Multi-token Mask-filling and Implicit Discourse Relations

Meinan Liu[1] Yunfang Dong[2] Xixian Liao[3] Bonnie Webber[1]
[1]School of Informatics, University of Edinburgh
[2]School of Engineering, Westlake University
[3]Barcelona Supercomputing Center
liumeinan99@outlook.com, dongyunfang@westlake.edu.cn,
xixian.liao@bsc.es, bonnie.webber@ed.ac.uk

Abstract

Previous work has shown that simple mask-filling can provide useful information about the discourse informativeness of syntactic structures. Dong et al. (2024) first adopted this approach to investigating preposing constructions. The problem with single token mask fillers was that they were, by and large, ambiguous. We address the issue by adapting the approach of Kalinsky et al. (2023) to support the prediction of multi-token connectives in masked positions. Our first experiment demonstrates that this multi-token mask-filling approach substantially outperforms the previously considered single-token approach in recognizing implicit discourse relations. Our second experiment corroborates previous findings, providing additional empirical support for the role of preposed syntactic constituents in signaling discourse coherence. Overall, our study extends existing mask-filling methods to a new discourse-level task and reinforces the linguistic hypothesis concerning the discourse informativeness of preposed structures.

1 Introduction

Previous work has shown that simple mask-filling can provide useful information about discourse relations—in particular, about whether a preposed syntactic construction (e.g., a prepositional phrase moved from its canonical post-verbal position to the sentence-initial position) in one sentence can help identify its sense relation to its immediately preceding sentence (Dong et al., 2024).

The discourse relations considered earlier were those that hold between adjacent English sentences that lack an explicit discourse connective providing information about how they are related (we call them *implicit inter-sentential relations* hereafter). Since the previous work used BERT (Devlin et al., 2019), which is limited to predicting one token per [MASK], the mask fillers considered were restricted to single-token explicit discourse connectives such as *but*, *so*, *however*, etc. As shown in Ex. (1), the connective *but* is a BERT-predicted mask filler inserted before the preposed prepositional phrase (PP; shown in bold) and was not present in the original sentence. In this case, *but* can signal a *Comparison* relation between the two discourse arguments.

(1) *The paper reflected the truth.*$_{\text{Arg1}}$ [inserted: but] **For the leadership**$_{\textbf{PP}}$, that was too painful to bear.$_{\text{Arg2}}$ [wsj_1603, PDTB-3]

A limitation of restricting connectives to such single-token forms is that it can introduce ambiguity, as a single-token connective may convey various discourse relations, also known as senses. For example, *but* can convey up to 8 different relations, according to the Penn Discourse Treebank 3.0 Annotation Manual (Webber et al., 2019), whereas multi-token connectives such as *in contrast* is only mapped to a single sense (see more in Section 3), making them much less ambiguous.

To address this limitation, we adapt the Extended-Matrix decoder approach of Kalinsky et al. (2023) to enable masked language models (MLMs) to predict multi-token connectives as mask fillers. Our contributions are twofold: (1) we present a system that directly generates multi-token discourse connectives in masked positions to help identify the discourse relation between text segments, and (2) we provide new empirical evidence supporting the role of preposed syntactic structures in signaling discourse relations. These results not only extend prior work but also demonstrate how language models can be used to evaluate and inform linguistic theories.

Specifically, we conduct two experiments. The first evaluates our multi-token completion approach against single-token mask filling on all implicit inter-sentential relations in the Penn Discourse TreeBank (Webber et al., 2019). Results show that multi-token completion shows better performance

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12546–12560
November 4-9, 2025

in discourse relation recognition. In the second experiment, we implement our multi-token completion model on a subset of implicit inter-sentential relations where the second argument starts with a preposed syntactic structure. With less ambiguous fillers, our results provide further evidence for the discourse informativeness of preposed syntactic constituents.

2 Background

2.1 Discourse relation recognition

Discourse relations specify relationships that hold between text segments. As shown in Ex. (2), a discourse relation can be marked with an explicit connective such as "but", which can serve as strong (albeit ambiguous) signals of the senses that hold between segments (Pitler and Nenkova, 2009). Linguistic resources such as the web-based multilingual Connective-Lex (Stede et al., 2019) and the Penn Discourse Treebank (PDTB-3) (Prasad et al., 2019) provide lists of these connectives. However, in both spoken and written communication, it is often the case that no connective is explicitly provided, as in Ex. (3). These are called implicit discourse relations, and there can be several possible relations linking the two segments. Yet listeners or readers can easily infer the relation between two segments of text.

(2) John left **but**$_{\textbf{explicit connective}}$ Bob stayed.

(3) John left, **[but/so/because]**$_{\textbf{implicit connective}}$ Bob stayed.

Early research on discourse relation classification employed basic machine learning strategies like Naïve Bayes, which required hand-crafted features (Xiang and Wang, 2023). However, such feature engineering requires costly, time-consuming expert knowledge, with the possibility that relevant features might not have been noticed. Recently, research has increasingly turned to neural networks or deep learning methods for relation classification. Input to these methods consists of word embeddings—numerical representations of the linguistic information of a token and its context. Using these methods, some studies focus on directly classifying implicit discourse relations (Qin et al., 2016), while other studies, recognizing the significance of discourse connectives in signaling discourse relations, leverage discourse connectives for sense classification (Xu et al., 2012; Qin et al., 2017; Shi and Demberg, 2019; Kishimoto et al., 2020).

2.2 Syntactic preposing

Das and Taboada (2019) show that discourse relations can be signaled by a variety of cues. One such cue, discussed in Ward and Birner (2006), is non-canonical syntactic structures. This hypothesis was validated by Dong et al. (2024), who found that an MLM more often chooses as mask-filler, a discourse connective that could express the manually-annotated sense when the second text span in the relation (Arg2) starts with a preposed constituent, compared to when that preposed constituent is moved rightward to its canonical position within the sentence. Dong et al. (2024) only considered two types of preposed constituents: prepositional phrases (PP) and noun phrases (NP), as in Ex. (4) and Ex. (5), respectively.

