Syntactic Preposing and Discourse Relations

Yunfang Dong[♣], **Xixian Liao**[♠], **Bonnie Webber**[♡]
[♣] School of Philosophy, Psychology & Language Science, University of Edinburgh
[♠] Department of Translation and Language Sciences, Universitat Pompeu Fabra
[♡] ILCC, School of Informatics, University of Edinburgh
yunfang.dong@outlook.com
xixianliao@gmail.com
bonnie.webber@ed.ac.uk

Abstract

Over 15 years ago, Ward and Birner (2006) suggested that non-canonical constructions in English can serve both to mark information status and to structure the information flow of discourse. One such construction is *preposing*, where a phrasal constituent appears to the left of its canonical position, typically sentence-initially. But computational work on discourse has, to date, ignored non-canonical syntax. We take account of non-canonical syntax by providing quantitative evidence relating NP/PP preposing to discourse relations. The evidence comes from an LLM mask-filling task that compares the predictions when a mask is inserted between the arguments of an implicit inter-sentential discourse relation — first, when the right-hand argument (**Arg2**) starts with a preposed constituent, and again, when that constituent is in canonical (post-verbal) position. Results show that (1) the top-ranked mask-fillers in the preposed case agree more often with "gold" annotations in the Penn Discourse TreeBank (Webber et al., 2019) than they do in the latter case, and (2) preposing in **Arg2** can affect the distribution of discourse-relational senses.

1 Introduction

While sentences in discourse are organized in a coherent manner, there are different ways of indicating how a clause and/or sentence relates to its neighbors — with an explicit discourse connective (as in (1))[1], or a lexico-syntactic construction (such as the *so Adjective* construction in (2)), or an alternative lexicalization of an explicit connective, such as *provided* conveying the sense as *if* in (3), or with punctuation or other text structuring devices discussed in Das and Taboada (2018).

(1) *Output will be gradually increased* until **it reaches about 11,000 barrels a day**. [wsj_0024]

(2) *The fit is so good*, **we see this as a time of opportunity**. [wsj_0317]

(3) *The prepaid plans may be a good bet*, **provided the guarantee of future tuition is secure**. [wsj_1569]

But readers/listeners can recognize such relations, even when such evidence is absent, as in (4), where the second sentence is taken to be more detailed about the claim in the first. (This has been called an *implicit discourse relation*.)

(4) *But the market is changing*. **The government is funding several projects to push PC use**. [wsj_0445]

The Penn Discourse Treebank 3.0, abbreviated PDTB-3 (Webber et al., 2019), is a large manually annotated corpus of discourse relations annotated over the Wall Street Journal section of the Penn Treebank (PTB) (Marcus et al., 1993). As shown in Figure 1, while connectives could be directly extracted from the text for explicit relations, for implicit relations, human annotators were required to first insert a connective to aid in annotating the relation and then identify its sense. Annotators were allowed to insert two connectives and their senses if they felt that more than one sense held between the two arguments.

In their summary of work on non-canonical syntax in English, Ward and Birner (2006) observed that linguists had identified two functions that preposed constituents serve: signalling information status and structuring information flow. We take the latter to include coherence relations between

[1] The parts of a discourse relation are indicated by underlining explicit connectives, italicizing the first argument to the relation and bolding the second. While our focus here is on English, explicit discourse connectives have been identified in many languages including Chinese, Czech, French, German, Lithuanian, Polish, Portuguese, Russian and Turkish (Zeyrek et al., 2019; Özer et al., 2022).

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2790–2802
March 17-22, 2024 ©2024 Association for Computational Linguistics

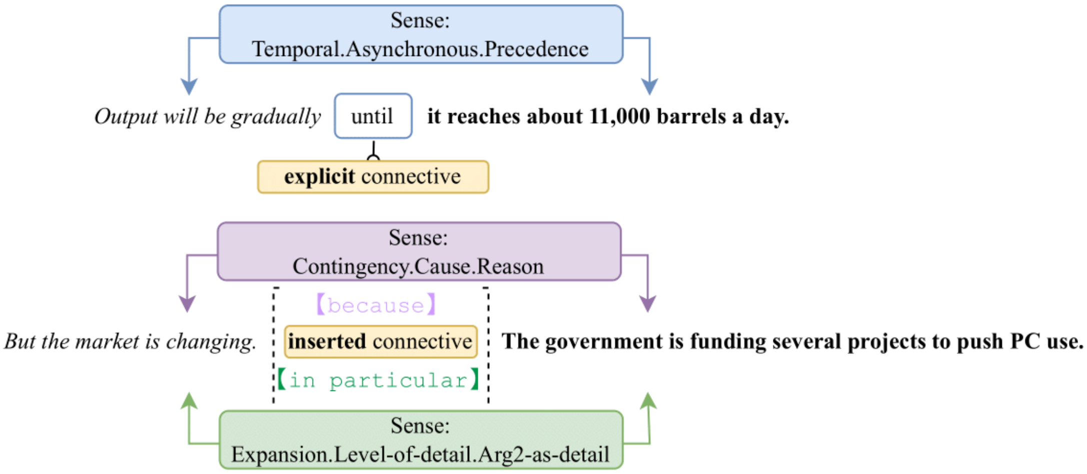
Sense:
Temporal.Asynchronous.Precedence
Output will be gradually
until
it reaches about 11,000 barrels a day.
explicit connective
Sense:
Contingency.Cause.Reason
【because】
But the market is changing.
inserted connective
The government is funding several projects to push PC use.
【in particular】
Sense:
Expansion.Level-of-detail.Arg2-as-detail

Figure 1: Examples of explicit (upper) and implicit (bottom) relations annotated in the PDTB-3 corpus

discourse units, also called *discourse relations*, and complement their study with quantitative work relating preposing and discourse relations. The work makes two contributions: It shows that (1) when **Arg2** of an inter-sentential implicit discourse relation begins with a preposed constituent, that relation is signalled more strongly than when the constituent is in canonical position, and (2) preposing in **Arg2** changes the distribution of discourse relational senses that hold between its arguments.

