příkladem toho, že
ARG2-AS-INST	for example	zum Beispiel	par exemple	například
ARG1-AS-DETAIL	in short	um es kurz zu machen	bref	zkrátka
ARG2-AS-DETAIL	in more detail	genauer gesagt	plus précisement	konkrétně
ARG1-AS-MANNER	thereby	hiermit	de cette manière	tímto způsobem
ARG2-AS-MANNER	as if	mittels	comme si způsobem:	následujícím
ARG1-AS-SUBST	rather than	anstatt, dass	plutôt que	místo, aby
ARG2-AS-SUBST	instead	stattdessen	au lieu de	místo toho
NO RELATION	(no direct relations)	(keine direkte Beziehung)	(pas de relation directe)	(bez přímého vztahu)

Table 3: The choices of linking phrases provided to the crowd workers for all relation types in DiscoGeM 2.0.

original: EN translation: DE,FR,CS
Aninal Farm
Great Gatsby
Harry Potter and the Philosopher's Stone
The Hitchhiker's Guide to the Galaxy
The Hobbit
original: DE translation: EN
Perfume
The Clown
The Glass Beads
The Magic Mountain
The Tin Drum
original: FR translation: EN
Arsene Lupin
In Search of Lost Time
Phantom of the Opera
The Reunion
The Stranger
original: CS translation: EN
Good Soldier Svejik
Love and Garbage
The Power of the Powerless
The Unbearable Lightness of Being
War with the Newts

Table 4: List of books and the language versions included in the literature genre

EXAMPLE 1:
Original German text: Du sollst aber nie vergessen, was ich dir so oft gesagt habe: unsere Bestimmung ist, die Gegensätze richtig zu erkennen, erstens nämlich als Gegensätze, dann aber als die Pole einer Einheit. // So ist es auch mit dem Glasperlenspiel.

Translation by Deep Translate: *But you should never forget what I have told you so often : our destiny is to recognize the contrasts correctly, first of all as contrasts, but then as the poles of a unity. // So it is with the Glass Bead Game.*

Translated English text: But never forget what I have told you so often: our mission is to recognize contraries for what they are: first of all as contraries, but the opposite poles of a unity. // Such is the nature of the Glass Bead Game.

- **Annotated labels on German**:
  SIMILARITY (5), REASON (2), EQUIVALENCE (2),CONTRAST (1)
- **Annotated labels on English**:
  ARG1-AS-DETAIL (6), RESULT (3), CONJUNCTION (1)

EXAMPLE 2:
Original German text: Ich hatte sie noch nie mit Hut gesehen, sie hatte sich immer geweigert, einen aufzusetzen. Der Hut veränderte sie sehr. // Sie sah wie eine junge Frau aus. Ich dachte, sie mache einen Ausflug, obwohl es eine merkwürdige Zeit für Ausflüge war.

Translation by Deep Translate: *I had never seen her in a hat before, she had always refused to wear one. The hat changed her a lot. // She looked like a young woman. I thought she was going on an outing, although it was a strange time for outings.*

English translated text: I had never seen her in a hat before, she had always refused to wear one. The hat altered her very much. // She looked like a young woman. I thought she must be going on an outing, though it was a strange time for outings. But in those days the schools were capable of anything.

- **Annotated labels on German**:
  REASON (5), EQUIVALENCE (2), ARG1-AS-DETAIL (1), ARG2-AS-GOAL (1), NO RELATION (1)
- **Annotated labels on English**:
  RESULT (7), CONTRAST (1), ARG2-AS-INSTANCE (1), PRECEDENCE (1)

Figure 9: Examples of annotation difference between the original and translated text

Bibliographical References

Toloka AI. 2023. Zero-based skill. https://crowd-kit.readthedocs.io/en/latest/classification/#crowdkit.aggregation.classification.ZeroBasedSkill. Accessed: 2024-03-01.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. The change that matters in discourse parsing: Estimating the impact of domain shift on parser error. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845.

Peter Bourgonje. 2021. *Shallow discourse parsing for German*, volume 351. IOS Press.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Bruno Cartoni and Thomas Meyer. 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.

Zhendong Chu, Jing Ma, and Hongning Wang. 2021. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5832–5840.

S Cortez and Cassandra L Jacobs. 2023a. Incorporating annotator uncertainty into representations of discourse relations. *Proceedings of the the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

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

Ludivine Crible, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, and Šárka Zikánová. 2019. Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

William A. Gale, Kenneth Ward Church, et al. 1994. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

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 (NAACL-HLT)*, pages 1120–1130, Denver, CO.

René Knaebel and Manfred Stede. 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. Association for Computational Linguistics.

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

Majid Laali and Leila Kosseim. 2017. Improving discourse relation projection to build discourse annotated corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416.