(4) *The paper reflected the truth.*$_{\text{Arg1}}$ **For the leadership**$_{\textbf{PP}}$, that was too painful to bear.$_{\text{Arg2}}$ [wsj_1603, PDTB-3]

(5) *Just days after the 1987 crash, major brokerage firms rushed out ads to calm investors.*$_{\text{Arg1}}$ **This time around**$_{\textbf{NP}}$, they're moving even faster.$_{\text{Arg2}}$ [wsj_2201, PDTB-3]

The experiments in Dong et al. (2024) involved two sets of discourse relations: a **preposed set**, where Arg2 of the relation starts with a non-subject NP or PP, as in Ex. (4) above, and a **canonical set**, where the NP/PP is right-moved to the end of the first main clause in Arg2 to create a canonical sentence structure as shown in Ex. (6) below (Dong et al., 2024).

(6) *The paper reflected the truth.*$_{\text{Arg1}}$ That was too painful to bear **for the leadership**.$_{\text{Arg2}}$

(7) *The paper reflected the truth.*$_{\text{Arg1}}$ [MASK], that was too painful to bear **for the leadership**.$_{\text{Arg2}}$

A [MASK] token is inserted at the beginning of Arg2 during data preprocessing, as shown in Ex. (7), and the model is tasked with predicting a single token to fill the mask. Of interest was when the mask was filled with a single-token discourse connective such as "but" or "so". The results in Dong et al. (2024) showed that BERT's mask-filling was better on the preposed set than the canonical set, making this study the first to

empirically validate that preposing can help signal discourse relations.

2.3 Multi-token mask-filling

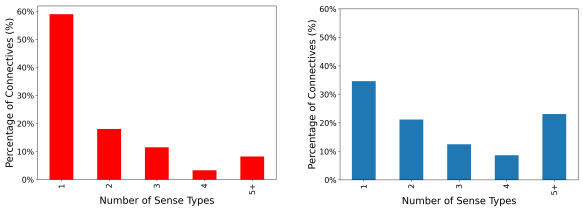
In general, it is not enough to allow only single-token mask-fillers, since so many relevant phrases involve multiple tokens. To address this, Joshi et al. (2020) introduced a sequence of contiguous [MASK] tokens to predict a text span. However, fixing the number of [MASK] tokens is not really suitable for predicting discourse connectives, as it fixes the possible fillers before prediction. Another approach (Raffel et al., 2020) treats multi-word phrases as sequences of single tokens during preprocessing. However, this approach increases tokenization time and requires the MLM's entire weight matrix be adapted to the new input.

A third approach to enabling an MLM to predict multi-token fillers is the EMAT (Extended-Matrix) decoder (Kalinsky et al., 2023), originally developed for question-answering. An EMAT decoder extends an MLM's output vocabulary to include multi-token phrases by augmenting the model's decoding matrix (the output projection layer) with new embedding vectors representing these phrases. Unlike previous methods, this approach does not require updating the entire pre-trained model. Rather, it assigns embedding vectors to newly added phrases and trains only the associated parameters, largely reducing computational costs. In the following section, we describe how EMAT was adapted to predicting multi-token discourse connectives as mask-fillers.

It is worth noting that Liu and Strube (2023) employs similar approach for generating multi-token connectives. However, key difference exists in the motivation of our studies. Specifically, their work focuses solely on implicit discourse relation classification, without addressing the distinct roles of multi- and single-token connectives in conveying discourse relations, particularly the ambiguity of connective senses, which is a central concern of our study.

3 Ambiguity in single- vs. multi-token connectives

Before presenting our main experiments, we begin by providing empirical support for the observation that multi-token connectives tend to be less ambiguous than single-token ones. To this end, we extracted all inter-sentential single-token and multi-token connectives from the PDTB-3, and for each of them, we also extracted its sense relations and the count of each relation. There are 104 types of single-token connectives and 61 types of multi-token connectives in the PDTB-3. As shown in Figure 1, around 60% of multi-token connectives are unambiguous (i.e., are mapped to only one sense relation), compared to just 35% of single-token connectives. Among those associated with multiple senses, over 20% of single-token connectives are mapped to more than 5 senses, versus less than 10% for multi-token connectives.



(a) Distribution of sense counts per multi-token connective.

(b) Distribution of sense counts per single-token connective

Figure 1: Sense distribution: single token connectives vs. multi-token connectives.

While the number of senses associated with a connective reflects its potential ambiguity, it does not capture how these senses are distributed. Some connectives may have multiple possible senses but are strongly associated with only one or two in practice. To capture this variability, we compute the entropy of the sense distribution for each connective based on annotated frequencies in PDTB-3, where **lower values indicate less ambiguity**.

To summarize connective-level ambiguity across the dataset, we report the average entropy. As shown in Table 1, the average entropy of sense type distribution for multi-token connectives (0.3374) is considerably less than that for single-token connectives (0.5661), further supporting that multi-token connectives are less ambiguous.

# of sense types	Entropy	
	Multi-token	Single token
All senses	**0.3374**	0.5661

Table 1: Average entropy of sense type distributions for multi-token vs. single-token connectives, computed over all explicit and implicit relations in PDTB-3. Full details of the calculation are provided in Appendix A.

4 Multi-token mask-filling model

Sections 4 and 5 provide the experimental setup for the experiments discussed in Sections 6 and 7.

4.1 Model architecture

Figure 2 shows the EMAT-based architecture of the model used to predict mask-fillers for each argument pair.

We first used the MLM encoder from BERT (Devlin et al., 2019) to obtain contextual embeddings for each of our multi-token connectives (see Section 4.2.1), which were then fed into the EMAT decoder. We trained the EMAT decoder on the entire training dataset, and mapped all word vectors, including the embeddings of new phrases, to the output prediction matrix. Note that the new vectors were added only to the prediction matrix, not the base model vocabulary, so BERT did not need to be retrained. During inference, each formatted input was fed into the model. The MLM encoder computed the contextual embedding for the masked token, and the EMAT decoder generated predictions along with their probabilities. A prediction corresponding to a multi-token connective from our pre-defined vocabulary was then mapped to its associated senses and compared to the gold sense. For instance, in Figure 2, the sense *Comparison.Contrast* can be signaled by the connective "In fact" (highlighted in red), indicating a match with the gold sense.