In what follows, Section 2 provides background motivation. Sections 3 and 4 describe the data and methodology used in the experiments, with results presented in Section 5. Section 6 discusses the results, while the Limitations section identifies limitations of the current work that should be addressed in the future.

2 Background

2.1 Discourse relation recognition

The discourse relations that hold between segments of texts can provide useful information for NLP tasks such as information extraction (e.g., Cimiano et al., 2005). Explicit relations can be accurately identified using a straightforward frequency-based classification approach that maps explicit connectives to senses (Xue et al., 2016).

For identifying implicit discourse relations, some recent studies have used prompt learning techniques to guide pre-trained language models to predict connectives between argument pairs and subsequently map them to corresponding discourse relations (Xiang et al., 2022; Zhou et al., 2022). Viewed this way, discourse relation recognition resembles a *cloze task* in which a portion of text is masked (here, the position between the two arguments, where a discourse connective could be made explicit) and where the respondent (here, the language model) must fill in the mask. While cloze tasks are generally easy for people to solve, implicit discourse relation recognition remains a significant challenge, so can benefit from any information that may have been ignored.

2.2 Preposing

Non-canonical syntactic constructions in English have been characterized in terms of both form and function. One such construction is *preposing*, where a constituent appears to the left of its canonical position, usually sentence-initially (Ward and Birner, 2006). The constituent that is preposed is called the *preposed constituent*, and it can take various forms, including a noun phrase (NP), a prepositional phrase (PP), a verb phrase (VP), or an adjective phrase (AP). Ex. (5) and Ex. (6) illustrate sentences with a preposed PP or NP.

(5) *We think there will be positive as well as negative reactions.* **On balance**$_{PP}$**, we think it will be positive**. [wsj_0277]

(6) *Some researchers have charged that the administration is imposing new ideological tests for top scientific posts.* **Earlier this week**$_{NP}$**, Dr. Sullivan tried to defuse these charges...**. [wsj_0047]

Preposing has long been discussed in linguistics as an indicator of topicalization in information structure. Yet previous research has also suggested that, in addition to marking information status, preposing can also structure the information flow of the discourse (Ward and Birner, 2006). We take the information flow of discourse to include the discourse relations. We thereby hypothesize that preposing might serve to indicate discourse relations.

In order to explore this hypothesis quantitatively, we use pre-trained large language models to investigate inter-sentential implicit discourse relations whose right-hand argument (**Arg2**)[2] contains a preposed NP or PP. We do this by asking the LLM to predict what fills a mask inserted between the arguments of inter-sentential implicit relations and, when a discourse connective is predicted, examining what discourse relational sense is conveyed (Section 3.3). The results of this study not only have implications for linguistic theory but also offer insights for improving discourse relation recognition.

2.3 Masked language models

In Section 2.1, we noted that implicit inter-sentential relation recognition can resemble a *cloze task*, in which the break between the sentential arguments is viewed as a gap that should be filled with a discourse connective, before positing its sense. The current study uses the off-the-shelf pre-trained language model BERT (Devlin et al., 2019) to propose fillers for this gap-filling cloze task, even though traditionally, a human subject has this role (Taylor, 1953). BERT is appropriate to use here since it is pre-trained on masked language modeling as a way of learning contextual word representations. It is also trained on next-sentence prediction, making a binary choice of whether two sentences are in sequence. This enhances BERT's comprehension of the relationships between sentences and longer-term dependencies across sentences.

Masked language models like BERT have been shown to exhibit biases consistent with human behavior (at least for English). This is consistent with evidence suggesting fundamental connections between deep language models and human language processing (e.g., Linzen and Baroni, 2021; McClelland et al., 2020; Hasson et al., 2020; Goldstein et al., 2022). As such, predictions from BERT-like models have been adopted as proxies for human predictions when addressing linguistic questions (e.g., Davis and van Schijndel, 2021; Aina et al., 2021; Irwin et al., 2023).

Since cloze tasks can be expensive when performed by human participants, using approximations from language models like BERT allows for a large-scale, but relatively inexpensive investigation. For instance, Pimentel et al. (2020) used BERT to calculate the surprisal of a masked word based on its left and right context, as a proxy for word predictability. Analogously, they used a BERT-based estimate of lexical ambiguity, found to correlate with the number of human-annotated senses of a word. Both uses of BERT allowed the experiments to be done on a large number of languages.

3 Method

This section describes (1) the process for extracting inter-sentential implicit relations in the PDTB-3 whose **Arg2** starts with a preposed constituent, (2) the process for creating the two datasets whose mask fillers will be compared, and (3) the mask-filling task we conduct using BERT.