Appen Limited. 2023. Worker agreement with aggregate. https://success.appen.com/hc/en-us/articles/202703205. Accessed: 2024-03-01.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

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

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021b. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Qianqian Ma and Alex Olshevsky. 2020. Adversarial crowdsourcing through robust rank-one matrix completion. *Advances in Neural Information Processing Systems*, 33:21841–21852.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of the SIGDIAL 2011 Conference*, pages 194–203.

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

Emily Pitler and Ani Nenkova. 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. Association for Computational Linguistics.

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

Valentina Pyatkin, Frances Yung, Merel C. J. 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.

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

Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267.

Dongyu Ru, Lin Qiu, Xipeng Qiu, Yue Zhang, and Zheng Zhang. 2023. Distributed marker representation for ambiguous discourse markers and entangled relations. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Ted Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning*, pages 105–114. University of Toulouse-le-Mirail Toulouse.

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

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of german. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1008–1013.

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

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

Merel C. J. Scholman, Valentina Pyatkin, Frances Yung, Ido Dagan, Reut Tsarfaty, and Vera Demberg. 2022b. 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).

Merel C.J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 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)*.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1044–1050, Marseille, France. European Language Resources Association.

Manfred Stede. 2002. Dimlex: a lexical approach to discourse markers.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2018. Connective-Lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

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

Dmitry Ustalov, Nikita Pavlichenko, Vladimir Losev, Iulian Giliazev, and Evgeny Tulin. 2021. A general-purpose crowdsourcing computational quality control toolkit for python. In *The Ninth AAAI Conference on Human Computation and Crowdsourcing: Works-in-Progress and Demonstration Track (HCOMP 2021)*.

Bang Wang, Zhenglin Wang, Wei Xiang, and Yijun Mo. 2023. Adaptive prompt learning with distilled connective knowledge for implicit discourse relation recognition. *arXiv preprint arXiv:2309.07561*.

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

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22.

Wei Xiang, Chao Liang, and Bang Wang. 2023. Teprompt: Task enlightenment prompt learning for implicit discourse relation recognition. *Findings of the Annual Meeting of the Association for Computational Linguistics*.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. Connprompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911.

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.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation

annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.

Frances Yung, Merel C. J. Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Vera Demberg. 2023. Investigating explicitation of discourse connectives in translation using automatic annotations. In *Proceedings of the 24th Annual SIGdial Meeting on Discourse and Dialogue*.

Deniz Zeyrek, Amália Mendes, Giedrė Valūnaitė Oleškevičienė, and Sibel Özer. 2022. An exploratory analysis of ted talks in english and lithuanian, portuguese and turkish translations: Results from the analysis of an annotated multilingual corpus. *Contrastive Pragmatics*, 3(3):452–479.

Šárka Zikánová, Jiří Mírovskỳ, and Pavlína Synková. 2019. Explicit and implicit discourse relations in the Prague Discourse Treebank. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, pages 236–248. Springer.

Language Resource References

Afantenos, Stergos and Asher, Nicholas and Benamara, Farah and Bras, Myriam and Fabre, Cécile and Ho-Dac, Lydia-Mai and Le Draoulec, Anne and Muller, Philippe and Péry-Woodley, Marie-Paule and Prévot, Laurent and others. 2012. *ANNODIS Resource, A discourse-level annotated corpus for French*. Laboratoire Cognition, Langues, Langage, Ergonomie (CLLE) UMR 5263. PID http://redac.univ-tlse2.fr/corpus/annodis/annodis_en.html.

Bourgonje, Peter and Stede, Manfred. 2020. *The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing*. PID http://angcl.ling.uni-potsdam.de/resources/pcc.html.

Mírovskỳ, Jiří and Synková, Pavlína and Rysová, Magdaléna and Poláková, Lucie. 2021. *CzeDLex 1.0*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL). PID http://hdl.handle.net/11234/1-4595.

Roze, Charlotte and Danlos, Laurence and Muller, Philippe. 2012. *LEXCONN: a French lexicon of discourse connectives*. PID http://www.linguist.univ-paris-diderot.fr/ croze/D/Lexconn.xml.

Scholman, Merel C. J. and Dong, Tianai and Yung, Frances and Demberg, Vera. 2022. *DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations*. PID https://github.com/merelscholman/DiscoGeM.

Synková, Pavlína and Rysová, Magdaléna and Mírovský, Jiří and Poláková, Lucie and Sheller, Veronika and Zdeňková, Jana and Zikánová, Šárka and Hajičová, Eva. 2022. *Prague Discourse Treebank 3.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zeyrek, Deniz and Mendes, Amália and Grishina, Yulia and Kurfalı, Murathan and Gibbon, Samuel and Ogrodniczuk, Maciej. 2020. *TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style*. PID https://github.com/MurathanKurfali/Ted-MDB-Annotations.