4.2 Dataset curation

Since EMAT was originally trained to recognize named entities (Kalinsky et al., 2023), it had to be adapted for multi-token discourse connectives. In particular, while a multi-token named entity is typically located within a single sentence, multi-token discourse connectives require consideration of a pair of adjacent sentences.

4.2.1 Multi-token discourse connectives

We collected a total of 69 multi-token connectives from two sources: (i) Appendices A and C of the PDTB-3 Annotation Manual (Webber et al., 2019), and (ii) the English Connective-Lex (Stede et al., 2019). From this set, we excluded subordinating conjunctions (e.g., *so that*), since they express relations between clauses within a single sentence (intra-sentential), as well as multi-token connectives listed in Connective-Lex (Stede et al., 2019) that lack a Level 3 sense annotation. For instance, some connectives under *Expansion.Substitution* do not specify whether the substitution occurs in Arg1 or Arg2 (e.g., *Arg1-as-subst* vs *Arg2-as-subst*), and were thus excluded due to incomplete labeling. The final list of connectives was added to the model's output vocabulary for use in EMAT.

Following Dong et al. (2024), we mapped each connective to all its associated senses (see Appendix F) and considered a mask-filler correct if it can signal the human-annotated sense of a relation. For example, according to the PDTB-3 Annotation Manual (Webber et al., 2019), "by contrast" is associated with both *Comparison.Concession.Arg2-as-denier* and *Comparison.Contrast*. It will therefore be treated as a correct mask-filler if the human-annotated sense is *Comparison.Contrast*, even if the connective inserted by the annotator was "by comparison" or "in contrast."

4.2.2 Training and development datasets

To train the model on multi-token connectives and to fill the [MASK] token with them, we extracted ~838K sentence-pairs from the Wikipedia English dataset (20220301.en) available on Huggingface (Foundation, 2024).[1] In these pairs, the second sentence (Arg2)[2] starts with a multi-token connective.

We replace the multi-token connective in each Arg2 with a [MASK] token. For instance, "By contrast, Bob left." is transformed to "[MASK], Bob left." Combined with the first sentence (Arg1), this forms an input tuple: (Arg1, masked Arg2), as illustrated in Figure 2. Special tokens (e.g., the sentence separator [SEP]) are automatically added during tokenization.

The dataset is randomly split into training (80%) and development (20%) sets, with ~671K and ~168K samples respectively.

Although neither training nor development data are annotated with discourse relations, the test set uses annotated discourse relations, in order to examine whether the predicted connectives can signal the annotated senses (see Sections 6.1 and 7.1).

4.3 Model training

We trained the EMAT decoder for 5 epochs on the full training dataset. The model has approximately 22.9 million trainable parameters. Training and inference were conducted on a single NVIDIA L4 GPU (24GB memory).

[1] https://huggingface.co/datasets/legacy-datasets/wikipedia

[2] Note that the arguments collected here are sentences, not the *Argument* whose strict definition is that it is a text segment with at least a predicate.

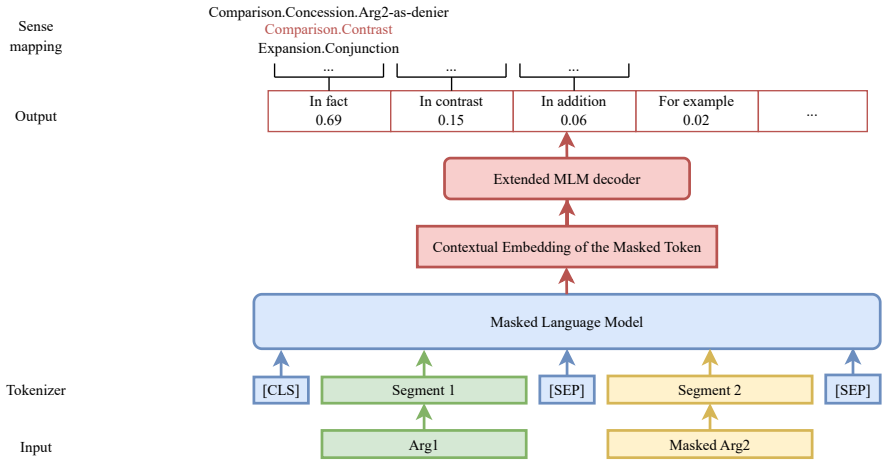


Figure 2: Illustration of mask-filling and mapping between predicted fillers and senses

5 Evaluation metrics

Accuracy and precision As mentioned in Sec 4.1, we consider a predicted connective correct if our sense-mapping dictionary indicates that it can signal the human-annotated sense. Following Dong et al. (2024), we first computed both accuracy and precision for the model's top N predictions. The model's average accuracy for the top N predictions in a dataset, denoted as $a@N$, is calculated using the equation:

$$a@N = \frac{1}{k}\sum_{i=1}^{k} \max_{x \in \text{pred}_i^N} \left(\mathbb{1}_{\{sense(x)=gold_i\}}\right), \quad (1)$$

where the model's top N predictions for sample i are represented as pred_i^N. A prediction x for sample i is considered correct if it can convey the gold sense $gold_i$ as per the sense-mapping dictionary $sense(x)$.

If any of the top N predictions is correct, the entire prediction for sample i is deemed correct. We compute dataset accuracy by averaging over its k samples.

Precision measures the proportion of correct predictions relative to the total number of predictions. $p@N$ refers to the precision within the model's top N predictions. For instance, if the top 2 predictions are correct, $p@2$ is 100%; if only one of them is correct, then $p@2$=50%. Average precision of the model's top N predictions is computed as:

$$p@N = \frac{1}{k}\sum_{i=1}^{k} \frac{\sum_{x \in \text{pred}_i^N} \mathbb{1}_{\{sense(x)=gold_i\}}}{N} \quad (2)$$

where items are similarly denoted as in Eq.(1), for a@N. For each sample i we count how many predictions are correct out of the top N predictions $pred_i^N$ and divide this by N. This calculation is averaged over all k samples.