3.1 Extracting discourse relations with preposing in Arg2

We use Tregex (Levy and Andrew, 2006), a tool developed by the Stanford NLP group for finding parse structures that match specified syntactic patterns in the PTB, on parse trees from the Wall Street Journal (WSJ) section of the PTB to extract sentences starting with a preposed NP or PP.[3] To help in the next step (aligning parse structures with the PDTB-3, where the arguments to discourse relations are identified by their byte position in the raw text), we use a version of the PTB whose parse nodes are annotated with the byte span of their projection onto the raw text.

In total, parse trees of 4988 sentences are matched and extracted, along with their corresponding byte spans.

As noted, we focus on inter-sentential implicit relation tokens, particularly those where a preposed phrase is extracted from the beginning of its right-hand argument (**Arg2**) such as Ex. (7). (But see the Limitations section for other types of examples that could be included in subsequent studies.) So as the next step, we extract from the 4988 sentences only those that start the right-hand argument (**Arg2**) of an implicit inter-sentential relation token in the PDTB-3 corpus using its file number and corresponding byte spans. Specifically, we do it by mapping the start span of a preposing sentence

[2] In implicit inter-sentential relations in the PDTB-3, *Arg1* always precedes **Arg2**.

[3] The Tregex pattern we use for matching and extracting a preposed NP/PP is: $(@PP|NP > 2(S! >> /S.*/)\&\$+ +(/NP-SBJ.*/ > (S! >> /S.*/)))$. (But see the Limitations section and Appendix A.1 for more details.)

to Field 31 of the PDTB-3 relation token which specifies the start span of **Arg2** of an implicit token. (Appendix A.2 provides more information about the methods we use for extracting these relations.)

(7) Expansion.Level-of-detail.Arg2-as-detail: *South Korea has different concerns.* [inserted: specifically] **In Seoul$_{PP}$, officials began visiting about 26,000 cigarette stalls to remove illegal posters and signboards advertising imported cigarettes**. [wsj_0037]

Of the 4988 sentences containing a preposed NP or PP, 1441 occurs in **Arg2** of an implicit inter-sentential relation. We also create a separate set comprising all inter-sentential implicit relations in the PDTB-3 that don't belong to the preposed set (14116 relation tokens in total). This we call the **complement set**. Its use is described in Section 5.2.

3.2 Data preprocessing

Using the 1441 extracted relation tokens, we create two distinct sets as input to the mask-filling task. In the first set, we concatenate each argument pair to form a continuous passage, insert a [MASK] token after the end of *Arg1* and before the start of **Arg2**, and then add the sentence boundary tokens [CLS] and [SEP] commonly used for the Next Sentence Prediction task in BERT pre-training, as shown in Ex. (8). We call this the **preposed set**.

In the second set, we concatenate *Arg1* with a version of **Arg2** in which the preposed phrase has been moved to its canonical position, which we take to be the end of the first sentence in **Arg2** that starts with the preposed constituent. As with elements of the **preposed set**, we insert a [MASK] token after the end of *Arg1* and before the start of the now modified **Arg2**, and then add sentence boundary tokens [CLS] and [SEP], as shown in Ex. (9). We call this the **canonical set**. (There are also a few special cases, which we describe in Appendix B.)

(8) [CLS] *We think there will be positive as well as negative reactions* [SEP] [MASK] **On balance$_{PP}$, we think it will be positive** [SEP]

(9) [CLS] *We think there will be positive as well as negative reactions* [SEP] [MASK] **we think it will be positive on balance$_{PP}$** [SEP]

3.3 Mask-filling

We use the off-the-shelf masked language model (MLM) BERT-base (Devlin et al., 2019), with its 12 hidden layers of 768 units and 12 attention heads to predict the inserted [MASK] token in each item from the two datasets. Using off-the-shelf BERT-base is sufficient because our objective is not to find the best possible mask fillers, but rather to show that more of the high-confidence mask fillers predicted by a competent MLM correlate with sense-appropriate discourse connectives when **Arg2** begins with a preposed constituent than when that constituent appears in its canonical position.

We extract the top 5 model predictions for each [MASK] token, along with their probabilities. If a predicted token matches either one of the one-word explicit connectives annotated in the PDTB-3 or one inserted by an annotator annotating an implicit relation, it is mapped to all relation senses it is associated with in the PDTB-3, as illustrated in Figure 2. This approach is akin to the connective-cloze task that has been employed in implicit discourse relation recognition (cf. Section 2.1).

While BERT predicts only a single token for each masked token, some connectives such as "as a result" span multiple words. However, all relation senses in our datasets can be conveyed by single-word connectives, which also account for over 80% of the connectives most frequently inserted in the PDTB-3 for implicit relations. As such, we believe that our focus on single-token prediction does not compromise either our objectives or the results of our experiments. On the other hand, we recognize the benefit of having multi-token completions and see it as a promising avenue for future exploration (cf. the Limitations section).

4 Analysis of predicted fillers

Before turning to preposing and discourse relations, we first summarize what BERT chooses as mask fillers. We focus on BERT's top 5 predictions because their average probabilities run from 0.41 to 0.15, 0.08, 0.05 and 0.03, resulting in their having a cumulative probability of 72%. Given that the remaining probability mass is distributed across a long tail of predictions with low probability, we have not considered these predictions any further.