Surprisal and entropy In addition to assessing how correct our model is, we also want to know how confident it is, to understand whether preposing influences model uncertainty. To quantify uncertainty, we use two information-theoretic measures: surprisal and entropy (Shannon, 1948). Surprisal measures how unexpected the model finds the human-annotated sense. A low surprisal value indicates that the model assigns high probabilities to connectives that convey the gold sense, suggesting confidence in its correct predictions. The average surprisal is calculated from the model's probability distribution over the entire vocabulary and is defined as the averaged sum of the negative log-likelihood (NLL) across all samples in the dataset:

$$\overline{NLL} = -\frac{1}{k}\sum_{i=1}^{k}\sum_{x \in V} \log p_i(x) \cdot \mathbb{1}_{\{sense(x)=gold_i\}} \quad (3)$$

where x denotes a lexical entry within the vocabulary V, and the summation extends over all the k samples in the dataset.[3] A lexical entry x is considered correct if it is a connective and can convey the annotated implicit discourse relation.

Entropy measures a model's certainty over all its predictions, regardless of their correctness. It re-

[3]Here, we do not restrict the predictions to top N, but instead use predictions over the entire vocabulary to compute surprisal and entropy.

flects how spread out the probability distribution is over the entire vocabulary (i.e. all the possible predictions). A higher entropy value suggests a more dispersed or even probability distribution, indicating greater uncertainty. Conversely, a lower entropy value indicates a more concentrated distribution, where top predictions (not necessarily correct) have much higher probabilities, implying greater confidence. The average entropy for a dataset is calculated as:

$$\overline{H} = -\frac{1}{k}\sum_{i=1}^{k}\sum_{x\in V} p_i(x)\log p_i(x) \qquad (4)$$

with parameters defined as in (Eq. 3). The overall entropy for the dataset is obtained by summing across all k samples.

6 Experiment 1: Multi-token vs. single-token mask filling

To evaluate whether our multi-token mask-filling model can achieve comparable performance on the implicit discourse relation recognition task while reducing ambiguity, we compare it with the single-token baseline model based on off-the-shelf BERT (Dong et al., 2024).

6.1 Test data

We use all implicit inter-sentential relations in the PDTB-3 as our test data. In total, we collected 15,555 argument pairs. Although the PDTB-3 corpus annotates up to two different connectives for each argument pair (when annotators infer more than one sense between a pair of spans), we considered only the sense of the first annotated connective as the gold label in the test set. [4]

6.2 Results

Before analyzing these predictions, we assume that all lexical entries that could serve as connectives are indeed connectives, even if they might also hold other syntactic roles.

The evaluation metric used for this experiment is $a@N$, as defined in Eq. (1). Table 2 shows that multi-token mask-filling model achieves significantly better performances compared to single-token model across three levels. We report only $a@1$ and $a@2$, as these results already suggest an exceptional performance.

[4]Future work could aim at recognizing multiple senses simultaneously, for example, by considering all predictions with probabilities above a specified threshold as valid.

We further examined the accuracy of each sense type at each level of the hierarchy. Results are reported in Appendix B, Tables 5–7. We find that predicting single-token connectives outperforms predicting multi-token connectives for *Comparison.Concession* and *Temporal.Asynchronous*, which is expected because these senses are predominantly conveyed by single-token connectives like "but", "however", "before", and "after". Whereas, the opposite holds for *Arg2-as-detail* and *Arg2-as-instance*. This is largely due to that few multi-token connectives have either of these senses as compared to *Expansion.Level-of-detail.Arg2-as-detail* and *Expansion.Instantiation.Arg2-as-instance*, for which there are several common multi-token connectives, such as "for example", "for instance", "in fact", "in particular", and "that is". Both the sparsity of multi-token instances of *Concession* and *Asynchronous* and the amount of multi-token instances expressing *Arg2-as-detail* and *Arg2-as-instance* are apparent in Appendix F.

7 Experiment 2: Preposed vs. canonical syntax

We then use our model to investigate whether multi-token connective mask-filling also provides evidence that preposed constituents help signal discourse coherence. This expands the empirical base of Dong et al. (2024) from single-token connectives to multi-token connectives, which tend to be less ambiguous (i.e., associated with fewer senses).

7.1 Test data

As in Dong et al. (2024), our test data consists of two datasets: a preposed set and its corresponding canonical set. The datasets were derived from the PDTB-3 (comprising primarily news articles) and DiscoGeM 1.0 (containing political speeches, literature, and Wikipedia texts as well).

We first extracted instances with a preposed structure from the entire set of implicit discourse relations in PDTB-3 and DiscoGeM 1.0, and then constructed the corresponding canonical set for comparison. While the canonical set for PDTB-3 relations was created by moving the preposed NP/PP in Arg2 to the end of the first main clause of Arg2 (see Sec. 2.2, examples (4) and (6)), a different method was needed for DiscoGeM 1.0, which is annotated at the sentence level without specific boundary markers for arguments (see Appendix C). Here, we created a canonical counter-

Model	a@N	Level 1 Accuracy (%)	Level 2 Accuracy (%)	Level 3 Accuracy (%)
BERT single-token mask-filling	a@1	52.10	44.78	40.58
	a@2	70.93	63.93	60.55
Our multi-token mask-filling	a@1	74.48	62.37	58.96
	a@2	87.72	80.02	73.77

Table 2: Model accuracy comparison between single-token mask-filling and multi-token mask-filling across three sense levels in PDTB-3

part to a preposed relation by simply moving the preposed constituent to the end of the second sentence (see Appendix D for more details). There are 156 entries in both the preposed and canonical datasets for DiscoGeM 1.0. For the PDTB-3, we could take advantage of the preposed and canonical datasets used in Dong et al. (2024).[5] These are all implicit inter-sentential relations.

As a result, each complete preposed and canonical set, combining data from both PDTB-3 and DiscoGeM 1.0, contains a total of 1,595 samples.

7.2 Results

Accuracy and precision Table 3a compares the model's predictions for the preposed and canonical sets using accuracy ($a@N$) from Eq.(1), and precision ($p@N$) from Eq.(2). As with the results in Dong et al. (2024) using single-token mask-filling, the multi-token mask-filling model consistently achieves higher accuracy on the preposed set across all N values. This suggests that preposed structures provide information that enables the model to more accurately align its predictions with human annotations. As for precision, the model also consistently performs better on the preposed set than on the canonical set. In general, as N increases, the $p@N$ for both preposed and canonical sets decreases. This is because of the limited number of connectives associated with each specific sense. Therefore, with larger values of N, there are fewer appropriate connectives left to predict, leading to a decrease in $p@N$.

Surprisal and entropy Table 3b presents average surprisal and entropy for both preposed and canonical sets, calculated according to Eq.(3) and Eq.(4).

In line with the accuracy and precision results, surprisal and entropy are also lower for the preposed set, indicating that the model is both more confident and more accurate in these cases than in their canonical counterparts.

Sense types Following Dong et al. (2024), we compare the model's correct predictions for each sense type. While Dong et al. (2024) focused on the top 5 predictions, we consider only the top 1 prediction, as the model achieves very high accuracy on both sets when considering the top 5 (91.29% and 89.72%). Chi-square tests were conducted on the 8 sense types with over 100 instances in the test set; results are presented in Table 4.

The results in Table 4 show significant differences between the preposed and canonical sets for two sense types (*Arg2-as-instance* and *Reason*), at a significance level of 0.05. They are largely consistent with results presented by Dong et al. (2024), who reported significant differences in predicting four sense types: the two mentioned above, plus *Conjunction* and *Arg2-as-detail*. A plausible explanation for the absence of significant differences in *Conjunction* and *Arg2-as-detail* in our results is that training on less ambiguous multi-token connectives has substantially improved the model's understanding of discourse relations. This improvement appears to hold even when the syntactic structure is altered—for example, when a preposed phrase is repositioned to the end of sentence.

Ambiguity of predicted connectives To assess whether multi-token connectives reduce ambiguity in model predictions, we compare the average number of discourse senses associated with the top-1 predicted connectives from our model and from Dong et al. (2024). Focusing on test items where both models predicted the correct sense, we observe that our model's predictions are associated with substantially fewer senses, specifically, 4.93 vs. 13.79 in the preposed setting, and 4.76 vs. 13.48 in the canonical setting. This suggests that our multi-token mask-filling approach yields significantly less ambiguous predictions than the single-token baseline. Full details and results are provided in Appendix E.

[5]We had to remove two instances whose gold sense was *Contingency.Cause+SpeechAct.Reason+SpeechAct*, as we lacked a multi-token connective that could convey this sense.

	Preposed Set		Canonical Set	
N	a@N	p@N	a@N	p@N
1	**60.31%**	**60.31%**	56.05%	56.05%
2	**77.12%**	**56.43%**	73.86%	52.60%
3	**84.70%**	**53.15%**	81.76%	49.55%
4	**89.34%**	**49.92%**	86.83%	47.26%
5	**91.29%**	**47.18%**	89.72%	44.97%

(a) Accuracy@N (a@N) and Precision@N (p@N).

Metric	Preposed Set	Canonical Set
Average surprisal	**1.108**	1.182
Average entropy	**1.996**	2.154

(b) Average surprisal and entropy

Table 3: Preposed set vs. Canonical set: comparing a@N and p@N (left), and average surprisal and entropy (right)

Sense Type	**N**	**Preposed**	**Canonical**	χ^2	**p**
Expansion.Conjunction	340	201	211	0.50	0.48
Expansion.Level-of-detail.Arg2-as-detail	241	218	210	1.02	0.31
Expansion.Instantiation.Arg2-as-instance	191	172	140	16.81	*
Contingency.Cause.Reason	191	119	90	8.28	*
Contingency.Cause.Result	184	86	96	0.88	.35
Comparison.Contrast	139	87	78	0.95	.33
Temporal.Asynchronous.Precedence	131	12	15	0.17	0.68
Comparison.Concession.Arg2-as-denier	101	52	40	2.42	.12

Table 4: Correct top 1 predictions for senses (with more than 100 samples) in preposed set vs. canonical set: counts, and χ^2 test results. N is the frequency of each sense type in the dataset.

8 Conclusions and discussion

As noted in Section 3, single-token mask filling (Dong et al., 2024), despite performing fairly well, is limited by the high sense ambiguity of single-token connectives. The current study addresses this in part, by using the efficient approach to multi-token mask-filling developed by Kalinsky et al. (2023) (see Section 2.3). We adapted their approach to discourse relation recognition by considering multi-token discourse connectives, which we show to be significantly less ambiguous than their single-token counterparts. The results of the first experiment provide evidence for the effectiveness of simple mask-filling with multi-token connectives for discourse relation recognition. We then assessed the effectiveness of our adaptation on the task of sense recognition of implicit discourse relations whose Arg2 starts with a preposed syntactic structure.

Experimental results for the preposing task show that our multi-token mask-filling model achieves higher and more confident performance on recognizing discourse relations on the preposed set than the canonical set, thereby confirming the previous results from Dong et al. (2024) and validating that preposing indeed provides evidence for some (but not all) discourse relational senses. Specifically, preposing significantly helps to indicate two sense types: *Expansion.Instantiation.Arg2-as-instance* and *Contingency.Cause.Reason.* Our results are largely consistent with Dong et al. (2024). Regarding why preposing helps to signal certain discourse relations rather than others, we have an initial hypothesis. According to Ward and Birner (2006), preposing in English is felicitous when the information conveyed by the preposed constituent form an anaphoric link to the prior discourse—that is, either discourse-old (i.e., explicitly evoked in the prior discourse) or inferable from the prior discourse through partial ordering (e.g., type/subtype, entity/attribute, part/whole, etc.). We hypothesize that preposing helps to signal discourse relations when the linking relation between discourse entities also supports the coherence between the preposed sentence and the preceding discourse. In such cases, preposing serves as evidence for a discourse relation. Otherwise, it doesn't. This hypothesis, however, requires further studies to testify.