Among the top 5 predictions are (1) connectives ($\sim$60% of the predictions in the preposed set and $\sim$55% in the canonical set); (2) stance adverbs such as "allegedly", "surely", "hopefully" (Biber

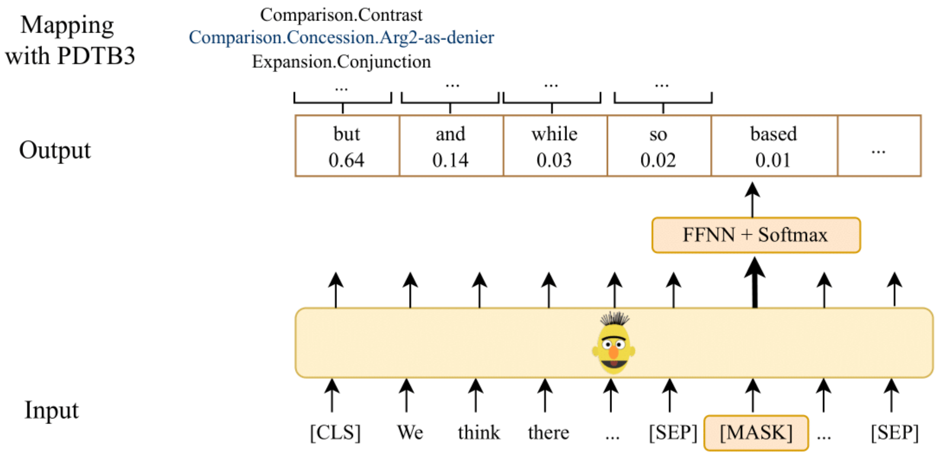

Figure 2: Illustration of mask-filling using BERT and mapping between predicted tokens and senses annotated in the PDTB-3

Dataset	Conn.	Adverbials				Focus Particles	Discourse Markers	Other	Total
		Stance	Frequency	Locational	Temporal				
Preposed	4238	36	17	20	29	320	7	2538	7205
Canonical	3993	81	47	28	186	86	21	2763	7205

Table 1: The number of predicted fillers of different types

and Finegan, 1988); (3) frequency adverbs such as "often", "sometimes", "usually" (Bass et al., 1974; Kennedy, 1987); (4) locational adverbs such as "here" and "there"; (5) temporal adverbs such as "today", "Friday", "currently"; (6) focus particles such as "even", "only" and "just" (König, 2002); and (7) discourse markers such as "well", "now", "anyway" (Schiffrin, 1987).[4] The number of different lexical tokens in each category for both the preposed and the canonical set is shown in Table 1.

The preposed set has more discourse connectives than the canonical set, and also more focus particles, since discourse connectives are often modified by focus particles as in "even when", "only after", and "just because".

In order to demonstrate the main claims of the paper (cf. Section 1), we focus on the explicit connectives predicted by BERT and the senses associated with them.

5 Results

Before presenting our results, we first note the evaluation measures we use in assessing BERT's predictions.

Evaluation measures. We use two measures to evaluate the model predictions from two different perspectives. The first is *accuracy*, which has the value 1 ($accuracy(N)$=1) if any of BERT's top N predictions $pred_i^N$ for item i in the dataset is an explicit connective that can convey the sense annotated in the PDTB-3 ($gold_i$), otherwise 0.[5] As defined in Eq. (1), we calculate and report the mean accuracy(N) for all items in a dataset.

$$\text{accuracy}(N) = \frac{1}{k}\sum_{i=1}^{k}\begin{cases}1 & n(pred_i^N \cap gold_i) > 0\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

Given that the top 5 predictions for each item are ordered by probability, the second measure we use is Precision@N, also written P@N, which indicates the proportion of the top-N predictions that are correct. That is, P@1 indicates whether the top-1 predicted token is correct (P@1=100%) or incorrect (P@1=0%), while P@2 reports on the top-2 predictions: P@2=100% if both of the top-2 tokens are correct; P@2=50% if one of them is correct, while P@2=0% if neither is correct. Similarly, for P@3, P@4 and P@5. Since implicit relations in the PDTB-3 can be taken to have more than one sense, if any of the senses associated with a predicted token agree with any of the gold senses[6], the

[4] The category of "Other" mainly comprises content words, numbers and punctuation.

[5] When a prediction is a connective, one of whose senses agrees with that annotated in the PDTB-3, we say that the prediction is *correct*.

[6] Note that we allow up to two implicit connectives, each

prediction will be considered correct.

$$\text{precision}(N) = \frac{1}{k} \sum_{i=1}^{k} \frac{n(pred_i^N \cap gold_i)}{N} \quad (2)$$

5.1 Preposed set vs. Canonical set

Accuracy and P@N. Table 2 compares BERT's predictions for the preposed and the canonical set in terms of accuracy and P@N. It shows that BERT achieves consistently higher accuracy and P@N on mask-filling predictions for the preposed set compared to the canonical set, irrespective of how many of the top 5 predictions are considered.

To confirm that the observed accuracy and precision are a consequence of the different forms of **Arg2** in the preposed and canonical sets, we conduct a variant of the mask-filling task in which *Arg1* is removed from the pattern. That is, what is submitted to BERT are instances of the patterns [CLS] [MASK] [preposed Arg2] [SEP] and [CLS] [MASK] [canonical Arg2] [SEP]. Table 3 illustrates the model performance in the variants of the preposed and the canonical set.