Our results extend the usefulness of mask-filling, to assessing the possible discourse relevance of other non-canonical syntactic structures, such as post-posing and right extraposition (e.g., "There was a man outside, wearing a plastic raincoat.")

Limitations

The current work is limited in several ways that can and should be addressed in future work.

First, since the amount of annotated implicit discourse relation data is limited, we used explicit discourse relations to construct the training set and test our model on implicit discourse relations. We know from Sporleder and Lascarides (2008) (and more recently, from Liu et al. 2024) that implicit discourse relations are not simply explicit discourse relations that lack an explicit connective. The alternative method used in DiscoGeM 1.0 (Scholman et al., 2022) of crowd-sourcing implicit discourse relation annotation may provide more useful resources in the future.

Also, our test dataset is limited to ~16K samples. The Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017), released in 2017, is another widely used corpus for the implicit discourse relation recognition task. However, it is annotated in the Rhetorical Structure Theory (RST) style (Mann and Thompson, 1987). It is possible that converting this annotation to something more similar to the PDTB will also provide additional annotated data for inference.

The model for predicting multi-token connectives is trained on a predefined set of connectives and corresponding training data. In this study, we chose to use the distribution of connectives occurring in a naturalistic language dataset—specifically, Wikipedia, rather than artificially balancing their frequency within the training data. As a result, some connectives or their senses may be underrepresented, while others could be overrepresented, due to their varying natural occurrence in the data.

Moreover, in choice of the gold label, a single sense was preferred in our experiments. This means we did not explore the ambiguity inherent in discourse relations, where multiple senses may co-occur or be equally plausible (Costa and Kosseim, 2024; Yung et al., 2022). However, the PDTB-3 (Prasad et al., 2019) allowed annotators to record two connectives and senses if they exist in the discourse. This could therefore be used in the future to consider more cases rather than limiting consideration to a single sense.

Ethical Considerations

Our study used two well-established corpora in NLP research. The Penn TreeBank has been used for over 30 years, and DiscoGeM 1.0 was made available in 2022 for general public use. They should therefore present no ethical concerns.

Acknowledgements

We would like to thank Zizhe Wang for early assistance on Experiment 1; Oren Kalinsky and Guy Kushilevitz at Amazon for guidance on using the published dataset via S3 links; and the EMNLP reviewers for highlighting related work from the CODI workshop.

References

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

Debopam Das and Maite Taboada. 2019. Multiple signals of coherence relations. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yunfang Dong, Xixian Liao, and Bonnie Webber. 2024. Syntactic Preposing and Discourse Relations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2790–2802.

Wikimedia Foundation. 2024. Wikimedia downloads. `https://dumps.wikimedia.org`.

Oren Halvani. 2024. Constituent Treelib - A Lightweight Python Library for Constructing, Processing, and Visualizing Constituent Trees. `https://github.com/Halvani/constituent-treelib`.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. 2023. Simple and Effective Multi-Token Completion from Masked Language Models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Stephen Wan, and Michael Strube. 2024. What causes the failure of explicit to implicit discourse relation recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2738–2753.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

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.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn discourse treebank version 3.0. *LDC2019T05*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Wei Shi and Vera Demberg. 2019. 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. Association for Computational Linguistics.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(3):369–416.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2019. 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).

Gregory Ward and Betty Birner. 2006. Information structure and non-canonical syntax. *The handbook of pragmatics*, pages 152–174.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse Treebank 3.0 Annotation Manual*. Available from the Linguistics Data Consortium, https://catalog.ldc.upenn.edu/docs/LDC2019T05/.

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

Yu Xu, Man Lan, Yue Lu, Zheng Yu Niu, and Chew Lim Tan. 2012. Connective prediction using machine learning for implicit discourse relation classification. In *The 2012 international joint conference on neural networks (ijcnn)*, pages 1–8. IEEE.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

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

Sense Type	a@1 (%)		a@2 (%)	
	multi	single	multi	single
Expansion	86.64	48.72	95.41	67.54
Contingency	72.10	54.63	88.85	74.26
Comparison	52.22	58.68	73.62	76.64
Temporal	31.43	51.94	49.15	72.43

Table 5: Accuracy for sense type at Level 1: analysis using top 1 and top 2 predictions

Sense Type	a@1 (%)		a@2 (%)	
	multi	single	multi	single
Expansion.Level-of-detail	82.84	44.62	93.18	64.02
Expansion.Conjunction	60.75	48.72	85.07	66.23
Expansion.Instantiation	76.66	24.55	87.90	44.94
Contingency.Cause	68.58	53.66	87.36	73.79
Comparison.Concession	42.93	45.72	64.47	66.52
Comparison.Contrast	55.36	53.89	74.68	70.80
Temporal.Asynchronous	23.16	37.83	35.74	64.96

Table 6: Accuracy for sense type at Level 2: analysis using top 1 and top 2 predictions (relations with more than 500 occurrences)

A Entropy-based measure of connective ambiguity

Let C denote the set of discourse connectives. For each connective $c \in C$, let $\mathcal{S}_c = \{s_1, s_2, \ldots, s_n\}$ be the set of distinct senses annotated for c. Let $f(c, s)$ denote the frequency with which connective c is annotated with sense $s \in \mathcal{S}_c$.

We define the empirical probability distribution over senses for each connective c as:

$$P_c(s) = \frac{f(c, s)}{\sum_{s' \in \mathcal{S}_c} f(c, s')}$$

The entropy of connective c is then given by:

$$H(c) = -\sum_{s \in \mathcal{S}_c} P_c(s) \cdot \log_2 P_c(s)$$

This entropy $H(c)$ measures the degree of ambiguity of connective c with respect to its senses: the more concentrated the senses are distributed in the dataset, the lower the entropy, indicating lower ambiguity. To obtain an overall measure of connective ambiguity across the dataset, we compute the average entropy over all connectives:

$$\bar{H} = \frac{1}{|C|} \sum_{c \in C} H(c)$$

B Accuracy at different levels

We present the accuracy at Level 1 in Table 5, Level 2 in Table 6, and Level 3 in Table 7.