A comparison between the results in Table 2 and Table 3 shows that while both accuracy and P@N drop over the two sets, the drop is considerably more for the canonical set, suggesting that the preposed constituent appears to provide evidence for what discourse relation the speaker intends.

To assess what BERT finds hard to predict correctly when only given **Arg2**, we count the occurrence of six most frequent one-word connectives ("and", "but", "because", "then", "so" and "with") that convey senses with a sample size of at least 95 [7] in the preposed and canonical sets. (See Table 4.) The only connective whose predictions drop dramatically compared to when both arguments are present is "because", which is most commonly used to convey "Reason". This makes sense since "Reason" holds when **Arg2** provides a reason for the event or situation described in *Arg1*. With no *Arg1*, it is difficult to interpret **Arg2** as the reason for it holding. On the other hand, the drop is more salient in the canonical set than in the preposed set, suggesting that preposing might carry information for the "Reason" relation.

Model confidence. Since there are 1052 cases where BERT predicts a sense-appropriate connective in both datasets, we want to see if there is any difference in its predictions. Focussing on prediction confidence, we find that in more than half of the cases (604, 57%), the top sense-appropriate predicted connective has a higher probability in the preposed set than it does in the canonical set. This suggests that BERT's superior performance on the preposed set across all measures serves as empirical support for the hypothesis that preposed constituents provide evidence for the discourse relation holding between the arguments.

For which senses does preposing most help BERT's predictions? We examine the sense types BERT correctly predicted in the preposed and canonical sets when all top 5 predictions are considered. We find that BERT achieves an accuracy of ∼90% or higher for the senses *Expansion.Conjunction*, *Expansion.Level-of-detail.Arg2-as-detail*, *Contingency.Cause.Reason*, *Contingency.Cause.Result* and *Comparison.Concession.Arg2-as-denier* in the preposed set. In contrast, accuracy for the senses *Comparison.Contrast* and *Temporal.Asynchronous.Precedence* is equally high (∼80%) in both sets.

To determine if the performance difference is significant, we conduct Chi-square tests on those sense types with a sample size of at least 95. (The significance level is chosen to be 0.05.) The results are presented in Table 5. The results show that the four senses *Expansion.Conjunction*, *Expansion.Level-of-detail.Arg2-as-detail*, *Expansion.Instantiation.Arg2-as-instance* and *Contingency.Cause.Reason* are the senses in which BERT makes significantly better predictions. For other senses, performance is comparable. This is in line with the common intuition that discourse relations can be signaled by various cues. While preposing provides evidence for some discourse relations, other relations may be signaled by additional cues present in both datasets, leading to similar accuracy for these sense types in both datasets.

5.2 Preposed set vs. Complement set

To determine whether discourse relations with a preposed constituent in **Arg2** may differ in their distribution of sense types from that of implicit inter-sentential relations more generally, we com-

with up to two senses.

[7] See Table 5. We choose 95 as the threshold since the count drops drastically below this threshold.

	Preposed Set		Canonical Set	
N	Acc.	P@N	Acc.	P@N
1	**0.44**	**0.44**	0.39	0.39
2	**0.67**	**0.45**	0.58	0.38
3	**0.79**	**0.43**	0.69	0.38
4	**0.86**	**0.41**	0.75	0.37
5	**0.89**	**0.39**	0.80	0.36

Table 2: Comparison of accuracy and Precision@N for the proposed and canonical sets

	Preposed Set		Canonical Set	
N	Acc.	P@N	Acc.	P@N
1	**0.36**	**0.36**	0.31	0.31
2	**0.61**	**0.39**	0.48	0.31
3	**0.74**	**0.38**	0.58	0.31
4	**0.81**	**0.37**	0.65	0.30
5	**0.85**	**0.35**	0.68	0.29

Table 3: Comparison of accuracy and Precision@N for the proposed and canonical sets with Arg2 alone

	And	But	Because	Then	So	With
Preposed set	1149	977	191	84	203	0
Canonical set	941	678	198	92	199	12
Preposed set with Arg2 alone	991	921	28	180	185	2
Canonical set with Arg2 alone	821	662	18	207	148	3

Table 4: The counts of “and”, “but”, “because”, “then”, “so” and “with” predicted by BERT in the preposed and canonical sets with and without Arg1

pare the distribution of sense types in the preposed set and the complement set (cf. Section 3.1). All sense types with a sample size of at least 95 in the preposed set are included in the analysis. We perform a Chi-square test to assess whether there is an association between preposing NP/PP and sense types within inter-sentential implicit instances. The analysis reveals a significant association between two ($\chi^2(7) = 159.67, p < 0.001$), indicating that the distribution of senses differs between the preposed set and the complement set, i.e., preposed NPs/PPs tend to be more frequently observed than expected with certain sense types and less with other.

We conduct a Chi-square post-hoc test to determine which specific sense types are driving the significant association. The results are presented in Table 6. It is clear from Table 6 that the three senses —*Comparison.Contrast*, *Expansion.Instantiation. Arg2-as-instance*, and *Temporal.Asynchronous.Precedence*—occur more frequently in the preposed set (that is, with an **Arg2** with a preposed NP/PP) than in general (i.e., in the complement set of inter-sentential implicit relations in general), as indicated by the positive residuals. This suggests that preposed NPs/PPs may occur more frequently in **Arg2** when one of these sense types is being conveyed.