Sense Type	a@1 (%)		a@2 (%)	
	multi	single	multi	single
Expansion.Level-of-detail.Arg2-as-detail	82.51	43.75	92.72	63.16
Expansion.Instantiation.Arg2-as-instance	76.64	23.60	87.89	43.10
Contingency.Cause.Result	61.47	54.17	80.38	72.86
Contingency.Cause.Reason	53.81	30.02	72.17	58.77
Comparison.Concession.Arg2-as-denier	43.09	45.43	64.48	65.83
Temporal.Asynchronous.Succession	18.92	27.67	25.68	52.20
Temporal.Asynchronous.Precedence	12.50	37.67	20.63	65.39

Table 7: Accuracy for sense type at Level 3: analysis using top 1 and top 2 predictions (relations with more than 500 occurrences)

C Extracting Data from DiscoGem 1.0

For the DiscoGeM 1.0 dataset, preposed structures were identified using the spaCy, NLTK, and constituent treelib libraries (Halvani, 2024) in Python. We treat each sentence in the corpus as an individual argument. For instance, as shown in Ex.(8), the sentence is parsed to generate a constituency tree, which outlines the syntactic structure of the sentence by organizing it into hierarchical components such as S (sentence), PP, and etc. The constituency tree reveals that the phrase "All through that summer" is a PP located at the beginning of the sentence. Dependency parsing is subsequently applied to determine if the preposed NP or PP serves as the grammatical subject of the argument, identified by labels such as "nsubj," "nsubjpass," or "expl." If the phrase does not function as the subject (as in this example), it is classified as a preposed phrase. This method effectively isolates non-subject phrases that have been fronted in the sentence, often for emphasis or to provide context.

(8) **Argument**: All through that summer the work of the farm went like clockwork. [Animal Farm, DiscoGeM 1.0]

Constituency Tree:

```
(S
  (PP (ADVP (DT All)) (IN through) (NP (DT that) (NN summer)))
  (NP (NP (DT the) (NN work)) (PP (IN of) (NP (DT the) (NN farm))))
  (VP (VBD went) (PP (IN like) (NP (NN clockwork))))
  (. .))
Preposed phrase: (PP All through that summer)
```

D More details about the preposed and canonical sets

Similar to the training data, each masked text in the preposed set was formatted as a tuple, "(Arg1, masked Arg2)." As illustrated in the preposed Ex.(9), the PP "by the light of the match" is sentence-initial in Arg2, while in the canonical Ex.(10), the canonical masked text is constructed by right-moving the preposed phrase (either NP or PP) to the end of Arg2. In addition to the masked

text and the annotated sense, metadata for each sample was recorded, including corpus, data source, genre, the inserted connective, and the preposed phrase.

(9) ('He heard a slight groan.', '[MASK], **by the light of the match**$_{\text{preposed PP}}$ he saw a heavy shape moving slightly on the floor.' [Animal Farm, DiscoGeM 1.0]

(10) ('He heard a slight groan.', '[MASK], he saw a heavy shape moving slightly on the floor **by the light of the match**$_{\text{canonical PP}}$.')

E Ambiguity analysis of predicted connectives

One motivation for extending Dong et al. (2024)'s work to include multi-token connectives is the hypothesis that multi-token connectives may reduce sense ambiguity compared to single-token connectives.

To verify this assumption in Section 3, we conduct a following analysis of sense ambiguity. We analyze the average number of sense types that each predicted connective can be mapped to. Since both our model and Dong's model are evaluated on PDTB-3, we base our comparison on the 1,439 inter-sentential implicit relations in the test set (excluding DiscoGem 1.0 instances). We further restrict the analysis to the intersection of instances where both models' top 1 predictions are correct. This results in 388 instances for the preposed set and 332 for the canonical set. As shown in Table 8, our model achieves correct top-1 predictions on multi-token connectives that are associated with significantly fewer sense types than those of Dong's model—4.93 vs. 13.79 in the preposed setting, and 4.76 vs. 13.48 in the canonical setting. This suggests that multi-token connectives, which are more prevalent in our model's predictions (all top 5 predictions are multi-token connectives), are inherently less ambiguous, as reflected by their lower average sense count.

Test Set	Average Sense Count	
	Ours	Dong et al. (2024)
Preposed	**4.93**	13.79
Canonical	**4.76**	13.48

Table 8: Average number of senses for connectives where both models' top 1 predictions are correct.

F Multi-token connectives and their senses (with counts)

This appendix provides each inter-sentential multi-token connective and its senses with counts in PDTB-3 and Connective-Lex (Stede et al., 2019). The counts show how often each connective-sense pair appears in PDTB-3 (Webber et al., 2019), including both implicit and explicit relations. "con-lex" means the sense is listed in Connective-Lex (Stede et al., 2019), but has no frequency data.

Connective	**Sense (Count)**
after all	Contingency.Cause+Belief.Reason+Belief (1) Expansion.Conjunction (con-lex) Expansion.Level-of-detail.Arg2-as-detail (1)
after that	Temporal.Asynchronous.Succession (con-lex)
along with	Expansion.Conjunction (2)
and then	Expansion.Disjunction (1)
as a consequence	Contingency.Cause.Result (2)
as a result	Contingency.Cause+Belief.Result+Belief (6) Contingency.Cause.Result (838) Expansion.Level-of-detail.Arg2-as-detail (1)
as an alternative	Expansion.Disjunction (2)
as it turns out	Contingency.Cause.Result (1) Expansion.Conjunction (1)
as part of that	Expansion.Instantiation.Arg2-as-instance (2)
as such	Contingency.Cause+Belief.Result+Belief (2) Contingency.Cause.Result (5)
as well	Comparison.Similarity (6) Expansion.Conjunction (12)
at that point	Temporal.Synchronous (con-lex)
at that time	Temporal.Synchronous (3)
at the same time	Expansion.Conjunction (1) Temporal.Synchronous (98)
at the time	Temporal.Synchronous (22)
because of that	Contingency.Cause.Result (4)
before that	Temporal.Asynchronous.Succession (1)
but then	Comparison.Concession.Arg2-as-denier (3)
but then again	Comparison.Concession.Arg2-as-denier (1)
by comparison	Comparison.Concession.Arg2-as-denier (2) Comparison.Contrast (198) Expansion.Conjunction (2)
by contrast	Comparison.Concession.Arg2-as-denier (2) Comparison.Contrast (146)
by doing so	Expansion.Manner.Arg1-as-manner (1)
by the way	Comparison.Contrast (con-lex) Expansion.Conjunction (con-lex)
by then	Temporal.Asynchronous.Succession (6) Temporal.Asynchronous.Succession|Contingency.Cause.Reason (1)
despite this	Comparison.Concession.Arg2-as-denier (3)
during that time	Temporal.Synchronous (1)
even before	Temporal.Asynchronous.Precedence|Comparison.Concession.Arg1-as-denier (14)
even before then	Temporal.Asynchronous.Succession|Comparison.Concession.Arg2-as-denier (1)
even then	Temporal.Asynchronous.Precedence|Comparison.Concession.Arg2-as-denier (2)
for example	Contingency.Cause.Reason (1)