6 Discussion and Conclusion

Our study is the first to empirically validate Ward and Birner (2006)’s claim based on qualitative linguistic evidence that preposing serves to structure information flow through a discourse.

To this end, we show that preposing provides evidence for the discourse-relational sense(s) that human annotators have ascribed to the relation tokens. The evidence comes from comparing BERT’s performance, in terms of accuracy and confidence, across datasets used in mask-filling tasks. BERT demonstrates higher accuracy and confidence when predicting connectives for implicit relations where **Arg2** begins with a preposed NP/PP than when that NP/PP appears in its canonical position. To validate the method, we feed in two variant patterns of the preposed and canonical sets that consist of only **Arg2**. A more dramatic drop of both accuracy and P@N in the canonical set with **Arg2** alone sug-

Sense Type	**N**	**N/% Preposed**	**N/% Canonical**	χ^2	***p***
Expansion.Conjunction	292	266/0.91	240/0.82	9.25	*
Expansion.Level-of-detail.Arg2-as-detail	237	209/0.88	178/0.75	12.67	*
Expansion.Instantiation.Arg2-as-instance	189	153/0.81	125/0.66	9.91	*
Contingency.Cause.Reason	188	176/0.94	153/0.81	11.77	*
Contingency.Cause.Result	158	145/0.92	136/0.86	2.06	.15
Comparison.Contrast	137	109/0.80	108/0.79	0	1
Temporal.Asynchronous.Precedence	104	86/0.83	83/0.80	0.13	.72
Comparison.Concession.Arg2-as-denier	95	87/0.92	77/0.81	3.61	.06

Table 5: Correct predictions for each sense type (with a sample size of at least 95) in preposed vs. canonical sets: counts, proportions, and χ^2 test results

	Concession	**Contrast**	Reason	Result	Conjunction	**Instance**	Detail	**Precedence**
Preposed	-3.17 (*)	**6.96 (*)**	-3.61 (*)	-0.82 (1)	-3.97 (*)	**4.63 (*)**	-0.65 (1)	**8.02 (*)**
Complement	3.17 (*)	**-6.96 (*)**	3.61 (*)	0.82 (1)	3.97 (*)	**-4.63 (*)**	0.65 (1)	**-8.02 (*)**

Table 6: Residuals from post hoc chi-square test results (*p* values in parentheses): Absolute residual magnitude indicates deviation from expected frequencies, with positive or negative signs indicating lower or higher frequencies than expected

gests that higher performance can be attributed to preposing.

We further examine the preposed and canonical sets to determine those senses where preposing helps BERT's prediction. The evidence from Chi-square tests (Table 5) suggests that BERT is significantly better at predicting *Expansion.Conjunction*, *Expansion.Instantiation.Arg2-as-instance*, *Expansion.Level-of-detail.Arg2-as-detail*, and *Contingency.Cause.Reason* on the preposed set, whereas its performance on other senses is comparable with-/without preposing. While preposing significantly improves BERT's prediction on these four senses, other senses might be predicted based on a combination of discourse cues. This is compatible with the claim by Das and Taboada (2019) that discourse relations can be signalled by multiple cues such as syntactic, semantic, lexical, morphological features. For instance, Ex. (10) demonstrates that the "Comparison" relation between the argument pair is simultaneously signalled by the explicit connective "while" and the syntactically parallel constructions "X has a Y".

(10) *Tele-Communications has a 21.8% stake*, while **Time Warner has a 17.8% stake**. [wsj_1190]

We also compare the distribution of sense types between the preposed set and the complement set. We provide quantitative evidence (Table 6) that the three discourse relations *Comparison.Contrast*, *Expansion.Instantiation.Arg2-as-instance* and *Temporal.Asynchronous.Precedence* are more frequently signalled when **Arg2** contains a preposed NP/PP. This is compatible with the claim in Ward and Birner (2006) that information conveyed by a preposed constituent can be linked to the previous discourse in any way that can be construed as a partial ordering. This would include temporal ordering (as with *Temporal.Asynchronous.Precedence*), ordering by inclusion (e.g., a set and its members, as is the case with *Expansion.Instantiation.Arg2-as-instance*), and alternatives ordered with respect to an inferred set, as is the case with *Comparison.Contrast*. Therefore, syntactic preposing can be regarded as a signal that increases the likelihood of classifying specific relation senses. In practical terms, this means that for studies employing conventional Machine Learning approaches for implicit relation recognition, it may be beneficial to consider incorporating a syntactic feature that indicates whether a sentence contains a preposed NP/PP.

Limitations

While the current study supports the claim that non-canonical syntax can provide evidence for the existence of discourse relations and the senses they convey, further work is suggested by limitations of the study, including (1) in the methodology – both computational and linguistic; and (2) data noise.

Starting with computational methodology, while we have a valid rationale for using single-token mask-filling (see Section 3.3), future research could benefit from exploring multi-token fillers, given that they are (in general) less ambiguous. There are several possibilities with multi-word mask-fillers. For instance, (i) multiple words that make up a single phrase, like "in short", "in summary", etc. (ii) multiple connectives, like "but instead", "and then". Both suggest more specific sense relations than a single token could convey.