	Expansion.Instantiation.Arg2-as-instance (986) Expansion.Level-of-detail.Arg2-as-detail (73)
for instance	Expansion.Conjunction (1) Expansion.Instantiation.Arg2-as-instance (703) Expansion.Level-of-detail.Arg2-as-detail (40)
for one	Expansion.Instantiation.Arg2-as-instance (1)
for one thing	Contingency.Cause.Reason (1) Expansion.Conjunction (1) Expansion.Instantiation.Arg2-as-instance (13) Expansion.Level-of-detail.Arg2-as-detail (8)
for that purpose	Contingency.Purpose.Arg1-as-goal (1)
for that reason	Contingency.Cause.Result (2)
in addition	Expansion.Conjunction (413) Expansion.Level-of-detail.Arg2-as-detail (1)
in any case	Comparison.Concession.Arg2-as-denier (3)
in any event	Expansion.Conjunction (con-lex) Expansion.Level-of-detail.Arg1-as-detail (con-lex)
in comparison	Comparison.Contrast (5)
in contrast	Comparison.Contrast (209)
in fact	Comparison.Concession.Arg2-as-denier (5) Comparison.Contrast (9) Contingency.Cause+Belief.Reason+Belief (6) Contingency.Cause+Belief.Result+Belief (1) Contingency.Cause.Reason (3) Contingency.Cause.Result (2) Expansion.Conjunction (470) Expansion.Equivalence (5) Expansion.Instantiation.Arg2-as-instance (20) Expansion.Level-of-detail.Arg1-as-detail (7) Expansion.Level-of-detail.Arg2-as-detail (389)
in general	Expansion.Level-of-detail.Arg1-as-detail (3)
in more detail	Expansion.Level-of-detail.Arg2-as-detail (1)
in other words	Comparison.Similarity (1) Contingency.Cause.Reason (1) Contingency.Cause.Result (1) Expansion.Conjunction (3) Expansion.Equivalence (247) Expansion.Level-of-detail.Arg1-as-detail (25) Expansion.Level-of-detail.Arg2-as-detail (16)
in particular	Expansion.Conjunction (1) Expansion.Instantiation.Arg2-as-instance (73) Expansion.Level-of-detail.Arg2-as-detail (666)
in response	Contingency.Cause.Result (2) Expansion.Conjunction (1)
in short	Contingency.Cause+SpeechAct.Result+SpeechAct (1) Contingency.Cause.Reason (2) Contingency.Cause.Result (1) Expansion.Conjunction (6) Expansion.Equivalence (20) Expansion.Level-of-detail.Arg1-as-detail (83)

	Expansion.Level-of-detail.Arg2-as-detail (18)
in sum	Expansion.Conjunction (6) Expansion.Equivalence (4) Expansion.Level-of-detail.Arg1-as-detail (28) Expansion.Level-of-detail.Arg2-as-detail (1)
in the end	Comparison.Concession.Arg2-as-denier (1) Comparison.Contrast (1) Contingency.Cause.Result (6) Expansion.Conjunction (21) Expansion.Equivalence (1) Expansion.Level-of-detail.Arg1-as-detail (5) Expansion.Level-of-detail.Arg2-as-detail (5) Temporal.Asynchronous.Precedence (8)
in the meantime	Temporal.Asynchronous.Succession (2) Temporal.Synchronous (12) Temporal.Synchronous|Comparison.Contrast (1)
in the meanwhile	Temporal.Synchronous (1)
in this case	Expansion.Instantiation.Arg2-as-instance (1)
in this way	Contingency.Cause.Result (con-lex)
in turn	Contingency.Cause.Result (con-lex) Expansion.Conjunction (con-lex) Temporal.Asynchronous.Precedence (con-lex)
later on	Temporal.Asynchronous.Precedence (2)
more accurately	Expansion.Substitution.Arg2-as-subst (1)
more specifically	Expansion.Level-of-detail.Arg2-as-detail (18)
more to the point	Expansion.Level-of-detail.Arg2-as-detail (1)
no matter	Comparison.Concession.Arg1-as-denier (8)
on the contrary	Comparison.Contrast (11) Expansion.Level-of-detail.Arg2-as-detail (1)
on the other hand	Comparison.Concession.Arg2-as-denier (4) Comparison.Contrast (62)
on the whole	Expansion.Conjunction (10) Expansion.Level-of-detail.Arg1-as-detail (19) Expansion.Level-of-detail.Arg2-as-detail (8)
prior to this	Temporal.Asynchronous.Succession (1)
since then	Temporal.Asynchronous.Precedence (7)
that is	Contingency.Cause.Reason (1) Contingency.Cause.Result (3) Expansion.Conjunction (2) Expansion.Equivalence (30) Expansion.Level-of-detail.Arg1-as-detail (6) Expansion.Level-of-detail.Arg2-as-detail (51)
to this end	Contingency.Cause.Result (1)
what's more	Expansion.Conjunction (1)