With respect to linguistic methodology, the current study does not distinguish among the different types of preposing such as identity linking with the prior discourse, proposition affirmation, focus preposing or topicalization discussed in Ward and Birner (2006). Different types of proposing may be related to different discourse relations, so should be a focus of future work. In addition, preposing is not the only form of non-canonical syntax. Ward and Birner (2006) also discussed postposing, which places a constituent (often the subject) to the right of its canonical position. Postposing too may serve to signal discourse relations.

There are also several limitations associated with noise in the data. While these issues represent a very small proportion of the data and don't invalidate our results, addressing them could merit effort for more robust future results. The first limitation comes from the Tregex pattern used to extract a preposed NP/PP in the PTB corpus. The pattern specifies the preposed NP/PP being the left-most daughter of the top-level S node. This ignores cases where punctuation precedes the preposed PP/NP which are not annotated under the scope of the preposed NP/PP. The Tregex pattern also extracts sentences with multiple NPs/PPs that occur before the matrix subject, which necessitates moving all preposed NPs/PPs to their canonical positions.

The second data limitation is associated with adjacency. While *Arg1* and **Arg2** of inter-sentential implicit relations in the PDTB-3 may not actually be adjacent because of attribution following *Arg1* (Prasad et al., 2008), we have considered two arguments to be adjacent even when attribution intervenes, as in Ex. (11), where the attribution "Banxquote said" separates *Arg1* from **Arg2**.

(11) *The average six-month yield on a jumbo CD was at 7.90%, down from 7.93%*, Banxquote said. **For longer-term CDs, yields were up**. [wsj_0238]

In addition, there are parsing inconsistencies in the PTB, and labelling inconsistencies in the PDTB. The former leads to problems in identifying instances of non-canonical syntax, which may not all have been parsed in the same way, while the latter leads to problems in interpreting what discourse relation(s) may be associated with a particular discourse connective.

Next, we identify additional data that could be investigated in the future to broaden the scope of our current analysis. Firstly, our study focuses on preposed NPs/PPs in **Arg2** of paragraph-internal adjacent sentences. Future work can extend to ones that occur elsewhere, such as paragraph-initially. Work on annotating cross-paragraph implicit discourse relations (Prasad et al., 2017) should enable future exploration of cross-paragraph cases.

Secondly, we examine inter-sentential relations in the PDTB-3 and do not consider non-adjacent arguments. Future work could explore non-adjacent cases using analyses such as in Prasad et al. (2011). While this study has also excluded intra-sentential discourse relations, future work can look at these cases in the PDTB-3, which is annotated with several thousand such relations (Prasad et al., 2018).

Thirdly, while we focus only on preposed NPs/PPs, other syntactic categories can be preposed, including adverb phrases and verb phrases (cf. Section 2.2). One needs to investigate whether preposing of these categories also correlates with discourse relations. Future work can also take into account finer-grained functional distinctions between preposed constituents of the same category, such as temporal PPs (tagged PP-TMP), locational PPs (tagged PP-LOC). It is likely that different tags will correlate with different discourse relations.

Lastly, the present study only considers text from the Wall Street Journal. In the future, preposing can be analyzed in other news corpora (which will have their own style sheets), as well as other genres, to assess whether preposing is used in the same way or with the same distribution.

Ethical Considerations

Our study uses a well-established corpus in NLP research for over 30 years, thereby presenting no ethical concerns.

Acknowledgements

We want to thank Yunong Liu, for help with extracting the arguments to discourse relations based on their spanlists; Hannah Rohde, for help in developing this paper based on the first author's MSc dissertation; Bhaavanaa Thumu, for help in extracting implicit discourse relations from the PDTB-3; and Lori Levin, for calling our attention to the work of Ward & Birner on non-canonical syntax.

References

Laura Aina, Xixian Liao, Gemma Boleda, and Matthijs Westera. 2021. Does referent predictability affect the choice of referential form? a computational approach using masked coreference resolution. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 454–469, Online. Association for Computational Linguistics.

Bernard M Bass, Wayne F Cascio, and Edward J O'Connor. 1974. Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology*, 59(3):313.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse processes*, 11(1):1–34.

Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. 2005. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering*, 55(1):59–83.

Debopam Das and Maite Taboada. 2018. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55:743–770.

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).

Forrest Davis and Marten van Schijndel. 2021. Uncovering constraint-based behavior in neural models via targeted fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1159–1171, Online. Association for Computational Linguistics.

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. 2023. *Probing the Predictions of Language Models about Discourse Coherence*. Master's thesis, University of Edinburgh.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Uri Hasson, Samuel A Nastase, and Ariel Goldstein. 2020. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434.

Tovah Irwin, Kyra Wilson, and Alec Marantz. 2023. BERT shows garden path effects. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3220–3232, Dubrovnik, Croatia. Association for Computational Linguistics.

Graeme D Kennedy. 1987. Expressing temporal frequency in academic english. *TESOL Quarterly*, 21(1):69–86.

Ekkehard König. 2002. *The meaning of focus particles: A comparative perspective*. Routledge.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974.

Sibel Özer, Murathan Kurfali, Deniz Zeyrek, Amália Mendes, and Giedre Valunaite Oleskeviciene. 2022. Linking discourse-level information and the induction of bilingual discourse connective lexicons. *Semantic Web*, 13(6):1081–1102.

Tiago Pimentel, Rowan Hall Maudslay, Damian Blasi, and Ryan Cotterell. 2020. Speakers fill lexical semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015, Online. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Katherine Forbes Riley, and Alan Lee. 2017. Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16, Saarbrücken, Germany. Association for Computational Linguistics.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12:1–18.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

Wilson L Taylor. 1953. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Gregory Ward and Betty Birner. 2006. Information structure and non-canonical syntax. In Laurence Horn and Gregory Ward, editors, *The handbook of pragmatics*, pages 152–174. Wiley Online Library.

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, 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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Deniz Zeyrek, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–27.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Data extraction

A.1 Identifying sentences with preposed NPs or PPs

Characterized syntactically, a preposed NP/PP is one that is the first child of the top level S node and a left sister of the matrix subject. This can be expressed with the Tregex pattern:

```
@PP|NP>2(S!>>/S.*/) &
$++(/NP-SBJ.*/>(S!>>/S.*/))
```

In the Penn TreeBank (PTB), function tags can be attached to syntactic labels. For example, “NP-SBJ” (for the NP subject of a clause), or “PP-LOC” (for locative prepositional phrases). The above Tregex pattern captures both bare syntactic labels and syntactic labels with function tags, through the use of an “@” symbol preceding the syntactic label. “S!»/S.*/” specifies S is the top-level S node that is not dominated by any other S node. “&” is an operator on relations that signals that two relations are satisfied simultaneously. “$++” stands for “a left sister of”. In the PTB corpus, “NP-SBJ” can have an index followed to indicate coindexation. Here we use regular expression to match any label that starts with “NP-SBJ”. Two node descriptions that identify PP/NPs separately can be combined with disjunction.

Finally, because attached to each tree nodes in the version of the Penn TreeBank we are using is an indication of the byte span in the raw text covered by the node, we need to indicate in the Tregex pattern that this byte span specification should be ignored in pattern matching. The "2" before "(S!»/S.*/)" refers to the second child of the top level S node because the byte span of the top level S node occupies the first node dominated by the S node, thus making the NP or PP a second child.

A.2 Selecting sentences that are Arg2 of inter-sentential implicit relations

After extracting all sentences in the corpus that start with a preposed NP/PP, the next step is to retain only those that start Arg2 of an inter-sentential implicit relation. We do this by mapping the start of the span list of the preposed NP/PP to Field 31 of the inter-sentential implicit relation tokens in the PDTB-3.[8]

This filters out tokens where the preposed constituent is in **Arg2** of an explicit relation, either the preposed constituent is followed by an explicit connective or what is preposed is a PP that itself serves as an explicit connective such as "as a result".

Since the selected tokens also include those with relations types “Hypophora”, “EntRel” and “NoRel”, we select only implicit tokens by choosing only tokens with “Implicit” in Field 0 which specifies relation type (Webber et al., 2019). Ex. (12) illustrates an implicit relation token in the PDTB-3.

(12)

```
Implicit|||||||as  a  result|Co
ntingency.Cause.Result||||||3042
..3142||||||3144..3222|||||||||||
3144|PDTB2::wsj_0003::3144::SAME|
```

[wsj_0003]

The majority of the relations in our data now are inter-sentential as we constrain the preposed constituent to appear at the start of both the top-level sentences and **Arg2**. Yet there is one exception, which brings in intra-sentential relations: prepositional clausal subordination with a prepositional phrase itself as **Arg2** preceding Arg1 (Prasad et al., 2018), as is shown in Ex. (13).

(13) **Without admitting or denying wrongdoing**, *they consented to findings of violations of escrow and record-keeping rules.* [wsj_0096]

To exclude such cases from our dataset, we check the relative position of *Arg1* and **Arg2** and remove those of which **Arg2** is to the left of *Arg1*. Only one case is detected and the total number of relations in our preposed set is 1441.

B Data preprocessing

B.1 Argument concatenation

The next step involves extracting the argument pair of the extracted relations. This is realized by consulting Field 20 and Field 14 of the relation token for the span list of **Arg2** and *Arg1* respectively, and extracting the text identified by the span list in the decorated PTB corpus.

Concatenating the two arguments to an implicit inter-sentential discourse relation is a very detailed procedure. It involves dealing with special cases with attribution, relative clauses, complementizer, missing punctuations, etc. For the details, readers are referred to Dong (2023).

[8] Field 31 specifies “the Conn SpanList of Explicit/AltLex/AltLexC tokens or the start point of the Arg2 of an Implicit/Hypophora/EntRel/NoRel tokens” (Webber et al., 2019).

B.2 Details of moving preposed phrases to canonical position

The method to create the canonical dataset is to map the preposed NP/PP in **Arg2** against the one that is extracted from the PTB corpus, re-extract it and then move it to its canonical position.

One issue that arises when constructing the canonical set is that sometimes the preposed constituent extracted from the PTB is not entirely the same as the preposed NP/PP in **Arg2**. It is illustrated in Ex. (14) where the relative clause in the preposed PP extracted from the PTB is not incorporated as a part of **Arg2** indicated in the PDTB-3. This is because the minimality principle applied in the annotation of the PDTB corpus requires only including in the annotation what is actually needed for the sense relation to be recognized. In this case, we take only the part of the preposed PP that is present in **Arg2** as the preposed constituent that will be right-moved.

(14) Arg2: In the so-called two-stroke engines; each piston goes up and down only once to provide power
Preposed PP: In the so-called two-stroke engines, which are expected to get sharply higher gas mileage [wsj_0956]